ON CONVERGENCE PROPERTIES OF MESSAGE-PASSING ESTIMATION ALGORITHMS

Justin Dauwels

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

ABSTRACT

The convergence and stability properties of several message-passing estimation algorithms are investigated, i.e., (generalized) expectation maximization, gradient methods, and iterated conditional modes. Results are presented for cycle-free and cyclic factor graphs.

1. INTRODUCTION

In this paper, we focus on the following generic problem. Assume that we wish to find

$$\hat{\theta}_{\max} \triangleq \underset{\theta}{\operatorname{argmax}}\, f(\theta), \tag{1}$$

where θ takes values in a subset Ω of $\mathbb{R}^n$. In particular, we are interested in solving (1) in the case where $f(\theta)$ is a "marginal" of a real-valued function $f(x, \theta)$:

$$f(\theta) = \int_x f(x, \theta)\, dx, \tag{2}$$

where $\int_x$ denotes either summation or integration over the whole range of x. We will assume $f(x, \theta) > 0$ for all x and all θ. The maximization (1) lies at the heart of many estimation problems. As an example, we consider parameter estimation in state space models (as, e.g., hidden Markov models, stochastic dynamical systems); the variables x and θ are then vectors, and the function $f(x, \theta)$ is given by

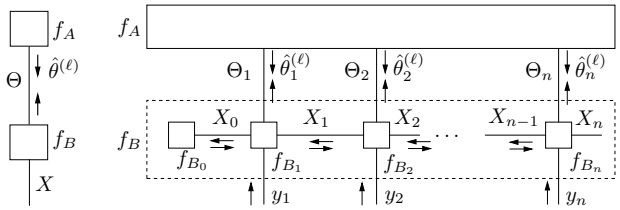
$$f(x, \theta) \triangleq f_A(\theta) f_B(x, \theta), \tag{3}$$

$$\triangleq f_A(\theta) f_{B_0}(x_0) f_{B_1}(x_0, x_1, y_1, \theta_1) f_{B_2}(x_1, x_2, y_2, \theta_2) \cdots f_{B_n}(x_{n-1}, x_n, y_n, \theta_n), \tag{4}$$

where X_k denotes the state at time k, 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 (3) and (4) is shown in Fig. 1(a) and Fig. 1(b) respectively (see [4] for a tutorial on factor graphs); the box f_B in Fig. 1(a) is detailed in Fig. 1(b) (dashed box). In model (4), the naive computation of (1) and/or (2) is often not feasible.

We will now assume that a factor graph for $f(x, \theta)$ is available. It may then be possible to compute $f(\theta)$ (2) and $\hat{\theta}_{\max}$ (1) by sum-product message passing and by max-product message passing, respectively [4]. Unfortunately, this approach 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 (2).



(a) Graph of (3). (b) Graph of (4).

Fig. 1. Factor graphs.

2. The max-product rule may lead to an intractable expression; in this case, we cannot compute (1).

If the maximization (1) is intractable, one may resort to standard optimization techniques such as iterative conditional modes (ICM, a.k.a. "coordinate ascent/descent") [2] or gradient methods [3]. In earlier work, we have applied those two methods to (1) (2); we described them as message-passing algorithms operating on a factor graph of $f(x, \theta)$ [10] [5]. An alternative way to compute $\hat{\theta}_{\max}$ (1) (exactly or approximately) is expectation maximization (EM) [1]. A message-passing view of EM is developed in [6] [7] [8]. The maximization step of EM is sometimes intractable, and one may then resort to ICM or some gradient method; such modifications are referred to as "generalized EM algorithms" [1]. In [10, Section 4.9.5] [5] [8] we described various generalized EM algorithms in the context of factor graphs, in particular, in the setting of problem (1) (2).

In this paper, we analyze the convergence and stability properties of the above mentioned classes of iterative estimation algorithms; we consider the standard formulation of those algorithms in addition to various extensions in the context of message passing on graphs. This paper is structured as follows. In the following section, we briefly describe ICM, gradient ascent, and (generalized) EM in the context of factor graphs. We outline our convergence and stability results in Section 3.

2. MESSAGE-PASSING ESTIMATION ALGORITHMS

As in [7], we first consider the trivial factorization

$$f(x, \theta) = f_{\mathrm{A}}(\theta) f_{\mathrm{B}}(x, \theta), \tag{5}$$

where $f_{\mathrm{A}}(\theta)$ may be viewed as encoding the a priori information about Θ (cf. Fig. 1(a)). In each of the four classes of algorithms, one tries to determine the mode (1) by recomputing an upwards and a downwards message, as indicated in Fig. 1(a). In Section 2.2 to 2.5 we list the explicit form of the upwards and downwards message for each of the four iterative algorithms. For the sake of definiteness, we will discuss gradient EM as an illustration of a generalized EM algorithm. In Section 2.6, we will

consider non-trivial factorizations. In Section 2.7, we will elaborate on various message-passing extensions of the standard formulation of the four iterative estimation algorithms. As a starting point of our discussion of the four iterative estimation algorithms, we describe a non-iterative (and often impractical) algorithm that (naively) computes (2) by the sum-product rule.

2.1. Non-Iterative Algorithm

Upwards message:

$$\mu(\theta) \triangleq \int_x f_B(x,\theta)\,dx. \tag{6}$$

Downwards message:

$$\hat{\theta} = \operatorname*{argmax}_\theta \big[\log \mu(\theta) + \log f_A(\theta)\big] \tag{7}$$

$$\triangleq \operatorname*{argmax}_\theta \log f(\theta). \tag{8}$$

The upwards message $\mu(\theta)$ is a standard sum-product message [4].

2.2. Iterative Conditional Modes

Upwards message:
Select an index $i \in \{1,2,\ldots,n\}$ and compute

$$\mu(\theta_i) \triangleq \int_x f_B(x,\hat{\theta}^{(\ell)}(\theta_i))\,dx. \tag{9}$$

where $\hat{\theta}^{(\ell)}(\theta_i) \triangleq (\hat{\theta}_1^{(\ell)},\ldots,\hat{\theta}_{i-1}^{(\ell)},\theta_i,\hat{\theta}_{i+1}^{(\ell)},\ldots,\hat{\theta}_n^{(\ell)})$.
Downwards message:

$$\hat{\theta}_j^{(\ell+1)} = \begin{cases} \operatorname{argmax}_{\theta_i} \log f_i(\theta_i) & \text{if } j = i \\ \hat{\theta}_j^{(\ell)} & \text{otherwise,} \end{cases} \tag{10}$$

where $f_i(\theta_i) \triangleq f(\hat{\theta}^{(\ell)}(\theta_i))$
The upwards message $\mu(\theta_i)$ is also here a standard sum-product message [4].

2.3. Gradient Ascent

Upwards message:

$$\nabla_\theta \log \mu(\hat{\theta}^{(\ell)}) \triangleq \frac{\int_x \nabla_\theta f_B(x,\theta)\big|_{\hat{\theta}^{(\ell)}}\,dx}{\int_x f_B(x,\hat{\theta}^{(\ell)})\,dx}. \tag{11}$$

Downwards message:

$$\hat{\theta}^{(\ell+1)} = \hat{\theta}^{(\ell)} + \lambda_k\Big(\nabla_\theta \log \mu(\hat{\theta}^{(\ell)}) + \nabla_\theta \log f_A\big((\hat{\theta}^{(\ell)})\big)\Big). \tag{12}$$

The upwards message is the gradient of (the log of) a standard sum-product message [5].

2.4. Expectation Maximization (EM)

Upwards message:

$$h(\theta) = \frac{\int_x f_{\mathrm{B}}(x,\hat{\theta}^{(\ell)}) \log f_{\mathrm{B}}(x,\theta)\,dx}{\int_x f_{\mathrm{B}}(x,\hat{\theta}^{(\ell)})\,dx} \tag{13}$$

$$= \mathrm{E}_{p_{\mathrm{B}}}[\log f_{\mathrm{B}}(X,\theta)]\,, \tag{14}$$

where $\mathrm{E}_{p_{\mathrm{B}}}$ denotes the expectation with respect to the probability distribution

$$p_{\mathrm{B}}(x|\hat{\theta}^{(\ell)}) \triangleq \frac{f_{\mathrm{B}}(x,\hat{\theta}^{(\ell)})}{\sum_x f_{\mathrm{B}}(x,\hat{\theta}^{(\ell)})}. \tag{15}$$

Downwards message:

$$\hat{\theta}^{(\ell+1)} = \operatorname*{argmax}_\theta \big(h(\theta) + \log f_{\mathrm{A}}(\theta)\big). \tag{16}$$

The upwards $h(\theta)$ is an "E-log message" [7] [8], which is *not* a standard sum-product message.

2.5. Gradient EM

Upwards message:

$$\nabla_\theta h(\hat{\theta}^{(\ell)}) = \mathrm{E}_{p_{\mathrm{B}}}\big[\nabla_\theta \log f_{\mathrm{B}}(X,\theta)\big|_{\hat{\theta}^{(\ell)}}\big]\,. \tag{17}$$

Downwards message:

$$\hat{\theta}^{(\ell+1)} = \hat{\theta}^{(\ell)} + \lambda_k\Big(\nabla_\theta h(\hat{\theta}^{(\ell)}) + \nabla_\theta \log f_A\big((\hat{\theta}^{(\ell)})\big)\Big). \tag{18}$$

The upwards message is the gradient of an E-log message [5]; it is identical to the message (11) (occurring in gradient ascent). Gradient EM is almost identical to gradient ascent with the marginal $\log f(\theta)$ as objective function; the only difference lies in the update schedule: gradient EM tries to find the maximum (16), and the marginal p_B remains fixed until that point (or an other stationary point of $h(\theta) + \log f_A(\theta)$) has been attained; in gradient ascent, the marginal p_B is updated at each iteration.

2.6. Non-Trivial Factorizations

The message-passing formulation outlined in the previous subsections naturally carries over to non-trivial factorizations; as an illustration, we will shortly discuss the factorization (4) (see [6] [7] [8] [10] [5] for more information). Suppose that we try to find (1) by means of ICM, more precisely, by alternating (low-dimensional) maximizations w.r.t. the individual components Θ_k (cf. (10)). It is easily verified that in order to perform those maximizations, we need the upward sum-product messages (cf. (9))

$$\mu(\theta_k) \triangleq \int_x f_{B_k}(x_{k-1},x_k,\theta_k)\,\vec{\mu}\,(x_{k-1})\,\overleftarrow{\mu}\,(x_k)\,dx_{k-1}\,dx_k. \tag{19}$$

Let us now consider EM; the term $h(\theta)$ in (16) decomposes into the sum

$$h(\theta) = \sum_{k=1}^{n} h_k(\theta_k). \tag{20}$$

The upwards messages $h_k(\theta_k)$ ("E-log messages") [7] [8] are given by:

$$h_k(\theta_k) = \mathrm{E}_{p_{\mathrm{B}}}[\log f_{\mathrm{B}k}(X_{k-1},X_k,\theta_k)]\,, \tag{21}$$

where the marginals p_B are obtained as:

$$p_B(x_{k-1},x_k;\hat{\theta}^{(\ell)}) \propto f_{B_k}(x_{k-1},x_k,\hat{\theta}_k^{(\ell)})\,\vec{\mu}\,(x_{k-1})\,\overleftarrow{\mu}\,(x_k). \tag{22}$$

A similar decomposition occurs in gradient ascent/gradient EM: the term $\nabla_\theta h(\theta)$ in (18) decomposes into the sum

$$\nabla_\theta h_k(\theta) = \sum_{k=1}^{n} \nabla_{\theta_k} h_k(\theta_k). \tag{23}$$

The upwards messages $\nabla_{\theta_k} h_k(\theta_k)\big|_{\hat\theta^{(\ell)}}$ are given by:

$$\begin{aligned}\nabla_{\theta_k} h(\theta_k)\big|_{\hat\theta^{(\ell)}} &= \mathrm{E}_{p_\mathrm{B}}\left[\nabla_{\theta_k} \log f_{\mathrm{B}k}(X_{k-1}, X_k, \theta_k)\big|_{\hat\theta^{(\ell)}}\right] &(24)\\ &= \nabla_{\theta_k} \log \mu(\theta_k)\big|_{\hat\theta^{(\ell)}} &(25)\\ &= \frac{\int_x \nabla_{\theta_k} f_{B_k}(x_{k-1}, x_k, \theta_k)\big|_{\hat\theta^{(\ell)}} \overrightarrow{\mu}(x_{k-1}) \overleftarrow{\mu}(x_k)\, dx}{\int_x f_{B_k}(x_{k-1}, x_k, \hat\theta^{(\ell)}) \overrightarrow{\mu}(x_{k-1}) \overleftarrow{\mu}(x_k)\, dx}. &(26)\end{aligned}$$

(It is straightforward to extend the message computation rules (21) and (24) to general node functions $f_{B_k}(x, \theta_k)$ [7] [8] [5].)

Due to the decomposition (20), the maximization in (16) can be performed by means of local computations, i.e., it can in principle be carried out by max-product message passing [7] [8]. Similarly, as a consequence of (23), the message computation rules (12) and (18) can be computed locally [5].

2.7. Message-Passing Extensions

The message-passing viewpoint suggest a.o. the following extensions of the standard algorithms:

1. We are free to choose the *order* in which message are updated ("message update schedule"). In the standard setting of the four estimation algorithms, all messages are recomputed at each iteration ("standard message update schedule"). Obviously, one may prefer *not* to update some messages during several iterations, especially, if they are computationally complex. For example, in the case of gradient ascent applied to (4) (Section 2.3), one may iterate (11) and (12) while the sum-product messages $\overrightarrow{\mu}(x_{k-1})$ and $\overleftarrow{\mu}(x_k)$ in (26) are kept fixed (see [10, Remark 4.9]).
2. It is straightforward to combine several of the above mentioned message update rules within one algorithm; at each node f_{B_k}, one can in principle choose one the update rules. Moreover, one may also mix several rules at one and the same node f_{B_k}. For example, as an alternative to (21), one may first eliminate the variable X_k (by means of the sum-product rule), and then apply an E-log rule similar to (21) [9, Section 5.6.9] [10, Section 4.9.4]:

   $$\tilde h_k(\theta_k) = \mathrm{E}_{p_\mathrm{B}}[\log g_{B\!k}(X_k, \theta_k)]\,, \tag{27}$$

   where

   $$g_{B\!k}(x_k, \theta_k) \triangleq \int_{x_{k-1}} \overrightarrow{\mu}(x_{k-1}) f_{B\!k}(x_{k-1}, x_k, \theta_k)\, dx_{k-1}, \tag{28}$$

   and

   $$p_{B\!k}(x_k; \hat\theta_k^{(\ell)}) \propto \overrightarrow{\mu}(x_k) \overleftarrow{\mu}(x_k). \tag{29}$$

   Note that the rule (27) is a low-complexity alternative to (21), which motivates our interest in (27); the E-log rule (21) requires the joint marginal in X_{k-1} and X_k, whereas (27) involves the (simpler) marginal $p_{B_k}(x_k; \hat\theta^{(\ell)})$ (see [9, Section 5.6.9] [10, Section 4.9.4] for concrete examples).

3. CONVERGENCE AND STABILITY

First we review some well-known (and perhaps some less familiar) properties of the standard formulation of the four algorithms. Then we consider the message-passing extensions discussed in Section 2.7; we present results for cycle-free (Section 3.2) and cyclic (Section 3.3) subgraphs $f_B(x, \theta)$.

3.1. Basic Properties of the Standard Formulation

1. The fixed points $\hat\theta_f$ of all four algorithms are stationary points ("zero-gradient" points) $\hat\theta_\mathrm{stat}$ of the marginal $f(\theta)$ (2), i.e.,

   $$\nabla_\theta f(\theta)\big|_{\hat\theta_f} = 0. \tag{30}$$

2. The saddle points and local minima of $f(\theta)$ are unstable fixed points, i.e., a slight perturbation will cause $\hat\theta^{(\ell)}$ to diverge from those points. The estimate $\hat\theta^{(\ell)}$ remains in a local minimum if it is accidentally initialized in a local minimum.
3. All four algorithms have the property

   $$f(\hat\theta^{(\ell+1)}) \geq f(\hat\theta^{(\ell)}), \tag{31}$$

   with equality if the algorithm has reached a fixed point. Note that in order to guarantee (31) for the gradient ascent and gradient EM algorithm, the step size λ_k in (12) and (18) needs to be appropriately chosen (e.g., according to the Armijo rule [3]).
4. Property (31) does in general not imply that, for *any* initial $\hat\theta^{(0)}$, the sequence $\hat\theta^{(\ell)}$ converges to a stationary point of $f(\theta)$ ("global convergence"). EM converges globally under some additional (weak) regularity conditions. For the other algorithms, global convergence is guaranteed if, besides some (weak) regularity assumptions, the following conditions are fulfilled.
   ICM: $f_i(\theta_i)$ has a unique (global) maximizer [12].
   Gradient ascent: appropriate λ_k, e.g., Armijo rule [3].
   Generalized EM: $\hat\theta^{(\ell+1)}$ is stationary point of $q(\theta)$ [1], which is achieved in gradient EM, if one iterates (18) (with fixed $q(\theta)$) until convergence; alternative condition: $f(\hat\theta^{(\ell+1)}) > f(\hat\theta^{(\ell)})$ for all $\hat\theta^{(\ell)}$ that are not fixed points.
5. The four algorithms each define a mapping $\hat\theta^{(\ell+1)} \triangleq M(\hat\theta^{(\ell)})$. The Jacobian $J \triangleq \nabla_\theta M(\theta)\big|_{\hat\theta_f}$ determines the convergence properties of the sequence $\hat\theta^{(\ell)}$ in the neighborhood of a fixed point $\hat\theta_f$. The latter is locally stable if all eigenvalues of J are smaller than one; the convergence rate in the neighborhood of $\hat\theta_f$ is given by the largest eigenvalue of J. We list J for the four estimation algorithms at hand:
   ICM: $J \triangleq F \triangleq \Delta_{\theta\theta} \log f(\theta)\big|_{\hat\theta_f}$ and $\Delta_{uv} \triangleq \nabla_u \nabla_v^T$ [12].
   Gradient ascent: $J \triangleq I + \lambda_k F$ [3].
   EM: $J \triangleq I - G^{-1}F$
   with $G \triangleq \int_x f(x, \hat\theta_f) \Delta_{\theta\theta} \log f(x, \theta)\big|_{\hat\theta_f}\, dx$ [1].
   Generalized EM: Same as EM, as long as $\hat\theta^{(\ell+1)}$ is a stationary point of $q(\theta)$ [1].

3.2. Cycle-Free Subgraph $f_B(x, \theta)$

As we pointed out in Section 2, the four iterative estimation algorithms can be viewed as message-passing algorithms with a particular message update schedule, i.e., all messages are updated at

each iteration (see [6] [7] [8] [10] [5]). An interesting question now is: If we *modify* this particular message-passing scheme (as outlined in Section 2.7), how does that affect the convergence and stability properties listed in Section 3.1?

3.2.1. Alternative Update Schedule

It is well known that the choice of the message update schedule does not influence the fixed points of a message-passing algorithm [11]. Depending on the schedule, however, the algorithm may or may not converge, i.e., the update schedule affects global convergence. In case the algorithm converges, scheduling determines to *which* fixed point the algorithm actually converges [11]. It also affects the local stability and convergence rate, since in general, the Jacobian J depends on the schedule; it is straightforward to derive J for a given schedule (starting from the definition of J).

As an illustration, let us consider the update schedule we mentioned in Section 2.7 (Issue 1); suppose that in the update (26), we use the sum-product messages $\overrightarrow{\mu}(x_{k-1};\hat{\theta}^{(\ell')})$ and $\overleftarrow{\mu}(x_k;\hat{\theta}^{(\ell')})$ computed in a previous iteration ℓ' (with $\ell' < \ell$). We will show that this schedule amounts to the same Jacobian J as the standard schedule, i.e., $J = I + \lambda_k F$. Note first of all, that since we do not recompute the sum-product messages at each iteration, we do not longer perform gradient ascent (in θ) on the global function $\log f(\theta)$ but (minus) the Gibbs free energy $\mathcal{G}$ instead [11]:

$$\mathcal{G}(\theta,b) \triangleq -\sum_k \int_x \log f_B(x_{k-1},x_k,\theta_k) b(x_{k-1},x_k;\theta,\theta')\,dx + \sum_k \int_x \log b(x_{k-1},x_k;\theta,\theta') b(x_{k-1},x_k;\theta,\theta')\,dx, \quad (32)$$

where the approximate marginals ("beliefs") b_k are given by

$$b(x_{k-1},x_k;\theta,\theta') \propto f_{B_k}(x_{k-1},x_k,\theta)\,\overrightarrow{\mu}(x_{k-1};\theta')\,\overleftarrow{\mu}(x_k;\theta'), \quad (33)$$

with $\overrightarrow{\mu}$ and $\overleftarrow{\mu}$ sum-product messages obtained for $\Theta = \theta'$. We will use the short-hand notation $b_{\theta\theta'} \triangleq b(\cdot\,;\theta,\theta')$. Note that we have $\mathcal{G}(\theta, b_{\theta\theta}) = -\log f(\theta)$; moreover, since for cycle-free graphs, the fixed-points of the sum-product algorithm are zero-gradient points of the Gibbs free energy [11], it follows

$$\nabla_b\,\mathcal{G}(\theta,b)\big|_{b_{\theta\theta}} = 0 \qquad (\text{for all } \theta). \quad (34)$$

The global gradient ascent update can be written as

$$\hat{\theta}^{(\ell+1)} = \hat{\theta}^{(\ell)} + \lambda_k \nabla_\theta \mathcal{G}(\theta, b_{\hat{\theta}^{(\ell)}\hat{\theta}^{(\ell')}})\big|_{\theta=\hat{\theta}^{(\ell)}}. \quad (35)$$

In order to obtain J, we start from the Taylor expansion:

$$\nabla_\theta \mathcal{G}(\theta,b)\big|_{\theta=\hat{\theta}^{(\ell)}} = \nabla_\theta \mathcal{G}(\theta, b_{\hat{\theta}_f\hat{\theta}_f})\big|_{\hat{\theta}_f} + \Delta_{\theta\theta}\mathcal{G}(\theta, b_{\hat{\theta}_f\hat{\theta}_f})\big|_{\hat{\theta}_f}(\hat{\theta}^{(\ell)} - \hat{\theta}_f) + \Delta_{\theta'\theta}\mathcal{G}(\theta, b_{\hat{\theta}_f\theta'})\big|_{(\hat{\theta}_f,\hat{\theta}_f)}(\hat{\theta}^{(\ell')} - \hat{\theta}_f) + \mathcal{O}(\varepsilon^2), \quad (36)$$

where $\varepsilon \triangleq |\hat{\theta}^{(\ell)} - \hat{\theta}^{(\ell')}|$. We have that:

$$\Delta_{\theta'\theta}\mathcal{G}(\theta, b_{\hat{\theta}_f\theta'})\big|_{(\hat{\theta}_f,\hat{\theta}_f)} = \Delta_{b\theta}\mathcal{G}(\theta,b)\big|_{(\hat{\theta}_f, b_{\hat{\theta}_f\hat{\theta}_f})} \nabla_{\theta'} b_{\hat{\theta}_f\theta'}\big|_{\hat{\theta}_f} \quad (37)$$

and

$$\Delta_{b\theta}\mathcal{G}(\theta,b)\big|_{(\hat{\theta}_f, b_{\hat{\theta}_f\hat{\theta}_f})} = \Delta_{\theta b}\mathcal{G}(\theta,b)\big|_{(\hat{\theta}_f, b_{\hat{\theta}_f\hat{\theta}_f})} \quad (38)$$

$$= \nabla_\theta\big[\nabla_b^T \mathcal{G}(\theta,b)\big|_{b_{\hat{\theta}_f\hat{\theta}_f}}\big]\Big|_{b_{\hat{\theta}_f\hat{\theta}_f}} = 0, \quad (39)$$

where we used (34), and therefore

$$\nabla_\theta \mathcal{G}(\theta,b)\big|_{\theta=\hat{\theta}^{(\ell)}} \approx \Delta_{\theta\theta}\mathcal{G}(\theta, b_{\hat{\theta}_f\hat{\theta}_f})\big|_{\hat{\theta}_f}(\hat{\theta}^{(\ell)} - \hat{\theta}_f), \quad (40)$$

where we also used the fact that $\nabla_\theta \mathcal{G}(\theta, b_{\hat{\theta}_f\hat{\theta}_f})\big|_{\hat{\theta}_f} = 0$; from (40) follows $J = I + \lambda_k F$. Since the non-standard schedule has the same fixed-points and Jacobian as the standard-schedule, both schedules amount to the same local convergence and stability properties. In particular, both algorithms asymptotically converge at the same rate (which is in agreement with the experimental results); note, however, that the non-standard schedule is significantly less complex than the standard schedule, since it avoids updating certain sum-product messages at each iteration. In other words, the convergence rate per flop is usually much larger in the non-standard schedule than in the standard one. We can extend this first-order analysis to study the global convergence of the non-standard schedule. Along the lines of (36)–(40), one can show that

$$f(\hat{\theta}^{(\ell)}) - f(\hat{\theta}^{(\ell')}) = A + \mathcal{O}(\varepsilon^2), \quad (41)$$

with first-order term $A \geq 0$ (and $A = 0$ only if a fixed point has been reached). As a consequence, if the difference $\Delta^{(\ell)} \triangleq |\hat{\theta}^{(\ell)} - \hat{\theta}^{(\ell-1)}|$ between two subsequent estimates $\hat{\theta}^{(\ell)}$ is sufficiently small (in addition to some regularity conditions), the algorithm is guaranteed to converge; this statement can be made more precise by standard arguments, we omit the details here. If one wishes to guarantee global convergence (without any restriction on $\Delta^{(\ell)}$), one may verify after each sequence of updates (35) whether

$$f(\hat{\theta}^{(\ell)}) - f(\hat{\theta}^{(\ell')}) > 0. \quad (42)$$

If that condition is not fulfilled, (i) one omits the estimates $\hat{\theta}^{(\ell'+m)}$ with $m > 1$; (ii) one recomputes the sum-product messages with $\Theta = \hat{\theta}^{(\ell'+1)}$; (iii) one determines the estimate $\theta^{(\ell'+2)}$ by means of the rule (18). If one chooses an appropriate step size λ_k, the condition (42) is in practice usually met; this is due to (41).

The above reasoning can straightforwardly be extended to (i) the other estimation algorithms at hand, i.e., (generalized) EM, ICM, and combinations; (ii) other update schedules.

3.2.2. Combining Update Rules

We pointed out that the message-passing viewpoint enables us to combine various local update rules (in particular, (6)–(26)) within one algorithm; those rules may also be applied at one and the same node (cf. (27)). It can be shown that the fixed points of the resulting message-passing estimation algorithms are stationary points of the marginal $f(\theta)$. We will briefly demonstrate this for the state space model (4); the extension to general models $p(x,y;\theta)$ is straightforward (see [10] for more details). Let us first consider an EM-type algorithm with upward messages $\tilde{h}_k(\theta_k)$ (27) (E-step) and downward message (M-step):

$$\hat{\theta}^{(\ell+1)} = \operatorname*{argmax}_\theta \Big(\sum_{k=1}^n \tilde{h}_k(\theta_k) + \log f_{\mathrm{A}}(\theta)\Big). \quad (43)$$

(In the following, we will write $\tilde{h}_k(\theta_k,\hat{\theta}_k)$ instead of $\tilde{h}_k(\theta_k)$ if we need to make clear that $\tilde{h}_k$ depends on $\hat{\theta}_k$, cf. (27) (29).) The fixed points of this EM-type algorithm fulfill the equality:

$$\sum_{k=1}^n \nabla_{\theta_k} \tilde{h}_k(\theta_k,\hat{\theta}_k^f)\big|_{\hat{\theta}_k^f} + \nabla_\theta \log f_{\mathrm{A}}(\theta)\big|_{\hat{\theta}^f} = 0. \quad (44)$$

Since

$$\nabla_{\theta_k}\tilde{h}_k(\theta_k,\hat{\theta}_k)\big|_{\hat{\theta}_k} = \nabla_{\theta_k}h_k(\theta_k,\hat{\theta}_k)\big|_{\hat{\theta}_k} \quad (45)$$

$$= \nabla_{\theta_k}\log\mu(\theta_k)\big|_{\hat{\theta}_k}, \quad (46)$$

we can rewrite the LHS of (44) as

$$\sum_{k=1}^{n}\nabla_{\theta_k}\log\mu(\theta_k)\big|_{\hat{\theta}_k^f} + \nabla_\theta\log f_{\mathrm{A}}(\theta)\big|_{\hat{\theta}^f} \quad (47)$$

$$= \nabla_\theta\log f_{\mathrm{B}}(\theta)\big|_{\hat{\theta}^f} + \nabla_\theta\log f_{\mathrm{A}}(\theta)\big|_{\hat{\theta}^f} \quad (48)$$

$$= \nabla_\theta\log f(\theta)\big|_{\hat{\theta}^f}, \quad (49)$$

which shows that the fixed points $\hat{\theta}^f$ of the EM-type algorithm (43) are stationary points of $f(\theta)$ (cf. (30)). From (45) (46), it follows (i) that the fixed points of ICM (cf. (19)), the EM algorithm (cf. (21)), and the gradient EM algorithm (cf. (24)) also fulfill (44) and hence (49); (ii) that if one combines all those approaches (by combining the corresponding local update rules), the fixed points of the resulting algorithm fulfill (44).

Also in this setting, it is straightforward to derive an expression for J. As an illustration, we will derive J for the EM-type algorithm with upward messages $\tilde{h}_k(\theta_k)$ (27) (E-step) and downward message (43) (M-step). Consider the first-order Taylor approximation of $\nabla_\theta\log f(\theta)$ around $\Theta = \theta_f$:

$$\nabla_\theta\log f(\theta) \approx \nabla_\theta\log f(\theta)\big|_{\theta_f} + \Delta_{\theta\theta}\log f(\theta)\big|_{\theta_f}(\theta-\theta_f) \quad (50)$$

$$= \Delta_{\theta\theta}\log f(\theta)\big|_{\theta_f}(\theta-\theta_f) \triangleq F\,(\theta-\theta_f), \quad (51)$$

where we used the fact that the fixed points of EM are stationary points of $f(\theta)$. From (43), it follows:

$$\sum_{k=1}^{n}\nabla_{\theta_k}\tilde{h}_k(\theta_k,\hat{\theta}_k^{(\ell)})\big|_{\hat{\theta}_k^{(\ell+1)}} + \nabla_\theta\log f_{\mathrm{A}}(\theta)\big|_{\hat{\theta}^{(\ell+1)}} = 0. \quad (52)$$

Combining the following first-order Taylor approximation of the first term in (52)

$$\nabla_{\theta_k}\tilde{h}_k(\theta_k,\hat{\theta}_k^{(\ell)})\big|_{\hat{\theta}_k^{(\ell+1)}} \approx \nabla_{\theta_k}\tilde{h}_k(\theta_k,\hat{\theta}_k^{(\ell)})\big|_{\hat{\theta}_k^{(\ell)}} + \Delta_{\theta_k\theta_k}\tilde{h}_k(\theta_k,\hat{\theta}_k^{(\ell)})\big|_{\hat{\theta}_k^{(\ell)}}(\hat{\theta}^{(\ell+1)}-\hat{\theta}^{(\ell)}), \quad (53)$$

and a similar expansion for the second term in (52) with (46) and (51) results in $J = I - \tilde{G}^{-1}F$, where

$$\tilde{G} \triangleq \sum_{k=1}^{n}\Delta_{\theta\theta}\tilde{h}_k(\theta_k,\hat{\theta}^f)\big|_{\hat{\theta}^f} + \Delta_{\theta\theta}f_A(\theta)\big|_{\hat{\theta}^f}. \quad (54)$$

As we pointed out earlier, in the standard EM algorithm, $J = I - G^{-1}F$; the matrix G is obtained from (54) by replacing $\tilde{h}_k$ with h_k. The non-standard EM algorithm (43) and the standard EM algorithm both have the same fixed points, but their Jacobian matrix J differs; as a result, their local stability and convergence properties are in principle different. Experimental results show, however, that both algorithms have about the same convergence rate [9]. This may in part be explained by the fact that the matrices G and $\tilde{G}$ become identical as the messages $\vec{\mu}$ and $\overleftarrow{\mu}$ tend towards Dirac deltas (high-SNR limit). The EM-type algorithm (43) is not guaranteed to converge globally, since the property (31) does not hold; in practice, however, it always seems to converge [9].

Although it seems hard to establish global convergence for the example (43), this does not necessarily apply to other message-passing estimation algorithms that combine various update rules; for example, one may easily prove convergence of algorithms that estimate certain variables by means of EM and other variables by gradient ascent or ICM; one needs to combine the convergence conditions for the individual approaches (i.e., EM/gradient ascent/ICM). One can readily obtain the Jacobian J of such algorithms by blending the Jacobians of the individual approaches. Due to space constraints, we omit the details here.

3.3. Cyclic Subgraph $f_B(x,\theta)$

On cyclic graphs, the sum-product algorithm is not guaranteed to converge [4], and this obviously carries over to the four classes of iterative estimation algorithms applied on cyclic subgraphs $f_B(x,\theta)$. Yedidia et al. [11] have shown that the fixed points of the sum-product algorithm applied on a cyclic graph are stationary points of the Bethe free energy of the system at hand (regarded as a functional of the "beliefs"). Following the above line of thought, it is straightforward to show that the fixed points of the four estimation algorithms applied on a cyclic subgraph $f_B(x,\theta)$ are stationary point of the Bethe free energy (regarded as a function in θ) (see [10]). This also holds if one combines several update rules within one algorithm. Note that it is straightforward to compute J also in this setting (following the above example); the Hessian of the Bethe free energy now plays an important role (instead of the Hessian of $\log f(\theta)$, cf. Section 3.1, Property 5).

4. CONCLUSION

We have derived convergence and stability properties of message-passing estimation algorithms operating on general graphs.

In the statistics literature, the asymptotic convergence properties of standard estimation algorithms have intensively been analyzed. It turns out that in many applications, the effective convergence rate and stability of those algorithms is strongly dependent on the Jacobian J; in other words, the asymptotic analysis is of significant practical relevance. In future work, we will further experimentally verify whether this also holds for the presented asymptotic analysis of message-passing estimation algorithms.

5. REFERENCES

[1] 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.

[2] 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.

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

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

[5] 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.

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

[7] 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.

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

[9] S. Korl, *A Factor Graph Approach to Signal Modelling, System Identification, and Filtering*, PhD. Thesis at ETH Zurich, Diss. ETH No 16170, July 2005.

[10] 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`.

[11] 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.

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