On Variational Message Passing on Factor Graphs

Justin Dauwels*
RIKEN Brain Science Institute, Saitama, Japan
email: justin@dauwels.com

Abstract

Variational methods are frequently used for performing inference in graphical models. The sum-product algorithm is often intractable for systems with continuous variables, and variational methods are then an interesting alternative; moreover, variational methods are guaranteed to converge (both on cycle-free and cyclic graphs), whereas the sum-product algorithm is in general not guaranteed to converge on cyclic graphs. In this paper, it is shown how (naive and structured) variational algorithms may be derived from a factor graph of the system at hand by mechanically applying generic message computation rules; in this way, one can bypass error-prone variational calculus. In prior work by Bishop et al., Xing et al., and Geiger, directed and undirected graphical models have been used for this purpose. The factor graph notation amounts to simpler generic variational message computation rules. By means of factor graphs, variational methods can straightforwardly be combined with various other message-passing algorithms, e.g., Kalman filters and smoothers, iterated conditional modes, expectation maximization (EM), gradient methods, and particle filters. Some of those combinations have been explored in the literature, others seem to be new. Generic message computation rules for such combinations are formulated. The connection between the variational message-passing algorithm and the message-passing formulation of EM is investigated.

1 Introduction

Variational techniques have a long history and they are currently applied in various research fields. They have been used for decades in quantum and statistical physics [1][2], where they are called "mean-field approximations". They allow physicists to compute macroscopic physical properties of many-particle systems. Variational methods have also been adopted in statistics [3] and machine learning (see [4]–[13] and references therein), in particular, for statistical inference. In this paper, we consider the following generic statistical inference problem. Assume that we are given a multivariate probabilistic model $f(x, \theta, y)$ with observed random variables Y and hidden random variables X and Θ. The latter takes values in a subset Ω of $\mathbb{R}^n$. We will assume that $f(x, \theta, y)$ is continuous (w.r.t. θ) in Ω and differentiable (w.r.t. θ) in the interior of Ω. Suppose that we are interested in X but *not* in Θ ("nuisance variable"), and that we wish to compute the marginal

$$f(x, y) \triangleq \int_{\theta} f(x, \theta, y) d\theta, \tag{1}$$

where $\int_\theta$ denotes either summation or integration over the whole range of Θ.

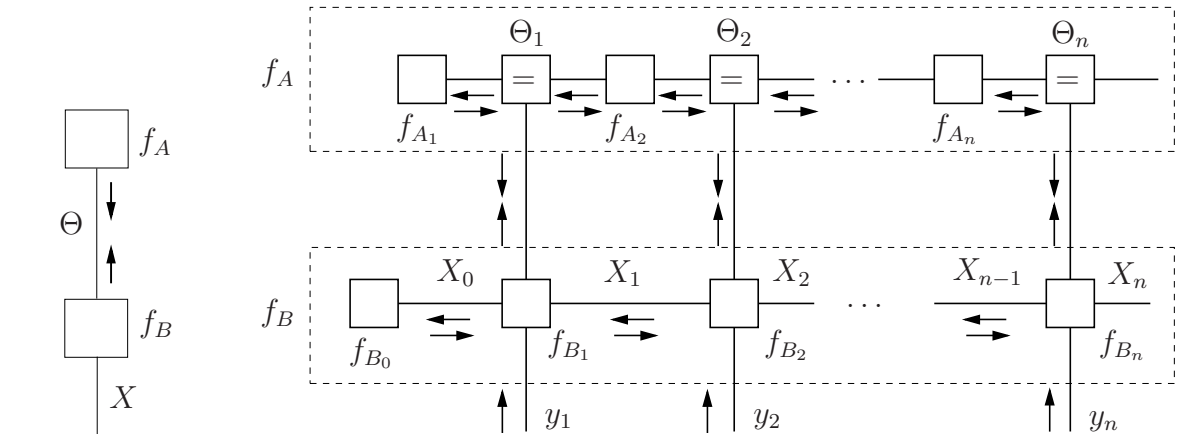
*The author is supported by a Post-Doctoral Fellowship (No. PE05060) from the Japanese Society for the Promotion of Science (JSPS).

The described problem arises, for example, in the context of estimation in state space models. In such a context, the variables X and Θ are random vectors, and the function $f(x, \theta, y)$ is given by

$$f(x, \theta, y) \triangleq f_A(\theta) f_B(x, \theta, y), \tag{2}$$

$$\triangleq f_{A_1}(\theta_1) f_{A_2}(\theta_1, \theta_2) \dots f_{A_n}(\theta_{n-1}, \theta_n) f_{B_0}(x_0) f_{B_1}(x_0, x_1, y_1, \theta_1) \cdot f_{B_2}(x_1, x_2, y_2, \theta_2) \dots f_{B_n}(x_{n-1}, x_n, y_n, \theta_n), \tag{3}$$

where X_k denotes the (unknown) state at time k, Y are the observed random variables, Θ are the (unknown) parameters of the state space model, $f_A(\theta)$ is the prior on Θ, and $f_{B_0}(x_0)$ is the prior on the initial state X_0. A factor graph of (2) and (3) is shown in Fig. 1(a) and Fig. 1(b) respectively (see [14] for a tutorial on factor graphs); the boxes f_A and f_B in Fig. 1(a) are detailed in Fig. 1(b) (dashed boxes). We consider the situation where we wish to estimate the state X and we are not interested in the parameters Θ. In model (3), the integration over Θ (1) is often infeasible.



(a) Graph of (2). (b) Graph of (3).

Figure 1: Factor graphs.

We will now assume that a factor graph for $f(x, \theta, y)$ is available. It may be possible to compute $f(x, y)$ (1) by sum-product message passing [14]. Unfortunately, this naive approach is often impractical: the variable Θ is supposed to be continuous, and the sum-product rule may lead to intractable integrals. In such situations, variational methods become an attractive alternative, since they often lead to simple message computation rules (especially if the model $f(x, \theta, y)$ belongs to the conjugate-exponential family) [33] [8] [13]. The naive and structured variational method have been formulated as message-passing algorithms by Bishop et al. [13], Xing et al. [42] and Geiger [41] in the notation of directed and undirected graphical models; variational message-passing algorithms have also been derived by means of factor graphs for certain specific cases [12, pp. 256–258] [15]. In this paper, we describe the generic (naive and structured) variational method as message-passing algorithms on factor graphs; the factor graph notation allows a simpler formulation of variational message passing. Moreover, once the variational method is cast as message passing on factor graphs, we can mix the variational method with other message-passing algorithms. For instance, structured variational algorithms compute besides variational messages also sum-product messages. The latter may for example be represented as Gaussian distributions or particle lists. This amounts to algorithms such as variational Kalman filters and smoothers [45] [8] and variational particle filters and smoothers. If the variational messages are intractable, they may be represented as particle lists, amounting to particle-based algorithms such as variational Markov Chain Monte Carlo methods [35][36].

Alternatively, if the integral in (1) is intractable, one often makes the (sometimes unsatisfactory) approximation

$$f(x,y) \approx \hat{f}(x,y) \triangleq f(x,\hat{\theta},y), \tag{4}$$

where $\hat{\theta}$ is a point estimate of Θ, typically the mode

$$\hat{\theta}^{\max} \triangleq \operatorname*{argmax}_{\theta} f(\theta,y), \tag{5}$$

where

$$f(\theta,y) \triangleq \int_x f(x,\theta,y)dx. \tag{6}$$

It may be possible to compute $f(\theta,y)$ (6) by sum-product message passing and $\hat{\theta}^{\max}$ (5) by max-product message passing [14]. Also this approach, however, is often impractical:

1. If the variable X is continuous, the sum-product rule may lead to intractable integrals, whereas if X is discrete, the sum-product rule may lead to an unwieldy sum; in both cases, we cannot compute (6).
2. The max-product rule may lead to an intractable expression; in this case, we cannot compute (5).

If X is discrete, the first problem may be solved by iterative sum-product message passing on a cyclic graph of $f(x,\theta,y)$, resulting in an approximate marginal ("belief") $b(\theta,y)$. However, this approach usually does not solve the problem of intractable integrals. Also here, variational techniques may be helpful. The mode (5) is then approximated by

$$\hat{\theta}^{\mathrm{var}} \triangleq \operatorname*{argmax}_{\theta} q(\theta,y), \tag{7}$$

where $q(\theta,y)$ is an approximate marginal obtained by a variational method.

We now address the second problem in the above list. If the maximization (5) is intractable, one may resort to standard optimization techniques such as iterative conditional modes (ICM, a.k.a. "cyclic maximization" or "coordinate ascent/descent") [16] or gradient methods [17]. In earlier work, we have described those two methods for solving (5) (6) as message-passing algorithms operating on a factor graph of $f(x,\theta)$ [18] [19]. Similarly, if the maximization (7) is intractable, one needs to resort to numerical optimization techniques such as ICM or gradient methods; this leads to message-passing algorithms that mix variational methods with ICM and/or gradient methods.

An alternative procedure to compute $\hat{\theta}^{\max}$ (5) (exactly or approximately) is expectation maximization (EM) [20]. A message-passing view of EM is developed in [21][22][23]; the message-passing EM algorithm computes sum-product messages (besides other messages). If those sum-product messages are intractable, they may be replaced by variational messages, amounting to a message-passing formulation of "variational EM" [5].

The maximization step of EM is sometimes intractable, and again, one may then apply ICM or some gradient method; such modifications are referred to as "generalized EM algorithms" [24]. In [18, Section 4.9.5] [19][23] we described various generalized EM algorithms as message-passing algorithms operating on factor graphs. Sum-product messages may also in this context be replaced by variational messages.

This paper is structured as follows. In the following section, we review the naive variational method (closely following [4]–[13]). In Section 3, we describe the naive variational method as a message-passing algorithm and formulate the generic naive variational message computation rule. In Section 4, we investigate the combination of naive variational methods with (generalized) EM, gradient methods, and ICM; in Section 5, we consider structured variational message passing. Some concluding remarks are offered in Section 6.

2 Review of the Naive Variational Method

Assume that we are given a generic multivariate function $f(z_1, \ldots, z_m)$ (not necessarily normalized), and suppose that we wish to compute its marginals

$$f(z_k) \triangleq \int f(z_1, \ldots, z_m) dz_1 dz_2 \ldots dz_{k-1} dz_{k+1} \ldots dz_m \quad (k = 1, \ldots, m), \tag{8}$$

where $\int_z$ denotes either summation or integration over the whole range of z. The idea behind variational methods is to find a sufficiently "simple" function $q(z_1, \ldots, z_m)$ (belonging to a family $\mathcal{Q}$ of trial functions) that is as "close" as possible to $f(z_1, \ldots, z_m)$, i.e.,

$$q^* \triangleq \underset{q \in \mathcal{Q}}{\operatorname{argmin}} D\left(f, q\right), \tag{9}$$

where $D(f, q)$ is a measure for the distance between f and q. The marginals $f(z_k)$ (8) are then approximated by the marginals of q^*. The family $\mathcal{Q}$ can be chosen in many ways, the only constraint is that the marginals of the functions $q \in \mathcal{Q}$ should be tractable.

Several distance measures $D\left(f, q\right)$ may be used. If f is normalized, a popular measure is the Kullback-Leibler divergence $D\left(q \| f\right)$ defined as

$$D\left(q \| f\right) \triangleq \int_x q(x) \log \frac{q(x)}{f(x)} dx. \tag{10}$$

Obviously, there are many alternatives to the Kullback-Leibler divergence, such as Amari's α-divergence [25] or Csiszár's f-divergence [26]; such extensions have been explored in [27] [28] [29] [30]. In this paper, however, we only consider the Kullback-Leibler divergence $D\left(q \| f\right)$.

A widely used family $\mathcal{Q}$ is the set of fully factorized functions

$$q(z_1, \ldots, z_m) \triangleq \prod_{k=1}^{m} q(z_k), \tag{11}$$

which amounts to the so-called "naive mean-field" approximations in statistical and quantum physics. Note that the marginals of $q(z_1, \ldots, z_m)$ are simply the factors $q(z_k)$. With this choice of D and $\mathcal{Q}$, the variational method tries to find

$$q^* \triangleq \underset{q \in \mathcal{Q}}{\operatorname{argmin}} D\left(q \| f\right), \tag{12}$$

and the marginals $f(z_k)$ (8) are approximated by $q^*(z_k)$. Note that the objective function $D\left(q \| f\right)$ (12) is in general non-convex in the factors $q(z_k)$. By variational calculus, one can easily verify that $q^*(z_k)$ fulfills the equality

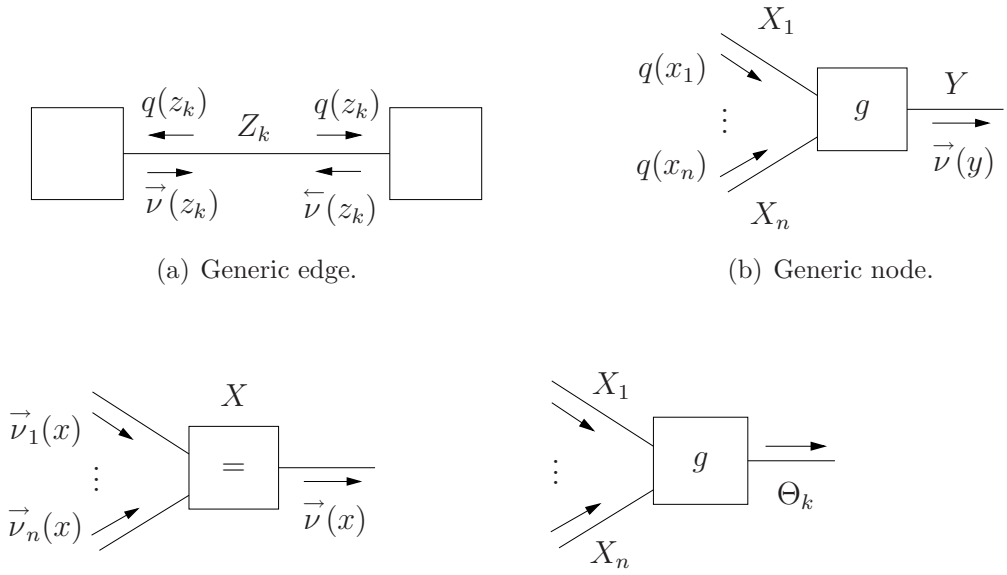
$$q^*(z_k) \triangleq \frac{1}{\gamma_k} \exp\Big(\int q^*(z_1) \ldots q^*(z_{k-1}) q^*(z_{k+1}) \ldots q^*(z_m) \log f(z_1, \ldots, z_m) \\ dz_1 dz_2 \ldots dz_{k-1} dz_{k+1} \ldots dz_m \Big), \tag{13}$$

where the constant γ_k ensures that $\int q^*(z_k) dz_k = 1$. The equality (13) suggests to determine (12) by iterating the update rule

$$q^{(\ell+1)}(z_k) \triangleq \frac{1}{\gamma_k^{(\ell)}} \exp\Big(\int q^{(\ell)}(z_1) \ldots q^{(\ell)}(z_{k-1}) q^{(\ell)}(z_{k+1}) \ldots q^{(\ell)}(z_m) \log f(z_1, \ldots, z_m) \\ dz_1 dz_2 \ldots dz_{k-1} dz_{k+1} \ldots dz_m \Big), \tag{14}$$

where $q^{(\ell)}(z_k)$ $(k = 1, \ldots, m)$ are the trial marginals at the ℓ-th iteration. This is precisely what is done by the naive variational method. It can be shown that at each iteration, the Kullback-Leibler divergence $D(q\|f)$ decreases, unless the algorithm has reached a fixed point; the method is guaranteed to convergence to a local minimum of $D(q\|f)$ (see [4]–[13]).

As was shown by Yedidia et al. [31] (see also [32]), also the sum-product algorithm can be interpreted from the viewpoint of Kullback-Leibler divergence minimization (cf. (12)). However, there is an important difference: in the sum-product algorithm, the factorization of the trial function q is more complex, and the Kullback-Leibler divergence $D(q\|f)$ is intractable. The sum-product algorithm tries to minimize an approximation of $D(q\|f)$; it it does not decrease that approximation at each iteration, however, and consequently, the sum-product algorithm is in general not guaranteed to converge.



(a) Generic edge.

(b) Generic node.

(c) Equality constraint node.

(d) Generic node.

Figure 2: Variational message passing.

3 Naive Variational Message Passing

Now we turn our attention back to the naive variational method. If the function f factorizes, the update (14) can be carried out by local computations. In particular, those computations can be cast as message passing on a factor graph that represents the factorization of f. A message-passing formulation of the naive variational method was proposed by Bishop and Winn [12] [13] in the setting of directed and undirected graphical models. Winn also formulated the naive variational message computation rule in the notation of factor graphs for the particular case of conjugate-exponential models [12, pp. 256–258]; Nissilä et al. considered the particular case of factorial hidden Markov models with conditionally Gaussian distributed observations [15]. We will now formulate the generic variational message computation rule in the notation of factor graphs. As we will outline in Section 4, this will a.o. allow us to combine and mix that rule with with other message computation rules on one and the same factor graph, e.g., the sum-product rule [14], gradient sum-product rule [19], E-log rule [22] [23], and gradient E-log rule [19]. In this fashion, one can straightforwardly combine and mix variational methods with other inference methods.

As is easily verified from (14), the variational method may be formulated as the following message-passing algorithm:

1. Initialize all messages q and ν, e.g., $q(\cdot) \propto 1$ and $\nu(\cdot) \propto 1$.
2. Select an edge z_k in the factor graph of $f(z_1, \ldots, z_m)$ (see Fig. 2(a)).
3. Compute the two messages $\vec{\nu}(z_k)$ and $\overleftarrow{\nu}(z_k)$ by applying the generic rule (see Fig. 2(b))

$$\vec{\nu}(y) \propto \exp \int q(x_1)q(x_2)\ldots q(x_n) \log g(x_1, \ldots, x_n, y) dx_1 \ldots dx_n \tag{15}$$

$$\propto \exp \mathrm{E}_q\big[\log g(X_1, \ldots, X_n, y)\big]. \tag{16}$$

4. Compute the marginal $q(z_k)$ (see Fig. 2(a))

$$q(z_k) \propto \vec{\nu}(z_k)\overleftarrow{\nu}(z_k), \tag{17}$$

   and send it to the two nodes connected to the edge X_k.
5. Iterate 2–4 until convergence.

Some remarks:

- Interestingly, the rule (15) is often simpler than the sum-product rule [14], especially if the model $f(x, \theta, y)$ belongs to the conjugate-exponential family [33] [8] [13].
- Along each edge z_k, four messages are propagated (cf. Fig. 2(a)): two messages arriving *from* the two incident nodes, i.e., $\vec{\nu}(z_k)$ and $\overleftarrow{\nu}(z_k)$, and two messages propagating *towards* those nodes, both equal to $q(z_k)$. The message arriving at the edge z_k from the right (i.e., $\overleftarrow{\nu}(z_k)$) is not directly sent to the left neighboring node (and vice versa). Instead the two incoming messages $\vec{\nu}(z_k)$ and $\overleftarrow{\nu}(z_k)$ are further processed, resulting in $q(z_k)$; the latter is then sent to both incident nodes. In the sum-product algorithm, only two messages propagate along each edge; the message arriving at the edge z_k from the right is directly sent to the left incident node (and vice versa), i.e., without further processing.
- The approximate marginals $q(z_k)$ propagate in the graph as messages (cf. Fig. 2(a) and 2(b)). In the sum-product algorithm, the approximate marginals are computed from sum-product messages; they are not propagated as messages in the graph.
- The rule (15) can not be applied to deterministic node functions g, i.e., node functions g that are Dirac or Kronecker deltas. At an equality constraint node (see Fig. 2(c)), the following rule applies:

$$\vec{\nu}(x) \propto \vec{\nu}_1(x)\vec{\nu}_2(x)\ldots\ \vec{\nu}_n(x). \tag{18}$$

  Other deterministic nodes can often (but not always!) be handled by combining them with neighboring non-deterministic nodes; the same procedure is applied in message-passing EM [18, p. 145] [23].
- One may evaluate the generic update rule (15) for often occurring mode functions g; this has been done by Beal [33] and by Bishop and Winn [12] [13] in the setting of directed and undirected graphical models and conjugate-exponential families.
- On the other hand, if the messages $\vec{\nu}(z_k)$ and $\overleftarrow{\nu}(z_k)$ are intractable, the marginal $q(z_k)$ may be represented as a particle list. The latter may be iteratively updated by Markov Chain Monte Carlo methods (MCMC) [34] with target function (17), leading to variational MCMC [35][36].

Let us now look back at the model $f(x, \theta, y)$ of Section 1. If (1) can not be computed by applying the sum-product algorithm on a factor graph of $f(x, \theta, y)$ (cf., e.g., Fig. 1(b)), we may apply variational message passing on the graph of $f(x, \theta, y)$ with trial function (cf. (11))

$$q(x, \theta) \triangleq \prod_k q(x_k) \prod_\ell q(\theta_\ell). \tag{19}$$

In the case of model (3), the naive variational method amounts to computing variational messages $\nu(\theta_k)$ and $\nu(x_k)$, and marginals $q(\theta_k)$ and $q_k(x_k)$ in the subgraphs $f_A(\theta)$ and $f_B(x,\theta)$ respectively. The marginal (1) is then approximated by $q(x) \triangleq q(x_1)\dots q(x_n)$.

4 Naive Variational Message Passing for Estimation

The naive variational method is also relevant for computing the mode (5). If the marginal $f(\theta, y)$ (6) cannot be computed (exactly or approximately) by the sum-product algorithm, one may compute approximative marginals $q(\theta_k)$ by naive variational message-passing. If the mode of $q(\theta_k)$ is not available in closed form, one may resort to standard optimization techniques such as ICM [16] ("variational ICM") and gradient methods [17] ("variational gradient algorithms").

Alternatively, one may determine the mode (5) by EM [20]. If the E-step is intractable, one may approximate the E-step by naive variational methods ("variational EM") [5]; the intractable marginals required in the E-step are then replaced by approximate marginals q. If in addition the M-step is intractable, one may apply ICM or a gradient method. This also leads to variational ICM and variational gradient methods; the latter can thus be derived in two different ways: (i) by applying ICM or a gradient method to determine the mode of the variational marginals $q(\theta_k)$; (ii) by approximating the M-step in variational EM by ICM or a gradient method. Recently, those three procedures for determining (5), i.e., ICM [18], gradient methods [19], and (generalized) EM [21][22][23], were described as message passing algorithms operating on a factor graph of $f(x,\theta,y)$. By slightly modifying those message-passing algorithms, one obtains a message-passing formulation of variational ICM, variational gradient methods, and variational EM: one simply needs to adapt certain messages, as we briefly outline in the following.

Solving (5) by ICM involves sum-product messages; in variational ICM, those messages are replaced by variational messages (15). The E-step in (standard) EM involves the computation of E-log messages [21][22][23] (see Fig. 2(d))

$$h(\theta_k) = \int p\big(x_1, x_2, \dots, x_n|\hat{\theta}\big) \log g(x_1, \dots, x_n, \theta_k) dx_1 \dots dx_n \tag{20}$$

$$= \mathrm{E}_p\big[\log g(X_1, \dots, X_n, \theta_k)|\hat{\theta}\big]. \tag{21}$$

In the E-step of naive variational EM, those messages are replaced by log-variational messages (cf. (15))

$$\log \nu(\theta_k) = \int q\big(x_1|\hat{\theta}\big) q\big(x_2|\hat{\theta}\big) \dots q\big(x_n|\hat{\theta}\big) \log g(x_1, \dots, x_n, \theta_k) dx_1 \dots dx_n \tag{22}$$

$$= \mathrm{E}_q\big[\log g(X_1, \dots, X_n, \theta_k)|\hat{\theta}\big]. \tag{23}$$

Gradient EM involves gradients of E-log messages [19] (see Fig. 2(d))

$$\nabla_{\theta_k} h(\theta_k) = \int p\big(x_1, x_2, \dots, x_n|\hat{\theta}\big) \log \nabla_{\theta_k} g(x_1, \dots, x_n, \theta_k) dx_1 \dots dx_n \tag{24}$$

$$= \mathrm{E}_p\big[\nabla_{\theta_k} \log g(X_1, \dots, X_n, \theta_k)|\hat{\theta}\big]. \tag{25}$$

In naive variational gradient methods, those messages are replaced by gradients of log-variational messages

$$\nabla_{\theta_k} \log \nu(\theta_k) = \int q\big(x_1|\hat{\theta}\big) q\big(x_2|\hat{\theta}\big) \dots q\big(x_n|\hat{\theta}\big) \nabla_{\theta_k} \log g(x_1, \dots, x_n, \theta_k) dx_1 \dots dx_n \tag{26}$$

$$= \mathrm{E}_q\big[\nabla_{\theta_k} \log g(X_1, \dots, X_n, \theta_k)\big]. \tag{27}$$

Gradient ascent methods for solving (5) involve gradients of logarithmic sum-product messages [19] (see Fig. 2(d))

$$\nabla_{\theta_k} \log \mu(\theta_k) = \frac{\int \mu\big(x_1|\hat{\theta}\big)\mu\big(x_2|\hat{\theta}\big)\ldots\mu\big(x_n|\hat{\theta}\big)\nabla_{\theta_k} g(x_1,\ldots,x_n,\theta_k)dx_1\ldots dx_n}{\int \mu\big(x_1|\hat{\theta}\big)\mu\big(x_2|\hat{\theta}\big)\ldots\mu\big(x_n|\hat{\theta}\big) g(x_1,\ldots,x_n,\theta_k)dx_1\ldots dx_n}, \tag{28}$$

where $\mu\big(x_k|\hat{\theta}\big)$ $(k = 1,\ldots,n)$ are sum-product messages. Note that the message (28) is identical to the message (25). In the context of problem (5), gradient EM is almost identical to gradient ascent: the only difference lies in the update schedule. The same obviously holds for variational gradient ascent and variational gradient EM.
Some remarks:

- The fixed points of variational ICM/gradient methods/EM are stationary points of a solution $q^*(\theta) \triangleq \prod_{k=1}^m q^*(\theta_k)$ of (12), i.e., they are not the stationary points of the true marginal $f(\theta, y)$.
- Variational EM is guaranteed to convergence under some weak regularity conditions. Variational gradient methods are guaranteed to convergence if in addition the step size is chosen appropriately (e.g., Armijo rule [17]). Variational ICM converges globally if the messages $\nu(\theta_i)$ are unimodal; this follows from the theory of [37].
- It is well known that EM, variational EM and the variational method in general often suffer from slow convergence [38]. Gradient methods as for instance variational gradient methods or (conjugate-)gradient EM [39] usually converge faster.
- The generic message computation rules (26) and (22) for variational gradient methods and variational EM respectively could be computed and tabulated for "standard" node functions g.

5 Structured Variational Message Passing

So far, we have considered fully factorized trial functions (cf. (11)). In this section, we consider more structured factorizations, leading to "structured" variational algorithms [4][40]–[42]. Structured variational methods have been formulated as message-passing algorithms by Bishop et al. [13], Xing et al. [42] and Geiger [41] in the notation of directed and undirected graphical models. Here we use the notation of factor graphs, which will lead to simpler generic message computation rules; it will also allow us to compare structured variational message passing to the message-passing formulation of EM [21] [22] [23].

5.1 A First Example

Suppose that we wish to improve the naive variational method for computing (1) for the system depicted in Fig. 1(b). To this end, let us now use the trial function

$$q(x,\theta) \triangleq q(x)q(\theta), \tag{29}$$

where $q(x)$ and $q(\theta)$ are *not* further factorized, in contrast to (19). The upper and lower dotted box in Fig. 3 correspond to the factors $q(\theta)$ and $q(x)$ respectively of the trial function (29); the upper and lower box contains the edges Θ and X respectively. Based on the trial (29), one may derive a "structured" variational method [4][40]–[42] following the procedure of Section 2: through variational calculus one obtains an equality similar to (13) and an update rule similar to (14). Iterating that update rule amounts to a "structured" variational method [4][40]–[42]; it can be formulated as a message-passing algorithm that iterates the following two steps:

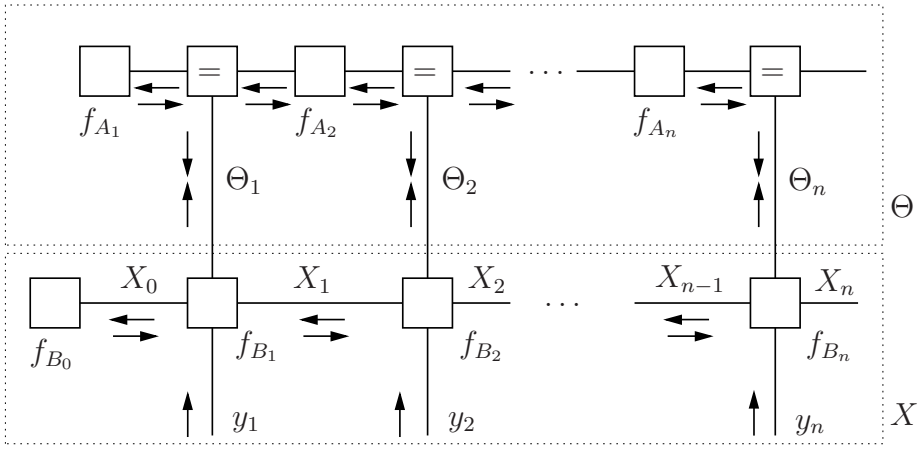


Figure 3: Partitioning corresponding to the trial (29).

Update $q(x)$
Perform the forward recursion

$$\vec{\mu}(x_k) \propto \int \vec{\mu}(x_{k-1}) \exp\Big[\int q(\theta_k) \log f_{B_k}(x_{k-1}, x_k, y_k, \theta_k) d\theta_k\Big] dx_{k-1}, \tag{30}$$

and the corresponding backward recursion with messages $\overleftarrow{\mu}(x_k)$.
Update:

$$q(x_{k-1}, x_k) \propto \vec{\mu}(x_{k-1}) \exp\Big[\int q(\theta_k) \log f_{B_k}(x_{k-1}, x_k, y_k, \theta_k) d\theta_k\Big] \overleftarrow{\mu}(x_k). \tag{31}$$

Update $q(\theta)$
Compute the upward messages

$$\nu\uparrow(\theta_k) \propto \exp \int q(x_{k-1}, x_k) \log f_{B_k}(x_{k-1}, x_k, y_k, \theta_k) dx_{k-1} dx_k. \tag{32}$$

Perform the forward recursion

$$\vec{\mu}'(\theta_k) \propto \int \vec{\mu}(\theta_{k-1}) f_{A_k}(\theta_{k-1}, \theta_k) d\theta_{k-1} \qquad \text{and} \qquad \vec{\mu}(\theta_k) \propto \vec{\mu}'(\theta_k) \nu\uparrow(\theta_k), \tag{33}$$

and the corresponding backward recursion with messages $\overleftarrow{\mu}(\theta_k)$ and $\overleftarrow{\mu}'(\theta_k)$.
Compute the downward messages

$$\mu\downarrow(\theta_k) \stackrel{\triangle}{\propto} \vec{\mu}'(\theta_k) \overleftarrow{\mu}(\theta_k) = \vec{\mu}(\theta_k) \overleftarrow{\mu}'(\theta_k). \tag{34}$$

Update:

$$q(\theta_k) \propto \nu\uparrow(\theta_k) \mu\downarrow(\theta_k). \tag{35}$$

Some remarks:

- Since the above message-passing scheme is a (structured) variational algorithm, it is guaranteed to converge to a local minimum of the divergence (12) with trial function (29) [4][40]–[42].

- The marginals $q(x_k)$ are computed as $q(x_k) = \overrightarrow{\mu}(x_k)\overleftarrow{\mu}(x_k)$.
- The updates (33) and (34) are instances of the sum-product rule [14]. This is also the case for the updates (30) and (31) if one considers the exponential factors in (30) and (31) as new node functions. Those observations were first made by MacKay [44] in the context of Hidden Markov Models, and later by Ghahramani and Beal [10] and Xin et al. [42] in the setting of directed and undirected graphical models.
- The messages $\overrightarrow{\mu}(x_k)$, $\overleftarrow{\mu}(x_k)$ and/or $\overrightarrow{\mu}(\theta_k)$, $\overleftarrow{\mu}(\theta_k)$ may be represented (exactly or approximately) as Gaussian distributions; Step 1 and/or Step 4 then involves Kalman smoothing, resulting in "variational Kalman smoothing" [45] [8]. Alternatively, those messages may be represented as particle lists; Step 1 and/or Step 4 then involves particle smoothing [46]–[50] ("variational particle smoother").
- Readers familiar with the problem of parameter estimation in state space models probably have noticed that the above structured variational message-passing algorithm resembles an EM algorithm. Indeed, approximating $q(\theta)$ in (30)–(35) by a Dirac delta results in an EM algorithm for estimating Θ:

  E-step
  In the subgraph $f_B(x, \theta)$, perform the sum-product forward sweep (cf. (30))

$$\overrightarrow{\mu}(x_k) \propto \int \overrightarrow{\mu}(x_{k-1}) f_{B_k}(x_{k-1}, x_k, \hat{\theta}_k^{(\ell)}) dx_{k-1}, \tag{36}$$

  and the corresponding backward sweep with messages $\overleftarrow{\mu}(x_k)$. Compute the upward messages (cf. (32))

$$\begin{aligned} \exp(h(\theta_k)) \propto \exp \int & p(x_{k-1}, x_k; \hat{\theta}^{(\ell)}) \\ & \cdot \log f_{B_k}(x_{k-1}, x_k, \theta_k) dx_{k-1} dx_k, \end{aligned} \tag{37}$$

  where (cf. (31))

$$p(x_{k-1}, x_k; \hat{\theta}^{(\ell)}) \propto \overrightarrow{\mu}(x_{k-1}) f_{B_k}(x_{k-1}, x_k, \hat{\theta}_k^{(\ell)}) \overleftarrow{\mu}(x_k). \tag{38}$$

  M-step (cf. (33)–(35))

$$\hat{\theta}^{(\ell+1)} = \underset{\theta}{\operatorname{argmax}} \Big[f_A(\theta) \exp(h(\theta_1)) \dots \exp(h(\theta_n)) \Big]. \tag{39}$$

  We formulated this EM algorithm as a message-passing algorithm operating on the factor graph of Fig. 1(b). Note that the message $h(\theta_k)$ (37) is a particular instance of the generic E-log message (20) [21] [22] [23]. The message $\exp(h(\theta_k))$ (37) is closely related to $\nu\uparrow(\theta_k)$ (32): the marginal $p(x_{k-1}, x_k; \hat{\theta}^{(\ell)})$ in (37) is replaced by a variational marginal $q(x_{k-1}, x_k)$ in (32). Since we started from a non-factorized trial function $q(x)$ (cf. (29)), we obtained the standard EM algorithm; a factorized trial $q(x)$ leads to a structured variational EM algorithm (see Section 5.3). Note also that the EM algorithm yields a point estimate $\hat{\theta}$ of Θ, whereas the structured variational algorithm computes an approximate posterior density in Θ.

5.2 A Second Example

In order to gain more insight in structured variational message-passing algorithms, we now consider the more general system

$$f(x,\theta,z,y) \stackrel{\triangle}{=} f_A(\theta,z) f_B(x,\theta,y), \tag{40}$$

$$\stackrel{\triangle}{=} f_{A_0}(z_0) f_{A_1}(z_0,z_1,\theta_1) \ldots f_{A_n}(z_{n-1},z_n,\theta_n) f_{B_0}(x_0) f_{B_1}(x_0,x_1,y_1,\theta_1) \cdot f_{B_2}(x_1,x_2,y_2,\theta_2) \ldots f_{B_n}(x_{n-1},x_n,y_n,\theta_n). \tag{41}$$

The model $f(x,\theta,y)$ (3) can be considered as a specific instance of (41). We wish to derive structured variational message-passing algorithms for statistical inference in model (41). There are several natural candidates for structured trial functions. Let us first investigate the trial function (see Fig. 4)

$$q(x,z,\theta) \stackrel{\triangle}{=} q(x) q(z,\theta). \tag{42}$$

The latter amounts to a structured variational message-passing algorithm that is similar to the algorithm (30)–(34). The update of $q(x)$ is identical (cf. (30)(31)), since the factor $f_B(x,\theta)$ remained unchanged; the update of $q(z,\theta)$ is similar to the update of $q(\theta)$ (cf. (32)–(35)). The forward recursion (33) is replaced by

$$\vec{\mu}(z_k) \propto \int \vec{\mu}(z_{k-1}) f_{A_k}(z_{k-1},z_k,\theta_k)\, \nu\uparrow(\theta_k) d\theta_k dz_{k-1}, \tag{43}$$

and similarly the backward recursion with messages $\overleftarrow{\mu}(z_k)$. The downward message $\mu\downarrow(\theta_k)$ (cf. (34)) is now computed as

$$\mu\downarrow(\theta_k) \propto \int f_{A_k}(z_{k-1},z_k,\theta_k) \vec{\mu}(z_{k-1}) \overleftarrow{\mu}(z_k) dz_{k-1} dz_k. \tag{44}$$

An alternative to (42) is the trial function

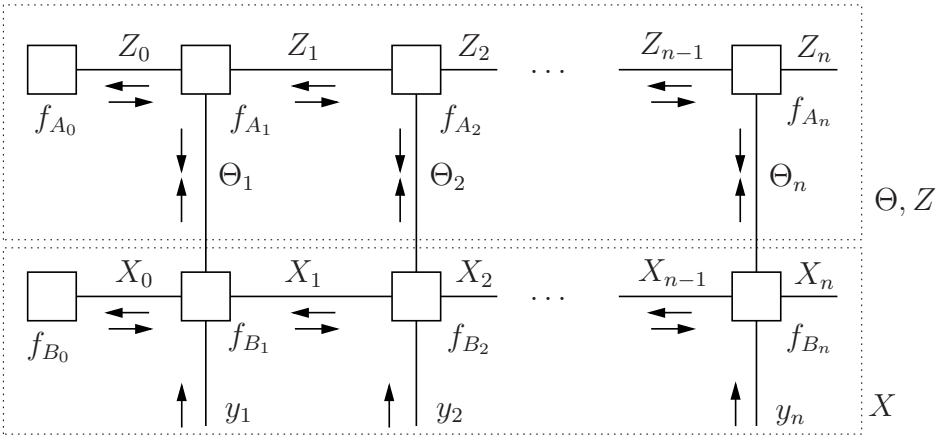


Figure 4: Partitioning corresponding to the trial (42).

$$q(x,z,\theta) \stackrel{\triangle}{=} q(x,\theta) q(z). \tag{45}$$

The corresponding structured variational message-passing algorithm (see Fig. 5) is obtained from the previous one by swapping the functions f_A and f_B and the variables X_k and Z_k (cf. Fig. 4 and 5). For instance, the (approximate) marginal $q(\theta_k)$ (cf. (35)) is now computed as

$$q(\theta_k) \propto \mu\uparrow(\theta_k) \nu\downarrow(\theta_k), \tag{46}$$

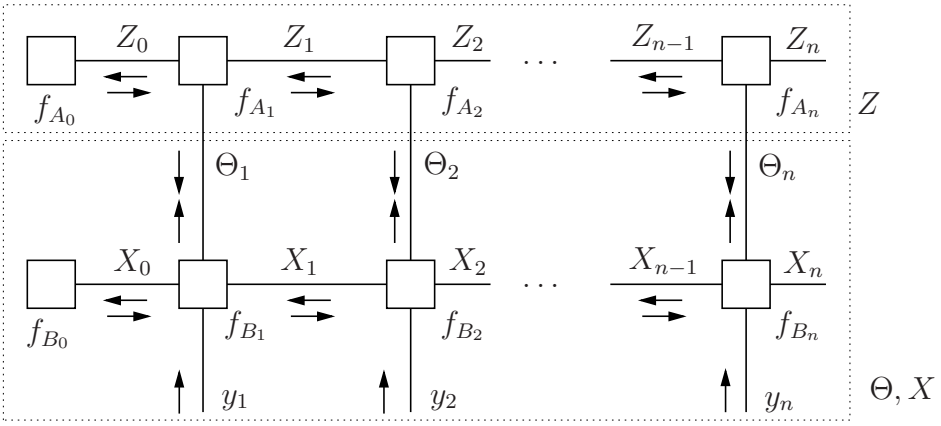


Figure 5: Partitioning corresponding to the trial (45).

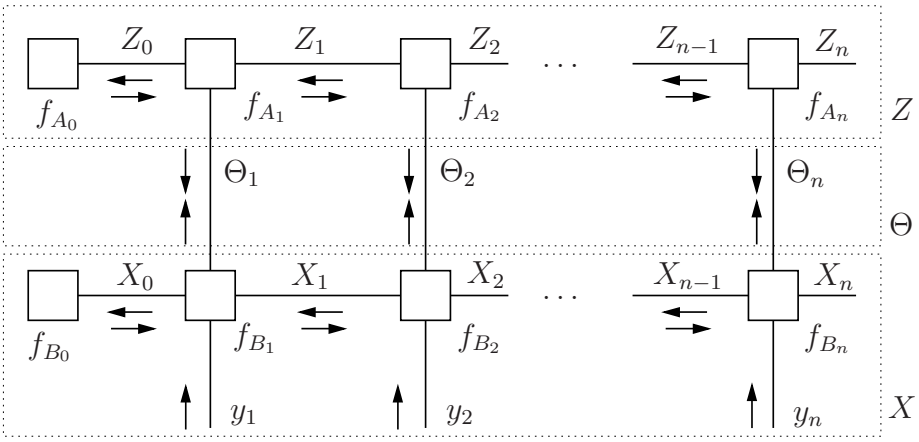


Figure 6: Partitioning corresponding to the trial (49).

where

$$\nu\downarrow(\theta_k) \propto \exp \int q(z_{k-1}, z_k) \log f_{A_k}(z_{k-1}, z_k, y_k, \theta_k) dz_{k-1} dz_k, \tag{47}$$

and

$$\mu\uparrow(\theta_k) \propto \int f_{B_k}(x_{k-1}, x_k, \theta_k) \vec{\mu}(x_{k-1}) \overleftarrow{\mu}(x_k) dx_{k-1} dx_k. \tag{48}$$

The trial functions (42) and (45) are asymmetric, and the same holds for the resulting structured variational message-passing algorithms: they treat the variables X and Z in a different manner. The trial function (see Fig. 6)

$$q(x, z, \theta) \stackrel{\triangle}{=} q(x) q(\theta) q(z) \tag{49}$$

leads to a structured variational message-passing algorithm that is symmetric in the variables X and Z; it iterates the following three steps:

Update $q(x)$

Perform the forward recursion

$$\vec{\mu}(x_k) \propto \int \vec{\mu}(x_{k-1}) \exp \Big[\int q(\theta_k) \log f_{B_k}(x_{k-1}, x_k, y_k, \theta_k) d\theta_k \Big] dx_{k-1}, \tag{50}$$

and the corresponding backward recursion with messages $\overleftarrow{\mu}(x_k)$.

Update:

$$q(x_{k-1}, x_k) \propto \overrightarrow{\mu}(x_{k-1}) \exp\Big[\int q(\theta_k) \log f_{B_k}(x_{k-1}, x_k, y_k, \theta_k) d\theta_k\Big] \overleftarrow{\mu}(x_k). \tag{51}$$

Update $q(z)$
Perform the forward recursion

$$\overrightarrow{\mu}(z_k) \propto \int \overrightarrow{\mu}(z_{k-1}) \exp\Big[\int q(\theta_k) \log f_A(z_{k-1}, z_k, \theta_k) d\theta_k\Big] dz_{k-1}, \tag{52}$$

and the corresponding backward recursion with messages $\overleftarrow{\mu}(\theta_k)$.
Update:

$$q(z_{k-1}, z_k) \propto \overrightarrow{\mu}(z_{k-1}) \exp\Big[\int q(\theta_k) \log f_{A_k}(z_{k-1}, z_k, \theta_k) d\theta_k\Big] \overleftarrow{\mu}(z_k). \tag{53}$$

Update $q(\theta)$
Compute the upward messages

$$\nu\uparrow(\theta_k) \propto \exp \int q(x_{k-1}, x_k) \log f_{B_k}(x_{k-1}, x_k, y_k, \theta_k) dx_{k-1} dx_k. \tag{54}$$

Compute the downward messages

$$\nu\downarrow(\theta_k) \propto \exp \int q(z_{k-1}, z_k) \log f_A(z_{k-1}, z_k, \theta_k) dz_{k-1} dz_k. \tag{55}$$

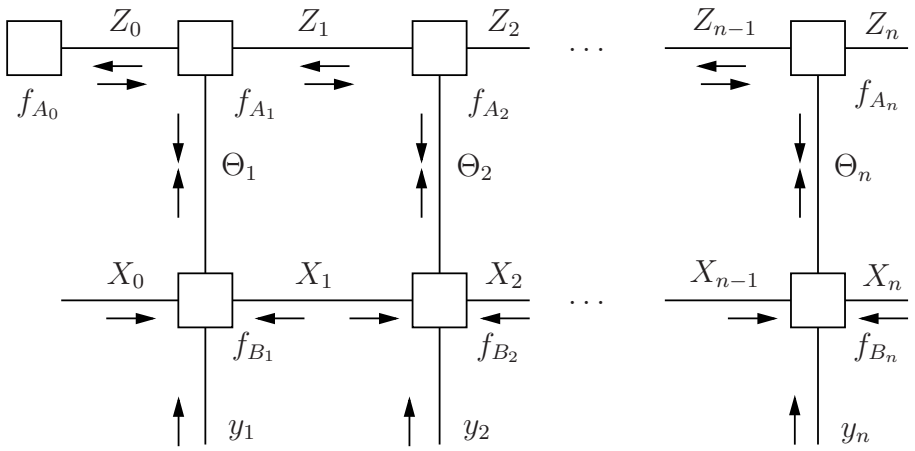
Update:

$$q(\theta_k) \propto \nu\uparrow(\theta_k) \nu\downarrow(\theta_k). \tag{56}$$

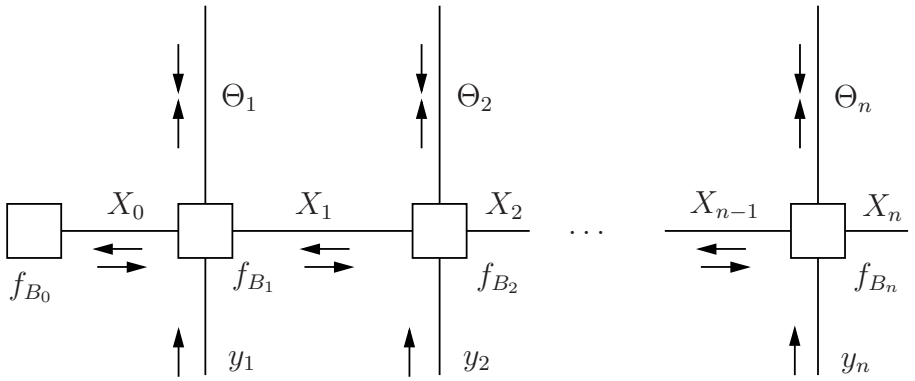
5.3 Generic formulation

From the previous examples, it is straightforward to formulate a general recipe to derive structured variational algorithms from factor graphs. Let f be a multivariate function, and assume that a factor graph $\mathcal{G}$ of f is available. As a first step, we partition the set $\mathcal{E}$ of edges of $\mathcal{G}$ in non-overlapping subsets $\mathcal{E}_\ell$ such that each edge belongs to one subset $\mathcal{E}_\ell$. For example, the trial function (42) corresponds to the partitions $\mathcal{E}_1 = (Z, \Theta)$ and $\mathcal{E}_2 = X$. Note that the edges connected to an equality constraint node correspond to the same variable (e.g., X in Fig. 2(c)), and they are supposed to belong to the same $\mathcal{E}_\ell$. We associate a subgraph $\mathcal{G}_\ell \subseteq \mathcal{G}$ to each subset $\mathcal{E}_\ell$ consisting of (i) all nodes of $\mathcal{G}$ connected to edges of $\mathcal{E}_\ell$; (ii) all edges of $\mathcal{G}$ connected to those nodes. Note that $\mathcal{G}_\ell$ may contain edges that do not belong to $\mathcal{E}_\ell$. As an illustration, Fig. 7 depicts the subgraphs $\mathcal{G}_1$ and $\mathcal{G}_2$ associated to the subsets $\mathcal{E}_1 = X$ and $\mathcal{E}_2 = (Z, \Theta)$ respectively, corresponding to the trial function (42). Edges of $\mathcal{G}_\ell$ that do not belong to $\mathcal{E}_\ell$ are referred to as "external edges"; the other edges of $\mathcal{G}_\ell$ are called "internal edges". In subgraph $\mathcal{G}_1$ (see Fig. 8(a)), the edges Z and Θ are internal, and the edges X are external. In subgraph $\mathcal{G}_2$ (see Fig. 8(b)), the edges X are internal, and the edges Θ are external. In the following, we will assume that the subgraphs $\mathcal{G}'_\ell \subseteq \mathcal{G}$, obtained from $\mathcal{G}_\ell$ by removing the external edges, are cycle-free. Fig. 8 shows the subgraphs $\mathcal{G}'_1$ and $\mathcal{G}'_2$ obtained from $\mathcal{G}_1$ and $\mathcal{G}_2$ respectively (cf. Fig. 7); they are both cycle-free. A generic node g of $\mathcal{G}_\ell$ is depicted in Fig. 9. The edges $X_1, \ldots, X_n$ are internal edges, the edges $V_1, \ldots, V_r$ are external. The edges $V_2, \ldots, V_r$ are assumed to belong to the same subset $\mathcal{E}_\ell$, and V_1 is assumed not to belong to $\mathcal{E}_\ell$.

The generic structured variational message-passing algorithm iterates the following steps:



(a) Subgraph $\mathcal{G}_1$.



(b) Subgraph $\mathcal{G}_2$.

Figure 7: Subgraphs $\mathcal{G}_\ell$ of the factor graph $\mathcal{G}$ depicted in Fig. 4.

1. Select a subgraph $\mathcal{G}_\ell$.
2. Update the messages along internal edges of $\mathcal{G}_\ell$ according to the rule (see Fig. 9)

$$\vec{\mu}(x_n) \propto \int \vec{\mu}(x_1) \dots \vec{\mu}(x_{n-1}) \exp\Big[\int q(v_1) q(v_2, \dots, v_r) \log g(x_1, \dots, x_n, v_1, \dots, v_r) dv \Big] dx_1 \dots dx_{n-1}, \tag{57}$$

3. At nodes g connected to external edges, compute

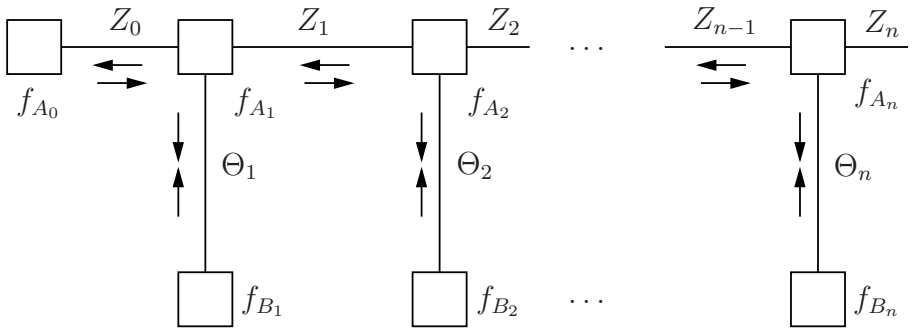
$$q(x_1, \dots, x_n) \propto \vec{\mu}(x_1) \dots \vec{\mu}(x_{n-1}) \overleftarrow{\mu}(x_n) \exp\Big[\int q(v_1) q(v_2, \dots, v_r) \log g(x_1, \dots, x_n, v_1, \dots, v_r) dv \Big] \tag{58}$$

$$\triangleq \vec{\mu}(x_1) \dots \vec{\mu}(x_{n-1}) \overleftarrow{\mu}(x_n) \mu\downarrow(x_1, \dots, x_n). \tag{59}$$

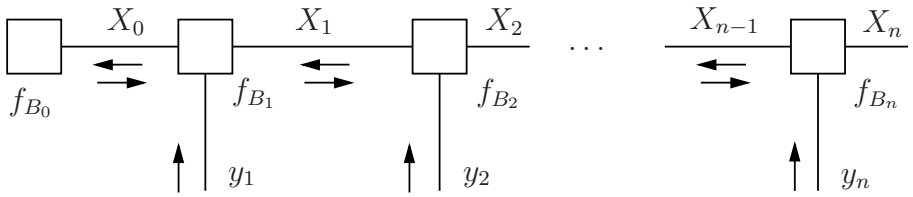
4. Iterate 1–3.

Some remarks:

- In order to keep the notation simple, we considered a particular factorization of $q(v_1, \dots, v_r)$ in (57) and (58). Obviously, both rules can easily be formulated for *any* factorization of the marginal $q(v_1, \dots, v_r)$.
- If all subgraphs $\mathcal{G}'_\ell$ are cycle-free, the above algorithm is a structured variational algorithm, and it is then guaranteed to convergence. The messages (57) are then updated according to the standard schedule for cycle-free graphs [51]. If one or more subgraph(s) $\mathcal{G}'_\ell$ is cyclic, the above



(a) Subgraph $\mathcal{G}'_1$.



(b) Subgraph $\mathcal{G}'_2$.

Figure 8: Subgraphs $\mathcal{G}'_\ell$ of the factor graph $\mathcal{G}$ depicted in Fig. 4.

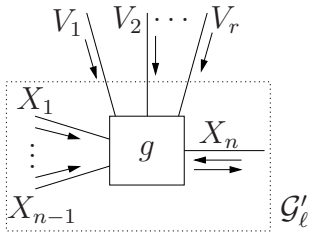


Figure 9: Structured variational message passing.

message-passing algorithm is no longer a variational algorithm, and there is no guarantee for convergence. One may first convert the cyclic subgraphs $\mathcal{G}'_\ell$ into cycle-free ones by clustering or stretching [51], and then apply the structured variational message-passing algorithm.

- One may also consider overlapping subsets $\mathcal{E}_\ell$ [40] [12], which leads to non-trivial modifications of the above message-passing scheme.
- If the node g (cf. Fig. 9) is only connected to internal edges of $\mathcal{G}_\ell$, the rule (57) boils down to the generic sum-product rule [14]. On the other hand, if the node g is connected to *one* internal edge X (i.e., $n = 1$ and $X \triangleq X_1$) and one or more external edge(s) $V_1, \ldots, V_r$, the rule (57) becomes

$$\vec{\mu}(x) \propto \exp\Big[\int q(v_1)q(v_2,\ldots,v_r)\log g(x,v_1,\ldots,v_r)dv\Big]. \tag{60}$$

  The update rule (60) is similar to the naive variational message computation rule (15). The difference lies in the factorization of the marginal $q(v_1,\ldots,v_r)$: in the naive variational rule (15), $q(v_1,\ldots,v_r)$ is fully factorized, whereas in the structured variational rule (60), it can be arbitrarily factorized.
- In the naive variational approach, all subsets $\mathcal{E}_\ell$ consist of single edges. The structured variational message-passing algorithm reduces to the naive variational message-passing algorithm of Section 3.

- It is easily verified that the structured variational algorithms we derived for the models (2) and (40) are particular instances of the above generic message-passing scheme.
- Structured variational message passing can also be used to determine the mode (5), similarly as naive variational message passing. Obviously, one may also combine structured variational message passing with other message-passing algorithms such as EM, ICM, gradient methods etc. The generic message computation rules of such combinations are similar to the ones we presented in Section 4. The fully factorized marginals $q(x|\hat{\theta}) = q(x_1|\hat{\theta}) \dots q(x_n|\hat{\theta})$ (cf., e.g., (26)(22)) are replaced by more structured factorizations.

6 Conclusion

We elaborated on variational message passing and have outlined various extensions in the context of factor graphs. We have underlined the tight connection between variational message passing and the message-passing formulation of EM. There are several topics for further investigation. For instance, it would be interesting to clarify the connection between variational message passing and expectation propagation in the setting of factor graphs.

Acknowledgments

The author wishes to thank Shin Ishii, Hans-Andrea Loeliger, Shin-ichi Maeda, Shigeyuki Oba, and Jonathan Yedidia for inspiring discussions.

References

[1] G. Parisi, *Statistical Field Theory,* Perseus Books, 1988.

[2] R. Feynman, *Statistical Mechanics: A Set of Lectures,* Perseus Books Group, 1998.

[3] J. Rustagi, *Variational Methods in Statistics,* Academic Press, New York, 1976.

[4] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Machine Learning,* 37:183–233, 1999.

[5] R. M. Neal and G. E. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants," in *Learning in Graphical Model,* M.I. Jordan (editor), MIT Press, 1998.

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

[7] H. J. Kappen and W. Wiegerinck, "Mean field theory for graphical models," *Advanced Mean Field Methods–Theory and Practice,* pp. 37–49, MIT Press, Cambridge, MA, 2001.

[8] M. J. Beal, *Variational Algorithms for Approximate Bayesian Inference,* PhD. Thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

[9] Z. Ghahramani and M. J. Beal, "Graphical Models and Variational Methods," in *Advanced Mean Field methods–Theory and Practice,* eds. D. Saad and M. Opper, MIT Press, 2000.

[10] Z. Ghahramani and M. J. Beal, "Propagation Algorithms for Variational Bayesian Learning," *Advances in Neural Information Processing Systems (NIPS),* vol. 13, MIT Press, 2001.

[11] C. M. Bishop, *Pattern Recognition and Machine Learning* (Chapter 10), Springer, 2006.

[12] J. Winn, *Variational Message Passing and its Applications,* PhD. Thesis, Cambridge University, 2003.

[13] J. Winn and C. Bishop, "Variational Message Passing," *Journal of Machine Learning Research,* Vol. 6, pp. 661-694, 2005.

[14] H.-A. Loeliger, "An Introduction to Factor Graphs," *IEEE Signal Proc. Mag.,* Jan. 2004, pp. 28–41.

[15] M. Nissilä and S. Pasupathy, "Reduced-complexity turbo receivers for single and multi-antenna systems via variational inference in factor graphs," in *Proc. IEEE International Conference on Communications (ICC'04),* 20–24 June, 2004, Paris, France, pp. 2767–2771.

[16] P. Stoica and Y. Selén, "Cyclic Minimizers, Majorization Techniques, and the Expectation-Maximization Algorithm: a Refresher," *IEEE Signal Proc. Mag.,* January 2004, pp. 112–114.

[17] D. P. Bertsekas, *Nonlinear Programming,* Athena Scientific, Belmont, MA, 1995.

[18] J. Dauwels, *On Graphical Models for Communications and Machine Learning: Algorithms, Bounds, and Analog Implementation,* PhD. Thesis at ETH Zurich, Diss. ETH No 16365, December 2005. Available from `www.dauwels.com/PhD.htm`.

[19] J. Dauwels, S. Korl, and H.-A. Loeliger, "Steepest Descent on Factor Graphs," *Proc. IEEE Information Theory Workshop,* Rotorua, New Zealand, Aug. 28–Sept. 1, 2005, pp. 42–46.

[20] A. P. Dempster, N. M. Laird, and D. B. Rubin "Maximum Likelihood From Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society,* B 39, pp. 1–38, 1977.

[21] A. W. Eckford and S. Pasupathy, "Iterative multiuser detection with graphical modeling" *IEEE International Conference on Personal Wireless Communications,* Hyderabad, India, 2000.

[22] J. Dauwels, S. Korl, and H.-A. Loeliger, "Expectation Maximization as Message Passing", *Proc. Int. Symp. on Information Theory (ISIT),* Adelaide, Australia, Sept. 4–9, 2005, pp. 583–586.

[23] J. Dauwels, A. W. Eckford, S. Korl, and H. A. Loeliger, "Expectation Maximization on Factor Graphs," in preparation.

[24] G. McLachlan and T. Krishnan, *The EM Algorithm and Extensions,* Wiley, 1997.

[25] S. Amari and H. Nagaoka, *Methods of Information Geometry,* Oxford university Press, 2000.

[26] I. Csiszár, "Information type measures of difference of probability distributions and indirect observations," *Studia Scientiarium Mathematicarum Hungarica,* 2:299–318, 1967.

[27] S. Amari, S. Ikeda, and H. Shimokawa, "Information geometry and mean field approximation: the alpha-projection approach," in Manfred Opper and David Saad, editors, *Advanced Mean Field Methods–Theory and Practice,* Chapter 16, pp. 241–257, MIT Press, Cambridge, MA, 2001.

[28] T. Toyoizumi and K. Aihara, "Mean-field and Variational Methods for alpha-families," *Trans. Instit. Electron.,* 86-D2 (in Japanese), pp. 959–965, 2003.

[29] B. J. Frey and N. Jojic, "A Comparison of Algorithms for Inference and Learning in Probabilistic Graphical Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 27, no. 9, Sept. 2005.

[30] T. Minka, "Divergence measures and message passing," Microsoft Research Technical Report (MSR-TR-2005-173), 2005.

[31] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Constructing Free-Energy Approximations and Generalized Belief Propagation Algorithms," *IEEE Trans. Information Theory*, vol. 51, no. 7, pp. 2282–2312, July 2005.

[32] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," Technical Report 649, Department of Statistics, University of California, Berkeley, 2003.

[33] M. J. Beal, "Variational Bayesian quick-reference sheet," available from `http://www.cse.buffalo.edu/faculty/mbeal/papers/vbqref/vbqref.html`, 2000.

[34] C. Robert and G. Casella, *Monte Carlo Statistical Methods,* Springer Texts in Statistics, 2nd ed., 2004.

[35] N. de Freitas, P. Højen-Sørensen, M. I. Jordan, and S. Russell, "Variational MCMC," *Proc. 17th Uncertainty in Artificial Intelligence (UAI),* 2001.

[36] F. Forbes and G. Fort, "Combining simulation and mean-field like methods for inference in Hidden Markov Random Fields," INRIA Technical Report no. 5721, January 2006.

[37] J. Bezdek and R. Hathaway, "Some Notes on Alternating Optimization," *Proc. AFSS Int. Conference on Fuzzy Systems,* Calcutta, India, February 3–6, 2002.

[38] K. B. Peterse, O. Winther, and L. K. Hansen "On the Slow Convergence of EM and VBEM in Low-Noise Linear Models," *Neural Computation,* 17, pp. 1921–1926, 2005.

[39] R. Salakhutdinov, S. Roweis, and Z. Ghahramani, "Optimization with EM and Expectation-Conjugate-Gradient", *Proc. Intl. Conf. on Machine Learning (ICML '03),* Washington DC, USA, August 21–24, 2003, vol. 20, pp. 672–679.

[40] W. Wiegerinck, "Variational approximations between mean field theory and the junction tree algorithm," *Proc. 16-th UAI-2000,* pp. 626–633.

[41] D. Geiger, "Structured Variational Inference Procedures and their Realizations," *Proc. Tenth International Workshop on Artificial Intelligence and Statistics,* Barbados, January 6–8, 2005.

[42] E. P. Xing, M. I. Jordan, and S. Russell, "A generalized mean field algorithm for variational inference in exponential families," *Proc. Uncertainty in Artificial Intelligence (UAI2003),* Morgan Kaufmann Publishers, pp. 583–591, 2003.

[43] E. P. Xing, M. I. Jordan, and S. Russell, "Graph partition strategies for generalized mean field inference," *Proc. Uncertainty in Artificial Intelligence 20 (UAI2004),* AUAI Press, pp. 602–610, 2004.

[44] D. J. C. MacKay, "Ensemble Learning for Hidden Markov Models," available from `http://wol.ra.phy.cam.ac.uk/mackay/`, 1997.

[45] M. J. Beal and Z. Ghahramani, "The Variational Kalman Smoother," Gatsby Unit Technical Report TR01-003, 2003.

[46] A. Doucet, J. F. G. de Freitas, and N. J. Gordon, eds., *Sequential Monte Carlo Methods in Practice.* New York: Springer-Verlag, 2001.

[47] M. Isard, "Pampas: Real-Valued Graphical Models for Computer Vision," *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '03),* Madison, Wisconsin, USA, June 16–22, 2003, vol. 1, pp. 613–620.

[48] E. Sudderth, A. Ihler, W. Freeman, and A. Willsky "Non-Parametric Belief Propagation," *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '03),* Madison, Wisconsin, USA, June 16–22, 2003.

[49] M. Briers, A. Doucet and S. S. Singh, "Sequential Auxiliary Particle Belief Propagation," *Proc. 7th Intern. Conf. on Information Fusion,* 2005.

[50] J. Dauwels, S. Korl, and H.-A. Loeliger, "Particle Methods as Message Passing", *Proc. Int. Symp. on Information Theory (ISIT),* Seattle, USA, July 9–14, 2006, pp. 2052–2056.

[51] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor Graphs and the Sum-Product Algorithm," *IEEE Trans. Information Theory,* vol. 47, pp. 498–519, Feb. 2001.