Variational Methods for Inference and Estimation in Graphical Models

by

Tommi S. Jaakkola[1]

Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, MA 02139

Abstract

Graphical models enhance the representational power of probability models through qualitative characterization of their properties. This also leads to greater efficiency in terms of the computational algorithms that empower such representations. The increasing complexity of these models, however, quickly renders exact probabilistic calculations infeasible. We propose a principled framework for approximating graphical models based on variational methods.

We develop variational techniques from the perspective that unifies and expands their applicability to graphical models. These methods allow the (recursive) computation of upper and lower bounds on the quantities of interest. Such bounds yield considerably more information than mere approximations and provide an inherent error metric for tailoring the approximations individually to the cases considered. These desirable properties, concomitant to the variational methods, are unlikely to arise as a result of other deterministic or stochastic approximations.

The thesis consists of the development of this variational methodology for probabilistic inference, Bayesian estimation, and towards efficient diagnostic reasoning in the domain of internal medicine.

Thesis Supervisor: Michael I. Jordan
Title: Professor

[1] E-mail: `tommi@psyche.mit.edu`.

Contents

Chapter 1

Introduction

The presence of uncertainty in modeling and knowledge representation is ubiquitous, whether arising from external factors or emerging as a result of lack of model precision. Probability theory provides a widely accepted framework for representing and manipulating uncertainties. Much effort has been spent in recent years in developing probabilistic knowledge representations (Pearl 1988) and the computational techniques necessary for realizing their power (Pearl 1988, Lauritzen & Spiegelhalter 1988, Jensen et al. 1990, see also Shachter et al. 1994). This has lead to a separation of qualitative and quantitative features of probabilities, one represented via a graph and the other through tables or functions containing actual probability values. It is the qualitative graphical representation that holds the key to the success, both in terms of sparseness and utility of the probabilistic representations and the efficiency of computational operations on them. (For readers who will find the remaining introduction hard to follow in terms of its content on graphical models, we suggest an introductory book by Jensen, 1996 as a suitable background reading.)

In graphical representation of probabilities the nodes in the graph correspond to the variables in the probability model and the edges connecting the nodes signify dependencies. The power of such graph representation derives from interpreting separation properties in the graph as illustrations of independence properties constraining or structuring the underlying probability model. For this reason it is the missing edges in the graph that are important as they give rise to separation properties and consequently imply conditional independencies about the probability model.

Separation in graphical representations is defined relative to their topology, where the topology results from the types of nodes and edges and the connectivity of the graph in particular. While the nodes we consider are of the same type, the edges come in two varieties: directed or undirected. The directed edges (arrows connecting *parents* to their *children*) signify an asymmetric relation while the *adjacent* or *neighboring* nodes linked via an undirected edge assume a symmetric role. For undirected graphs containing only undirected edges (a.k.a. Markov random fields), the separation that postulates conditional independences in the underlying probability model corresponds to the ordinary graph separation[1]. For directed graphs possessing

[1] One set of nodes blocking all the "paths" between the other two sets

only directed edges or for chain graphs (see e.g. Whittaker 1990, Lauritzen 1996) containing both directed and undirected edges, the separation corresponds to a more involved d-separation criterion (Pearl 1988) or its variants (Lauritzen et al. 1990). We note that these separation criteria that allow us to code independence statements are mostly faithful in only one direction: all the separations in the graph hold as independencies in the underlying model[2] (Pearl 1988). The converse is rarely true. Graphs are therefore not perfect representations of independencies but nevertheless useful in characterizing the underlying probability distributions.

The utility of a graph is not limited to facilitating structured knowledge representation but it also holds a huge computational value. Any independencies in the probability model constitute licenses to ignore in manipulating the probabilities leading to greater efficiency. The ability of a graph to elucidate these independencies thus determines the feasibility of the computations on the associated probability model. All exact algorithms for probabilistic calculations on graphical models make an extensive use of the graph properties (Pearl 1988, Lauritzen & Spiegelhalter 1988, Jensen et al. 1990).

Any graph representation of the underlying probability model is not necessarily well-suited for computational purposes. It is typically the case that these graph representations do not remain invariant to ordinary manipulations of the underlying probability model such as marginalization or conditioning. The invariance property is highly desirable, however, and is particular to graphs known as decomposable graphs (see e.g. Whittaker 1990, Lauritzen 1996). Other graphs can be turned into decomposable ones through procedures consisting of prudently adding dependencies (edges). Decomposable graphs consist of cliques[3] that determine the complexity of ensuing calculations on the underlying model. The complexity grows exponentially in the size (the number of nodes or variables) of each clique. Decomposable graphs can be further written in a form of junction trees (hyper-graphs) that serve as the final platforms for exact probabilistic calculations (Jensen et al., 1990). The computations in junction trees proceed in a message passing fashion and are well described in an introductory book by Jensen (1996).

The framework for exact calculations is limited by the clique sizes[4] of the decomposable graphs that it relies on. Large clique sizes, on the other hand, may emerge as a result of the transformation into decomposable graphs in certain graph structures or as a result of dense connectivity in the original graphical representation (only few independencies are represented/available). In either case, the use of approximate techniques becomes necessary.

One avenue for approximation is to severe the graph by eliminating edges, i.e., forcing additional independencies on the underlying distribution (Kjaerulff 1994).

[2] Graphical representations satisfying this property are called I-maps of the underlying probability model

[3] Cliques are subsets of nodes where all nodes are connected to all others. The term is often used in the sense of maximal such sets, i.e., a clique couldn't be a proper subset of another clique.

[4] The probability model underlying a clique is a joint distribution over the variables (nodes) in the clique and contains all possible interactions among these variables. It is therefore exponentially costly (in the number of variables) to maintain or handle such a distribution.

This method reduces the clique sizes and can consequently benefit from the previously described exact algorithms that exploit the sparseness in the graphical representation. Another approach in the same spirit is to combine sampling techniques with exact methods (Kjaerulff 1995). These approximation methods are perhaps more suitable in settings where the large clique sizes arise later, when they are transformed into decomposable graphs.

Graphical models that are inherently densely connected not only lead to large clique sizes limiting the applicability of exact methods but they also involve a large number of parameters that need to be assessed or estimated. The latter is often addressed by resorting to compact forms of dependencies among the variables as is the case with noisy-OR networks (Pearl 1988) and models more often studied in the neural networks literature such as Boltzmann machines (see e.g. Hertz et al. 1991) or Sigmoid belief networks (Neal 1992). Such compact dependency structure, however, is not exploited in the exact algorithms nor in the approximate methods mentioned earlier. These models nevertheless have important applications, most prominently in the form of the QMR-DT belief network formulated for diagnostic reasoning in the domain of internal medicine (Shwe et al. 1991, chapter 5).

We develop approximation methods for graphical models with several goals in mind. First, we require these methods to be principled and controlled. In other words, they should be reasonably general and it should be possible to assess their reliability. Ideally, we look for upper and lower bounds on the quantities of interest since the approximation accuracy in this case is readily measured as the width of the interval separating the bounds. The approximation techniques should also be integrable with exact algorithms to the extent that such algorithms are/can be made applicable in large graphical models. Finally, the methods should exploit the compactness of the dependency structure that appears in many densely connected models. The approximation techniques we develop to meet these goals are known generally as variational methods.

Variational methods have long been used as approximate techniques in the physics literature, particularly in mechanics in the form of calculus of variations. Calculus of variations pertains to extrema of integral functionals, where the solution is obtained through fixed point equations (the Euler-Lagrange equations). In quantum physics bounds on energies can be achieved through formulating a variational optimization problem (Sakurai 1985). The computation of the system free energy can also be cast as a variational problem and this formulation contains the wide-spread special case known as the mean field approximation (Parisi, 1988). We note that modern use of variational methods is not restricted to physics and these techniques have been found useful and developed also in statistics (Rustagi, 1976) among other fields.

In the context of graphical models, it is the variational formulation of free energy that has received the most attention. This fact derives from the direct analogy between free-energy (or log-partition function) in physics and the log-likelihood of data in a probabilistic setting. From another but equivalent perspective, Peterson & Anderson (1987) formulated mean field theory for Boltzmann machines in order to speed up parameter estimation over sampling techniques used previously in these models. Ghahramani (1995) employed mean field approximation to facilitate the E-step of the

EM-algorithm in density models with factorial latent structure. Dayan et al. (1995) and Hinton et al. (1995) developed learning algorithms for layered bi-directional density models called the Helmholtz machines. The algorithms are explicitly derived from the variational free energy formulation. Subsequently, Saul et al. (1996) calculated rigorous (one-sided) mean field bounds for these and other models in the class of sigmoid belief networks. Jaakkola et al. (1996) derived related results for the class of cumulative Gaussian networks. The more recent work on variational techniques in the context of graphical models will be introduced in the following chapters when relevant.

The work in this thesis considers variational methods from another perspective, one that expands and unifies their development for graphical models. The variational techniques we consider involve the use of dual representations (transformations) of convex/concave functions. These representations will be introduced in detail in the following section. Let us motivate these methods here in comparison to the goals set earlier for approximate methods replacing exact calculations in graphical models. It is the nature of the dual representations that they cast the function in question as an optimization problem over simpler functions. Such representations can be easily turned into approximations by simply relaxing the optimizations involved. This relaxation has two main consequences: (i) it simplifies the problem and (ii) it yields a bound, both of which are desirable properties. We use the dual representations in place of the original functions in probabilistic calculations and turn them into approximations or bounds in this manner. The ensuing bound on the target quantity has two important roles. First, it naturally provides more information about the desired quantity than a mere approximation. Second, it allows the approximation to be further tuned or readjusted so as to make the bound tighter. For example, if the approximation yields a lower bound, then by maximizing the lower bound necessarily improves the approximation. We may also obtain complementary bounds and consequently get interval estimates of the quantities of interest. Similarly to one-sided bounds, the intervals can be refined or reduced by minimizing/maximizing the complementary bounds. We note that the ability of variational methods to provide bounds in addition to their inherent "error metric" makes them quite unlike other deterministic or stochastic approximation techniques.

Graphical models, rife with convexity properties, are well-suited as targets for the abovementioned approximations. We may, for example, use the transformations to simplify conditional probabilities in the joint distribution underlying the graph representation. In the presence of compact representations of dependencies, these transformations can uncover the dependency structure in the conditional probabilities and lead to simplification not predictable from the graph structure alone. A particularly convenient form of simplification is factorization as it reduces dependent problems into a product of independent ones that can be handled separately. We may continue substituting the variational representations for the conditional probabilities, while consistently maintaining either upper or lower bound on the original distribution (to preserve the error metric). When the remaining (transformed) distribution becomes feasible it can be simply handed over to exact algorithms. We can equally well use these transformations to simplify marginalization operations with analogous effects.

We now turn to the more technical development of the ideas touched in this introduction and provide first a tutorial on dual representations of convex functions. These representations will be used frequently in the thesis. The overview of the chapters will follow.

1.1 Convex duality and variational bounds

We derive here the dual or conjugate representations for convex functions. The presentation is tutorial and is intended to provide a basic understanding of the foundation for the methods employed in later chapters. The Appendix 1.A can be skipped without a loss of continuity; it is included for completeness of the more technical content appearing later in the thesis. We refer the reader to e.g. Rockafellar (1970) for a more rigorous and extensive treatment concerning convex duality (excluding the material in the Appendix).

Let $f(x)$ be a real valued and convex (i.e. convex up) function defined in some convex set X (for example, $X = R^n$). For simplicity, we assume that f is a well-behaving (differentiable) function. Consider the graph of f, i.e., the points $(x, f(x))$ in an $n+1$ dimensional space. The fact that the function f is convex translates into convexity of the set $\{(x,y) : y \geq f(x)\}$ called the epigraph of f and denoted by $epi(f)$ (see figure 1-1). Now, it is an elementary property of convex sets that they can be represented as the intersection of all the half-spaces that contain them (see figure 1-1). Through parameterizing these half-spaces we obtain the dual representations of convex functions. To this end, we define a half-space by the condition:

$$\text{all } (x,y) \text{ such that } x^T\xi - y - \mu \leq 0 \tag{1.1}$$

where ξ and μ parameterize all (non-vertical) half-spaces. We are interested in characterizing the half-spaces that contain the epigraph of f. We require therefore that the points in the epigraph must satisfy the half-space condition: for $(x,y) \in epi(f)$, we must have $x^T\xi - y - \mu \leq 0$. This holds whenever $x^T\xi - f(x) - \mu \leq 0$ as the points in the epigraph have the property that $y \geq f(x)$. Since the condition must be satisfied by all $x \in X$, it follows that

$$\max_{x\in X}\{\, x^T\xi - f(x) - \mu \,\} \leq 0, \tag{1.2}$$

as well. Equivalently,

$$\mu \;\; \geq \;\; \max_{x\in X}\{\, x^T\xi - f(x) \,\} \quad \equiv \quad f^*(\xi) \tag{1.3}$$

where $f^*(\xi)$ is now the dual or conjugate function of f. The conjugate function, which is also a convex function, defines the critical half-spaces (those that are needed) for the intersection representation of $epi(f)$ (see figure 1-1). To clarify the duality, let us drop the maximum and rewrite the inequality as

$$x^T \xi \le f(x) + f^*(\xi) \tag{1.4}$$

The roles of the two functions are interchangeable and we may suspect that also

$$f(x) = \max_{\xi \in \Xi} \{\, x^T \xi - f^*(\xi) \,\} \tag{1.5}$$

which is indeed the case. This equality states that the dual of the dual gives back the original function.

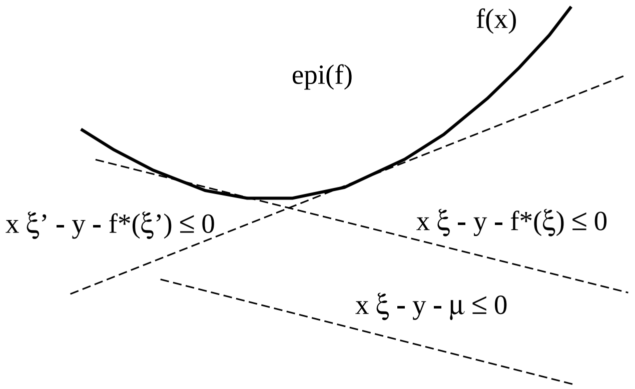

Figure 1-1: Half-spaces containing the convex set *epi*(f). The conjugate function $f^*(\xi)$ defines the critical half-spaces whose intersection is *epi*(f), or, equivalently, it defines the tangent planes of $f(x)$.

Let us now consider the nature of the space Ξ over which the conjugate function f^* is defined. The point x that attains the maximum in Eq. (1.3) for a fixed ξ is obtained through

$$\nabla_x \{\, x^T \xi - f(x) \,\} = \xi - \nabla_x f(x) = 0 \tag{1.6}$$

Thus for each ξ in Ξ there is a point x such that $\xi = \nabla_x f(x)$ (this point is unique for strictly convex functions). The conjugate space is therefore the gradient space of f. If ξ represents gradients of f, why not write this explicitly? To do this, substitute $\nabla_{x'} f(x')$ for ξ in Eq. (1.5) and find

$$f(x) \;=\; \max_{x' \in X} \{\, x^T \, \nabla_{x'} f(x') - f^*(\, \nabla_{x'} f(x') \,) \,\} \tag{1.7}$$

We can simplify the form of the conjugate function $f^*(\, \nabla_{x'} f(x') \,)$ in this representation by using its definition from Eq. (1.3). For $\xi = \nabla_{x'} f(x')$, the point x that attains the maximum in Eq. (1.3) must be x': we have already seen that the maximizing point must satisfy $\xi - \nabla_x f(x) = 0$, which implies that $\nabla_{x'} f(x') = \nabla_x f(x)$. For strictly convex functions this gives $x = x'$. By setting $x = x'$ we get

$$x'^T \, \nabla_{x'} f(x') - f(x') = f^*(\, \nabla_{x'} f(x') \,) \tag{1.8}$$

giving a more explicit form for the conjugate function. Putting this form back into the representation for f yields

$$f(x) \quad = \quad \max_{x' \in X} \left\{ \, (x - x')^T \, \nabla_{x'} f(x') + f(x') \, \right\} \tag{1.9}$$

which is a maximum over tangent planes for f. Importantly, this means that each tangent plane is a lower bound on f. The maximum in the above representation is naturally attained for $x' = x$ (the tangent plane defined at that point). We note that while the brief derivation assumed strict convexity, the obtained result (Eq. (1.9)) nevertheless holds for any (well-behaving) convex function. Whether this explicit tangent representation is more convenient than the one in Eq. (1.5) depends on the context.

Finally, to lower bound a convex function $f(x)$ we simply drop the maximum in Eq. (1.5) or Eq. (1.9) and get

$$f(x) \geq x^T \, \xi - f^*(\xi) \equiv f_\xi(x) \tag{1.10}$$

where the affine family of bounds $f_\xi(x)$ parameterized by ξ consists of tangent planes of f. The parameter ξ is referred to as the variational parameter. There is an inherent error metric associated with these bounds, one that can be defined solely in terms of the variational parameter. This is discussed in Appendix 1.A with examples. The fact that the bounds are affine regardless of the non-linearities in the function f (as long as it remains convex), provides a considerable reduction in complexity. We will make a frequent use of this simplifying aspect of these variational bounds in the chapters to follow.

We note as a final remark that for concave (convex down) functions the results are exactly analogous; just replace max with min, and lower bounds with upper bounds.

1.2 Overview

This thesis consist of the derivation of principled approximation methods for graphical models basing them on variational representations obtained from convex duality. The chapters contain a development of this methodology from inference (chapters 2 and 3) to estimation (chapter 4), and finally to an application in medical diagnosis (chapter 5).

The purpose of chapter 2 is to develop methods for computing upper and lower bounds on probabilities in graphical models. The emphasis is on a particular class of networks we call *log-concave* models containing, for example, sigmoid belief networks and noisy-OR networks as special cases. The emphasis is not in full generality but in demonstrating the use of dual representations in transforming conditional probabilities into computationally usable forms. The chapter is based on Jaakkola & Jordan (1996) "Computing upper and lower bounds in intractable networks".

In chapter 3 we formulate a recursive approximation framework through which upper and lower bounds are obtained without imposing restrictions on the model topology. Directed, undirected, and chain graphs are considered. The results in this chapter are (partly complementary) generalizations of those found in chapter 2. The chapter is extended from Jaakkola & Jordan (1996) "Recursive algorithms for approximating probabilities in graphical models".

Chapter 4 constitutes a shift of emphasis from inference to estimation. In this chapter we exemplify the use of variational techniques in the context of Bayesian parameter estimation. We start from a simple Bayesian regression problem and extend the obtained results to more general graphical models. Of particular interest is the difficult problem of estimating graphical models from incomplete cases. The chapter is expanded from Jaakkola & Jordan (1996) "A variational approach to Bayesian logistic regression problems and their extensions".

In chapter 5 we return to inference and consider the application of variational methods towards efficient diagnostic reasoning in internal medicine. The probabilistic framework is given by the QMR-DT belief network, which is a densely connected graphical model embodying extensive statistical and expert knowledge about the domain. The variational techniques we apply and extend in this setting are those described in chapter 2 for noisy-OR networks.

Guide for the reader

The basic content of the thesis can be understood by reading the introduction along with chapters 2 and 3, in that order. If any interest remains or to initially gain some, we recommend reading chapter 5. For the reader interested in Bayesian estimation at the expense of the approximation methodology for inference, we note that chapter 4 can be read without necessarily going through chapters 2 and 3 first. If, on the other hand, the reader only wishes to explore the application side of the variational methods in this thesis, we suggest reading chapter 5 as a stand alone article; the previous chapters can be consulted later for a more comprehensive understanding of variational methods.

References

T. Cover and J. Thomas (1991). *Elements of information theory.* John Wiley & Sons, Inc.

P. Dayan, G. Hinton, R. Neal, and R. Zemel (1995). The helmholtz machine. *Neural Computation.* **7**:889–904.

A. Dempster, N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B* **39**:1-38.

Z. Ghahramani (1995). Factorial learning and the EM algorithm. In *Advances of Neural Information Processing Systems 7.* MIT press.

J. Hertz, A. Krogh and R. Palmer (1991). *Introduction to the theory of neural computation.* Addison-Wesley.

G. Hinton, P. Dayan, B. Frey, and R. Neal (1995). The wake-sleep algorithm for unsupervised neural networks. *Science* **268**: 1158–1161.

T. Jaakkola, L. Saul, and M. Jordan (1996). Fast learning by bounding likelihoods in sigmoid-type belief networks. In *Advances of Neural Information Processing Systems 8.* MIT Press.

T. Jaakkola and M. Jordan (1996). Computing upper and lower bounds on likelihoods in intractable networks. In *Proceedings of the twelfth Conference on Uncertainty in Artificial Intelligence.*

T. Jaakkola and M. Jordan (1996). Recursive algorithms for approximating probabilities in graphical models. In *Advances of Neural Information Processing Systems 9.*

T. Jaakkola and M. Jordan (1996). A variational approach to Bayesian logistic regression problems and their extensions. In *Proceedings of the sixth international workshop on artificial intelligence and statistics.*

F. Jensen (1996). *Introduction to Bayesian networks.* Springer.

F. Jensen, S. Lauritzen, and K. Olesen (1990). Bayesian updating in causal probabilistic networks by local computations. *Computational Statistics Quarterly* **4**: 269-282.

U. Kjaerulff (1994). Reduction of Computational Complexity in Bayesian Networks through Removal of Weak Dependences. In *Proceedings of the Tenth Annual Conference on Uncertainty in Artificial Intelligence.*

U. Kjaerulff (1995. HUGS: Combining Exact Inference and Gibbs Sampling in Junction Trees. In *Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence.*

S. Lauritzen (1996). *Graphical Models.* Oxford University Press.

S. Lauritzen and D. Spiegelhalter (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society B* **50**:154-227.

R. Neal. Connectionist learning of belief networks (1992). *Artificial Intelligence* **56**: 71-113.

R. Neal and G. Hinton. A new view of the EM algorithm that justifies incremental and other variants. University of Toronto technical report.

J. Pearl (1988). *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann.

C. Peterson and J. R. Anderson (1987). A mean field theory learning algorithm for neural networks. *Complex Systems* **1**: 995–1019.

R. Rockafellar (1970). *Convex Analysis.* Princeton Univ. Press.

J. Rustagi (1976). *Variational Methods in Statistics.* Academic Press.

L. Saul, T. Jaakkola, and M. Jordan (1996). Mean field theory for sigmoid belief networks. *Journal of Artificial Intelligence research* **4**: 61-76.

R Shachter, S. Andersen and P. Szolovits (1994). Global Conditioning for Probabilistic Inference in Belief Networks. In *Proceedings of the Tenth Annual Conference on Uncertainty in Artificial Intelligence.*

M. Shwe, B. Middleton, D. Heckerman, M. Henrion, E. Horvitz. H. Lehmann, G. Cooper (1991). Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base: Part-I. *Methods of Information in Medicine* **30**: 241-255.

J. Whittaker (1990). *Graphical models in applied multivariate statistics.* John Wiley & Sons.

1.A Bounds and induced metrics

As we have seen the affine family of bounds $f_\xi(x) = x^T\,\xi - f^*(\xi)$ parameterized by ξ consists of tangent planes of f. We would like to characterize the accuracy of these bounds in terms of the variational parameter that defines them. We can, in fact, define a metric in the conjugate space $\xi \in \Xi$ that specifies the accuracy of the bound $f_\xi(x)$ in the original space X. To this end, we note that for strictly convex functions each $\xi' \in \Xi$ has a unique point $x' \in X$ such that $f(x') = f_{\xi'}(x')$; x' and ξ' will be called the corresponding points. Put another way, ξ' is the point in the conjugate space that attains the maximum in $f(x') = \max_\xi f_\xi(x')$. We define the distance from ξ to ξ' to be the difference $f(x') - f_\xi(x')$, which measures the approximation error in the bound if at x' a suboptimal parameter ξ is used instead of the optimal ξ'. Thus (see figure 1-2)

$$\begin{aligned} D_{f^*}(\xi \to \xi') &= f_{\xi'}(x') - f_\xi(x') && (1.11)\\ &= x'^T(\xi' - \xi) + f^*(\xi) - f^*(\xi') && (1.12)\\ &= (\nabla f^*)(\xi')^T\,(\xi' - \xi) + f^*(\xi) - f^*(\xi') && (1.13) \end{aligned}$$

where we have used the fact that $x' = \nabla_{\xi'} f^*(\xi')$ (see convex duality).

We now consider a few properties of this metric definition. Locally, the metric reduces to a quadratic distance metric relative to the Hessian of the conjugate function f^*, which is positive definite[5]. Globally, however, the definition can be asymmetric but nevertheless satisfies the uniqueness and positivity conditions. To define a similar metric in the original space we would start from Eq. (1.3) and proceed analogously. Consequently, we have two metrics, $D_{f^*}(\xi \to \xi')$ and $D_f(x \to x')$, induced in their respective spaces. How are these metrics related? If we let (x, ξ) and (x', ξ') be two

[5] For strictly convex functions

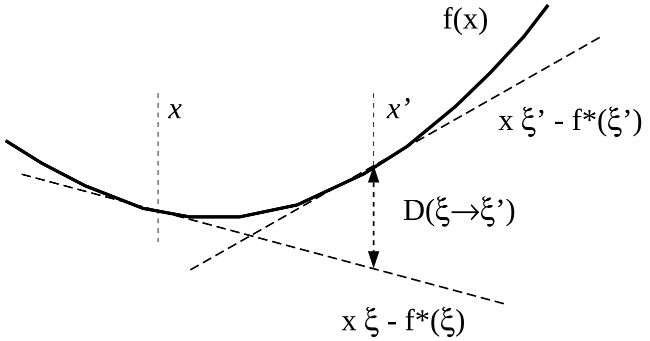

Figure 1-2: The definition of the dual (semi)metric.

pairs of corresponding points, then we have the equality

$$D_f(x \to x') = D_{f*}(\xi' \to \xi) \tag{1.14}$$

This indicates that the induced (directional) distance in the original space is equal to the induced distance in the dual space but in the reverse direction. More generally, the length of a path in the original space is equal to the length of the reverse path in the dual space. The definition of D above thus generates a meaningful dual metric.

Let us consider a few examples. We start from a simple one where $f(x) = x^2/2$, which is a convex function. The conjugate of f can be found from Eq. (1.3)

$$f^*(\xi) = \max_x \{ \, x\,\xi - \frac{1}{2}x^2 \, \} = \frac{1}{2}\xi^2 \tag{1.15}$$

and it has the same quadratic form. According to Eq. (1.13), this representation induces a metric in the original x space (note that unlike before we are now considering the bounds on the conjugate rather than the original function). This metric is

$$\begin{aligned} D_f(x \to x') &= (\nabla f)^T(x')\,(x'-x) + f(x) - f(x') & (1.16) \\ &= x'(x'-x) + \frac{1}{2}x^2 - \frac{1}{2}x'^2 = \frac{1}{2}(x'-x)^2 & (1.17) \end{aligned}$$

which is the usual Euclidean metric. This result can be generalized to quadratic forms defined by a positive definite matrix.

As another example, let us consider the negative entropy function $f(p) = -H(p)$, which is convex in the space of probability distributions (see e.g., Cover & Thomas 1991). The conjugate to this function is given by

$$\begin{aligned} f^*(\phi) &= \max_p \{ \, p^T \, \phi - (-H(p)) \, \} & (1.18) \\ &= \max_p \{ \, \sum_i p_i \phi_i - \sum_i p_i \log p_i \, \} & (1.19) \\ &= \max_p \{ \, \sum_i p_i \log \frac{e^{\,\phi_i}}{p_i} \, \} \; = \; \log \sum_i e^{\,\phi_i} & (1.20) \end{aligned}$$

where we have carried out the maximization analytically; the maximum is attained for $p_i \propto e^{\phi_i}$ which is the Boltzmann distribution. We note that the conjugate function – the log partition function – has to be a convex function of ϕ (the conjugates of convex functions are convex). We are now able to ask what the induced metric is in the p space. Using the definition above we obtain

$$\begin{aligned} D_f(p \to p') &= (\nabla f)^T(p')\,(p' - p) + f(p) - f(p') && (1.21) \\ &= \sum_i (p'_i - p_i) \log p'_i - H(p) + H(p') && (1.22) \\ &= \sum_i p_i \log \frac{p_i}{p'_i} \;\;=\;\; KL(p, p') && (1.23) \end{aligned}$$

which is the KL-divergence between the distributions p and p'. We note that log-partition functions will appear frequently in the thesis and thus whenever the representation Eq. (1.18) is used to lower bound them, the associated metric will be the KL-distance. The associated metrics for all other transformations appearing in the thesis can be obtained analogously.

Chapter 2

Upper and lower bounds[1]

2.1 Introduction

While Monte Carlo methods approximate probabilities in a stochastic sense, we develop deterministic methods that yield strict lower and upper bounds for marginal probabilities in graphical models. These bounds together yield interval bounds on the desired probabilities. Although the problem of finding such intervals to predescribed accuracy is NP-hard in general (Dagum and Luby 1993), bounds that can be computed efficiently may nevertheless yield intervals that are accurate enough to be useful in practice.

Previous work on interval approximations include Draper and Hanks (1994) (see also Draper 1995) where they extended Pearl's polytree algorithm (Pearl 1988) to propagate interval estimates instead of exact probabilities in polytrees. Their method, however, relies on the polytree property for efficiency and does not (currently) generalize feasibly to densely connected networks. We focus in this work particularly on network models with dense connectivity.

Dense connectivity leads not only to long execution times but may also involve exponentially many parameters that must be assessed or learned. The latter issue is generally addressed via parsimonious representations such as the logistic sigmoid (Neal 1992) or the noisy-OR function (Pearl 1988). In order to retain the compactness of these representations throughout our inference and estimation algorithms, we dispense with the moralization and triangulation procedures of the exact (clustering) framework and operate on the given graphical model directly.

We develop our interval approximation techniques for a class of networks that we call *log-concave* models. These models are characterized by their convexity properties and include sigmoid belief networks and noisy-OR networks as special cases. Previous work on sigmoid belief networks (Saul et al. 1996) provided a rigorous *lower* bound; we complete the picture here for these models by deriving the missing *upper* bound. We also present upper and lower bounds for the more general class of log-concave

[1] The basic content of this chapter has previously appeared in "T. Jaakkola and M. Jordan (1996). Computing upper and lower bounds on likelihoods in intractable networks. In *Proceedings of the twelfth Conference on Uncertainty in Artificial Intelligence*".

models focusing on noisy-OR networks in particular. While the lower bounds we obtain are applicable to generic network structures, the upper bounds we derive here are restricted to two-level (or bipartite) networks. The extension of the upper bounds to more general networks will be addressed in chapter 3. There are nevertheless many potential applications for the restricted bounds, including the probabilistic reformulation of the QMR knowledge base (Shwe et al. 1991) that will be considered in detail in chapter 5. The emphasis of this chapter is on techniques of bounding and their analysis rather than on all–encompassing inference algorithms. Merging the bounding techniques with exact calculations can yield a considerable advantage, as will be evident in chapter 5 (see chapter 3 for more general methodology).

The current chapter is structured as follows. Section 2.2 introduces the log-concave models and develops the techniques for upper and lower bounds. Section 2.3 contains an application of these techniques to sigmoid networks and gives a preliminary numerical analysis of the accuracy of the obtained bounds. Section 2.4 is devoted analogously for noisy-OR networks. We then summarize the results and describe some future work.

2.2 Log-concave models

We consider here a class of acyclic probabilistic networks defined over binary (0/1) variables $S_1, \ldots, S_n$. The joint probability distribution over these variables has the usual decompositional structure:

$$P(S_1, \ldots, S_n|\theta) = \prod_i P(S_i|S_{pa_i}, \theta_i) \tag{2.1}$$

where S_{pa_i} is the set of parents of S_i. The conditional probabilities, however, are assumed to have the following restricted form

$$P(S_i|S_{pa_i}, \theta) = G_{S_i}\left(\theta_{i0} + \sum_{j \in \mathrm{pa}_i} \theta_{ij} S_j\right) \tag{2.2}$$

where $\log G_0(x)$ and $\log G_1(x)$ are both *concave* functions of x. The parameters specifying these conditional probabilities are the real valued "weights" θ_{ij}. We refer to networks conforming to these constraints as *log-concave* models. While this class of models is restricted, it nevertheless contains sigmoid belief networks (Neal, 1992) and noisy-OR networks (Pearl, 1988) as special cases. The particulars of both sigmoid and noisy-OR networks will be considered in later sections.

In the remainder of this section we present techniques for computing upper and lower bounds on marginal probabilities in log-concave networks. We note that any successful instantiation of evidence in these networks relies on the ability to estimate such marginals. The upper bounds that we derive are restricted to two-level (bipartite) networks while the lower bounds are valid for arbitrary network structures.

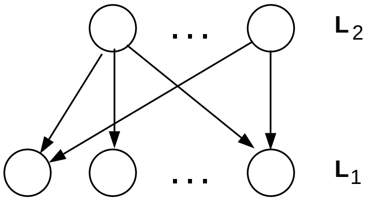

Figure 2-1: Two level (bipartite) network.

2.2.1 Upper bound for log-concave networks

We restrict our attention here to two-level directed architectures. For clarity we divide the variables in these two levels as observable (level 1 or L_1) and latent variables (level 2 or L_2). See figure 2-1 for an example. The bipartite structure of these networks implies (i) that the latent variables are marginally independent, and (ii) the observable variables are conditionally independent. The joint distribution for log-concave models with this special architecture is given by

$$P(S_1, \ldots, S_n|\theta) = \prod_{i \in L_1} G_{S_i}\left(\theta_{i0} + \sum_{j \in pa_i} \theta_{ij} S_j\right) \prod_{j \in L_2} P(S_j|\theta_j) \tag{2.3}$$

where L_1 and L_2 signify the two layers (observable and latent). Note that the connectivity is from L_2 (latent variables) to L_1 (observables).

To compute the marginal probability of a set of variables in these networks we note that (i) any latent variable included in this desired marginal set only reduces the complexity of the calculations, and (ii) the form of the architecture makes those observable variables (in L_1) that are excluded from the desired marginal set inconsequential. We will thus adopt a simplifying notation in which the marginal set consists of all and only the observable variables (i.e., those in L_1). The goal is therefore to compute

$$P(\{S_i\}_{i \in L_1}|\theta) = \sum_{\{S_j\}_{j \in L_2}} P(S_1, \ldots, S_n|\theta) \tag{2.4}$$

Given our assumption that computing such marginal probability is intractable, we seek an upper bound instead. The goal is to simplify the joint distribution such that the marginalization across L_2 or latent variables can be accomplished efficiently, while maintaining at all times a rigorous upper bound on the desired marginal probability. We develop variational transformations for this purpose. The transformations we consider come from convex duality as explained previously in section 1.1.

To use such variational bounds in the context of probabilistic calculations, we recall the concavity property of our conditional probabilities (on a log-scale). According to convex duality, we can find a variational bound of the form

$$\log G_{S_i}\left(\theta_{i0} + \sum_{j \in pa_i} \theta_{ij} S_j\right) \le \xi_i \left(\theta_{i0} + \sum_{j \in pa_i} \theta_{ij} S_j\right) - f^*(\xi_i) \tag{2.5}$$

where the definition of the conjugate function f^* is conditional on the value of S_i. We can recover the original log-conditional probabilities by minimizing the bound with respect to ξ. The setting of the variational parameter involves a trade-off between accuracy and simplicity. We choose to fix ξ thereby gaining feasibility but possibly loosing in accuracy. Now, by exponentiating both sides of the above inequality, we obtain a bound on the conditional probabilities:

$$G_{S_i}\left(\theta_{i0}+\sum_{j\in pa_i}\theta_{ij}S_j\right) \leq e^{\xi_i\left(\theta_{i0}+\sum_{j\in pa_i}\theta_{ij}S_j\right)-f^*(\xi_i)} \tag{2.6}$$

$$= e^{\xi_i\theta_{i0}-f^*(\xi_i)}\prod_{j\in pa_i}\left[e^{\xi_i\theta_{ij}}\right]^{S_j} \tag{2.7}$$

The important property of this bound is that it factorizes over the latent variables S_j in L_2 associated with the particular observable S_i. We have explicated this factorization by rewriting the bound in a product form. Since products of factorized forms remain factorized, it follows that if we replace each of the conditional probabilities in the joint distribution (see Eq. (2.3)) with the factorized upper bounds, then the resulting upper bound on the joint remains factorized as well. Computing marginal probabilities from such factorized distributions is very efficient (scaling linearly in the number of variables to be marginalized over). To develop this more quantitatively, consider the joint distribution with the variational bounds substituted for the conditional probabilities:

$$P(S_1,\ldots,S_n|\theta) \leq \left[\prod_{i\in L_1} e^{\xi_i\left(\theta_{i0}+\sum_{j\in pa_i}\theta_{ij}S_j\right)-f^*(\xi_i)}\right]\prod_{j\in L_2}P(S_j|\theta_j) \tag{2.8}$$

$$= \left[\prod_{i\in L_1} e^{\xi_i\theta_{i0}-f^*(\xi_i)}\right]\times\left[\prod_{i\in L_1} e^{\xi_i\sum_{j\in pa_i}\theta_{ij}S_j}\right]\prod_{j\in L_2}P(S_j|\theta_j) \tag{2.9}$$

$$= \left[\prod_{i\in L_1} e^{\xi_i\theta_{i0}-f^*(\xi_i)}\right]\times\prod_{j\in L_2}\left[e^{\sum_{i\,:\,j\in pa_i}\xi_i\theta_{ij}}\right]^{S_j}P(S_j|\theta_j) \tag{2.10}$$

where the first equality follows simply from pulling out the terms that do not depend on the latent variables S_j. To get the last equality, we have replaced the product over $i \in L_1$ as a sum in the exponent while conversely turning the sum over the latent variables into a product; the resulting product was combined with the product over the prior probabilities. The notation $i : j \in pa_i$ stands for "those i that have j as a parent" (i.e. the "children" of j). As a result, we have a factorized bound on the joint distribution that can be marginalized efficiently over any set of latent variables. We therefore obtain a feasible bound on the desired marginal

$$P(\{S_i\}_{i\in L_1}|\theta) = \sum_{\{S_j\}_{j\in L_2}} P(S_1,\ldots,S_n|\theta) \tag{2.11}$$

$$\leq \left[\prod_{i \in L_1} e^{\,\xi_i \theta_{i0} - f^*(\xi_i)}\right] \times \prod_{j \in L_2} \left[e^{\sum_{i\,:\,j \in pa_i} \xi_i \theta_{ij}} P_j + 1 - P_j\right] \tag{2.12}$$

where for clarity we have used the notation $P_j = P(S_j = 1|\theta_j)$. Before touching the question of how to optimally set the variational parameters, we make a few general observations about this bound. First, the bound is never vacuous, i.e., there always exists a setting of the variational parameters such that the bound is less than or equal to one. To see this, let $\xi_i = 0$ for all $i \in L_1$. The bound in this case becomes simply

$$\prod_{i \in L_1} e^{\,-f^*(0)} \tag{2.13}$$

We can now make use of the property that for every[2] ξ there exist an x such that the bound is exact at that point. If we let x_0 be this point for $\xi = 0$, then $\log G_{S_i}(x_0) = x_0\, 0 - f^*(0) = -f^*(0)$. Putting this back into the reduced bound gives

$$\prod_{i \in L_1} e^{\,\log G_{S_i}(x_0)} = \prod_{i \in L_1} G_{S_i}(x_0) \tag{2.14}$$

As all $G_{S_i}(x_0)$ are conditional probabilities, the product has to remain less than or equal to one. Second, we note (without a proof) that in the limit of vanishing interaction parameters θ_{ij} between the layers, the bound becomes exact. Also, since the variational bounds can be viewed as tangents, small variations in θ_{ij} around zero are (always) captured accurately.

We now turn to the question of how to optimize the variational parameters. Due to the products involved, the bound is easier to handle in log-scale, and we write

$$\log P(\{S_i\}_{i \in L_1}|\theta) \leq \sum_{i \in L_1} [\xi_i \theta_{i0} - f^*(\xi_i)] + \sum_{j \in L_2} \log \left[e^{\sum_{i\,:\,j \in pa_i} \xi_i \theta_{ij}} P_j + 1 - P_j\right] \tag{2.15}$$

In order to make this bound as tight as possible, we need to minimize the bound with respect to ξ. Importantly, this minimization problem is convex in the variational parameters ξ (see below). The implication is that there are no local minima, and the optimal parameter settings can be always found through standard procedures such as Newton-Raphson. Let us now verify that the bound is indeed a convex function of the variational parameters ξ. The convexity of $-f^*(\xi_i)$ follows directly from the concavity of the conjugate functions f^*. The convexity of

$$\log \left[e^{\sum_{i\,:\,j \in pa_i} \xi_i \theta_{ij}} P_j + 1 - P_j\right] \tag{2.16}$$

follows from a more general convexity property well-known in the physics literature:

$$f(r) = \log E_X \left\{e^{r(X)}\right\} \tag{2.17}$$

[2] The variational parameter may have a restricted set of possible values, but in the context of the log-concave models considered here, the value zero is in the admissible set.

is a convex functional of r (see Appendix 2.A for a simple proof). The expectation pertains to the random variable X. In our case the expectation is over S_j taking values in $\{0,1\}$ with probabilities $1-P_j$ and P_j. The function r corresponds to $\sum_{i\,:\,j\in pa_i}\xi_i\theta_{ij}S_j$, and since affine transformations do not alter convexity properties, convexity in r implies convexity in ξ.

2.2.2 Generic lower bound for log-concave networks

Methods for finding lower bounds on marginal likelihoods were first presented by Dayan et al. (1995) and Hinton et al. (1995) in the context of a layered network known as the "Helmholtz machine". Saul et al. (1996) subsequently provided a more rigorous calculation of these lower bounds (by appeal to mean field) in the case of generic sigmoid networks. Unlike the method presented earlier for obtaining upper bounds presented in the previous section, this lower bound methodology poses no constraints on the network structure. The restriction on the upper bounds will be removed in chapter 3.

We provide here an alternative derivation of the lower bounds, one that establishes the connection to convex duality similarly to the upper bounds. Consider therefore the problem of computing a log-marginal probability $\log P(S^*)$:

$$\log P(S^*) = \log\sum_S P(S,S^*) = \log\sum_S e^{\,\log P(S,S^*)} = \log\sum_S e^{\,r(S)} = f(r) \tag{2.18}$$

where S and S^* denote two sets of variables, and the summation is over the possible configurations of the variables in the set S. The function r is defined as $r(S) = \log P(S,S^*)$. We can now use the previously established result that $f(r)$ is convex in r. Analogously to upper bounds, convex duality now implies that we can find a bound

$$\begin{aligned} f(r) = \log\sum_S e^{\,r(S)} &\geq \sum_S q(S)r(S) - f^*(q) && (2.19)\\ &= \sum_S q(S)\log P(S,S^*) - f^*(q) && (2.20)\\ &= \sum_S q(S)\log P(S,S^*) + H(q) && (2.21)\end{aligned}$$

which is exact if maximized over the variational distribution[3] q. The conjugate function f^* turns out to be the negative entropy $-H(q)$ of the q distribution (see Appendix 1.A). Unlike with upper bounds, the variational parameters are now quite complex, i.e., distributions over a set of variables. The optimal variational distribution $q^*(S) = P(S|S^*)$, the one that attains the maximum, is often infeasible to use in the bound. The choice of the variational distribution therefore involves a trade-off between feasibility and accuracy, where the accuracy is characterized by an inherent error metric associated with these convexity bounds (see the appendix of chapter 1).

[3] The fact that the variational parameters define a probability distribution is specific to the case we are considering.

In this case, the metric specifying the loss in accuracy due to a suboptimal choice for q is the Kullback-Leibler distance between q and the optimal distribution q^* (see Appendix 1.A). To emphasize feasibility we consider a particularly tractable (but naturally suboptimal) class of variational distributions, that of completely factorized distributions:

$$q(S) = \prod_i q_i(S_i) \tag{2.22}$$

We can insert this form into the lower bound of Eq. (2.21) and consequently adjust the parameters in q (the component distributions) to make the bound as tight as possible. The resulting approximation is known as the mean field approximation (see e.g. Saul, et al., 1996). We note that in contrast to the upper bound, the mean field lower bound is not guaranteed to be free of local maxima. Furthermore, even in the case of mean field distributions, it may not be trivial to evaluate the associated bound. While the entropy term in this case reduces to a manageable expression given by

$$H(q) = \sum_i H(q_i) = -\sum_i \sum_{S_i} q_i(S_i) \log q_i(S_i), \tag{2.23}$$

the same is not necessarily true for the first term in Eq. (2.21). For example, in the context of our log-concave models, the lower bound is given by

$$\begin{aligned}
&\sum_i \sum_S q(S) \log G_{S_i}\left(\theta_{i0} + \sum_{j \in pa_i} \theta_{ij} S_j\right) + H(q) \\
&\quad = \sum_i E_q \, \log G_{S_i}\left(\theta_{i0} + \sum_{j \in pa_i} \theta_{ij} S_j\right) + H(q)
\end{aligned} \tag{2.24}$$

where the sum over i follows from the product decomposition of the joint distribution, and the expectation is with respect to the mean field distribution q. The non-linearities in the log-conditional probabilities (i.e. in $\log G_{S_i}$) may render the expectation over the "latent" variables S exponentially costly in the number of parents of S_i (almost regardless of the form of the distribution q). Additional approximations or bounds are therefore often necessary even with the mean field approximation. These approximations, however, make use of the particulars of the models and we will consider them separately for sigmoid and noisy-OR networks. Note that the simplicity offered by the mean field distribution greatly facilitates the derivation of these additional approximations.

2.3 Sigmoid belief networks

Sigmoid belief networks belong to the class of log-concave models considered earlier. The conditional probabilities for sigmoid networks take the form

$$P(S_i|S_{pa_i},\theta) \;=\; g\left(\theta_{i0}+\sum_{j\in pa_i}\theta_{ij}S_j\right)^{S_i}\left[1-g\left(\theta_{i0}+\sum_{j\in pa_i}\theta_{ij}S_j\right)\right]^{1-S_i} \qquad (2.25)$$

where $g(x) = 1/(1+\exp(-x))$ is the logistic function (also called a "sigmoid" function based on its graphical shape; see Figure 2-6). The required log-concavity property follows from the fact that both $\log g(x)$ and $\log(1-g(x)) = \log g(-x)$ are concave functions of x. We note that the choice of the above dependency model is not arbitrary but is rooted in logistic regression in statistics (McCullagh & Nelder, 1983). Furthermore, this form of dependency corresponds to the assumption that the odds from each parent of a node combine multiplicatively; the weights θ_{ij} in this interpretation bear a relation to log-odds.

We now adapt the generic results for log-concave models to sigmoid networks and evaluate them numerically. We start with the upper bound.

2.3.1 Upper bound for sigmoid network

In order to obtain an upper for the class of two-level sigmoid networks we need to specify the exact form of the variational transformation used in Eq. (2.5) and consequently in Eq. (2.15) for log-concave models. For sigmoid networks this means to quantify the transformation of $\log g(x)$ (cf. the probability model). The derivation of this transformation is presented in Appendix 2.B. As a result we obtain

$$\log G_{S_i}\left(\theta_{i0}+\sum_{j\in pa_i}\theta_{ij}S_j\right) \;\leq\; (S_i-\xi_i)\left(\theta_{i0}+\sum_{j\in pa_i}\theta_{ij}S_j\right) - H(\xi_i) \qquad (2.26)$$

where $H(\cdot)$ is the binary entropy function. Note that we have explicated how the transformation depends on the value of S_i (unlike in the case of the generic transformation of Eq. (2.5), where the dependence was left implicit). We may now use the above transformation in Eq. (2.15) to get the desired bound on the log-marginal probability:

$$\begin{aligned}\log P(\{S_i\}_{i\in L_1}|\theta) \;\leq\; & \sum_{i\in L_1}\left[(S_i-\xi_i)\theta_{i0} - H(\xi_i)\right] \\ & + \sum_{j\in L_2}\log\left[e^{\sum_{i\,:\,j\in pa_i}(S_i-\xi_i)\theta_{ij}}P_j + 1 - P_j\right] \qquad (2.27)\end{aligned}$$

As described earlier the optimal setting of the variational parameters ξ can be found simply via the standard Newton-Raphson method.

2.3.2 Generic lower bound for sigmoid network

Our derivation of lower bounds for sigmoid networks deviates from those presented in Saul et al. (1996) in terms of the type of additional approximations used to facilitate the evaluation of the lower bound. Let us therefore describe briefly the techniques that we use as they correspond to the numerical results presented in the following section. We refer the reader to Saul et al. for their method. To this end, recall first that the evaluation of the lower bound involves performing averages over log-conditional probabilities with respect to the mean field distribution (see Eq. (2.24)). For sigmoid networks this means averages of the form

$$E_q \ \log G_{S_i}(X_i) = E_q \left\{ \ S_i X_i + \log g(-X_i) \ \right\} \tag{2.28}$$

where we have used the notation $X_i = \theta_{i0} + \sum_{j \in pa_i} \theta_{ij} S_j$ and the identity

$$\begin{aligned} G_{S_i}(X_i) &= g(X_i)^{S_i}(1-g(X_i))^{1-S_i} = \left[\frac{g(X_i)}{1-g(X_i)}\right]^{S_i}(1-g(X_i)) && (2.29) \\ &= e^{\,S_i X_i}\, g(-X_i) && (2.30) \end{aligned}$$

particular to sigmoid networks. Given that the mean field distribution q is factorized, the only difficulty in evaluating the expectations above comes from the term $E_q \log g(-X_i)$. To alleviate this problem we resort to an additional variational transformation of $\log g(-X_i)$, one that yields a lower bound (cf. the upper bound transformation presented earlier). This transformation is derived in Jaakkola & Jordan (1996) (see Appendix 3.B) and is given by

$$\log g(-X_i) \geq -(X_i - \xi_i)/2 - \lambda(\xi)(X_i^2 - \xi_i^2) + \log g(-\xi_i) \tag{2.31}$$

where $\lambda(\xi) = \tanh(\xi/2)/(4\xi)$, and the transformation is exact whenever $\xi_i = X_i$. This additional lower bound depends on X_i only quadratically and can be readily averaged over the factorized mean field distribution.

2.3.3 Numerical experiments for sigmoid network

In testing the accuracy of the developed bounds we used $8 \rightarrow 8$ networks (complete bipartite graphs as in Figure 2-1 with 8 nodes in each level), where the network size was chosen to be small enough to allow exact computation of the marginal probabilities for purposes of comparison. The method of testing was as follows. The parameters for the $8 \rightarrow 8$ networks were drawn from a Gaussian prior distribution and a sample from the resulting joint distribution of the variables was generated. The variables in the "receiving" layer of the bipartite graph were set according to the sample. The true marginal probability as well as the upper and lower bounds were computed for this setting. The resulting bounds were assessed by employing the relative error in log-likelihood, i.e. $(\log P_{\text{Bound}}/\log P - 1)$, as a measure of accuracy.

More precisely, the prior distribution over the parameters was taken to be

$$P(\theta) = \prod_i \prod_{j \in pa_i} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\theta_{ij}^2} \tag{2.32}$$

where the overall variance σ^2 allows us to vary the degree to which the resulting parameters make the two layers of the network dependent. For small values of σ^2 the layers are almost independent whereas larger values make them strongly interdependent. To make the situation worse for the bounds[4] we enhanced the coupling of the layers by setting $P(S_j|\theta_j) = 1/2$ for the variables not in the desired marginal set, i.e., making them maximally variable.

In order to make the accuracy of the bounds commensurate with those for the noisy-OR networks reported below, we summarize the results via a measure of inter-layer dependence. This dependence was estimated by

$$\sigma_{std} = \max_{i \in L_1} \sqrt{\mathrm{Var}\{P(S_i|S_{pa_i})\}} \tag{2.33}$$

i.e., the maximum variability of the conditional likelihoods. Here S_i was fixed in the $P(S_i|S_{pa_i})$ functional according to the initial sample and the variance was computed with respect to the joint distribution[5].

Figure 2-2 illustrates the accuracy of the bounds as measured by the relative log-likelihood as a function of σ_{std}[6]. In terms of probabilities, a relative error of ϵ translates into a $P^{1+\epsilon}$ approximation of the true likelihood P. Note that the relative error is always positive for the upper bound and negative for the lower bound, as guaranteed by the theory. The figure indicates that the bounds are accurate enough to be useful. In addition, we see that the the upper bound deteriorates faster with increasingly coupled layers.

Let us now briefly consider the scaling properties of the bounds as the network size increases. We note first that the evaluation time for the bounds increases approximately linearly with the number of parameters θ in these two-level networks[7]. The accuracy of the bounds, on the other hand, needs experimental illustration.

In large networks it is not feasible to compute σ_{std} nor the true marginal likelihood. We may, however, calculate the relative error between the upper and lower bounds. To maintain approximately same level of σ_{std} across different network sizes we plotted the errors against $\sigma\sqrt{n}$ (for fully connected n by n two-level networks), where σ is the overall standard deviation in the prior distribution. Figure 2-3 shows that the relative errors are invariant to the network size in this scaling.

[4] Both the upper and lower bounds are exact in the limit of lightly coupled layers.

[5] Note that $P(S_i|S_{pa_i})$ with S_i fixed is just some function of the variables in the network whose variance can be computed.

[6] Note that the maximum value for σ_{std} is 1/2.

[7] The amount of computation needed for sequentially optimizing each variational parameter once scales linearly with the number of network parameters. Only a few such iterations are needed.

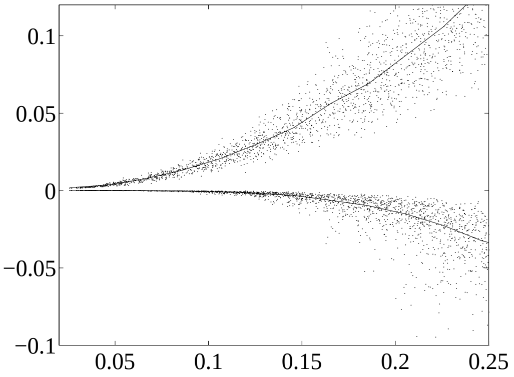

Figure 2-2: Sigmoid networks. Accuracy of the bounds for 8 by 8 two-level networks. The solid lines are the median relative errors in log-likelihood as a function of σ_{std}. The upper and lower curves correspond to the upper and lower bounds respectively.

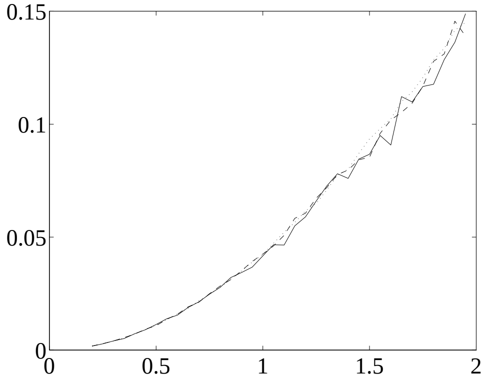

Figure 2-3: Sigmoid networks. Median relative errors between the upper and lower bounds (log scale) as a function of $\sigma n^{\frac{1}{2}}$ for n by n two-level networks. Solid line: $n = 8$; dashed line: $n = 32$; dotted line: $n = 128$.

2.4 Noisy-OR network

Noisy-OR networks – like sigmoid networks – are log-concave models. In defining the conditional probabilities for noisy-OR networks we deviate somewhat from the standard notation to reveal the relation to log-concave models and for clarity of the forthcoming expressions. In our notation the conditional probabilities are written as

$$P(S_i|S_{pa_i},\theta) \;=\; \left(1 - e^{-[\theta_{i0}+\sum_{j\in pa_i}\theta_{ij}S_j]}\right)^{S_i}\left(e^{-[\theta_{i0}+\sum_{j\in pa_i}\theta_{ij}S_j]}\right)^{1-S_i} \qquad (2.34)$$

where all the weights θ_{ij} are non-negative. The connection to the standard notation is via $\theta_{ij} = -\log(1-q_{ij})$, where q_{ij} is the probability corresponding to the proposition that the j^{th} parent alone can turn S_i on. q_{i0} (related to the bias in our notation) has a special interpretation as a leak probability, i.e., the probability that something other than the parents can independently turn S_i on. The log-concavity feature of this probability model stems from the concavity of $\log(1-e^{-x})$.

As with sigmoid networks we now adapt the general results to noisy-OR networks and verify the accuracy of the bounds numerically.

2.4.1 Upper bound for noisy-OR network

To specify the variational transformation of the noisy-OR conditional probabilities, we only need to consider the transformation of $\log G_1(x) = \log(1-e^{-x})$ since $\log G_0(x)$ already has a linear form (see the probability model). Details are given in Appendix 2.C. By combining the transformation and the linear form, we find

$$\log G_{S_i}\left(\theta_{i0} + \sum_{j\in pa_i} \theta_{ij} S_j\right) \;\leq\; (S_i\xi_i + S_i - 1)\left(\theta_{i0} + \sum_{j\in \mathrm{pa}_i} \theta_{ij} S_j\right) - S_i f^*(\xi_i) \qquad (2.35)$$

where $f^*(\xi_i) = \xi_i \log \xi_i - (\xi + 1)\log(\xi_1 + 1)$. We have again written explicitly the dependency on the value of S_i. This transformation leads to the following log-marginal bound (see Eq. (2.15)):

$$\begin{aligned} \log P(\{S_i\}_{i\in L_1}|\theta) \;&\leq\; \sum_{i\in L_1} [(S_i\xi_i + S_i - 1)\theta_{i0} - S_i f^*(\xi_i)] \\ &\quad + \sum_{j\in L_2} \log\left[e^{\sum_{i\,:\,j\in pa_i}(S_i\xi_i+S_i-1)\theta_{ij}} P_j + 1 - P_j\right] \qquad (2.36)\end{aligned}$$

The log-bound is convex in the variational parameters and can be optimized by appeal to standard methods.

2.4.2 Generic lower bound for noisy-OR network

The earlier work on lower bounds by Saul et al. was restricted to sigmoid networks; we extend that work here by deriving a lower bound for generic noisy-OR networks. We refer to section 2.2.2 for the framework and commence from the noisy-OR counterpart of eq. (2.21). For clarity, we use the notation $X_i = \theta_{i0} + \sum_{j\in pa_i} \theta_{ij} S_j$ and find

$$\begin{aligned} \log P \;&\geq\; \sum_i E_q \log G_{S_i}(X_i) + H(q) \qquad (2.37) \\ &=\; \sum_i E_q \left\{\, S_i \log(1-e^{-X_i}) \,\right\} + \sum_i E_q \left\{\, -(1-S_i)X_i \,\right\} + H(q) \qquad (2.38)\end{aligned}$$

which comes from using the form of the conditional probabilities for noisy-OR networks; $\log P$ is a shorthand for the log-marginal we are trying to compute. While the second (mean field) expectation in eq. (2.38) simply corresponds to replacing the binary variable S_i and those in the linear form X_i with their "means" q_i [8], the first expectation lacks a closed form expression. To compute this expectation efficiently

[8] For simplicity we denote $q_i(S_i = 1)$ by q_i.

we make use of the following expansion:

$$1 - e^{-x} = \prod_{k=0}^{\infty} g(2^k x) \tag{2.39}$$

where $g(\cdot)$ is the sigmoid function (see Appendix 2.D for details). This expansion converges exponentially fast and thus only a few terms need to be included in the product for good accuracy. By carrying out this expansion in the bound above and explicitly using the form of the sigmoid function we get

$$\begin{aligned} \log P \;\; &\geq \;\; \sum_i \sum_k E_q \Big\{ -S_i \log(1 + e^{-2^k X_i}) \Big\} \\ &\quad - \sum_i (1 - q_i) \sum_j \theta_{ij} q_j + H(q) \end{aligned} \tag{2.40}$$

Now, as the parameters θ_{ij} are non-negative, $X_i = \theta_{i0} + \sum_{j \in pa_i} \theta_{ij} S_j \geq 0$, and

$$e^{-2^k X_i} \in [0, 1] \tag{2.41}$$

We may therefore use the smooth convexity properties of $-\log(1 + x)$ (for $x \in [0, 1]$) to bring the expectations in eq. (2.40) inside the log. This results in

$$\begin{aligned} \log P \;\; &\geq \;\; -\sum_{ik} q_i \, \log \left[1 + e^{-2^k \theta_{i0}} \prod_{j \in pa_i} (q_j e^{-2^k \theta_{ij}} + 1 - q_j) \right] \\ &\quad - \sum_i (1 - q_i) \sum_{j \in pa_i} \theta_{ij} q_j + H(q) \end{aligned} \tag{2.42}$$

A more sophisticated and accurate way of computing the expectations in eq. (2.40) is discussed in Appendix 2.E.

2.4.3 Numerical experiments for noisy-OR network

The method of testing used here was, for the most part, identical to the one presented earlier for sigmoid networks (section 2.3.3). The only difference was that the prior distribution over the parameters defining the conditional probabilities was chosen to be an exponential instead of a Gaussian:

$$\theta_{ij} \sim \lambda \, e^{-\lambda \theta_{ij}} \tag{2.43}$$

(this corresponds to a Dirichlet distribution over the parameters in the standard noisy-OR notation). For large λ, θ_{ij} stays small and the layers of the bipartite network are only weakly connected; smaller values of λ, on the other hand, make the layers strongly dependent. We thus used λ to vary (on average) the interdependence between the two layers. To facilitate comparisons with the bounds derived for sigmoid networks we used σ_{std} (see eq. (2.33)) as a measure of dependence between the layers.

Figure 2-4 illustrates the accuracy of the computed bounds as a function of σ_{std}[9]. The samples with zero relative error are from the upper/lower bounds in cases where all the variables in the desired marginal are zero since the bounds become exact whenever this happens. The lower bound is slightly worse than the one for sigmoid networks most likely due to the symmetry and smoother nature of the sigmoid function. As with the sigmoid networks the upper bound becomes less accurate more quickly.

We now turn to the effects of increasing the network size. Analogously to the sigmoid networks the evaluation times for the bounds vary approximately linearly with the number of parameters in these two-level networks, albeit with slightly larger coefficients (for the lower bound). As for the accuracy of the bounds, Figure 2-5 shows the relative errors[10] between the bounds across different network sizes. The errors are plotted against $\sqrt{n}/\lambda$ for n by n two-level networks, where λ defines the exponential prior distribution for the parameters. In the chosen scale the network size has little effect on the errors[11].

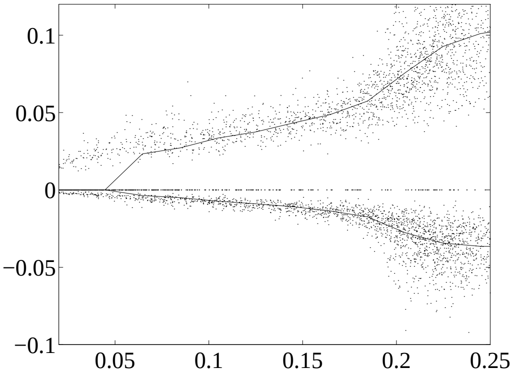

Figure 2-4: Noisy-OR network. Accuracy of the bounds for 8 by 8 two-level networks. The solid lines are the median relative errors in log-likelihood as a function of σ_{std}. The upper and lower curves correspond to the upper and lower bounds respectively.

2.5 Discussion and future work

Applying probabilistic methods to real world inference problems can lead to the emergence of cliques that are prohibitively large for exact algorithms (for example, in medical diagnosis). We focused on dealing with such large (sub)structures in the context of a class of networks we call log-concave models with an emphasis on the sigmoid and noisy-OR realizations. For these networks we developed techniques for

[9] The slight unevenness of the samples are due to the non-linear relationship between the parameter λ for the exponential distribution and σ_{std}.

[10] The errors are for the worst case marginal, i.e., for $P(\{S_i = 1\}_{i \in L_1})$.

[11] The 8 by 8 network is too small to be in the desired asymptotic regime.

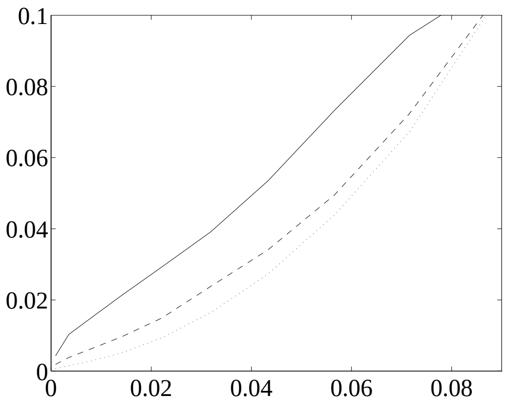

Figure 2-5: Noisy-OR network. Median relative errors between the upper and lower bounds (in log scale) as a function of $n^{\frac{1}{2}}/\lambda$ for n by n two-level networks. Solid line: $n = 8$; dashed line: $n = 32$; dotted line: $n = 128$.

computing upper and lower bounds on marginal probabilities. The bounds serve as an alternative to sampling methods in the presence of intractable structures. They can define interval bounds for the marginals and can be used to improve the accuracy of decision making in intractable networks.

Toward extending the work presented in this chapter we note that both the upper and lower bounds can be improved by considering a mixture partitioning (Jaakkola & Jordan 1996) of the space of marginalized variables instead of using a completely factorized approximation. The restriction of the upper bounds for two-level networks can be overcome by introducing the variational bounds recursively; this recursive generalization will be considered in detail in the next chapter. There we will also quantify how the variational techniques can be merged with exact probabilistic calculations whenever they are/become feasible (see also chapter 5, Draper 1994, Saul & Jordan 1996).

References

P. Dayan, G. Hinton, R. Neal, and R. Zemel (1995). The Helmholtz machine. *Neural Computation* **7**: 889-904.

P. Dagum and M. Luby (1993). Approximate probabilistic reasoning in Bayesian belief networks is NP-hard. *Artificial Intelligence* **60**: 141-153.

D. Draper (1995). *Localized partial evaluation of belief networks.* PhD thesis, University of Washington.

D. Draper & S. Hanks (1994). Localized Partial Evaluation of Belief Networks. In *Proceedings of the Tenth Annual Conference on Uncertainty in Artificial Intelligence.*

B. D'Ambrosio (1994). Symbolic probabilistic inference in large BN20 networks. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence.* Morgan

Kaufmann.

G. Hinton, P. Dayan, B. Frey, and R. Neal (1995). The wake-sleep algorithm for unsupervised neural networks. *Science* **268**: 1158-1161.

D. Heckerman (1989). A tractable inference algorithm for diagnosing multiple diseases. In *Proceedings of the Fifth Conference on Uncertainty in Artificial Intelligence.* Morgan Kaufmann.

T. Jaakkola, L. Saul, and M. Jordan (1996). Fast learning by bounding likelihoods in sigmoid-type belief networks. To appear in *Advances of Neural Information Processing Systems 8.* MIT Press.

T. Jaakkola and M. Jordan (1996). Mixture model approximations for belief networks. Manuscript in preparation.

F. Jensen, S. Lauritzen, and K. Olesen (1990). Bayesian updating in causal probabilistic networks by local computations. *Computational Statistics Quarterly* **4**: 269-282.

S. Lauritzen and D. Spiegelhalter (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society B* **50**:154-227.

P. McCullagh & J. Nelder (1983). *Generalized linear models.* Chapman and Hall.

R. Neal. Connectionist learning of belief networks (1992). *Artificial Intelligence* **56**: 71-113.

J. Pearl (1988). *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann.

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

L. Saul and M. Jordan (1996). Exploiting tractable substructures in intractable networks. To appear in *Advances of Neural Information Processing Systems 8.* MIT Press.

M. Shwe, B. Middleton, D. Heckerman, M. Henrion, E. Horvitz. H. Lehmann, G. Cooper (1991). Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base. *Methods of Information in Medicine* **30**: 241-255.

2.A Convexity

We show that

$$f(r) = \log E_X\left\{e^{r(X)}\right\} = \log \sum_i p_i\, e^{\,r(X_i)} \tag{2.44}$$

is a convex functional of r. For clarity we have assumed that X takes values in a discrete set $\{X_1, X_2, \ldots\}$ with probabilities p_i. Taking the gradient with respect to $r_k = r(X_k)$ gives

$$\frac{\partial}{\partial r_k} f(r) = \frac{p_k e^{\,r_k}}{\sum_i p_i \, e^{\,r_i}} = P_k \tag{2.45}$$

where P_k defines a probability distribution. The convexity is revealed by a positive semi-definite Hessian $\mathcal{H}$, whose components in this case are

$$\mathcal{H}_{kl} = \frac{\partial^2}{\partial r_k \partial r_l} f(r) = \delta_{kl} P_k - P_k P_l \tag{2.46}$$

To see that $\mathcal{H}$ is positive semi-definite, consider

$$x^T \mathcal{H} x = \sum_k P_k x_k^2 - (\sum_k P_k x_k)(\sum_l P_l x_l) = Var_P\{x\} \geq 0 \tag{2.47}$$

where $Var_P\{x\}$ is the variance of x_i with respect to the distribution P_i.

2.B Sigmoid transformation

Owing to the concavity of $\log g(x)$ we can write

$$\log g(x) = \min_{\xi} \{ \, \xi x - f^*(\xi) \, \} \leq \xi x - f^*(\xi) \tag{2.48}$$

where $g(x) = 1/(1 + e^{-x})$ is the sigmoid function. It remains to specify the conjugate function $f^*(\xi)$. We can obtain this function by reversing the roles of log-sigmoid and its conjugate in the above representation:

$$f^*(\xi) = \min_{x} \{ \, \xi x - \log g(x) \, \} \tag{2.49}$$

To carry out the maximization we set

$$\frac{\partial}{\partial x} \{ \, \xi x - \log g(x) \, \} = \xi - g(-x) = 0 \tag{2.50}$$

giving $x = -g^{-1}(\xi) = \log(1 - \xi)/\xi$. This point attains the maximum in the representation for the conjugate function and therefore at that point

$$\begin{aligned} f^*(\xi) = \xi x - \log g(x) &= \xi \log \frac{1 - \xi}{\xi} - \log(1 - \xi) && (2.51) \\ &= -\xi \log \xi - (1 - \xi) \log(1 - \xi) = H(\xi) && (2.52) \end{aligned}$$

where we have used the fact that $\log g(x) = \log(1 - g(-x)) = \log(1 - \xi)$; $H(\xi)$ is the binary entropy function. Consequently, we have the bound

$$\log g(x) \leq \xi x - H(\xi) \tag{2.53}$$

This log-bound implies a bound on the sigmoid function given by

$$g(x) \leq e^{\,\xi x - H(\xi)} \tag{2.54}$$

which is the form actually used in the probabilistic calculations in the text. The geometry of this bound for a fixed ξ is illustrated in figure 2-6. The point x where the bound is exact corresponds to the maximizing point $x = \log(1 - \xi)/\xi$ obtained earlier.

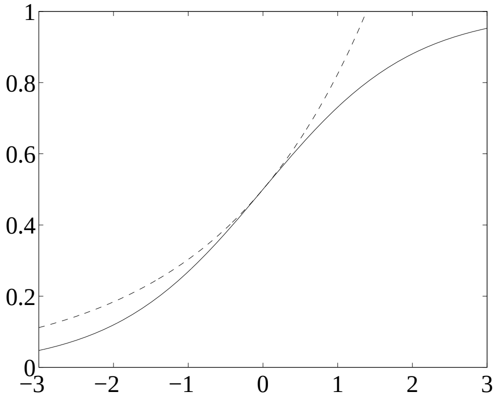

Figure 2-6: Geometry of the sigmoid transformation. The dashed curve plots $\exp\{\xi x - H(\xi)\}$ as a function of x for a fixed ξ (=0.5).

2.C Noisy-OR transformation

As already stated, $\log(1 - e^{-x})$ is a concave function of x and therefore

$$\log(1 - e^{-x}) = \min_{\xi} \{\, \xi x - f^*(\xi) \,\} \leq \xi x - f^*(\xi) \tag{2.55}$$

To specify the conjugate function f^* we can proceed as in Appendix 2.B for the sigmoid function:

$$f^*(\xi) = \max_{x} \{\, \xi x - \log(1 - e^{-x})\} \tag{2.56}$$

The point x that attains the maximum is solved from

$$\frac{\partial}{\partial x} \{\, \xi x - \log(1 - e^{-x})\} = \xi - \frac{e^{-x}}{1 - e^{-x}} = 0 \tag{2.57}$$

giving $x = \log(1+\xi)/\xi$. The conjugate function therefore becomes

$$f^*(\xi) = \xi x - \log(1 - e^{-x}) = \xi \log \xi - (1+\xi)\log(1+\xi) \tag{2.58}$$

Turning the resulting bound on $\log(1-e^{-x})$ into a bound on the noisy-OR conditional probability $1 - e^{-x}$ gives

$$1 - e^{-x} \leq e^{\,\xi x - f^*(\xi)} \tag{2.59}$$

The geometric behavior of this bound can be seen in figure 2-7 where we have plotted $1 - e^{-x}$ and the associated bound for a fixed value of ξ. The point where the bound is exact is again given by the maximizing point $x = \log(1+\xi)/\xi$.

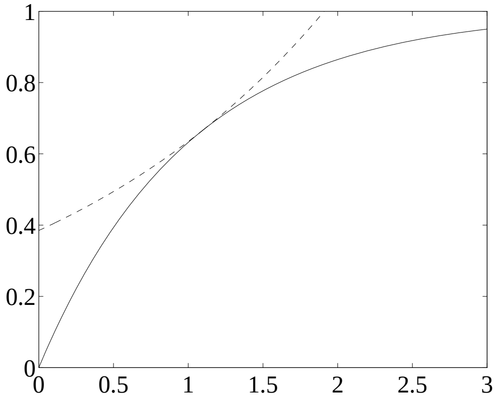

Figure 2-7: Geometry of the noisy-OR transformation. The dashed curve gives $\exp\{\xi x - f^*(\xi)\}$ as a function of x when ξ is fixed at 0.5.

2.D Noisy-OR expansion

The noisy-OR expansion

$$1 - e^{-x} = \prod_{k=0}^{\infty} g(2^k x) \tag{2.60}$$

follows simply from

$$\begin{aligned} 1 - e^{-x} &= \frac{(1+e^{-x})(1-e^{-x})}{1+e^{-x}} = g(x)(1-e^{-2x}) \\ &= g(x)\frac{(1+e^{-2x})(1-e^{-2x})}{1+e^{-2x}} = g(x)g(2x)(1-e^{-4x}) \end{aligned} \tag{2.61}$$

and induction. For $x > 0$ the accuracy of the expansion is governed by $1 - e^{-2^k x}$ which goes to one exponentially fast. Also since $g(2^k 0) = 1/2$, the expansion becomes $(\frac{1}{2})^N$ at $x = 0$, where N is the number of terms included. As this approaches $1 - e^{-0} = 0$ exponentially fast, we conclude that the rapid convergence is uniform. Figure 2-8 illustrates the accuracy of the expansion for small N.

A word of caution is needed here. When the expansion is used as an approximation, i.e., only a few terms are included, it actually gives an upper bound on $1 - e^{-x}$ (see the figure or use the fact that $1 - e^{-4x}$ is less than one in Eq. (2.61)). The expansion is nevertheless used in the text to obtain lower bounds. It is therefore essential that sufficiently many terms are included in the expansion or otherwise we run the risk of compromising the lower bound we are after. To determine an appropriate number of terms to include, we assume that there is an $\epsilon > 0$ such that $\epsilon \leq x$ with high probability. For the expansion to be accurate regardless of the distribution over x, we must have $2^k\epsilon \gg 1$. This implies that $k \gg -\log_2 \epsilon$ many terms need to be included. Adding many terms to the expansion, however, can be costly for other reasons and therefore in cases where ϵ is very small the expansion may not be appropriate.

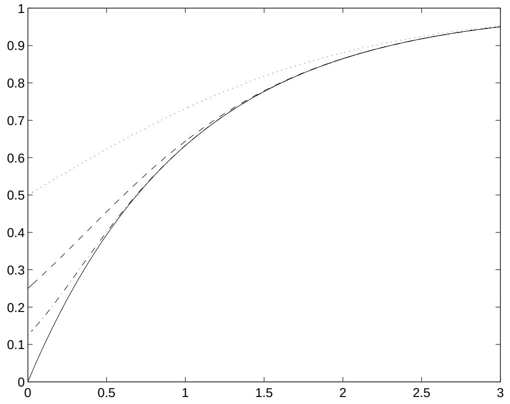

Figure 2-8: Accuracy of the noisy-OR expansion. Dotted line: $N = 1$; dashed line: $N = 2$; dotdashed: $N = 3$. N is the number of terms included in the expansion.

2.E Quadratic bound

For $X \in [0, 1]$ we can bound $-\log(1 + X)$ by a quadratic expression:

$$-\log(1 + X) \geq a(X - x)^2 + b(X - x) + c \tag{2.62}$$

where $c = -\log(1 + x)$, $b = -1/(1 + x)$, and $a = -[(1 - x)b + c + \log 2]/(1 - x)^2$. The coefficients can be derived by requiring that the quadratic expression and its derivative are exact at $X = x$, and by choosing the largest possible a such that the expression remains a bound. The resulting approximation is good for all $x \in [0, 1]$ and can be optimized by setting $x = E\{X\}$.

Let us now use this quadratic bound in eq. (2.40) to better approximate the expectations. To simplify the ensuing formulas we use the notation

$$E_q\left\{ e^{-2^k[\theta_{i0}+\sum_{j\in \mathrm{pa}_i}\theta_{ij}S_j]} \right\} = e^{-2^k\theta_{i0}} \prod_{j\in \mathrm{pa}_i} \left(q_j e^{-2^k[\theta_{ij}} + 1 - q_j \right) = X_i^{(k)} \tag{2.63}$$

With these we straightforwardly find

$$\begin{aligned}\log P(\{S_i\}_{i\in L}|\theta) \;\; \geq \;\; & \sum_{ik} q_i a_{ik} \left[X_i^{(k+1)} - 2X_i^{(k)} x_i^{(k)} + (x_i^{(k)})^2\right] \\ & + \sum_{ik} q_i \left[b_{ik}(X_i^{(k)} - x_i^{(k)}) + c_{ik}\right] \\ & - \sum_i (1-q_i) \sum_j \theta_{ij} q_j + H(q) \qquad (2.64)\end{aligned}$$

which is optimized with respect to $x_i^{(k)}$ simply by setting $x_i^{(k)} = X_i^{(k)}$. The simpler bound in eq. (2.42) corresponds to ignoring the quadratic correction, i.e., using $a_{ik} = 0$ above.

Chapter 3

Recursive algorithms[1]

3.1 Introduction

In this chapter we complement and generalize the results of chapter 2 by extending the bounding techniques to other types of graphical models and dispensing with the rather strong two-layer topological constraints on the upper bounds that we assumed in the previous chapter. In addition, the framework developed in this chapter will allow the variational methods to be straightforwardly integrated with exact probabilistic calculations. The key difference making these generalizations possible is the recursive introduction of the simplifying variational transformations; the computation of upper and lower bounds on marginal probabilities in this formalism can proceed in a node-elimination fashion. This elimination formalism applies to a powerful class of networks known as chain graphs (see e.g. Whittaker 1990, Lauritzen 1996), and although the type of chain graphs that we consider in detail in this chapter are somewhat restricted, Boltzmann machines and sigmoid belief networks are nevertheless included as special cases.

While we attain generality with recursive bounds, we also have to rely on assumptions that guarantee a feasible continuation of the recursion steps. These assumptions will constrain the results of this chapter to the extent that the recursive formalism does not entail all the methods of chapter 2 but is to some degree complementary. This is particularly true with noisy-OR networks as will be explained later.

The methodology proposed in this chapter nevertheless comes close to achieving the objectives set previously (in the introductory chapter 1) for approximate methods: they are controlled, reasonably general, and mergeable with exact probabilistic calculations. We start by developing the general recursive formalism. We then apply the formalism to Boltzmann machines, sigmoid belief networks, and to chain graphs.

[1] This chapter extends the material in the previously appeared article "T. Jaakkola and M. Jordan (1996). Recursive algorithms for approximating probabilities in graphical models. In *Advances of Neural Information Processing Systems 9*".

3.2 Recursive formalism

The approximation methodology that we derive will be applicable to graphical models whether undirected, directed or chain graphs (currently some restrictions apply). Let us first introduce our notation by defining the probability models for these graphs and clarify the relevant differences between them.

3.2.1 Probability models

Undirected graphical models (a.k.a. Markov random fields) are defined in terms of potential functions that specify a Gibbs' distribution over the network variables. The normalization is global, and the joint distribution is given by

$$P(S) = \frac{1}{Z}\, e^{\,\phi(S)} \tag{3.1}$$

where $S = \{S_1, \ldots, S_n\}$ is the set of variables or nodes in the graph. The (constant) partition function $Z = \sum_S e^{\,\phi(S)}$ performs the normalization. If a graphical representation of the joint distribution is given, then the potential functions in the above joint distribution cannot be arbitrary but must conform to the constraints inherent in the graph. The Markov properties of the graph are consistent with the joint whenever the potential functions are linear combinations of clique potentials. Clique potentials depend only on the variables in each clique (see Besag 1974 for details).

Unlike the global normalization in undirected models, directed graphical models presume a local normalization. Each variable S_i in these models is normalized with respect to a subset of variables called the parents of S_i (denoted by S_{pa_i}). A valid probability model is obtained when the variables and their parents conform to some global ordering, where the variable must follow all its parents. The local normalizations give rise to conditional probabilities (a variable given its parents) and the joint distribution is simply a product of these conditionals. In a potential form we have

$$P(S) = \prod_i P(S_i|S_{pa_i}) = \prod_i \frac{1}{Z_{i|S_{pa_i}}}\, e^{\,\phi_i(S_i|S_{pa_i})} \tag{3.2}$$

where the local partition functions $Z_{i|S_{pa_i}} = \sum_{S_i} e^{\,\phi_i(S_i|S_{pa_i})}$ guarantee the normalization. Note that these are in general random variables (functions of random variables S_{pa_i}). The directed graphical representation fixes only the set of parents for each variable and therefore leaves the potential functions in the above representation unconstrained.

The chain graphs (Whittaker 1990, Lauritzen 1996) are combinations of undirected and directed graphs and therefore the most general of the three classes. They can be viewed as directed graphs over clusters of variables, where the structure of each cluster is specified by an undirected graphical model. Analogously to directed graphs, we can define the probability model on the level of clusters:

$$P(S) = \prod_c P(S_c|S_{pa_c}) = \prod_c \frac{1}{Z_{c|S_{pa_c}}} \, e^{\,\phi_c(S_c|S_{pa_c})} \tag{3.3}$$

where S_c refers to the set of variables comprising the cluster c. Chain graphs have therefore a local normalization on the level of clusters, whereas within each cluster the normalization is global. Unlike in directed graphical models, the potential functions $\phi_c(S_c|S_{pa_c})$ are constrained; the structure in the undirected graphs corresponding to each cluster imposes restrictions on the possible dependencies in the potential functions, just as with ordinary undirected models. We note finally that directed graphs are special cases of chain graphs, where each cluster contains only one variable. The undirected models, on the other hand, have only one cluster.

3.2.2 The formalism

The objective here is to be able to compute marginal probabilities in graphical models regardless of their type, size, or complexity. Naturally, approximations are needed to achieve this and the formalism we develop makes use of variational transformations. The transformations we consider have their origin in convex duality as described previously in section 1.1. The advantage of variational methods compared to other approximate techniques is that they are controlled, i.e., they yield upper or lower bounds on the quantities of interest. The approximation framework consists of three phases: 1) transformation, 2) recursive elimination, and 3) exact computation. These phases overlap to some extent and the division is intended to clarify the main differences. The main function of the transformation phase is to rerepresent directed and chain graphs in terms of undirected graphical models. We will not moralize or triangulate the graph in this phase to perform a strict transformation of directed (or chain graphs) to undirected models. These procedures will be employed in the exact computation phase, which is used whenever it becomes applicable. Instead, we retain the compactness of the conditional probabilities and the change of representation from directed to undirected models involves approximations (esp. for chain graphs) that will facilitate the subsequent marginalization task. In the recursive marginalization phase, we successively eliminate or marginalize over the variables that are not in the desired marginal set. The recursivity of this procedure allows the introduction of simplifying transformations at various stages (consistently yielding upper or lower bounds) that guarantee the feasible continuation of the marginalization procedure. Exact algorithms will take over whenever they become feasible. We will now consider these phases in more detail.

3.2.3 Transformation

For directed and chain graphs it is often their local normalizations or partition functions that make them unwieldy for marginalization. To assist the marginalization process it is therefore desirable to transform these graphical models into undirected ones. While such "transformation" is superficially achieved by rewriting the conditional probabilities in terms of potentials, i.e.,

$$P(S_c|S_{pa_c}) = \frac{1}{Z_{S_{pa_c}}} e^{\,\phi(S_c|S_{pa_c})} = e^{\,\phi(S_c|S_{pa_c}) - \log Z_{S_{pa_c}}} \tag{3.4}$$

the resulting undirected model would be just as infeasible. The transformations therefore must involve a certain degree of simplification. To achieve this in a principled way, we employ variational methods to transform the local partition functions. The objective of these transformations is to bound the partition functions in terms of single Boltzmann factors[2], the potentials in which are feasible for further calculations. More precisely,

$$Z_{S_{pa_c}} = \sum_{S_c} e^{\,\phi(S_c|S_{pa_c})} \quad \begin{smallmatrix}<\\>\end{smallmatrix} \quad e^{\,\tilde{\phi}(S_{pa_c})} \tag{3.5}$$

where $\tilde{\phi}(S_{pa_c})$ is an *effective* potential. In order for the effective potential to posses a sufficient degree of simplification we require that it should belong to the same class of functions as the original potentials $\phi(S_c|S_{pa_c})$. The motivation for this requirement will become clear later. We note that the reduction of (local) partition functions into effective potentials is a form of renormalization[3], which in our case is carried out with bounds rather than exactly. This, however, is exactly the goal of the recursive marginalization phase considered below. We will thus defer the details of the process by which the effective potential is obtained, and instead consider some implications. Substituting the bounding Boltzmann factor for the partition function of the conditional probability in Eq. (3.4) gives

$$\frac{1}{Z_{S_{pa_c}}} e^{\,\phi(S_c|S_{pa_c})} = e^{\,\phi(S_c|S_{pa_c}) - \log Z_{S_{pa_c}}} \quad \begin{smallmatrix}<\\>\end{smallmatrix} \quad e^{\,\phi(S_c|S_{pa_c}) - \tilde{\phi}(S_{pa_c})} = e^{\,\tilde{\phi}(S_c, S_{pa_c})} \tag{3.6}$$

where the last equality follows from our assumption that the effective potential remains in the same function class as the original potentials and therefore the two can be merged. The resulting potential function that characterizes the desired undirected model may have an altered dependency structure. Such "graphical" changes will be considered later in conjunction with the effective potential. While the notation used in this section was for chain graphs, the results apply equally well to directed models (as special cases).

We note finally that sometimes the complexity of the potential functions themselves can preclude efficient marginalization and, in this case, variational techniques can be introduced to simplify them as well. This is the case with noisy-OR networks, for example, for which the local partition functions are unity and the potentials cor-

[2] A Boltzmann factor is a term of the form $e^{\,\phi(S)}$.

[3] More precisely we refer to position space renormalization group. For a subset of lattice models such renormalization can be performed exactly; for others approximate methods have been developed. These approximations do not yield upper and lower bounds on the original quantity, however, and are therefore unsuitable for our purposes.

respond to the log-conditional probabilities. The transformations of chapter 2 are applicable in this setting.

Examples of networks directly covered by the formalism will be considered later.

3.2.4 Recursive marginalization

In light of the previous section, we assume that the probability model of interest is an undirected graphical model and that the objective is to compute the marginal distribution over a subset of variables, which we denote by S^*. This is equivalent to finding the partition function

$$Z^* = \sum_S e^{\,\phi(S,S^*)} \tag{3.7}$$

where S refers to the set of variables to be marginalized over. In the following we will drop the explicit reference to S^*. We propose to carry out the required marginalization recursively:

$$Z = \sum_S e^{\,\phi(S)} = \sum_{S \setminus S_k} \left[\sum_{S_k} e^{\,\phi(S)} \right] \tag{3.8}$$

where after each marginalization step, the resulting expression (in brackets) is transformed so as to be able to continue with the recursion over the remaining variables. The transformations we consider for this purpose are of the form

$$\sum_{S_k} e^{\,\phi(S)} \;\; \begin{matrix} < \\ > \end{matrix} \;\; e^{\,\tilde{\phi}(S \setminus S_k)} \tag{3.9}$$

where the Boltzmann factor corresponding to the effective potential $\tilde{\phi}(S \setminus S_k)$ must remain amenable to further marginalization. If we require the effective potential to be in the same class of functions as the original potentials, then the complexity of the subsequent marginalizations will remain at most at the same level (instead of increasing exponentially). This property is reminiscent of exact renormalization techniques in the context of graphical models possessing a particularly feasible structure, as is the case with decimatable graphs[4]. In our case each renormalization step (consistently) introduces either an upper or lower bound instead of yielding the exact result. We note that decimatable models or any other types of graphs that permit a feasible application of exact methods would be treated accordingly in our framework as well. We have therefore made a tacit assumption that at least at this stage of the recursive algorithm, the graph does not possess any such structure.

We now turn to defining the variational transformations allowing the bounds in Eq. (3.9). As mentioned earlier the variational transformations we will consider come

[4] Decimation is an exact renormalization technique well-known in the physics literature (see e.g. Itzykson & Drouffe 1989). Saul & Jordan (1994) give an introduction to decimation as well as an application to "Boltzmann trees".

from convex duality, which are illustrated in detail in section 1.1. We start with a lower bound transformation.

Lower bounds

In appendix 3.A we derive the following lower bound transformation:

$$\log \sum_{S_k} e^{\phi(S)} \;\; \geq \;\; \sum_{S_k} q(S_k)\phi(S) + H(q) \tag{3.10}$$

where $H(q)$ is the entropy of the variational distribution $q(S_k)$. Raising both sides to the exponent, we see that this lower bound gives the desired effective potential $\tilde{\phi}(S \setminus S_k)$. Whenever the original potentials are polynomials or any other functions closed under affine transformations then the resulting effective potential remains in the same function class, as required. The above transformation is precisely the mean field approximation applied selectively to only one variable. There are several possible extensions towards more accurate transformations. For example, we may easily use a mixture distribution in place of the q distribution above, where the scope of the mixture components (mean field approximations) can be extended over several recursions. Put another way, each mixture component can be defined as a sequence of mean field approximations thereby providing the tools for recapturing the dependencies that the naive mean field would ignore. We note that a correct application of the mixture extension introduces an additional mutual information term into the bound. There are nevertheless efficient methods for dealing with the added complexity (Jaakkola & Jordan 1996).

We now consider the graphical implications of the simple lower bound transformation in comparison to those resulting from the exact marginalization. For any two variables that are connected in the original graph, there has to be a term in the potential $\phi(S)$ that depends on both variables. The correct marginalization over S_k would introduce such terms between all the variables adjacent to S_k in the (undirected) graph. In the approximate marginalization, however, where the effective potential is a linear combination of the existing potentials, none will be introduced. Graphically, therefore the lower bound marginalization amounts to simply eliminating the variable from the graph. The elimination nevertheless comes with quantitative changes to the existing potentials albeit that qualitatively they remain unchanged.

Upper bounds

Upper bound transformations for generic networks are currently under development and we demonstrate only the availability of such bounds for a particular class of binary networks known as Boltzmann machines (see e.g. Hertz et al. 1991). These are graphical models with pairwise potentials mediating the interactions between the binary variables in the model. More specifically, the potential $\phi(S)$ for these models consists of pairwise terms $\phi_{ij}(S_i, S_j) \equiv J_{ij} S_i S_j$ (symmetric) and $\phi_i(S_i) \equiv h_i S_i$, where the notation in terms of the weights J_{ij} and biases h_i is more typical. In appendix

3.B we derive the following upper bound transformation

$$\log \sum_{S_k \in \{0,1\}} e^{\phi(S)} \leq (\phi_1 + \phi_0)/2 + \lambda\,(\phi_1 - \phi_0)^2 - f^*(\lambda) \tag{3.11}$$

where we have used the shorthand $\phi_s = \phi(S)_{|S_k=s}$. λ is a variational parameter, and f^* a conjugate function whose form is specified in the appendix. This upper bound is the effective potential we are after. To confirm that it remains in the class of pairwise potential functions, let us switch to the notation in terms of weights and biases. The difference $\phi_1 - \phi_0$ becomes

$$\phi_1 - \phi_0 = h_k + \sum_{j \in \mathrm{adj}_k} J_{kj} S_j \tag{3.12}$$

in which the summation is over the variables that are adjacent to S_k in the original graph. The square of the difference $\phi_1 - \phi_0$, the highest order term in the effective potential, clearly contains no higher than pairwise dependencies.

In terms of the graphical consequences of this upper bound transformation, we note that the squared term $(\phi_1 - \phi_0)^2$ will create a dependence between each pair of variables adjacent to the one being marginalized over. Graphically these changes in the connectivity correspond precisely to those of exact marginalization. Correct marginalization, however, would accompany dependencies (potentials) of order higher than two but such differences are implicit in the graphical representation.

3.2.5 Exact computation

Each iteration in the recursive marginalization procedure yields an additional bound and consequently deteriorates the accuracy of the final bound. It is therefore necessary to stop introducing further bounds when the remaining undirected model becomes amenable to exact calculations. With exact calculations, we mean any method from decimation (see e.g. Saul & Jordan 1994) to the clustering algorithm (Lauritzen & Spiegelhalter 1988, Jensen et al. 1990) that does not involve approximations. The interface between the recursive bounds and exact algorithms is straightforward: we simply pass on the remaining potential to the exact algorithm. The variational parameters (distributions) introduced throughout the recursive marginalization can be optimized relative to the remaining exact calculations.

The approximate methods can be also used to facilitate the applicability of exact algorithms. For example, since graphically the simple lower bound recursion eliminates the variables from the model, we can use it to uncover feasible substructure. Selective use of such elimination can make exact algorithms quickly applicable. This objective is illustrated in figure 3-1. We note that this idea bears a connection to the structured mean field method studied by Saul & Jordan (1996). Their approach (non-recursive) suggests using exact methods for tractable substructures while applying mean field approximation for the variables mediating these structures. Translated into our framework, this would mean eliminating the mediating variables through the recursive lower bound transformation with a subsequent appeal to exact methods.

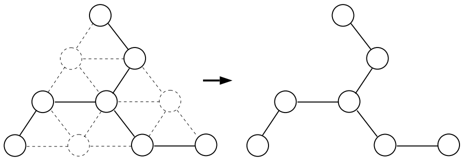

Figure 3-1: Enforcing tractable networks. Each variable in the graph can be removed (in any order). The elimination of the dotted nodes reveals a simplified graph underneath.

The real problem, however, is that of finding the key variables whose elimination would render the remaining graph tractable. This can be achieved, for example, as follows. A recursive elimination of all the variables from the network yields a variational expression for the partition function (the order of the recursions is immaterial, see the examples below). The correct marginalizations can be subsequently re-established by replacing the variational transformations with their associated exact marginalizations (i.e. moving recursively backwards). This procedure amounts to building amenable substructure from scratch. Look-ahead algorithms can be devised to guide the search towards larger but manageable structures. The details will be left for future work.

3.3 Examples

3.3.1 Boltzmann machines

We begin by considering Boltzmann machines with binary (0/1) variables. We assume that the joint probability distribution for the variables $S = \{S_1, \ldots, S_n\}$ is given by

$$P(S|h, J) = \frac{1}{Z(h, J)} e^{\,\phi(S|h,J)} \tag{3.13}$$

where h and J are the vector of biases and the weight matrix, respectively. The potential $\phi(S|h, J)$ in this setting consists of pairwise interaction terms (and biases)

$$\phi(S|h, J) = \sum_i h_i S_i + \frac{1}{2} \sum_{i,j} J_{ij} S_i S_j \tag{3.14}$$

where the weight matrix is symmetric and $J_{ii} = 0$. We assume that $J_{ij} = 0$ whenever there is no direct dependence between the variables S_i and S_j in the undirected graph. The partition function $Z(h, J) = \sum_S e^{\,\phi(S|h,J)}$ normalizes the distribution. The Boltzmann distribution defined in this manner is tractable insofar as we are able to compute the partition function; indeed, all marginal distributions can be reduced to ratios of partition functions in different contexts.

We now turn to methods for computing the partition function. This is exactly the problem considered earlier in recursive marginalization. We may therefore apply both the lower and upper bound recursive procedures to obtain lower and upper bounds on the desired partition function. Let us consider these in detail.

Lower bound recursion

As explained earlier, an approximate marginalization over S_k involves finding the effective potential $\tilde{\phi}(S \setminus S_k|\tilde{h}, \tilde{J})$ through variational transformations. In the context of Boltzmann machines, the transformation in Eq. (3.10) reduces to

$$\begin{aligned} \log \sum_{S_k \in \{0,1\}} e^{\,\phi(S|h,J)} &\geq \sum_{S_k} q(S_k)\phi(S|h,J) + H(q) &(3.15)\\ &= \sum_{i\neq k} h_i S_i + \frac{1}{2}\sum_{i,j\neq k} J_{ij} S_i S_j \\ &\quad + h_k q_k + \sum_{i\neq k} q_k J_{ki} S_i + H(q_k) &(3.16) \end{aligned}$$

where the second line corresponds to the terms that do not depend on S_k and therefore remain unchanged; in the average over the variational distribution we have used the simplified notation $q(S_k = 1) = q_k$. By collecting the weights and biases in this bound we find that the parameters of the effective potential $\tilde{\phi}(S \setminus S_k|\tilde{h}, \tilde{J})$ are given by

$$\begin{aligned} \tilde{h}_i &= h_i + q_k J_{ki} &(3.17)\\ \tilde{J}_{ij} &= J_{ij} &(3.18) \end{aligned}$$

for $i, j \neq k$. The effective potential also includes an additional constant factor $h_k q_k + H(q_k)$. It is evident that the effective potential remains in the class of Boltzmann machines, and the recursion can continue.

Let us analyze the recursion a bit further. First, as noted earlier, the approximate marginalization step does not change the interaction weights J_{ij}; it only affects the biases h_i. Marginalization in this approximation therefore leads to elimination with modified biases for the remaining variables. Second, the resulting effective potential is independent of the order in which the variables are marginalized over. To see this, observe first that the cumulating biases remain linear (i.e., are of the form $\tilde{h}_i = h_i + \sum_k q_k J_{ki}$ after any number of iterations) and are therefore unaffected by the ordering. It remains to show that the cumulating constant factor is order independent as well. After each recursion the emerging constant factor has the form $h_k^{k-1} q_k + H(q_k)$, where the notation h_k^{k-1} refers to the bias for the k^{th} variable after $k-1$ recursive marginalizations. The quantity of interest to us is the sum of these factors:

$$\begin{aligned} \sum_k \left\{\tilde{h}_k^{k-1} q_k + H(q_k)\right\} &= \sum_k \left\{\left(h_k + \sum_{k'<k} q_{k'} J_{k'k}\right) q_k + H(q_k)\right\} &(3.19)\\ &= \sum_k \{h_k q_k + H(q_k)\} + \sum_{k'<k} J_{k'k} q_{k'} q_k &(3.20)\\ &= \sum_k \{h_k q_k + H(q_k)\} + \frac{1}{2}\sum_{k',k} J_{k'k} q_{k'} q_k &(3.21) \end{aligned}$$

where the first equation follows from rewriting the recursion biases in terms of the original ones, and the last from the symmetry of the weight matrix J. In the final expression the order independence is evident. As a final observation, we claim that the recursive approximation is equal to ordinary mean field approximation for Boltzmann machines in the limit of applying the approximate marginalizations for all the variables in the model. In this limit, all that remains from the recursions are the constant factors introduced after each recursion step. The equation above collects all these factors assuming we extend the sum over k over all the variables. The result is exactly the mean field free energy for Boltzmann machines (see e.g. Hertz et al. 1991).

Upper bound recursion

We have already introduced the upper bound variational transformation applicable to Boltzmann machines. We consider this transformation and its properties here in more detail. Using the notation in terms of weights and biases, the upper bound from Eq. (3.11) becomes

$$\begin{aligned}\log \sum_{S_k \in \{0,1\}} e^{\phi(S|h,J)} \quad \leq \quad & \sum_{i \neq k} h_i S_i + \frac{1}{2} \sum_{i,j \neq k} J_{ij} S_i S_j + (h_k + \sum_j J_{kj} S_j)/2 \\ & \lambda_k \left(h_k + \sum_j J_{kj} S_j \right)^2 - f^*(\lambda_k) \qquad (3.22)\end{aligned}$$

where only the terms pertaining to S_k are affected. Recall that we have assumed that $J_{ij} = J_{ji} = 0$ whenever i and j are not adjacent in the original graph. The parameters $\tilde{h}_i$ and $\tilde{J}_{ij}$ characterizing the effective potential $\tilde{\phi}(S \setminus S_k|\tilde{h}, \tilde{J})$ can be found by rewriting the above bound in the form of Eq. (3.14). This gives

$$\begin{aligned}\tilde{h}_i &= h_i + J_{ki}/2 + 2\lambda_k h_k J_{ki} + \lambda_k J_{ki}^2 \qquad (3.23)\\ \tilde{J}_{ij} &= J_{ij} + 2\lambda_k J_{ik} J_{kj} \qquad (3.24)\end{aligned}$$

for $i, j \neq k$. The constant factor emerging from the transformation is given by $h_k/2 + \lambda_k h_k^2 - f^*(\lambda_k)$

Similarly to the lower bounds we make a few observations about the transformation. First, the new weight matrix $\tilde{J}_{ij}$ contains pairwise potentials connecting all the variables adjacent to S_k as in correct marginalization. We would therefore expect this transformation to yield more accurate bounds than the simpler lower bound. This will be quantified in the next section. Second, unlike with the simple lower bound transformation the recursion order matters in this case. Third, if we complete the recursive marginalization for all the variables in the model, then the resulting effective potential is the sum of the introduced constant factors:

$$\sum_k \left\{ h_k^{k-1} + \lambda_k (h_k^{k-1})^2 - f^*(\lambda_k) \right\} \qquad (3.25)$$

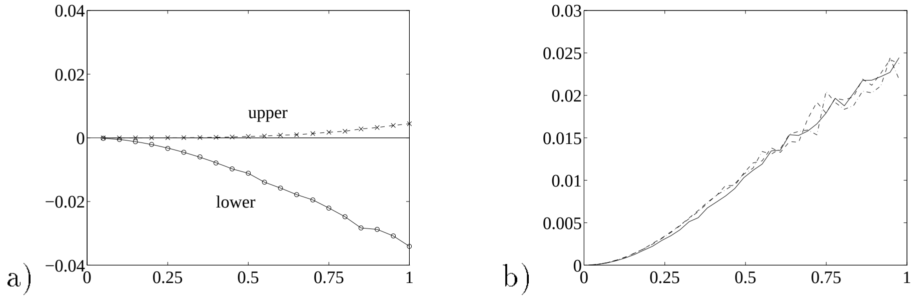

Figure 3-2: a) The mean relative errors in the log-partition function as a function of the scale of the random weights (uniform in $[-d, d]$). Solid line: lower bound recursion; dashed line: upper bound. b) Mean relative difference between the upper and lower bound recursions as a function of $d(n/8)^{1/2}$, where n is the network size. Solid: $n = 8$; dashed: $n = 64$; dotdashed: $n = 128$.

where h_k^{k-1} is the bias for the variable S_k after $k-1$ recursion steps. From the update equations above we see that the biases h_k^{k-1} remain linear functions of the original biases. The resulting potential (the equation above) is therefore only a quadratic function of the original biases. It is not feasible, however, to unravel the result in terms of the original weights J_{ij}. We conclude that the upper bound transformation is inherently recursive, quite unlike the lower bound method (see the previous section). We suspect that the lower bound approximation can be improved by making it more sophisticated to the extent that it remains at least recursively computable.

Accuracy

To substantiate the previously made claim that the upper bound recursion yields a more accurate bound we tested the recursions in randomly generated fully connected Boltzmann machines with 8 binary variables[5]. The weights in these models were chosen uniformly in the range $[-d, d]$ and all the initial biases were set to zero. Figure 3-2a plots the relative errors in the log-partition function estimates for the two recursions as a function of the scale d.

The upper and lower bound recursions would ideally be used in combination to yield interval estimates of the partition functions of interest. The tightness of such intervals as well as their scaling as a function of the network size is illustrated in figure 3-2b. The figure shows how the relative difference between the two bounds varies with the network size. In the illustrated scale the size has little effect on the difference. We note that the difference is mainly due to the lower bound recursion as is evident from figure 3-2a.

[5] The small network size was chosen to facilitate comparisons with exact results.

3.3.2 Sigmoid belief networks

Sigmoid belief networks are directed graphical models over binary variables for which the joint probability has the usual product form:

$$P(S) = \prod_i \frac{1}{Z_{i|S_{pa_i}}} e^{\,\phi_i(S_i|S_{pa_i})} = \prod_i e^{\,\phi_i(S_i|\,S_{pa_i})-\log Z_{i|S_{pa_i}}} \tag{3.26}$$

The potentials, however, take on a particular form

$$\phi_i(S_i|S_{pa_i}) = S_i(h_i + \sum_{j\in pa_i} J_{ij}S_j) \tag{3.27}$$

Consequently, the local partition functions are given by

$$Z_{i|S_{pa_i}} = \sum_{S_i\in\{0,1\}} e^{\,S_i(h_i+\sum_{j\in pa_i} J_{ij}S_j)} \tag{3.28}$$

Now, to transform this model into an undirected graph we compute the effective potentials corresponding to the log-partition functions via upper/lower bound variational transformations. These transformations have already been introduced in the context of recursive marginalization. From Eq. (3.10) and Eq. (3.11) we get

$$\begin{aligned} &\log \sum_{S_i\in\{0,1\}} e^{\,S_i(h_i+\sum_{j\in pa_i} J_{ij}S_j)} \\ &\qquad \geq\ q_i(h_i + \sum_{j\in pa_i} J_{ij}S_j) + H(q_i) && (3.29) \\ &\qquad \leq\ (h_i + \sum_{j\in pa_i} J_{ij}S_j)/2 + \lambda_i\,(h_i + \sum_{j\in pa_i} J_{ij}S_j)^2 - f^*(\lambda_i) && (3.30) \end{aligned}$$

where $q_i = q(S_i = 1)$. Inserting these back into the probability model of Eq. (3.26) gives upper and lower bounds on the joint distribution in terms of Boltzmann machines. The recursive marginalization procedure presented earlier for Boltzmann machines is therefore applicable and can be used to find the desired marginal probability. Note that in this conversion the upper/lower bounds on the log-partition function are reversed as bounds on the joint distribution (see Eq. (3.26)). The upper bound on the joint (corresponding to the simple lower bound) replaces the log-partition functions with linear combinations of existing potentials. Such removal of normalization amounts in graphical terms to simply substituting undirected links for the original directed ones. The lower bound on the joint, in contrast, introduces pairwise potentials between the parents of each variable. By connecting the parents this transformation therefore correctly moralizes the graph.

3.3.3 Chain graphs

The joint probability distribution for the binary chain graphs we consider here is given by

$$P(S_c|S_{pa_c}) = \frac{1}{Z_{S_{pa_c}}} e^{\,\phi(S_c|S_{pa_c})} = e^{\,\phi(S_c|S_{pa_c}) - \log Z_{S_{pa_c}}} \tag{3.31}$$

where the potentials $\phi(S_c|S_{pa_c})$ corresponding to each cluster of variables S_c have the following restricted form:

$$\begin{aligned} \phi(S_c|S_{pa_c}) &= \sum_{i\in c} \left\{ S_i h_i + S_i \sum_{j\in pa_c} J_{ij}^{c,out} S_j \right\} + \frac{1}{2} \sum_{i,j\in c} J_{ij}^c S_j S_j && (3.32) \\ &= \sum_{i\in c} S_i h_i^{eff} + \frac{1}{2} \sum_{i,j\in c} J_{ij}^c S_j S_j && (3.33) \end{aligned}$$

Here the weight matrix $J^{c,out}$ mediates the influence of outside cluster variables on the variables within the cluster c. The nature of this influence is to bias the within cluster variables while leaving the interaction structure intact. For this reason we have used the compact notation h_i^{eff} for the (effective) biases that contain the outside influence. As to the availability of our approximation framework it is important (currently) that h^{eff} depends linearly on the outside cluster variables.

To transform the chain graphs into undirected models we no longer can use the individual variational transformations to achieve the conversion. Instead, we need to consider the recursive framework for finding the effective potentials corresponding to the cluster log-partition functions $\log Z_{S_{pa_c}}$. These effective potentials consequently characterize the resulting undirected model. It is important, therefore, that these potentials will have a feasible dependence on the outside cluster variables S_{pa_c}. If this dependence is at most quadratic, then the recursive marginalization procedure for Boltzmann machines becomes applicable as with sigmoid networks. This will indeed be the case. To verify this, note first that the outside cluster variables can appear in the effective potentials only through h_i^{eff}, which are linear functions. We know from previous results (see section 3.3.1) that the effective potential from the lower bound recursion depends only linearly on the biases. The upper bound recursion, on the other hand, yields potentials that are at most quadratic in the biases. The desired property therefore holds.

In sum, we have shown how the joint distribution for chain graphs can be bounded by Boltzmann machines to which the recursive approximation formalism is again applicable.

3.4 Discussion

In this chapter we have developed a recursive node-elimination formalism for rigorously approximating intractable networks. The formalism applies to a large class of networks known as chain graphs and can be straightforwardly integrated with exact probabilistic calculations whenever they are applicable. Furthermore, the formalism provides rigorous upper and lower bounds on the desired quantities (e.g., the variable means).

The generality of our approach remains constrained, however, by the available transformations (upper bounds) and also by the assumptions needed to guarantee the feasible continuation of the recursion steps. The latter refers to the requirement for the effective potentials to remain in the same function class as the original potentials. The effects of this requirement can be mitigated by employing the methods of chapter 2, particularly in case of noisy-OR networks, to reduce the complexity of the original potentials.

We note finally that this chapter completes the development of generic variational methods for inference in graphical models (apart from the combination of exact and approximate methods for noisy-OR networks developed in chapter 5). The next chapter will shift the emphasis from inference to Bayesian estimation and does not rely on the results obtained so far, except for its use of the same variational transformations.

References

J. Besag (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statististical Society B* **2**:192-236.

J. Hertz, A. Krogh and R. Palmer (1991). *Introduction to the theory of neural computation.* Addison-Wesley.

C. Itzykson and J. Drouffe (1989). *Statistical field theory.* Cambridge University Press.

F. Jensen, S. Lauritzen, and K. Olesen (1990). Bayesian updating in causal probabilistic networks by local computations. *Computational Statistics Quarterly* **4**: 269-282.

S. Lauritzen (1996). *Graphical Models.* Oxford: Oxford University Press.

S. Lauritzen and D. Spiegelhalter (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society B* **50**:154-227.

T. Jaakkola and M. Jordan (1996). Mixture model approximations for belief networks. Manuscript in preparation.

R. Neal. Connectionist learning of belief networks (1992). *Artificial Intelligence* **56**: 71-113.

L. Saul and M. Jordan (1996). Exploiting tractable substructures in intractable networks. In *Advances of Neural Information Processing Systems 8.* MIT Press.

L. Saul and M. Jordan (1994). Learning in Boltzmann trees. *Neural Computation* **6** (6): 1174-1184.

J. Whittaker (1990). *Graphical models in applied multivariate statistics.* John Wiley & Sons.

3.A Lower bound

Consider the function

$$f(\vec{\phi}) = \log \sum_{S_k} e^{\,\phi(S)} \tag{3.34}$$

where $\vec{\phi}$ is a vector with components equal to $\phi(S)$ for different values of S_k while the remaining variables are kept fixed. We claim that $f(\vec{\phi})$ is a convex function of $\vec{\phi}$. This can be verified simply by computing the Hessian of f and observing that it is positive semi-definite (see Appendix 2.A), or alternatively, by noting that it is a conjugate function of the negative entropy function (see Appendix 1.A). According to section 1.1, f must have a representation of the form

$$f(\vec{\phi}) \;\; = \;\; \max_q \left\{ \, q^T \vec{\phi} - f^*(q) \, \right\} \tag{3.35}$$

The variational parameter q in this case is a probability distribution (gradient space of f consist of probability distributions) and the conjugate function is the negative entropy function $-H(q)$. By recalling the definition of $\vec{\phi}$ as $\phi(S)$ we get

$$f(\vec{\phi}) \;\; \geq \;\; \sum_{S_k} q(S_k) \phi(S) + H(q) \tag{3.36}$$

We have also (notationally) constrained the variational distribution q to depend only on S_k and not (as it should) on the remaining variables that were assumed to be fixed during the transformation. When the other variables are allowed to vary, this restriction amounts to an additional approximation, known as the mean field approximation.

3.B Upper bound

Here we derive a variational upper bound for

$$\begin{aligned} \log \sum_{S_k \in \{0,1\}} e^{\,\phi(S)} \;\; &= \;\; \log \left(e^{\,\phi_0} + e^{\,\phi_1} \right) && (3.37) \\ &= \;\; (\phi_0 + \phi_1)/2 + \log \left(e^{\,(\phi_1 - \phi_0)/2} + e^{\,-(\phi_1 - \phi_0)/2} \right) && (3.38) \end{aligned}$$

where we have used the notation $\phi_s = \phi(S)_{|S_k = s}$. For clarity let us define $x = \phi_1 - \phi_0$ and let

$$\tilde{f}(x) = \log \left(e^{\,x/2} + e^{\,-x/2} \right) \tag{3.39}$$

which is the function we need to find the transformation for. Now, $\tilde{f}(x)$ is a symmetric function of x, and a concave function of x^2. Let $f(x^2) = \tilde{f}(x)$, or, in words, f treats $\tilde{f}$ as a function of x^2. Since now $f(x^2)$ is a concave function of x^2, convex duality implies a representation of the form

$$\tilde{f}(x) = f(x^2) \;\; = \;\; \min_{\lambda} \Big\{ \lambda\, x^2 - f^*(\lambda) \Big\} \leq \lambda\, x^2 - f^*(\lambda) \tag{3.40}$$

where f^* is the conjugate function of f. To specify the form of the conjugate function, it is easier to reparameterize the bound in terms of tangent planes; see section 1.1 and Eq. (1.9) in particular. Take ξ^2 to be the new variational parameter indicating the location of the tangent, so that $\lambda = \lambda(\xi) = \nabla_{\xi^2} f(\xi^2)$, and find

$$f(x) \;\; \leq \;\; \lambda(\xi)\, x^2 - f(\xi^2) - \xi^2 \lambda(\xi) \tag{3.41}$$

where

$$\lambda(\xi) = \nabla_{\xi^2} f(\xi^2) = \nabla_{\xi^2} \tilde{f}(\xi) = \frac{1}{4\xi} \, \tanh(\xi/2) \tag{3.42}$$

Recall that the bound is exact whenever $\xi^2 = x^2$ (tangent defined at that point).

Chapter 4

Bayesian parameter estimation[1]

4.1 Introduction

The Bayesian formalism is well suited for representing uncertainties in the values of variables, model parameters, or in the model structure. The formalism further allows ready incorporation of prior knowledge and the combination of such knowledge with statistical data (Bernardo & Smith 1994, Heckerman et al. 1995). The rigorous semantics, however, often comes with a sizable computational cost of evaluating multidimensional integrals. This cost precludes the use of exact Bayesian methods even in relatively simple settings, such as generalized linear models (McCullagh & Nelder 1983). We concern ourselves in this chapter with a particular generalized linear model—logistic regression—and show how variational approximation techniques can restore the computational feasibility of the Bayesian formalism.

Variational methods should be contrasted with sampling techniques (Neal 1993) that have become standard in the context of Bayesian calculations. While surely powerful in evaluating complicated integrals, sampling techniques do not guarantee monotonically improving approximations nor do they yield explicit bounds. It is precisely these issues that are important in the current chapter.

The chapter is organized as follows. First we develop a variational approximation method that allows the computation of posterior distributions for the parameters in Bayesian logistic regression models. This is followed by a brief evaluation of the method's accuracy along with a comparison to other methods. We then extend the framework to belief networks, considering the case of incomplete data. Finally, we consider the dual of the regression problem – a density estimation problem – and show that our techniques lead to exactly solvable EM updates.

[1] This chapter is based on "T. Jaakkola and M. Jordan (1996). A variational approach to Bayesian logistic regression problems and their extensions. In *Proceedings of the sixth international workshop on artificial intelligence and statistics*".

4.2 Bayesian logistic regression

We begin with a logistic regression model given by

$$P(S|X,\theta) = g\left((2S-1)\theta^T X\right) \tag{4.1}$$

where $g(x) = (1+e^{-x})^{-1}$ is the logistic function, S the binary response variable, and $X = \{X_1, \ldots, X_n\}$ the set of explanatory variables. We represent the uncertainty in the parameter values θ via a prior distribution $P(\theta)$ which we assume to be a Gaussian with possibly full covariance structure. Our predictive distribution is therefore

$$P(S|X) = \int P(S|X,\theta)P(\theta)d\theta \tag{4.2}$$

In order to utilize this distribution we need to be able to compute the posterior parameter distribution $P(\theta|D^1, \ldots, D^T)$, where we assume that each $D^t = \{S^t, X_1^t, \ldots, X_n^t\}$ is a complete observation. To compute this posterior exactly, however, is not feasible[2]. It is nevertheless possible to find an accurate variational transformation of the conditional probability $P(S|X,\theta)$ such that the desired posterior can be computed in closed form. Let us next introduce the transformation and show how the posterior can be computed based on a single observation D. We will see that under the variational approximation the parameter posterior remains Gaussian, and thus the full posterior can be obtained by sequentially absorbing the evidence from each of the observations.

The variational transformation we use is given by (see Appendix 3.B)

$$\begin{aligned} P(S|X,\theta) = g(H_s) \;&\geq\; g(\xi)\, \exp\left\{(H_S-\xi)/2 - \lambda(\xi)(H_S^2-\xi^2)\right\} && (4.3)\\ &=\; P(S|X,\theta,\xi) && (4.4)\end{aligned}$$

where $H_S = (2S-1)\sum_j \theta_j X_j$ and $\lambda(\xi) = \tanh(\xi/2)/(4\xi)$. We may verify the transformation by a direct maximization with respect to the variational parameter ξ, and recover the original conditional distribution. The value of ξ at the maximum is simply H_S.

The posterior $P(\theta|D)$ can be computed by normalizing the left hand side of

$$P(S|X,\theta)P(\theta) \geq P(S|X,\theta,\xi)P(\theta) \tag{4.5}$$

Given that this normalization is not feasible in practice we normalize the variational distribution instead. The variational distribution has the convenient property that it depends on the parameters θ only quadratically in the exponent (eq. 4.4). Consequently, as the prior distribution is a Gaussian with mean μ and covariance matrix Σ, computing the variational posterior – absorbing the (variational) evidence – amounts to only updating the prior mean and the covariance matrix. Omitting the algebra

[2] Even without the prior distribution iterative schemes are need and such methods are intractable in our case.

this update yields

$$\begin{aligned}\Sigma_{post}^{-1} &= \Sigma^{-1} + 2\lambda(\xi)\, XX^T &(4.6)\\ \mu_{post} &= \Sigma_{post}\left[\Sigma^{-1}\mu + (S-1/2)X\right] &(4.7)\end{aligned}$$

where $X = [X_1 \ldots X_n]^T$. Now, the posterior covariance matrix depends on the variational parameter ξ through $\lambda(\xi)$ and thus its value needs to be specified. We obtain ξ by optimizing the approximation in eq. (4.5). Using the fact that the approximation is in fact a lower bound we may devise a fast EM algorithm to perform this optimization (see appendix 4.A). This leads to a closed form update for ξ given by

$$\xi^2 = E\left\{(\sum_j \theta_j X_j)^2\right\} = X^T\Sigma_{post}\, X + (X^T\mu_{post})^2 \qquad (4.8)$$

where the expectation is taken with respect to $P(\theta|D, \xi^{old})$, the variational posterior distribution based on the previous value of ξ. Alternating between the ξ update and those of the parameters monotonically improves the posterior approximation of eq. (4.5). The convergence of this procedure is very fast; roughly only two iterations are needed. The accuracy of the resulting variational approximation is considered in the next section.

In summary the variational approach allows us to obtain the posterior predictive distribution

$$P(S|X,\mathcal{D}) = \int P(S|X,\theta)P(\theta|\mathcal{D})d\theta \qquad (4.9)$$

where the posterior distribution $P(\theta|\mathcal{D})$ comes from sequentially absorbing each (complete) observation D^t in the data set $\mathcal{D} = \{D^1, \ldots, D^T\}$. The predictive likelihoods $P(S^t|X^t,\mathcal{D})$ for any complete observation D^t have the form

$$\log P(S^t|X^t,\mathcal{D}) = \log g(\xi_t) - \xi_t/2 + \lambda(\xi_t)\xi_t^2 - \frac{1}{2}\mu^T\Sigma^{-1}\mu + \frac{1}{2}\mu_t^T\Sigma_t^{-1}\mu_t + \frac{1}{2}\log\frac{|\Sigma_t|}{|\Sigma|} \quad (4.10)$$

where μ and Σ signify the parameters in $P(\theta|\mathcal{D})$ and the subscript t refers to the posterior $P(\theta|\mathcal{D}, D^t)$ found by absorbing the evidence in D^t.

4.3 Accuracy of the variational method

Figure 4-1a compares the variational form of eq. (4.4) to the logistic function for a fixed value of ξ (here $\xi = 2$). We note that this is the optimized variational approximation in cases where $E\left\{(\sum_j \theta_j X_j)^2 | \xi = 2\right\} = 2^2$ since this condition is the fixed point of the update equation (4.8).

To get an indication of the quality of the variational approximation in the context of Bayesian calculations we numerically computed the approximation errors in the simple case where there is only one explanatory variable and the observation is $D = \{S = 1, X = 1\}$. Figure 4-1b shows the accuracy of the variational predictive

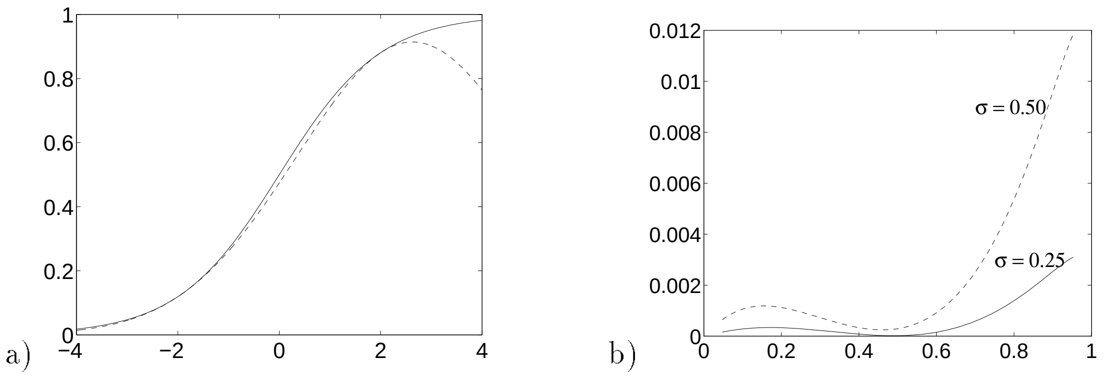

Figure 4-1: a) The logistic function (solid line) and its variational form (dashed line) when ξ is kept fixed at $\xi = 2$. b) The difference between the predictive likelihood $P(S = 1|X) = \int g(\theta)P(\theta)d\theta$ and its variational approximation as a function of $g(\mu)$; here $P(\theta)$ is Gaussian with mean μ and variance σ^2.

likelihood as a function of different prior distributions. The evaluation of the posterior accuracy is deferred to the next section where comparisons are made to other related methods. In practice, we expect the accuracy of the posterior to be more important than that of the predictive likelihood since errors in the posterior run the risk of accumulating in the course of the sequential estimation procedure.

4.4 Comparison to other methods

Other sequential approximation methods have been proposed to yield closed form posterior parameter distributions in logistic regression models. The most closely related appears to be that of Spiegelhalter and Lauritzen (1990) (referred to as the S-L approximation in this chapter). Their method is based on making a local quadratic approximation to the complete log-likelihood centered at the prior mean μ (also known as the Laplace approximation). Similarly to the variational updates of eq. (4.6-4.7), the S-L approximation changes the prior distribution according to

$$\Sigma_{post}^{-1} = \Sigma^{-1} + \hat{p}(1-\hat{p})\, XX^T \tag{4.11}$$

$$\mu_{post} = \mu + (S - \hat{p})\Sigma_{post}\, X \tag{4.12}$$

where $\hat{p} = g(\mu^T X)$. Since there are no additional adjustable parameters in this approximation, it is simpler than the variational method. For the same reason, however, it can be expected to yield less accurate posterior estimates.

We compared the accuracy of the posterior estimates in the simple case where there is only one explanatory variable $X = 1$. The posterior of interest was $P(\theta|S = 1)$, computed for various settings of the prior mean μ and standard deviation σ. The correct second order statistics for the posterior were obtained numerically. Figures 4-2 and 4-3 illustrate the accuracy of the posterior for the two approximation methods. We used simple (signed) errors in comparing the obtained posterior means to the correct ones; relative errors were used for the posterior standard deviations. The error measures were left signed to reveal any systematic biases. Based on figures 4-2a and 4-3a the variational method yields more accurate estimates of the posterior means. When the prior variance is small (figure 4-2b), the S-L estimate of the posterior variance appears to be at least as good as the variational estimate. For larger prior variances, however, the S-L approximation degrades more rapidly. We note that the variational method consistently underestimates the true posterior variance – a fact that could also have been predicted theoretically (and could be used to refine the approximation). Finally, in terms of the KL-divergences between the approximate and true posteriors, the variational method seems to (slightly) outperform the S-L approximation, again the more clearly the larger the prior variance. This is shown in Figure 4-4.

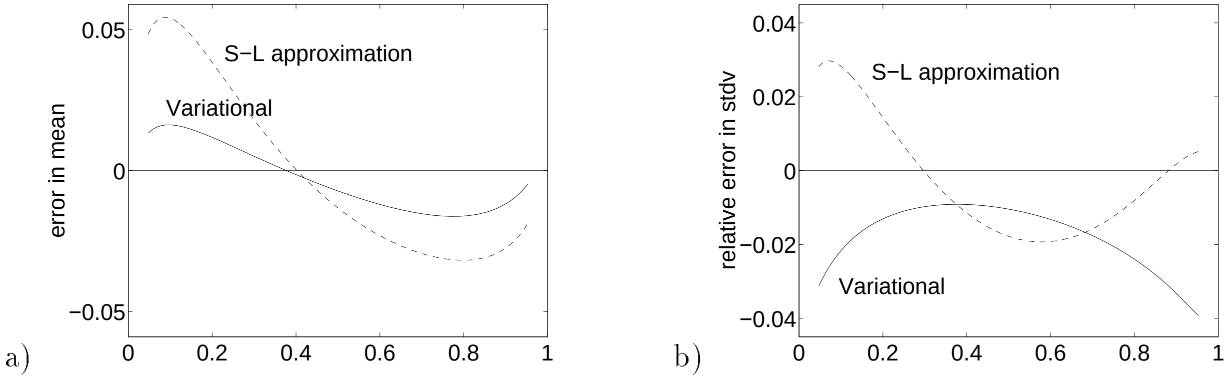

Figure 4-2: a) The errors in the posterior means as a function of $g(\mu)$, where μ is the prior mean. Here $\sigma = 1$ for the prior. b) The relative errors in the posterior standard deviations as a function of $g(\mu)$. $\sigma = 1$ for the prior distribution.

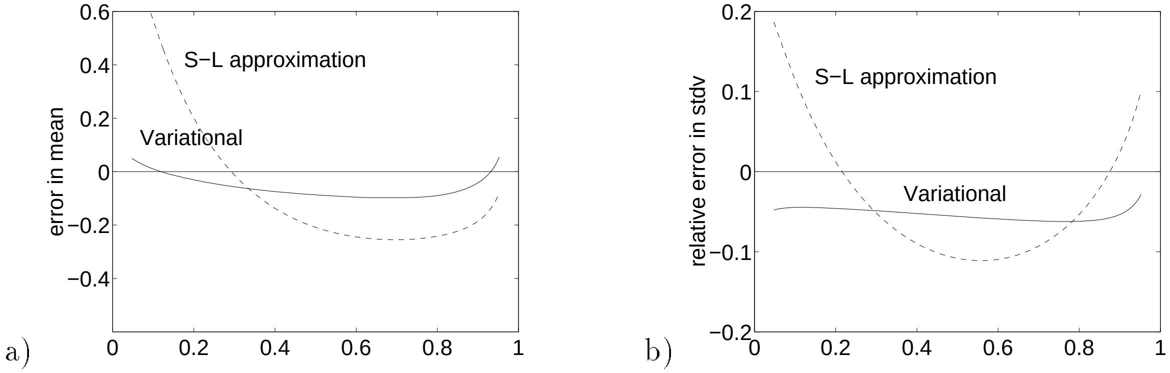

Figure 4-3: As figure 3 but now $\sigma = 2$ for the prior distribution.

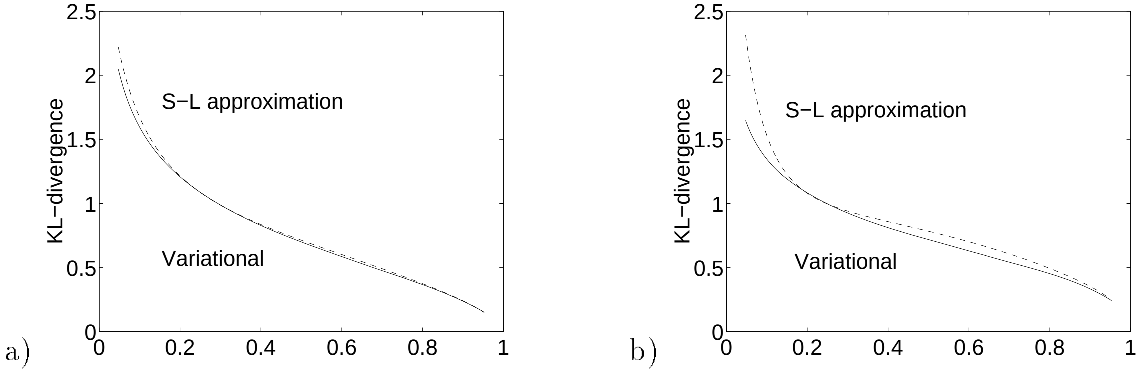

Figure 4-4: KL-divergences between the approximate and the true posterior distribution as a function of $g(\mu)$. a) $\sigma = 2$ for the prior. b) $\sigma = 3$. The two approximation methods have (visually) identical curves for $\sigma = 1$.

4.5 Extension to belief networks

A belief network can be constructed from logistic regression models that define conditional probabilities of a variable given its parents[3]. The predictive joint distribution for this belief network takes the usual product form

$$P(S_1, \ldots, S_n) = \prod_i P(S_i | S_{pa_i}) \tag{4.13}$$

where S_{pa_i} is the set of parents of S_i. We note that this is an extension of sigmoid belief networks (Neal 1992) due to the prior distributions over the parameters. In order to be able to use the techniques we have described for computing the posterior distributions on parameters in this setting, we assume first that the observations are complete, i.e., contain a value assignment for all the variables in the network. Complete observations contain all the Markov blankets[4] for the parameter distributions defining the conditional models (see figure 4-5a). Given value assignments for the variables in the Markov blanket, i.e., for the variables in the associated conditional model, the parameter distributions become independent of everything else in the network. Consequently, the posterior distributions over the parameters for each conditional model remain independent from each other as well. To estimate the parameters distributions based on complete observations therefore reduces to n independent subproblems of estimating each of the conditional regression models, which can be done as before.

4.5.1 Incomplete cases

[3] The sets of parents for the variables must be consistent with some global ordering of the variables.

[4] Markov blanket for a variable consists of any set of variables such that, conditionally on the variables in the Markov blanket, the variable in question becomes independent from the remaining variables in the probability model.

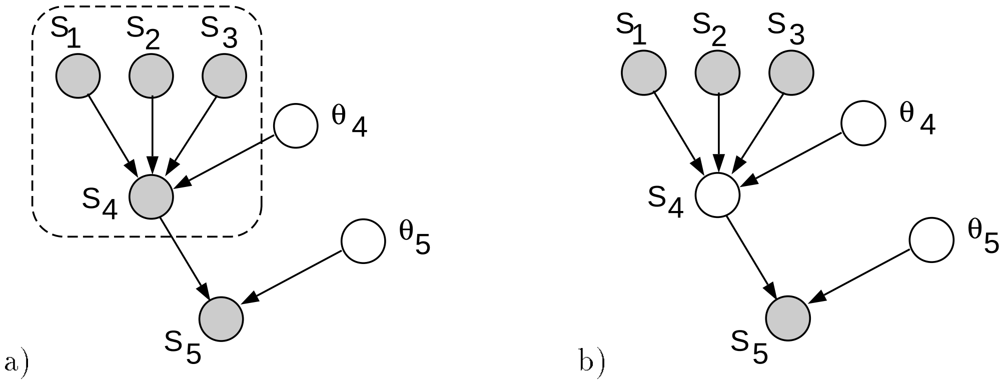

Figure 4-5: a) A complete observation (shaded variables) and the Markov blanket (dashed line) associated with the parameters θ_4. b) An observation where the value of S_4 is missing (unshaded in the figure).

In many practical situations the assumption of complete cases is quite unrealistic. The presence of missing values, however, means that we no longer have all the Markov blankets for the parameters in the network. Thus dependencies can arise between the parameter distributions across different conditional models. Let us consider this in more detail. A missing value implies that the observations pertain to the marginal distribution rather than the full joint distribution. The marginal distribution is obtained by summing or marginalizing over the possible missing values of the relevant variables, such as S_4 in figure 4-5b. This figure illustrates, for example, that the posterior distributions for the parameters θ_4 and θ_5 would depend on the value of S_4 had it been included in the observation. Thus when we marginalize over S_4, the posterior distribution over the parameters θ_4 and θ_5 becomes a mixture distribution, where each mixture component corresponds to a possible value of S_4, and the mixture weighting is according to the posterior distribution for S_4 (given the observations). In such a mixture posterior, the parameters associated with the two conditional models become dependent and the distribution will be multimodal. More precisely, the posterior distribution for these parameters is given by

$$\begin{aligned} P(\theta_4, \theta_5|\text{obs.}) &\propto \sum_{S_4} P(\text{obs.}, S_4|\theta_4, \theta_5) P(\theta_4) P(\theta_5) & (4.14) \\ &\propto \sum_{S_4} P(S_4|S_{pa_4}, \theta_4) P(S_5|S_{pa_5}, \theta_5) P(\theta_4) P(\theta_5) & (4.15) \end{aligned}$$

where $S_4 \in S_{pa_5}$, i.e, one of the parents of S_5. For a more thorough discussion of posterior parameter distributions see Spiegelhalter & Lauritzen (1990). We note here that the new dependencies arising from the missing values in the observations can make the network quite densely connected (a missing value effectively connects its neighboring nodes in the graph). The dense connectivity, on the other hand, leaves little structure to be exploited in the exact probabilistic computations in these networks and tends to make the model infeasible.

Importantly, the simplifying variational transformations introduced earlier for the logistic regression models do not remove the emerging dependencies among the posterior parameter distributions. Thus unlike in the regression case, where the posterior parameter distributions remained simple Gaussians, the use of these variational transformations with missing values yields not a single Gaussian posterior but a (large) mixture of them containing parameter distributions for other conditional models as well. To avoid such connectivity and multimodality, further approximations are necessary. What should the nature of these approximations be? Ideally, we would like to return to the setting of complete observations, where the parameter distributions can be maintained locally. Such a transition, however, would presume having "filled-in" the missing values. We propose to perform this fill-in in a principled manner by resorting to an additional variational technique known as the mean field transformation.

In the context of belief networks, mean field transformation can be characterized as a form of relaxed marginalization. The correct marginalization is a global operation in the sense that it affects synchronously all the conditional models that depend on the variable being marginalized over. Mean field, on the other hand, is a local operation by acting individually on the relevant conditional models. In other words, mean field approximation simply transforms separately all the conditional models in the belief network that pertain to the marginalized variable. Importantly, this locality preserves the factorization of the joint distribution as with complete observations thus performing the desired "filled-in" of the missing value. More precisely, the transformation of the (relevant) conditional models is given by

$$P(S|S_{pa},\theta) \rightarrow \prod_{S'} P(S|S_{pa},\theta)^{q(S')} \tag{4.16}$$

where S' is the variable with a missing value assignment and $q(S')$ is a mean field distribution over the missing values. The variational parameter in this transformation is the distribution q. The transformed conditional probabilities are thus geometrically averaged over the missing values (with respect to the mean field distribution). Note that if the particular conditional probability does not depend on the missing value the above transformation will not change it. While the transformations are carried out separately for each relevant conditional model, the mean field distribution associated with any particular missing value remains the same across such transformations. We note that the transformation does include a multiplicative constant that depends on q (see Appendix 4.B for more details).

The use of mean field with in maintaining the Bayesian parameter distributions is straightforward. When absorbing evidence from observations with missing values we first apply the mean field transformation to fill-in the missing values. The resulting joint distribution factorizes as with complete observations but now contains transformed conditional probabilities. The posterior parameter distributions therefore can be obtained independently for the parameters associated with different (transformed) conditionals. Two issues remain to be considered. First, the transformed conditional probabilities are now products of logistic functions and therefore more complicated than before. The introduction of the variational transformations for the logistic functions, however, turns them into exponentials with quadratic dependence on the parameters, and the product transformation of Eq. (4.16) retains this general form. Thus the evidence will again be "Gaussian" and if the prior is a multivariate Gaussian so will the posterior. The second issue is the dependence of the posterior parameter distributions on the mean field distribution q. The metric for optimizing q comes from the fact that mean field transformation yields a lower bound on the true marginalization. We therefore set q to the value that maximizes this lower bound. This optimization is carried out in conjunction with the optimization of the ξ parameters for the logistic transformations, which are also lower bounds. We can again devise an EM-algorithm to perform this maximization, the details of which are given in Appendix 4.B.3. The resulting sequential updating equations for the parameter distributions are similar to eq. (4.6-4.7):

$$\Sigma_{post_i}^{-1} = \Sigma_i^{-1} + 2\lambda(\xi_i)\, E\left\{S_{pa_i} S_{pa_i}^T\right\} \tag{4.17}$$

$$\mu_{post_i} = \Sigma_{post_i}\left[\, \Sigma^{-1}\mu_i + E\left\{(S_i - 1/2) S_{pa_i}\right\}\,\right] \tag{4.18}$$

where S_{pa_i} is the vector of parents of S_i, and the expectations are with respect to the mean field distributions. When the observations in the database are complete the expectations simply vanish.

4.6 The dual problem

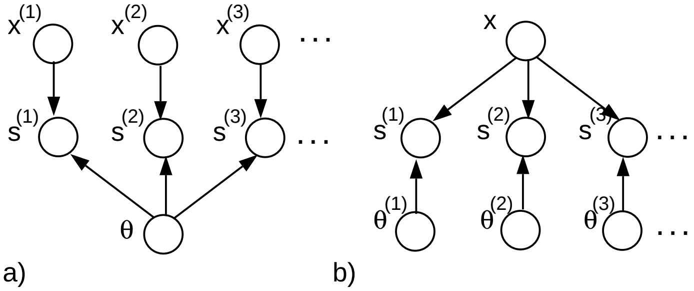

Figure 4-6: a) Bayesian regression problem. b) The dual problem.

The dual of the regression problem (eq. (4.1)) is found by switching the roles of the explanatory variables x and the parameters θ. In the dual problem, we have fixed parameters x and explanatory variables θ. Unlike before, distinct values of θ may explain different observations while the parameters x remain the same for all the observations, as shown in figure 4-6. In order to make the dual problem of figure 4-6b useful as a density model we generalize the binary output variables s to vectors $S = [S_1, \ldots, S_n]^T$ where each component S_i has a distinct set of parameters $X_i = [X_{i1} \ldots X_{im}]^T$ associated with it. The explanatory variables θ remain the same for all components. Consequently, the dual of the regression problem becomes a latent variable density model with a joint distribution given by

$$P(S_1, \dots, S_n|X) = \int \left[\prod_i P(S_i|X_i, \theta) \right] P(\theta) d\theta \tag{4.19}$$

where the conditional probabilities are logistic regression models

$$P(S_i|X_i, \theta) = g\left((2S_i - 1) \textstyle\sum_j X_{ij} \theta_j \right) \tag{4.20}$$

We would like to use the EM- algorithm for parameter estimation in this latent variable density model. To achieve this, we again make use of the variational transformations. These transformations are introduced for each of the conditional probabilities in the above joint distribution and optimized separately for the observations $D^t = \{S_1^t, \dots, S_n^t\}$ in the database (note that the observations now consist of only the values for the binary output variables). As in the logistic regression case, the transformations change the unwieldy conditional models into simpler ones that depend on the parameters only quadratically in the exponent (they become "Gaussian"). The variational evidence, which is a product of the transformed conditional probabilities, retains the same property. We are thus able to obtain the posterior distribution corresponding to such "Gaussian" evidence and, analogously to the regression case, the mean and the covariance of this posterior are given by

$$\Sigma_t^{-1} = \Sigma^{-1} + \sum_i 2\lambda(\xi_i^t)\, X_i X_i^T \tag{4.21}$$

$$\mu_t = \Sigma_t \left[\Sigma^{-1} \mu + \sum_i (S_i^t - 1/2) X_i \right] \tag{4.22}$$

The variational parameters ξ_i^t associated with each observation and the conditional model can be updated using eq. (4.8) assuming X_{pa} is replaced with X_i, the vector of parameters associated with the i^{th} conditional model. After we have computed the posterior distributions for all the observations in the data set, we can solve the M-step exactly. Omitting the algebra we get the following updates

$$\Sigma \leftarrow \frac{1}{T} \sum_t \Sigma_t \tag{4.23}$$

$$\mu \leftarrow \frac{1}{T} \sum_t \mu_t \tag{4.24}$$

$$X_i \leftarrow A_i^{-1} b_i \tag{4.25}$$

where

$$A_i = \sum_t 2\lambda(\xi_i^t)\, (\Sigma_t + \mu_t \mu_t^T) \tag{4.26}$$

$$b_i = \sum_t (S_i^t - 1/2) \mu_t \tag{4.27}$$

We note finally that since the variational transformations are lower bounds, these updates result in a monotonically increasing *lower bound* on the log-likelihood of the observations. This desirable monotonicity property is unlikely to arise with other types of approximation methods, such as the Laplace approximation.

4.7 Technical note: ML estimation

The standard maximum likelihood procedure for estimating the parameters in logistic regression uses an iterative Newton-Raphson method to find the parameter values. While the method is fast, it is not monotonic; i.e., the likelihood of the observations is not guaranteed to increase after an iteration. We show here how to derived a monotonic, fast estimation procedure for logistic regression by making use of the variational transformation in eq. (4.4). Let us denote $H_t = (2S^t - 1)\sum_j \theta_j X_j^t$ and write the log-likelihood of the observations as

$$\begin{aligned}\mathcal{L}(\theta) = \sum_t \log P(S^t|X^t,\theta) &= \sum_t \log g(H_t) \\ \geq & \sum_t \log g(\xi_t) + (H_t - \xi_t)/2 - \lambda(\xi_t)\left(H_t^2 - \xi_t^2\right) \\ = & \mathcal{L}(\theta,\xi) \qquad (4.28)\end{aligned}$$

The variational lower bound is exact whenever $\xi_t = H_t$ for all t. Although the parameters θ cannot be solved easily from $\mathcal{L}(\theta)$, $\mathcal{L}(\theta,\xi)$ allows a closed form solution for any fixed ξ, since the variational log-likelihood is a quadratic function of θ. The parameters θ that maximize $\mathcal{L}(\theta,\xi)$ are given by $\theta' = A^{-1}b$ where

$$A = \sum_t 2\lambda(\xi_t)\, H_t {H_t}^T \text{ and } b = \sum_t (S^t - 1/2)H_t \qquad (4.29)$$

Successively solving for θ and updating ξ yields the following chain of inequalities:

$$\mathcal{L}(\theta) = \mathcal{L}(\theta,\xi) \leq \mathcal{L}(\theta',\xi) \leq \mathcal{L}(\theta',\xi') = \mathcal{L}(\theta') \qquad (4.30)$$

where the prime signifies an update and we have assumed that $\xi_t = H_t$ initially. The combined update thus leads to a monotonically increasing likelihood. In addition, the closed form θ-updates make this procedure comparable in speed to the standard Newton-Raphson alternative.

4.8 Discussion

We have exemplified the use of variational techniques in a Bayesian estimation problem. We found that variational methods can be employed to obtain closed form expressions that approximate the posterior distributions for the parameters in logistic regression and associated belief networks even in the case of missing values. Furthermore, our variational techniques lead to an exactly solvable EM algorithm for a type of latent variable density model—the dual of the regression problem.

References

J. Bernardo and A. Smith (1994). *Bayesian theory.* New York: Wiley.

D. Heckerman, D. Geiger, and D. Chickering (1995). Learning Bayesian networks: the combination of knowledge and statistical data. *Machine Learning* **20** No. 3: 197.

J. Hertz, A. Krogh and R. Palmer (1991). *Introduction to the theory of neural computation.* Addison-Wesley.

P. McCullagh & J. A. Nelder (1983). *Generalized linear models.* London: Chapman and Hall.

R. Neal (1992). Connectionist learning of belief networks. *Artificial Intelligence* **56**: 71-113.

R. Neal (1993). Probabilistic inference using Markov chain Monte Carlo methods. Technical report CRG-TR-93-1, University of Toronto.

G. Parisi (1988). *Statistical field theory.* Addison-Wesley.

D. Spiegelhalter and S. Lauritzen (1990). Sequential updating of conditional probabilities on directed graphical structures. *Networks* **20**: 579-605.

4.A Optimization of the variational parameters

To optimize the variational approximation of eq. (4.5) in the context of an observation $D = \{S, X_1, \ldots, X_n\}$ we formulate an EM algorithm to maximize the predictive likelihood of this observation with respect to ξ. In other words, we find ξ that maximizes the right hand side of

$$\int P(S|X,\theta)P(\theta)d\theta \geq \int P(S|X,\theta,\xi)P(\theta)d\theta \tag{4.31}$$

In the EM formalism this is achieved by iteratively maximizing the expected complete log-likelihood given by

$$Q(\xi|\xi^{old}) = E\left\{\log P(S|X,\theta,\xi)P(\theta)\right\} \tag{4.32}$$

where the expectation is over $P(\theta|D,\xi^{old})$. Taking the derivative of Q with respect to ξ and setting it to zero leads to

$$\frac{\partial}{\partial\xi}Q(\xi|\xi^{old}) = -\frac{\partial\lambda(\xi)}{\partial\xi}\left[E\,(\sum_j \theta_j X_j)^2 - \xi^2\right] = 0 \tag{4.33}$$

As $\lambda(\xi)$ is a monotonically decreasing function[5] the maximum is obtained at

$$\xi^2 = E\,(\sum_j \theta_j X_j)^2 \tag{4.34}$$

By substituting ξ for ξ^{old} above, the procedure can be repeated. Each such iteration yields a better approximation in the sense of eq. (4.31).

4.B Parameter posteriors through mean field

Here we consider in detail the use of mean field in filling-in the missing values to facilitate the computation of posterior parameter distributions. We start by introducing the mean field approximation from the point of view that will be most convenient for our purposes. We note that several different formulations exist for mean field (Parisi 1988); the one presented here is tailored for our purposes.

4.B.1 Mean field

As mentioned in the text, mean field can be viewed as an approximation to marginalization. Consider therefore the problem of marginalizing over the variable S' when the joint distribution is given by

$$P(S_1,\ldots,S_n|\theta) = \prod_i P(S_i|S_{pa_i},\theta_i) \tag{4.35}$$

If we performed the marginalization exactly, then the resulting distribution would not retain the same factorization as the original joint (assuming S' is involved in more than one of the conditionals then) as can be seen from

$$\sum_{S'}\prod_i P(S_i|S_{pa_i},\theta_i) = \left[\prod_{i''} P(S_i|S_{pa_i},\theta_i)\right]\sum_{S'}\prod_{i'} P(S_i|S_{pa_i},\theta_i) \tag{4.36}$$

where we have partitioned the product over the conditionals according to whether they depend on S' (indexed by i') or not (indexed by i''). Marginalization is therefore not a local operation. Locality is a desirable property for computational reasons, however, and we attempt to preserve locality under approximate marginalization. The approximation we use for this purpose is a variational transformation based on the fact that any geometric average[6] is always less than or equal to the usual average. By applying this property to the marginalization, we find

$$\sum_{S'} P(S_1,\ldots,S_n|\theta) \;\; = \;\; \sum_{S'} q(S')\left[\frac{P(S_1,\ldots,S_n|\theta)}{q(S')}\right] \tag{4.37}$$

[5] This holds for $\xi \geq 0$. However, since $P(S|X,\theta,\xi)$ is a symmetric function of ξ, assuming $\xi \geq 0$ has no effect on the quality of the approximation.

[6] A geometric average over x_i, $i \in \{1,\ldots,n\}$ with respect to the distribution q_i is given by $\prod_i x_i^{q_i}$.

$$\geq \prod_{S'} \left[\frac{P(S_1, \ldots, S_n|\theta)}{q(S')} \right]^{q(S')} \tag{4.38}$$

$$= C(q) \prod_i \left[\prod_{S'} P(S_i|S_{pa_i}, \theta_i)^{q(S')} \right] \tag{4.39}$$

where the inequality comes from transforming the average over the bracketed term (with respect to the distribution q) into a geometric average. The third line follows from plugging in the form of the joint distribution and exchanging the order of the products. The multiplicative constant $C(q)$ relates to the entropy of the variational distribution q

$$C(q) = \prod_{S'} \left[\frac{1}{q(S')} \right]^{q(S')} \quad \text{and therefore} \quad \log C(q) = -\sum_{S'} q(S') \log q(S') \tag{4.40}$$

Let us now make a few observations about the result in Eq. (4.39). First, the choice of the variational distribution q always involves a trade-off between feasibility and accuracy. For example, if q were chosen to be the posterior distribution for S' given the remaining variables in the network, then the transformation would be exact rather than a lower bound. While attaining perfect accuracy, this choice would amount to no simplification. We emphasize feasibility and consequently assume that the variational distribution q depends only on the variable being marginalized over (i.e., S'). The resulting lower bound is the mean field approximation. Second, for any fixed q and for mean field in particular, the transformed marginalization (in Eq. (4.39)) follows the factorization of the original joint distribution. The conditional probabilities, however, have been modified in this approximation (mean field) according to

$$P(S_i|S_{pa_i}, \theta_i) \rightarrow \prod_{S'} P(S_i|S_{pa_i}, \theta_i)^{q(S')} \tag{4.41}$$

Unlike correct marginalization, this transformation of conditional probabilities is a local operation. Note that the transformation has no effect on the conditional probabilities that do not depend on S'. We note that the mean field distribution q is the same in all transformations corresponding to the same missing value.

Let us lastly explicate the case where the missing variable S' is in fact a set of variables, i.e., $S' = \{S'_1, \ldots, S'_m\}$. It might not be feasible in this case to allow the variational distribution q to vary independently for each configuration of the variables in S'; nor would it be a mean field approximation. Indeed, mean field approximation in this case would consist of successive applications of the conditional transformations for each of the variables in S'. It can be easily verified that this is equivalent to transforming the conditional probabilities for the whole set S' using a product or factorized distribution

$$q(S') = \prod_{i=1}^{m} q_i(S'_i) \tag{4.42}$$

4.B.2 Parameter posteriors

To fill-in the missing values in an observation $D_t = \{s_i^t, \ldots\}$ we use the mean field approximation. As a result, the joint distribution factorizes as with complete observations. Thus the posterior distributions for the parameters remain independent across the different conditional models and can be computed separately. Accordingly,

$$P(\theta_i|D_t, q) \propto \left[\prod_{S'} P(S_i|S_{pa_i}, \theta_i)^{q(S')} \right] P(\theta_i) \tag{4.43}$$

Even with the mean field approximation, however, this posterior remains at least as unwieldy as the Bayesian logistic regression problem considered earlier. Similarly to that case, we introduce the logistic transformations from Eq. (4.4) for the conditional probabilities and obtain

$$\begin{aligned}
P(\theta_i|D_t, q, \xi) &\propto \left[\prod_{S'} P(S_i|S_{pa_i}, \theta_i, \xi_i)^{q(S')} \right] P(\theta_i) && (4.44) \\
&= \left[\prod_{S'} \left\{ g(\xi)\, e^{(H_{S_i} - \xi)/2 - \lambda(\xi)(H_{S_i}^2 - \xi^2)} \right\}^{q(S')} \right] P(\theta_i) && (4.45) \\
&= \left[g(\xi)\, e^{\sum_{S'} q(S') \left\{ (H_{S_i} - \xi)/2 - \lambda(\xi)(H_{S_i}^2 - \xi^2) \right\}} \right] P(\theta_i) && (4.46) \\
&= \left[g(\xi)\, e^{(E\{H_{S_i}\} - \xi)/2 - \lambda(\xi)(E\{H_{S_i}^2\} - \xi^2)} \right] P(\theta_i) && (4.47) \\
&\equiv P(S_i|S_{pa_i}, \theta_i, \xi_i, q) P(\theta_i) && (4.48)
\end{aligned}$$

where $H_{S_i} = (2S_i - 1)\theta_i^T S_{pa_i}$, and S_{pa_i} is the vector of parents of S_i. The expectations are with respect to the mean field distribution q and, for simplicity, we have used the same variational parameter ξ for all the conditionals involved. The latter is naturally suboptimal but may be necessary if the number of missing values is large. Now, since H_{S_i} is linear in the parameters θ_i, the exponent in Eq. (4.47), consisting of averages over H_{S_i} and its square, stays at most quadratic in the parameters θ_i. This property implies that if the prior distribution is a multivariate Gaussian so will the posterior. The mean μ_{post_i} and covariance Σ_{post_i} of such posterior are given by (we omit the algebra)

$$\begin{aligned}
\Sigma_{post_i}^{-1} &= \Sigma_i^{-1} + 2\lambda(\xi_i)\, E\left\{ S_{pa_i} S_{pa_i}^T \right\} && (4.49) \\
\mu_{post_i} &= \Sigma_{post_i} \left[\Sigma_i^{-1} \mu_i + E\left\{ (S_i - 1/2) S_{pa_i} \right\} \right] && (4.50)
\end{aligned}$$

The expectations here are with respect to the mean field distribution q. Note that this posterior depends both on the distribution q and the parameters ξ. The optimization of these parameters is shown in Appendix 4.B.3.

4.B.3 Optimization of the variational parameters

We have introduced two variational "parameters": the mean field distribution q for the missing values, and the ξ parameters corresponding to the logistic transformations. The metric for optimizing the parameters comes from the fact that both of the transformations associated with these parameters introduce a lower bound on the likelihood of the observations. Thus by maximizing this lower bound we find the parameter values that yield the most accurate approximations. We therefore attempt to maximize the right hand side of

$$\begin{aligned}
\log P(D_t) &\geq \log P(D_t|\xi, q) && (4.51)\\
&= \log \int P(D_t|\theta, \xi, q) P(\theta) d\theta && (4.52)\\
&= \log \prod_i \int P(S_i|S_{pa_i}, \theta_i, \xi_i, q) P(\theta_i) d\theta_i + \log C(q) && (4.53)
\end{aligned}$$

where D_t is contains the observed variable settings. We have used the fact that the joint distribution in our approximation factorizes as with complete cases. Similarly to the case of Bayesian logistic regression considered previously, we can devise an EM-algorithm to maximize the variational log-likelihood of observations with respect to the parameters q and ξ; the parameters θ are considered as latent variables in this formulation. The E-step of the EM-algorithm, i.e., finding the posterior distribution over the latent variables, has already been described in Appendix 4.B.2. Here we will consider in detail only the M-step. For simplicity, we solve the M-step in two phases: one where the mean field distribution is kept fixed and the maximization is over ξ, and the other where these roles have been reversed. We start with the first phase.

As the variational joint distribution factorizes, the problem of finding the optimal ξ parameters separates into independent problems concerning each of the transformed conditional. Thus the optimization becomes analogous to the simple Bayesian logistic regression considered earlier. Two differences exist: first, the posterior over each θ_i is now obtained from Eq. (4.48); second, we have an additional expectation with respect to the mean field distribution q. As to the second difference, we note that the mean field expectations appear only in the exponents of the transformed conditionals (see Eq. (4.48) above). The significance of this feature lies in the fact that such expectations can always be incorporated into the E-step simply as additional averages. Reinterpreting the E-step in this manner reduces the problem to the simple regression explained earlier. For this reason we don't repeat the resulting EM-algorithm of Appendix 4.A here.

The latter part of our two-stage M-step is new, however, and will be considered in detail. The purpose here is to hold the parameters ξ fixed and instead optimize over q. In this case we do not include the mean field averages into the E-step, but more traditionally compute

$$\begin{aligned}
Q(q|q^{old}) &= E\{\log P(D_t, \theta|\xi, q)\} && (4.54)\\
&= \sum_i E_i \left\{\log P(S_i|S_{pa_i}, \theta_i, \xi_i, q) P(\theta_i)\right\} + \log C(q) && (4.55)
\end{aligned}$$

where the first expectation is with respect to $P(\theta|\xi, q^{old})$, which factorizes across the conditional models as explained previously; the expectations E_i are over the components $P(\theta_i|\xi_i, q^{old})$, obtained directly from Eq. (4.48). Let us now insert the form of the transformed conditional probabilities, $P(S_i|S_{pa_i}, \theta_i, \xi_i, q)$, into the above definition of the Q function. For clarity we will omit all the terms that have no dependence on the mean field distribution q as they will be irrelevant to our M-step. After some algebra we obtain:

$$\begin{aligned} Q(q|q^{old}) &= \sum_i E_i \Big\{ E_q\{H_{S_i}\}/2 - \lambda(\xi_i)\, E_q\{H_{S_i}^2\} \Big\} + \log C(q) + \dots \quad (4.56) \\ &= E_q \sum_i \Big\{ E_i\{H_{S_i}\}/2 - \lambda(\xi_i)\, E_i\{H_{S_i}^2\} \Big\} + H(q) + \dots \quad (4.57) \end{aligned}$$

where E_q refers to the expectation with respect to the mean field distribution q. The second equation follows by exchanging the order of the expectations E_i and E_q, which are mutually independent. We have also used the fact that $\log C(q)$ is the entropy $H(q)$ of the mean field distribution q (see Appendix 4.B above). Recall the notation $H_{S_i} = (2S_i - 1)\theta_i^T S_{pa_i}$, where S_{pa_i} is a binary vector of parents of S_i. Before proceeding to maximize the Q function with respect to q, we explicate the averages E_i in the above formula:

$$\begin{aligned} E_i\{H_{S_i}\} &= (2S_i - 1)\mu_{post_i}^T S_{pa_i} \quad (4.58) \\ E_i\{H_{S_i}^2\} &= \left(\mu_{post_i}^T S_{pa_i}\right)^2 + S_{pa_i}^T \Sigma_{post_i} S_{pa_i} \quad (4.59) \end{aligned}$$

Here μ_{post_i} and Σ_{post_i} are the mean and the covariance, respectively, of the posterior $P(\theta_i|\xi_i, q^{old})$ for the parameters of the i^{th} conditional model. Simply putting these back into the expression for the Q function we get

$$\begin{aligned} Q(q|q^{old}) = E_q \sum_i \Big\{ & (2S_i - 1)\mu_{post_i}^T S_{pa_i}/2 - \lambda(\xi_i) \left(\mu_{post_i}^T S_{pa_i}\right)^2 \\ & - \lambda(\xi_i)\, S_{pa_i}^T \Sigma_{post_i} S_{pa_i} \Big\} + H(q) + \dots \quad (4.60) \end{aligned}$$

Now, some of the binary variables s have a value assignment based on the observation D_t and the remaining variables will be averaged over the mean field distribution q. A reader familiar with Boltzmann machines will notice that since the binary variables appear only quadratically in the above formula, the Q function can be viewed as a mean field approximation for a type of Boltzmann machine (see e.g., Hertz et al. 1991). The (locally) maximizing distribution q can therefore be obtained through sequentially solving

$$\frac{\partial}{\partial q_k} Q(q|q^{old}) = 0 \quad (4.61)$$

for each of the component distributions (here $q_k = q_k(S_k = 1)$). The solutions for these can be found in closed form and each update leads to a monotonically increasing $Q(q|q^{old})$. The latter property follows from the fact $Q(q|q^{old})$ is a concave function of each of the component distributions (the expectation is linear and the entropy concave) implying a single attainable maximum. Note that the concavity property does not necessarily hold globally since the mean field expectations over the quadratic terms yield bi-linear forms with respect to these component distributions; we may not therefore be able to obtain the globally optimal mean field distribution.

Chapter 5

Variational methods in medical diagnosis

5.1 Introduction

The wealth and complexity of information in modern medicine underlies the tendency towards automated processing. In addition to low level processing aids, automated techniques can serve as powerful tools in more cognitive tasks such as in diagnosis. An example of such diagnostic tool, and the one considered in this chapter, is the Quick Medical Reference (QMR) knowledge base, compiled for internal medicine. The knowledge base embeds a combination of statistical and expert knowledge of about 600 significant diseases and their associated findings (about 4000). The intended use of this knowledge base is as an interactive decision tool for a practicing doctor in internal medicine. For the purpose of diagnosis, an early version of QMR included a heuristic reasoning mechanism for suggesting underlying diseases based on known findings (laboratory results or physicians' observations), or to outline informative tests to be carried out in order to verify a disease hypothesis, among other things. The reliance of heuristic tools for diagnostic reasoning, however, leaves much to be desired. The employed reasoning methods, for example, may be inconsistent and they can hide the assumptions underlying the success or failure of obtained diagnoses.

QMR-DT (Shwe et al. 1991), a decision theoretic reformulation of the QMR knowledge base, provides a more rigorous basis for automated inference. The relationships between the diseases and findings in this formulation take the form of a belief network (see e.g. Pearl 1988, Jensen 1996). Such probabilistic format enhances interpretability by allowing relevant assumptions to be explicated and, furthermore, endows the system with consistent rules of inference that are readily available for diagnostic (and other) decision making. The semantic rigor brought by the use of probability theory, however, often comes with a sizable computational cost of evaluating the conclusion implied by the probabilistic framework. Probabilistic inference, after all, has been shown to be an NP-hard problem in general (Cooper 1990). The size and complexity of the QMR belief network in particular appears to render exact evaluation methods inapplicable. To reap the benefits from the probabilistic formulation of the QMR network, we are forced to consider approximate methods for carrying out the inferences.

Several objectives can be set for approximate methods replacing exact calculations. Such methods first of all should be consistent in that, given enough computational resources, they would yield the result implied by the probabilistic model. Second, we should be able to assess the reliability of the results when such computational resources are not available. In the context of the QMR-belief network, Shwe et al. (1991) (see also Shwe & Cooper 1991) proposed the use of sampling techniques to obtain estimates of the posterior probabilities for the diseases (diagnoses in probabilistic terms). Sampling methods meet the first of our objectives since, in the limit of a large number of samples, the estimates will converge to the correct ones under mild conditions. The second objective, however, remains largely unattained by these techniques.

In this chapter we apply and extend an alternative approximation framework, that of variational methods, to compute the posterior probabilities for the diseases in the QMR belief network. Variational methods have a long history as approximation techniques in the physics literature. Unlike sampling methods variational techniques yield deterministic approximations that are adapted to each case separately. We consider here a particular type of variational methods, those most applicable to the QMR belief network. These techniques can be readily merged with exact evaluation methods in the QMR setting and therefore allowing us to let the available computational resources to determine the extent to which approximations are introduced. Towards the second objective for approximate methods, the variational methods yield explicit expressions for the posterior probabilities of the diseases. These expressions in turn can be subjected to further analysis concerning the accuracy and sensitivity to the various aspects of each case under consideration.

We start by defining the QMR belief network and the inference problem we are trying to solve through the use of variational methods. We then introduce and develop the variational techniques used in the chapter. Finally, we report numerical results on the accuracy and usefulness of variational techniques in diagnostic reasoning.

5.2 The QMR-DT belief network

The structure of the QMR-DT belief network currently conforms to the class of two-level or bi-partite networks (see figure 5-1). The diseases and findings in this model occupy the nodes on the two levels of the network, respectively, and the conditional probabilities specifying the dependencies between the levels are assumed to be noisy-OR gates. The bi-partite network structure encodes the assumption that, in the absence of findings, the diseases appear independently from each other with their respective prior probabilities (i.e. marginal independence). Also evident from the structure is that conditionally on the states of the diseases the findings are independent of each other (conditional independence). For a more thorough discussion regarding the medical validity of these and other assumptions embedded into the QMR-DT belief network, we refer the reader to Shwe et al. (1991).

To state more precisely the probability model implied by the QMR-DT belief

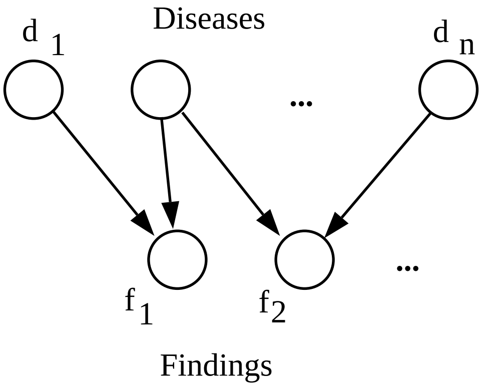

Figure 5-1: The QMR belief network is a two-level network where the dependencies between the diseases and their associated findings have been modeled via noisy-OR gates.

network, we write the joint probability of diseases and findings as

$$P(f,d) \;=\; P(f|d)P(d) = \left[\prod_i P(f_i|d)\right]\left[\prod_j P(d_j)\right] \tag{5.1}$$

where d and f are binary (1/0) vectors referring to presence/absence states of the diseases and the positive/negative states or outcomes of the findings, respectively. The conditional probabilities $P(f_i|d)$ for the findings given the states of the diseases, are assumed to be a noisy-OR models:

$$\begin{aligned} P(f_i = 0|d) &= P(f_i = 0|L) \prod_{j \in pa_i} P(f_i = 0|d_j) && (5.2) \\ &= (1 - q_{i0}) \prod_{j \in pa_i} (1 - q_{ij})^{d_j} = e^{-\theta_{i0} - \sum_{j \in pa_i} \theta_{ij} d_j} && (5.3) \end{aligned}$$

Here pa_i ("parents" of i) is the set of diseases pertaining to finding f_i. $q_{ij} = P(f_i = 0|d_j = 1)$ is the probability that the disease j, if present, could alone cause the finding to have a positive outcome (i.e. 1). $q_{i0} = P(f_i = 0|L)$ is the "leak" probability, i.e., the probability that the finding is caused by means other than the diseases included in the belief network model. The noisy-OR probability model encodes the causal independence assumption (Shwe et al., 1991), i.e., that the diseases act independently to cause the outcome of the findings. The exponentiated notation with $\theta_{ij} = -\log(1 - q_{ij})$ will be used later in the chapter for reasons of clarity.

5.3 Inference

To carry out diagnostic inferences in the QMR belief network involves computing posterior (marginal) probabilities for the diseases given a set of observed positive ($f_i = 1$) and negative ($f_{i'} = 0$) findings. Note that these sets are considerably smaller than the set of possible findings; the posterior probabilities for the diseases are affected only by findings whose states we have observed. For brevity we adopt the notation where f_i^+ corresponds to the event $f_i = 1$, and similarly f_i^- refers to $f_i = 0$ (positive and negative findings respectively). Thus the posterior probabilities of interest are $P(d_j|f^+, f^-)$, where f^+ and f^- are the vectors of positive and negative findings. The computation of these posterior probabilities exactly is in the worst case exponentially costly in the number of positive findings (Heckerman, 1988; D'Ambrosio 1994); the negative findings f^-, on the other hand, can be incorporated in linear time (in the number of associated diseases and in the number of negative findings). In practical diagnostic situations the number of positive findings often exceeds the feasible limit for exact calculations.

Let us consider the inference calculations more specifically. To find the posterior probability $P(d|f^+, f^-)$, we first absorb the evidence from negative findings, i.e., compute $P(d|f^-)$. This is just $P(f^-|d)P(d)$ with normalization. Since both $P(f^-|d)$ and $P(d)$ factorize over the diseases (see Eq. (5.2) and Eq. (5.1) above), the posterior $P(d|f^-)$ must factorize as well. The normalization of $P(f^-|d)P(d)$ therefore reduces to independent normalizations over each disease and can be carried out in time linear in the number of diseases (or negative findings). In the remainder, we will concentrate solely on the positive findings as they pose the real computational challenge. Unless otherwise stated, we will assume that the prior distribution over the diseases already contains the evidence from the negative findings. In other words, we presume that the updates $P(d_j) \leftarrow P(d_j|f^-)$ have already been made.

We now turn to the question about how to compute $P(d_j|f^+)$, the posterior marginal probability based on the positive findings. Formally, to obtain such a posterior involves marginalizing $P(f^+|d)P(d)$ over all the remaining diseases, i.e.

$$P(d_j|f^+) \propto \sum_{d\backslash d_j} P(f^+|d)P(d) \tag{5.4}$$

In the QMR belief network $P(f^+|d)P(d)$ has the form

$$P(f^+|d)P(d) = \left[\prod_i P(f_i^+|d)\right]\left[\prod_j P(d_j)\right] = \left[\prod_i \left(1 - e^{-\theta_{i0} - \sum_j \theta_{ij} d_j}\right)\right]\left[\prod_j P(d_j)\right] \tag{5.5}$$

which follows from the notation in Eq. (5.3) and the fact that $P(f_i^+|d) = 1 - P(f^-|d)$. To perform the summation in Eq. (5.4) over the diseases, we would have to multiply out the terms $1 - e^{\{\cdot\}}$ corresponding to the conditional probabilities for each positive finding. The number of resulting terms would be exponential in the number of positive findings and is not feasible. There is, however, local structure in the dependencies in the QMR belief network that can be exploited to speed up the computation (D'Ambrosio, 1994). For example, some findings are associated with only a few diseases. The gain from exploiting the local structure nevertheless does not appear to be sufficient to render the inference problem tractable in a practical setting.

5.4 Variational methods

5.4.1 A brief introduction

The variational approximation techniques we use and extend in this chapter are based on those developed in chapter 2 for noisy-OR networks. The extension considered in the current chapter pertains to the integration of these methods with exact probabilistic calculations rather than to the transformations themselves. For readers unfamiliar with the results of chapter 2 we have included a brief introductory presentation below. Other readers may wish to skip this material and commence from around Eq. (5.13).

The objective of our approximation methods is to simplify a complicated joint distribution such as the one in Eq. (5.5) above through variational transformations of conditional probabilities. The transformations themselves rerepresent the conditional probabilities in terms of optimization problems. Such representations are naturally turned into approximations by relaxing the optimizations involved. The fact that these approximations come from optimization problems implies that they have an inherent error metric associated with them, which is quite uncharacteristic of other deterministic or stochastic approximation methods. The use of this metric is to allow the approximation to be readjusted once they have been used, for example, in computation of marginal probabilities.

For the origin of the variational methods considered here we refer the reader to the introductory section 1.1 of chapter 1. For the purposes of this chapter, we recall that these methods transform any concave function f into an optimization problem by

$$f(x) = \min_{\xi}\{\, \xi^T x - f^*(\xi)\,\} \tag{5.6}$$

where f^* is the dual or conjugate function and ξ is a variational parameter. If we relax the minimization above and fix the the variational parameter ξ, we obtain a bound

$$f(x) \leq \xi^T x - f^*(\xi) \tag{5.7}$$

which is how the transformations are naturally used as approximations. We note that the reduction in complexity ensuing the substitution of the linear bound for f can be huge. What is gained in simplicity may, however, be lost in accuracy but this is not necessarily so. The loss in accuracy is contingent on the smoothness of f: the smoother f is, the more accurate the bound is.

5.4.2 Variational methods for QMR

Let us now return to the problem of computing the posterior probabilities in the QMR belief network. Recall that it is the conditional probabilities corresponding to the positive findings that need to be simplified. To this end, we write

$$P(f_i^+|d) \;=\; 1 - e^{-\theta_{i0} - \sum_j \theta_{ij} d_j} = e^{\log(1-e^{-x})} \tag{5.8}$$

where $x = \theta_{i0} + \sum_j \theta_{ij} d_j$. Consider the exponent $f(x) = \log(1 - e^{-x})$. For noisy-OR, as well as for many other conditional models involving compact representations (e.g. logistic regression), the exponent $f(x)$ is a concave function of x. Based on the introduction we know that there must exist a variational upper bound for this function that is linear in x:

$$f(x) \leq \xi x - f^*(\xi) \tag{5.9}$$

The conjugate function $f^*(\xi)$ for noisy-OR is given by (see chapter 2 for a derivation)

$$f^*(\xi) = -\xi \log \xi + (\xi + 1) \log(\xi + 1) \tag{5.10}$$

The desired bound or simplification of the noisy-OR conditional probabilities is found by putting the above bound back into the exponent (and recalling the definition $x = \theta_{i0} + \sum_j \theta_{ij} d_j$):

$$\begin{aligned}
P(f_i^+|d) &= e^{f(x)} \qquad\qquad &(5.11)\\
&\leq e^{\xi_i(\theta_{i0} + \sum_j \theta_{ij} d_j) - f^*(\xi_i)} &(5.12)\\
&= e^{\xi_i \theta_{i0} - f^*(\xi_i)} \prod_j \left[e^{\xi_i \theta_{ij}} \right]^{d_j} &(5.13)\\
&\equiv P(f_i^+|d, \xi_i) &(5.14)
\end{aligned}$$

where we have rewritten the bound as a product over the associated diseases to make explicit the fact that it factorizes over such diseases. Importantly, any evidence possessing this factorization can be absorbed efficiently (in time and space) just as with negative findings. Thus unlike the correct evidence $P(f_i^+|d)$ from the positive findings, the "variational" evidence $P(f_i^+|d, \xi_i)$ can be incorporated efficiently into the posterior.

We are now ready to outline the variational approximation framework for obtaining efficient estimates of the posterior marginal probabilities for the diseases. The first step is to reduce the complexity of handling the positive findings by introducing the transformations

$$P(f_i^+|d) \rightarrow P(f_i^+|d, \xi_i) \tag{5.15}$$

Not all the positive findings need to be transformed, however, and we use these transformations only to the extent that is necessary to reduce the computational load to a manageable (or practical) level. The posterior estimates can be consequently obtained from the transformed probability model.

Two issues need to be clarified for this framework. The posterior estimates will depend on the variational parameters ξ which we need to set and adjust to the current diagnostic context. This issue is resolved in Appendix 5.A; the adjustment of the variational parameters reduces to convex optimization that can be carried out efficiently (there are no local minima). The second issue is the question of which conditional probabilities (or positive findings) should be transformed and which left unchanged. This will be considered next.

The order of transformations

The decision to transform or treat exactly any of the conditional probabilities $P(f_i^+|d)$ corresponding to the positive findings is a trade-off between efficiency and accuracy. To maintain a maximal level of accuracy while not sacrificing efficiency, we introduce the transformations by starting from the conditional for which the variational form is the most accurate and proceed towards less accurate transformations. When it is manageable to treat the remaining conditionals exactly we stop introducing any further transformations. How then do we measure the accuracy of the transformations? The metric for this comes from the fact that the transformations are indeed bounds.

Each transformation introduces an upper bound on the exact conditional probability. Thus the likelihood of the observed (positive) findings $P(f^+)$ is also upper bounded by its variational counterpart $P(f^+|\xi)$:

$$P(f^+) = \sum_d P(f^+|d)P(d) \quad \leq \quad \sum_d P(f^+|d,\xi)P(d) = P(f^+|\xi) \tag{5.16}$$

The better the variational approximations are, the tighter this bound is. We can assess the accuracy of each variational transformation as follows. First we introduce and optimize the variational transformations for all the positive findings. Then for each positive finding we replace the variational transformation with the exact conditional and compute the difference between the corresponding bounds on the likelihood of the observations:

$$\delta_i = P(f^+|\xi) - P(f^+|\xi \setminus \xi_i) \tag{5.17}$$

where $P(f^+|\xi \setminus \xi_i)$ is computed without transforming the i^{th} positive finding. The larger the difference δ is, the worse the i^{th} transformation is. We should therefore introduce the transformations in the ascending order of δs. Put another way, we should treat exactly the findings for which δ is large.

Figure 5-2 illustrates the significance of using the proposed ordering for introducing the variational transformations as opposed to a random ordering. The two plots correspond to representative diagnostic cases, and show the log-likelihoods for the observed findings as a function of the number of positive findings that were treated exactly. We emphasize that the plots are on a log-scale and therefore the observed differences are huge. We also note that the curves for the proposed ordering are convex, i.e., the bound improves the less the more findings have already been treated exactly. This is because the exact conditionals first replace the worst transformations and the differences among the better transformations are smaller. It is for this reason that we can expect the variational posterior estimates to converge close to the true values after a reasonably small fraction of the positive findings have been treated exactly. We finally note that the δ measure for determining the ordering favors variational transformations for conditional probabilities that are diagnostically the least relevant. This is because the variational transformations are more accurate for positive findings that are not surprising, i.e., are likely to occur, or when there is less impetus for explaining them (the leak probability is large). See Appendix 5.B for details.

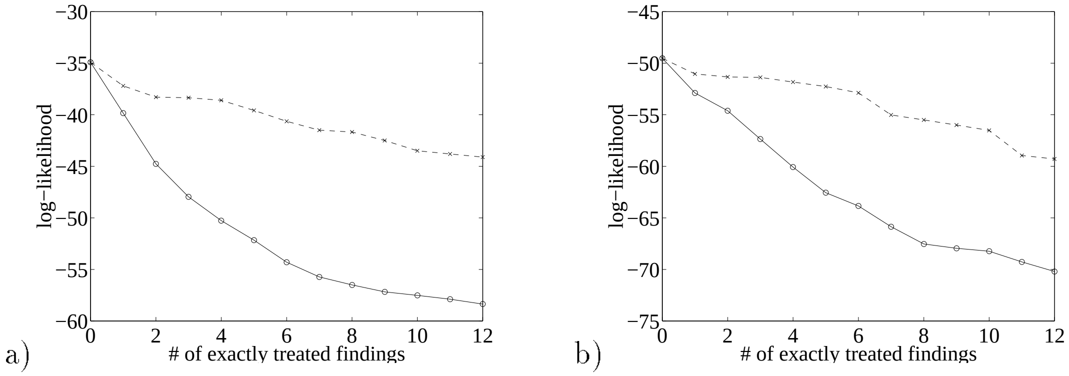

Figure 5-2: The log-likelihood of the observed findings as a function of the number of positive findings treated exactly. The solid line corresponds to the proposed ordering and the dashed line is for a random ordering.

5.5 Towards rigorous posterior bounds

The variational methods we have described earlier provide *upper* bounds on the likelihood of the observed findings. Other variational methods can be used to obtain *lower* bounds on the same likelihoods. The ability to obtain upper and lower bounds on the likelihoods provides the means for achieving such bounds on the desired posterior marginals as well. To see why, note first that by Bayes' rule the posterior marginals can be written in terms of likelihoods:

$$P(d_j = 1|f^+) = \frac{P(d_j = 1, f^+)}{P(d_j = 1, f^+) + P(d_j = 0, f^+)} \tag{5.18}$$

where, for example, $P(d_j = 1, f^+)$ is the joint likelihood of the findings and the disease j to be present. Let us assume that there are methods available for finding likelihood bounds and use the notation $\underline{P}(\cdot)$ for lower bounds and analogously $\overline{P}(\cdot)$ for upper bounds. It follows from standard results (see e.g. Draper, 1994) that the posterior disease marginals in this case are bounded by

$$P(d_j = 1|f^+) \;\; \geq \;\; \frac{\underline{P}(d_j = 1, f^+)}{\underline{P}(d_j = 1, f^+) + \overline{P}(d_j = 0, f^+)} \tag{5.19}$$

$$P(d_j = 1|f^+) \;\leq\; \frac{\overline{P}(d_j = 1, f^+)}{\overline{P}(d_j = 1, f^+) + \underline{P}(d_j = 0, f^+)} \tag{5.20}$$

Note that both upper and lower bounds are needed to obtain either an upper or lower bound on the posterior disease marginal. The techniques described in Jaakkola & Jordan (1996a) provide the missing lower bounds on the marginal likelihoods in the QMR setting. While these methods can produce quite accurate lower bounds for many networks, the specifics of the QMR belief network nevertheless precludes their use. The reason for this discrepancy comes from the mathematical expansions that these lower bounds rely on: for very small leak probabilities such expansions become unreliable. This is indeed the case with the QMR belief network whose leak probabilities are typically quite small (can be as low as 10^{-7}).

We introduce here an alternative technique towards obtaining rigorous bounds on the posterior marginals. Although we only use the upper bound variational technique, we make a more extensive use of its approximation properties. To this end, we adopt the notation $r(d_j = 1, f^+)$ to denote the ratio of the variational upper bound to the correct marginal likelihood, i.e.,

$$r(d_j = 1, f^+) = \frac{\overline{P}(d_j = 1, f^+)}{P(d_j = 1, f^+)} \tag{5.21}$$

where $r(d_j = 0, f^+)$ is defined analogously. We may now write the correct marginal likelihoods as $\overline{P}(d_j = 1, f^+)/r(d_j = 1, f^+)$. Substituting these for the correct likelihoods in Eq. (5.18) and multiplying by the ratio $r(d_j = 1, f^+)$ gives

$$P(d_j = 1|f^+) \;=\; \frac{\overline{P}(d_j = 1, f^+)}{\overline{P}(d_j = 1, f^+) + \overline{P}(d_j = 0, f^+)\frac{r(d_j=1,f^+)}{r(d_j=0,f^+)}} \tag{5.22}$$

$$\geq\; \frac{\overline{P}(d_j = 1, f^+)}{\overline{P}(d_j = 1, f^+) + \overline{P}(d_j = 0, f^+)} \tag{5.23}$$

where the inequality follows from the conjecture (partially proved in Appendix 5.B) that the approximation in terms of the ratio is better for the case $d_j = 1$. The key here is that the variational parameters are optimized individually to $\overline{P}(d_j = 1, f^+)$ and $\overline{P}(d_j = 0, f^+)$; otherwise, Eq. (5.23) would correspond exactly to the variational posterior estimates discussed earlier and figures 5-3 and 5-4 below would refute the conjecture in that regard. Now, the effect of the instantiation $d_j = 1$ is to increase the leak probabilities selectively for its associated positive findings (see the probability model). In Appendix 5.B we show that an increase in the leak probability (or the bias term θ_{i0}) for a single finding necessarily makes the variational approximation for that finding more accurate in terms of the ratio. We emphasize that the proof presented in the appendix is not sufficient to warrant the inequality in Eq. (5.23) and the question remains open for future work.

5.6 Results

The diagnostic cases that we used in evaluating the performance of the variational techniques were cases abstracted from clinicopathologic conference (abbrev. CPC) cases. These are considered clinically more difficult cases involving multiple diseases underlying the observed findings. These are also cases in which Middleton et al. (1991) did not find their importance sampling method to work satisfactorily. Four of the 48 CPC-cases included in our evaluation turned out to have a sufficiently small number of positive findings ($\leq$ 20) to allow an exact computation of the posterior marginals for the purposes of comparison[1]. We begin assessing the quality of the variational estimates from these cases. For the remaining cases, we don't have an exact reference posterior distribution to compare against; alternative measures of accuracy will be considered.

5.6.1 Comparison to exact posterior marginals

Here we have chosen out of the CPC-cases those that have positive findings less than or equal to 20 so as to be able to evaluate the correct posterior marginals. Table 5.1 contains a description of these "admissible" cases.

case	# of pos. findings	# of neg. findings
1	20	14
2	10	21
3	19	19
4	19	33

Table 5.1: Description of the cases for which we evaluated the correct posterior marginals.

Figures 5-3 and 5-4 illustrate the correlation between the approximate and the true posterior marginals. If the approximate marginals were in fact correct then the points in the figures should align along the diagonals as shown by the dotted lines. The plots are obtained by first extracting the 10 highest posterior marginals from each (admissible) case and then computing the approximate posterior marginals for the corresponding diseases. In the approximate solutions we varied the number of positive findings that were treated exactly in order to elucidate the rate by which the approximate marginals approach the correct ones. Figure 5-5 reveals more quantitatively the rate of convergence of the posterior marginals. The plots show the fraction of all posterior marginal estimates (10 largest from each admissible case) whose error exceeds the specified threshold as a function of the number of positive findings that were treated exactly. We may loosely interpret these level curves as probabilities that,

[1] One of the cases with $\leq$ 20 positive findings had to be excluded due to vanishing numerical precision in the exact evaluation of the corresponding posterior marginals.

in a hypothetical case, the error in a posterior marginal estimate would exceed the specified limit. Figure 5-5a is in terms of the relative error in the posterior marginals; figure 5-5b on the other hand uses the absolute error.

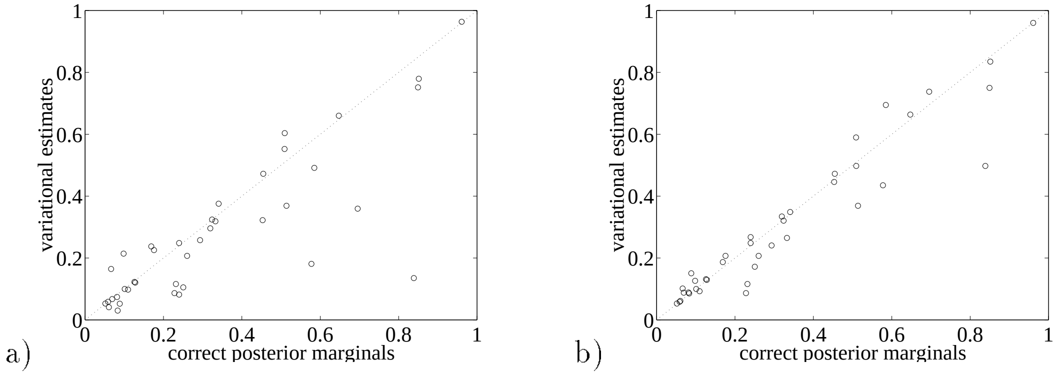

Figure 5-3: Correlation between the variational posterior estimates and the correct marginals. In a) 4 and in b) 8 positive findings were treated exactly.

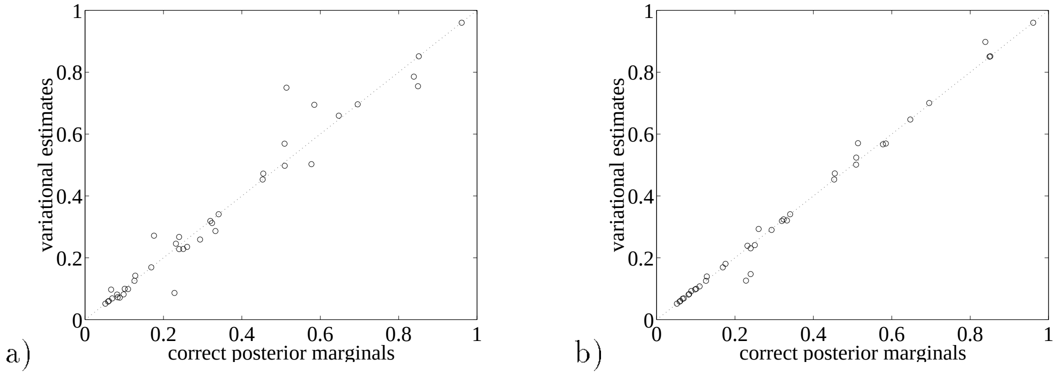

Figure 5-4: Correlation between the variational posterior estimates and the correct marginals. In a) 12 and in b) 16 positive findings were treated exactly.

5.6.2 Comparison to Gibbs' sampling estimates

In this section we will compare the accuracy of the variational posterior estimates to those obtained through stochastic procedures. Our goal here is merely to demonstrate that the variational methods are competetive with sampling techniques. For this reason, we chose to use a straightforward Gibbs' sampling technique in our comparisons. We emphasize, however, that many (often superior) sampling methods are available (see Shwe & Cooper 1991, Neal 1993) and our results should be viewed in this light.

In our Gibbs' sampling method the posterior disease marginals were obtained from

$$\hat{P}(d_i) = \frac{1}{T}\sum_t P(\, d_i \mid f,\, d^t \setminus d^t_i \,) \tag{5.24}$$

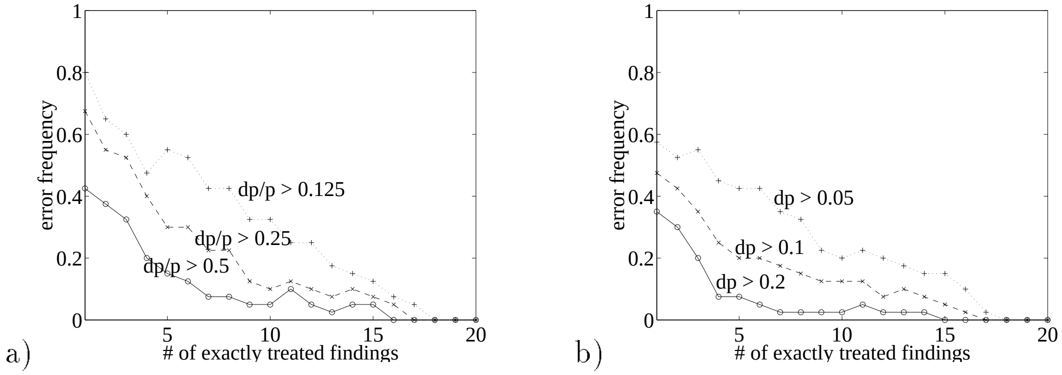

Figure 5-5: The fraction of posterior marginal estimates exceeding the specified error limits as a function of the number of positive findings that were treated exactly. The error measures used were a) the relative error, and b) the absolute error.

where each new disease configuration d^t was computed from the previous one d^{t-1} through sequentially resampling the disease states with replacement. The order for the updates was chosen randomly at each stage. For good performance only every 5^{th} such d^t sample was included in the above sum. The initial configuration d^0 was drawn from the prior distribution over the diseases[2]. While discarding (a lot of) early samples is generally profitable, it seemed to only deteriorate the results in our case. The accuracy gained through the use of later samples was offset by the loss in computation time spent for discarding the early samples (cf. the time/accuracy plot of figure 5-6 below). Consequently no early samples were excluded.

To be able to reliably assess the accuracy of the posterior estimates we used the four tractable cases described in the previous section. Figure 5-6 shows the mean correlations (across these admissable cases) between the approximate estimates and the correct posterior marginals as a function of the execution time needed for computing the estimates. The correlation measures for the stochastic method were averaged across 20 independent runs for each admissable case, and across these cases for the final measure. The error bars in the figure were obtained by averaging the standard deviations computed for each admissable case from the 20 different runs; the error bars therefore reflect how much the correlations would typically vary over several runs on the same case, i.e., they capture the repeatability of the stochastic estimates. Note that the variational estimates are deterministic and vary only across cases. We conclude from the figure that more sophisticated sampling techniques are needed to achieve a level of performance comparable to variational methods.

5.6.3 Posterior accuracy across cases

In the absence of the exact posterior marginals for reference, we have to find a way to assess the accuracy of the estimated posterior marginals. We perform this assessment through a measure of variability of these marginals. Recall first that in the varia-

[2] The most likely initial configuration was therefore the one with all the diseases absent.

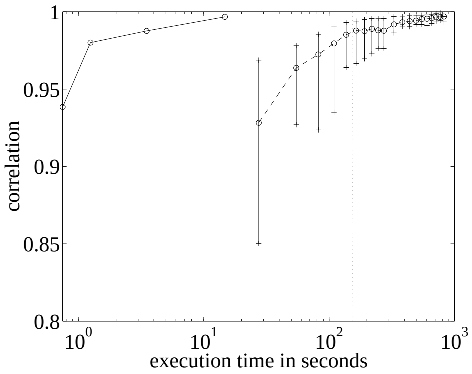

Figure 5-6: The mean correlation between the approximate and exact posterior marginals as a function of the execution time (seconds). Solid line: variational estimates; dashed line: Gibbs' sampling. The dotted line indicates the average time for the exact calculation of the posterior marginals in the admissable cases.

tional approximation some of the conditional probabilities for the positive findings are treated exactly while the remaining conditionals are replaced with their variational counterparts. The posterior marginals will naturally depend on which conditionals received exact treatment and which were approximated. Conversely, a lack of such dependence implies that we have the correct posterior marginals. We can therefore use this dependence as a way to assess the validity of the current posterior estimates. Let $\hat{P}_i(d_i = 1)$ be the i^{th} largest posterior marginal probability based on the variational method, and let $\hat{P}_i^{+k}(d_i = 1)$ be a refined estimate of the same marginal, where the refinement comes from treating the k^{th} positive finding exactly. As a measure of accuracy of the posterior estimates we use the variability of $\hat{P}_i^{+k}(d_i = 1)$ around $\hat{P}_i(d_i = 1)$, where k varies in the set of positive findings whose conditional probabilities have been transformed in obtaining $\hat{P}_i(d_i = 1)$. Several definitions can be given for this variability and we consider two of them below.

Mean squared variability

For each disease we define the variability of its posterior probability estimates according to

$$\hat{\sigma}_i{}^2 = \frac{1}{K}\sum_k \left(\hat{P}_i(d_i = 1) - \hat{P}_i^{+k}(d_i = 1)\right)^2 \tag{5.25}$$

which is the mean squared difference between $\hat{P}_i(d_i = 1)$ and its possible refinements $\hat{P}_i^{+k}(d_i = 1)$. The sum goes over the positive findings for which we have introduced a variational transformation in computing $\hat{P}_i(d_i = 1)$. As an overall measure of variability for any particular diagnostic case, we use

$$\hat{\sigma} = \max_{i \le 10} \hat{\sigma}_i \tag{5.26}$$

The decision to include only 10 largest posterior marginals is inconsequential but convenient. We note that the $\hat{\sigma}$ measure is scale dependent, i.e., it assigns a higher variability to the same relative difference when the probabilities involved are larger. The measure therefore puts more emphasis on the posterior marginals that are likely to be diagnostically most relevant.

Before adopting the variability measure $\hat{\sigma}$ for further analysis we will first provide some empirical justification for it. To this effect, we can use the admissible CPC cases considered in section 5.6.1 for which the exact posterior disease marginals can be computed. We would expect the variability measure to reflect the true mean squared error between the variational posterior estimates and the correct posterior marginals. As shown in figure 5-7, the correlation between these measures is indeed quite good. Naturally, only the four reference cases included in the figure are not sufficient to establish the credibility of the variability measure but they are indicative. The a) and b) parts of the figure, respectively, correspond to treating 8 and 12 findings exactly.

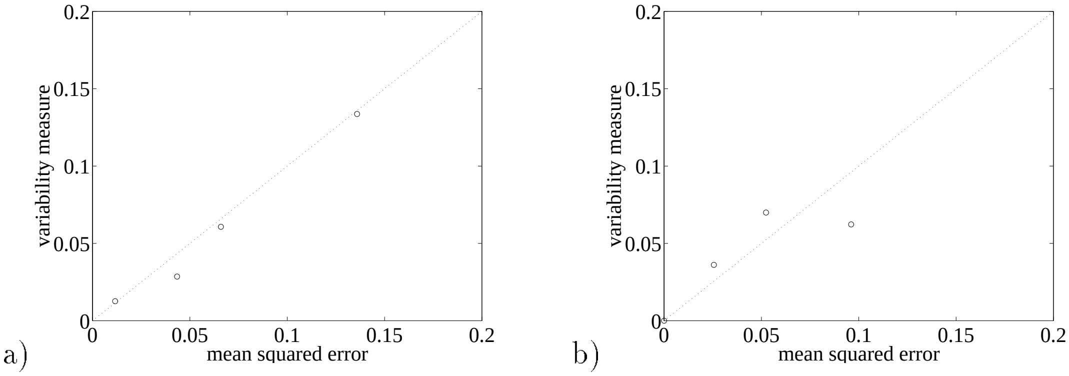

Figure 5-7: The correlation between $\hat{\sigma}$ and σ, where $\hat{\sigma}$ is the variability measure and σ is the true mean squared error between the variational estimates and the correct marginals. 10 most likely posterior marginals were included from each admissible case. The number of exactly treated findings was 8 in figure a) and 12 in b).

Figure 5-8 illustrates how the accuracy or the variability $\hat{\sigma}$ of the posterior estimates depends on the number of positive and negative findings across the CPC-cases. Eight conditional probabilities were treated exactly in each of the CPC-cases. Figure 5-9 is analogous except that the number of exactly treated findings was 12. As expected, the variational approximation is less accurate for larger numbers of positive findings (see the regression lines in the figures). Since the number of exactly treated findings (or conditionals) was fixed, the more positive findings a case has, the more variational transformations need to be introduced. This obviously deteriorates the posterior accuracy and this is seen in the figures. The figures also seem to indicate that the variational approximations become better as the number of negative findings increases. This effect, however, has to do with the scale dependent measure of accuracy. To see why, note first that the negative findings reduce the prior probabilities for the diseases, although somewhat selectively. Smaller prior probabilities nevertheless decrease the posteriors marginals. The scale dependent $\hat{\sigma}$ therefore decreases without any real improvement of the variational accuracy. The figure 5-9 is included in comparison to indicate that indeed the error measure is consistently lower when more findings have been treated exactly. We note finally that the squared error measure for the posterior marginals is generally quite small; large deviations from the true marginals are therefore rare.

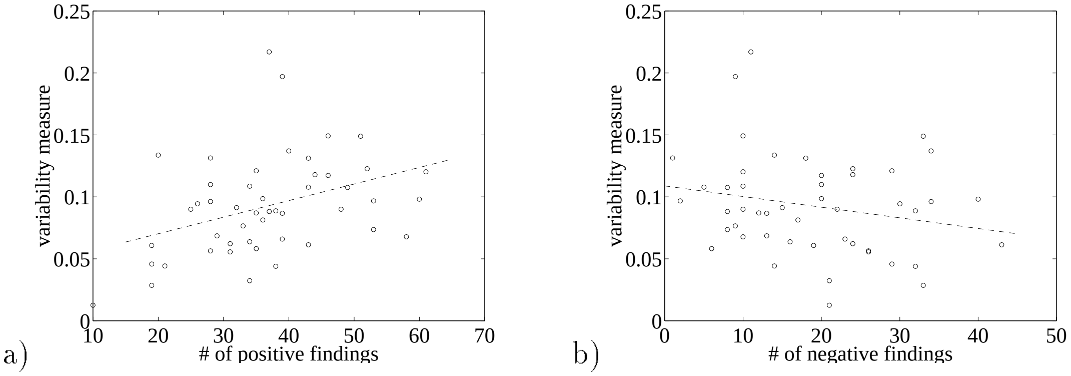

Figure 5-8: a) The variability $\hat{\sigma}$ of the posterior marginals as a function of the number positive findings in the CPC-cases. b) The same variability measure $\hat{\sigma}$ but now as a function of the number of negative findings. 8 positive findings were treated exactly for this figure.

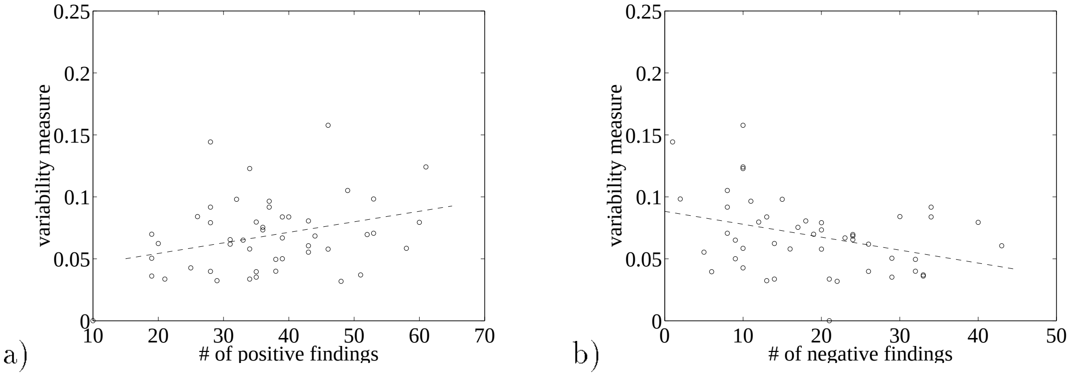

Figure 5-9: a) The variability $\hat{\sigma}$ of the posterior marginals as a function of the number positive findings in the CPC-cases. b) The same variability measure $\hat{\sigma}$ but now as a function of the number of negative findings. Now 12 positive findings were handled exactly.

Min/Max variability across cases

While the squared error captures the mean variability in the posterior estimates, it is important to find out the limits of how much the true posterior marginals could deviate from our estimates. We use the bounds $\min_k \hat{P}_i^{+k}(d_i = 1)$ and $\max_k \hat{P}_i^{+k}(d_i = 1)$

as indicators of this deviation. While these variability bounds do not provide rigorous bounds on the posterior disease marginals they nevertheless come close to doing so in practice. To substantiate this claim, we used the four CPC cases considered in section 5.6.1. Figure 5-10 illustrates the accuracy of these bounds for 10 most likely posterior marginals from each of the four cases. Although few of the posterior marginals do fall outside of these bounds, the discrepancies are quite minor and even in these cases the bounds provide an indication of the direction of the error between the correct and the estimated posterior marginals. Moreover, when the bounds are tight, the true posterior marginals appear within or very close to the bounds.

While the bounds provide a measure of accuracy for individual posterior estimates, we employ a correlation measure to indicate the overall accuracy. In other words, we use the correlation between the variational posterior estimates and the min/max bounds of the refined marginals as the overall measure. A high degree of correlation indicates that the posterior probabilities are very accurate; otherwise, at least one of the positive findings should influence the refined posterior marginals and consequently the bounds thereby deteriorating the correlation. Recall that each positive finding is treated exactly in one of the refined marginals. We note that this correlation measure is deterministic.

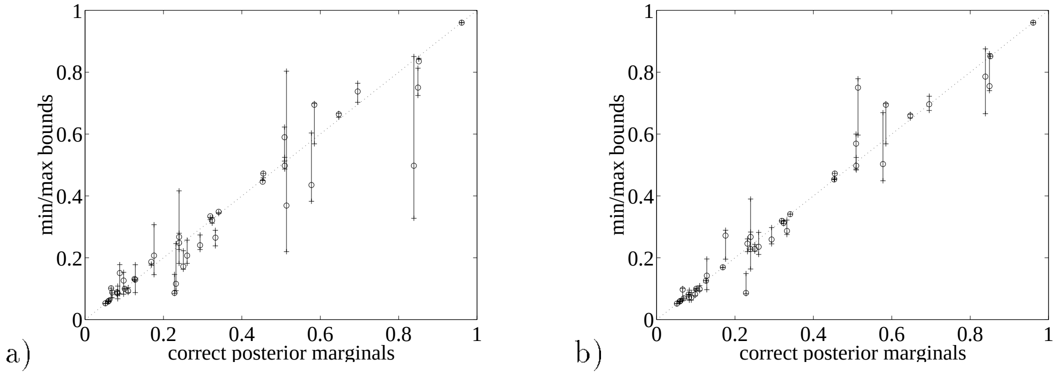

Figure 5-10: The correlation between the min/max bounds and the true posterior marginals. 10 most likely posterior marginals were included from each admissible case. In a) 8 findings were treated exactly and in b) 12.

Figure 5-11a illustrates the correlation between the variational posterior marginals and the min/max bounds of the refined marginals for the CPC cases. The correlation coefficients when 8 findings were treated exactly for each diagnostic case were 0.953/0.879 between the approximate marginals and those of the min/max bounds, respectively. When 12 findings were included exactly these coefficients rose to 0.965/0.948 (see figure 5-11b). The dependence of the correlation coefficients on the number of exactly treated positive findings is illustrated in figure 5-12a. The high monotonic increase in the correlation is mainly due to the proper ordering of the findings to be treated exactly (see section 5.4.2). In comparison, figure 5-12b shows the development of the correlations for a random ordering.

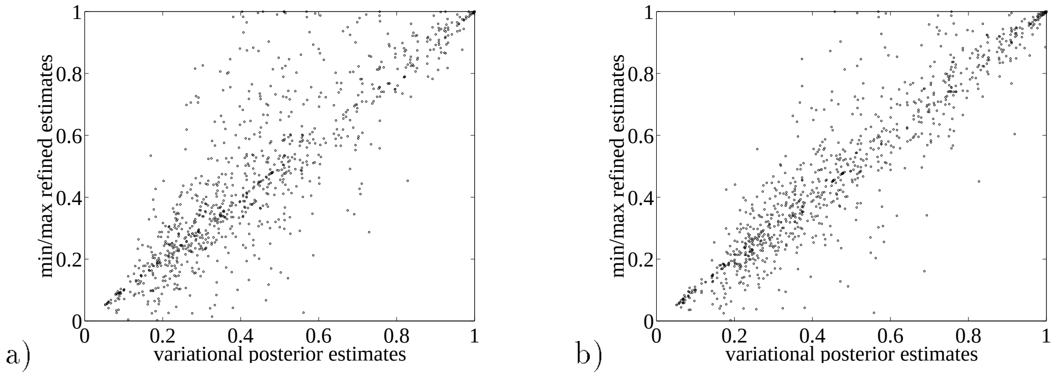

Figure 5-11: a) Correlation between the estimated posterior marginals and the min/max refined marginals. There were 8 positive findings considered exactly. b) as before but now the number of findings treated exactly was 12.

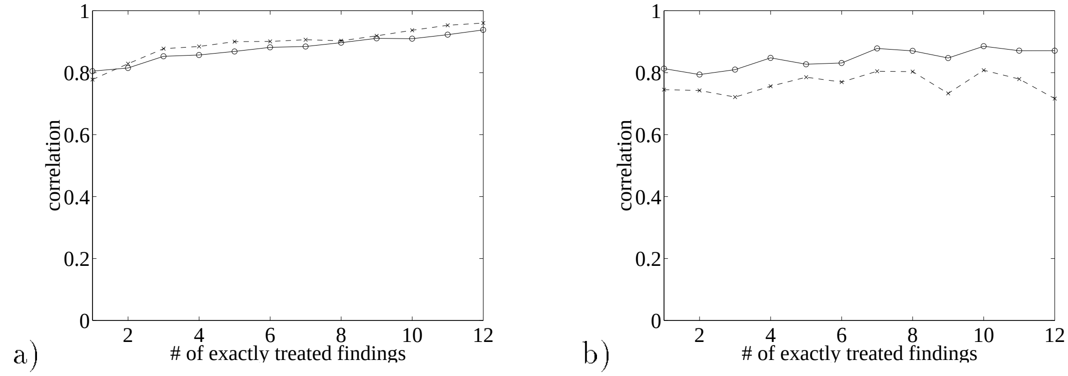

Figure 5-12: Mean correlation between the approximate posterior marginals and the min/max bounds as a function of the number of positive findings that were treated exactly. Solid line: correlation with the max- bound; dashed line: correlation with the min- bound. Figure a) is for the case where a proper ordering was used to select the findings to be treated exactly and in b) a random ordering was used.

5.6.4 Execution times

We report here additional results on the feasibility of the variational techniques in terms of computation time. The results were obtained on a Sun Sparc 10 workstation.

The execution times for obtaining the variational posterior estimates are overwhelmingly dominated by the number of exactly treated positive findings, i.e., the amount of exact probabilistic calculations. The time needed for the variational optimization using the shortcut discussed at the end of Appendix 5.A is insignificant. The maximum time across the CPC-cases was less than 2 seconds whenever at most 12 positive findings were treated exactly.

More relevant in a practical setting is the combined time of obtaining the posterior estimates and carrying out the verifying analyses as described in the previous sections. Figure 5-13 plots the mean and the maximum execution times across the CPC-cases for such analyses as a function of the number of positive findings that were treated exactly. The mean time when 12 findings were treated exactly was about 1 minute and the maximum about 2 minutes.

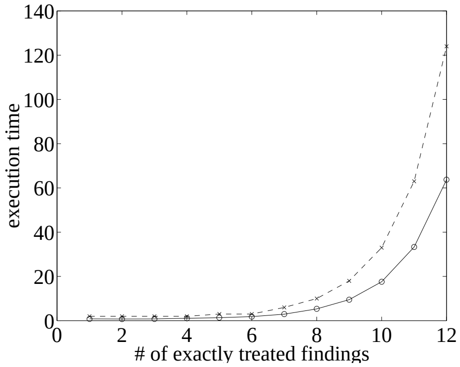

Figure 5-13: Mean (solid line) and maximum (dashed line) execution times in seconds for the analyses of section 5.6.2 across the CPC-cases as a function of the number of exactly treated positive findings.

5.7 Discussion

The benefits of probabilistic models and reasoning comes from the rigorous semantics that underlies them. Increasingly complex probabilistic models such as the QMR-DT belief network, however, necessitate the use of approximate techniques in evaluating the probabilities or quantities implied by the probabilistic framework. We have demonstrated in this chapter that variational methods can be employed as viable alternatives to sampling techniques that are more traditionally used towards such goals. We maintain, however, that our results are preliminary and resort to one out of many possible techniques. For example, the feasibility of obtaining rigorous bounds on the posterior disease marginals through variational methods was only partially exploited in the current work. Such bounds become perhaps even more important in fully decision theoretic systems, towards which QMR-DT is being developed. This chapter therefore characterizes merely the potential of variational methods for probabilistic reasoning in important application domains such as in medicine.

Acknowledgments

We would like to thank the University of Pittsburgh and Randy Miller for the use of the QMR database.

References

B. D'Ambrosio (1994). Symbolic probabilistic inference in large BN20 networks. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence.* Morgan Kaufmann.

J. Bucklew (1990). *Large deviation techniques in decision, simulation, and estimation.* John Wiley & Sons.

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

F. Jensen (1996). *Introduction to Bayesian networks.* Springer.

B. Middleton, M. Shwe, D. Heckerman, M. Henrion, E. Horvitz. H. Lehmann, G. Cooper (1991). Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base: Part-II. Stanford University technical report.

R. Neal (1993). Probabilistic inference using Markov chain Monte Carlo methods. Technical report CRG-TR-93-1, University of Toronto.

M. Shwe and G. Cooper (1991). An empirical analysis of Likelihood – weighting simulation on a large, multiply connected medical belief network. *Computers and Biomedical Research* **24**:453-475.

M. Shwe, B. Middleton, D. Heckerman, M. Henrion, E. Horvitz. H. Lehmann, G. Cooper (1991). Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base: Part-I. *Methods of Information in Medicine* **30**: 241-255.

5.A Optimization of the variational parameters

The metric for optimizing the variational parameters comes from the bounding properties of the individual variational transformations introduced for the conditional probabilities. Each transformation is an upper bound on the corresponding conditional and therefore also the resulting joint distribution is an upper bound on the true joint; similarly all marginals such as the likelihood of the positive findings that are computed from the new joint will be upper bounds on the true marginals. Thus

$$P(f^+) = \sum_d P(f^+|d)P(d) \quad \leq \quad \sum_d P(f^+|d,\xi)P(d) = P(f^+|\xi) \tag{5.27}$$

and we may take the accuracy of $P(f^+|\xi)$, the variational likelihood of observations, as a metric. To simplify the ensuing notation we assume that the first m of the positive findings have been transformed (and therefore need to be optimized) while the remaining conditional probabilities will be treated exactly. In this notation $P(f^+|\xi)$

is given by

$$P(f^+|\xi) = \sum_d \left[\prod_{i\leq m} P(f_i^+|d,\xi_i)\right]\left[\prod_{i>m} P(f_i^+|d)\right]\prod_j P(d_j) \tag{5.28}$$

$$\propto E\left\{\prod_{i\leq m} P(f_i^+|d,\xi_i)\right\} \tag{5.29}$$

where the expectation is over the posterior distribution for the diseases given those positive findings that we plan to treat exactly. Note that the proportionality constant does not depend on the variational parameters; it is the likelihood of the exactly treated positive findings. We now insert the explicit forms of the transformed conditional probabilities (see Eq. (5.12)) into Eq. (5.29) and find

$$P(f^+|\xi) \propto E\left\{\prod_{i\leq m} e^{\xi_i(\theta_{i0}+\sum_j \theta_{ij}d_j)-f^*(\xi_i)}\right\} \tag{5.30}$$

$$= e^{\sum_{i\leq m}(\xi_i\theta_{i0}-f^*(\xi_i))} E\left\{e^{\sum_{j,\,i\leq m}\xi_i\theta_{ij}d_j}\right\} \tag{5.31}$$

where we have simply the products over i into sums in the exponent and pulled out the terms that are independent of the expectation. On a log-scale, the proportionality becomes an equivalence up to a constant:

$$\log P(f^+|\xi) = C + \sum_{i\leq m}(\xi_i\theta_{i0} - f^*(\xi_i)) + \log E\left\{e^{\sum_{j,i\leq m}\xi_i\theta_{ij}d_j}\right\} \tag{5.32}$$

Several observations are in order. Recall that $f^*(\xi_i)$ is the conjugate of the concave function f (the exponent), and is therefore also concave; for this reason $-f^*(\xi_i)$ is convex. We claim that the remaining term

$$\log E\left\{e^{\sum_{j,i\leq m}\xi_i\theta_{ij}d_j}\right\} \tag{5.33}$$

is also a convex function of the variational parameters. The justification for this can be found either in Appendix 1.A of chapter 1 or Appendix 2.A of chapter 2 (it is also well-known in the physics literature). Since any sum of convex functions stays convex, we conclude that $\log P(f^+|\xi)$ is a convex function of the variational parameters. Importantly, this means that there are no local minima and the optimal or minimizing ξ can be always found. We may safely employ the standard Newton-Raphson procedure to solve $\nabla \log P(f^+|\xi) = 0$. For simplicity, we may equivalently iteratively optimize the individual variational parameters, i.e., for each ξ_k solve $\partial/\partial\xi_k \log P(f^+|\xi) = 0$. In this case, the relevant derivatives consists of (algebra omitted):

$$\frac{\partial}{\partial\xi_k}\log P(f^+|\xi) = \theta_{k0} + \log\frac{\xi_k}{1+\xi_k} + E\left\{\sum_j \theta_{kj}d_j\right\} \tag{5.34}$$

$$\frac{\partial^2}{\partial^2 \xi_k} \log P(f^+|\xi) \;=\; \frac{1}{\xi_k} - \frac{1}{1+\xi_k} + Var\left\{ \sum_j \theta_{kj} d_j \right\} \qquad (5.35)$$

Here the expectation and the variance are with respect to the same posterior distribution as before, and both derivatives can be computed in time linear in the number of associated diseases for the finding. We note that the benign scaling of the variance calculations comes from exploiting the special properties of the noisy-OR dependence and the marginal independence of the diseases. The convexity of the log-likelihood (with respect to ξ_k) is evident from the expression for the second derivative (i.e., it is always positive).

To further simplify the optimization procedure, we can simply set the variational parameters to values optimized in the context where all the positive findings have been transformed. While such setting is naturally suboptimal for cases where there are exactly treated positive findings, the incurred loss in accuracy is typically quite small. The gain in computation time can, however, be considerable especially when a large number of positive findings are treated exactly; the expectations above can be exponentially costly in the number of such positive findings (see Eq. (5.5)). The simulation results reported in this chapter have been obtained using this shortcut unless otherwise stated.

5.B Accuracy and the leak probability

Here we prove that any increase in the leak probability improves the accuracy of the variational transformation. Such increase of the leak probability can happen, for example, as a result of conditioning the likelihood of the observed findings on the presence of some of the diseases (see the calculation of the posterior marginal probabilities in the text). Now, for the accuracy to improve as a function of the leak probability (or the bias term in the exponentiated notation) means that the following log-likelihood ratio has to be a decreasing function of the bias ϵ:

$$\text{ratio}(\epsilon) = \log \frac{\min_\xi \; E\left\{ e^{\xi(\sum_j \theta_{ij} d_j + \epsilon) - F(\xi)} \right\}}{E\left\{ 1 - e^{-\sum_j \theta_{ij} d_j - \epsilon} \right\}} \;=\; \log \frac{\min_\xi \; E\left\{ e^{\xi(x+\epsilon) - F(\xi)} \right\}}{E\left\{ 1 - e^{-x-\epsilon} \right\}} \qquad (5.36)$$

where the expectations are with respect to the posterior distribution over the diseases based on all other observed findings except for the positive finding i. We have adopted the simplified notation $x = \sum_j \theta_{ij} d_j$ since the linear form plays no role in the proof. We note that it suffices to show that the derivative of ratio(ϵ) is negative at $\epsilon = 0$. The negativity of this derivative for other values of ϵ follows simply by redefining x (i.e, $x \leftarrow x + \epsilon$), and the fact that the proof does not depend on the distribution over x.

To fix ideas, we start from the observation that since the variational transformation is optimized with respect to ξ at $\epsilon = 0$ we must have

$$\frac{\partial}{\partial \xi} E\left\{e^{\xi x - F(\xi)}\right\} = e^{-F(\xi)}\left[E\left\{x e^{\xi x}\right\} - \frac{\partial F(\xi)}{\partial \xi} E\left\{e^{\xi x}\right\}\right] = 0 \tag{5.37}$$

The fact that $\partial F(\xi)/\partial \xi = -\log(\xi/(\xi+1))$ allows us to obtain an implicit solution for ξ from the above equation giving

$$\xi = \frac{e^{-\tilde{x}}}{1 - e^{-\tilde{x}}}, \text{ where } \tilde{x} = \frac{E\left\{x e^{\xi x}\right\}}{E\left\{e^{\xi x}\right\}} \tag{5.38}$$

$\tilde{x}$ can be seen to be the mean of x with respect to the twisted or *tilted* distribution. It is well-known in the large deviation literature that $\tilde{x} \geq E\{x\}$ (see e.g. Bucklew, 1990). To use this property we first differentiate ratio(ϵ) with respect to ϵ and find:

$$\frac{\partial}{\partial \epsilon}\text{ratio}(\epsilon) = \xi - \frac{E\left\{e^{-x}\right\}}{1 - E\left\{e^{-x}\right\}} \tag{5.39}$$

where the implicit derivatives vanish due to the optimization of ξ. As $z/(1-z)$ is an increasing function of z and since e^{-x} is a convex function of x, i.e., $E\left\{e^{-x}\right\} \geq e^{-E\{x\}}$, it follows that

$$\frac{E\left\{e^{-x}\right\}}{1 - E\left\{e^{-x}\right\}} \geq \frac{e^{-E\{x\}}}{1 - e^{-E\{x\}}} \tag{5.40}$$

Finally, combining the results from Eq. (5.38), Eq. (5.39), and Eq. (5.40) with the fact that e^{-x} is a decreasing function gives

$$\frac{\partial}{\partial \epsilon}\text{ratio}(\epsilon) = \frac{e^{-\tilde{x}}}{1 - e^{-\tilde{x}}} - \frac{E\left\{e^{-x}\right\}}{1 - E\left\{e^{-x}\right\}} \leq \frac{e^{-\tilde{x}}}{1 - e^{-\tilde{x}}} - \frac{e^{-E\{x\}}}{1 - e^{-E\{x\}}} \leq 0 \tag{5.41}$$

This completes the proof.

Chapter 6

Discussion

Graphical models remain useful as knowledge representations to the extent that we are able to carry out inferences in these models or learn on the basis of observations. The increasing complexity of the models in important application areas, however, necessitate the use of approximate techniques to reap the benefits from the probabilistic representations. We have provided in this thesis a framework for approximating graphical models in a principled way through variational transformations. The preceding chapters contain the development of this approximation methodology for probabilistic inference, Bayesian estimation, and for diagnostic reasoning in the context of the QMR-DT belief network.

Probabilistic inference: In chapter 2 we laid the foundation for obtaining upper and lower bounds on probabilities and subsequently in chapter 3 removed the topological constraints from such bounds by introducing the relevant transformations recursively. The recursive methods proposed in chapter 3 are currently constrained by the available transformations. The framework nevertheless provides the basis for further development of these transformations.

Relevant to this framework is the role that the variational bounds play in determining the approximation accuracy. If the interval estimate specified by the complementary bounds is tight then necessarily the approximation is reliable. The converse, however, is not true in general. By this we mean that the marginal estimates[1] obtained through one of the bounds need not be inaccurate even if the interval estimate is too wide to be useful. The aspect that is lost due to wide intervals is the guarantee of reliability, not the accuracy directly. In the absence of useful intervals, the bounds will nevertheless retain their error metric (boundedness), and they can be refined according to this metric. It is only by mixing upper bounds with lower bounds that we lose the metric as well and, consequently, all the benefits accompanying variational methods over other approximation techniques.

Another issue to consider is the nature of the variational transformations we have used. Is the convexity framework too restrictive? Many variational transformations producing strict bounds can be viewed as manifestations of underlying convexity

[1] Note that the marginal estimate based on either upper or lower bound are found simply by normalizing the expression for the corresponding bound.

representations through reparametrization (cf. the mean field derivation in Appendix 4.39). More to the point is the question whether it is convenient to establish the relation – assuming it exists – to convex duality. This, of course, is not always the case.

We have also largely disregarded the issue of possible semantic losses resulting from the use of variational approximation techniques: which independence properties will be lost and which are imposed by the variational framework? While important, the question as posed is somewhat ambiguous since the answer depends on the optimization schedule adopted for the variational parameters. For example, if the optimization is performed conditionally on each configuration of the variables in the probability model, we retain the exact model. The way to investigate this issue is to concentrate on actual bounds on marginal probabilities, where the variational parameters are optimized conditionally on each marginal configuration. Normalizing such a bound yields a marginal distribution whose independence properties can be characterized.

Bayesian estimation: The Bayesian formalism for parameter estimation can be particularly useful in the extreme cases of scarce data or when the amount of data is overwhelming, both of which are likely to arise in modern environments. In the former case, the benefit comes from the ease of embedding prior knowledge into the parameter values, while in the latter the advantage relies on the sequentiality of the estimation procedure. Computationally, however, the Bayesian framework is quite costly and often infeasible. We demonstrated in chapter 4 that variational methods can provide efficient and principled approximations to re-establish the utility of the Bayesian formalism, particularly in the presence of missing values.

Applications: The preliminary application of the variational methods to the QMR-DT belief network illustrates the potential in these techniques for large scale applications. We did not fully utilize the power of the methodology, however, and the calculation of rigorous bounds on the posterior marginals in the QMR setting remains an issue for future work.

We have deferred the discussion about sampling methods in comparison to the variational ones here as such considerations seem most appropriate in the context of applications. While sampling techniques enjoy wider (or at least more immediate) applicability, they also appear to be less able to benefit from the structure of a particular model; nor do they attain bounds or analyzable expressions for the quantities of interest. It is also straightforward and profitable to combine variational methods with exact calculations, whereas with sampling techniques it is less apparent how this can be done efficiently. Extensive comparisons between these methods will be needed to resolve the domain characteristics in which one is preferable to the other. We note finally that the two frameworks need not be mutually exclusive: the variational estimates could be used, for example, as importance sampling distributions, which has been suggested by many people.