MIT Computational Cognitive Science Technical Report 9701

Variational methods and the QMR-DT database

Tommi S. Jaakkola and Michael I. Jordan
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, MA 02139
{tommi,jordan}@psyche.mit.edu

Abstract

We describe variational approximation methods for efficient probabilistic reasoning, applying these methods to the problem of diagnostic inference in the QMR-DT database. The QMR-DT database is a large-scale belief network based on statistical and expert knowledge in internal medicine. The size and complexity of this network render exact probabilistic diagnosis infeasible for all but a small set of cases. This has hindered the development of the QMR-DT network as a practical diagnostic tool and has hindered researchers from exploring and critiquing the diagnostic behavior of QMR. In this paper we describe how variational approximation methods can be applied to the QMR network, resulting in fast diagnostic inference. We evaluate the accuracy of our methods on a set of standard diagnostic cases and compare to stochastic sampling methods.

1 Introduction

Bayesian belief networks provide an elegant unifying formalism for probabilistic modeling (see, e.g., Pearl 1988, Jensen 1996). Given a set of random variables represented as nodes in a directed acyclic graph, and given a conditional probability distribution for each node, the formalism defines the joint probability distribution of the variables as the product of the node probabilities. General algorithms have been developed that calculate arbitrary conditional probabilities under this joint distribution; these algorithms can be used to perform a wide variety of inferential calculations.

For large-scale problems, however, the exact algorithms can be infeasible computationally. The algorithms must essentially sum over all combinations of values of nodes that are not in the conditioning set, and, roughly speaking, the number of terms in these sums scales exponentially with the number of variables that are stochastically dependent. Even efficient methods for performing these sums become infeasible in large, dense networks. Consider, for example, the Quick Medical Reference (QMR) knowledge base, compiled for internal medicine. The QMR knowledge base consists of a combination of statistical and expert knowledge for approximately 600 significant diseases and their associated findings (about 4000). In the probabilistic formulation of the database (QMR-DT; Shwe et al. 1991), the diseases and findings are arranged in a bi-partite graph, and the diagnosis problem is to infer a probability distribution for the diseases given a subset of findings. Given that each finding is generally relevant to a wide variety of diseases, the graph underlying the QMR-DT is dense, reflecting high-order stochastic dependencies. These dependencies are infeasible to handle exactly; indeed, for the more difficult diagnosis problems that we consider below we estimate that exact algorithms would require approximately 50 years to run on current computers.

An alternative to exact methods is provided by stochastic sampling methods (see, e.g., Gelfand & Smith 1990). These methods are readily implemented for general belief networks and provide theoretical assurance of convergence to exact answers. It can be difficult in practice, however, to diagnose convergence and to assess the reliability of results obtained over finite sampling intervals. Sampling methods can also be slow, and for problem such as medical diagnosis in which on-line, interactive use of the inference system is envisaged, the convergence rate of the inference algorithm is a serious consideration.

In this paper we present variational methods for performing approximate inference. We apply these methods to the problem of diagnosis in the QMR-DT setting. Variational methods have a long history as approximation techniques in physics and applied mathematics. Unlike sampling methods, variational techniques yield deterministic approximations that are adapted to each case separately. Moreover, these techniques can readily be merged with exact techniques—we show this explicitly in the QMR-DT setting. This fact allows us to develop "anytime" algorithms in which available computational resources determine the extent to which approximations are introduced. Moreover, the variational methods yield explicit expressions for the posterior probabilities of the diseases; these expressions can be subjected to analysis concerning the accuracy and sensitivity to the various aspects of each case under consideration.

We begin by defining the QMR-DT belief network and the diagnostic inference problem. We then introduce and develop the variational techniques used in the paper. Finally, we report numerical results.

2 The QMR-DT belief network

The QMR-DT belief network is a two-level or bi-partite network (see figure 1). The diseases and findings 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 (cf. Pearl 1988). 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). (Note that diseases are *not* assumed to be mutually exclusive; a patient can have multiple diseases). Also evident from the structure is that conditional on the states of the diseases the findings are independent of each other (conditional independence). For a discussion regarding the medical validity of these and other assumptions embedded into the QMR-DT belief network, see Shwe et al. (1991).

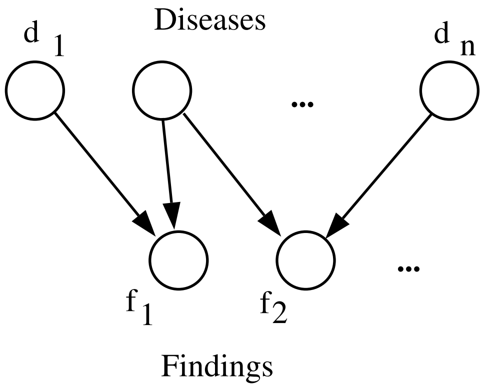


Figure 1: The QMR belief network is a two-level network where the dependencies between the diseases and their associated findings have been modeled via noisy-OR gates.

To state more precisely the probability model implied by the QMR-DT belief network, we write the joint probability of diseases and findings as

$$P(f,d) \quad = \quad P(f|d)P(d) = \left[\prod_i P(f_i|d)\right]\left[\prod_j P(d_j)\right] \qquad (1)$$

where d and f are binary (1/0) vectors refering 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:

$$P(f_i = 0|d) \quad = \quad P(f_i = 0|L) \prod_{j \in pa_i} P(f_i = 0|d_j) \qquad (2)$$

$$= (1 - q_{i0}) \prod_{j \in pa_i} (1 - q_{ij})^{d_j} = e^{-\theta_{i0} - \sum_{j \in pa_i} \theta_{ij} d_j}, \tag{3}$$

where 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, and $q_{i0} = P(f_i = 0|L)$ 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); that is, 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 paper for reasons of clarity.

3 Inference

Carrying 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, however, 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. (2) and Eq. (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.

$$P(d_j|f^+) \propto \sum_{d \backslash d_j} P(f^+|d)P(d) \tag{4}$$

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

$$P(f^+|d)P(d) = \left[\prod_i P(f_i^+|d)\right] \left[\prod_j P(d_j)\right] = \left[\prod_i \left(1 - e^{-\theta_{i0} - \sum_j \theta_{ij} d_j}\right)\right] \left[\prod_j P(d_j)\right] \tag{5}$$

which follows from the notation in Eq. (3) and the fact that $P(f_i^+|d) = 1 - P(f^-|d)$. To perform the summation in Eq. (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 this calculation is not feasible.

4 Variational methods

4.1 A brief introduction

The objective of variational methods is to simplify a complicated joint distribution such as the one in eq. (5) through variational transformations of the node probabilities. The transformations rerepresent the conditional node probabilities in terms of optimization problems. Such representations are 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 the variational transformations have been introduced.

How do we find appropriate transformations? The variational methods we consider here come from convex duality. Let us first consider methods for obtaining upper bounds. It is well-known in convex analysis that any concave (i.e. convex down) function can be rerepresented in terms of its dual or conjugate function (see Appendix A):

$$f(x) = \min_{\xi} \{ \, \xi^T x - f^*(\xi) \, \} \tag{6}$$

where $f^*(\xi)$ is the conjugate function of $f(x)$. The roles of f and f^* are interchangeable in this transformation (hence the duality). This representation of f as an optimization problem over a family of linear functions is a variational transformation. The additional parameter ξ parameterizing this family is known as the variational parameter. If we relax the minimization above and fix the the variational parameter, we obtain a bound

$$f(x) \leq \xi^T x - f^*(\xi) \tag{7}$$

which holds for each value of the variational parameter.

For convex functions the dual representation is expressed in terms of a maximization; relaxing the maximization yields lower bounds.

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

$$P(f_i^+|d) \;=\; 1 - e^{-\theta_{i0} - \sum_j \theta_{ij} d_j} = e^{\log(1-e^{-x})} \tag{8}$$

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 discussion in the previous section, we know that there must exist a variational upper bound for this function that is linear in x:

$$f(x) \leq \xi x - f^*(\xi) \tag{9}$$

The conjugate function $f^*(\xi)$ for noisy-OR is given by

$$f^*(\xi) = -\xi \log \xi + (\xi + 1) \log(\xi + 1) \tag{10}$$

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

$$\begin{aligned} P(f_i^+|d) &= e^{f(x)} && (11)\\ &\leq e^{\xi_i(\theta_{i0} + \sum_j \theta_{ij} d_j) - f^*(\xi_i)} && (12)\\ &= e^{\xi_i \theta_{i0} - f^*(\xi_i)} \prod_j \left[e^{\xi_i \theta_{ij}} \right]^{d_j} && (13)\\ &\equiv P(f_i^+|d, \xi_i) && (14) \end{aligned}$$

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

$$P(f_i^+|d) \rightarrow P(f_i^+|d,\xi_i) \tag{15}$$

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 subsequently obtained from the transformed probability model.

Two issues need to be clarified within 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 B; the adjustment of the variational parameters reduces to a convex optimization problem that can be carried out efficiently and reliably (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.

4.2.1 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 must be based on 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 assessing this accuracy comes from the fact that the transformations are 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 $P(f^+|\xi)$:

$$P(f^+) = \sum_d P(f^+|d)P(d) \quad \leq \quad \sum_d P(f^+|d,\xi)P(d) = P(f^+|\xi) \tag{16}$$

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:

$$\delta_i = P(f^+|\xi) - P(f^+|\xi \setminus \xi_i) \tag{17}$$

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 those findings for which δ is large.

Figure 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 quite large. We also note that the curves for the proposed ordering are convex; thus the bound improves 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. For this reason we might expect the variational posterior estimates to become reasonably accurate after a reasonably small fraction of the positive findings have been treated exactly. We note finally 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 suprising, i.e., are likely to occur, or when there is less impetus for explaining them (the leak probability is large).

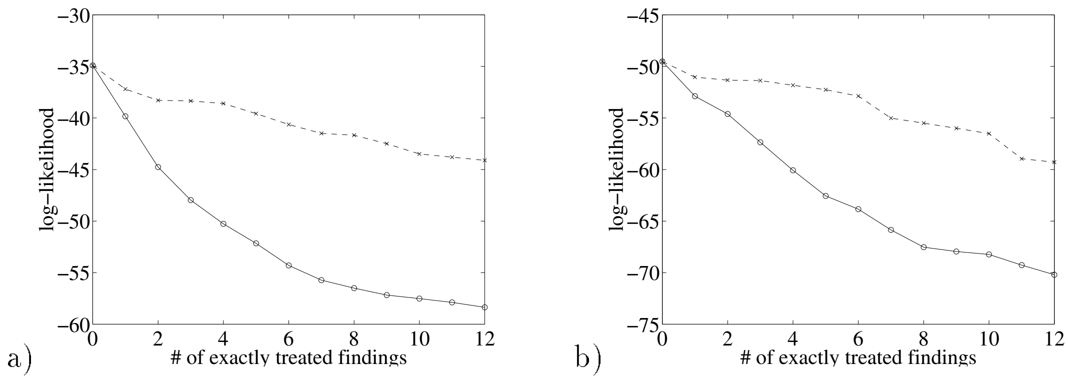


Figure 2: The log-likelihood of the observed findings as a function of the number of positive findings treated exactly. The solid line corresponds to the proposed ordering and the dashed line is for a random ordering.

5 Results

The diagnostic cases that we used in evaluating the performance of the variational techniques were cases abstracted from clinocopathologic conference ("CPC") cases. These cases involve multiple diseases underlying the observed findings and are considered clinically difficult cases. 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 (≤ 20) to allow an exact computation of the posterior marginals for the purposes of comparison[1]. We begin by 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.

5.1 Comparison to exact posterior marginals

In this section we discuss the CPC cases that have 20 or fewer positive findings. Table 1 contains a description of these "tractable" cases.

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

Table 1: Description of the cases for which we evaluated the correct posterior marginals.

Figures 3 and 4 illustrate the correlation between the true posterior marginals and the approximate marginals calculated under the variational distribution. 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 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 reveals quantitatively the rate of convergence of the posterior marginals. The plots show the fraction of all posterior marginal estimates (10 largest from each case)

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

whose error exceeds the specified threshold as a function of the number of positive findings that were treated exactly. We may loosely interprete these level curves as probabilities that, in a hyphothetical case, the error in a posterior marginal estimate would exceed the specified limit. Figure 5a is in terms of the relative error in the posterior marginals; figure 5b on the other hand uses the absolute error.

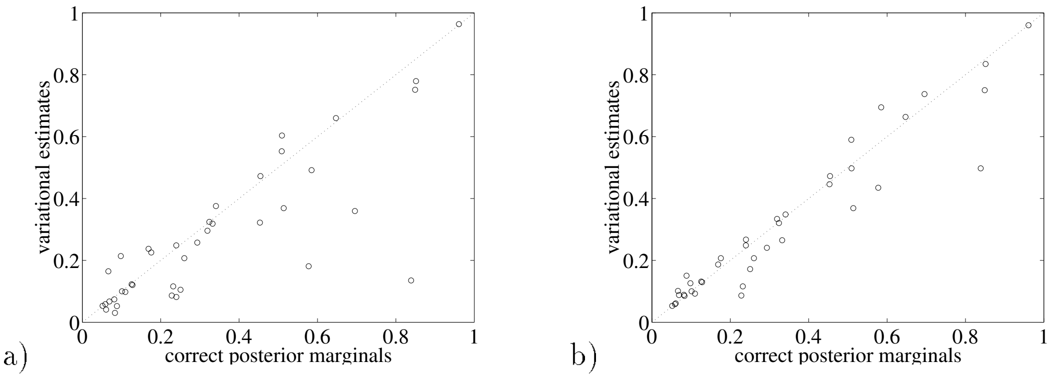


Figure 3: Correlation between the variational posterior estimates and the correct marginals. In a) 4 and in b) 8 positive findings were treated exactly.

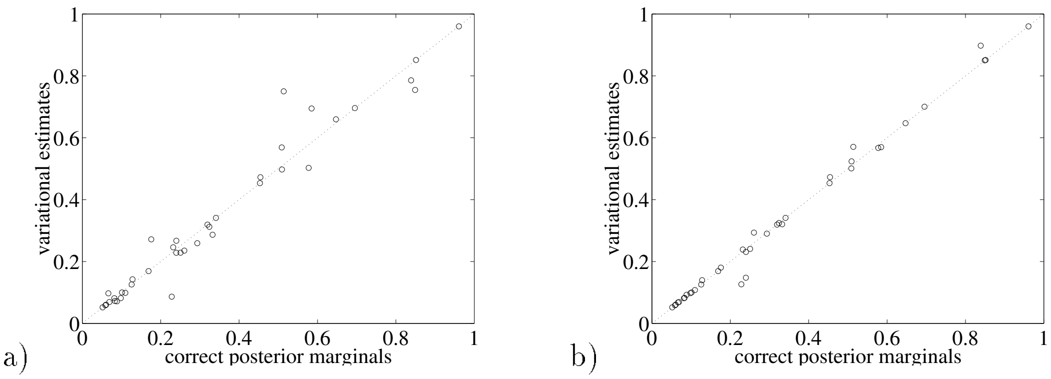
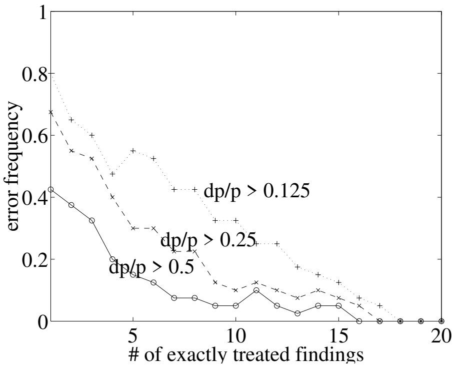


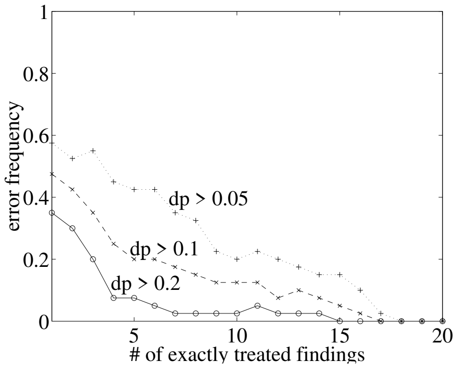
Figure 4: Correlation between the variational posterior estimates and the correct marginals. In a) 12 and in b) 16 positive findings were treated exactly.

5.2 Comparison to Gibbs' sampling

In this section we compare the accuracy and computation time associated with the variational posterior estimates to those obtained through stochastic sampling. We have implemented a simple Gibbs' sampler as a representative stochastic sampling technique. Our goal is not to present a conclusive comparison of variational methods



a)



b)

Figure 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.

and sampling methods, but rather to present experiments that indicate the kinds of results that can be achieved with off-the-shelf techniques. (For more sophisticated variational methods, see Jaakkola & Jordan, 1996b; for more sophisticated sampling methods, see, e.g., Neal, 1993; and, in the context of QMR-DT, Shwe & Cooper, 1991).

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

$$\hat{P}(d_i) = \frac{1}{T} \sum_t P(\, d_i \mid f,\, d^t \setminus d_i^t \,) \tag{18}$$

where each new disease configuration d^t was computed from the previous configuration d^{t-1} by sequentially resampling the disease states with replacement. The order for the updates was chosen randomly at each stage. Every fifth such d^t configuration was included in the sum; intervening samples were dropped. The initial configuration d^0 was drawn from the prior distribution over the diseases[2]. While discarding early samples is generally profitable in sampling implementations, such a maneuver only seemed to deteriorate the results in our case. In particular the accuracy gained from including only later samples was offset by the loss in computation time spent discarding early samples (cf. the time/accuracy plot of figure 6 below). Consequently no early samples were excluded.

To be able to assess the accuracy of the posterior estimates we restricted ourselves to the four tractable cases described in the previous section. Figure 6 plots the mean correlations (across the tractable cases) between the approximate estimates and the correct posterior marginals as a function of the computation time needed for obtaining

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

the estimates. The correlation measures for the stochastic method were averaged across 20 independent runs for each tractable 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 tractable case from the 20 different runs; the error bars therefore reflect how much the correlations would be expected to 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. The figure shows that to achieve roughly equivalent levels of accuracy, the Gibbs' sampler requires significantly more computation time than the variational method.

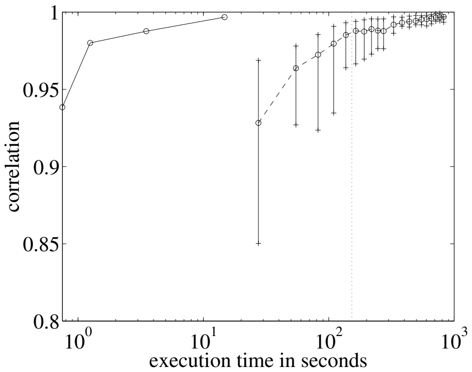


Figure 6: The mean correlation between the approximate and exact posterior marginals as a function of the execution time (seconds). Solid line: variational estimates; dashed line: Gibbs' sampling. The dotted line indicates the average time for the exact calculation of the posterior marginals in the tractable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

5.3 Posterior accuracy across cases

We now discuss our results for the cases with more than 20 positive findings. For these cases it is infeasible to compute the exact posterior marginals. In the absence of the exact reference values, we have to find a surrogate to assess the accuracy of the estimated marginals. We obtain such a surrogate via a measure of variability of these marginals. Recall first that in the variational 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 generally depend on which conditionals received exact treatment and which were approximated. A lack of such dependence suggests that we have the correct posterior marginals. We can therefore use this dependence as a surrogate 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.

5.3.1 Mean squared variability

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

$$\hat{\sigma_i}^2 = \frac{1}{K}\sum_k \left(\hat{P}_i(d_i = 1) - \hat{P}_i^{+k}(d_i = 1)\right)^2 \tag{19}$$

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

$$\hat{\sigma} = \max_{i \le 10} \hat{\sigma}_i \tag{20}$$

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 provide some empirical justification for it. We do this by using the tractable CPC cases considered in section 5.1 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 7, the correlation between these measures is indeed quite good, suggesting that the variability measure is a reasonable surrogate. indicative.

Figure 8 illustrates how the variability $\hat{\sigma}$ of the posterior estimates depends on the number of positive and negative findings across all of the CPC cases. Eight conditional probabilities were treated exactly in each of the CPC cases. Figure 9 is analogous except that the number of findings treated exactly 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 findings treated exactly 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

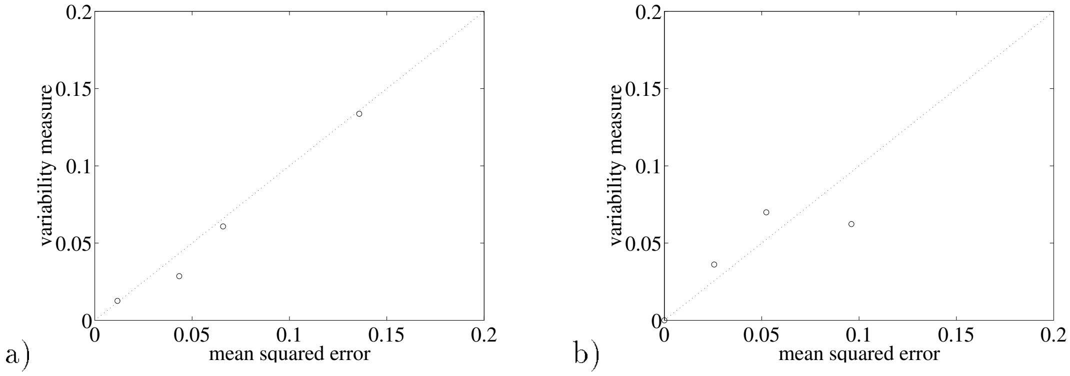


Figure 7: The correlation between $\hat{\sigma}$ and σ, where $\hat{\sigma}$ is the variability measure and σ is the true mean squared error between the variational estimates and the correct marginals. 10 most likely posterior marginals were included from each tractable case. The number of exactly treated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

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 generally reduce the prior probabilities for the diseases. Smaller prior probabilities decrease the posteriors marginals. The scale dependent $\hat{\sigma}$ therefore decreases without any real improvement in the variational accuracy. The figure 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 rare.

5.3.2 Min/Max variability across cases

While the squared error captures the mean variability in the posterior estimates, it is also important to ascertain how much the true posterior marginals can 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 5.1. Figure 10 illustrates the accuracy of these bounds for 10 most likely posterior marginals from each of the four cases. Although a few of the posterior marginals fall outside of these bounds, the discrepancies are quite minor. Moreover, when the bounds are tight, the correct 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,

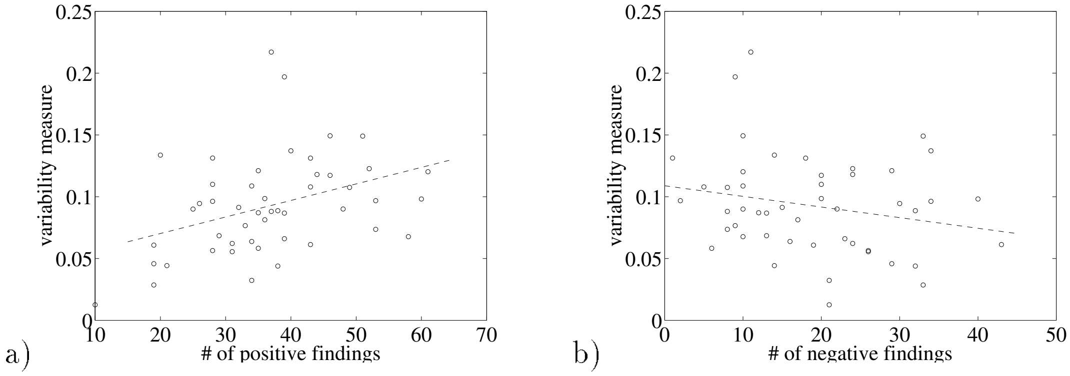


Figure 8: a) The variability $\hat{\sigma}$ of the posterior marginals as a function of the number positive findings in the CPC cases. b) The same variability measure $\hat{\sigma}$ but now as a function of the number of negative findings. 8 positive findings were treated exactly for this figure.

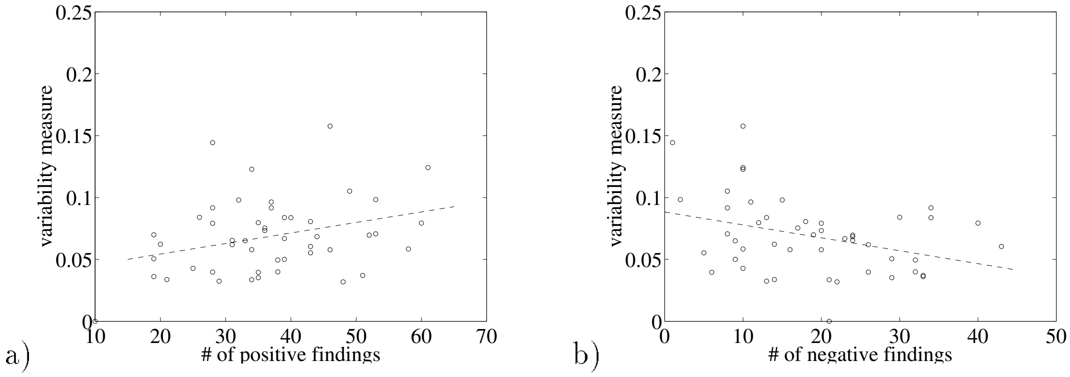


Figure 9: a) The variability $\hat{\sigma}$ of the posterior marginals as a function of the number positive findings in the CPC cases. b) The same variability measure $\hat{\sigma}$ but now as a function of the number of negative findings. Now 12 positive findings were handled exactly.

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.

Figure 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

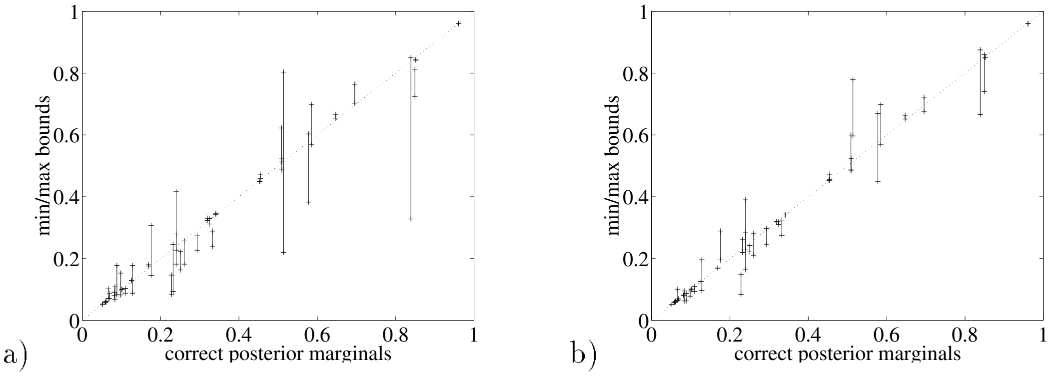


Figure 10: The correlation between the min/max bounds and the true posterior marginals. 10 most likely posterior marginals were included from each tractable case. In a) 8 findings were treated exactly and in b) 12.

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 11b). The dependence of the correlation coefficients on the number of exactly treated positive findings is illustrated in figure 12a. The high monotonic increase in the correlation is mainly due to the proper ordering of the findings to be treated exactly (see section 4.2.1). In comparison, figure 12b shows the development of the correlations for a random ordering.

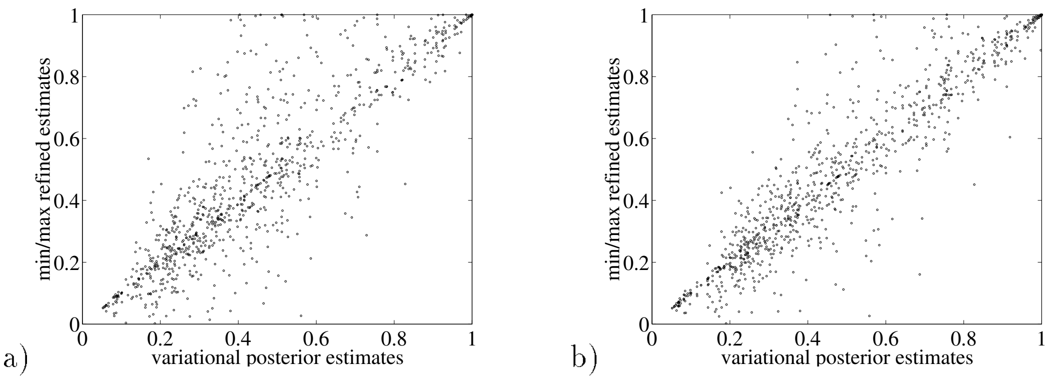


Figure 11: a) Correlation between the estimated posterior marginals and the min/max refined marginals. There were 8 positive findings considered exactly. b) as before but now the number of findings treated exactly was 12.

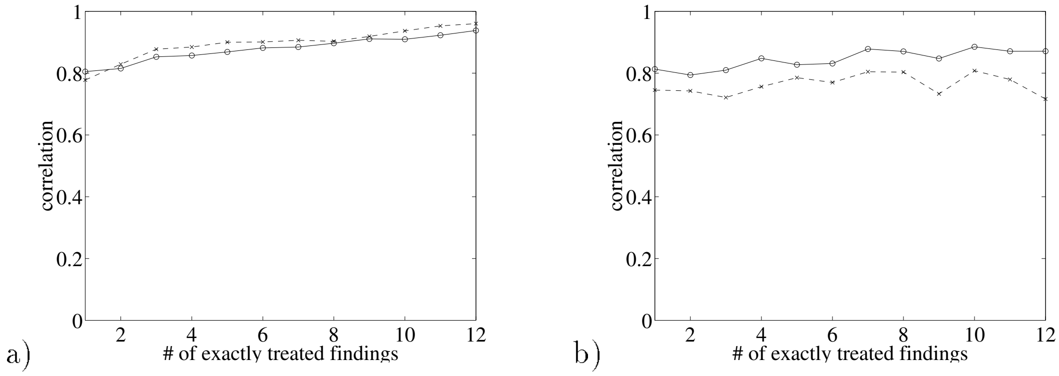


Figure 12: Mean correlation between the approximate posterior marginals and the min/max bounds as a function of the number of positive findings that were treated exactly. Solid line: correlation with the max- bound; dashed line: correlation with the min- bound. Figure a) is for the case where a proper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

5.4 Computation time

In this section we provide additional details on the computation time required to run the variational algorithms. The times that we report are obtained from runs on a Sun Sparc 10 workstation.

The execution times are overwhelmingly dominated by the number of positive findings treated exactly. (The time required for the variational optimization, using the shortcut discussed at the end of Appendix B, is insignificant). When 12 or fewer findings were treated exactly, the maximum time across the CPC cases was less than 2 seconds. Thus the variational posterior estimates can be obtained in real time.

In a practical setting it would be important to carry out the verifying analyses described in the previous sections. Thus the total time for obtaining the variational estimates and verifying them is the relevant timing figure for practical implementation. Figure 13 plots the mean and the maximum execution times for verification across the CPC cases, 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.

6 Discussion

Our work is preliminary in several respects. Note in particular that all of the analyses that we report here are based on *upper* bounds on the likelihood. The optimizing variational distribution that we obtain does indeed provide an upper bound on the probability of the findings, but the distribution does not necessarily provide upper

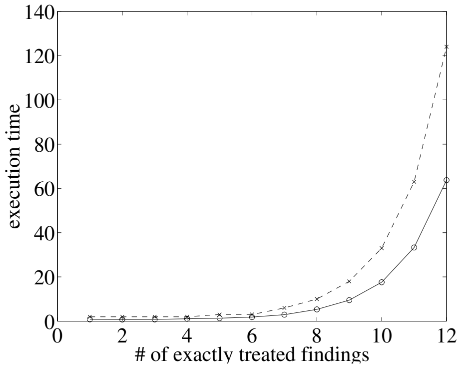


Figure 13: Mean (solid line) and maximum (dashed line) execution times in seconds for the analyses of section 5.3 across the CPC cases as a function of the number of exactly treated positive findings.

bounds on the posterior marginals. To obtain upper bounds on the latter quantities would require that we also obtain *lower* bounds on the likelihood. Although variational machinery naturally provides lower bounds as well as upper bounds on the likelihood (cf. Jaakkola & Jordan 1996a), in our preliminary work on the QMR-DT database, however, the bounds that we have obtained are not sufficiently tight. We are currently exploring alternative variational transformations for the node probabilities in order to tighten these bounds.

Our work also has not yet exploited the power of variational methods to study the sensitivity of the calculations. The fact that the variational methodology yields explicit expressions for the posterior probabilities of the diseases provides us with a tool to perform such sensitivity analyses, and we are currently exploring this possibility. Finally, note that we have utilized only the simplest variational transformations in the current paper, in particular those based on linear convexity bounds. It is worth exploring the speed/accuracy tradeoffs obtained by using more sophisticated bounds (cf. Jaakkola & Jordan 1996b).

Our results are nonetheless quite promising. We have presented an algorithm which runs in real time on a large-scale belief network for which exact algorithms are entirely infeasible. While further work is required to verify the accuracy of the results obtained by this algorithm, our comparisons with exact algorithms on the tractable cases, and our surrogates on the intractable cases, suggest that the results are quite accurate. We also showed that our variational method is significantly faster than simple sampling methods, although further work is required to flesh out this comparison.

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.

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

A. Gelfand and A. Smith (1990). Sampling-based approaches to calculating marginal Densities. *Journal of the American Statistical Association* **85**(410):398-409.

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

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

B. Middleton, M. Shwe, D. Heckerman, M. Henrion, E. Horvitz, H. Lehmann, and G. Cooper (1990). Probabilistic Diagnosis Using a Reformulation of the INTERNIST-1/QMR Knowledge Base II. Evaluation of Diagnostic Performance. Section on Medical Informatics Technical report SMI-90-0329. Stanford University.

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

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

R. Rockafellar (1972). *Convex Analysis.* Princeton Univ. Press.

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, and G. Cooper (1991). Probabilistic Diagnosis Using a Reformulation of the INTERNIST-1/QMR Knowledge Base I. The Probabilistic Model and Inference Algorithms. *Methods of Information in Medicine* **30**:241-255, 1991.

A Duality

We discuss here the dual or conjugate representations for convex functions. We refer the reader to Rockafellar (1970) for a more extensive treatment of convex duality.

Let $f(x)$ be a real valued and convex function defined on 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 14). 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 14). Through parameterizing these half-spaces we obtain the dual representations of convex functions. To this end, we define a half-space by the condition:

$$\text{all } (x, y) \text{ such that } x^T\xi - y - \mu \leq 0 \tag{21}$$

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

$$\max_{x \in X} \{ \, x^T\xi - f(x) - \mu \, \} \leq 0, \tag{22}$$

as well. Equivalently,

$$\mu \;\; \geq \;\; \max_{x \in X} \{ \, x^T\xi - f(x) \, \} \quad \equiv \quad f^*(\xi) \tag{23}$$

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 14). To clarify the duality, let us drop the maximum and rewrite the inequality as

$$x^T\xi \leq f(x) + f^*(\xi) \tag{24}$$

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

$$f(x) = \max_{\xi \in \Xi} \{ \, x^T\xi - f^*(\xi) \, \} \tag{25}$$

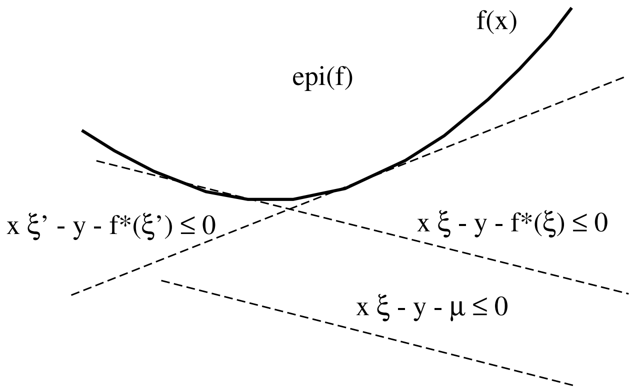


Figure 14: Half-spaces containing the convex set $epi(f)$. The conjugate function $f^*(\xi)$ defines the critical half-spaces whose intersection is $epi(f)$, or, equivalently, it defines the tangent planes of $f(x)$.

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

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

B 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

$$P(f^+) = \sum_d P(f^+|d)P(d) \quad \leq \quad \sum_d P(f^+|d,\xi)P(d) = P(f^+|\xi) \tag{26}$$

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

$$\begin{aligned} P(f^+|\xi) &= \sum_d \left[\prod_{i\leq m} P(f_i^+|d,\xi_i)\right]\left[\prod_{i>m} P(f_i^+|d)\right]\prod_j P(d_j) && (27) \\ &\propto E\left\{\prod_{i\leq m} P(f_i^+|d,\xi_i)\right\} && (28) \end{aligned}$$

where the expectation is over the posterior distribution for the diseases given those positive findings that we plan to treat exactly. Note that the propotionality 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. (12)) into Eq. (28) and find

$$
\begin{aligned}
P(f^+|\xi) &\propto E\left\{\prod_{i\leq m} e^{\,\xi_i(\theta_{i0}+\sum_j \theta_{ij}d_j)-f^*(\xi_i)}\right\} && (29)\\
&= e^{\sum_{i\leq m}(\xi_i\theta_{i0}-f^*(\xi_i))}\, E\left\{e^{\sum_{j,\, i\leq m}\xi_i\theta_{ij}d_j}\right\} && (30)
\end{aligned}
$$

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 propotionality becomes an equivalence up to a constant:

$$
\log P(f^+|\xi) = C + \sum_{i\leq m}\left(\xi_i\theta_{i0} - f^*(\xi_i)\right) + \log E\left\{e^{\sum_{j,i\leq m}\xi_i\theta_{ij}d_j}\right\} \qquad (31)
$$

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

$$
\log E\left\{e^{\sum_{j,i\leq m}\xi_i\theta_{ij}d_j}\right\} \qquad (32)
$$

is also a convex function of the variational parameters. Appendix C below provides the necessary justification for this claim. Now, 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):

$$
\begin{aligned}
\frac{\partial}{\partial\xi_k}\log P(f^+|\xi) &= \theta_{k0} + \log\frac{\xi_k}{1+\xi_k} + E\left\{\sum_j \theta_{kj}d_j\right\} && (33)\\
\frac{\partial^2}{\partial^2\xi_k}\log P(f^+|\xi) &= \frac{1}{\xi_k} - \frac{1}{1+\xi_k} + Var\left\{\sum_j \theta_{kj}d_j\right\} && (34)
\end{aligned}
$$

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.

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)). The simulation results reported in this paper have been obtained using this shortcut unless otherwise stated.

C Convexity

We note first that affine transformations do not change convexity properties. Thus convexity in $X = \sum_{j,i \leq m} \xi_i \theta_{ij} d_j$ implies convexity in the variational parameters ξ. It remains to show that

$$\log E\left\{e^{X}\right\} = \log \sum_i p_i \, e^{X_i} = f(\vec{X}) \tag{35}$$

is a convex function of the vector $\vec{X} = \{X_1 \ldots X_n\}^T$; here we have indicated the different discrete the values that the random variable X can take by X_i and denoted the probability measure on such values by p_i. Taking the gradient of f with respect to X_k gives

$$\frac{\partial}{\partial X_k} f(\vec{X}) = \frac{p_k e^{X_k}}{\sum_i p_i \, e^{X_i}} = P_k \tag{36}$$

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

$$\mathcal{H}_{kl} = \frac{\partial^2}{\partial X_k \partial X_l} f(\vec{X}) = \delta_{kl} P_k - P_k P_l \tag{37}$$

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

$$\vec{Z}^T \mathcal{H} \vec{Z} = \sum_k P_k Z_k^2 - (\sum_k P_k Z_k)(\sum_l P_l Z_l) = Var\{Z\} \geq 0 \tag{38}$$

where $Var\{Z\}$ is the variance of the discrete random variable Z assuming the values Z_i with probability P_i.