$$\nu_{i\to j}^{(t+1)}(x_i) \cong \prod_{l\in\partial i\setminus j} \sum_{x_l} \psi_{il}(x_i,x_l)\, \nu_{l\to i}^{(t)}(x_l)\,. \tag{14.31}$$

Simplified expressions can be derived in this case for the joint distribution of several variables (see eqn (14.18)), as well as for the free entropy.

Exercise 14.7 Show that, for pairwise models, the free entropy given in eqn (14.27) can be written as $\mathbb{F}_*(\underline{\nu}) = \sum_{i\in V}\mathbb{F}_i(\underline{\nu}) - \sum_{(ij)\in E}\mathbb{F}_{(ij)}(\underline{\nu})$, where

$$\mathbb{F}_i(\underline{\nu}) = \log\left[\sum_{x_i}\prod_{j\in\partial i}\left(\sum_{x_j}\psi_{ij}(x_i,x_j)\nu_{j\to i}(x_j)\right)\right]\,,$$
$$\mathbb{F}_{(ij)}(\underline{\nu}) = \log\left[\sum_{x_i,x_j}\nu_{i\to j}(x_i)\psi_{ij}(x_i,x_j)\nu_{j\to i}(x_j)\right]\,. \tag{14.32}$$

14.3 Optimization: Max-product and min-sum

Message-passing algorithms are not limited to computing marginals. Imagine that you are given a probability distribution $\mu(\,\cdot\,)$ as in eqn (14.13), and you are asked to find a configuration $\underline{x}$ which maximizes the probability $\mu(\underline{x})$. Such a configuration is called a **mode** of $\mu(\,\cdot\,)$. This task is important in many applications, ranging from MAP estimation (e.g. in image reconstruction) to word MAP decoding.

It is not hard to devise a message-passing algorithm adapted to this task, which correctly solves the problem on trees.

14.3.1 Max-marginals

The role of marginal probabilities is played here by the **max-marginals**

$$M_i(x_i^*) = \max_{\underline{x}}\{\mu(\underline{x}) : x_i = x_i^*\}\,. \tag{14.33}$$

In the same way as the tasks of sampling and of computing partition functions can be reduced to computing marginals, optimization can be reduced to computing max-marginals. In other words, given a black box that computes max-marginals, optimization can be performed efficiently.

Consider first the simpler case in which the max-marginals are non-degenerate, i.e., for each $i \in V$, there exists an x_i^* such that $M_i(x_i^*) > M_i(x_i)$ (strictly) for any $x_i \neq x_i^*$. The unique maximizing configuration is then given by $\underline{x}^* = (x_1^*,\dots,x_N^*)$.

In the general case, the following 'decimation' procedure, which is closely related to the BP-guided sampling algorithm of Section 14.2.4, returns one of the maximizing configurations. Choose an ordering of the variables, say $(1,\dots,N)$. Compute $M_1(x_1)$, and let x_1^* be one of the values maximizing it: $x^* \in \arg\max M_1(x_1)$. Fix x_1 to take this

value, i.e. modify the graphical model by introducing the factor $\mathbb{I}(x_1 = x_1^*)$ (this corresponds to considering the conditional distribution $\mu(\underline{x}|x_1 = x_1^*)$). Compute $M_2(x_2)$ for the new model, fix x_2 to one value $x_2^* \in \arg\max M_2(x_2)$, and iterate this procedure, fixing all the x_i's sequentially.

14.3.2 Message passing

It is clear from the above that max-marginals need only to be computed up to a multiplicative normalization. We shall therefore stick to our convention of denoting equality between max-marginals up to an overall normalization by $\cong$. Adapting the message-passing update rules to the computation of max-marginals is not hard: it is sufficient to replace sums with maximizations. This yields the following **max-product** update rules:

$$\nu_{i\to a}^{(t+1)}(x_i) \cong \prod_{b\in\partial i\setminus a} \widehat{\nu}_{b\to i}^{(t)}(x_i)\,, \tag{14.34}$$

$$\widehat{\nu}_{a\to i}^{(t)}(x_i) \cong \max_{\underline{x}_{\partial a\setminus i}} \left\{ \psi_a(\underline{x}_{\partial a}) \prod_{j\in\partial a\setminus i} \nu_{j\to a}^{(t)}(x_j) \right\}\,. \tag{14.35}$$

The fixed-point conditions for this recursion are called the **max-product equations**. As in BP, it is understood that, when $\partial j \setminus a$ is an empty set, $\nu_{j\to a}(x_j) \cong 1$ is the uniform distribution. Similarly, if $\partial a \setminus j$ is empty, then $\widehat{\nu}_{a\to j}(x_j) \cong \psi_a(x_j)$. After any number of iterations, an estimate of the max-marginals is obtained as follows:

$$\nu_i^{(t)}(x_i) \cong \prod_{a\in\partial i} \widehat{\nu}_{a\to i}^{(t-1)}(x_i)\,. \tag{14.36}$$

As in the case of BP, the main motivation for the above updates comes from the analysis of graphical models on trees.

Theorem 14.4. (the max-product algorithm is exact on trees) *Consider a tree-graphical model with diameter t_*. Then:*

1. *Irrespective of the initialization, the max-product updates (14.34) and (14.35) converge after at most t_* iterations. In other words, for any edge (i, a) and any $t > t_*$, $\nu_{i\to a}^{(t)} = \nu_{i\to a}^*$ and $\widehat{\nu}_{a\to i}^{(t)} = \widehat{\nu}_{a\to i}^*$.*
2. *The max-marginals are estimated correctly, i.e., for any variable node i and any $t > t_*$, $\nu_i^{(t)}(x_i) = M_i(x_i)$.*

The proof follows closely that of Theorem 14.1, and is left as an exercise for the reader.

Exercise 14.8 The crucial property used in both Theorem 14.1 and Theorem 14.4 is the distributive property of the sum and the maximum with respect to the product. Consider, for instance, a function of the form $f(x_1, x_2, x_3) = \psi_1(x_1, x_2)\psi_2(x_1, x_3)$. Then one can decompose the sum and maximum as follows:

$$\sum_{x_1,x_2,x_3} f(x_1, x_2, x_3) = \sum_{x_1} \left[\left(\sum_{x_2} \psi_1(x_1, x_2) \right) \left(\sum_{x_3} \psi_2(x_1, x_3) \right) \right] , \tag{14.37}$$

$$\max_{x_1,x_2,x_3} f(x_1, x_2, x_3) = \max_{x_1} \left[\left(\max_{x_2} \psi_1(x_1, x_2) \right) \left(\max_{x_3} \psi_2(x_1, x_3) \right) \right] . \tag{14.38}$$

Formulate a general 'marginalization' problem (with the ordinary sum and product substituted by general operations with a distributive property) and describe a message-passing algorithm that solves it on trees.

The max-product messages $\nu_{i\to a}^{(t)}(\,\cdot\,)$ and $\widehat{\nu}_{a\to i}^{(t)}(\,\cdot\,)$ admit an interpretation which is analogous to that of sum–product messages. For instance, $\nu_{i\to a}^{(t)}(\,\cdot\,)$ is an estimate of the max-marginal of variable x_i with respect to the modified graphical model in which factor node a is removed from the graph. Along with the proof of Theorem 14.4, it is easy to show that, in a tree-graphical model, fixed-point messages do indeed coincide with the max-marginals of such modified graphical models.

The problem of finding the mode of a distribution that factorizes as in eqn (14.13) has an alternative formulation, namely as minimizing a cost (energy) function that can be written as a sum of local terms:

$$E(\underline{x}) = \sum_{a\in F} E_a(\underline{x}_{\partial a}) \, . \tag{14.39}$$

The problems are mapped onto each other by writing $\psi_a(\underline{x}_{\partial a}) = \mathrm{e}^{-\beta E_a(\underline{x}_{\partial a})}$ (with β some positive constant). A set of message-passing rules that is better adapted to the latter formulation is obtained by taking the logarithm of eqns (14.34) and (14.35). This version of the algorithm is known as the **min-sum** algorithm:

$$E_{i\to a}^{(t+1)}(x_i) = \sum_{b\in\partial i\setminus a} \widehat{E}_{b\to i}^{(t)}(x_i) + C_{i\to a}^{(t)} \, , \tag{14.40}$$

$$\widehat{E}_{a\to i}^{(t)}(x_i) = \min_{\underline{x}_{\partial a\setminus i}} \left[E_a(\underline{x}_{\partial a}) + \sum_{j\in\partial a\setminus i} E_{j\to a}^{(t)}(x_j) \right] + \widehat{C}_{a\to i}^{(t)} \, . \tag{14.41}$$

The corresponding fixed-point equations are also known in statistical physics as the **energetic cavity equations**. Notice that, since the max-product marginals are relevant only up to a multiplicative constant, the min-sum messages are defined up to an overall additive constant. In the following, we shall choose the constants $C_{i\to a}^{(t)}$ and $\widehat{C}_{a\to i}^{(t)}$ such that $\min_{x_i} E_{i\to a}^{(t+1)}(x_i) = 0$ and $\min_{x_i} \widehat{E}_{a\to i}^{(t)}(x_i) = 0$, respectively. The

analogue of the max-marginal estimate in eqn (14.36) is provided by the following log-max-marginal:

$$E_i^{(t)}(x_i) = \sum_{a \in \partial i} \widehat{E}_{a \to i}^{(t-1)}(x_i) + C_i^{(t)} \, . \tag{14.42}$$

In the case of tree-graphical models, the minimum energy $U_* = \min_{\underline{x}} E(\underline{x})$ can be immediately written in terms of the fixed-point messages $\{E^*_{i \to a}, \widehat{E}^*_{i \to a}\}$. We obtain, in fact,

$$U_* = \sum_a E_a(\underline{x}^*_{\partial a}) \, , \tag{14.43}$$

$$\underline{x}^*_{\partial a} = \arg\min_{\underline{x}_{\partial a}} \left\{ E_a(\underline{x}_{\partial a}) + \sum_{i \in \partial a} \widehat{E}^*_{i \to a}(x_i) \right\} . \tag{14.44}$$

In the case of non-tree graphs, this can be taken as a prescription to obtain a max-product estimate $U_*^{(t)}$ of the minimum energy. One just needs to replace the fixed-point messages in eqn (14.44) with the messages obtained after t iterations. Finally, a minimizing configuration $\underline{x}^*$ can be obtained through the decimation procedure described in the previous subsection.

Exercise 14.9 Show that U_* is also given by $U_* = \sum_{a \in F} \epsilon_a + \sum_{i \in V} \epsilon_i - \sum_{(ia) \in E} \epsilon_{ia}$, where

$$\epsilon_a = \min_{\underline{x}_{\partial a}} \left[E_a(\underline{x}_{\partial a}) + \sum_{j \in \partial a} E^*_{j \to a}(x_j) \right] , \qquad \epsilon_i = \min_{x_i} \left[\sum_{a \in \partial i} \widehat{E}^*_{a \to i}(x_i) \right] ,$$
$$\epsilon_{ia} = \min_{x_i} \left[E^*_{i \to a}(x_i) + \widehat{E}^*_{a \to i}(x_i) \right] . \tag{14.45}$$

[Hints: (*i*) Define $x_i^*(a) = \arg\min \left[\widehat{E}^*_{a \to i}(x_i) + E^*_{i \to a}(x_i) \right]$, and show that the minima in eqn (14.45) are achieved at $x_i = x_i^*(a)$ (for ϵ_i and ϵ_{ai}) and at $\underline{x}^*_{\partial a} = \{x_i^*(a)\}_{i \in \partial a}$ (for ϵ_a). (*ii*) Show that $\sum_{(ia)} \widehat{E}^*_{a \to i}(x_i^*(a)) = \sum_i \epsilon_i$.]

14.3.3 Warning propagation

A frequently encountered case is that of constraint satisfaction problems, where the energy function just counts the number of violated constraints:

$$E_a(\underline{x}_{\partial a}) = \begin{cases} 0 & \text{if constraint } a \text{ is satisfied,} \\ 1 & \text{otherwise.} \end{cases} \tag{14.46}$$

The structure of messages can be simplified considerably in this case. More precisely, if the messages are initialized in such a way that $\widehat{E}^{(0)}_{a \to i} \in \{0, 1\}$, this condition is preserved by the min-sum updates (14.40) and (14.41) at any subsequent time. Let us