Expectation Maximization as Message Passing

Justin Dauwels, Sascha Korl, and Hans-Andrea Loeliger
Dept. of Information Technology and Electrical Engineering, ETH, CH-8092 Zürich, Switzerland.

***Abstract*— Based on prior work by Eckford, it is shown how expectation maximization (EM) may be viewed, and used, as a message passing algorithm in factor graphs.**

I. Introduction

Graphical models [1] such as factor graphs [2] are tools both for system modeling and for the development of algorithms for detection and estimation, cf. [3], [4]. In addition to the basic sum-product and max-product (or min-sum) algorithms, which dominate coding applications, signal processing techniques including LMMSE/Kalman filtering, gradient algorithms, and particle filters can be naturally viewed and used as message passing in factor graphs [3], [4].

Expectation maximization (EM) [5] [6] has also become a standard technique for parameter estimation in graphical models [7] [8]. In particular, Eckford showed how EM can be viewed, and used, as a technique for breaking cycles in factor graphs [9], [10]. However, it is not obvious if and how EM can be described as a *message passing* algorithm with local message update rules.

In the present paper, we develop EM as a message passing technique. The standard "global" view of EM is thus replaced by a "local" message passing view with a new (local) message computation rule for continuous variables. The new message computation rule can often be used in cases where the standard sum-product (integral-product) rule yields impractical expressions for the messages.

II. Review of EM Algorithm

We begin by reviewing the expectation maximization (EM) algorithm in a setting which is suitable for the purpose of this paper. Suppose we wish to find

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

We assume that $f(\theta)$ is the "marginal" of some real-valued function $f(x, \theta)$:

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

where $\int_x g(x)$ denotes either integration or summation of $g(x)$ over the whole range of x. The function $f(x, \theta)$ is assumed to be nonnegative:

$$f(x, \theta) \geq 0 \quad \text{for all } x \text{ and all } \theta. \tag{3}$$

We will also assume that the integral (or the sum) $\int_x f(x, \theta) \log f(x, \theta')$ exists for all θ, θ'. The EM algorithm attempts to compute (1) as follows:

1) Make some initial guess $\hat{\theta}^{(0)}$.

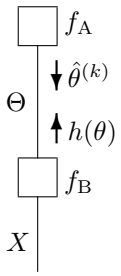

Fig. 1. Factor graph corresponding to (7).

2) Expectation step: evaluate

$$f^{(k)}(\theta) \triangleq \int_x f(x, \hat{\theta}^{(k)}) \log f(x, \theta). \tag{4}$$

3) Maximization step: compute

$$\hat{\theta}^{(k+1)} \triangleq \operatorname*{argmax}_{\theta} f^{(k)}(\theta). \tag{5}$$

4) Repeat 2–3 until convergence or until the available time is over.

The main property of the EM algorithm is

$$f(\hat{\theta}^{(k+1)}) \geq f(\hat{\theta}^{(k)}). \tag{6}$$

For completeness, a proof of (6) is given in the appendix.

III. Message Passing Interpretation

We now rewrite the EM algorithm in message passing form. In this section, we will assume a trivial factorization

$$f(x, \theta) = f_{\mathrm{A}}(\theta) f_{\mathrm{B}}(x, \theta), \tag{7}$$

where $f_{\mathrm{A}}(\theta)$ may be viewed as encoding the a priori information about Θ. More interesting factorizations (i.e., models with internal structure) will be considered in the next section.

We will use Forney-style factor graphs as in [3], where edges represent variables and nodes represent factors. As in [3], we will use capital letters for model variables and small letters for values of such variables. The factor graph of (7) is shown in Fig. 1. In this setup, the EM algorithm amounts to iterative recomputation of the following messages:

Upwards message $h(\theta)$:

$$h(\theta) = \frac{\int_x f_{\mathrm{B}}(x, \hat{\theta}^{(k)}) \log f_{\mathrm{B}}(x, \theta)}{\int_x f_{\mathrm{B}}(x, \hat{\theta}^{(k)})} \tag{8}$$

$$= \mathrm{E}_{p_{\mathrm{B}}} [\log f_{\mathrm{B}}(X, \theta)], \tag{9}$$

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

$$p_\mathrm{B}(x|\hat{\theta}^{(k)}) \triangleq \frac{f_\mathrm{B}(x,\hat{\theta}^{(k)})}{\int_{x'} f_\mathrm{B}(x',\hat{\theta}^{(k)})}. \tag{10}$$

Downwards message $\hat{\theta}^{(k)}$:

$$\hat{\theta}^{(k+1)} = \operatorname*{argmax}_\theta \left(\log f_\mathrm{A}(\theta) + h(\theta)\right) \tag{11}$$

$$= \operatorname*{argmax}_\theta \left(f_\mathrm{A}(\theta)\cdot e^{h(\theta)}\right). \tag{12}$$

The equivalence of this message passing algorithm with (4) and (5) may be seen as follows. From (4) and (5), we have

$$\hat{\theta}^{(k+1)} = \operatorname*{argmax}_\theta \int_x f(x,\hat{\theta}^{(k)}) \log f(x,\theta) \tag{13}$$

$$= \operatorname*{argmax}_\theta \int_x f_\mathrm{A}(\hat{\theta}^{(k)}) f_\mathrm{B}(x,\hat{\theta}^{(k)}) \log\big(f_\mathrm{A}(\theta) f_\mathrm{B}(x,\theta)\big) \tag{14}$$

$$= \operatorname*{argmax}_\theta \int_x f_\mathrm{B}(x,\hat{\theta}^{(k)}) \Big(\log f_\mathrm{A}(\theta) + \log f_\mathrm{B}(x,\theta)\Big) \tag{15}$$

$$= \operatorname*{argmax}_\theta \left(\log f_\mathrm{A}(\theta) + \frac{\int_x f_\mathrm{B}(x,\hat{\theta}^{(k)}) \log f_\mathrm{B}(x,\theta)}{\int_{x'} f_\mathrm{B}(x',\hat{\theta}^{(k)})}\right), \tag{16}$$

which is equivalent to (8) and (11).

Some remarks:

1) The computation (8) or (9) is *not* an instance of the sum-product algorithm.
2) The message $h(\theta)$ may be viewed as a "log-domain" summary of f_B. In (12), the corresponding "probability domain" summary $e^{h(\theta)}$ is consistent with the factor graph interpretation.
3) A constant may be added to (or subtracted from) $h(\theta)$ without affecting (11).
4) If $f_\mathrm{A}(\theta)$ is a constant, the normalization in (8) can be omitted. More generally, the normalization in (8) can be omitted if $f_\mathrm{A}(\theta)$ is constant for all θ such that $f_\mathrm{A}(\theta) \neq 0$. However, in contrast to most standard accounts of the EM algorithm, we explicitly wish to allow more general functions f_A.
5) Nothing changes if we introduce a known observation (i.e., a constant argument) y into f such that (7) becomes $f(x,y,\theta) = f_\mathrm{A}(y,\theta) f_\mathrm{B}(x,y,\theta)$.

IV. Nontrivial Factor Graphs

The algorithm of the previous section still applies if both $\Theta = (\Theta_1,\ldots,\Theta_n)^T$ and $X = (X_0,\ldots,X_n)^T$ are vectors. However, opportunities to simplify the computations may arise if f_A and f_B have "nice" factorizations. For example, assume that f_B factors as

$$f_\mathrm{B}(x,y,\theta) = f_0(x_0) f_1(x_0,x_1,y_1,\theta_1)\cdots f_n(x_{n-1},x_n,y_n,\theta_n), \tag{17}$$

where $y = (y_1,\ldots,y_n)^T$ is some known (observed) vector. Such factorizations arise from classical trellis models and state space models. The factor graph corresponding to (17) is shown in Fig. 2.

The upwards message $h(\theta)$ (9) splits into a sum with one term for each node in the factor graph:

$$h(\theta) = \mathrm{E}\Big[\log\Big(f_0(x_0) f_1(x_0,x_1,y_1,\theta_1)\cdots \cdots f_n(x_{n-1},x_n,y_n,\theta_n)\Big)\Big] \tag{18}$$

$$= \mathrm{E}[\log f_0(X_0)] + \mathrm{E}[\log f_1(X_0,X_1,y_1,\theta_1)] + \ldots \ldots + \mathrm{E}[\log f_n(X_{n-1},X_n,y_n,\theta_n)] \tag{19}$$

Each term

$$h_k(\theta_k) \triangleq \mathrm{E}[\log f_k(X_{k-1},X_k,y_k,\theta_k)] \tag{20}$$

may be viewed as the message out of the corresponding node, as indicated in Fig. 2. The constant term $\mathrm{E}[\log f_0(X_0)]$ in (19) may be omitted (cf. Remark 3 in Section III). As in (9), all expectations are with respect to the probability distribution p_B, which we here denote by $p_\mathrm{B}(x|y,\hat{\theta})$. Note that each term (20) requires only $p_\mathrm{B}(x_{k-1},x_k|y,\hat{\theta})$, the joint distribution of X_{k-1} and X_k:

$$h_k(\theta_k) = \int_{x_{k-1}} \int_{x_k} p_\mathrm{B}(x_{k-1},x_k|y,\hat{\theta}) \log f_k(x_{k-1},x_k,y_k,\theta_k). \tag{21}$$

These joint distributions may be obtained by means of the standard sum-product algorithm (belief propagation) [2] [3]: from elementary factor graph theory, we have

$$p_\mathrm{B}(x_{k-1},x_k|y,\hat{\theta}) \propto f_k(x_{k-1},x_k,y_k,\hat{\theta})\, \mu_{X_{k-1}\to f_k}(x_{k-1}) \cdot \mu_{X_k\to f_k}(x_k), \tag{22}$$

where $\mu_{X_{k-1}\to f_k}$ and $\mu_{X_k\to f_k}$ are the messages of the sum-product algorithm towards the node f_k and where "$\propto$" denotes equality up to a scale factor that does not depend on x_{k-1}, x_k. It follows that

$$p_\mathrm{B}(x_{k-1},x_k|y,\hat{\theta}) = \frac{f_k(x_{k-1},x_k,y_k,\hat{\theta})\,\mu_{X_{k-1}\to f_k}(x_{k-1})\,\mu_{X_k\to f_k}(x_k)}{\int_{x_{k-1}}\int_{x_k} f_k(x_{k-1},x_k,y,\hat{\theta})\,\mu_{X_{k-1}\to f_k}(x_{k-1})\,\mu_{X_k\to f_k}(x_k)}. \tag{23}$$

Note that, if the sum product messages $\mu_{X_{k-1}\to f_k}$ and $\mu_{X_k\to f_k}$ are computed without any scaling, then the denominator in (23) equals $p_\mathrm{B}(y|\hat{\theta})$, which is independent of k.

The downwards message $\hat{\theta}$ (11) is

$$(\hat{\theta}_1,\ldots,\hat{\theta}_n)^T = \operatorname*{argmax}_{\theta_1,\ldots,\theta_n} \big(\log f_\mathrm{A}(\theta) + h_1(\theta_1) + \ldots \ldots + h_n(\theta_n)\big) \tag{24}$$

$$= \operatorname*{argmax}_{\theta_1,\ldots,\theta_n} \Big(f_\mathrm{A}(\theta)\cdot e^{h_1(\theta_1)}\cdots e^{h_n(\theta_n)}\Big). \tag{25}$$

If f_A has itself a nice factorization, then (24) or (25) may be computed by the standard max-sum or max-product algorithm, respectively. This applies, in particular, for the standard case $\Theta_1 = \Theta_2 = \ldots = \Theta_n$, which is illustrated in Fig. 3.

The above derivations do not in any essential way depend on the specific example (17). In principle, any cut-set of edges in

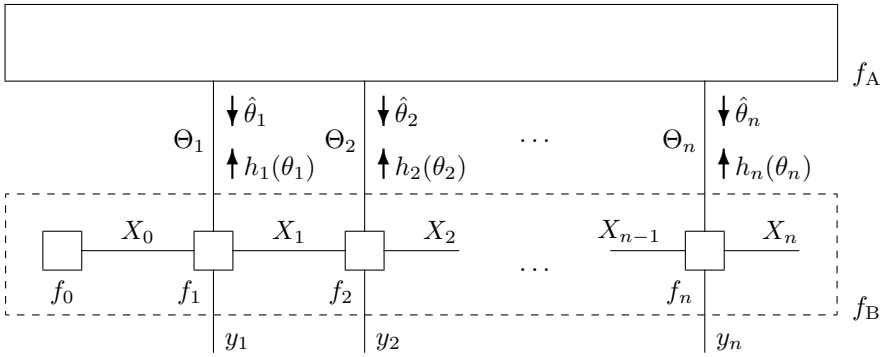

Fig. 2. Factor graph corresponding to (17).

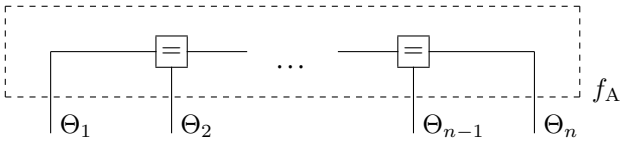

Fig. 3. Factor graph of $\Theta_1 = \Theta_2 = \ldots = \Theta_n$.

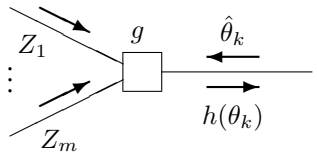

Fig. 4. h-message out of a generic node.

some factor graph may be chosen to be the vector Θ. However, the resulting subgraphs corresponding to f_{A} and f_{B} should be cycle-free in order to permit the computation of exact expectations (h-messages) and maximizations ($\hat{\theta}$-messages). The h-message out of a generic node $g(z_1, \ldots, z_m, \theta_k)$ (cf. Fig. 4) is

$$h(\theta_k) = \gamma^{-1} \int_{z_1} \ldots \int_{z_m} g(z_1, \ldots, z_m, \hat{\theta}_k)\, \mu(z_1) \cdots \mu(z_m) \cdot \log g(z_1, \ldots, z_m, \theta_k) \tag{26}$$

with

$$\gamma \triangleq \int_{z_1} \ldots \int_{z_m} g(z_1, \ldots, z_m, \hat{\theta}_k)\, \mu(z_1) \cdots \mu(z_m) \tag{27}$$

and where $\mu(z_1), \ldots, \mu(z_m)$ are the standard sum-product messages. Obviously, this message passing rule may also be applied to a (sub-) graph with cycles, but then there is no guarantee for (6).

V. Conclusion

Elaborating on prior work by Eckford, we have formulated EM in message passing form with a new message computation rule (26). In this setting, a main attraction of EM is that this message passing rule can be evaluated in some cases where the standard sum-product or max-product rules yield intractable expressions. It is likely that "local" use of this message computation rule can give good results even in situations where the "global" conditions required to guarantee (6) are not satisfied.

Appendix: Proof of (6)

The proof is standard (cf. [6]), but adapted to the slightly nonstandard setup of Section II.

Lemma: The function

$$\tilde{f}(\theta, \theta') \triangleq f(\theta') + \int_x f(x, \theta') \log \frac{f(x, \theta)}{f(x, \theta')} \tag{28}$$

satisfies both

$$\tilde{f}(\theta, \theta') \le f(\theta) \tag{29}$$

and

$$\tilde{f}(\theta, \theta) = f(\theta). \tag{30}$$

□

Proof: The equality (30) is obvious. The inequality (29) follows from eliminating the logarithm in (28) by the inequality $\log x \le x - 1$ for $x > 0$:

$$\tilde{f}(\theta, \theta') \le f(\theta') + \int_x f(x, \theta') \left(\frac{f(x, \theta)}{f(x, \theta')} - 1 \right) \tag{31}$$

$$= f(\theta') + \int_x f(x, \theta) - \int_x f(x, \theta') \tag{32}$$

$$= f(\theta). \tag{33}$$

■

To prove (6), we first note that (5) is equivalent to

$$\hat{\theta}^{(k+1)} = \operatorname*{argmax}_{\theta} \tilde{f}(\theta, \hat{\theta}^{(k)}). \tag{34}$$

We then obtain

$$f(\hat{\theta}^{(k)}) = \tilde{f}(\hat{\theta}^{(k)}, \hat{\theta}^{(k)}) \tag{35}$$

$$\le \tilde{f}(\hat{\theta}^{(k+1)}, \hat{\theta}^{(k)}) \tag{36}$$

$$\le f(\hat{\theta}^{(k+1)}), \tag{37}$$

where (35) follows from (30), (36) follows from (34), and (37) follows from (29).

REFERENCES

[1] M. I. Jordan and T.J. Sejnowski, eds., *Graphical Models: Foundations of Neural Computation.* MIT Press, 2001.

[2] 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.

[3] H.-A. Loeliger, "An introduction to factor graphs," *IEEE Signal Proc. Mag.*, Jan. 2004, pp. 28–41.

[4] H.-A. Loeliger, J. Dauwels, V. M. Koch, and S. Korl, "Signal processing with factor graphs: examples," *Proc. First Int. Symp. on Control, Communications and Signal Processing*, March 21–24, Hammamet, Tunisia, pp. 571–574.

[5] 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*, vol. 39, Series B, pp. 1–38, 1977.

[6] 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.

[7] S. Lauritzen, "The EM algorithm for graphical association models with missing data," *Computational Statistics and Data Analysis*, vol. 19, pp. 191–201, 1995.

[8] Z. Ghahramani, "Unsupervised Learning," in *Advanced Lectures on Machine Learning*, Bousquet et al., Eds., Springer Verlag 2004.

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

[10] A. W. Eckford, "Channel estimation in block fading channels using the factor graph EM algorithm," *22nd Biennial Symposium on Communications*, Kingston, Ontario, Canada, May 31 – June 3, 2004. http://www.comm.toronto.edu/ eckford/pubs/index.html