Variational methods for approximate reasoning in graphical models

Wim Wiegerinck, Bert Kappen, Martijn Leisink, David Barber,
Sybert Stroeve, Tom Heskes and Stan Gielen

RWCP, Theoretical and Algorithmic Foundation SNN,
University of Nijmegen,Nijmegen, The Netherlands
wimw@mbfys.kun.nl

Abstract

Exact inference in large and complex graphical models (e.g. Bayesian networks) is computationally intractable. Approximate schemes are therefore of great importance for real world computation. In this paper we consider a general scheme in which the original intractable graphical model is approximated by a model with a tractable structure. The approximating model is optimised by an iterative procedure, which minimises the Kullback-Leibler divergence between the two models. The procedure is guaranteed to converge to a local minimum of the Kullback-Leibler divergence. The scheme provides a bridge between naive mean-field theory and exact computation. Simulation results are provided to illustrate the method.

1. INTRODUCTION

Graphical models, such as Bayesian networks, Markov fields and Boltzmann machines provide a rich framework for probabilistic modelling and reasoning [1, 2, 3, 4]. 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. Undirected models, such as Markov fields and Boltzmann machines are useful for domains with correlated structures in which the causal direction is less obvious. The invention of algorithms for exact inference during the last decades has lead to the rapid increase in popularity of graphical models in modern AI. However, exact inference is NP-hard [5]. 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 [6].

In this paper, we address the problem of approximate inference in intractable graphical models. In this context, the variational methods gain increasingly interest [7, 8, 9, 10, 11, 12]. An advantage of these methods is that they provide bounds on the approximation error. This is in contrast to stochastic sampling methods [4, 9] 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 an intractable model using (simpler) approximating model that factorises according to a given structure. An iterative algorithm is presented to optimise the parameters of the approximating model 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. In section 4., we consider the approximation of extremely dense connected networks. For these networks, the optimisation of the approximating model 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. VARIATIONAL APPROXIMATION

2.1 Target models

Our starting point is a probabilistic model $P(x)$ on a set of discrete variables $x = x_1, \ldots, x_n$ in a finite domain, $x_i \in \{1, \ldots, n_i\}$. Our goal is to find its marginals $P(x_i)$ on single variables or small subsets of variables $P(x_i, \ldots, x_k)$. We assume that P can be written in the form

$$P(x) \;=\; \frac{1}{Z_p}\prod_\alpha \Psi_\alpha(d_\alpha) \;=\; \frac{1}{Z_p}\exp\sum_\alpha \psi_\alpha(d_\alpha) \quad (1)$$

in which Ψ_α are potential functions that depend on a small number of variables, denoted by the clusters d_α. Sometimes, we use the logarithmic form of the potentials, $\psi_\alpha = \log\Psi_\alpha$. Z_p is a normalisation factor that might be unknown. An example is a Boltzmann machine with binary units,

$$P(x) = \frac{1}{Z_p}\exp(\sum_{i<j} w_{ij}x_i x_j + \sum_k h_k x_k) \quad (2)$$

that fits in our form with $d_{ij} = (x_i, x_j), \quad i<j$, $d_k = x_k$ and potentials $\psi_{ij}(x_i,x_j) = w_{ij}x_i x_j$, $\psi_k(x_k) = h_k x_k$. Note that the potential representation is not unique. Another example of a model that fits in our framework is a Bayesian network given evidence e,

$$P_e(x) = P(x|e) = \frac{\prod_j P(x_j|\pi_j)}{P(e)} \quad (3)$$

which can be expressed in terms of the potentials $\Psi_j(d_j) = P(x_j|\pi_j)$, with $d_j = (x_j, \pi_j)$ and the normalisation $Z_p = P(e)$. This example shows that our inference problem includes the problem of computation of conditionals given evidence, since conditioning can be included by absorbing the evidence into the model definition via $P_e(x) = P(x,e)/P(e)$.

The computational complexity of computing marginals in P depends on the underlying graphical structure of the model, and is exponential in the maximal clique size of the triangulated moralised graph [2, 3, 4]. This may lead to intractable models, even if the clusters d_α are small. An example is a fully connected Boltzmann machine: the clusters contain at most two variables, while the model has one clique that contains all the variables in the model.

2.2 Approximating models

In the variational method [7, 9, 10, 11, 12], the intractable probability distribution $P(x)$ is approximated by a tractable distribution $Q(x)$. This distribution can be is used to compute the node probabilities $Q(x_i)$. In the standard (mean field) approach, Q is assumed to be completely factorised $Q(x) = \prod_i Q(x_i)$. We take the more general approach [10, 11], with Q being a tractable model that factorises according a given structure. By tractable we mean that marginals over small subsets of variables are computationally feasible.

To construct Q we first define its structure. In this paper, we consider two classes of factorisations for the approximating models. The first class are the 'undirected' factorisations,

$$Q(x) = \prod_\gamma \Phi_\gamma(c_\gamma) \quad (4)$$

in which c_γ are predefined clusters whose union contains all variables. $\Phi_\gamma(c_\gamma)$ are nonnegative potentials of the variables in the clusters. The only restriction on the potentials is the global normalisation

$$\sum_{\{x\}}\prod_\gamma \Phi_\gamma(c_\gamma) \;=\; 1\,. \quad (5)$$

The second class are the 'directed' factorisations. These can be written in the same form (4), but the clusters need to have an ordering $c_1, c_2, c_3, \ldots$. We define separator sets $s_\gamma = c_\gamma \cap \{c_1 \cup \ldots \cup c_{\gamma-1}\}$ and residual sets $r_\gamma = c_\gamma \backslash s_\gamma$. We restrict the potentials $\Phi_\gamma(c_\gamma) = \Phi_\gamma(r_\gamma, s_\gamma)$ to satisfy the local normalisation

$$\sum_{\{r_\gamma\}} \Phi_\gamma(r_\gamma, s_\gamma) \;=\; 1\,, \quad (6)$$

We can identify $\Phi_\gamma(r_\gamma, s_\gamma) = Q(r_\gamma|s_\gamma)$ and (4) can be written in the familiar directed notation

$$Q(x) = \textstyle\prod_\gamma Q(r_\gamma|s_\gamma)\,. \quad (7)$$

2.3 Variational optimisation

In the variational approach, the approximation Q is optimised such that the Kullback-Leibler (KL) divergence between Q and P,

$$D(Q,P) = \sum_{\{x\}} Q(x)\log\frac{Q(x)}{P(x)} \equiv \left\langle \log\frac{Q(x)}{P(x)}\right\rangle \quad (8)$$

is minimised. In this paper, $\langle\ldots\rangle$ denotes the average with respect to Q. The KL-divergence is related to the difference of the probabilities of Q and P,

$$\max_A |P(A) - Q(A)| \le \sqrt{\frac{1}{2}D(Q,P)} \quad (9)$$

for any event A in the sample space (see [13]). The KL-divergence satisfies $D(Q,P) \ge 0$, and

$D(Q,P) = 0 \Leftrightarrow Q = P$. Using the logarithmic potential representations of P and Q, with $\varphi_\gamma = \log \Phi_\gamma$, we can rewrite D,

$$D(Q,P) = \left\langle \sum_\gamma \varphi_\gamma(c_\gamma) - \sum_\alpha \psi_\alpha(d_\alpha) \right\rangle + \text{constant} \quad (10)$$

which shows that $D(Q,P)$ is tractable when Q is tractable and the clusters in P and Q are small.

To optimise Q under the normalisation constraints ((5) for undirected factorisations resp. (6) for directed factorisations), we do a constrained optimisation of the KL-divergence with respect to φ_γ using Lagrangian multipliers. In this optimisation, the other potentials φ_β, $\beta \neq \gamma$ remain fixed. This leads to the general solution $\varphi^*_\gamma(c_\gamma)$,

$$\varphi^*_\gamma(c_\gamma) = \left\langle \sum_{\alpha \in D_\gamma} \psi_\alpha(d_\alpha) - \sum_{\beta \in C_\gamma} \varphi_\beta(c_\beta) \right\rangle_{c_\gamma} - z \quad (11)$$

The average $\langle \ldots \rangle_{c_\gamma}$ is taken with respect to the conditional distribution $Q(x|c_\gamma)$. For undirected (resp. directed) approximations, D_γ is the set of clusters α in P that depend on c_γ (resp. r_γ). So for undirected approximations, $\alpha \notin D_\gamma$ implies $Q(d_\alpha|c_\gamma) = Q(d_\alpha)$, etc. Similarly, for undirected (resp. directed) approximations, C_γ is the set of clusters $\beta \neq \gamma$ that depend on c_γ (resp. r_γ). For undirected approximations, z is a constant that can be inferred from the normalisation (5), i.e.

$$z = \log \sum_{\{x\}} \exp \left[\sum_{\beta \neq \gamma} \varphi_\beta(c_\beta) + \right. \\ \left. + \left\langle \sum_{\alpha \in D_\gamma} \psi_\alpha(d_\alpha) - \sum_{\beta \in C_\gamma} \varphi_\beta(c_\beta) \right\rangle_{c_\gamma} \right] . \quad (12)$$

For directed approximations, z is a function of the separator s_γ, and can be inferred from (6), i.e.

$$z(s_\gamma) = \log \sum_{\{r_\gamma\}} \exp \left\langle \sum_{\alpha \in D_\gamma} \psi_\alpha(d_\alpha) \right. \\ \left. - \sum_{\beta \in C_\gamma} \varphi_\beta(c_\beta) \right\rangle_{c_\gamma} . \quad (13)$$

Since $Q(x|c_\gamma)$ is independent of the potential φ_γ, both expressions for z [(12) and (13)], are independent of φ_γ. Consequently, the right hand side of (11) is independent of φ_γ as well. So (11) provides a unique solution φ^*_γ to the optimisation of the potential of cluster γ. This solutions corresponds to the global minimum of $D(Q,P)$ given that the potentials of other clusters $\beta \neq \gamma$ are fixed. This means that in a sequence where at each step different potentials are selected and updated, the KL-divergence decreases at each step. Since $D(Q,P) \geq 0$, we conclude that this iteration over all clusters leads to a local minimum of $D(Q,P)$.

3. APPROXIMATING STRUCTURES

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. $Q(x) = \exp \sum_i \varphi(x_i)$. Then (11) reduces to the standard mean field equations [12]

$$\varphi^*_i(x_i) = \left\langle \sum_{\alpha \in D_i} \psi(d_\alpha) \right\rangle_{x_i} - z \quad (14)$$

The other extreme is to factorise Q according to a triangulated (moral) graph of P [2, 3, 4]. In this case, the approximating distribution Q converges to the target distribution P in finite time. Of course, this solution is only theoretically of interest, since the computational complexity of this approximation is equal to the complexity of the target distribution. However, it indicates that the variational approach using structure interpolates between the standard mean field approach and the exact solution.

In general one must choose a structure for Q that is a good compromise between approximation error and complexity. An important question is how to do this to get the best out of the approximation. In principle, the number of possible structures grows exponentially with the number of nodes. A heuristic is to try making the graphical overlap between Q and P as large as computationally feasible [10, 11]. One way to do this is to copy the target model P to Q, and subsequently, to split clusters in Q that cause too large cliques. On the other hand, some clusters in Q are already computationally harmless, and one could ask the question whether it is helpful to join these clusters into larger ones, assumed that this is computationally still efficient. The motivation is that these joint clusters may indirectly model relations in P that are not modelled in Q otherwise.

In some cases, one can infer from the structure of P and Q that it is useless to join two clusters

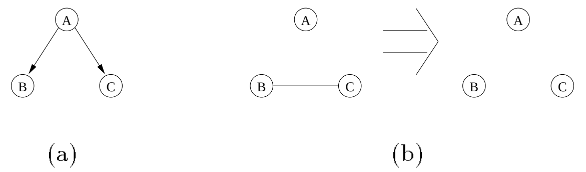


Figure 1: Example of redundant structure. (a): Graph of exact model $P(A)P(B|A)P(C|A)$. (b): Optimisation of an approximating model with structure $Q(A)Q(B,C)$ leads to a model with simpler structure $Q(A)Q(B)Q(C)$. The variables B and C become independent in Q, although they are marginally dependent in P (via A).

c_κ and c_λ into one joint cluster $c_{\kappa\cup\lambda} = c_\kappa \cup c_\lambda$. This is important information since the approximation with the joint cluster has more parameters to estimate and is more complex in computation. For example, in undirected approximations, joining clusters c_κ and c_λ is useless if for any of the remaining clusters $t \in \{d_\alpha, c_\beta, \beta \neq \kappa, \beta \neq \lambda\}$ at least one of the following independency relations

$$Q(t|c_\kappa \cup c_\lambda) = Q(t|c_\kappa) \quad (15)$$
$$Q(t|c_\kappa \cup c_\lambda) = Q(t|c_\lambda) \quad (16)$$

holds. In such a case one can show that an update of $\varphi_{\kappa\cup\lambda}$ in an approximation with the joint cluster leads to the same approximation as subsequent updates of φ_κ and φ_λ in an approximation with separate clusters. In directed approximations, similar results can be obtained. In fig. 1 a simple example is given.

In general, however, the optimal selection of the approximate structure is still an open problem.

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 $\langle \log P(x_i|\pi_i^p)\rangle$. This means that computational advantage can only be obtained if the parent size is much smaller than the clique size of P [2, 3, 4]. 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, 3] and weighted sigmoid functions [14]. For these parametrisations $\langle \log P(x_i|\pi_i^p)\rangle$ 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,

$$P(x_i = 1|\{x_k\}) = (1 + \exp(z_i))^{-1} \quad (17)$$

where z_i is the weighted input of the node, $z_i = \sum_k w_{ik}x_k + h_i$. In this case, the averaged log probability is intractable for large parent sets. To proceed we can use the approximation proposed in [7]

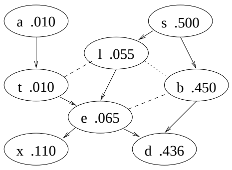
$$\langle \log(1 + e^{z_i})\rangle \leq$$
$$\xi_i \langle z_i \rangle + \log \left\langle e^{-\xi_i z_i} + e^{(1-\xi_i)z_i} \right\rangle \equiv \mathcal{E}_i(Q,\xi) \quad (18)$$

which is tractable if Q is tractable. Numerical optimisation of $\mathcal{L}(Q,\xi) \equiv \langle \log Q\rangle - \mathcal{E}(Q,\xi)$ 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.

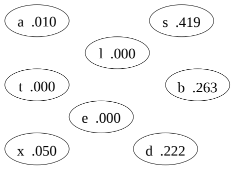
5. NUMERICAL RESULTS

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 [2] 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 $P(x_i)$ and $Q(x_i)$ is about 0.2).

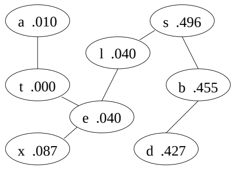
In the second toy problem we simulated approximate inference in a structure that has both tractable substructures and sigmoidal nodes with large parent sets. We generated models with graphical structure as in figure 3(a). The upper node is a mixture node with m mixture components. The next layer consists of $n + 1$ binary nodes. The third layer consists of n binary nodes x_i. Each of these nodes has two parents in the preceding layer. Up to this layer the network is



(a) Target model



(b) KL = 0.43

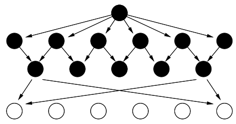


(c) KL = 0.03

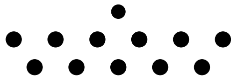
Figure 2: Chest clinic model (ASIA). (a): Exact model P with marginal probabilities. Dashed lines indicate its underlying cluster structure (moral graph). The dotted line indicates an extra fill needed to triangulate the graph. (b-c): Approximating models with approximated marginal probabilities. In (b) Q is fully factorised. In (c), Q is a tree. KL is the KL-divergence $D(Q, P)$ between the approximating model Q and the true model P.

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 x_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 (from top) given evidence on the observable nodes. Exact computation of these probabilities is intractable for large n.

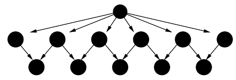
We choose $m = 10$, $n_v = 50$ and varied $n = 8, \ldots, 15$. Networks of this size are still tractable for exact computation. The values in the probability tables of $\mathcal{N}_1$ are drawn uniformly. The weights in the sigmoidal functions are drawn from the Gaussian distribution with zero mean and standard deviation $1/\sqrt{n}$. We computed exact and approximated marginals for the third layer x_i. As



(a)



(b)



(c)

Figure 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.

approximating models we used a factorised model and a model with the tractable structure $\mathcal{N}_1$ (fig. 3 (b-c)). In figure 4 we plotted the maximal error $\max_i |Q(x_i) - P(x_i|x_v)|$ 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 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 graphical models such as Bayesian networks is crucial if their application to large scale problems is to be realised. We have presented a general scheme to use a (simpler) approximating model that factorises according to a given structure. The scheme includes approximations with undirected, directed and chain graph models. The approximating model is tuned via minimisation of the Kullback-Leibler divergence. We have addressed the ques-

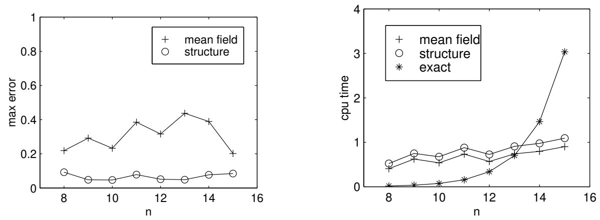


Figure 4: Left: The maximal error as a function of the network size. Right: CPU-time (arbitrary units) for exact and approximate inference as a function of the network size

tion of selecting the structure of the approximating model. Parametrised models with large parent sets can be dealt with by minimising an approximation of the KL divergence.

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. Approximations using structure give significant improvements. Our results seem to indicate that these improvements are 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.

References

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

[2] 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.

[3] F.V. Jensen. *An introduction to Bayesian networks.* UCL Press, 1996.

[4] E. Castillo, J. M. Gutierrez, and A. S. Hadi. *Expert Systems and Probabilistic Network Models.* Springer, 1997.

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

[6] 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.

[7] 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.

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

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

[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] 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, pages 183–189. MIT Press, 1999.

[12] M. Haft, R. Hofmann, and V. Tresp. Model-independent mean field theory as a local method for approximate propagation of information. *Network: Computation in Neural Systems*, 10:93–105, 1999.

[13] J. Whittaker. *Graphical models in applied multivariate statistics.* Wiley, Chichester, 1990.

[14] R. Neal. Connectionist learning of belief networks. *Artificial Intelligence*, 56:71–113, 1992.