Tutorial on variational approximation methods

Tommi S. Jaakkola
MIT Artificial Intelligence Laboratory
545 Technology Square
Cambridge, MA 02139

October 25, 2000

Abstract

We provide an introduction to the theory and use of variational methods for inference and estimation in the context of graphical models. Variational methods become useful as efficient approximate methods when the structure of the graph model no longer admits feasible exact probabilistic calculations. The emphasis of this tutorial is on illustrating how inference and estimation problems can be transformed into variational form along with describing the resulting approximation algorithms and their properties insofar as these are currently known.

1 Introduction

The term variational methods refers to a large collection of optimization techniques. The classical context for these methods involves finding the extremum of an integral depending on an unknown function and its derivatives. This classical definition, however, and the accompanying calculus of variation no longer adequately characterizes modern variational methods. Modern variational approaches have become indispensable tools in various fields such as control theory, optimization, statistics, economics, as well as machine learning. The finite element method for solving differential equations[44], for example, is inherently a variational approach as is maximum entropy estimation[25].

There are a number of qualitative features that are shared across variational formulations. The primary component is naturally an optimization problem. The problem of interest is either transformed into an optimization problem or directly formulated as such based on a principle as in maximum entropy estimation (our emphasis in this tutorial is on transforming various inference and estimation problems into variational problems). The quantity to be optimized is typically an unknown function which, in simple cases, may be reduced to a vector (function values at discrete points). The solution to variational problems is often given in terms of *fixed point equations* that capture necessary conditions for optimality (characterizing locally optimal solutions). These are analogous to setting the gradient to zero in ordinary function optimization. Mean field equations (e.g., [37]) and Euler-Lagrange equations are prime examples of these fixed point equations. A method that successively enforces individual fixed point equations provides a common way of finding solutions to variational problems whenever a closed form solution cannot be found.

In recent years, a number of variational approaches have been successfully used for inference and estimation in large densely connected graphical probability models for which exact probabilistic calculations are no longer feasible (see, e.g., [23]). Their success derives primarily from two insights: first, probabilistic inference problems lend themselves naturally to variational formulations and, second, the resulting variational optimization problems admit principled approximate solutions. While there is nothing inherently approximate about variational formulations, as optimization problems they naturally facilitate finding approximate solutions. For example, any extremum problem involving an unknown function can be solved approximately by restricting the space of admissible functions (e.g., in terms of a finite number of basis functions). Analogous restrictions (factorization) can be found in the context of probabilistic calculations.

The primary goal of this tutorial is to illustrate how inference and estimation problems can be transformed into variational form along with describing the resulting approximation algorithms and their properties insofar as these are currently known. This tutorial is not intended to be exhaustive but merely to highlight the mathematical structure and properties of a number of variational approaches for inference and estimation calculations.

The paper is organized as follows: we begin with a detailed handling of two examples of variational formulations emphasizing their general features. This is followed by a brief introduction to graphical models and a derivation of the variational mean field approximation in the context of graphical models. We then derive structured mean field approximation along with variational factorization methods closely related to large deviation techniques. The last two sections concern with variational methods for maximum likelihood and Bayesian estimation. We end with a discussion of open problems.

2 Examples of variational methods

Many variational methods have similar mathematical structure. We illustrate this by building on two simple examples of variational methods. The basic insights derived from these variational methods carry over to mean field approximation. Specifically, we wish to clarify the transformation of the problem of interest into a variational form and how the resulting variational formulations admit approximate solutions.

We start with a well-known variational formulation of a matrix inversion problem in an estimation context and subsequently derive finite element methods as a variational solution to Poisson differential equation.

2.1 Matrix inversion

Many estimation methods such as linear regression and Gaussian process models (e.g., [48]) involve the need to invert large matrices. For the purpose of illustration, we provide here a variational formulation of this problem.

To fix ideas, suppose we are given a set of input vectors of $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, $\mathbf{x}_i \in R^d$, and corresponding scalar output values $\{y_1, \ldots, y_n\}$. We wish to find the best linear predictor of the form $y = \beta^T \mathbf{x} = \sum_{i=1}^d \beta_i x_i$, where β is the vector of parameters. For simplicity, we will assume that the fitting criterion is least squares. The least squares optimal parameter setting β^* is given by $\beta^* = C^{-1} b$, where

$$C = \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T, \quad b = \sum_{i=1}^n y_i \mathbf{x}_i \tag{1}$$

As the dimension d of the input vectors increases, evaluating $\beta^* = C^{-1} b$ can become burdensome. We formulate here a variational approach to computing $C^{-1} b$ (see also [12]).

Variational problem starts with a transformation into an optimization problem. It is perhaps surprising that we can often start with a trivial transformation. Suppose therefore that we knew the solution to the above problem, i.e, we had already evaluated β^*. We can then certainly optimize

$$J(\beta) = \frac{1}{2} (\beta^* - \beta)^T C (\beta^* - \beta) \tag{2}$$

with respect to β to find β^*. The distance measure here is weighted with matrix C so that deviations of β from β^* count more in directions where input examples $\mathbf{x}$ vary considerably. While this is a variational formulation leading to β^*, it is important to realize that we couldn't yet evaluate $J(\beta)$

without first computing β^*. To avoid this apparent conflict, we proceed to expand this trivial objective function. We also make use of the fact that we know the form of the solution $\beta^* = C^{-1}b$:

$$J(\beta) = \frac{1}{2}\beta^{*T}C\beta^* - \beta^T C\beta^* + \frac{1}{2}\beta^T C\beta \quad (3)$$

$$= \frac{1}{2}b^T C^{-1}b - \beta^T b + \frac{1}{2}\beta^T C\beta \quad (4)$$

In the resulting expression, the first term is a constant as far as the parameters β are concerned and we can drop it. Even without the constant term, the minimum is attained at $\beta = \beta^*$. The new objective that we can actually evaluate without consulting β^* is given by

$$\tilde{J}(\beta) = -\beta^T b + \frac{1}{2}\beta^T C\beta \quad (5)$$

It is easy to verify that this is a *convex* function of β. You may find it helpful to interpret the first linear term as an energy and the second quadratic term as a potential term playing a role analogous to the entropy in physics.

We have now made some important progress. While we can obtain the optimal solution β^* by minimizing $\tilde{J}(\beta)$, we can also find an approximate solution; we simply perform a partial minimization of $\tilde{J}(\beta)$. This can be done, for example, by taking only a few conjugate gradient steps (taking d such steps would recover the exact solutionβ^*). The objective function $\tilde{J}(\beta)$ serves as a metric guiding the choice of the approximate solution without the need to evaluate β^* for reference.

The purpose of this initial exercise was to demonstrate two basic underlying ideas. First, we can transform the original problem into an optimization problem whose objective can be evaluated without reference to the solution being sought. While this transformation may require some creativity, we argue that in many cases it is quite natural. We will return to this point later on. The second idea is to seek for an approximate solution using the variational objective to guide the selection of simpler approximations.

2.2 Finite element methods

Many problems in physics can be reduced to solving differential equations. This includes, for example, finding the temperature distribution over a material or gauging material deformations. One of the simplest but nevertheless representative problems is the following one dimensional Poisson differential equation:

$$-u''(x) = f(x), \;\; \forall x \in (a, b) \quad (6)$$

where $u''(x)$ is the second derivative of $u(x)$ with respect to the scalar argument x and $f(x)$ is the "source". We assume that the solution $u(x)$ (e.g., deformation) satisfies homogeneous boundary conditions, $u(a) = u(b) = 0$. A number of techniques exist for solving this problem. The best known is perhaps *finite element method* (see, e.g., [44]) that can be viewed as a variational method. The associated variational problem possesses a number of exemplary properties and is the reason for why we are introducing it here.

As in the context of linear regression, we first transform the problem into an optimization problem and subsequently search for an approximate solution. How do we find the optimization problem? Let $u^*(x)$ denote the desired solution satisfying the appropriate boundary conditions. Since this function is forced to be zero at the boundary points we have no degrees of freedom left for a constant term in the function. An appropriate way to compare any estimate $u(x)$ to the optimal solution $u^*(x)$ can be done in terms of the L_2 norm of its derivative:

$$J(u) = \frac{1}{2}\int_a^b (u'(x) - u^{*\prime}(x))^2 dx \tag{7}$$

This indeed serves as a valid distance measure. While minimizing this objective surely recovers $u^*(x)$, it is of no use to us unless we already know the solution. So, as before, we turn this objective into a form that we can actually evaluate without reference to $u^*(x)$. We can do this by expanding the integrand, integrating by parts, and using the form of the solution $-u^*(x)'' = f(x)$:

$$\begin{aligned} J(u) &= \frac{1}{2}\int_a^b u^{*\prime}(x)^2 dx - \int_a^b u'(x)u^{*\prime}(x)dx + \frac{1}{2}\int_a^b u'(x)^2 dx \\ &= \text{const.} - \left[\Big/_a^b u'(x)u^*(x) - \int_a^b u(x)u^{*\prime\prime}(x)dx \right] + \frac{1}{2}\int_a^b u'(x)^2 dx \\ &= \text{const.} - \left[0 + \int_a^b u(x)f(x)dx \right] + \frac{1}{2}\int_a^b u'(x)^2 dx \end{aligned} \tag{8}$$

where we have also used the fact that $u^*(x)$ must vanish at the boundary points. If we drop the first constant term that depends only on the solution u^*, we have an objective that can be readily evaluated for any $u(x)$:

$$\tilde{J}(u) = -\int_a^b u(x)f(x)dx + \frac{1}{2}\int_a^b u'(x)^2 dx \tag{9}$$

Similarly to our previous example, $\tilde{J}(u)$ is convex in $u(x)$ (differential operator is linear; any linear transformation of the argument of a convex function preserves convexity). The solution is, of course, unique since minimization of $\tilde{J}(u)$ with respect to $u(x)$ is equivalent to minimizing the original $J(u)$.

As a result, we have transformed the differential equation into an optimization problem involving a function $u(x)$ and its derivative $u'(x)$. The transformation is exact in the sense that minimizing the objective recovers the solution. The main benefit of this variational formulation, however, comes from the need to find an approximate solution.

To begin with, we must choose the form of the approximate solution. A natural choice in this context is to find the best function in a linear subspace spanned by a set of basis functions $\phi_1(x), \ldots, \phi_k(x)$ (in finite element methods these basis functions are derived from local approximating functions within each discretization interval or *element*). In other words, we with to find the best solution of the form

$$\tilde{u}(x) = \sum_{i=1}^{k} \alpha_i \phi_i(x) \tag{10}$$

where the ranking of the solutions is based on the objective $\tilde{J}(u)$. Note that the basis functions must confirm to the boundary conditions for our solution attempt to be admissible. It suffices now to substitute this form of the solution back into the objective function $\tilde{J}(u)$ and minimize it with respect to the free parameters, the linear coefficients $\{\alpha_i\}$. If we omit the straightforward algebra for clarity, the resulting objective looks like

$$\tilde{J}(\tilde{u}) = -\sum_i \alpha_i \left[\int_a^b \phi_i(x) f(x) dx \right] + \frac{1}{2} \sum_{ij} \alpha_i \alpha_j \left[\int_a^b \phi_i'(x) \phi_j'(x) dx \right] \tag{11}$$

By defining $b_i = \int_a^b \phi_i(x) f(x) dx$ and $C_{ij} = \int_a^b \phi_i'(x) \phi_j'(x) dx$, for $i, j = 1, \ldots, k$, we can rewrite this optimization problem in a matrix form:

$$\tilde{J}(\alpha) = -\alpha^T b + \frac{1}{2} \alpha^T C \alpha \tag{12}$$

which is conveniently exactly the variational form of the matrix inversion problem discussed earlier (this is, of course, not generally true for variational methods).

The necessary (and in this case also sufficient) conditions for optimality within the space of functions we are considering are obtained by setting the partial derivatives with respect to the parameters $\{\alpha_i\}$ to zero. In this case, the resulting *fixed point equations* are

$$\frac{\partial}{\partial \alpha} \tilde{J}(\alpha) = -b + C\alpha = 0 \tag{13}$$

implying, as before, that $\alpha^* = C^{-1}b$. In the context of finite element methods, inverting C is typically somewhat easier since the basis functions $\phi_i(x)$ have by design only local support. The inner product matrix C is therefore band-diagonal.

We make here a few final observations concerning this example. First, to find an approximate solution within a variational approach, we must first specify the form of the solution we are after. Second, by substituting the desired solution form back into the objective function, we obtain another variational problem, this time over the remaining free parameters. Finally, we note that finding a closed form solution for the variational parameters is rather atypical; variational problems often have to be solved iteratively.

After a brief introduction to graphical models provided in the next section, we will use the intuition derived from these two examples to guide our derivation and understanding of mean fields and beyond.

3 A brief introduction to graphical models

The feasibility of working with probability models over a large number of variables depends on how dependent the variables are on each other. In a graphical model, the presence/absence of such dependencies between the variables are represented in terms of a graph. In the graphical representation, the nodes V in the graph G correspond to the variables in the probability model and the edges E connecting the nodes signify dependencies. The power of such graph representation arises from the rigorous connection between separation properties in the graph and independence statements pertaining to the underlying probability model.

There are two main types of graph models, *undirected* and *directed*. The distinction arises from the type of edges used in the graphs and implies a difference in their independence semantics. The key problem in graphical representation of probability models is to explicate the structure of any probability distribution consistent with all the independence properties we can derive from the graph.

Figure 1a) illustrates an *undirected* graph model [3, 46]) also known as a *Markov random field* or MRF for short. For undirected graph models the ordinary graph separation of nodes is isomorphic to conditional independence statements about the variables associated with the nodes. For example, the graph in Figure 1a) states that the variables y_1 and x_2 are conditionally independent given x_1.

Independence properties read from the graph impose factorization con-

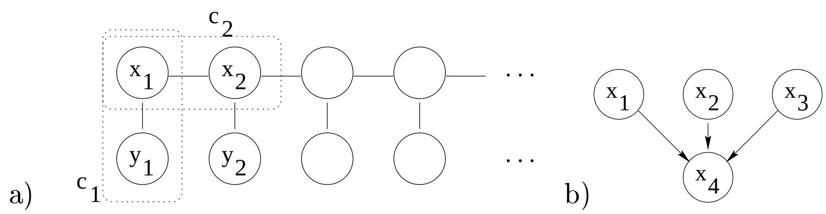

Figure 1: a) An undirected graph model (a Boltzmann chain [40, 42]). We have highlighted the first two cliques of the undirected graph with dotted lines. b) A simple directed graph model. Here x_1, x_2, and x_3 are marginally independent of each other while x_4 is dependent on the others. Knowing the value of "effect" x_4 renders the "causes" x_1, x_2, and x_3 dependent. This semantics cannot be captured with an undirected graph.

straints on any probability distribution consistent with the graph. In other words, the joint distribution must be expressed in terms of a product of non-negative *potential* functions $\Psi_c(\mathbf{x}_c)$, each depending on a specific subset of variables. The celebrated Hammersley-Clifford theorem (see, e.g., [3]) specifies the form of this factorization: the joint distribution must expressible as a product of potential function over *cliques* in the graph G:

$$P(\mathbf{x}) = \frac{1}{Z} \prod_{c \in C(G)} \Psi_c(\mathbf{x}_c) \tag{14}$$

where $C(G)$ is a collection of *cliques*[1] in the graph and $\mathbf{x}_c = \{x_i\}_{i \in c}$ is the set of variables corresponding to the nodes in clique c (in our notation here c is an index set of variables). Z is the normalization constant or partition function and plays an important role.

To exemplify these concepts we have indicated the first two cliques in Figure 1a). Any joint distribution consistent with the conditional independence properties we can derive from this chain-like structure must factor according to $P(\mathbf{x}) = \Psi_{c_1}(x_1, y_1)\Psi_{c_2}(x_1, x_2)\cdots$. It is important to realize that the probability distribution $P(\mathbf{x})$ may factor much more than this. For example, a distribution where all the variables are independent of each other, expressible as a product of potentials each depending on a single variable, is also consistent with the graph.

The computational cost of exact probabilistic inference calculations in undirected graph models depends on the size of the cliques. More precisely,

[1] A clique here is a maximal set of mutually connected nodes.

the cost is exponential in the size of the largest clique of a *triangulated*[2] graph (e.g., [28]). The cliques of a triangulated graph can be arranged in a tree structure (the junction tree) where computations can be carried out efficiently[29, 22]. The graph in Figure 1 is triangulated and its cliques already form a tree.

3.1 Directed graphical models

The second type of graph models, Bayesian networks, are based on *directed* graphs. In directed graphs, the edges signify asymmetric relations between the variables, loosely speaking the edges follow causal effects. Again, separation properties in the graph, correspond to independence statements about the underlying probability model. The separation criterion (the $d-$ separation criterion [38]) is a bit more involved but imposes a rather simple structure on the joint probability distribution. We must be able to write the joint distribution as a product of conditional probabilities[38] of x_i given its *parents* pa_i (the variables with directed arrows into x_i):

$$P(\mathbf{x}) = \prod_{i=1}^{n} P(x_i|\mathbf{x}_{pa_i}) \tag{15}$$

To ensure the joint distribution is well-defined, the directed graph must be *acyclic* (there are no directed cycles). Note that we don't need any normalization constant Z here – by design $Z = 1$.

We can always interpret the probability model $P(\mathbf{x})$ corresponding to a directed graph as an undirected model: we can set the potential functions equal to the conditional probabilities $\Psi_{v_i}(\mathbf{x}_{v_i}) = P(x_i|\mathbf{x}_{pa_i})$, $v_i = i \cup pa_i$ for $i = 1, \ldots, n$. In the corresponding undirected graph, each set of nodes v_i is fully connected. Such transformation into an undirected graph, known as *moralization*, hides some of the independence properties that were previously explicit in the directed graph. Directed graph models are, however, regularly transformed into undirected models as part of exact probabilistic calculations (see, e.g., [29]).

3.2 Additional structure in graphical models

Approximate inference methods rely on additional structure in the joint distribution beyond what is already explicated by the graph. For example,

[2]To triangulate the graph, we add edges so that any cycle of four or more nodes has a chord.

the probability model corresponding to a fully connected graph may factor into a product of *pairwise* potential functions depending only the variables associated with each undirected edge:

$$P(\mathbf{x}) = \prod_{e \in E} \Psi_e(\mathbf{x}_e) \tag{16}$$

where E is the collection of edges in the graph and we have absorbed the normalization constant into one of the potentials. Note that we can easily collect together the edge potentials into larger clique potentials. Mean field and other approximate inference algorithms heavily exploit this type of additional factorization structure.

The clique potentials or conditional probabilities may also possess useful additional parametric structure, other than factorization discussed above. Such parametric structure as in logistic regression models [31, 33], can be either directly exploited in approximate inference algorithms or used to impose additional factorization by breaking such conditionals into products of smaller ones. We will discuss variational methods for this purpose later in the tutorial.

4 Variational mean field method

We are now ready to apply the intuition from the two examples of varational methods to a probabilistic inference problem in graphical models. We start by defining the problem. Let G be the graph corresponding to a probability distribution $P(\mathbf{x})$ over n variables, $\mathbf{x} = \{x_1, \ldots, x_n\}$. Some of these variables are assumed observed or instantiated, $\mathbf{x}_v = \{x_i\}_{i \in v}$, while others remain hidden or unobserved, $\mathbf{x}_h = \{x_i\}_{i \in h}$. Here $\mathbf{x}_v$ is a shorthand for the instantiation of values of the variables $\{x_i\}_{i \in v}$. The two sets of variables are disjoint and $\mathbf{x} = \{\mathbf{x}_v, \mathbf{x}_h\}$. We also assume, for notational simplicity, that each variable x_i takes values in the finite set $\{0, \ldots, r-1\}$. The inference problem here is two fold: a) to evaluate the marginal probability of the observed data:

$$\log P(\mathbf{x}_v) = \log \sum_{\mathbf{x}_h} P(\mathbf{x}_v, \mathbf{x}_h) \tag{17}$$

where the summation is over the possible instantiations of the hidden variables $\mathbf{x}_h$, and 2) compute the posterior probability $P(\mathbf{x}_h|\mathbf{x}_v) = P(\mathbf{x}_v, \mathbf{x}_h)/P(\mathbf{x}_v)$ over the hidden variables. These goals are naturally tied; we can evaluate the posterior if we already have $P(\mathbf{x}_v)$. Exact computation of $P(\mathbf{x}_v)$, however, scales exponentially with the size of the largest clique in the induced (and triangulated) subgraph of G over the hidden variables or nodes. We will tacitly assume that this graph is too densely connected for exact computation to be practical.

Our first step here is to transform the problem into an optimization problem. We can do this in the following apparently silly way:

$$J(Q) = \log P(\mathbf{x}_v) - KL\left(Q_{\mathbf{x}_h} \,\|\, P_{\mathbf{x}_h|\mathbf{x}_v} \right) \tag{18}$$

where the Kullback-Leibler (KL) divergence is given by

$$KL\left(Q_{\mathbf{x}_h} \,\|\, P_{\mathbf{x}_h|\mathbf{x}_v} \right) = \sum_{\mathbf{x}_h} Q(\mathbf{x}_h) \log \frac{Q(\mathbf{x}_h)}{P(\mathbf{x}_h|\mathbf{x}_v)} \tag{19}$$

The KL-divergence is always positive and zero only if the *variational distribution* $Q(\mathbf{x}_h)$ over the hidden variables equals the true posterior probability $Q^*(\mathbf{x}_h) = P(\mathbf{x}_h|\mathbf{x}_v)$. Thus by maximizing $J(Q)$ with respect to Q we will always recover the log-probability of data $J(Q^*) = \log P(\mathbf{x}_v) - 0$. We conclude that our silly optimization problem indeed gives both the desired marginal, as the maximum value of $J(Q)$, and the posterior $Q^*(\mathbf{x}_h)$.

Note that the non-negativity of the KL-divergence also ensures us that for any variational distribution Q other than the posterior, we have a lower bound on the desired log-marginal probability

$$\log P(\mathbf{x}_v) = J(Q^*) \geq J(Q) \tag{20}$$

Moreover, it can be readily shown that $J(Q)$ is a concave (convex down) function of the variational distribution Q (see, e.g., [6]).

It remains to show that this trivial transformation into an optimization problem is at all useful. It is not even clear that we can evaluate the objective function for any choice of the variational distribution Q. To explicate this issue, we will rewrite the posterior probability appearing in the KL-divergence in terms of the joint distribution $P(\mathbf{x}_v, \mathbf{x}_h)$

$$\begin{aligned} J(Q) &= \log P(\mathbf{x}_v) - \sum_{\mathbf{x}_h} Q(\mathbf{x}_h) \log \frac{Q(\mathbf{x}_h)}{P(\mathbf{x}_h|\mathbf{x}_v)} && (21) \\ &= \log P(\mathbf{x}_v) - \sum_{\mathbf{x}_h} Q(\mathbf{x}_h) \log \frac{Q(\mathbf{x}_h) P(\mathbf{x}_v)}{P(\mathbf{x}_h, \mathbf{x}_v)} && (22) \\ &= -\sum_{\mathbf{x}_h} Q(\mathbf{x}_h) \log \frac{Q(\mathbf{x}_h)}{P(\mathbf{x}_h, \mathbf{x}_v)} && (23) \end{aligned}$$

$$
\begin{aligned}
&= -\sum_{\mathbf{x}_h} Q(\mathbf{x}_h) \log Q(\mathbf{x}_h) + \sum_{\mathbf{x}_h} Q(\mathbf{x}_h) \log P(\mathbf{x}_h, \mathbf{x}_v) && (24) \\
&= H(Q) + E_Q\{ \log P(\mathbf{x}_h, \mathbf{x}_v) \} && (25)
\end{aligned}
$$

where $H(Q)$ is the entropy of the variational distribution and $E_Q\{\cdot\}$ represents the expectation with respect to $Q(\mathbf{x}_h)$ (the observed variables $\mathbf{x}_v$ remain fixed to their instantiated values). Note that the variational distribution Q tries to balance two competing goals: assign values to the hidden variables $\mathbf{x}_h$ that have high probability under $P(\mathbf{x}_h, \mathbf{x}_v)$ (second term) and at the same time entertain a large number of distinct assignments (the entropy term).

Now, feasibility of evaluating $J(Q)$ depends on two types of structure. First, the graph structure (factorization) of the original probability model $P(\mathbf{x}_h, \mathbf{x}_v)$ and, second, any structure imposed on the variational distribution $Q(\mathbf{x}_h)$. We start by exploiting the structure in the original probability model: suppose, for simplicity, that $P(\mathbf{x}_h, \mathbf{x}_v)$ factorizes across the edges in the graph[3] as in equation (16). In this case, $\log P(\mathbf{x}_h, \mathbf{x}_v)$ in the above expectation reduces to a sum of simpler terms

$$
\begin{aligned}
J(Q) &= H(Q) + E_Q\{ \log \prod_{e \in E} \Psi(\mathbf{x}_e) \} && (26) \\
&= H(Q) + E_Q\{ \sum_{e \in E} \log \Psi(\mathbf{x}_e) \} && (27) \\
&= H(Q) + \sum_{e \in E} \sum_{\mathbf{x}_{e \cap h}} Q(\mathbf{x}_{e \cap h}) \log \Psi(\mathbf{x}_e) && (28)
\end{aligned}
$$

where $Q(\mathbf{x}_{e \cap h})$ is the variational marginal probability over the variables associated with edge e insofar as they are hidden. Note that for notational clarity we have dropped here explicit references to hidden/observed variables. The resulting objective above seems simpler than what we started from. However, we have merely transformed it and can still recover the exact solution if we maximize the objective with respect to the variational distribution Q. Again, the benefit arises from further constraining the solution or the variational distribution Q. This is the second type of structure that we need.

In the context of finite element methods (section 2.2), the approximation was in terms of a linear basis functions. In case of probability distributions, the appropriate simplification comes from independence properties. The

[3]Note that we may not be able to evaluate the partition function of such a joint. The variational objective $J(Q)$ will therefore be a constant away from the desired log-marginal.

simplest family of variational distributions is one where all the hidden variables $\{x_i\}_{i\in h}$ are independent of each other. More precisely, we assume that [37, 10, 7, 15, 41]:

$$Q(\mathbf{x}_h) = \prod_{i\in h} Q_i(x_i) \tag{29}$$

While this is a very simple class of distributions, we still have $|h|(r-1)$ degrees of freedom for adjusting the variational marginals $\{Q_i(x_i)\}_{i\in h}$.

Surely we should now be able to evaluate $J(Q)$? Indeed, by the fact that entropy is additive across independent variables, we get

$$J(Q) \;=\; \sum_{i\in h} H(Q_i) + \sum_{e\in E} \sum_{\mathbf{x}_{e\cap h}} Q(\mathbf{x}_{e\cap h}) \, \log \Psi(\mathbf{x}_e) \tag{30}$$

The evaluation of the first summation scales like $O(|h|r)$ where $|h|$ is the number of hidden variables and r is the number of distinct values each variable can take. Analogously, evaluating the second summation term scales like $O(|E|r^2)$ since each expectation over $\mathbf{x}_{e\cap h}$ involves (at most) two variables and there are $|E|$ edges. In our fully factored distribution Q, the marginal probability over the variables associated with each edge are obtained simply by picking the right two components from the product $\prod_{i\in h} Q_i(x_i)$. For more general distributions, obtaining such marginals may involve considerable effort. In particular, this is true by assumption for the posterior distribution $P(\mathbf{x}_h|\mathbf{x}_v)$.

4.0.1 Updating the mean field distribution

Having succeeded in evaluating the objective function for any (restricted) variable distribution Q, we still need to optimize the marginals. In the context of finite element methods, we could easily solve for the optimal linear coefficients. This is no longer true in our setting here and we have to resort to iterative methods for maximizing the objective function $J(Q)$ within the class of factored variational distributions Equation (29). Since the marginals in $Q(\mathbf{x}_h) = \prod_{i\in h} Q_i(x_i)$ can be adjusted independently, we can optimize $J(Q)$ one marginal component at a time.

We need a bit of notation. As before, let $E_Q\{\cdot\}$ stand for the expectation with respect to the variational distribution Q. Similarly, let $E_Q\{\cdot|x_k\}$ be the conditional expectation with respect to Q. Since we will make frequent use of such conditional expectations, we provide here a more explicit illustration:

$$E_Q\{\,\log P(\mathbf{x}_h, \mathbf{x}_v)\,|x_k\} = \sum_{\{x_i\}_{i\in h\setminus k}} \left[\prod_{i\in h\setminus k} Q_i(x_i)\right] \log P(\mathbf{x}_h, \mathbf{x}_v) \tag{31}$$

where, e.g., $h \setminus k$ is the set of hidden nodes other than k. Note that the expectation specifically does not depend on the variational marginal $Q_k(\cdot)$ over x_k; the result is, however, a function of the conditioning variable x_k.

To update the k^{th} variational marginal, we view $J(Q)$ as a function of $Q_k(\cdot)$ while keeping the remaining marginals fixed. To emphasize this, we may treat the entropy terms corresponding to remaining marginals as constants and appeal to the linearity of expectation $E_Q\{\cdot\} = \sum_{x_k} Q_k(x_k) E_Q\{\cdot|x_k\}$ to get

$$J(Q) = \text{const.} + H(Q_k) + \sum_{x_k} Q_k(x_k) E_Q\{ \log P(\mathbf{x}_v, \mathbf{x}_h) \mid x_k \} \tag{32}$$

where the dependence of $J(Q)$ on the marginal $Q_k(x_k)$ is explicit. It is easy to verify via straightforward calculation that maximizing this objective with respect to the marginal $Q_k(x_k)$ gives the standard Gibbs' distribution (cf. [13]):

$$Q_k(x_k) \leftarrow \frac{1}{Z_k}\, e^{E_Q\{\log P(\mathbf{x}_h, \mathbf{x}_v)|x_k\}} \tag{33}$$

for $x_k \in \{0, \ldots, r-1\}$. Here Z_k is the local normalization constant (partition function).

$$Z_k = \sum_{x_k} e^{E_Q\{\log P(\mathbf{x}_h, \mathbf{x}_v)|x_k\}} \tag{34}$$

These update equations, collectively for all k, are the *mean field equations* (cf. [41]). Successive application of the updates correspond to iteratively enforcing different mean field equations. Note that since each update is carried out in closed form, the updates monotonically increase the objective function $J(Q)$. We cannot, however, necessarily find the best factored variational approximation. This rather unfortunate property follows from the fact that although $J(Q)$ is concave in Q, it is not jointly concave in the new restricted parameterization in terms of the marginals $\{Q_i(x_i)\}_{i \in h}$. The order in which the iterative updates are carried out as well as the initialization of the marginals affect which of the locally optimal solution we arrive at.

Finally, let us briefly explicate in more detail the feasibility of evaluating the conditional expectations in the updates. For this purpose, let $P(\mathbf{x}_v, \mathbf{x}_h)$ factor across the edges in the graph, i.e., $P(\mathbf{x}_v, \mathbf{x}_h) = \prod_{e \in E} \Psi(\mathbf{x}_e)$, as before. Similarly to equation (30), we can write

$$E_Q\{\log P(\mathbf{x}_h, \mathbf{x}_v)|x_k\} = \sum_{e \in E} \sum_{\mathbf{x}_{e \cap \{h \setminus k\}} \mathbf{x}_{h \setminus k}} Q(\mathbf{x}_{e \cap \{h \setminus k\}}) \log \Psi_e(\mathbf{x}_e) \tag{35}$$

where $e \cap \{h \setminus k\}$ is either an empty set or refers to a single hidden node $k' \neq k$ associated with edge e. Thus, $Q(\mathbf{x}_{e\cap\{h\setminus k\}})$ is either one or the single marginal $Q_{k'}(x_{k'})$. Since there can be only n edges that pertain to node k, the complexity of evaluating the conditional expectation is at most $O(nr^2)$.

4.1 Quality of variational approximation

The variational mean field approximation we have explained above is arguably rough. It uses a completely factored distribution to approximate the posterior distribution $P(\mathbf{x}_h|\mathbf{x}_v)$ which may possess strong dependencies among the hidden variables. We explore here briefly the question of when this approximation is likely to be reasonable and when we can expect it to fail.

There are in fact two measures of accuracy that we can use. One is the tightness of the lower bound on the marginal probability of observed data that we set out to compute in the first place. In other words, we can take the difference $\log P(\mathbf{x}_v) - J(Q)$ as a figure of merit for the approximation. Without any constraints on the variational distribution Q, this difference would vanish but is unlikely to do so with the factored mean field distribution. The other measure we can use pertains to how closely the variational marginals $\{Q_i(x_i)\}$ match the true posterior marginals $P(x_i|\mathbf{x}_v)$. Since maximizing $J(Q)$ with respect to Q is equivalent to minimizing the KL-divergence between Q and the true posterior, it is reasonable to expect that the marginals aspire to be close as well. In the example below, however, we demonstrate that these two measures need not be strongly coupled.

We start by discussing in broad terms when we can expect the variational approximation to be accurate (cf. [23, 18]). Clearly, if in the posterior distribution the hidden variables are almost independent of each other, the variational approximation should be nearly perfect (we could, after all, closely represent the true posterior with a factored variational distribution). When this (strong) independence assumption no longer holds, we can expect either accuracy measure to degrade rapidly. Consider, for example, a mixture of two or more almost identical factored distributions. When the components become more distinct, the factored variational distribution can only represent one of the components, not the dependencies arising from switching between them.

A particularly important setting where almost factored distributions arise is a large densely connected graph model where the (pairwise) couplings between the variables are relatively weak. The net effect from a large number of fairly weak influences impinging on each variable converges, by the law of the large numbers, to a "mean effect". As a result, the variables become nearly independent of each other. This averaging effect underlies some of the success of mean field methods in large physical systems.

An important though rather undesirable property of the naive mean field approximation is that it exhibits *spontaneous symmetry breaking.* This happens when the optimal setting of the variational marginals is asymmetric even when the variables play a symmetric role in the posterior distribution. The symmetry breaking and more generally the selection of one of the posterior modes accounts for sometimes poor correspondence between the variational and true posterior marginals. The example below is specifically geared towards clarifying this issue.

4.1.1 Example

For simplicity, we assume a joint distribution over two binary (0/1) variables x_1 and x_2. Suppose, in addition, that both variables are hidden and there are no observed variables $\mathbf{x}_v$. In the variational formalism developed earlier, the "marginal probability" that we are trying to compute is in this case simply the normalization constant:

$$\log \sum_{\mathbf{x}_h} P(x_1, x_2) = \log \sum_{x_1, x_2} P(x_1, x_2) = \log 1 = 0 \tag{36}$$

While there's no reason to compute this value approximately, the fact that it's value does not depend on the properties of the joint distribution, permits us to easily evaluate the accuracy of the lower bound $J(Q)$ as a function of controlled changes in the joint.

We add structure to our representation of $P(x_1, x_2)$ by introducing a single parameter p that controls how dependent the two binary variables are. The probability table can be found in Table 1. In particular, the parameter p signifies the probability mass assigned to two configurations $(x_1 = 1, x_2 = 0)$ and $(x_1 = 0, x_2 = 1)$ that are consistent with the XOR operation. The remaining probability mass is divided equally among the leftover configurations. Note that at $p = 0.5$ the joint distribution is uniform and can be therefore captured by the factored variational distribution. At $p = 1$, only the two XOR configurations have non-zero probability and any factored distribution fails to capture such deterministic dependence between the variables. By varying p from 0.5 to 1 we can study how the variational approximation degrades with stronger dependencies.

To obtain $J(Q)$, we can simply substitute the simple distributions into

$P(0,0) = (1-p)/2$	$P(0,1) = p/2$
$P(1,0) = p/2$	$P(1,1) = (1-p)/2$

Table 1: Symmetric XOR-dominated joint distribution over binary variables x_1 and x_2; the probability mass falling on the two XOR configurations is controlled by parameter p.

the more general formulas we derived earlier. This gives

$$J(Q) = H(Q_1) + H(Q_2) + \sum_{x_1,x_2=0}^{1} Q_1(x_1)Q_2(x_2)\log P(x_1,x_2) \quad (37)$$

where the factored variational distribution is $Q(x_1,x_2) = Q_1(x_1)Q_2(x_2)$. Similarly, we can exploit the update equations (fixed point equations) derived earlier:

$$Q_1(x_1) \leftarrow \frac{1}{Z_1} e^{E_Q\{\log P(x_1,x_2)|x_1\}} \quad (38)$$

$$= \frac{1}{Z_1} e^{Q_2(0)\log P(x_1,0)+Q_2(1)\log P(x_1,1)} \quad (39)$$

where the right hand side is evaluated for $x_1 = 0, 1$ while the other marginal $Q_2(x_2)$ is held fixed. The update rule for $Q_2(x_2)$ is analogous. For any $p \in [0.5, 1]$, we can obtain a mean field solution by iteratively employing the above update rules. As discussed earlier, the solution may depend on the initial conditions. Here the variational marginals were initialized with uniform distributions subject to slight random perturbations.

Now, tracking the mean field solutions as a function of increasing p demonstrates spontaneous symmetry breaking. First, up to a critical value p^*, the variational marginals remain fixed at $Q_1(x_1 = 1) = Q_2(x_2 = 1) = 0.5$. These match the true marginals which, by symmetry, are $P(x_i = 1) = 0.5$ regardless of the parameter value p. Beyond the critical value $p = p^*$, the mean field solution undergoes a symmetry breaking: the objective $J(Q)$ prefers a solution with unequal marginals Q_1 and Q_2. This symmetry breaking arises entirely from the approximation as the true marginals remain fixed. As we can see in Figure 2a), this phase transition has an adverse effect on the quality of the variational marginals: after p^* the variational marginals suddenly and rapidly diverge from 0.5. The effect is less pronounced and to a degree opposite for the objective function $J(Q)$; indeed, after the symmetry breaking, the rapid degradation of the lower bound slows down (see figure 2b)). This symmetry breaking was, after all, forced upon us to improve the lower bound $J(Q)$.

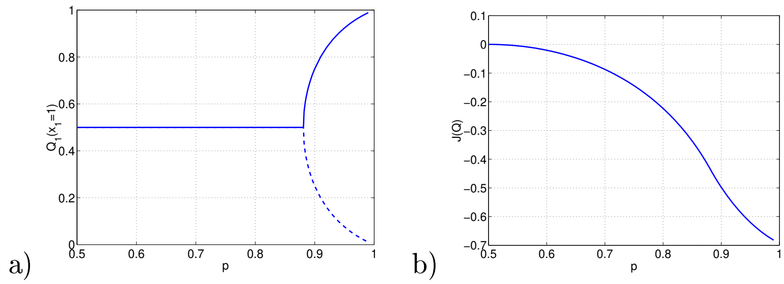

Figure 2: a) $Q_1(x_1 = 1)$ resulting from symmetry breaking as a function of the parameter p. The dashed line represents the alternative solution resulting from different initialization. b) the lower bound $J(Q)$ as a function of p.

While this example is simple and artificial it nevertheless provides us with some insight into larger problems as well. For example, note that the slope of the lower bound $J(Q)$ is zero when the joint distribution deviates from a factored distribution (p close to 0.5). Thus the naive mean field approximation appears insensitive to the introduction of weak dependencies. With larger deviations, however, the accuracy is lost at an accelerating pace.

The example also shows that it can be difficult to guarantee that the variational marginals $\{Q_i(x_i)\}$ reflect the true marginals. Even though in our simple case, it took fairly strong dependencies (large values of p) to induce the phase transition, more realistic problems with a large number of variables and associated dependencies offer considerably more ways of initiating such symmetry breaking. This effect is also not limited to symmetries but persists more generally when the posterior involves a number of competing modes; the variational marginals will typically reflect only the marginals of one of the modes.

The structured variational approach [43] discussed in the next section is less susceptible to these errors.

5 Structured variational approach

While the simple variational mean field approach is computationally attractive, it may not yield sufficiently accurate results. A natural approach to improving over the simple mean field method is to combine it with exact probabilistic calculations [43, 19, 24, 2, 47] (for other extensions see [20, 4]). In other words, we may be able to identify tractable substructures such as chains or trees within the larger graph model and these substructures could be readily handled with exact methods. A viable approach would be to impose a mean field approximation *between* the substructures while resorting to exact calculations *within* each substructure.

The first problem is to identify the substructures. This is a non-trivial problem for which no serious automated solutions have been proposed (cf. [19]). We will therefore assume that there are m tractable substructures identified by an expert or obtained via other means. Let the sets of nodes corresponding to these substructures be $h_1, \ldots, h_m$; the substructures are induced subgraphs over these sets. We assume also that the substructures create a disjoint partition of all the hidden variables: $h_i \cap h_j = \emptyset$ for $i \neq j$ and $h = h_1 \cup \ldots \cup h_m$.

The second problem is to ensure that we indeed apply exact probabilistic calculations within each subgraph in the variational framework. This is achieved by not introducing any constraints on the variational distribution Q within each substructure. In other words, the variational distribution must be composed of unconstrained components $\{Q_k(\mathbf{x}_{h_k})\}_{k=1,\ldots,m}$.

Finally, we wish to impose a mean field approximation across the substructures. This is equivalent to requiring that the variational distribution Q factors across the substructures. Consequently, we assume

$$Q(\mathbf{x}_h) = \prod_{k=1}^{m} Q_k(\mathbf{x}_{h_k}) \tag{40}$$

without any additional constraints.

5.1 Update equations

The update equations resulting from the structured approximation are exactly analogously to simple mean field. The intuition here is that we can always interpret the structured mean field method as a mean field approach over "mega variables" $\mathbf{x}_{h_k}$. Thus each variational marginal $Q_k(\mathbf{x}_{h_k})$ is updated according to

$$Q_k(\mathbf{x}_{h_k}) \leftarrow \frac{1}{Z_{h_k}}\, e^{E_Q\{\,\log P(\mathbf{x}_v,\mathbf{x}_h)\,|\,\mathbf{x}_{h_k}\,\}} \tag{41}$$

where the conditional expectation is defined and computed analogously to mean field. Can these updates be carried out efficiently? This depends on whether the joint distribution, $P(\mathbf{x}_v, \mathbf{x}_h)$, corresponding to a graph G

has tractable[4] induced subgraphs over the sets h_k. The following example illustrates this in more detail.

Suppose the probability model $P(\mathbf{x}_v, \mathbf{x}_h)$ consist of m coupled Markov (Boltzmann) chains as shown in Figure 3 (see [42, 11]). In a mean field approximation, the variables within and across each chain would be assumed to be independent of each other. Since each Markov chain individually is perfectly tractable, we can improve the mean field approximation considerably by decoupling only the variables across the chains. Whenever the chains in the original probability model are only loosely coupled, we would expect this structured mean field approach to be quite accurate.

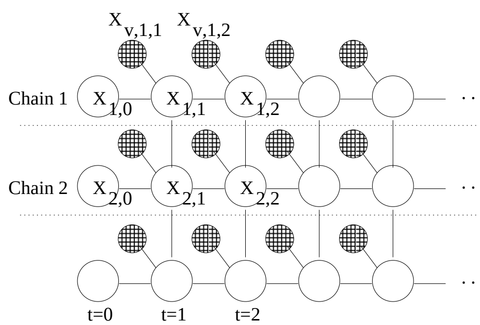

Figure 3: Coupled Boltzmann chains. The shaded smaller nodes denote observed variables.

To develop this further, let $\mathbf{x}_{v_k} = \{x_{v,k,0}, \ldots, x_{v,k,T}\}$ be the observation sequence for the k^{th} chain, and, collectively, $\mathbf{x}_v = \{\mathbf{x}_{v_1}, \ldots, \mathbf{x}_{v_m}\}$. Similarly, let $\mathbf{x}_{h_k} = \{x_{k,0}, \ldots, x_{k,T}\}$, be the sequence of hidden states corresponding to the k^{th} Markov chain or substructure. If the chains were not coupled, the probability distribution governing the variables within each chain would have the following familiar form

$$P(\mathbf{x}_{h_k}, \mathbf{x}_{v_k}) = \frac{1}{Z_k} \prod_{t=1}^{T} \Psi_k^h(x_{k,t-1}, x_{k,t}) \Psi_k^v(x_{k,t}, x_{v,k,t}) \tag{42}$$

where the potential $\Psi_k^h(x_{k,t-1}, x_{k,t})$ links the successive hidden variables in time while $\Psi_k^v(x_{k,t}, x_{v,k,t})$ connects the observation at time t, $x_{v,k,t}$, to the corresponding hidden state variable, $x_{k,t}$. For simplicity, we will refer to this tractable chain structure with a single potential function $\Psi_k(\mathbf{x}_{v_k}, \mathbf{x}_{h_k})$.

[4] In general, we would have to consider also the portion of the graph G connecting the substructures. We assume here that the coupling between the substructures is sparse.

Now, the joint distribution over all the chains and observations, including the couplings between the chains, is given by

$$P(\mathbf{x}_h, \mathbf{x}_v) = \frac{1}{Z} \left[\prod_{k=1}^{m} \Psi_k(\mathbf{x}_{v_k}, \mathbf{x}_{h_k}) \right] \left[\prod_{t=1}^{T} \prod_{k=2}^{m} \phi_{k-1,k}(x_{k-1,t}, x_{k,t}) \right] \quad (43)$$

Here the first term represents independent chains and the second product term quantifies the couplings between the state variables in neighboring chains.

To demonstrate that the structured mean field approach is tractable in this context, it remains to evaluate the conditional expectations $E_Q\{\log P(\mathbf{x}_v, \mathbf{x}_h)|\mathbf{x}_{h_k}\}$ in equation (41). In computing these expectations, we can safely ignore all the terms that do not depend on the conditioning variables $\mathbf{x}_{h_k}$; these terms will automatically vanish during normalization. For the k^{th} chain, the only relevant components of the joint distribution are the interactions within the k^{th} chain and the couplings between it and the neighboring chains $k-1$ and $k+1$. Thus,

$$\begin{aligned}
&E_Q\{\log P(\mathbf{x}_v, \mathbf{x}_h)|\mathbf{x}_{h_k}\} \\
&= \text{const.} + \log \Psi_k(\mathbf{x}_{v_k}, \mathbf{x}_{h_k}) \\
&\quad + \sum_t E_{Q_{k-1}}\{\log \phi_{k-1,k}(x_{k-1,t}, x_{k,t})\} \\
&\quad + \sum_t E_{Q_{k+1}}\{\log \phi_{k,k+1}(x_k, x_{k+1,t})\} && (44) \\
&= \text{const.} + \log \Psi_k(\mathbf{x}_{v_k}, \mathbf{x}_{h_k}) + \sum_t \log \tilde{\phi}_k(x_{t,k}) && (45)
\end{aligned}$$

where the expectations $E_{Q_{k-1}}\{\cdot\}$ and $E_{Q_{k+1}}\{\cdot\}$ are taken with respect to the variational marginals over the state variables in chains $k-1$ and $k+1$, respectively. In the last expression, we have collected together the contributions from the neighboring chains into effective terms $\log \tilde{\phi}_k(x_{t,k})$.

As a result, the structured mean field updates are given by

$$Q_k(\mathbf{x}_{h_k}) \leftarrow \frac{1}{Z_{h_k}} \Psi_k(\mathbf{x}_{v_k}, \mathbf{x}_{h_k}) \times \prod_t \tilde{\phi}_k(x_{t,k}) \quad (46)$$

where the additional terms beyond the original chain interactions provide independent evidence to individual state variables $x_{k,0}, \ldots, x_{k,T}$. This does not change the structure of the original distribution ($\tilde{\phi}_k(x_{t,k})$ could be simply absorbed into $\Psi_k^v(x_{k,t}, x_{v,k,t})$). No significant loss in tractability is therefore incurred due to the influence from the other chains in this structured mean field approximation.

We emphasize that the interactions within the substructures, i.e., $\Psi_k(\mathbf{x}_{v_k}, \mathbf{x}_{h_k})$, remained unaffected by the updates. Thus the optimal variational marginal within each substructure maintains the original strength of dependencies in addition to the interaction structure. The influences between the substructures are mediated by the effective potentials, $\tilde{\phi}$, which, in case of pairwise couplings between the substructures, appear as additional biases on the individual variables. For a related discussion, see [47].

6 Local variational approach

The variational mean field approximation that we introduced in previous sections relies on a suitable additional structure in the probability model. This additional structure was expressed in terms of additional factoring of the joint distribution beyond what is dictated by the graph (e.g., pairwise potential functions). In the absence of such factorization, we may still find useful structure in the probability model. For example, the conditional probabilities in the directed graph model or the potential functions in the undirected models may possess parametric structure that we can exploit in approximate inference calculations.

As an example, consider the noisy-OR probability model [38] over binary (0/1) variables, where the interactions between the variables are defined in terms of probabilistic generalizations of the OR function. The conditional probabilities in these directed graph models are given by

$$P(x_i|x_{pa_i}, \theta_i) = f_{x_i}\left(\theta_{i0} + \sum_{j \in pa_i} \theta_{ij} x_j\right) \tag{47}$$

In other words, we pass a linear combination of the parents, $z = \theta_0 + \sum_{j \in pa_i} \theta_{ij} x_j$ through an appropriate *transfer* function $f_{x_i}(z)$, where $f_0(z) = \exp(-z)$ and $f_1(z) = 1 - \exp(-z)$. Note that $f_0(z) + f_1(z) = 1$ for any input z as required. By setting $\theta_{i0} = 0$ and increasing all θ_{ij}, we recover the OR function in the limit: $f_1(z) \rightarrow OR(\{x_j\}_{j \in pa_i})$.

The local conditional probabilities (or potentials) $P(x_i|x_{pa_i})$ depend on $|pa_i| + 1$ variables. As the number of parents increases, these potentials cannot be used efficiently in the mean field approximation, at least not directly as stated above. The cost of dealing with such component potentials would be exponential in the number of variables they depend on, i.e., $|pa_i| + 1$. We attempt here to exploit the parametric form of these conditional probabilities to impose additional factorization. Ideally, we would like to get

$$P(x_i|x_{pa_i}) \approx \prod_{j \in pa_i} \Psi_{ij}(x_i, x_j) \tag{48}$$

since in the product form the parents of x_i are decoupled. Selective use of such factorization transformations may render the remaining (approximate) joint distribution tractable [16]. Alternatively, we may exploit the resulting factorization as a part of mean field or structured mean field approximation.

It remains to show how such factorization can be achieved. We will introduce a class of variational methods that are closely related to large deviation methods for this purpose (for a direct application of large deviation theory towards approximate inference see [26, 27]). The approximate factorization is provided in terms of upper or lower bounds rather than uncontrolled approximations. We start with an example from large deviation theory (see, e.g., [5]).

6.1 Large deviation example

Suppose we wish to derive a standard large deviation result for a sum of n independent and identically distributed binary (0/1) variables $x_1, \ldots, x_n$. The tails of the distribution governing the sum vanish exponentially fast. We wish to capture the probability that the sum deviates from its expected value np_0 by more than $n\epsilon$ for arbitrary $\epsilon > 0$. Here p_0 is the generative probability for the event that $x_i = 1$ for any i. Consider the following one-sided probability:

$$P\left(\sum_{i=1}^{n} x_i \geq n(p_0 + \epsilon)\right) = E_{p_0}\left\{ \text{step}\left(\sum_{i=1}^{n} x_i - n(p_0+\epsilon)\right)\right\} \tag{49}$$

where where the expectation is taken with respect to the product distribution over $x_1, \ldots, x_n$ and step $(\, z\,) = 1$ for $z \geq 0$ and zero otherwise. The step function inside the expectation captures the appropriate event. We can also interpret the step function as a transfer function $f_1(z) = \text{step}\,(\,(\,)\,z)$ analogously to the noisy-OR model discussed above. The above large-deviation probability can be therefore viewed as a marginal probability (marginalized over the parents) of a binary variable.

Even in this simple case, however, we are unable to obtain a closed form expression for this expectation. On the other hand, evaluating the expected value of any factored approximation $\prod_i \Psi(x_i - (p_0 + \epsilon))$ with respect to the product distribution could be done efficiently on a term by term basis (as a product of expectations with respect to individual binary variables). To turn the original expectation into such factored form, we will make use of the following variational transformation of the step function:

$$\text{step}\,(\,z\,) = \min_{\lambda \geq 0} \exp(\,\lambda z\,) \tag{50}$$

where λ serves as a variational parameter. To understand this transformation note that when $z < 0$, increasing λ decreases $\exp(\lambda z)$ since the exponent is negative. Letting $\lambda \to \infty$, results in $\exp(\lambda z) \to 0$, as desired. On the other hand, when $z \geq 0$, $\exp(\lambda z)$ is minimized by setting $\lambda = 0$. This gives $\exp(0 \cdot z) = 1$. Note that the optimal setting of the variational parameter is a function of z. For this function $\lambda^*(z)$, $\text{step}\,(\,z\,) = \exp(\,\lambda^*(z)\,z\,)$.

The above transformation is exact and therefore not yet useful to us. Similarly to other variational methods, however, we can obtain a controlled approximation by restricting the choice of the variational parameters. Here we require that the choice of the variational parameter as a function of z, i.e., $\lambda(z)$, must be a constant: $\lambda(z) = \hat{\lambda}$ for all values of z. This gives a simple upper bound on the step function[5]

$$\text{step}\,(\,z\,) \leq \exp(\,\hat{\lambda}\,z\,),\ \forall z \tag{51}$$

The usefulness of this bound is immediate in the large deviation context:

$$\begin{aligned}
\text{step}\left(\sum_{i=1}^{n} x_i - n(p_0+\epsilon)\right) &\leq \exp\left(\hat{\lambda}[\sum_{i=1}^{n} x_i - n(p_0+\epsilon)]\right) && (52)\\
&= \prod_{i=1}^{n} \exp\left(\hat{\lambda}[x_i - (p_0+\epsilon)]\right) && (53)\\
&= \exp(-n\hat{\lambda}(p_0+\epsilon)) \prod_{i=1}^{n} \exp(\,\hat{\lambda} x_i) && (54)
\end{aligned}$$

Since the variables x_i are independent we can evaluate the expectation of the right hand side with respect to the product distribution on a term by term basis. Moreover, all such expectations are identical since $x_1, \ldots, x_n$ are identically distributed. This gives

$$\begin{aligned}
P\left(\sum_{i=1}^{n} x_i \geq n(p_0+\epsilon)\right) &\leq \exp(-n\hat{\lambda}(p_0+\epsilon)) \left[E_{p_0} \exp(\,\hat{\lambda} x_i)\right]^n && (55)\\
&= \exp(-n\hat{\lambda}(p_0+\epsilon)) \left[p_0 \exp(\hat{\lambda}) + 1 - p_0\right]^n && (56)
\end{aligned}$$

where the last expression comes from taking the expectation with respect to a Bernoulli distribution $P(x_i = 1) = p_0$. We can improve this result by utilizing the degree of freedom that we have in choosing $\hat{\lambda}$. The optimal choice for $\hat{\lambda}$ is found by minimizing the resulting bound:

$$\log P\left(\sum_{i=1}^{n} x_i \geq n(p_0+\epsilon)\right)$$

$$\leq \min_{\hat{\lambda}\geq 0}\left(-n\hat{\lambda}(p_0+\epsilon)+n\log\left[p_0\,\exp(\hat{\lambda})+1-p_0\right]\right) \tag{57}$$

$$= -n\cdot\max_{\hat{\lambda}\geq 0}\left(\hat{\lambda}(p_0+\epsilon)-\log\left[p_0\,\exp(\hat{\lambda})+1-p_0\right]\right) \tag{58}$$

where in the last expression we pulled the negative sign from within the minimization, turning it into a maximization. The term obtained through the maximization is precisely the large deviation *rate function* (see, e.g., [5]). Basic information theoretic bounds (specifically, Chernoff bounds) result from such simple factorization transformations.

6.2 Representation theorem

To exploit such factorization transformations more generally in probabilistic inference calculations, we would need to find the appropriate variational transformation for any given situation. Do such transformations even exist for any given family of conditional probabilities? Perhaps surprisingly, this question can be answered affirmatively: the factorization transformation *always* exists. The following theorem makes this more precise

Theorem 1 *Let $P(x_i|\mathbf{x}_{pa_i})$ be a conditional probability model over x_i taking values in a finite set. We assume further that the number of possible instantiations of the parents $\mathbf{x}_{pa_i}$ is finite. Let λ be a variational parameter taking values in a finite or finitely dimensional set F. Then there exists non-negative pairwise potentials*

$$\overline{\Psi}_j(x_i,x_j|\lambda),\;\; \underline{\Psi}_j(x_i,x_j|\lambda)\;\;\forall j\in pa_i \tag{59}$$

such that

$$P(x_i|\mathbf{x}_{pa_i}) = \max_{\lambda\in F}\prod_{j\in pa_i}\underline{\Psi}(x_i,x_j|\lambda) = \min_{\lambda\in F}\prod_{j\in pa_i}\overline{\Psi}(x_i,x_j|\lambda) \tag{60}$$

for all $(x_i,\mathbf{x}_{pa_i})$.

We emphasize that this is merely an existence proof and does not mean that we can find any useful transformations, those that lead to efficient and accurate approximate inference. Finding a suitable transformation for any specific family of conditional probabilities (apart from the log-concave class of generalized linear models discussed below) remains an open problem.

6.3 Example: log-concave models

Useful variational transformations of conditional probabilities leading to additional factorization can be found systematically for a log-concave class of generalized linear models[16, 21, 17]. This family of conditional probabilities includes, e.g., noisy-OR and logistic regression models. More precisely, it is characterized by conditional probabilities of the form

$$P(x_i|x_{pa_i}, \theta_i) = f_{x_i}\left(\theta_{i0} + \sum_{j \in pa_i} \theta_{ij} x_j\right) \tag{61}$$

where the transfer function $f_{x_i}(\cdot)$ is log-concave: $\log f_{x_i}(z)$ is a concave function of its argument z for all values of x_i. We will exploit both the concavity property and the linear predictive structure.

We start by noting that the product decomposition in Equation (48) is equivalent to an additive decomposition on the log-scale. In other words, to achieve $P(x_i|x_{pa_i}, \theta_i) \approx \prod_{j \in pa_i} \Psi_{ij}(x_i, x_j)$, it suffices to find the following additive approximation in our context

$$\log f_{x_i}\left(\theta_{i0} + \sum_{j \in pa_i} \theta_{ij} x_j\right) \approx \sum_{j \in pa_i} \psi_{ij}(x_i, x_j) \tag{62}$$

(simply choose $\psi_{ij}(x_i, x_j) = \log \Psi_{ij}(x_i, x_j)$ to preserve equality). Now, since the argument of $\log f_{x_i}(\cdot)$ here already has the desired additive structure, we merely need to find a linear approximation to $\log f_{x_i}(\cdot)$. The fact that $\log f_{x_i}(z)$ is also concave guarantees that we can find a linear upper bound approximation via first order Taylor expansion. Figure 4 illustrates this for the log-logistic function. For example, expanding $\log f_1(z)$ around any point z_0 gives

$$\begin{aligned}
\log f_1(z) &\leq \frac{\partial \log f_1(z)}{\partial z}\Big|_{z=z_0} (z - z_0) + \log f_1(z_0) && (63)\\
&= \frac{\partial \log f_1(z)}{\partial z}\Big|_{z=z_0} z - \left[\frac{\partial \log f_1(z)}{\partial z}\Big|_{z=z_0} z_0 - \log f_1(z_0)\right] && (64)\\
&= \lambda_1 z - F_1(\lambda_1) && (65)
\end{aligned}$$

where $\lambda_1 = \partial \log f_1(z)/\partial z$. For concave (convex) differentiable functions, the offset in the brackets or $F_1(\lambda_1)$ can indeed be expressed in terms of the gradient λ_1 [5]. Note here that varying the point of expansion, z_0, is

[5]Note, for example, that for strictly concave differentiable functions, the gradient is a monotonically decreasing function and therefore invertible. Any point z_0 in our example can be expressed as a function of the gradient λ_1 evaluated at z_0.

equivalent to varying λ_1 in the gradient space. We may therefore take λ_1 as the variational parameter without explicitly referring to z_0. This simple explanation captures a more general duality property of concave (convex) functions[39]: any concave function such as $\log f_1(z)$ has a *conjugate* or *dual* function $F_1(z)$, also concave, such that

$$\log f_1(z) = \min_{\lambda_1}\{\lambda_1 z - F_1(\lambda_1)\} \tag{66}$$

where λ_1 takes values in the domain of $F_1(\cdot)$. The duality comes from the fact that $F_1(\lambda_1)$ as a concave function can be similarly expressed in terms of $\log f_1(z)$ (the conjugate of the conjugate function is the function itself).

Finally, substituting our linear upper bound from Equation (65) for the log-conditional probability (separately for each x_i) gives

$$\log f_{x_i}\left(\theta_{i0} + \sum_{j\in pa_i} \theta_{ij}x_j\right) \leq \lambda_{x_i}\left(\theta_{i0} + \sum_{j\in pa_i} \theta_{ij}x_j\right) - F_{x_i}(\lambda_{x_i}) \tag{67}$$

The additive expansion follows from identifying $\psi(x_i, x_j) = \lambda_{x_i}\theta_{ij}x_j$ and absorbing the remaining terms into one of such potentials. This is a variational transformation and comes with an adjustable parameter(s) λ_{x_i} that can be used to optimize the approximation in the appropriate context, just as in the large deviation example. Table 2 explicates such transformations for typical members of the log-concave family.

Name	$\log f(z)$	Conjugate function $F(\lambda)$	Domain for λ
Noisy-OR	$\log(1-\exp(-z))$	$(1+\lambda)\log(1+\lambda) - \lambda\log\lambda$	$[0,\infty]$
Logistic	$-\log(1+\exp(-z))$	$-\lambda\log\lambda - (1-\lambda)\log(1-\lambda)$	$[0,1]$

Table 2: Upper bound variational transformations for noisy-OR and logistic functions.

7 Parameter estimation with variational methods

We explain here how the variational lower bound on the marginal likelihood discussed earlier can be used for maximum likelihood (ML) parameter estimation. This variational approach leads to the standard EM-algorithm [8] with another maximization step taking the place of the original E-step. The variational approach remains applicable, however, even when the E-step in the EM-algorithm can no longer be computed exactly and guarantees monotonically increasing sequence of lower bounds on the log-likelihood.

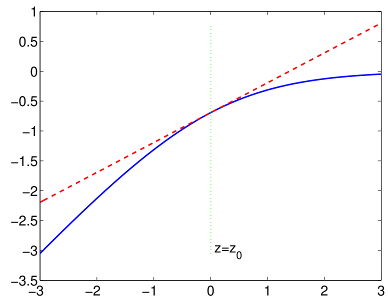

Figure 4: A concave function (log of the logistic function) and its linear (variational) upper bound.

To fix ideas, let $D = \{\mathbf{x}_v^1, \ldots, \mathbf{x}_v^T\}$ be a set of i.i.d observations. We assume for notational simplicity that the set of observed variables is the same throughout the observations. In other words, we can use the same division between observed and hidden variables $\mathbf{x} = \{\mathbf{x}_v, \mathbf{x}_h\}$ for all data points. Our goal is to maximize the log-likelihood of the data D:

$$J(\theta) = \sum_{t=1}^{T} \log P(\mathbf{x}_v^t|\theta) \tag{68}$$

where θ denotes the adjustable parameters in the joint distribution $P(\mathbf{x}_v, \mathbf{x}_h|\theta)$. We assume that the parameter estimation problem can be carried out efficiently when the observations are complete. To transform the above log-likelihood objective $J(\theta)$ into a form that involves only complete data, we introduce a separate variational transformation for each of the log-marginal probabilities in the above sum. This gives

$$\begin{aligned} J(\theta) &\geq \sum_{t=1}^{T} \left[\sum_{\mathbf{x}_h} Q_t(\mathbf{x}_h) \log P(\mathbf{x}_v^t, \mathbf{x}_h|\theta) + H(Q_t) \right] && (69) \\ &= \sum_{t=1}^{T} J(Q_t, \mathbf{x}_v^t; \theta) = J(Q_1, \ldots, Q_T; \theta) && (70) \end{aligned}$$

Recall that maximizing each $J(Q_t, \mathbf{x}_v^t; \theta)$ with respect to Q_t recovers the corresponding log-marginal likelihood or $\log P(\mathbf{x}_v^t|\theta)$. Thus by maximizing $J(Q_1, \ldots, Q_T; \theta)$ with respect to all the variational distributions $Q_1, \ldots, Q_T$, we recover the ML objective $J(\theta)$

$$\max_{Q_1,\ldots,Q_T} J(Q_1, \ldots, Q_T; \theta) = J(\theta) \tag{71}$$

Now, to take advantage of the variational formulation, we do not maximize $J(\theta)$ directly but instead maximize the variational objective $J(Q_1, \ldots, Q_T; \theta)$ in two alternating maximization steps[34]. In the first step, we maximize the variational objective with respect to the distributions $Q_1, \ldots, Q_T$ while keeping the parameters θ fixed. If no constraints are imposed on the variational distributions, we obtain $Q_t^*(\mathbf{x}_h) = P(\mathbf{x}_h|\mathbf{x}_v^t, \theta)$ for all t and the maximum value of the variational objective equals $J(\theta)$. In the second step, the variational distributions $Q_1, \ldots, Q_T$ remain fixed and we maximize the variational objective with respect to the parameters θ alone.

This two step max-max algorithm leads to a monotonically increasing log-likelihood of data. To see this, let's denote each maximization step by successively priming the corresponding parameters. We obtain the following chain of inequalities

$$\begin{aligned} J(\theta) = J(Q'_1, \ldots, Q'_T; \theta) &\leq J(Q'_1, \ldots, Q'_T; \theta') \\ &\leq J(Q''_1, \ldots, Q''_T; \theta') = J(\theta') \end{aligned} \tag{72}$$

Thus $J(\theta) \leq J(\theta')$, where the inequality is strict whenever either of the last two maximization steps could improve the variational objective $J(Q_1, \ldots, Q_T; \theta)$. If not, we have reached a local optimum.

The algorithm presented above is in fact precisely the standard EM-algorithm. The E-step of the EM-algorithm corresponds to the first maximization step with respect to the variational distributions $Q_1, \ldots, Q_T$. Indeed, this maximization step results in setting the variational distributions equal to the posterior probabilities over the hidden variables. Evaluation of the variational objective in Equation (69) with $Q_t(\mathbf{x}_h) = P(\mathbf{x}_h|\mathbf{x}_v^t, \theta)$ gives the expected complete log-likelihood of the data as in the E-step. The additional entropy terms in the variational objective are kept fixed during the second maximization step and are therefore inconsequential. See also [34].

Unlike the EM-algorithm, however, the variational formulation remains applicable even when we can no longer handle the posterior probabilities $P(\mathbf{x}_h|\mathbf{x}_v^t, \theta)$. Indeed, we can restrict the variational distributions $Q_1, \ldots, Q_T$ to be within, for example, a class of completely factored (mean field) distributions. The first maximization step will be therefore carried out incompletely, only within the restricted class. However, we can still guarantee a monotonically increasing lower bound $J(Q_1, \ldots, Q_T; \theta)$ on the log-likelihood $J(\theta)$ [10, 41]. Whether this guarantee suffices in practice depends on the accuracy of the (structured) mean field approximation.

8 Variational Bayesian methods

Parameter estimation within the Bayesian framework reduces to an inference problem, that of evaluating the posterior probability over the parameters given the observed data. One could therefore suspect that the variational framework we have developed earlier for approximate inference could be used in this context as well. While this is indeed the case, there are a couple of additional difficulties. First, the parameters (excluding the model structure) are typically continuous rather than discrete making it harder to represent the posterior probabilities. Second, each parameter setting needs to be evaluated across all the observed data, not merely in the context of a single observation. In computing the distribution over the parameters, the data points cannot be treated individually but rather as a set. Moreover, in the context of incomplete observations, it no longer suffices to infer the posterior probabilities over the hidden variables independently for each observation; the posteriors are contingent on a specific parameter setting and we must consider *all* such settings. Incomplete observations are therefore quite difficult to handle exactly within the Bayesian framework.

We start with the simpler setting where each observation is assumed to be complete, i.e., we have a value assignment for all the variables in the probability model. For a moment, we will drop the subindex v denoting the set of visible variables. The goal here is to evaluate the posterior probability over the parameters given the observed i.i.d. data:

$$P(\theta|D) = \frac{1}{P(D)} P(D|\theta)P(\theta) = \frac{1}{P(D)} \left[\prod_{t=1}^{T} P(\mathbf{x}^t|\theta) \right] P(\theta) \tag{73}$$

where $P(\theta)$ is the prior probability over the parameters and $P(D)$ is the marginal data likelihood:

$$P(D) = \int \left[\prod_{t=1}^{T} P(\mathbf{x}^t|\theta) \right] P(\theta) d\theta \tag{74}$$

Our ability to evaluate $P(D)$ determines whether the estimation problem is tractable. Computing $P(D)$ is the type of inference problem that we have already solved variationally. The relevant joint distribution is now $P(D,\theta) = P(D|\theta)P(\theta)$, which factors across the data points. Each component $P(\mathbf{x}^t|\theta)$ of this joint, must itself factor into smaller components for their product to remain tractable. When the observations are complete, this is indeed the case. If we assume, in addition, that we have distinct parameters associated with different factors, that such parameters are a priori independent of each other, and that the prior distributions are conjugate to the corresponding likelihoods, we can typically evaluate the marginal data likelihood in closed form (as in [14]). However, parameter independence and conjugate form for the priors may not reflect our prior knowledge. Other prior distributions and associated independence assumptions may necessitate approximate methods for evaluating the posteriors.

The typical approximate computations involve sampling methods [35]. While these are important and useful in various aspects of Bayesian calculations, we will not discuss them here. A number of excellent sources are available [36]. Our focus here is an alternative and to a degree complementary approach based on variational methods.

Formally, the application of the variational approach to a Bayesian parameter estimation problem is straightforward: we introduce a variational distribution $Q(\theta)$ over the parameters and evaluate a lower bound $J(Q)$ on the log-marginal likelihood of the data (cf. [30]):

$$\begin{aligned}\log P(D) &\geq H(Q_\theta) + \int Q(\theta) \log P(D|\theta) d\theta && (75)\\ &= H(Q_\theta) + \int Q(\theta) \log P(\theta) d\theta + \sum_t \int Q(\theta) \log P(\mathbf{x}^t|\theta) d\theta && (76)\end{aligned}$$

Without imposing any constraints on Q, however, we recover $\log P(D)$ by maximizing the lower bound $J(Q)$ with respect to the variational distribution. Moreover, at the maximum $Q^*(\theta) = P(\theta|D)$, as desired.

Additional factorization present in $P(\mathbf{x}^t|\theta)$ further simplifies the necessary expectations with respect to the variational distribution Q. For example, $P(\mathbf{x}|\theta)$ may factor according to a directed graph, permitting us to write it as $P(\mathbf{x}|\theta) = \prod_i P(x_i|x_{pa_i}, \theta_i)$, where each conditional probability depends on a distinct set of parameters θ_i. Now, so long as the prior distribution $P(\theta)$ factors across the parameters associated with the conditional probabilities, so does the posterior. We may therefore assume without loss of generality that $Q(\theta) = \prod_i Q_i(\theta_i)$. The variational lower bound reduces in this case to

$$\begin{aligned}\log P(D) \quad \geq \quad & \sum_i \Big[H(Q_i) + \int Q_i(\theta_i) \log P(\theta_i) d\theta_i \\ & + \sum_t \int Q_i(\theta_i) \log P(x_i^t|x_{pa_i}^t, \theta_i)\, d\theta_i \Big] \qquad (77)\end{aligned}$$

Of course, we can still recover the true marginal likelihood and the true posterior by maximizing this with respect to all the variational distributions $Q_i(\theta_i)$. In many cases, however, even the component posteriors $P(\theta_i|D)$

cannot be evaluated in closed form. This is, for example, the case with logistic regression models, where

$$P(x_i = 1|x^t_{pa_i}, \theta_i) = f_1\left(\theta_{i0} + \sum_{j \in pa_i} \theta_{ij} x_j\right) \tag{78}$$

and $f_1(z) = (1 + e^{-z})^{-1}$ is the logistic function. In this case we can still apply the variational formalism by constraining the variational posteriors $\{Q_i(\theta_i)\}$ to have simpler parametric forms such as multivariate Gaussian distributions. The variational lower bound $J(Q)$ can be evaluated in closed form if we combine this restriction with additional approximations of the following expectations

$$\begin{aligned} &\int Q_i(\theta_i) \log P(x^t_i|x^t_{pa_i}, \theta_i)\, d\theta_i \\ &= \; -\int Q_i(\theta_i) \log\left(1 + e^{-(2x^t_i - 1)(\theta_{i0} + \sum_{j \in pa_i} \theta_{ij} x_j)}\right) d\theta_i \end{aligned} \tag{79}$$

which can be efficiently lower bounded by taking the expectation inside the logarithm ($-\log(\cdot)$ is a convex function); see [41] and the references therein for a refined lower bound. We may also impose additional factorization of the logistic function (as alluded to earlier in this tutorial) or resort to transformations that are more specifically tailored to the logistic function[19, 18].

Bayesian estimation of parameters and hyper-parameters may also sometimes preclude exact computations. The prior distribution over the parameters $P(\theta_i)$ in this case is a marginal over some hyper-parameters α_i:

$$P(\theta_i) = \int P(\theta_i|\alpha_i) P(\alpha_i) d\alpha_i \tag{80}$$

and we wish to infer a posterior probability over both the parameters and hyper-parameters $P(\theta_i, \alpha_i|D)$. Whenever the marginal $P(\theta_i)$ cannot be evaluated in closed form, we may still rely on the variational approach provided that we restrict ourselves to factored variational distributions: $Q(\theta_i, \alpha_i) = Q(\theta_i)Q(\alpha_i)$ (see [9]). Our earlier assessment of the accuracy of the variational mean field approach applies to this case as well. We can expect this approach to be accurate whenever the parameters θ_i and the hyper-parameters α_i are only loosely coupled. However, as discussed earlier, it may be dangerous to use the resulting product of variational marginals $Q(\theta_i)Q(\alpha_i)$ as a proxy for the true posterior $P(\theta_i, \alpha_i|D)$, particularly if the true posterior contains multiple modes.

8.1 Incomplete cases

The situation becomes substantially more complex when there are incomplete cases in the data set. We start by making a few simplifying assumptions. First, we assume a fixed division between hidden and observed variables, $\mathbf{x} = \{\mathbf{x}_v, \mathbf{x}_h\}$, for all data points. We also refrain from discussing a) joint distributions $P(\mathbf{x}_v, \mathbf{x}_h|\theta)$ whose components are not in the exponential family as well as non-conjugate prior distributions. These aspects were discussed in the previous section and in the references therein. Finally, we will assume that for any fixed setting of the parameters θ, the posterior probabilities over the hidden variables $P(\mathbf{x}_h|\mathbf{x}_v^t, \theta)$ can be computed in a feasible manner (cf. [18, 1]).

Now, when the observed cases in the dataset are not complete, the likelihood term pertaining to the parameters still factors across the observations

$$P(D|\theta) = \prod_{t=1}^{T} P(\mathbf{x}_v^t|\theta) \tag{81}$$

but the components $P(\mathbf{x}_v^t|\theta) = \sum_{\mathbf{x}_h^t} P(\mathbf{x}_v^t, \mathbf{x}_h^t|\theta)$ may lack any further factorization[6]. The fact that we are force to infer both the posterior over the hidden configurations of variables and the parameters is a serious impediment. Even worse, the posteriors over the hidden variables corresponding to each observation depend on the specific setting of the parameters θ (i.e., $P(\mathbf{x}_h^t|\mathbf{x}_v^t, \theta)$). We can, however, still apply the variational framework so long as we explicitly remove such direct dependencies between the parameters and the hidden configurations. Put another way, we impose the following factored structure on the variational distribution[30, 19, 18, 1, 9]:

$$Q(\mathbf{x}_h^1, \ldots, \mathbf{x}_h^T, \theta) = Q_1(\mathbf{x}_h^t) \cdots Q_T(\mathbf{x}_h^T)\, Q(\theta) \tag{82}$$

The lower bound on the marginal data likelihood corresponding to this variational distribution can be obtained fairly easily. Since in the variational distribution the hidden variable configurations and the parameters are independent, we can introduce the variational lower bounds in two stages, first for the parameters and then for each of the marginals $\log P(\mathbf{x}_v^t|\theta)$. In other words,

$$\log P(D) \quad \geq \quad H(Q_\theta) + \int Q(\theta) \log P(\theta) d\theta + \sum_t \int Q(\theta) \log P(\mathbf{x}_v^t|\theta) d\theta$$

[6]Note that the hidden variables may affect only part of the model and therefore the marginal probabilities of each observation may still possess useful factorization [18].

$$
\begin{aligned}
\geq\ & H(Q_\theta) + \int Q(\theta) \log P(\theta) d\theta \\
& + \sum_t \left[H(Q_t) + \sum_{\mathbf{x}_h^t} \int Q_t(\mathbf{x}_h^t) Q(\theta) \log P(\mathbf{x}_v^t, \mathbf{x}_h^t | \theta) d\theta \right]
\end{aligned} \tag{83}
$$

The first lower bound comes from Equation (76) and the second as in mean field. We emphasize that by maximizing the resulting lower bound with respect to the variational distributions, we can no longer hope to recover the true marginal likelihood. This is because the true posterior over both the parameters and the hidden configurations cannot be represented within our restricted class of variational distributions.

To make use of the lower bound, we optimize it with respect to the variational distributions. This can be done by successively maximizing the bound with respect to one of the variational marginals while keeping all other marginals fixed. With only minor modifications, we can borrow the update equations from our earlier derivations (see section 4.0.1). First, we fix $Q(\theta)$ and update all $Q_t(\mathbf{x}_h^t)$ according to

$$
Q_t(\mathbf{x}_h^t) \leftarrow \frac{1}{Z_t}\, e^{E_\theta \{\log P(\mathbf{x}_v^t, \mathbf{x}_h^t | \theta)\}} \tag{84}
$$

for all $\mathbf{x}_h^t$ and $t = 1, \ldots, T$. The expectation is taken with respect to the current (fixed) estimate $Q(\theta)$. Note that the exponent in this update rule is a function of $\mathbf{x}_h^t$ only. Moreover, since we have removed the parameters as common correlates between the hidden variable configurations, the variational distributions $\{Q_t(\mathbf{x}_h^t)\}$ can be updated independent of each other.

In the second iterative step, we update the variational parameter distribution while keeping $\{Q_t(\mathbf{x}_h^t)\}$ fixed:

$$
Q(\theta) \leftarrow \frac{1}{Z}\, e^{\log P(\theta) + \sum_t E_{\mathbf{x}_h^t} \{\log P(\mathbf{x}_v^t, \mathbf{x}_h^t | \theta)\}} \tag{85}
$$

where the expectations in the exponent are taken with respect to each $Q_t(\mathbf{x}_h^t)$. Although we cannot find the true posterior distribution over the parameters (except in special cases), these updates nevertheless monotonically increase the lower bound on the marginal data likelihood.

We make here a few final observations about the accuracy of the variational Bayesian approach. First, the true posterior over the parameters in this case will almost surely contain multiple modes. These modes arise from different possible configurations of the hidden variables corresponding to each observation. The factored nature of our posterior approximation makes the previous analysis about the accuracy of variational mean field applicable. We suspect therefore that the variational posterior $Q(\theta)$ is likely to reflect only one of the posterior modes. The identity of the selected mode depends on the initialization of the variational distributions, the order in which the updates are carried out, as well as possible differences in the posterior weight of the modes.

9 Discussion

The focus of this tutorial has been on the formulation of variational methods for inference and estimation problems in graphical models along with the associated algorithms. Although the topics covered are diverse, this tutorial remains in many respects complementary to [23].

We have dispensed with discussing a number of variational approaches to inference and estimation. For example, mean field approximation and its higher order extensions can be viewed as recursive propagation algorithms[21, 19]. We may also go beyond the simple disjoint factorization assumption in the context of structured mean field approach and use, for example, directed graphical models as variational approximating distributions[2, 47] (see also [45]). Variational approximations can also be used for inference in mixed graphical models containing both continuous and discrete variables[32]. In terms of Bayesian estimation, variational methods lend themselves naturally to on-line approximation algorithms[18, 1] and remain applicable to structured Bayesian priors[9], which was briefly mentioned in the text.

Although we have treated variational methods in this tutorial as standalone approximation techniques, they can be naturally combined with other approximation techniques such as sampling methods. In [17] upper/lower bounds are used in a rejection sampling setting while [9] uses variational distributions as proposal distributions in the context of an importance sampling method. A number of other combinations and extensions are possible as well.

One of the main open problems in the use of variational approximation methods is characterizing their accuracy. We would like to obtain performance guarantees for specific classes of graphical models (upper/lower bounds that can be obtained from several variational formulations provide such guarantees only for specific instantiations of the inference problem and would not serve as *a priori* guarantees). Another open problem concerns focusing the inference calculations within the overall variational approach.

This is particularly important in the context of decision making.

Finally, we note that the graph structure of the relevant probability model is typically not fixed *a priori* in many estimation/inference problems. This leaves us the option of either using a simple graph model with exact inference algorithms or adopting more expressive models but with the cost of having to employ approximate inference methods. There has been little work in characterizing the conditions under which one approach is preferable to the other. Is the error from the simpler model class greater or less than the error resulting from approximate inference?

References

[1] Hagai Attias. Inferring parameters and structure of latent variable models by variational bayes. In *Proceedings of the Fifteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI–99)*, pages 21–30, San Francisco, CA, 1999. Morgan Kaufmann Publishers.

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

[3] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statististical Society B*, 2:192–236, 1974.

[4] C. M. Bishop, N. Lawrence, T. Jaakkola, and M. I. Jordan. Approximating posterior distributions in belief networks using mixtures. In Michael I. Jordan, Michael J. Kearns, and Sara A. Solla, editors, *Advances in Neural Information Processing Systems*, volume 10. The MIT Press, 1998.

[5] J. Bucklew. *Large deviation techniques in decision, simulation, and estimation.* John Wiley & Sons, 1990.

[6] T. Cover and J. Thomas. *Elements of information theory.* John Wiley & Sons, 1991.

[7] P. Dayan, G. E. Hinton, R. M. Neal, and R. S. Zemel. The Helmholtz machine. *Neural Computation*, 7(5):889–904, 1995.

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

[9] Z. Ghahramani and M. Beal. Variational inference for bayesian mixtures of factor analysers. In S. A. Solla, T. K. Leen, and K.-R. Mller, editors, *Advances in Neural Information Processing Systems*, volume 12. The MIT Press, 1999.

[10] Z. Ghahramani and M. I. Jordan. Supervised learning from incomplete data via an EM approach. In Jack D. Cowan, Gerald Tesauro, and Joshua Alspector, editors, *Advances in Neural Information Processing Systems*, volume 6, pages 120–127. Morgan Kaufmann Publishers, Inc., 1994.

[11] Z. Ghahramani and M. I. Jordan. Factorial hidden markov models. *Machine Learning*, 29:245, 1997.

[12] M. N. Gibbs and D. J. C. MacKay. Efficient implementation of Gaussian processes for interpolation. Unpublished manuscript, 1996.

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

[14] David Heckerman, Dan Geiger, and David M. Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20:197, 1995.

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

[16] T. Jaakkola and M. Jordan. Computing upper and lower bounds on likelihoods in intractable networks. In *Proceedings of the Twelfth Annual Conference on Uncertainty in Artificial Intelligence (UAI–96)*, pages 340–348, Portland, Oregon, 1996.

[17] T. Jaakkola and M. Jordan. Variational probabilistic inference and the qmr-dt database. *Journal of Artificial Intelligence Research*, 10:291–322, 1999.

[18] T. Jaakkola and M. Jordan. Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10:25–37, 2000.

[19] T. S. Jaakkola. *Variational methods for inference and learning in graphical models.* Ph.d. thesis, MIT, 1997.

[20] T. S. Jaakkola and M. I. Jordan. Improving the mean field approximation via the use of mixture distributions. In Michael I. Jordan, editor, *Proceedings of the NATO ASI on Learning in Graphical Models*. Kluwer, 1997.

[21] T. S. Jaakkola and M. I. Jordan. Recursive algorithms for approximating probabilities in graphical models. In Michael C. Mozer, Michael I. Jordan, and Thomas Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9, page 487. The MIT Press, 1997.

[22] F. Jensen, S. Lauritzen, and K. Olesen. Bayesian updating in causal probabilistic networks by local computations. *Computational Statistics Quarterly*, 4:269–282, 1990.

[23] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183, 1999.

[24] M. I. Jordan, Z. Ghahramani, and L. K. Saul. Hidden markov decision trees. In Michael C. Mozer, Michael I. Jordan, and Thomas Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9, page 501. The MIT Press, 1997.

[25] J. Kapur. *Maximum entropy models in science and engineering*. John Wiley & Sons, 1989.

[26] M. Kearns and L. Saul. Large deviation methods for approximate probabilistic inference. In *Proceedings of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI–98)*, pages 311–319, San Francisco, CA, 1998. Morgan Kaufmann Publishers.

[27] M. Kearns and L. Saul. Inference in multilayer networks via large deviation bounds. In M. S. Kearns, S. A. Solla, and D. A. Cohn, editors, *Advances in Neural Information Processing Systems*, volume 11. The MIT Press, 1999.

[28] S. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

[29] S. Lauritzen and D. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society B*, 50:154–227, 1988.

[30] D. J. C. MacKay. Ensemble learning for hidden Markov models. Unpublished manuscript, 1997.

[31] P. McCullagh and J. Nelder. *Generalized linear models.* Chapman and Hall, 1983.

[32] K. Murphy. A variational approximation for bayesian networks with discrete and continuous latent variables. In *Proceedings of the Fifteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI–99)*, pages 457–466, San Francisco, CA, 1999. Morgan Kaufmann Publishers.

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

[34] R. Neal and G. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In Michael I. Jordan, editor, *Proceedings of the NATO ASI on Learning in Graphical Models.* Kluwer, 1997.

[35] R. M. Neal. Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG–TR–93–1, Dept. of Computer Science, University of Toronto, 1993.

[36] R. M. Neal. *Bayesian Learning for Neural Networks.* Number 118 in Lecture Notes in Statistics. Springer, New York, 1996.

[37] G. Parisi. *Statistical field theory.* Addison-Wesley, 1988.

[38] J. Pearl. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann, 1988.

[39] R. Rockafellar. *Convex Analysis.* Princeton University Press, 1972.

[40] L. Saul and M. I. Jordan. Learning in Boltzmann trees. *Neural Computation*, 6(6):1174–1184, 1994.

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

[42] L. K. Saul and M. I. Jordan. Boltzmann chains and Hidden Markov Models. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems*, volume 7, pages 435–442. The MIT Press, 1995.

[43] L. K. Saul and M. I. Jordan. Exploiting tractable substructures in intractable networks. In David S. Touretzky, Michael C. Mozer, and

Michael E. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8, pages 486–492. The MIT Press, 1996.

[44] H. R. Schwarz. *Finite element methods.* Academic Press, 1988.

[45] A. Storkey. Dynamic trees: A structured variational method giving efficient propagation rules. In *Proceedings of the Sixteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI–00)*, San Francisco, CA, 2000. Morgan Kaufmann Publishers.

[46] J. Whittaker. *Graphical models in applied multivariate statistics.* John Wiley & Sons, 1990.

[47] W. Wiegerinck. Variational approximations between mean field theory and the junction tree algorithm. In *Proceedings of the Sixteenth Annual Conference on Uncertainty in Artificial Intelligence*, 2000.

[48] C. K. I. Williams and C. E. Rasmussen. Gaussian processes for regression. In David S. Touretzky, Michael C. Mozer, and Michael E. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8, pages 514–520. The MIT Press, 1996.