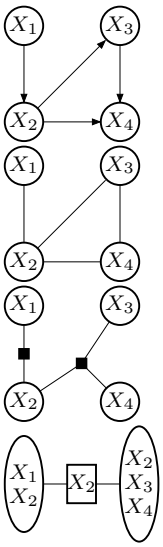
Lecture Notes: Factor graphs and belief propagation

Marc Toussaint
Machine Learning & Robotics group, TU Berlin
Franklinstr. 28/29, FR 6-9, 10587 Berlin, Germany

March 14, 2008

A graphical model is nothing but an illustration of an equation. This equation determins how a functional over multiple variables factorizes. We are mostly interested in functionals that represent a multivariate probability distribution – but graphical models and the related methods could be applied on any functional. Here are some basic examples for equations concerning the factorization of a functional P over variables $X_{1:4}$ and their graphical notation:



$$\Leftrightarrow \quad P(\boldsymbol{X}) = P(X_1)\, P(X_2 \,|\, X_1)\, P(X_3 \,|\, X_2)\, P(X_4 \,|\, X_3, X_2)$$
$$P(\boldsymbol{X}) = \textstyle\prod_i P(X_i \,|\, X_{\text{parents}(i)})$$

$$\Leftrightarrow \quad P(\boldsymbol{X}) = \phi(X_1, X_2)\, \phi(X_2, X_3)\, \phi(X_3, X_4)\, \phi(X_2, X_4)$$
$$P(\boldsymbol{X}) = \textstyle\prod_i \phi_i(X_i) \prod_{(ij)} \phi_{(ij)}(X_i, X_j)$$

$$\Leftrightarrow \quad P(\boldsymbol{X}) = f_{12}(X_1, X_2)\, f_{234}(X_2, X_3, X_4)$$
$$P(\boldsymbol{X}) = \textstyle\prod_C f_C(X_C)$$

$$\Leftrightarrow \quad P(\boldsymbol{X}) = \frac{\psi_{12}(X_1,X_2)\, \psi_{234}(X_2,X_3,X_4)}{\phi_2(X_2)}$$
$$P(\boldsymbol{X}) = \frac{\prod_C \psi_C(X_C)}{\prod_S \phi_S(X_S)}$$

The first example is called a (directed) graphical model or Bayes Net, the second is a pair-wise undirected graphical model, the third a factor graph, the last a junction tree (with clique potentials ψ_C and separator potentials ϕ_S).

Factor graphs (Kschischang, Frey, & Loeliger 2001) are structured probability distributions of the form

$$P(X_{1:n}) = \prod_C f_C(X_C)\,, \tag{1}$$

where $C \subseteq \{X_1, .., X_n\}$ indexes different cliques (subsets) of variables. Pictorially, such a structural property of the probability distribution can be captured in a (bi-partite) graph – as shown in Figure 1 – where the random variables X_i are of one type of nodes (circles), and the factors f_C are of another type (black boxes). We use the notation $\nu(i) := \{C : X_i \in C\}$ to denote all cliques which contain the random variable X_i.

Clearly, Bayesian Networks (which are distributions structured in the form $P(X_{1:n}) = \prod_i P(X_i \,|\, X_{\text{parents}(i)})$) and also Dynamic Bayesian Networks as well as undirected graphical models (e.g., Markov random fields) fall into this category.

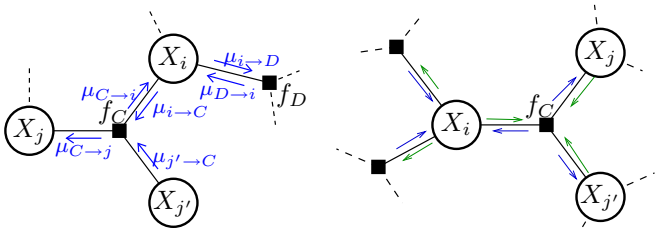


Figure 1: Part of a factor graph. Circle nodes represent random variables X_i, black boxes represent factor functionals $f_i = f_i(X_{C_i})$ where $C_i \subseteq \{X_1, .., X_n\}$ are subsets of variables, the graph as a whole represents the structure of the joint probability distribution $P(X_{1:n}) = \prod_i f_i(X_{C_i})$. Some messages illustrate the defining equations (4) and (6) of belief propagation.

Belief propagation passes messages from cliques to variables and from variables to cliques, as we define below. The point is that these messages multiply into the *belief* at each variable and each clique, such that in total the belief is always the product of the initial (prior) belief and all incoming messages

$$b_C(X_C) := f_C(X_C) \prod_{i \in C} \mu_{i\to C}(X_i)\,, \tag{2}$$

$$b_i(X_i) := \prod_{C \in \nu(i)} \mu_{C\to i}(X_i)\,. \tag{3}$$

The initial belief over a clique is just f_C and the initial belief over a variable is $b_i = 1$. As for the intuition, one should think of the belief as all the "information" currently available about a variable or clique, and this information is the product of the initial information (prior belief) and the information passed on from the messages. The way this information is fused is analogous to Bayes rule: one can think of the messages as likelihoods multiplying into a prior to get the posterior.

By definition, the clique-to-variable messages and

variable-to-clique messages are computed as follows

$$\mu_{C\to i}(X_i) = \frac{1}{\mu_{i\to C}(X_i)} \sum_{X_C \setminus X_i} b_C(X_C) \quad (4)$$

$$= \sum_{X_C \setminus X_i} f_C(X_C) \prod_{j\in C, j\neq i} \mu_{j\to C}(X_j)\ , \quad (5)$$

$$\mu_{i\to C}(X_i) = \frac{1}{\mu_{C\to i}(X_i)}\, b_i(X_i) \quad (6)$$

$$= \prod_{D\in\nu(i), D\neq C} \mu_{D\to i}(X_i)\ . \quad (7)$$

One can think of belief propagation as a method that tries to establish "consistence" between shared beliefs. *Consistence* means that we want to propagate messages until all the beliefs agree in terms of their marginals, i.e., for all C and $i \in C$:

$$b_C(X_i) = b_i(X_i)\ . \quad (8)$$

where $b_C(X_i)$ is the marginal of b_C over the variables X_i. However, we also want that these beliefs still represent our initial joint probability distribution $P(X_{1:n})$, which can be expressed as the following constraint

$$P(X_{1:n}) = \prod_C f_C(X_C) = \frac{\prod_C b(X_C)}{\prod_i b(X_i)^{|\nu(i)|-1}}\ . \quad (9)$$

Roughly speaking, the terms we divide by (which are related to "separators") ensure that the marginals over single variables X_i only appear once in the whole product (see (Yedidia, Freeman, & Weiss 2001)). These two goals, establishing consistency but faithfully representing the joint distribution motivate the message passing equations. We can observe that

1. when all messages are initialized with $\mu = 1$ then the faithfulness constraint (9) is initially fulfilled (using (2) in (9));
2. the message updates (4) and (6) will leave the faithfulness constraint (9) intact;
3. assuming we have reached the consistency (8), the message updates leave the messages unchanged, i.e., consistency is a fixed point of the message updates.

These three observations suggest (but do not prove!) that the message updates are a means to converge towards consistency while maintaining the faithful representation of the joint distribution. Beyond this insight, one can also prove that BP will compute beliefs $b_C(X_C)$ equal to the correct posterior marginal $\sum_{X_j: j\notin C} P(X_{1:n})$ on a tree structured factor graph fulfilling the running intersection property – simply by direct comparison to the elimination algorithm. For loopy graphs there is no unique order to resolve these recursive update equations but BP can still be applied iteratively – and in practice often is – but convergence is not guaranteed. If it converges then to a certain (Bethe) approximation of the true marginals (Yedidia, Freeman, & Weiss 2001). Further discussion of BP is beyond the scope of this paper, please refer to (Yedidia, Freeman, & Weiss 2001; Murphy 2002; Minka 2001) for more details.

The equations for the BP message updates can be simplified when each clique is only a pair-wise factor, i.e., depends only on two variables, $C = \{X_i, X_j\}$. In this case we can define variable-to-variable messages $\mu_{j\to i}(X_i) := \mu_{C\to i}(X_i)$ where $C = \{X_i, X_j\}$ is unique. Equations (4 & 6) simplify to

$$\mu_{j\to i}(X_i) = \sum_{X_j} f_C(X_i, X_j) \frac{b_j(X_j)}{\mu_{i\to j}(X_j)} \quad (10)$$

$$= \sum_{X_j} f_C(X_i, X_j) \prod_{k:k\neq i} \mu_{k\to j}(X_j)\ , \quad (11)$$

which is the standard message passing, for instance, on Markov random fields. Similarly, the equations can be simplified when each single variable X_i (or "separator") is contained in only two neighboring cliques, $|\nu(i)| = 2$. In the case we can define clique-to-clique messages $\mu_{D\to C}(X_i) := \mu_{i\to C}(X_i)$ where $i = C\cap D$ is unique. Equations (4 & 6) simplify to

$$\mu_{D\to C}(X_i) = \frac{1}{\mu_{C\to D}(X_i)} \sum_{X_D \setminus X_i} b_D(X_D) \quad (12)$$

$$= \sum_{X_D \setminus X_i} f_D(X_D) \prod_{E:E\neq C} \mu_{E\to D}(X_{E\cap D})\ , \quad (13)$$

which is the standard message passing on junction trees (where X_i is a separator fulfilling the running intersection property).

References

Kschischang, Frey, & Loeliger (2001). Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory* **47**.

Minka, T. (2001). A family of algorithms for approximate bayesian inference. PhD thesis, MIT.

Murphy, K. (2002). Dynamic bayesian networks: Representation, inference and learning. PhD Thesis, UC Berkeley, Computer Science Division.

Yedidia, J., W. Freeman, & Y. Weiss (2001). Understanding belief propagation and its generalizations.