ON VARIATIONAL MESSAGE PASSING ON FACTOR GRAPHS

Justin Dauwels

Amari Research Unit, RIKEN Brain Science Institute, Wako-shi, 351-0106, Saitama, Japan
email: justin@dauwels.com

ABSTRACT

In this paper, it is shown how (naive and structured) variational algorithms may be derived from a factor graph 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 compared to and combined with various other message-passing inference 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.

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], where they are called "mean-field approximations". Variational methods have also been adopted for statistical inference (see, e.g., [2]–[8]), which is the topic of this paper. We will consider the following generic inference problem: suppose 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

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

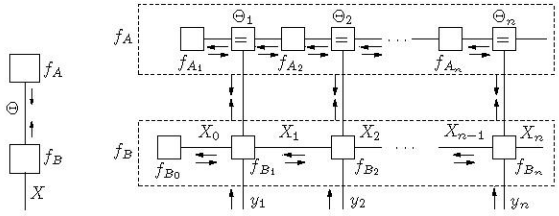
where $\int_\theta$ denotes either summation or integration over the whole range of Θ.

The 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

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

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

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; we use Forney-style factor graphs ("normal factor graphs") [9], where each edge corresponds to a variable and each node corresponds to a factor (see [10] 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).

Fig. 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 [10]. 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 [3] [7]). The naive and structured variational method have been formulated as message-passing algorithms by Bishop et al. [7], Xing et al. [23] and Geiger [22] 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 [6, pp. 256–258] [11]. 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 compare it to other message-passing algorithms; we may then also straightforwardly combine 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 [25] [3] and variational particle filters and smoothers. If the variational messages are intractable, they may

1-4244-1429-6/07/$25.00 ©2007 IEEE

be represented as particle lists, resulting in particle-based algorithms such as variational Markov Chain Monte Carlo methods [21].

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 [10]. 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).

Variational message passing may also be used to address those two problems, which results in variational estimation algorithms such as variational ICM, variational EM, and variational gradient methods.

This paper is structured as follows. In the following section, we review the naive variational method (closely following [2]–[7]). 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.

2. REVIEW OF THE NAIVE VARIATIONAL METHOD

Assume that we are given a generic multivariate function $f(z)$ (not necessarily normalized) with $z \in \mathbb{R}^m$, and suppose that we wish to compute its marginals

$$f(z_k) \triangleq \int f(z_1,\ldots,z_m)dz_1dz_2\ldots dz_{k-1}dz_{k+1}\ldots dz_m, \tag{7}$$

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 $q(z)$ (belonging to a family $\mathcal{Q}$ of trial functions) that is as "close" as possible to $f(z)$, i.e.,

$$q^* \triangleq \operatorname*{argmin}_{q\in\mathcal{Q}} D\left(f,q\right), \tag{8}$$

where $D(f,q)$ is a measure for the distance between f and q. The marginals $f(z_k)$ (7) 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.

If f is normalized, a popular measure is the Kullback-Leibler divergence $D\left(q\|f\right)$ defined as

$$D\left(q\|f\right) \triangleq \int_x q(x)\log\frac{q(x)}{f(x)}dx. \tag{9}$$

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

$$q(z_1,\ldots,z_m) \triangleq \prod_{k=1}^{m} q(z_k), \tag{10}$$

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

$$q^* \triangleq \operatorname*{argmin}_{q\in\mathcal{Q}} D\left(q\|f\right), \tag{11}$$

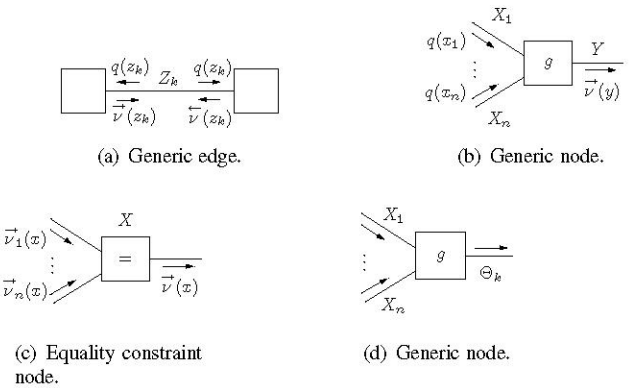
and the marginals $f(z_k)$ (7) are approximated by $q^*(z_k)$. Note that the objective function $D\left(q\|f\right)$ (11) is in general non-convex in the factors $q(z_k)$. By variational calculus, one can easily verify that $q^*(z_k)$ fulfills the equality

$$q^*(z_k) \overset{\triangle}{\propto} \exp\Big(\int q^*(z_1)\ldots q^*(z_{k-1})q^*(z_{k+1})\ldots q^*(z_m) \log f(z_1,\ldots,z_m)dz_1dz_2\ldots dz_{k-1}dz_{k+1}\ldots dz_m\Big). \tag{12}$$

The equality (12) suggests to determine (11) by iterating the update rule

$$q^{(\ell+1)}(z_k) \overset{\triangle}{\propto} \exp\Big(\int q^{(\ell)}(z_1)\ldots q^{(\ell)}(z_{k-1})q^{(\ell)}(z_{k+1})\ldots q^{(\ell)}(z_m) \cdot \log f(z_1,\ldots,z_m)dz_1dz_2\ldots dz_{k-1}dz_{k+1}\ldots dz_m\Big), \tag{13}$$

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 $D\left(q\|f\right)$ decreases, unless the algorithm has reached a fixed point; the method is guaranteed to convergence to a local minimum of $D\left(q\|f\right)$ (see [2]–[7]).



(a) Generic edge. (b) Generic node.

(c) Equality constraint node. (d) Generic node.

Fig. 2. Variational message passing.

3. NAIVE VARIATIONAL MESSAGE PASSING

If the function f factorizes, the update (13) 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 [6] [7] in the setting of directed 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 [6, pp. 256–258]; Nissilä et al. considered the particular case of factorial hidden Markov models with conditionally Gaussian distributed observations [11]. We will now formulate the generic variational message computation rule in the notation of factor graphs.

As is easily verified from (13), 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))

$$\vec{\nu}(y) \propto \exp \int q(x_1)q(x_2)\ldots q(x_n) \cdot \log g(x_1, \ldots, x_n, y) dx_1 \ldots dx_n \quad (14)$$

$$\stackrel{\triangle}{\propto} \exp \mathrm{E}_q\big[\log g(X_1, \ldots, X_n, y)\big]. \quad (15)$$

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

$$q(z_k) \propto \vec{\nu}(z_k)\overleftarrow{\nu}(z_k), \quad (16)$$

and send it to the two nodes connected to the edge X_k.

5. Iterate 2–4 until convergence.

Some remarks:

- Interestingly, the rule (14) is often simpler than the sum-product rule [10], especially if the model $f(x, \theta, y)$ belongs to the conjugate-exponential family [3] [7].
- 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 (14) 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:

$$\vec{\nu}(x) \propto \vec{\nu}_1(x)\vec{\nu}_2(x)\ldots\vec{\nu}_n(x). \quad (17)$$

  Other deterministic nodes can often (but not always!) be handled by combining them with non-deterministic nodes.
- 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) with target function (16), leading to variational MCMC [21].

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. (10))

$$q(x, \theta) \triangleq \prod_k q(x_k) \prod_\ell q(\theta_\ell). \quad (18)$$

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 $q(x) \triangleq q(x_1)\ldots q(x_n)$.

4. NAIVE VARIATIONAL 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. Eventually, the mode (5) is then approximated by the mode of $q(\theta) \triangleq q(\theta_1)\ldots q(\theta_n)$. If the mode of $q(\theta)$ is not available in closed form, one may resort to standard optimization techniques such as ICM [12] ("variational ICM") and gradient methods [13] ("variational gradient algorithms").

Alternatively, one may determine the mode (5) by EM [16]. If the E-step is intractable, one may approximate the E-step by naive variational methods ("variational EM") [4]; the intractable marginals required in the E-step are then replaced by approximate marginals q.

Recently, ICM [14], gradient methods [15], and EM [17][18][19], 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 just 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 (14). Gradient methods for solving (5) involve the gradient of logarithmic sum-product messages [15] (see Fig. 2(d))

$$\nabla_{\theta_k} \log \mu(\theta_k) = \frac{\int \mu(x_1; \hat{\theta}) \ldots \mu(x_n; \hat{\theta}) \nabla_{\theta_k} g(x, \theta_k) dx}{\int \mu(x_1; \hat{\theta}) \ldots \mu(x_n; \hat{\theta}) g(x, \theta_k) dx}, \quad (19)$$

where $\mu(x_k; \hat{\theta})$ $(k = 1, \ldots, n)$ are sum-product messages and $x = x_1, x_2, \ldots, x_n$. In naive variational gradient methods, those messages are replaced by the gradient of log-variational messages

$$\nabla_{\theta_k} \log \nu(\theta_k) = \int q(x_1; \hat{\theta}) \ldots q(x_n; \hat{\theta}) \nabla_{\theta_k} \log g(x, \theta_k) dx \quad (20)$$

$$= \mathrm{E}_q\big[\nabla_{\theta_k} \log g(X, \theta_k)\big]. \quad (21)$$

The E-step in (standard) EM involves the computation of E-log messages [17][18][19] (see Fig. 2(d))

$$h(\theta_k) = \int p(x; \hat{\theta}) \log g(x, \theta_k) dx \quad (22)$$

$$= \mathrm{E}_p\big[\log g(X, \theta_k); \hat{\theta}\big]. \quad (23)$$

In the E-step of naive variational EM, those messages are replaced by log-variational messages (cf. (14))

$$\log \nu(\theta_k) = \int q(x_1; \hat{\theta}) \ldots q(x_n; \hat{\theta}) \log g(x, \theta_k) dx \quad (24)$$

$$= \mathrm{E}_q\big[\log g(X, \theta_k); \hat{\theta}\big]. \quad (25)$$

5. STRUCTURED VARIATIONAL MESSAGE PASSING

So far, we have considered fully factorized trial functions (cf. (10)). In this section, we consider more structured factorizations, leading to "structured" variational algorithms [24] [2] [22][23]. Structured variational methods have been formulated as message-passing algorithms by Bishop et al. [7] [8], Xing et al. [23] and Geiger [22]

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 make the connection between structured variational message passing and the message-passing formulation of EM [17] [18] [19].

5.1. An 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

$$q(x,\theta) \triangleq q(x)q(\theta), \tag{26}$$

where $q(x)$ and $q(\theta)$ are *not* further factorized, in contrast to (18). Based on the trial (26), one may derive a "structured" variational method [2] [22][23]: through variational calculus one obtains an equality similar to (12) and an update rule similar to (13). Iterating that update rule amounts to a "structured" variational algorithm that can be formulated as the following message-passing procedure (see Fig. 1(b)):

Update $q(x)$
Perform the forward recursion

$$\overrightarrow{\mu}(x_k) \propto \int \overrightarrow{\mu}(x_{k-1}) \exp\Big[\int q(\theta_k) \log f_{B_k}(x_{k-1},x_k,\theta_k) d\theta_k\Big] dx_{k-1}, \tag{27}$$

and the corresponding backward recursion with messages $\overleftarrow{\mu}(x_k)$.
Update $q(x_{k-1},x_k)$:

$$q(x_{k-1},x_k) \propto \overrightarrow{\mu}(x_{k-1}) \exp\Big[\int q(\theta_k) \log f_{B_k}(x_{k-1},x_k,\theta_k) d\theta_k\Big] \overleftarrow{\mu}(x_k). \tag{28}$$

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

$$\nu\uparrow(\theta_k) \propto \exp \int q(x_{k-1},x_k) \log f_{B_k}(x_{k-1},x_k,y_k,\theta_k) dx_{k-1} dx_k. \tag{29}$$

Perform the two-step forward recursion

$$\overrightarrow{\mu}'(\theta_k) \propto \int \overrightarrow{\mu}(\theta_{k-1}) f_{A_k}(\theta_{k-1},\theta_k) d\theta_{k-1} \tag{30}$$

$$\overrightarrow{\mu}(\theta_k) \propto \overrightarrow{\mu}'(\theta_k) \nu\uparrow(\theta_k), \tag{31}$$

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

$$\mu\downarrow(\theta_k) \stackrel{\triangle}{\propto} \overrightarrow{\mu}'(\theta_k) \overleftarrow{\mu}(\theta_k) = \overrightarrow{\mu}(\theta_k) \overleftarrow{\mu}'(\theta_k). \tag{32}$$

Update $q(\theta_k)$:

$$q(\theta_k) \propto \nu\uparrow(\theta_k) \mu\downarrow(\theta_k). \tag{33}$$

Some remarks:

- Since the above message-passing scheme is a (structured) variational algorithm, it is guaranteed to converge [2] [22][23].
- The marginals $q(x_k)$ are computed as $q(x_k) \propto \overrightarrow{\mu}(x_k) \overleftarrow{\mu}(x_k)$.
- The updates (30)–(32) are instances of the sum-product rule [10].
- 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" [25]. Alternatively, those messages may be represented as particle lists (see, e.g., [26]–[27]); Step 1 and/or Step 4 then involves particle smoothing ("variational particle smoother"); this option does not seem to have been explored yet.
- 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 (27)–(33) 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. (27))

$$\overrightarrow{\mu}(x_k) \propto \int \overrightarrow{\mu}(x_{k-1}) f_{B_k}(x_{k-1},x_k,\hat{\theta}_k^{(\ell)}) dx_{k-1}, \tag{34}$$

and the corresponding backward sweep with messages $\overleftarrow{\mu}(x_k)$. Compute the upward messages (cf. (29))

$$\exp(h(\theta_k)) \propto \exp \int p(x_{k-1},x_k;\hat{\theta}^{(\ell)}) \cdot \log f_{B_k}(x_{k-1},x_k,\theta_k) dx_{k-1} dx_k, \tag{35}$$

where (cf. (28))

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

M-step (cf. (30)–(33))

$$\hat{\theta}^{(\ell+1)} = \operatorname*{argmax}_{\theta} \Big[f_A(\theta) \exp(h(\theta_1)) \ldots \exp(h(\theta_n))\Big]. \tag{37}$$

We formulated this EM algorithm as a message-passing algorithm operating on the factor graph of Fig. 1(b). Note that the message $h(\theta_k)$ (35) is a particular instance of the generic E-log message (22) [17] [18] [19]. The message $\exp(h(\theta_k))$ (35) is closely related to $\nu\uparrow(\theta_k)$ (29): the marginal $p(x_{k-1},x_k;\hat{\theta}^{(\ell)})$ in (35) is replaced by a variational marginal $q(x_{k-1},x_k)$ in (29). Since we started from a non-factorized trial function $q(x)$ (cf. (26)), we obtained the standard EM algorithm; a factorized trial $q(x)$ leads to a structured variational EM algorithm (see Section 5.2). 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. Generic Formulation

From the previous example, it is straightforward to formulate a general recipe to derive structured variational algorithms from factor graphs. Let f be 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 (26) corresponds to the partitions $\mathcal{E}_1 = \Theta$ and $\mathcal{E}_2 = X$. Note that

the edges connected to an equality constraint node correspond to the same variable (e.g., Θ_k in Fig. 1(b)), 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 $\mathcal{G}$ that are connected to edges of $\mathcal{E}_\ell$; (ii) all edges of $\mathcal{G}$ that are connected to those nodes. Note that $\mathcal{G}_\ell$ may contain edges that do not belong to $\mathcal{E}_\ell$; the latter are referred to as "external edges", the other edges of $\mathcal{G}_\ell$ are called "internal edges". 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. A generic node g of $\mathcal{G}_\ell$ is depicted in Fig. 3. The edges $X_1, \ldots, X_n$ are internal edges, the edges $V_1, \ldots, V_r$ are external. For the sake of definiteness, we assume that the edges $V_2, \ldots, V_r$ belong to the same subset $\mathcal{E}_\ell$, whereas V_1 is assumed not to belong that subset—our considerations are trivially extendable to other partitions.

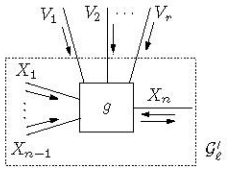


Fig. 3. Structured variational message passing.

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

1. Select a subgraph $\mathcal{G}_\ell$.
2. Update the messages along internal edges X_n of $\mathcal{G}_\ell$ according to the rule (see Fig. 3)

$$\vec{\mu}(x_n) \propto \int \vec{\mu}(x_1) \ldots \vec{\mu}(x_{n-1}) \cdot \exp\Big[\int q(v_1) q(v_2, \ldots, v_r) \log g(x, v) dv\Big] dx_1 \ldots dx_{n-1}, \tag{38}$$

3. At nodes g connected to external edges (cf. Fig. 3), compute

$$q(x_1, \ldots, x_n) \propto \vec{\mu}(x_1) \ldots \vec{\mu}(x_{n-1}) \overleftarrow{\mu}(x_n) \cdot \exp\Big[\int q(v_1) q(v_2, \ldots, v_r) \log g(x, v) dv\Big] \tag{39}$$

4. Iterate 1–3.

Some remarks:

- If all subgraphs $\mathcal{G}'_\ell$ are cycle-free, the above algorithm is a structured variational algorithm, and it is guaranteed to converge; otherwise, there is no guarantee for convergence. One may first convert the cyclic subgraphs $\mathcal{G}'_\ell$ into cycle-free subgraphs and then apply the structured variational message-passing algorithm.
- If the node g (cf. Fig. 3) is only connected to internal edges of $\mathcal{G}_\ell$, the rule (38) boils down to the generic sum-product rule [10]. 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 to one or more external edge(s) $V_1, \ldots, V_r$, the rule (38) becomes

$$\vec{\mu}(x) \propto \exp\Big[\int q(v_1) q(v_2, \ldots, v_r) \log g(x, v) dv\Big], \tag{40}$$

  which is similar to the naive variational message computation rule (14). The marginal $q(v_1, \ldots, v_r)$ is now not fully factorized, i.e., it may now be arbitrarily factorized.
- In the naive variational approach, each subset $\mathcal{E}_\ell$ contains either a single edge or all edges connected to a particular equality constraint node (e.g., the three edges connected to each equality constraint node Θ_k in Fig. 1(b)).
- It is easily verified that the example of Section 5.1 is a particular instance of the above generic message-passing scheme.
- Structured variational message passing can also be used to determine the mode (5). The generic message computation rules of such estimation algorithms are similar to the ones of Section 4. The fully factorized marginals $q(x; \hat{\theta}) = q(x_1; \hat{\theta}) \ldots q(x_n; \hat{\theta})$ (cf. (20) (24)) are replaced by more structured factorizations.

6. ACKNOWLEDGMENTS

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

7. REFERENCES

[1] G. Parisi, *Statistical Field Theory*, Perseus Books, 1988.
[2] 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.
[3] M. J. Beal, *Variational Algorithms for Approximate Bayesian Inference*, PhD. Thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.
[4] 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.
[5] C. M. Bishop, *Pattern Recognition and Machine Learning* (Chapter 10), Springer, 2006.
[6] J. Winn, *Variational Message Passing and its Applications*, PhD. Thesis, Cambridge University, 2003.
[7] J. Winn and C. Bishop, "Variational Message Passing," *Journal of Machine Learning Research*, Vol. 6, pp. 661-694, 2005.
[8] C. Bishop and J. Winn, "Structured Variational Distributions in VIBES," *Proc. Artificial Intelligence and Statistics*, Key West, Florida,USA, Jan. 3–6, 2003.
[9] G. D. Forney, Jr., "Codes on Graphs: Normal Realizations," *IEEE Trans. Information Theory*, vol. 47, no. 2, pp. 520–548, 2001.
[10] H.-A. Loeliger, "An Introduction to Factor Graphs," *IEEE Signal Proc. Mag.*, Jan. 2004, pp. 28–41.
[11] 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.
[12] 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.
[13] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, Belmont, MA, 1995.
[14] 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`.
[15] 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.
[16] 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.
[17] A. W. Eckford and S. Pasupathy, "Iterative Multiuser Detection with Graphical Modeling" *IEEE International Conference on Personal Wireless Communications*, Hyderabad, India, 2000.
[18] 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.
[19] J. Dauwels, A. W. Eckford, S. Korl, and H. A. Loeliger, "Expectation Maximization on Factor Graphs," in preparation.
[20] G. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*, Wiley, 1997.
[21] N. de Freitas, P. Højen-Sørensen, M. I. Jordan, and S. Russell, "Variational MCMC," *Proc. 17th Uncertainty in Artificial Intelligence (UAI)*, 2001.
[22] D. Geiger, "Structured Variational Inference Procedures and their Realizations," *Proc. Tenth International Workshop on Artificial Intelligence and Statistics*, Barbados, January 6–8, 2005.
[23] 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.
[24] D. J. C. MacKay, "Ensemble Learning for Hidden Markov Models," available from `http://wol.ra.phy.cam.ac.uk/mackay/`, 1997.
[25] M. J. Beal and Z. Ghahramani, "The Variational Kalman Smoother," Gatsby Unit Technical Report TR01-003, 2003.
[26] A. Doucet, J. F. G. de Freitas, and N. J. Gordon, eds., *Sequential Monte Carlo Methods in Practice*. New York: Springer-Verlag, 2001.
[27] 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.