Lecture Notes for Stat 375
Inference in Graphical Models

Andrea Montanari[1]

February 21, 2011

[1]Department of Electrical Engineering and Department of Statistics, Stanford University

Contents

Chapter 1

Probability and graphs

Graphical models are probability distributions that 'factor' according to a graph structure. The specific class of graph structures used and the precise meaning of 'factor' depend on the type of graphical model under consideration. Typically, factorization according to a graph encodes a specific class of conditional independence properties.

There are two fundamental reasons that make graphical models interesting:

1. The class of probability distributions that factors accorting to a suitably sparse graph is a low-dimensional subclass of the set of all probability distribution over a given domain. This enables concise representation, and efficien learnability.

2. Sparse graph structures often correspond to weak dependencies, or to highly structured dependencies in the corresponding distributions. This leads to efficient algorithms for statistical inference.

Specific families of graphical models include Bayesian networks, factor graphs, Markov random fields. These allow to encode in various ways indepndency statements, and the choice of the most suitable formalism can be important in applications. On the other hand, they are loosely 'reducible' to each other. By this we mean that a distribution that factors accordingly to a sparse graph structure in one formalism, also factors in other formalism according to a graph with the same sparsity.

This chapter provides a synthetic overview of the various formalisms. In the rest of the course, we will focus on factor graphs but it is important to know how to pass from one formalism to another.

1.1 Bayesian networks

A **Bayesian network** describes the joint distributions of variables associated to the vertices of a directed acyclic graph $G = (V, E)$, A directed graph is an ordinary graph with a direction (i.e. an ordering of the adjacent vertices) chosen on each of its edges. The graph is acyclic if it has no directed cycle. In such a graph, we say that a vertex $u \in V$ is a **parent** of v, and write $u \in \pi(v)$, if (u, v) is a (directed) edge of G. A random variable X_v is associated with each vertex v of the graph (for simplicity we assume all the variables to take values in the same finite set $\mathcal{X}$).

In a Bayesian network, the joint distribution of $\{X_v,\ v \in V\}$ is completely specified by the conditional probability kernels $\{p_v(x_v|\underline{x}_{\pi(v)})\}_{v\in V}$, where $\pi(v)$ denotes the set of parents of vertex v, and $\underline{x}_{\pi(v)} = \{x_u : u \in \pi(v)\}$. We also denote by $\pi(G)$ the set of vertices that have no parent in G.

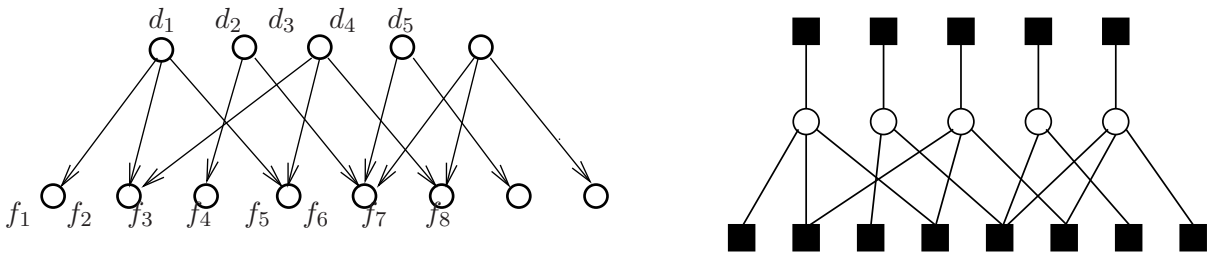


Figure 1.1: Left: toy example of QMR-DT Bayesian network. Right: factor graph representation of the conditional distribution of the diseases $d_1, \dots d_5$, given the findings $f_1, \dots f_8$.

Given such a collection of conditional probabilities indexed by the vertices of G, we can construct in a unique way the joint distribution of all the variables μ on $\mathcal{X}^V$.

This is done according to the following definition.

Definition 1. *Given a directed acyclic graph $G = (V, E)$, and a probability distribution μ over $\mathcal{X}^V$, we say that μ* factors according to the Bayes network structure G *(for short* factors on G*) if it can be written as*

$$\mu(\underline{x}) = \prod_{v \in \pi(G)} p_v(x_v) \prod_{v \in G \setminus \pi(G)} p_v(x_v | \underline{x}_{\pi(v)}) \,, \tag{1.1.1}$$

for a set of conditional probability kernels $\underline{p} \equiv \{p_v\}$.

If μ factors according to G, we say that the pair (G, μ) is a Bayesian network. Equivalently, we will say that the triple $(G, \underline{p}, \mathcal{X})$ is a Bayesian network.

Here is the conditional independence property encoded by G.

Proposition 1.1.1. *The probability distribution μ over $\mathcal{X}^V$ factors according to G if and only if, for each $v \in V$, and any $S \subseteq V$ such that $S \cap$ descendants$(v) = \emptyset$, X_v is conditionally independent of X_S, given $X_{\pi(v)}$.*

Here is an example showing the practical utility of Bayesian networks.

Example: The Quick Medical Reference–Decision Theoretic (QMR-DT) network is a two level Bayesian network developed for automatic medical diagnostic. A schematic example is shown in Fig. 1.1. Variables in the top level, denoted by $d_1, \dots, d_N$, are associated with *diseases*. Variables in the bottom level, denoted by $f_1, \dots, f_M$, are associated with symptoms or *findings*. Both diseases and findings are described by binary variables. An edge connects the disease d_i to the finding f_a whenever such a disease may be a cause for that finding. Such networks of implications are constructed on the basis of accumulated medical experience.

The network is completed with two types of probability distributions. For each disease d_i we are given an *a priori* occurrence probability $p(d_i)$. Furthermore, for each finding we have a conditional probability distribution for that finding given a certain disease pattern. This usually takes the so called 'noisy-OR' form:

$$p(f_a = 0 | d) = \frac{1}{z_a} \exp\left\{ - \sum_{i=1}^{N} \theta_{ia} d_i \right\} . \tag{1.1.2}$$

This network is to be used for diagnostic purposes. The findings are set to values determined by the observation of a patient. Given this pattern of symptoms, one would like to compute the marginal probability that any given disease is indeed present.

1.2 Pairwise graphical models

Pairwise graphical model are defined in terms of a simple graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with vertex set $\mathcal{V}$ and edge set $\mathcal{E}$. It is convenient to introduce a compatibility function $\psi_i : \mathcal{X} \to \mathbb{R}_+$ for each vertex $i \in \mathcal{V}$, and one $\psi_{ij} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ for each edge $(i, j) \in \mathcal{E}$. The joint distribution of $(X_1, \dots, X_n)$, $\mathbb{P}(\underline{X} = \underline{x}) = \mu(\underline{x})$ is then defined by

$$\mu(\underline{x}) = \frac{1}{Z} \prod_{(i,j)\in\mathcal{E}} \psi_{ij}(x_i, x_j) \prod_{i\in\mathcal{V}} \psi_i(x_i)\,. \tag{1.2.1}$$

The constant Z is called partition funtion and plays a quite important role. It is determined by the normalization condition on μ, which implies

$$Z = \sum_{\underline{x}\in\mathcal{X}^V} \prod_{(i,j)\in\mathcal{E}} \psi_{ij}(x_i, x_j) \prod_{i\in\mathcal{V}} \psi_i(x_i)\,. \tag{1.2.2}$$

Example: A classical example of pairwise model is the *Ising model* from statistical physics. In this case $\mathcal{X} = \{+1, -1\}$, and it is customary to parametrize the potentials in the form

$$\psi_{ij}(x_i, x_j) = \exp\left\{J_{ij} x_i x_j\right\}, \qquad \psi_i(x_i) = \exp\left\{h_i x_i\right\}. \tag{1.2.3}$$

The same model is popular in machine learning under the name of *Boltzmann machine* (in this case one often takes $x_i \in \{0, 1\}$. It includes as special cases some toy models for neural networks, such as the *Hopfield model*.

1.3 Factor graphs

A factor graph is a bipartite graph $G = (V, F, E)$, whereby V and F are two (finite) sets of vertices, and $E \subseteq V \times F$ a set of undirected edges. We will often identify $V = [n]$ (set of the first n integers), $F = [m]$. We will call *variable nodes* the vertices in V, and use for them letters $i, j, k, \dots$, and *function* (or *factor*) *nodes* the vertices in F to be denoted by $a, b, c \dots$.

Given $i \in V$, the set of its neighbors is denoted by $\partial i = \{a \in F : (i, a) \in E\}$. The neighborhood of $a \in F$, denoted by ∂a, is defined analogously.

A *factor graph model* (or *graphical model*) specifies the joint distribution of random variables $\underline{X} = (X_1, \dots, X_n) = \{X_i : i \in V\}$ taking value in a domain $\mathcal{X}$. As above, we shall assume[1] that $\mathcal{X}$ is a finite set. Indeed we shall mostly focus on the first case and write general equations in discrete notation. Finally, for any subset of variable nodes $A \subseteq V$, we write $\underline{x}_A = \{x_i : i \in A\}$.

[1] In principle, the formalism can be defined whenever $\mathcal{X}$ is a measure space, but I do not know of any example that is worth the effort.

Definition 2. *The joint distribution μ over $\underline{x} \in \mathcal{X}^V$ factors on the factor graph $G = (V, F, E)$ if there exists a vector of functions $\underline{\psi} = (\psi_1, \dots, \psi_m) = \{\psi_a : a \in F\}$, $\psi_a : \mathcal{X}^{\partial a} \to \mathbb{R}_+$, and a constant Z such that*

$$\mu(\underline{x}) = \frac{1}{Z} \prod_{a \in F} \psi_a(\underline{x}_{\partial a}) \,. \tag{1.3.1}$$

We then say that the pair (G, μ) is a factor graph model. *Equivalently, we will call the triple $(G, \underline{\psi}, \mathcal{X})$ a factor graph model.*

The ψ_a's are referred to as *potentials* or *compatibility functions*. The normalization constant is again called the *partition function* and is given by

$$Z \equiv \sum_{\underline{x}} \prod_{a \in F} \psi_a(\underline{x}_{\partial a}) \,. \tag{1.3.2}$$

Notice that, in order for the distribution (1.3.1) to be well defined, there must be at least one configuration $\underline{x} \in \mathcal{X}^V$ such that $\psi_a(\underline{x}_{\partial a}) > 0$ for all $a \in F$. Checking this property is already highly non-trivial, as shown by the following example.

Example. Let $\mathcal{X} = \{0, 1\}$, G be a factor graph with regular degree k at factor notes, and for any $a \in G$ let $(x_1^*(a), \dots, x_k^*(a)) \in \mathcal{X}^{\partial a}$ be given. Define

$$\psi_a(\underline{x}_{\partial a}) = \mathbb{I}(\underline{x}_{\partial a} \neq \underline{x}^*_{\partial a}) \,. \tag{1.3.3}$$

The problem of checking whether this specifies a factor graph model is the k-satisfiability problem which is NP-complete. If it does, the resulting measure is the uniform measure over solutions of the k-satisfiability formula.

1.3.1 Reduction from Bayesian networks to factor graphs

Given a Bayesian network G and a set of observed variable O, it is easy to obtain a factor graph representation of the conditional distribution $p(\underline{x}_{V\setminus O}|\underline{x}_O)$, by the following general rule is as follows: (i) associate a variable node with each non-observed variable (i.e. each variable in $\underline{x}_{V\setminus O}$); (ii) for each variable in $\pi(G)\setminus O$, add a degree 1 function node connected uniquely to that variable; (iii) for each non observed vertex v which is not in $\pi(G)$, add a function node and connect it to v and to all the parents of v; (iv) finally, for each observed variable u, add a function node and connect it to all the parents of u.

1.3.2 Reduction between to factor graphs

Pairwise models can be reduced to factor graphs and viceversa.

In order to get a reduction one has to construct, for any pairwise model $\mathcal{M}$, an associated factor graph model M which describes the same distribution $\mu(\,\cdot\,)$, and viceversa. Reduction from pairwise models to factor graphs is straightforward. Use $\mathcal{V}$ as the set of variable nodes, and associate to each edge $(i, j) \in \mathcal{E}$ and to each vertex $\mathcal{V}$ a factor node (in other words, set $V = \mathcal{V}$ and $F \simeq \mathcal{V} \cup \mathcal{E}$).

Reduction from factor graphs to pairwise models is slightly less trivial. The basic idea is to replace each factor node a by an ordinary vertex and associate to it a variable $x_a \in \mathcal{X}^{\partial a}$ which keeps track of the values $\{x_i :\in \partial a\}$. We will fill the details in class.

1.4 Markov random fields

In order to reduce factor graph models to pairwise models, we had to enlarge the alphabet $\mathcal{X}$. If you do not like this, but still you prefer graphs to factor graphs, Markov random fields (sometimes called Markov networks) are for you.

The underlying graph structure is an undirected graph $G = (V, E)$.

Definition 3. *The joint distribution μ over $\underline{x} \in \mathcal{X}^V$ is a Markov Random Field on $G = (V, F, E)$ if there exists a vector of functions $\underline{\psi} = \{\psi_C\}_C$ indexed by the cliques in G, $\psi_C : \mathcal{X}^C \to \mathbb{R}_+$, and a constant Z such that*

$$\mu(\underline{x}) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(\underline{x}_C) \,. \tag{1.4.1}$$

We then say that the pair (G, μ) (equivalently $(G, \underline{\psi}, \mathcal{X})$)) is a Markov Random Field.

Reduction to and from factor graphs is trivial.

The conditional independence property encoded in the graph structure is particularly crisp. Given sets of vertices $A, B, C \subseteq V$, we say that C *separates* A from B if every path in G connecting a vertex $i \in A$ to a vertex $j \in B$ has at least one vertex in C. We then have the following.

Theorem 1.4.1 (Hamersley, Clifford)**.** *If $\underline{X} = (X_i)_{i \in V}$ is distributed according to a Markov Random Field (G, μ), then for any $A, B, C \subseteq V$, such that C separates A from B, X_A is conditionally independent of X_A, given X_C.*

Viceversa, if $\underline{X} \sim \mu$, with $\mu(\underline{x}) > 0$ for all $\underline{x} \in \mathcal{X}^V$ is such that X_A is conditionally independent of X_A, given X_C, for any $A, B, C \subseteq V$, such that C separates A from B, then μ is a Markov Random Field on G.

Proving the direct part is fairly obvious. The converse instead uses a clever construction of the factors $\psi_C(\,\cdot\,)$ form the probability distribution μ. At the moment do not need this construction, but we will revisit Hammersley-Clifford theorem in Chapter 5.

1.5 Inference tasks

It is useful to keep in mind three inference problems that can be essentialy reduced to each other. We describe them for factor graphs but little needs to be changed for other types of graphical models.

Computing marginals. Given a (typically small) subset of vertices $A \subseteq V$, compute the probability $\mathbb{P}\{\underline{X}_A = \underline{x}_A\} = \mu(\underline{x}_A)$.

Computing conditional probabilities. Given two subsets A, $B \subseteq V$, compute the conditional probability $\mathbb{P}\{\underline{X}_A = \underline{x}_A | \underline{X}_B = \underline{x}_B\} = \mu(\underline{x}_A | \underline{x}_B)$.

Sampling. Sample a configuration of the random variables $\underline{X} \sim \mu$.

Partition function. Compute the partition function Z, as per Eq. (eq:PartFun).

Reducibility between this tasks has been studied in detail in the Monte Carlo Markov Chain literature within theoretical computer science: early references include [JVV86, JS89]. The ideas are straightforward to describe informally.

Reduction between marginals and conditional probabilities. Obviously marginals are a special case of conditional probabilities whereby $B = \emptyset$. In the opposite direction, computing the joint marginal $\mu(\underline{x}_{A\cup B})$ gives access to conditional probabilities via Bayes theorem.

A somewhat different way to obtain the second reduction consists in modifying the graphical model. Namely, given an assignment $\underline{x}^*_B$ to the variables in B construct the new model $\mu|_{B,\underline{x}^*_B}$ by adding for each node $i \in B$ a new factor node with potebtial $\psi_i(x_i) = \mathbb{I}(x_i = x_i^*)$. It is then easy to realize that $\mu|_{B,\underline{x}^*_B}(\underline{x}_A) = \mu(\underline{x}_A|\underline{x}^*_B)$. Therefore computing conditional probabilities in the original model is equivalent to computing marginals in the reduced model. We notice in passing that since the variables in B are fixed once and for all, we might as well eliminate them in the obvious way. We obtain a new model with vertex set $V \setminus B$ and factors $\psi_a|_{B,\underline{x}^*_B}$.

Reduction from marginals to sampling. If we can sample one configuration, we can as well sample many of them $\underline{X}^{(1)}, \underline{X}^{(2)}, \dots, \underline{X}^{(M)}$. Estimate tha marginals of $\mu(\underline{x}_A)$ with the empirical distribution of $\underline{x}_A$ in this sample. As long as A is bounded, precision ε can be achieved with probability latger than $(1-\delta)$ with $O(\varepsilon^{-1/2}\log(1/\delta))$ samples.

Reduction from sampling to marginals. Order the variable nodes arbitrarily, say 1, 2, 3, $\dots n$. Compute the marginal of the first variable $\mu(x_1)$ and sample from it (this is easy because the alphabet $\mathcal{X}$ is small). Let x_1^* be this sample. Reduce the model, by letting $V_1 = V \setminus \{1\}$ and $\psi_a^{(1)} = \psi_a|_{\{1\},x_1^*}$. Repeat.

Reduction from marginals to partition function. Let $Z|_{A,\underline{x}^*_A}$ be the partition function of the reduced model. Then we have the identity

$$\mu(\underline{x}^*_A) = \frac{Z|_{A,\underline{x}^*_A}}{Z}\,, \tag{1.5.1}$$

which immediately yields the desired reduction.

Reduction from marginals to partition function. Again order variables as 1, 2, 3, $\dots n$, and choose special values x_1^*, x_2^*, x_3^*, $\dots x_n^*$. Also let $A(\ell) = \{1, 2, \dots, \ell\} \subseteq V$. We then have

$$Z = \frac{Z}{Z|_{A(1),\underline{x}^*_{A(1)}}} \frac{Z|_{A(1),\underline{x}^*_{A(1)}}}{Z|_{A(2),\underline{x}^*_{A(2)}}} \dots \frac{Z|_{A(n-1),\underline{x}^*_{A(n-1)}}}{Z|_{A(n),\underline{x}^*_{A(n)}}} Z|_{A(n),\underline{x}^*_{A(n)}}\,. \tag{1.5.2}$$

The last term is trivial to compute since

$$Z|_{A(n),\underline{x}^*_{A(n)}} = \prod_{a\in F} \psi_a(\underline{x}_a)\,. \tag{1.5.3}$$

As for the n ratios they are just marginals. Indeed the identity (1.5.1) yields

$$\frac{Z|_{A(\ell+1),\underline{x}^*_{A(\ell+1)}}}{Z|_{A(\ell),\underline{x}^*_{A(\ell)}}} = \mu|_{A(\ell),\underline{x}^*_{A(\ell)}}(\underline{x}^*_{\ell+1})\,, \tag{1.5.4}$$

which implies the claimed reduction.

1.6 Continuous domains

In the case $\mathcal{X} = \mathbb{R}$, Eq. (1.3.1) defines the density of $(X_1, \ldots, X_n)$ with respect to the product Lebesgue measure (the ψ_a's have to be measurable in this case).

Chapter 2

Inference via message passing algorithms

Belief propagation (BP) is an umbrella term describing a family algorithms for approximate inference in graphical models. These algorithms are also collectively referred to as *message passing algorithms*. Both of these are somewhat loose terms, but generally convey the idea that algorithms in this class proceed by updating estimates of local marginals of the graphical model μ. Local marginals are updated by using information about marginals at neighboring nodes that is 'passed' along edges in the graph.

2.1 Preliminaries

Throughout this chapter we will consider the factor graph model

$$\mu(\underline{x}) = \frac{1}{Z} \prod_{a \in F} \psi_a(\underline{x}_a)\,, \tag{2.1.1}$$

defined on the factor graph $G = (V, F, E)$, and alphabet $\mathcal{X} \ni x_i$.

Consistently with our discussion of reduction between various inference tasks, we will focus on the problem of computing local marginals, i.e. marginal distributions of a small subset of variables, e.g. $\mu_A(\underline{x}_A) = \mathbb{P}_\mu\{\underline{X}_A = \underline{x}_A\}$. Explicitely

$$\mu_A(x_A) = \sum_{\underline{x}_{V \setminus A}} \mu(\underline{x})\,. \tag{2.1.2}$$

To be definite, you can think of simple 'one-point' marginals $A = \{i\}$. The most popular message-passing algorithm for computing marginals is known as the *sum-product* algorithm.

One can define a natural analogous of marginals when the problem is the one of computing the mode of μ. These are called, for lack of better ideas, 'max marginals'. For $A \subseteq V$, the max marginal is deined as

$$M_A(\underline{x}_A) = \max_{\underline{x}_{V \setminus A}} \mu(\underline{x})\,. \tag{2.1.3}$$

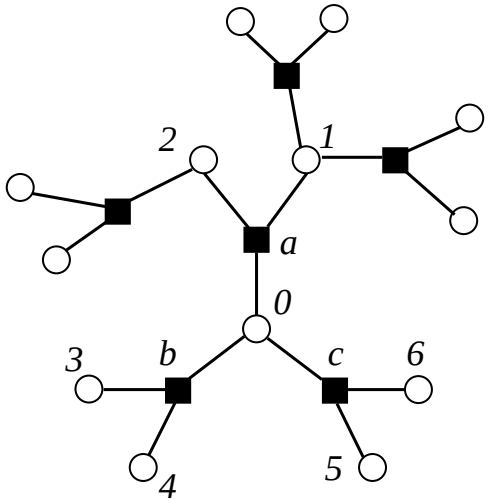


Figure 2.1: A small tree.

What is a max-marginal good for? Say that you computed $M_A(x_A)$. Then it is easy to see that, for any $\underline{x}^*_A \in \arg\max_{x_A} M_A(x_A)$, there exists an extension of $\underline{x}^*_A$ to a mode if μ. The analogous of the sum-product algorithm for computing max-marginals is known as the *max-product* algorithm (or sometimes *min-sum algorithm*).

In the following we will often have to write identites in which the overall normalization of both sides is not really interesting. In order to get rid of all these normalization floating around, it is conveniet to use a special notation. Throughout this course, the symbol $\cong$ is used to denote equality between functions up to a multiplicative normalization. Formally, we shall write $f(x) \cong g(x)$ if there exist a non-vanishing constant A such that $f(x) = Ag(x)$ for all x.

Sometimes the symbol $\propto$ is used for the same purpose. Since $\propto$ has actually is somewhat more vague, I will avoid this symbol.

2.2 Trees

Inference is easy if the underlying factor graph is a tree. By 'easy' we mean that it can be performed in time linear in the number of nodes, provided the factor nodes have bounded degree and the alphabet size is bounded as well. The algorthm that achieves this goal is a symple 'dynamic programming' procedure.

Let us see how this works on the simple example reproduced in Fig. **??**. We begin by assuming that we want to compute the marginal at node 0:

$$\mu(x_0) \cong \sum_{\underline{x}_{V\setminus 0}} \prod_{\ell \in F} \psi_\ell(\underline{x}_\ell)\,. \tag{2.2.1}$$

Node 0 is the root of three subtrees of G, that we can distinguish by the name of the factor node neighbor of 0 that they contain, namely $G_{a\to 0} = (V_{a\to 0}, F_{a\to 0}, E_{a\to 0})$, $G_{b\to 0} = (V_{b\to 0}, F_{b\to 0}, E_{b\to 0})$,

$G_{c\to 0} = (V_{c\to 0}, F_{c\to 0}, E_{c\to 0})$. We can rewrite the sum in Eq. (2.2.1) as (for such a small graph it is a bit of an overkill but bear with me)

$$\begin{aligned}
\mu(x_0) &\cong \sum_{\underline{x}_{V\setminus 0}} \prod_{\ell\in F_{a\to 0}} \psi_\ell(\underline{x}_\ell) \prod_{\ell\in F_{b\to 0}} \psi_\ell(\underline{x}_\ell) \prod_{\ell\in F_{c\to 0}} \psi_\ell(\underline{x}_\ell) \\
&\cong \Big\{ \sum_{\underline{x}_{V_{a\to 0}\setminus 0}} \prod_{\ell\in F_{a\to 0}} \psi_\ell(\underline{x}_\ell)\Big\} \cdot \Big\{ \sum_{\underline{x}_{V_{b\to 0}\setminus 0}} \prod_{\ell\in F_{b\to 0}} \psi_\ell(\underline{x}_\ell)\Big\} \Big\{ \sum_{\underline{x}_{V_{c\to 0}\setminus 0}} \prod_{\ell\in F_{c\to 0}} \psi_\ell(\underline{x}_\ell)\Big\} \\
&\cong \mu_{a\to 0}(x_0)\mu_{b\to 0}(x_0)\mu_{c\to 0}(x_0)\,.
\end{aligned}$$

Here in the second line we used the distributive property to 'push' the sums through the product of factor terms. In the last step we defined $\mu_{a\to 0}(x_0)$ that is the marginal with respect to the factor graph $G_{a\to 0}$.

Thus we reduced the problem of computing a marginal with respect to G to the ones of computing marginals with respect to subgraphs of G. We can repeat this recursively. Consider the subtree $G_{a\to 0}$. This can be decomposed into the factor node a, plus the subtrees $G_{1\to a}$ and $G_{2\to a}$. Using again the distributive property we have

$$\begin{aligned}
\mu_{a\to 0}(x_0) &\cong \sum_{\underline{x}_{V_{a\to 0}\setminus 0}} \prod_{\ell\in F_{a\to 0}} \psi_\ell(\underline{x}_\ell) \\
&\cong \sum_{x_1,x_2} \psi_a(\underline{x}_a) \Big\{ \sum_{\underline{x}_{V_{1\to a}\setminus 1}} \prod_{\ell\in F_{1\to a}} \psi_\ell(\underline{x}_\ell)\Big\} \Big\{ \sum_{\underline{x}_{V_{2\to a}\setminus 2}} \prod_{\ell\in F_{2\to a}} \psi_\ell(\underline{x}_\ell)\Big\} \\
&\cong \sum_{x_1,x_2} \psi_a(\underline{x}_a)\, \mu_{1\to a}(x_1)\, \mu_{2\to a}(x_2)\,.
\end{aligned}$$

It is quite clear that our arguments did not use the specific structure of the factor graph in Fig. 2.1. But they instead hold for any tree. Namely given a tree G and a directed edge $a \to i$ (factor-to-variable) or $i \to a$ (variable-to-factor) we can define the subgraphs $G_{a\to i}$ or $G_{i\to a}$ as above and the corresponding 'partial marginals' of the variable i: $\mu_{a\to i}(x_i)$ and $\mu_{i\to a}(x_i)$. We will also call these *messages*. We then have the following

Proposition 2.2.1. *For a tree graphical model, the 'partial marginals' are the unique solution of the equations*

$$\mu_{j\to a}(x_j) \cong \prod_{b\in\partial j\setminus a} \mu_{b\to j}(x_j)\,, \tag{2.2.2}$$

$$\mu_{a\to j}(x_j) \cong \sum_{\underline{x}_{\partial a\setminus j}} \psi_a(\underline{x}_{\partial a}) \prod_{k\in\partial a\setminus j} \mu_{k\to a}(x_k)\ , \tag{2.2.3}$$

For all $(i, a) \in E$.

Notice that both existence and uniqueness follow from the recursive argument discussed above. It should also be clear that marginals of μ can be computed easily in terms of partial marginals. We have for instance:

$$\mu_i(x_i) \cong \prod_{a\in\partial i} \mu_{a\to i}(x_i)\,, \tag{2.2.4}$$

$$\mu_a(\underline{x}_a) \cong \psi_a(\underline{x}_a) \prod_{i\in\partial a} \mu_{i\to a}(x_i)\,. \tag{2.2.5}$$

2.3 The sum-product algorithm

The sum product algorithm is an iterative message passing algorithm. The basic variables are 'messages' which are probability distributions over $\mathcal{X}$. These are also called 'beliefs'. Two such distributions are used for each edge in the graph $\nu_{i\to a}(\,\cdot\,)$ (variable to factor node) $\widehat{\nu}_{a\to i}(\,\cdot\,)$ (factor to variable node). We shall denote the vector of *messages* by $\underline{\nu} = \{\nu_{i\to a}, \widehat{\nu}_{a\to i}\}$: it is a vector of probability distributions, indexed by directed edges in G.

We shall indicate the iteration number by supercripts, e.g. $\nu^{(t)}_{i\to a}$, $\widehat{\nu}^{(t)}_{a\to i}$ are the messages value after t iterations. Messages are initialized to some non-informative values, typically $\nu^{(0)}_{i\to a}$, $\widehat{\nu}^{(0)}_{a\to i}$ are equal to the uniform distribution over $\mathcal{X}$.

Various update scheduling are possible, but for the sake of simplicity we will consider parallel updates, which read:

$$\nu^{(t+1)}_{j\to a}(x_j) \;\cong\; \prod_{b\in\partial j\setminus a} \widehat{\nu}^{(t)}_{b\to j}(x_j)\,, \tag{2.3.1}$$

$$\widehat{\nu}^{(t)}_{a\to j}(x_j) \;\cong\; \sum_{\underline{x}_{\partial a\setminus j}} \psi_a(\underline{x}_{\partial a}) \prod_{k\in\partial a\setminus j} \nu^{(t)}_{k\to a}(x_k)\;. \tag{2.3.2}$$

It is understood that, when $\partial j \setminus a$ is an empty set, $\nu_{j\to a}(x_j)$ is the uniform distribution.

After t iterations, one can estimate the marginal distribution $\mu(x_i)$ of variable i using the set of *all* incoming messages. The BP estimate is:

$$\nu^{(t)}_{i}(x_i) \cong \prod_{a\in\partial i} \widehat{\nu}^{(t-1)}_{a\to i}(x_i)\,. \tag{2.3.3}$$

The rationale for Eqs. (2.3.1) and (2.3.2) is easy to understand given the discussion in the previous section: we are trying to iteratively find solutions of Eqs. (2.2.2), (2.2.3). These exactly hold for partial marginals on trees. On general graphs, the resulting fixed points will not be necessarily marginals of μ, but te hope is that they are nevertheless a good approximation of the actual marginals. We use a different letter (ν instead of μ) to emphasize the fact that messages do not coincide in general with marginals.

We formalize the connection between the sum-product algorithm and tree graphical model as follows.

Proposition 2.3.1. *If G is a tree, then the sum-product algorithm converges in* $\mathrm{diam}(G)$ *iteration to its unique fixed point $\nu_{i\to a} = \mu_{i\to a}$, $\widehat{\nu}_{a\to i} = \mu_{a\to i}$. The resulting estimates for the marginals, cf. Eq. (2.3.3) are exact.*

2.4 The max-product algorithm

Max-product updates are used to approximately find the mode of a distribution specified by a factor graph model. Before discussing the algorithm, it is useful to show how mode computation can be effectively reduced to the computation of max-marginals. Recall that the mode of μ, cf. Eq. (2.1.1) is any assignment $\underline{x}^* \in \mathcal{X}^V$ of maximal probability, i.e.

$$\underline{x}^* \in \arg\max_{\underline{x}\in\mathcal{X}^V} \prod_{a\in F} \psi_a(\underline{x}_a)\,. \tag{2.4.1}$$

Order the variables arbitrarily, say 1, 2, ..., n. Compute the max-marginal of x_1, call it $M_1(x_1)$ and choose $x_1^* \in \arg\max_{x_1 \in \mathcal{X}} M_1(x_1)$. Reduce the model by , by letting $V_1 = V \setminus \{1\}$ and $\psi_a^{(1)} = \psi_a|_{\{1\},x_1^*}$. Compute the max-marginal of variable 2 in the reduced model $M_2^{(1)}(x_2)$ and repeat.

Vice-versa, computing max-marginals is easy if we have at our disposal a routine to compute the mode of a graphical model. Suppose that the variable of interest is x_i. We have

$$M_i(x_i) \cong \max_{\underline{x}_{V\setminus\{i\}}} \prod_{a\in F} \psi_a|_{\{i\},x_i}(\underline{x}_a)\,, \tag{2.4.2}$$

which is requires obtained from the computation of the mode of the reduced graphical model.

The derivation of the max-product algorithm closely parallels the one of the sum-prosuct algorithm. (Indeed the two can be put in a unified framework using the so-called 'generalized distributive property' [AM00].) Again consider first the case of a tree graphical model. One can define subtrees $G_{i\to a}$, $G_{a\to i}$ analogously to what we did for the sum-product algorithm, and the corresponding messages as

$$M_{i\to a}(x_i) \cong \max_{\underline{x}_{V_{i\to a}\setminus i}} \prod_{b\in F_{i\to a}} \psi_b(\underline{x}_b)\,, \tag{2.4.3}$$

$$M_{i\to a}(x_i) \cong \max_{\underline{x}_{V_{a\to i}\setminus i}} \prod_{b\in F_{a\to i}} \psi_b(\underline{x}_b)\,. \tag{2.4.4}$$

Proceeding as for the sum-product algorithm, it is then easy to show that these marginals are the unique solution of the equations

$$M_{j\to a}(x_j) \;\cong\; \prod_{b\in\partial j\setminus a} M_{b\to j}(x_j)\,, \tag{2.4.5}$$

$$M_{a\to j}(x_j) \;\cong\; \max_{\underline{x}_{\partial a\setminus j}} \Big\{\psi_a(\underline{x}_{\partial a}) \prod_{k\in\partial a\setminus j} M_{k\to a}(x_k)\Big\}\,, \tag{2.4.6}$$

with one equation per edge in G.

The above equations for a tree motivate the following iterative algorithm

$$\nu_{i\to a}^{(t+1)}(x_i) \;\cong\; \prod_{b\in\partial i\setminus a} \widehat{\nu}_{b\to i}^{(t)}(x_i)\,, \tag{2.4.7}$$

$$\widehat{\nu}_{a\to i}^{(t)}(x_i) \;\cong\; \max_{\underline{x}_{\partial a\setminus i}} \left\{\psi_a(\underline{x}_{\partial a}) \prod_{j\in\partial a\setminus i} \nu_{j\to a}^{(t)}(x_j)\right\}\,. \tag{2.4.8}$$

The max-marginals are then estimated as

$$\nu_i^{(t)}(x_i^*) \cong \prod_{b\in\partial i\setminus a} \widehat{\nu}_{b\to i}^{(t)}(x_i)\,,\,. \tag{2.4.9}$$

The name of the max-product algorithm has obvious origins. Sometimes, the same algorithm is also called min-sum because is written in a slightly different form. Instead of maximizing the

probability, one starts from the equivalent problem of minimizing a cost function that factorizes according to G:

$$H(\underline{x}) = \sum_{a \in F} H_a(\underline{x}_{\partial a})\,. \tag{2.4.10}$$

The min-sum update equations read

$$\mathsf{J}^{(t+1)}_{i\to a}(x_i) \;=\; \text{const.} + \sum_{b\in\partial i\setminus a} \widehat{\mathsf{J}}^{(t)}_{b\to i}(x_i)\,, \tag{2.4.11}$$

$$\widehat{\mathsf{J}}^{(t)}_{a\to i}(x_i) \;\cong\; \min_{\underline{x}_{\partial a\setminus i}} \left\{ H_a(\underline{x}_{\partial a}) + \sum_{j\in\partial a\setminus i} \mathsf{J}^{(t)}_{j\to a}(x_j) \right\}\,. \tag{2.4.12}$$

They can be obtained from the max-product updates via the identification $\nu^{(t)}_{i\to a}(x_i) \cong \exp\{-J^{(t)}_{i\to a}(x_i)\}$ and $\widehat{\nu}^{(t)}_{a\to i}(x_i) \cong \exp\{-\widehat{\mathsf{J}}^{(t)}_{a\to i}(x_i)\}$

An analogous of Proposition 2.3.1 holds for the max-product algorithm: On tree graphical models the max-product algorithm converges to the correct max-marginals after $\mathrm{diam}(G)$ iterations.

2.5 Existence

Let $\vec{E}$ denote the set of directed edges (whereby each edge (i,a) corresponds to two directed edges $i\to a$ and $a\to i$), and $\mathsf{M}(\mathcal{X})$ the set of probability distributions over $\mathcal{X}$ (i.e. the $(|\mathcal{X}|-1)$-dimensional simplex). Then the space of messages is $\mathsf{M}(\mathcal{X})^{\vec{E}} \ni \underline{\nu}$. The sum-product BP update defines a non-linear mapping

$$\begin{aligned} \mathsf{T}_G : \mathsf{M}(\mathcal{X})^{\vec{E}} &\to \mathsf{M}(\mathcal{X})^{\vec{E}}\,, \\ \underline{\nu} &\mapsto \mathsf{T}_G(\underline{\nu})\,. \end{aligned}$$

For the sake of completeness, let me copy here the definition of such a mapping. If $\underline{\nu}' = \mathsf{T}_G(\underline{\nu})$, then

$$\nu'_{j\to a}(x_j) \;\cong\; \prod_{b\in\partial j\setminus a} \widehat{\nu}_{b\to j}(x_j)\,, \tag{2.5.1}$$

$$\widehat{\nu}'_{a\to j}(x_j) \;\cong\; \sum_{\underline{x}_{\partial a\setminus j}} \psi_a(\underline{x}_{\partial a}) \prod_{k\in\partial a\setminus j} \nu_{k\to a}(x_k)\,. \tag{2.5.2}$$

Notice that a distribution $\mu(\underline{x})$ admits (in general) more than one decomposition of the form (**??**). On the other hand:

Remark 1. *The mapping T_G depends on the factor graph G and on the potentials ψ.*

If G is a tree (or a forest), T_G admits a unique fixed point. What happens on general graphs G? A large number of cases is covered by the following definition.

Definition 4. *We say that the pair (G,ψ) is* permissive *if, for each $i\in V$, there exists $x_i^*\in\mathcal{X}$ such that*

$$\psi_a(x_i^*, \underline{x}_{\partial a\setminus i}) \geq \psi_{\min} > 0\,, \tag{2.5.3}$$

for each $\underline{x}_{\partial a\setminus i}\in\mathcal{X}^{\partial a\setminus i}$ and $a\in\partial i$.

Then we have the following simple result.

Proposition 2.5.1. *If the factor graph model (G, ψ) is permissive, then the BP operator T_G admits at least one fixed point.*

Proof. We will only sketch the main ideas, leaving details to the reader. The proof indeed a direct application of the following well known result in topology.

Theorem 2.5.2 (Brouwer's fixed point theorem)**.** *Let $f : B_N \to B_N$ be a continuous mapping from the N-dimensional closed unit ball $B_N = \{\underline{z} \in \mathbb{R} : \ ||\underline{z}|| \le 1\}$ to itself. Then f admits at least one fixed point.*

Since this is a theorem about continuous functions, it generalizes to any domain D that is homeomorphic to B_N. (Indeed, if $\varphi : B_N \to D$ is such an homeomorphism, and $g : D \to D$ is the mapping of interest, set $f = \varphi^{-1} \circ g \circ \varphi$.)

The theorem is then applied to $\mathsf{T}_G : \mathsf{M}(\mathcal{X})^{\vec{E}} \to \mathsf{M}(\mathcal{X})^{\vec{E}}$. To finish the proof, one has to check that: (i) $\mathsf{M}(\mathcal{X})^{\vec{E}}$ is homeomorphic to a unit ball; (ii) T_G is continuous. In the second step, one uses the permissivity assumption. □

2.6 An example: Group testing

Let us reconsider the group testing example. Recall that the we want to make inference on variables $\underline{x} = (x_1, \dots, x_n)$, $x_i \in \{0, 1\}$ (causes), given observations $\underline{y} = (y_1, \dots, y_m)$, $y_a = \vee_{i \in \partial a} x_i$ (symptoms). We assume that, a priori, the x_i's are iid Bernoulli(p). As usual, we consider the factor graph $G = (V, F, E)$ where $V = [n]$, $F = [m]$, and $(i, a) \in E$ if and only if $i \in V$ is among the causes of $a \in F$. We further let $F = F_0 \cup F_1$, where $F_{0/1} \equiv \{a \in F : \ y_a = 0/1\}$.

$$\mu(\underline{x}) = \frac{1}{Z} \prod_{a \in F_0} \mathbb{I}(\underline{x}_{\partial a} = \underline{0}) \prod_{a \in F_1} \mathbb{I}(\underline{x}_{\partial a} \neq \underline{0}) \prod_{i \in V} p^{x_i} (1-p)^{1-x_i} \, . \tag{2.6.1}$$

We could write down the BP update rules for this model (and the reader is indeed invited to do so) but it is more convenient to simplify it a bit. If $y_a = 0$, then we know that $x_i = 0$ for each $i \in \partial a$. We can thus reduce the factor graph by eliminating all f-nodes in F_0, and all adjacent v-nodes. Next if $a \in F_1$ has only one adjacent node $i \in V$, then we know that $x_i = 1$. We can than recursively eliminate all such nodes. At the end of such process we are left with a subgraph $G' = (V', F', E')$ with $V' \subseteq V$, $F' \subseteq F_1$, and each node $a \in F'$ having degree $|\partial a| \ge 2$. The joint distribution

$$\mu(\underline{x}) = \frac{1}{Z} \prod_{a \in F'} \mathbb{I}(\underline{x}_{\partial a} \neq \underline{0}) \prod_{i \in V'} p^{x_i} (1-p)^{1-x_i} \, . \tag{2.6.2}$$

Is this graphical model permissive?

Messages are distributions over binary variables. Each term $p^{x_i}(1-p)^{1-x_i}$ can be described by a factor node of degree one. The corresponding message directed towards i is fixed to $\overline{\nu}$, with $\overline{\nu}(0) = 1 - p$, $\overline{\nu}(1) = p$. Since this message does not change over time, we will only write the BP

update equations for other edges. They read

$$\nu_{i\to a}^{(t+1)}(x_i) \;\cong\; \overline{\nu}(x_i) \prod_{b\in\partial i\setminus a} \widehat{\nu}_{b\to i}^{(t)}(x_i)\,, \tag{2.6.3}$$

$$\widehat{\nu}_{a\to i}^{(t)}(x_i) \;\cong\; \begin{cases} 1-\prod_{j\in\partial a\setminus i}\nu_{j\to a}^{(t)}(0) & \text{if } x_i=0,\\ 1 & \text{if } x_i=1. \end{cases} \tag{2.6.4}$$

Distributions over binary variables can be parametrized by a single real number. For instance, we can use the probability of 0. If we let $v_{i\to a}\equiv\nu_{i\to a}(0)$ and $\widehat{v}_{a\to i}\equiv\widehat{\nu}_{a\to i}(0)$, we obtain the equations

$$v_{i\to a}^{(t+1)} \;=\; \frac{(1-p)\prod_{b\in\partial i\setminus a}\widehat{v}_{b\to i}^{(t)}}{(1-p)\prod_{b\in\partial i\setminus a}\widehat{v}_{b\to i}^{(t)}+p\prod_{b\in\partial i\setminus a}(1-\widehat{v}_{b\to i}^{(t)})}\,, \tag{2.6.5}$$

$$\widehat{v}_{a\to i}^{(t)} \;=\; 1-\frac{1}{2-\prod_{j\in\partial a\setminus i}v_{j\to a}^{(t)}}\,. \tag{2.6.6}$$

Using these messages, one can estimate the a posteriori probability that cause i is present as

$$\nu_i^{(t+1)}(x_i=1)=\frac{p\prod_{b\in\partial i}(1-\widehat{v}_{b\to i}^{(t)})}{(1-p)\prod_{b\in\partial i}\widehat{v}_{b\to i}^{(t)}+p\prod_{b\in\partial i}(1-\widehat{v}_{b\to i}^{(t)})}\,. \tag{2.6.7}$$

2.7 Pairwise graphical models

Since pairwise graphical models are reducible to factor fraph (and indeed are a special type of factor graph whereby all factors have degree 2) we could in principle skip this section altogether. On the other hand, they are an important subclass, and it is instructive to write equations explicitely in this case.

We consider the model

$$\mu(\underline{x})=\frac{1}{Z}\prod_{(i,j)\in E}\psi_{ij}(x_i,x_j)\prod_{i\in V}\psi_i(x_i)\,. \tag{2.7.1}$$

Adapting Eqs. (2.3.1), (2.3.2) is straightforward. We can associate one factor node a to each edge (i,j) and hence will have messages $\nu_{i\to(i,j)}$ and $\widehat{\nu}_{(i,j)\to j}$. However, the latter is a simple function of the former:

$$\widehat{\nu}_{(i,j)\to j}^{(t)}(x_j)\cong\sum_{x_i}\psi_{ij}(x_i,x_j)\nu_{i\to(i,j)}^{(t)}(x_i)\,. \tag{2.7.2}$$

We thus can eliminate $\widehat{\nu}_{(i,j)\to j}^{(t)}$ completely, and write simple updates for $\nu_{i\to(i,j)}^{(t)}$, that we will denote as $\nu_{i\to j}^{(t)}$. Notice that in principle we should also introduce a factor node a for each singleton term $\psi_i(x_i)$, but the corresponding message would be $\widehat{\nu}_{a\to i}(x_i)\cong\psi_i(x_i)$ and we can eliminate it as well.

We finally obtain the update rules

$$\nu_{i\to j}^{(t+1)}(x_i)\cong\psi_i(x_i)\prod_{k\in\partial i\setminus j}\Big\{\sum_{x_k}\psi_{ik}(x_i,x_k)\nu_{k\to i}(x_k)\Big\}\,. \tag{2.7.3}$$

2.8 Another example: Ising model

Recall that the Ising model is a pairwise model on the graph $G = (V, E)$ whose distribution takes the form

$$\mu(\underline{x}) = \frac{1}{Z} \exp\Big\{ \sum_{(i,j)\in E} J_{ij} x_i x_j + \sum_{i\in V} h_i x_i \Big\}, \tag{2.8.1}$$

with $x_i \in \mathcal{X} \equiv \{+1, -1\}$ (any pairwise nmodel on binary variables can be written in this form).

It is immesiate to adapt the sum-product update (2.7.3) to the present case. We get

$$\nu_{i\to j}^{(t+1)}(x_i) \cong e^{h_i x_i} \prod_{k\in\partial i\setminus j} \Big\{ \sum_{x_k\in\{+1,-1\}} e^{J_{ik} x_i x_k} \nu_{k\to i}(x_k) \Big\}. \tag{2.8.2}$$

Since x_i takes two values, it is sufficient to specify one real parameter for message. It is customary in many applications to use log-likelihood ratios, i.e.

$$h_{i\to j}^{(t)} \equiv \frac{1}{2} \log\Big\{ \frac{\nu_{i\to j}^{(t)}(+1)}{\nu_{i\to j}^{(t)}(-1)} \Big\}. \tag{2.8.3}$$

In terms of these variables, the update becomes (after a tedious calculus exercise)

$$h_{i\to j}^{(t+1)} = h_i + \sum_{k\in\partial i\setminus j} \operatorname{atanh}\big\{ \tanh(J_{ik}) \tanh(h_{k\to i}^{(t)}) \big\}. \tag{2.8.4}$$

Where tanh and atanh are hyerbolic tangent and arctangent.

2.9 Monotonicity

Consider again the group testing example. Notice that, according to Eq. (2.6.5), the output of a function node is monotonically decreasing in its inputs. On the other hand, by Eq. (2.6.6), the output of a function node is monotonically increasing in its inputs. It follows that the BP iteration is *anti-monotone* in the following sense.

Consider the vector of variable-to-function node messages, and denote it, with an abuse of notation, by $\underline{\nu} = \{\nu_{i\to a}(\,\cdot\,) : (i, a) \in E\}$. Let T_G^* be the mapping degined by one full BP iteration applied to this set of messages. We write $\underline{\nu} \preceq \underline{\nu}'$ if, for each $(i, a) \in E$, $\nu_{i\to a}(1) \le \nu'_{i\to a}(1)$. This is a partial ordering on the space of message sets.

The above remarks imply that, if $\underline{\nu} \preceq \underline{\nu}'$, then $\mathsf{T}_G^*(\underline{\nu}) \succeq \mathsf{T}_G^*(\underline{\nu}')$. This remarks has some interesting consequences. Imagine to initialize messages in such a way that $\nu_{i\to a}^{(0)}(0) = 1$ for each directed edge $i \to a$. Then $\underline{\nu}^{(1)} \succeq \underline{\nu}^{(0)}$ necessarily. By applying $(\mathsf{T}_G^*)^t$ to this inequality, we get that $\underline{\nu}^{(t+1)} \succeq \underline{\nu}^{(t)}$ for t even, and $\underline{\nu}^{(t+1)} \preceq \underline{\nu}^{(t)}$ for t odd: messages 'toggle'. By the same token $\underline{\nu}^{(2)} \succeq \underline{\nu}^{(0)}$. Applyting $(\mathsf{T}_G^*)^t$, we get that the sequence of messages at even times is monotone increasing, and the one at odd times is monotone decreasing. In particular The two sequences converge to a pair of fixed points of $(\mathsf{T}_G^*)^2$.

Another example of the utility of monotonicity arguments is provided by the Ising model, discussed in section 2.8. If the interaction parameters are all non-negative, $J_{ij} \ge 0$, then this update is monotone increasing. It is then easy to show that BP always converges if initialized, for instance, with $h_{i\to j}^{(0)} = 0$.

2.10 Hidden Markov Models

A (homogeneous) Markov Chain over the state space $\mathcal{X}$ is a sequence of random variables $\underline{X} = (X_0, X_1, \dots, X_n)$ whose joint distribution takes the form

$$\mathbb{P}\{\underline{X} = \underline{x}\} = p_0(x_0) \prod_{i=0}^{n-1} p(x_{i+1}|x_i)\,, \tag{2.10.1}$$

for some initial distribution p_0 and conditional probability kernel $p(x_{i+1}|x_i)$.

An Hidden Markov Model arises when partial/noisy observations of the chain are available. The joint distribution of the chain $\underline{X}$ and of the observations $\underline{Y}$ reads

$$\mathbb{P}\{\underline{X} = \underline{x}, \underline{Y} = \underline{y}\} = p_0(x_0) \prod_{i=0}^{n-1} p(x_{i+1}|x_i) \prod_{i=0}^{n} q(y_i|x_i)\,, \tag{2.10.2}$$

where q describes the observation process.

A common computational task is the one of inferring the sequence of states $\underline{x}$ from the observations. For this task, it makes sense to consider the conditional probability distribution of $\underline{X}$ given the observations, which by Bayes theorem takes the form

$$\mathbb{P}\{\underline{X} = \underline{x}|\underline{Y} = \underline{y}\} \cong p_0(x_0) \prod_{i=0}^{n-1} p(x_{i+1}|x_i) \prod_{i=0}^{n} q(y_i|x_i)\,. \tag{2.10.3}$$

This is easily represented as a pairwise graphical model whose underlying graph is a chain. We get indeed (assuming $\underline{y}$ fixed once and for all) $\mathbb{P}\{\underline{X} = \underline{x}|\underline{Y} = \underline{y}\} = \mu(\underline{x})$, where

$$\begin{aligned}
\mu(\underline{x}) &= \frac{1}{Z} \prod_{i=0}^{n-1} \psi_i(x_i, x_{i+1})\,, & (2.10.4)\\
\psi_0(x_0, x_1) &= p_0(x_0) p(x_1|x_0)\, q(y_0|x_0)\,, & (2.10.5)\\
\psi_i(x_i, x_{i+1}) &= p(x_{i+1}|x_i)\, q(y_i|x_i) \quad i \geq 1\,. & (2.10.6)
\end{aligned}$$

The sum-product update is straightforwarly written for this case

$$\begin{aligned}
\nu_{i\to i+1}(x_i) &\cong \sum_{x_{i-1}} \psi_{i-1}(x_{i-1}, x_i) \nu_{i-1\to i}(x_{i-1})\,, & (2.10.7)\\
\nu_{i\to i-1}(x_i) &\cong \sum_{x_{i+1}} \psi_i(x_i, x_{i+1}) \nu_{i+1\to i+}(x_i)\,. & (2.10.8)
\end{aligned}$$

Notice that since the graph is a tree, the algorithm is guaranteed to converge, and it is sufficient a single pass forward and a single pass backwards. Indeed the algorithm has been well-known for a long time as the *forward-backward algorithm*.

Chapter 3

Mixing

We will begin by introducing the Monte Carlo Markov Chain method. This is a randomized algorithm that can be use to compute marginals of a graphical model. We will see that MCMC is guaranteed to work when distinct variables in the graphical model are not too strongly correlated. We will then introduce the computation tree: a very useful construction for analyzing message passing algorithms. In particular, the behavior of such algorithms is related to the strenght of correlations on such trees.

3.1 Monte Carlo Markov Chain method

Given a model defined on the factor graph $G = (V = [n], F, E)$:

$$\mu(\underline{x}) = \frac{1}{Z} \prod_{a \in F} \psi_a(\underline{x}_{\partial a}) \,, \tag{3.1.1}$$

where $\underline{x} \in \mathcal{X}^V$, the Monte Carlo method tries to estimate its marginal by sampling $\underline{x}^{(1)}, \dots \underline{x}^{(N)}$ approximately iid from $\mu(\,\cdot\,)$. In order to obtain iid samples, we introduce an irreducible and aperiodic Markov Chain that has μ as its stationary measure.

To be concrete, we will focus on a siggle example throughout this lecture (but wath we shall say can indeed be generalized). Given $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, sample independent sets with probability

$$\mu(\underline{x}) = \frac{1}{Z} \prod_{(i,j) \in \mathbb{E}} \mathbb{I}\big((x_i, x_j) \neq (1,1)\big) \prod_{i \in \mathcal{V}} \lambda^{x_i} \tag{3.1.2}$$

where $x_i \in \mathcal{X} = \{0, 1\}$.

Metropolis dynamics allows to define a Markov chain with μ as its stationary measure. The chain is identified by the matrix of transition probabilities $\{p(\underline{x}, \underline{y})\}_{\underline{x}, \underline{y} \in \mathcal{X}^{\mathcal{V}}}$, whereby $p(\underline{x}, \underline{y})$ is the probability that the configuration at time $t+1$ is $\underline{y}$, given that at time t it is $\underline{x}$. In the case of Metropolis dynamics, they satisfy the so-called *reversibility condition*

$$\mu(\underline{x}) p(\underline{x}, \underline{x}') = \mu(\underline{x}') p(\underline{x}', \underline{x}) \,. \tag{3.1.3}$$

This in particular implies that μ is a stationary measure for the chain.

For our example, Metropolis dynamics can be described as follows:

- Given the current configuration $\underline{x}$, first choose $i \in \mathcal{V}$ uniformly at random.
- Set $x'_j = x_j \forall j \neq i$, and choose $x'_i \in \{0,1\}$ uniformly at random. This is the "proposed" move.
- The proposal is accepted with probability

$$\pi = \min\{1, \lambda^{\hat{x}_i - x_i}\}$$

  if all the neighbors of i are empty. Otherwise, $\pi = 0$.

The transition rule can also be specified as follows.
If $x_i = 1$,

$$x'_i = \begin{cases} 0 & \text{with prob } \frac{1}{2}\min(1, \lambda^{-1}) \\ x_i & \text{otherwise} \end{cases}$$

If $x_i = 0$,

$$x'_i = \begin{cases} 1 & \text{if } x_j = 0 \ \forall j \sim i \text{ and with prob } \frac{1}{2}\min(1, \lambda) \\ x_i & \text{otherwise} \end{cases}$$

Notice that with probability half, $\underline{x}$ does not change. It is also immediate to see that the chain is irreducible: for any two independent sets $\underline{x}$, $\underline{y}$, there exists a sequence of transitions with non vanishing probability that bring from $\underline{x}$ to $\underline{y}$. You are invited to check that the chain is indeed reversible with respect to $\mu(\,\cdot\,)$. This implies that $\mu(\,\cdot\,)$ is indeed the unique stationary measure.

3.1.1 Mixing time

After defining a Markov chain with μ as its stationary measure, we start from an arbitrary configuration $\underline{x}^{(0)}$, eg. an empty independent set, and run the transitions for t steps. We "pretend" that the configuration $\underline{x}^{(t)}$ is distributed according to $\mu(\,\cdot\,)$ to compute the desired marginal. If we denote by $\mu^{(t)}$ the distribution of $\underline{x}^{(t)}$, we would like

$$\mu^{(t)}(\underline{x}) \approx \mu(\underline{x})$$

In order to make this idea more precise, we define

Total variation distance

$$\|\mu^{(t)} - \mu\|_{TV} \equiv \frac{1}{2} \sum_{\underline{x}} |\mu^{(t)}(\underline{x}) - \mu(\underline{x})|$$

and

Mixing time

$$\tau_{mix}(\epsilon) = \sup_{\underline{x}_0} \inf\{\ \tau\ : \|\mu^{(t)} - \mu\|_{TV} \leq \epsilon \, \forall t \geq \tau\}$$

Notice that

$$\|\mu^{(t)} - \mu\|_{TV} \leq \epsilon \Rightarrow |\sum_{\underline{x}} f(\underline{x})\mu^{(t)}(\underline{x}) - \sum_{\underline{x}} f(\underline{x})\mu(\underline{x})| \leq 2\epsilon \sup_{\underline{x}} |f(\underline{x})|$$

In particular, singular variable marginals are well aproximated

$$|\mu_i^{(t)}(x_i) - \mu(x_i)| \leq \epsilon$$

3.1.2 Bounding mixing time using coupling

The next question is: How can we upper bound the mixing time?

One method that is easy to apply is Path Coupling, developed by Bubley and Dyer. There are many other techniques some of which are more powerful, but Path Coupling is easily applicable to many examples.

Given two rv. X, Y on different probability spaces, a coupling is a rv. $(\tilde{X}, \tilde{Y})$, such that $\tilde{X}$ is distributed as X and $\tilde{Y}$ as Y.

We prove a lemma that serves as the foundation of general coupling method.

Lemma 3.1.1. *Given $X_1 \sim \mu_1$, $X_2 \sim \mu_2$ and coupling (X_1, X_2), we have*

$$\|\mu_1 - \mu_2\|_{TV} \leq \mathbb{P}(X_1 \neq X_2)$$

Proof.

$$\begin{aligned}
\mathbb{P}(X_1 \neq X_2) &= \sum_x (\mathbb{P}(X_1 = x) - \mathbb{P}(X_1 = x, X_2 = x)) \\
&\geq \sum_x (\mathbb{P}(X_1 = x) - \min[\mathbb{P}(X_1 = x), \mathbb{P}(X_2 = x)]) \\
&= \sum_x \max(\mathbb{P}(X_1 = x) - \mathbb{P}(X_2 = x), 0) \\
&= \frac{1}{2} \sum_x |\mathbb{P}(X_1 = x) - \mathbb{P}(X_2 = x)|
\end{aligned}$$

□

Corollary 3.1.2. *Let $\underline{x}^{(1,t)}$, $\underline{x}^{(2,t)}$ be two realizations of the Markov Chain st $\underline{x}^{(1,0)} = \underline{x}^{(0)}$, $\underline{x}^{(2,t)} \sim \mu$. Then, for any coupling of $\{\underline{x}^{(1,t)}\}$, $\{\underline{x}^{(2,t)}\}$,*

$$\left\| \mu^{(t)}_{\underline{x}^{(0)}} - \mu \right\|_{TV} \leq \mathbb{P}(\underline{x}^{(1,t)} \neq \underline{x}^{(2,t)})$$

In order to see how we can apply this corollary, we will go back to our example on independent sets. First we define the distance of two configurations $\underline{x}$, $\hat{\underline{x}}$.

Definition 5.

$$\mathcal{D}(\underline{x}, \underline{x}') = [\textit{minimal number of ``allowed'' moves to go from } \underline{x} \textit{ to } \underline{x}']$$

where 'allowed' moves are all the transitions with positive probability in the Markov chain. Notice that $\mathcal{D}(\underline{x}, \hat{\underline{x}}) \leq 2n$.

Suppose we are able to prove

$$\mathbb{E}[\mathcal{D}(\underline{x}^{(1,t+1)}, \underline{x}^{(2,t+1)}) | \underline{x}^{(1,t)}, \underline{x}^{(2,t)}] \leq \beta \mathcal{D}(\underline{x}^{(1,t)}, \underline{x}^{(2,t)}) \tag{3.1.4}$$

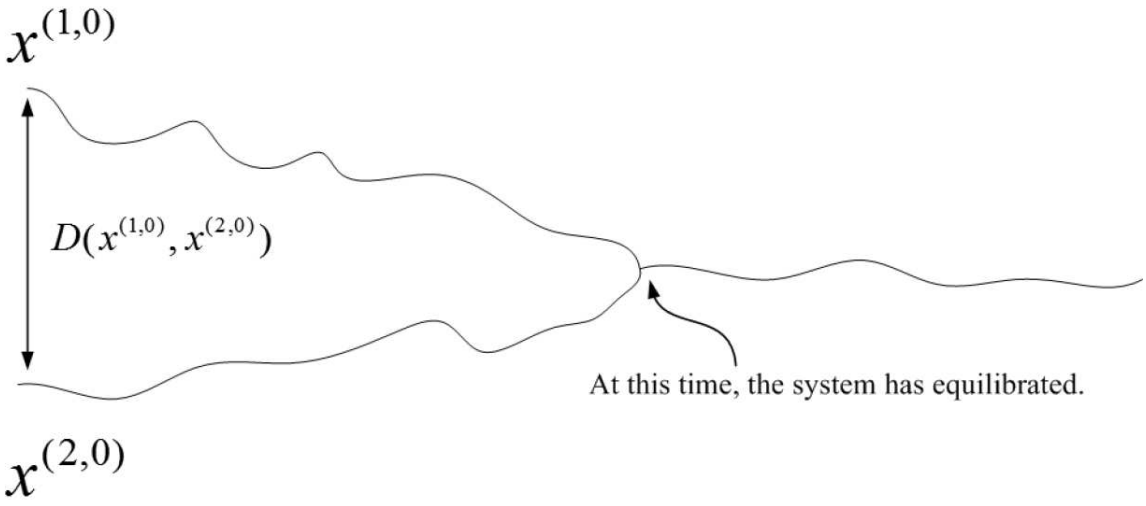


Figure 3.1: The mental picture for path coupling.

for some $\beta < 1$, then

$$\begin{aligned} & \mathbb{E}\mathcal{D}(\underline{x}^{(1,t)}, \underline{x}^{(2,t)}) \leq \beta^t \cdot 2N \\ \Longrightarrow \quad & \mathbb{P}(\underline{x}^{(1,t)} \neq \underline{x}^{(2,t)}) \leq 2N\beta^t \\ \Longrightarrow \quad & \mathbb{P}(\underline{x}^{(1,t)} \neq \underline{x}^{(2,t)}) < \epsilon \quad \text{for} \quad t \geq \left(\frac{\log \frac{2N}{\epsilon}}{\log \frac{1}{\beta}} \right) \\ \Longrightarrow \quad & \tau_{mix}(\epsilon) \leq \frac{\log \frac{2N}{\epsilon}}{\log \frac{1}{\beta}} \end{aligned}$$

Refer to figure (3.1) for a picture of coupling.

3.1.3 Proof of the inequality (3.1.4)

Idea: Consider the path between $\underline{x}$ and $\underline{y}$. If we prove that each step in the path decreases in expectation by a factor β, we get inequality (3.1.4). Hence we have to prove that if $\mathcal{D}(\underline{x}, \underline{y}) = 1$, then

$$\mathbb{E}[\mathcal{D}(\underline{x}', \underline{y}')|\underline{x}, \underline{y}] \leq \beta$$

Figure (3.2) illustrates this idea.

Assume $\underline{y}$ is obtained from $\underline{x}$ by flipping the variable at i. We define the coupling in detail as follows:

- Pick a vertex j same for the two system
- Pick $z \in \{0, 1\}$ same for the two system
- Let $\pi_j(\underline{x})$ and $\pi_j(\underline{y})$ as defined before and draw $W \in [0, 1]$.

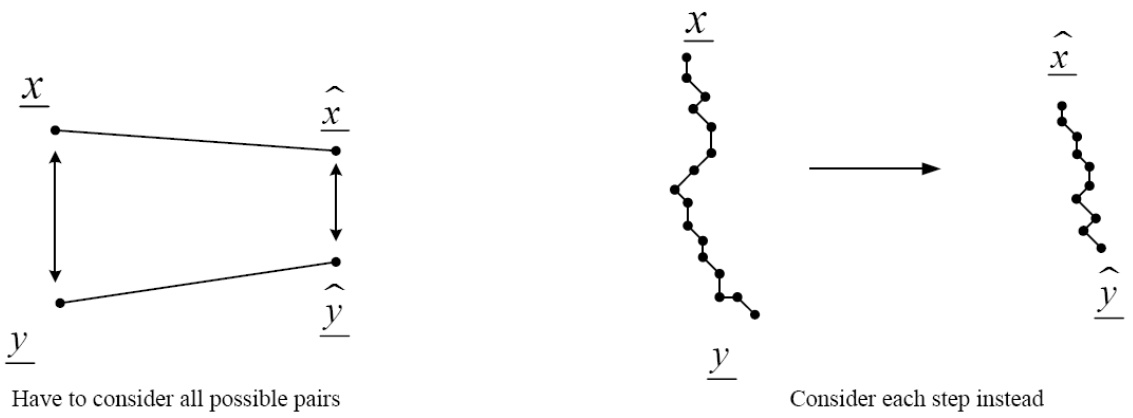


Figure 3.2: Breaking the path down into steps

- Set $x'_j = z$ if $W \leq \pi_j(\underline{x})$ and $y'_j = z$ if $W \leq \pi_j(\underline{y})$.

Let us assume for the sake of analysis that $\mathcal{G}$ has uniform degree k, and compute $\mathbb{E}\{\mathcal{D}(\underline{x}', \underline{y}')\}$. We consider three cases:

- If $j \neq i$, $j \notin \partial i$ (with probability $1 - \frac{k+1}{n}$)then $\mathcal{D}(\underline{x}', \underline{y}') = \mathcal{D}(\underline{x}, \underline{y}) = 1$.
- If $j \neq i$, $j \in \partial i$ (with probability $\frac{k}{n}$) then
  $\mathcal{D}(\underline{x}', \underline{y}') = 2$ with probability $\alpha = |\pi_j(\underline{x}) - \pi_j(\underline{y})|$.
  $\mathcal{D}(\underline{x}', \underline{y}') = 1$ otherwise.
- If $j = i$ (with probability $\frac{1}{n}$) then
  $\mathcal{D}(\underline{x}', \underline{y}') = 1$ with probability $\gamma = |\pi_j(\underline{x}) - \pi_j(\underline{y})|$
  $\mathcal{D}(\underline{x}, \underline{y}) = 0$ otherwise.

We now compute a worst-case upper bound for α. A filled circle in the pictures indicates $x_i = 1$ and an unfilled circle indicates $x_i = 0$.

In figure 3.3, x_j has to be 0 since $y_i = 0$.

$$\pi_j(\underline{x}) = \min(1, \lambda^{1-0}) = \lambda, \text{ assuming} \lambda < 1$$

$$\pi_j(\underline{y}) = 0$$

Hence $\alpha \leq \lambda$.

Similarly, we compute γ.

- Case 1: $z' = 1$.
  $\pi_j(\underline{x}) = \min(1, \lambda^{1-0}) = \lambda$, assuming $\lambda < 1$.
  $\pi_j(\underline{y}) = 1$.
- Case 2: $z' = 0$.
  $\pi_j(\underline{x}) = 1$.
  $\pi_j(\underline{y}) = \min(1, \lambda^{0-1}) = 1$, assuming $\lambda < 1$.

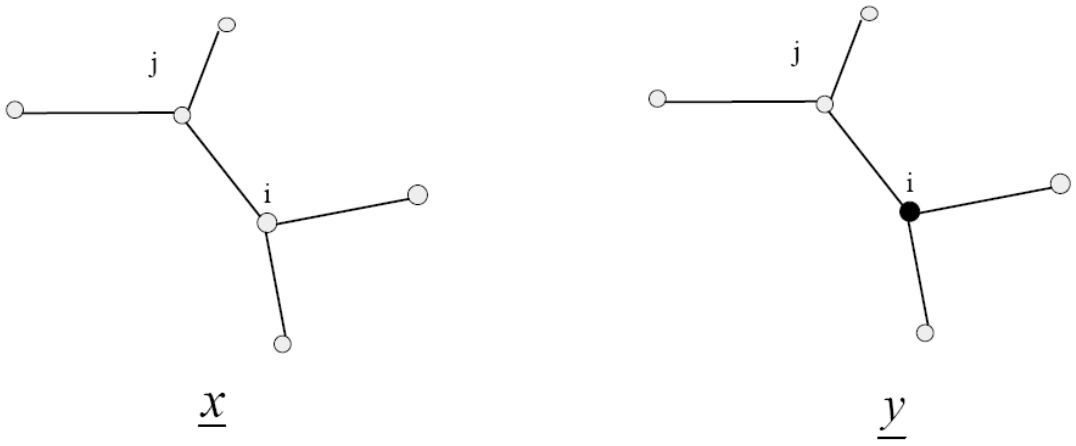


Figure 3.3: The picture for computing α

Figure 3.4: The picture for computing γ

Hence

$$\gamma \leq \frac{1}{2}(1-\lambda) + \frac{1}{2} \cdot 0 \leq \frac{1}{2}$$

Together,

$$\begin{aligned} \mathbb{E}\mathcal{D}(\underline{x}', \underline{y}') &\leq 1 \cdot \left(1 - \frac{k+1}{n}\right) + \frac{k}{n}(1-\lambda+2\lambda) + \frac{1}{n} \cdot \frac{1}{2} \\ &= 1 - \frac{1}{2n} + \frac{k\lambda}{n} \end{aligned}$$

Hence

$$\beta \leq 1 - \frac{1}{2n}(1-2k\lambda)$$

Which implies, for $\lambda < \frac{1}{2k}$, the chain is rapidly mixing.

Note that we were very lousy in computing α, we could have got $\lambda < \frac{1}{k}$

3.1.4 What happens at larger λ?

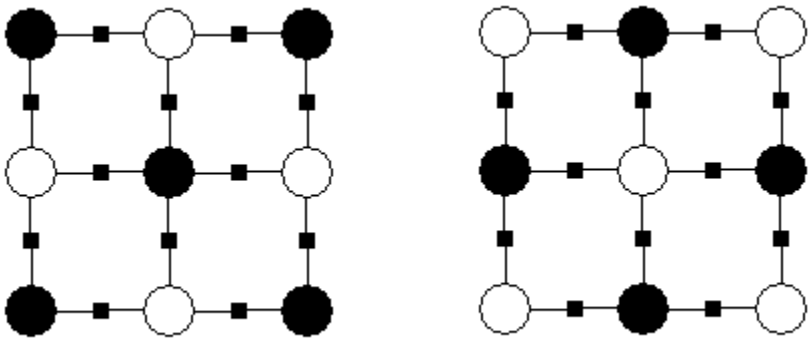


Figure 3.5: For large λ, two likely independent sets. Transitions between them are unlikely.

We did not discuss what happens when $\lambda \gg \frac{1}{k}$, nor what can be said about lower bounds of τ_{min}. In fact, we can have situations where $\tau_{min} = \exp \Theta(n)$. As an example, consider Figure (3.5), which shows a likely independent set on such a grid, and its likely complement. Transitions between the two, however, are rather improbable.

3.2 Computation tree and spatial mixing

3.2.1 Computation tree

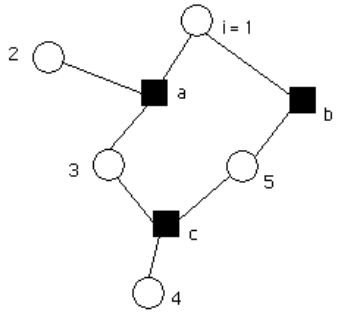


Figure 3.6: An example factor graph G.

Consider a factor graph $G = (V, F, E)$ and vertex $i \in V$ (our example graph is Figure (3.6)). The *computation tree* $T_i(G)$ is pictured below in Figure (3.7) as an example:

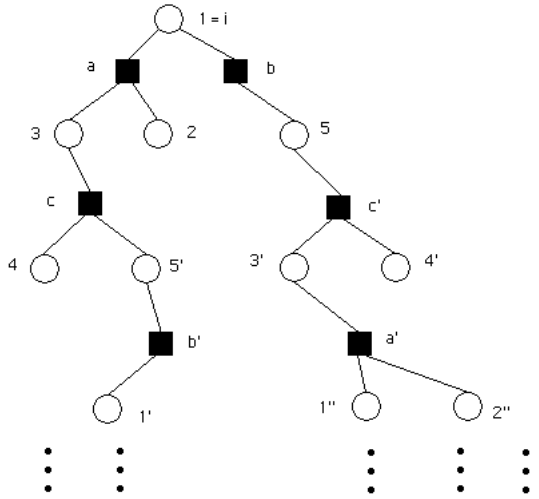


Figure 3.7: The computation tree $T_i(G)$ for the graph in Figure (3.6).

Formally, $T_i(G)$ is the tree formed by all *non-reversing* paths on G that start at i. It is endowed with a graph structure in the natural way: i is the root of the tree, and a node appears above another node in the tree iff its path is a subpath of the other node. Figure (3.6) shows two such paths, which will be neighbors in the computation tree. Note that several paths may come to be identified with the same node in the original graph – in this case, in the computation tree the node is copied and given a new label (for example, i maybe be copied to obtain i', i'', etc.).

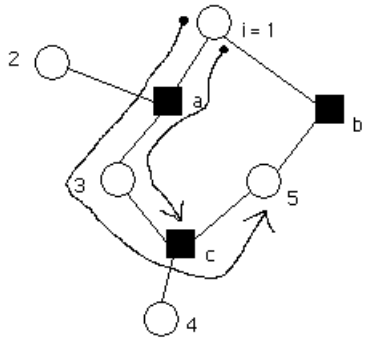


Figure 3.8: Two paths corresponding to adjacent nodes in the computation tree.

So then $T_i(G) = (V_i, F_i, E_i)$ is an *infinite* factor graph. It is a *graph covering* of G – that is, we have a mapping $\pi : V_i \to V, F_i \to F$ that is onto and such that $(j, a) \in E_i$ iff $(\pi(j), \pi(a)) \in E$.

Denote by $T_i^{(t)}(G)$ the tree obtained by truncating $T_i(G)$ after its first t generations, where in determining generations we are counting layers of variable nodes. For example, see Figure (3.9), which shows $T_i^{(2)}(G)$.

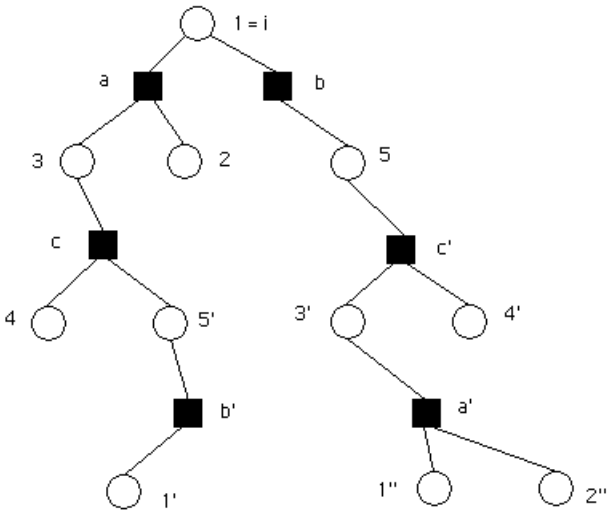


Figure 3.9: $T_i^{(2)}$, the truncated computation tree with depth 2.

With such graphs we can associate a graphical model, with a specific *boundary condition* to be applied to the bottom-most layer of variable nodes.

We are putting the same compatibility functions as before:

$$\mu^{(t_i)}(\underline{x}) = \frac{1}{Z_t} \prod_{a \in F_i} \psi_{\pi(a)}(\underline{x}_{\partial a}). \tag{3.2.1}$$

For j, a variable node at the t-th generation of $T_i^{(t)}(G)$, let a(j) be its unique adjacent function node, i.e. its parent in that rooted tree.

A boundary condition is a collection of distribution over χ into the following:

$$\{\eta_{j \to a(j)}(x_j) : j \in \partial T_i^{(t)}(G)\}. \tag{3.2.2}$$

We then define the graphical model with boundary condition corresponding to η as below:

$$\mu^{(t,i)}(\underline{x}) = \frac{1}{Z_t} \prod_{a \in T_i^{(t)}(G)} \psi_{\pi(a)}(\underline{x}_{\partial a}) \prod_{j \in \partial T_i^{(t)}(G)} \eta_{j \to a(j)}(x_j), \tag{3.2.3}$$

where $\underline{x} = \{x_j : j \in T_i^{(t)}(G)\}$.

Proposition 3.2.1. *Let $\overline{\nu}_i^{(t)}(.)$ be the BP estimate for the marginals w.r.t $\mu_G(.)$ after t BP iterations on G. If the boundary condition on $T_i^{(t_1)}(G)$ is taken to be $\nu_{j\to a(j)}^{(t_0)}(.) = \eta_{j\to a(j)}(.)$, then for $t_1 \geq 1, t_0 \geq 0$ we have*

$$\overline{\nu}_i^{(t_0+t_1)}(x_i) = \mu^{(t_1,i)}(x_i), \tag{3.2.4}$$

where $\mu^{(t_1,i)}(x_i)$ is naturally the marginal corresponding to the root in $T_i^{(t_1)}(G)$ [1].

Proof. Let j be at level $t_1 - s$ on $T_i(G)$, a its parent, and call $\mu_{j\to a}^{T(j)}(x_j)$ the marginal for x_j w.r.t the graphical model in the subtree $T(j)$ (see figure 3.10). We prove by induction that

$$\mu_{j\to a}^{T(j)}(x_j) = \overline{\nu}_{j\to a}^{(t_0+s)}(x_j). \tag{3.2.5}$$

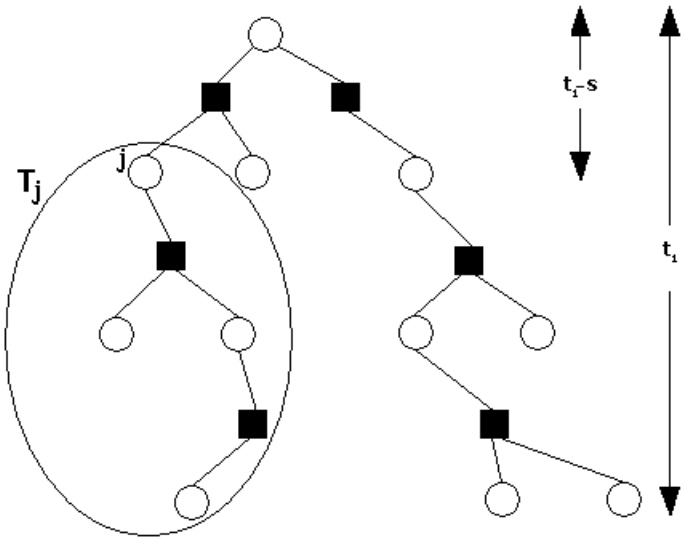


Figure 3.10: The subtree rooted at j, namely $T(j)$.

For $s = 0$ it is just a consequence of the choice of boundary conditions.

If it is true for some s, assume that the level of j is $t_1 - (s+1)$. Then by the induction hypothesis we have the following (see figure 3.11).

$$\begin{aligned}
\mu_{j\to a}^{T(j)}(x_j) &\propto \prod_{b\in\partial j\setminus a}\left[\sum_{x_{\partial b\setminus j}} \psi_b(\underline{x}_{\partial b}) \prod_{l\in\partial b\setminus j} \mu_{l\to b}^{T(j)}(x_l)\right] \\
&\propto \prod_{b\in\partial j\setminus a}\left[\sum_{x_{\partial b\setminus j}} \psi_b(\underline{x}_{\partial b}) \prod_{l\in\partial b\setminus j} \overline{\nu}_{l\to b}^{(t_0+s)}(x_l)\right] \\
&\propto \overline{\nu}_{j\to a}^{(t_0+s+1)}(x_l).
\end{aligned} \tag{3.2.6}$$

[1] Note that here one can see a slight abuse of notation, where we should use $\eta_{\pi(j)\to\pi(a(j))}$ instead of $\eta_{j\to a(j)}$ to be completely accurate. We will use the simpler -though not very accurate form- in the rest of lecture.

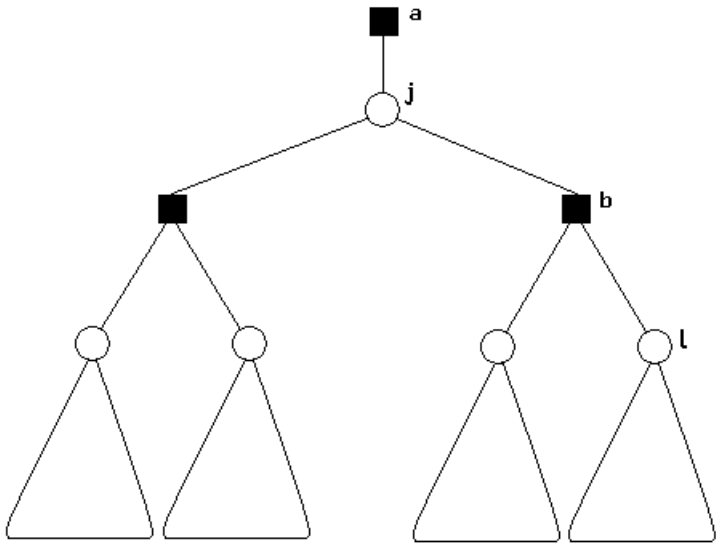


Figure 3.11: used in proof of proposition 3.2.1

Now, repeat the same argument for the root and it completes the proof. □

Corollary 3.2.2. *If*

$$\sup_{\underline{x}^{(t)},\underline{x}^{(t)'}} \left|\mu^{(t,i)}(x_i|\underline{x}^{(t)}) - \mu^{(t,i)}(x_i|\underline{x}^{(t)'})\right| \leq \delta(t) \tag{3.2.7}$$

then, for any $t_1, t_2 \geq t$ *we have*

$$\left|\overline{\nu}_i^{(t_1)}(x_i) - \overline{\nu}_i^{(t_2)}(x_i)\right| \leq \delta(t). \tag{3.2.8}$$

In particular, if $\delta(t) \to 0$ *as* $t \to \infty$*, then belief propagation converges.*

Proof. By the previous proposition we have:

$$\begin{aligned}
&\left|\overline{\nu}_i^{(t_1)}(x_i) - \overline{\nu}_i^{(t_2)}(x_i)\right| = \\
= &\left|\sum_{\underline{x}^{(t)}} \mu^{(t)}(x_i|\underline{x}^{(t)})\overline{\nu}_i^{(t_1-t)}(\underline{x}^{(t)}) - \sum_{\underline{x}^{(t)}} \mu^{(t)}(x_i|\underline{x}^{(t)})\overline{\nu}_i^{(t_2-t)}(\underline{x}^{(t)})\right| \\
= &\left|\sum_{\underline{x}^{(t)},\underline{x}^{(t)\prime}} \left[\mu^{(t)}(x_i|\underline{x}^{(t)}) - \mu^{(t)}(x_i|\underline{x}^{(t)\prime})\right] \overline{\nu}_i^{(t_1-t)}(\underline{x}^{(t)})\overline{\nu}_i^{(t_2-t)}(\underline{x}^{(t)\prime})\right| \\
= &\ \delta(t). \qquad (3.2.9)
\end{aligned}$$

□

Now, let $B_i(t)$ be the subgraph of G induced by the vertices whose distance from i is at most t.

Corollary 3.2.3. *If $B_i(t)$ is a tree and inequality 3.2.7 holds, then*

$$\left|\underbrace{\mu_i(x_i)}_{\text{actual marginal}} - \underbrace{\overline{\nu}_i^{(t)}(x_i)}_{\text{BP estimate}}\right| \leq \delta(t). \qquad (3.2.10)$$

In particular, if g is the girth[2] of G, then we will have

$$\left|\mu_i(x_i) - \overline{\nu}_i^{(t)}(x_i)\right| \leq \delta(\frac{g-1}{2}). \qquad (3.2.11)$$

Proof. Observe that

$$\mu_i(x_i) = \sum_{\underline{x}^{(t)}} \mu(x_i|\underline{x}^{(t)})\mu(\underline{x}^{(t)}), \qquad (3.2.12)$$

where $\underline{x}^{(t)}$ is the set of vertices at distance t from i in G. Also notice that $\mu_G(x_i|\underline{x}^{(t)}) = \mu_T(x_i|\underline{x}^{(t)})$. Now, proceed as in the previous proof.

□

3.3 Dobrushin uniqueness criterion

In this section we consider a general factor graph model of the type (3.1.1) on the factor graph $G = (V, F, E)$. We estabilish a general condition that is easy to check and implies correlation decay. Applied to the computation tree, this approach allow to obtain sufficient conditions for the convergence of belief propagation, and for its correctness on locally tree-like graphs.

The condition developed here was initially proposed by Roland Dobrushin in his study of Gibbs measures. While the initial results concerned traslation invariant models on regular grids, its generalization to arbitrary factor graphs is quite immediate. Dobrushin criterion measures the strength of interactions as follows.

[2]The girth of an undirected graph G is defined by the length of the shortest cycle in it.

Definition 6. *Given vertices $i, j \in V$, the* influence *of j on i is defined as*

$$C_{ij} = \max_{\underline{x},\underline{x}'} \Big\{ \|\mu(x_i = \cdot\,|\underline{x}_{V\setminus i}) - \mu(x_i = \cdot\,|\underline{x}'_{V\setminus i})\|_{\mathrm{TV}} \;:\; x_l = x'_x \,\forall l \neq j \Big\}$$

Notice that by definition $0 \le C_{ij} \le 1$ and $C_{ij} = 0$ unless $d(i,j) = 1$ (i.e. unless i, j are neighbors of the same factor node). The following theorem shows that small influence implies correlation decay. Here, for a vertex i, we let $\mathsf{B}_i(t)$ denote the ball of radius t around i, and $\overline{\mathsf{B}}_i(t)$ its complement.

Theorem 3.3.1 (Dobrushin, 1968)**.** *Assume*

$$\gamma \equiv \sup_{i\in V} \Big(\sum_{j\in V\setminus i} C_{ij} \Big) .$$

Then for any $i \in V$, any $t \ge 0$, letting $\overline{\mathsf{B}} = \overline{\mathsf{B}}_i(t)$, we have

$$\max_{\underline{x},\underline{x}'} \|\mu(x_i = \cdot\,|\underline{x}_{\overline{\mathsf{B}}}) - \mu(x_i = \cdot\,|\underline{x}'_{\overline{\mathsf{B}}})\|_{\mathrm{TV}} \le \frac{\gamma^t}{1-\gamma} .$$

In other words, this theorem estabilishes that correlations decay exponential in the graph distance with a rate that is controlled by the parameter γ defined in the statement.

Proof. With a slight abuse of notation, let us denote by μ, μ' the conditional measure on $\underline{x}_{\mathsf{B}_i(t)}$ given configurations $\underline{x}_{\overline{\mathsf{B}}}$, $\underline{x}'_{\overline{\mathsf{B}}}$ on $\overline{\mathsf{B}}_i(t)$. We will construct a coupling $\widehat{\mu}$ of μ, μ' such that

$$\widehat{\mu}(x_i \neq x'_i) \le \frac{\gamma^t}{1-\gamma} .$$

This coupling can be constructed recursively as follows:

1: Start from $\widehat{\mu} = \mu \times \mu'$.

2: Repeat

(a) Choose $j \in \mathsf{B}_i(t)$;

(b) Define a new coupling $\widehat{\mu}^{\mathrm{new}}$ as follows:

(b1) Sample $\underline{x}, \underline{x}' \sim \widehat{\mu}$;

(b2) Resample (x_j, x'_j) from the optimal coupling between $\mu(x_j|\underline{x}_{V\setminus j})$ and $\mu(x_j|\underline{x}'_{V\setminus j})$;

(c) Let $\widehat{\mu} \leftarrow \widehat{\mu}^{\mathrm{new}}$;

The vertex j in step (a) is selected in such a way that the vertices in $\mathsf{B}_i(t)$ are swept over.

The proof consists in analyzing this recursion. Let $\widehat{\mu}^{(k)}$ be the coupling produced after k sweeps, and assume we are given numbers $a_j^{(k)}$ such that for all $j \in \mathsf{B}_i(t)$

$$\widehat{\mu}^{(k)}(x_j \neq x'_j) \le a_j^{(k)} .$$

Then it is easy to show that at the next iteration analogous upper bounds can be obtained by letting

$$a_j^{(k+1)} = \sum_{l\in V\setminus j} C_{jl} a_l^{(k)} , \tag{3.3.1}$$

where $a_l^{(k)} = 1$ by definition for $l \in \overline{\mathsf{B}}_i(t)$. The thesis follows by controlling the recursion (3.3.1). [**A few more details to be written.**] □

Chapter 4

Variational methods

The basic idea in variational approaches to formulate the inference problem as an optimization problem. Once this is done, all sorts of ideas from standard optimization theory can be applied to solve or approximately solve the optimization problem. It turns out that the whole approach is particularly simple and clean when (guess what) the inference problem was an optimization problem to start with, e.g. in the mode computation.

The whole idea of connecting inference (more precisely, computation of expected values of high-dimensional probability distributions) to optimization dates back to physics. It is a known fact in thermodynamics that systems in thermal equilibrium with their environment tend to optimize their free energy. A mathematical foundation for this physical law was provided by statistical mechanics with Gibbs' variational principle. The implications of these ideas to algorithms and graphical models where first discussed by Yedidia and coworkers [YFW05], and subsequently extended in a number of ways (see for instance [WJW05] and the review [WJ08]).

This chapter is organized as follows. The basic connection between inference and optimization is introduced in Section 4.1. One basic way to use this connection is the so-called naive mean field approximation, discussed in Section 4.2. We then develop Bethe approximation to the Gibbs free energy and show its connection to belief propagation in Section 4.3. Finally, we explain two ways of modifying Bethe approximation, namely region-based approximations in Sections 4.4 and 4.6, together with the algorithms based on these approximations.

4.1 Free Energy and Gibbs Free Energy

Throughout this chapter we shall use the graphical model formalism and hence assume that a model $(G, \underline{\psi})$ is given, with $G = (V, F, E)$ a factor graph and $\underline{\psi} = \{\psi_a\}_{a \in F}$ a collection of compatibility functions. We define the joint distribution as

$$\mu(\underline{x}) = \frac{1}{Z} \prod_{a \in F} \psi_a(\underline{x}_{\partial a}) \, .$$

The concepts discussed in this section do not depend on the factorization of μ accordinbg to G but only on the total weight $\psi(\underline{x}) \equiv \prod_{a \in F} \psi_a(\underline{x}_{\partial a})$ We will therefore consider a generic un-normalized weigth $\psi : \mathcal{X}^V \to \mathbb{R}_+$, and use this weight to compute a probability distribution

$$\mu(\underline{x}) = \frac{1}{Z} \psi(\underline{x}) \, . \tag{4.1.1}$$

We begin define the Helmotz free energy and Gibbs free energy of a distribution.

Definition 7 (Helmotz Free Energy)**.** *Given a model $(G, \underline{\psi})$, its Helmotz free energy is defined as*

$$\Phi = \log Z = \log \Big\{ \sum_{\underline{x}} \prod_{a \in \mathcal{F}} \psi_a(\underline{x}_{\partial a}) \Big\} .$$

A few remarks are in order. First of all, physicists normally call 'free energy' the quantity[1] $-\log Z$, but the minus sign is usually misterious for non-physicists and we will therefore drop it. Second, the 'Helmotz' qualifier is not that common. We will drop it most of the times. Sometimes, the same quantity is called mor simply log-partition function.

The importance of Φ can be understood by recalling that computing marginals of μ is equivalent to computing differences $\Phi - \Phi'$ for modified models. Also sampling from μ can be reduced to computing Φ for a sequence of models.

We next introduce the notion of Gibbs free energy. Again, the factorization structure is not crucial here, but we can consider any model of the form (4.1.1). (Recall that $H(p)$ is Shannon's entropy of the distribution p.)

Definition 8 (Gibbs free energy)**.** *Given a model $(G, \underline{\psi})$, its Gibbs free energy is a function $\mathbb{G} : \mathsf{M}(\mathcal{X}^V) \to \mathbb{R}$. For a distribution $\nu \in \mathsf{M}(\mathcal{X}^V)$, the corresponding Gibbs free energy is defined as*

$$\begin{aligned} \mathbb{G}(\nu) &= H(\nu) + \mathbb{E}_\nu \log \psi(\underline{x}) && (4.1.2) \\ &= -\sum_{\underline{x}} \nu(\underline{x}) \log \nu(\underline{x}) + \sum_{\underline{x}} \nu(\underline{x}) \log \psi(\underline{x}) . && (4.1.3) \end{aligned}$$

The connection between these concepts is provided by the following result.

Proposition 4.1.1. *The function $\mathbb{G} : \mathsf{M}(\mathcal{X}^V) \to \mathbb{R}$ is strictly concave on the convex domain $\mathsf{M}(\mathcal{X}^V)$. Further its unique maximum is achieved at $\nu = \mu$, with $\mathbb{G}(\mu) = \Phi$.*

Proof. Convexity is just a consequence of the fact that $z \mapsto z \log z$ is convex on $\mathbb{R}_+$. The unique minimum can be found by introducing the Lagrangian

$$\mathcal{L}(\nu, \lambda) = \mathbb{G}(\nu) - \lambda \left(\sum_{\underline{x}} \nu(\underline{x}) - 1 \right) . \qquad (4.1.4)$$

Setting to zero the derivative with respect to $\nu(\underline{x})$, one gets $\nu(\underline{x}) = \psi(\underline{x})/Z$. It is immediate to check that the value at the minimum is indeed Φ. □

Remark 2. *Gibbs free energy can be re-expressed as follows:*

$$G[\nu] = \Phi - D(\nu || \mu)$$

where $D(\,\cdot\,)$ is the Kullback-Leibler divergence between ν and μ [CT91].

[1] More precisely, they call free enregy the quantity $-$(Temp.) $\times \log Z$ with Temp. the system temperature.

We therefore reformulated the inference problem as a convex optimization problem, and we know that convex optimization is tractable! The problem of course is that the search space is very high-dimensional. The dimension of $\mathsf{M}(\mathcal{X}^V)$ is $(|\mathcal{X}|^n - 1)$ which scales exponentially in the model size. For reasonably large models even storing such a vector is impossible.

The basic idea of variational inference is look for approximate solutions of this problem. The following tricks are at the core of most approach: (i) Optimize $\mathbb{G}(\nu)$ only within a low-dimensional subset of $\mathsf{M}(\mathcal{X}^V)$; (ii) Construct low-dimensional projections of $\mathsf{M}(\mathcal{X}^V)$, and approximate $\mathbb{G}(\nu)$ at a point ν by some function of the projection.

4.2 Naive mean field

Naive mean field amounts to applying trick (i) in the above list, whereby the subset is just the set of measure of product form, i.e. such that

$$\nu(\underline{x}) = \prod_{i \in V} \nu_i(x_i)\,. \tag{4.2.1}$$

If we write $\underline{\nu} = (\nu_1, \nu_2, \dots, \nu_n) \in \mathsf{M}(\mathcal{X}) \times \cdots \times \mathsf{M}(\mathcal{X}) \equiv \mathsf{M}(\mathcal{X})^V$ for a vector of one-variable marginals, we can denote formally the above embedding of factorized distributions into $\mathsf{M}(\mathcal{X}^V)$, $\mathcal{K} : \mathsf{M}(\mathcal{X})^V \to \mathsf{M}(\mathcal{X}^V)$, by

$$\mathcal{K}(\underline{\nu})(\underline{x}) = \prod_{i \in V} \nu_i(x_i)\,. \tag{4.2.2}$$

The Naive Mean Field free energy is then the function $\mathbb{G}_{\mathrm{MF}} : \mathsf{M}(\mathcal{X})^V \to \mathsf{M}(\mathcal{X}^V)$ defined by

$$\mathbb{G}_{\mathrm{MF}}(\underline{\nu}) = \mathbb{G}(\mathcal{K}(\underline{\nu}))\,. \tag{4.2.3}$$

Plugging this in the definition of Gibbs free energy , we obtain the following explicit expression:

$$\begin{aligned} \mathbb{G}_{\mathrm{MF}}(\underline{\nu}) &= H(\mathcal{K}(\nu)) + \mathbb{E}_\nu[\log \psi(\underline{x})] \\ &= \sum_{i \in V} H(\nu_i) + \sum_{a \in F} \sum_{\underline{x}_{\partial a}} \prod_{i \in \partial a} \nu_i(x_i) \log \psi_a(\underline{x}_{\partial a})\,. \end{aligned}$$

It is an immediate consequence of Proposition 4.1.1 that the naive mean field approximation provides a lower bound on the free energy, namely

$$\Phi \geq \max_{\underline{\nu} \in \mathsf{M}(\mathcal{X})^V} \mathbb{G}_{\mathrm{MF}}(\underline{\nu}) \tag{4.2.4}$$

The nice fact about this expression is that the resulting optimization problem has much lower dimensionality than for Gibbs variational principle: we passed from $(|\mathcal{X}|^n - 1)$ to $n(|\mathcal{X}| - 1)$. The price to pay is that $\mathbb{G}_{\mathrm{MF}}$ is no longer a concave function and therefore solving the optimization problem can be hard.

It is instructive to write down the stationarity conditions for $\mathbb{G}_{\mathrm{MF}}$. We introduce Lagrange multipliers λ_i to constrain the beliefs ν_i to be normalized, that is, $\sum_{x_i} \nu_i(x_i) = 1$, giving us the

following Lagrangian and derivatives:

$$\begin{aligned}\mathcal{L}(\underline{\nu},\lambda) &= \sum_{i\in V} H(\nu_i) + \sum_{a\in F}\sum_{\underline{x}_{\partial a}}\prod_{i\in\partial a}\nu_i(x_i)\log\psi_a(\underline{x}_{\partial a}) + \sum_{i\in V}\lambda_i\left(\sum_{x_i}\nu_i(x_i)-1\right),\\ \frac{\partial\mathcal{L}}{\partial\nu_i(x_i)} &= -1-\log\nu_i(x_i) + \sum_{a\in\partial i}\sum_{x_j:j\in\partial a\setminus i}\prod_{j\in\partial a\setminus i}\nu_j(x_j)\log\psi_a(\underline{x}_{\partial a})+\lambda_i\,.\end{aligned}$$

Solving, we obtain the stationarity conditions

$$\begin{aligned}\nu_i(x_i) &= \exp\left(\sum_{a\in\partial i}\sum_{x_j:j\in\partial a\setminus i}\prod_{j\in\partial a\setminus i}\nu_j(x_j)\log\psi_a(\underline{x}_{\partial a})+\lambda_i-1\right)\\ &\cong \exp\left(\sum_{a\in\partial i}\sum_{x_j:j\in\partial a\setminus i}\log\psi_a(\underline{x}_{\partial a})\prod_{j\in\partial a\setminus i}\nu_j(x_j)\right) \qquad (4.2.5)\end{aligned}$$

where we solved for λ_i to normalize the ν's. These are somewhat more transparent if we introduce messages $\widehat{\nu}_{a\to i}(x_i)$ thus getting the following equations

$$\nu_i(x_i) \cong \prod_{a\in\partial i}\widehat{\nu}_{a\to i}(x_i)\,, \qquad (4.2.6)$$

$$\widehat{\nu}_{a\to i}(x_i) \cong \exp\left(\sum_{x_j:j\in\partial a\setminus i}\log\psi_a(\underline{x}_{\partial a})\prod_{j\in\partial a\setminus i}\nu_j(x_j)\right). \qquad (4.2.7)$$

A simple greedy algorithm for finding a stationary point consists in updating the ν's by iterating the above equations until convergence. Of course, standard first orded methods can be used as well to reach a stationary point of $\mathbb{G}_{\rm MF}$.

4.2.1 Pairwise graphical models and the Ising model

Consider the special case of a pairwise model

$$\mu(\underline{x}) = \frac{1}{Z}\prod_{(i,j)\in E}\psi_{ij}(x_i,x_j)\prod_{i\in V}\psi_i(x_i)\,. \qquad (4.2.8)$$

Then the naive mean field equations read

$$\nu_i(x_i) \cong \psi_i(x_i)\exp\left(\sum_{j\in\partial i}\sum_{x_j}\log\psi_{ij}(x_i,x_j)\nu_j(x_j)\right). \qquad (4.2.9)$$

As an example we will consider again Ising models, which we recall are pairwise graphical models with binary variables $x_i\in\mathcal{X}=\{+1,-1\}$. For the sake of simplicity we shall assume G to be a d-dimensional discrete torus of linear size L. This is obtained by letting the vertex set be $V=\{1,\dots,L\}^d\subseteq\mathbb{Z}^d$, and the edge set $(i,j)\in E$ if and only if $j=(i\pm e_k) \mod L$ where e_k, $k=1,\dots,d$ is the canonical basis. This graph is reproduced in Fig. **??**, for $d=2$ dimensions.

To have complete symmery among vertices, we also assume that potentials are all equal and that they are symmetric ander change of sign of the variables, i.e.

$$\psi_{ij}(x_i, x_j) = e^{\beta x_i x_j}, \quad \text{i.e.} \quad \psi_{\text{ij}} = \begin{pmatrix} e^{\beta} & e^{-\beta} \\ e^{-\beta} & e^{\beta} \end{pmatrix}. \tag{4.2.10}$$

This defines the joint probability distribution

$$\mu(\underline{x}) = \frac{1}{Z} \exp\Big\{\beta \sum_{(i,j)\in E} x_i, x_j\Big\}. \tag{4.2.11}$$

The mean field free energy takes the form

$$\mathbb{G}_{\text{MF}}(\underline{\nu}) \;\; = \;\; \sum_{i\in V} H(\nu_i) + \beta \sum_{(i,j)\in E} \sum_{x_i x_j} \nu_i(x_i)\nu_j(x_j) x_i x_j \,. \tag{4.2.12}$$

Since variables are binary, a single real number is sufficient to pearametrize each marginal. We chose this number to be the expectation:

$$\nu_i(x_i) = \frac{1 + m_i x_i}{2}, \qquad \mathbb{E}_{\nu_i}[x_i] = m_i \,.$$

We have of course $m_i \in [-1, +1]$. By substituting in Eq. (4.2.12), we get the explicit expression (with some abuse of notation)

$$\mathbb{G}_{\text{MF}}(\underline{m}) \;\; = \;\; = \sum_{i\in V} h\big((1+m_i)/2\big) + \beta \sum_{(i,j)\in\Gamma} m_i m_j \,.$$

where $h(x) \equiv -x \log x - (1-x)\log(1-x)$.

The mean field equationtion (4.2.9) reduces to

$$\nu_i(x_i) \cong \exp\Big\{\beta \sum_{j\in\partial i} x_i \sum_{x_j} \nu_j(x_j) x_j\Big\} \tag{4.2.13}$$

and we thus have the fllowing equations for the means

$$\begin{aligned} m_i \;\; &= \;\; \frac{\exp\Big(\beta \sum_{j\in\Gamma(i)} m_j\Big) - \exp\Big(-\beta \sum_{j\in\Gamma(i)} m_j\Big)}{\exp\Big(\beta \sum_{j\in\Gamma(i)} m_j\Big) + \exp\Big(-\beta \sum_{j\in\Gamma(i)} m_j\Big)} \\ &= \;\; \tanh\left(\beta \sum_{j\in\Gamma(i)} m_j\right). \end{aligned} \tag{4.2.14}$$

Recall the hyperbolic tangent is a strictly increasing function bounded by $[-1, 1]$.

Convergence of iterative mean field updates for Ising models

In this section, we will be considering the convergence properties of the iterative updates we have derived for our mean-field approximation of Ising models. More precisely, we shall consider the iteration

$$m_i^{(t+1)} \quad = \quad \tanh\left(\beta \sum_{j\in\partial i} m_j^{(t)}\right). \tag{4.2.15}$$

It turns that this iteration always converges if a uniform initialization is used. On the other hand, the accuracy of the result depends strongly on the strength of interactions which is tuned via the parameter β.

First, it follows from Eq. (4.2.15) that, if we initialize $m_i^{(0)} = m^{(0)}$ for all $i \in V$, we will have $m_i^{(t)} = m^{(t)}$ for all $t \geq 0$. This common value evolves according to the one-dimensional recursion

$$m^{(t+1)} \quad = \quad \tanh\left(2d\beta m^{(t)}\right). \tag{4.2.16}$$

This recursion can be easily studied graphically, cf. Fig. 4.1. The function $x \mapsto \tanh(x)$ is odd and has maximum derivative at $x = 0$, where $\tanh(x) = x + O(x^3)$. Hence for $\beta < \frac{1}{2d}$, the slope of the mapping in Eq. (4.2.16) is everywhere smaller than 1, and therefore the mapping is a contraction. No matter the value of $m^{(0)}$, the iterations will converge to $\lim_{t\to\infty} m^{(t)} = 0$. An illustration of this is in Fig. 4.1(a).

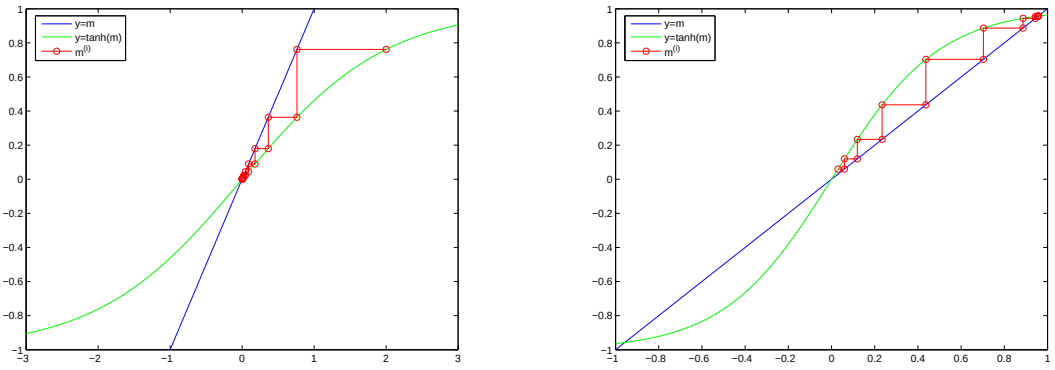
If $\beta > \frac{1}{2d}$, on the other hand, the slope of the hyperbolic tangent will be greater than 1 at $m = 0$, and there are three points that the line $y = m$ will intersect $y = \tanh(\beta \sum m)$. Thus, if we begin the iterations with $m^{(0)} > 0$, $\lim_{t\to\infty} m_i^{(t)} = +m^*$. If $m^{(0)} < 0$, then the iterations will converge to $-m^*$, and in the degenerate case that $m^{(0)} = 0$, we will converge to the saddlepoint 0. An illustartion of this is in Fig. 4.1(b).

Marginals in the Ising Model

Up ot this point, we have ignored the marginals of the actual distribution of the Ising model, focusing instead on the partition function and maximizing the free energy. Is the naive mean field approximation reproducing the correct marginals.

It is easy to see that the real marginals $\mu_i(x_i)$ should be uniform i.e. $\mu_i(+1) = \mu_i(-1) = 1/2$, because the factors are completely symmetric under exchange of signs of the variables $\underline{x}$. In other words for each configuration $\underline{x}$, te flipped configuration $-\underline{x}$ has exactly the same probability under μ. This apperars to be captured by the naive mean field approximation only for 'weak' factors, namely for $\beta \leq 1/(2d)$.

For $\beta > 1/(2d)$, the fixed point $m = 0$ becomes unstable, and the mean field iteration converges to the either of the fixed points $m = \pm m^*$. This appears at first sight completely incorrect, and just a pathology of the naive mean field approximation. In fact this is not the hole story. At large enough β, we have that the model is equally likely to have most of its variables either $+1$ or -1. The distribution μ becomes strongly bimodal. This can be revealed for instance by considering the sum of variables $\sum_{i\in V} x_i$. For $d \geq 2$ and any $\beta > \beta_{rmc}(d)$ the distribution of this quantity concentrates around values $\pm m_0|V|$ as $L \to \infty$. This is a typical 'phase transition' phenomenon. Naive mean field



(a) (b)

Figure 4.1: (a) Updates to $m^{(t)}$ with $\beta < \frac{1}{2d}$. Path is down and left. (b) Updates to $m^{(t)}$ with $\beta > \frac{1}{2d}$. Path is up and right.

captures this phenomenon with te appearence of two fixed points, although this is only a 'cartoon' of the actual behavior.

Nevertheless this cartoon is more than a simple coincidence. This can be understood considering the joint distribution of multiple variables. Within naive mean field this is postulated to factorize. For $\beta \le \beta_c(d)$ the actual joint distribution does not factorize, but it does approximately for far apart vertices. More explicitly, if we consider $L \gg 1$ and vertices i, j, k far apart in the graph, we have

$$\mu(x_i, x_j, x_k) \approx \mu(x_i)\mu(x_j)\mu(x_k)\,. \tag{4.2.17}$$

For $\beta > \beta_c(d)$ the actual joint distribution does not factorize even for far apart vertices. Nevertheless, it is asymptotically well approximated by a convex combination of two factorized distributions:

$$\mu(x_i, x_j, x_k) \approx \frac{1}{2}(\mu^+(x_i)\mu^+(x_j)\mu^+(x_k) + \mu^-(x_i)\mu^-(x_j)\mu^-(x_k)).$$

Here μ^+ corresponds to positive configurations and μ^- to negative ones.

4.3 Bethe Free Energy

The naive mean field approach approximates the joint distribution of $\underline{x} = (x_1, x_2, \dots, x_n)$ with the product of one-variable marginals:

$$\mu(\underline{x}) \;\approx\; \prod_{i\in V} \mu_i(x_i)\,.$$

In other words, variables x_i and x_j are treated as independent for $i \neq j$. There are cases in which this is grossly incorrect. In particular, it neglects strong correlations among variables in $x_{\partial a}$ that are adjacent to a common factor node a.

As an example, consider the graphical model with a single factor node a and three variables x_1, x_2, $x_3 \in \{0, 1\}$, whose joint distribution is given by

$$\mu(x_1, x_2, x_3) = \frac{1}{Z}\, \mathbb{I}(x_1 \oplus x_2 \oplus x_3 = 0)\,.$$

Here $\oplus$ denotes sum modulo 2, and we obviously have $Z = 4$. It is a simple exercise to show that the naive mean field yields the disappointing estimate $Z_{\rm MF} = 1$.

Bethe-Peierls approximation improves in a crucial way, by accounting for correlations induced by factor nodes. Instead of parametrizing the entire distribution in terms of single variable marginals $\mu_i(x_i)$. Bethe-Peierls approximation instead keeps track of joint distributions $\mu_a(\underline{x}_a)$. We will start by building intuition in the case of tree factor graphs, then define the Bethe free energy in the case of general graphs, and finally discuss the connection with belief propagation.

4.3.1 The case of tree factor graphs

The variational approximation we are going to construct will approximate the Gibbs free energy as a function of single variable marginals $\mu_i(x_i)$, and of joint distributions at a factor node $\mu_a(\underline{x}_a)$. Valuable insight can be gained by considering the case of tree factor graphs. In this case we have the following important structural result.

Lemma 4.3.1. *If $(G, \underline{\psi})$ is a tree factor graph model, then the corresponding joint distribution μ is given by*

$$\begin{aligned} \mu(\underline{x}) &= \prod_{a\in F} \left(\frac{\mu_a(\underline{x}_{\partial a})}{\prod_{i\in\partial a}\mu_i(x_i)} \right) \prod_{i\in V} \mu_i(x_i) \qquad (4.3.1)\\ &= \prod_{a\in F} \mu_a(\underline{x}_{\partial a}) \prod_{i\in V} \mu_i(x_i)^{1-|\partial i|}\,. \end{aligned}$$

Proof. The proof will proceed by induction on $|F|$. For the base case, assume $|F| = 0$. The expression (4.3.1) then reduces to $\mu(\underline{x}) = \prod_{i\in V}\mu_i(x_i)$ which is correct, since for an empty factor graph the variables $x_1, \ldots, x_n$ are independent.

Now assume that the claim holds for all tree graphs such that $|F| \le m$. We need to now check the expression (4.3.1) for $|F| = m + 1$. Consider a tree G with m factor nodes, and at variable node i, we append a factor node a (along with other variable nodes connected to a), to get a tree G' with $m + 1$ factor nodes. Indeed any graph G' eith $m + 1$ factor nodes can be decomposed in this way, see Fig. 4.2.

Note that $|\partial i|$ is the number of factor nodes connected to variable node i in graph G'. Thus number of factor nodes connected to i in graph G is $|\partial i| - 1$. Further notice that the joint distribution of the variables in V, factors according to the graph G. We therefore have, by the induction hypothesis,

$$\mu(\underline{x}_V) = \prod_{b\in G} \mu_b(\underline{x}_{\partial b}) \prod_{j\in G\setminus i} \mu_j(x_j)^{1-|\partial j|} \mu_i(x_i)^{1-(|\partial i|-1)}\,.$$

By Bayes rule $\mu(\underline{x}_{V'}) = \mu(\underline{x}_V)\mu(\underline{x}_{\partial a\setminus i}|\underline{x}_V)$. On the other hand the vertex i separates $\partial a \setminus i$ from $V \setminus \{i\}$. By the global Markov property we have

$$\mu(\underline{x}_{\partial a\setminus i}|\underline{x}_V) = \mu(\underline{x}_{\partial a\setminus i}|x_i)\,.$$

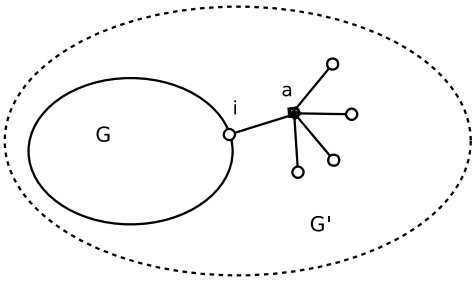


Figure 4.2: Addition of a single factor node to graph G

Using therefore the induction hypothesis we get

$$\begin{aligned}
\mu(\underline{x}_{V'}) &= \mu(\underline{x}_V)\mu(\underline{x}_{\partial a\setminus i}|x_i) \\
&= \prod_{b\in G}\mu_b(\underline{x}_{\partial b})\prod_{j\in G\setminus i}\mu_j(x_j)^{1-|\partial j|}\mu_i(x_i)^{1-(|\partial i|-1)}\frac{\mu(\underline{x}_{\partial a\setminus i},x_i)}{\mu_i(x_i)} \\
&= \prod_{b\in G'}\mu_b(\underline{x}_{\partial b})\prod_{j\in G'}\mu_j(x_j)^{1-|\partial j|}\,,
\end{aligned}$$

where the last equality follows from the fact that $\mu(\underline{x}_{\partial a\setminus i},x_i)=\mu_a(\underline{x}_{\partial a})$ and for variable nodes $j\in\partial a\setminus i$ the number of neighbouring factor nodes $|\partial j|=1$. This completes the induction step, thereby proving the lemma. □

As an example, consider a Markov chain. The expression for the distribution now reduces to

$$\begin{aligned}
\mu(\underline{x}) &= \prod_{a=1}^{n-1}\mu_a(x_a,x_{a+1})\prod_{i=2}^{n-1}\mu_i(x_i)^{-1} \\
&= \mu_1(x_1,x_2)\prod_{i=2}^{n-1}\frac{\mu_i(x_i,x_{i+1})}{\mu_i(x_i)} \\
&= \mu_1(x_1)\mu_1(x_2|x_1)\prod_{i=2}^{n-1}\mu_i(x_{i+1}|x_i) \\
&= \mu_1(x_1)\prod_{i=1}^{n-1}\mu_i(x_{i+1}|x_i)
\end{aligned}$$

Thus we obtain the usual formula for the distribution of a Markov Chain.

Calculating the Gibbs energy for trees, one obtains,

$$\begin{aligned}
\mathbb{G}(\mu) &= H(\mu) + \sum_{a\in F} \mathbb{E}_\mu \log(\psi_a(\underline{x})) \\
&= -\sum_{\underline{x}} \mu(\underline{x}) \log \mu(\underline{x}) + \sum_{a\in F} \sum_{\underline{x}_{\partial a}} \mu_a(\underline{x}_{\partial a}) \log \psi_a(\underline{x}_{\partial a}) \\
&= -\sum_{\underline{x}} \mu(\underline{x}) \log \left(\prod_{a\in F} \mu_a(\underline{x}_{\partial a}) \prod_{i\in V} \mu_i(x_i)^{1-|\partial i|} \right) + \sum_{a\in F} \sum_{\underline{x}_{\partial a}} \mu_a(\underline{x}_{\partial a}) \log \psi_a(\underline{x}_{\partial a})) \\
&= -\sum_{a\in F} \sum_{\underline{x}_{\partial a}} \mu_a(\underline{x}_{\partial a}) \log \mu_a(\underline{x}_{\partial a}) - \sum_{i\in V} \sum_{x_i} (1-|\partial i|) \mu_i(x_i) \log \mu_i(x_i) + \sum_{a\in F} \sum_{\underline{x}_{\partial a}} \mu_a(\underline{x}_{\partial a}) \log \psi_a(\underline{x}_{\partial a}) \\
&= \sum_{a\in F} H(\mu_a) - \sum_{i\in V} (|\partial i| - 1) H(\mu_i) + \sum_{a\in F} \sum_{\underline{x}_{\partial a}} \mu_a(\underline{x}_{\partial a}) \log \psi_a(\underline{x}_{\partial a})
\end{aligned}$$

Here $H(\,\cdot\,)$ is the Shannon entropy. We therefore proved the following.

Corollary 4.3.2. *If $(G, \underline{\psi})$ is a tree factor graph model, then the corresponding Gibbs free energy is*

$$\mathbb{G}(\mu) = \sum_{a\in F} H(\mu_a) - \sum_{i\in V} (|\partial i| - 1) H(\mu_i) + \sum_{a\in F} \sum_{\underline{x}_{\partial a}} \mu_a(\underline{x}_{\partial a}) \log \psi_a(\underline{x}_{\partial a})$$

4.3.2 General graphs and locally consistent marginals

For the general case, one can have marginals $\{b_i, b_a\}$ for each factor node $a \in F$ and variable node $i \in V$, which define the Gibbs free energy. We see that such marginals have to satisfy the following conditions to be a valid set of marginals.

$$\begin{aligned}
b_i(x_i) &\geq 0, \quad \forall x_i, \quad \forall i \in V \\
b_a(\underline{x}_{\partial a}) &\geq 0, \quad \forall \underline{x}_{\partial a}, \quad \forall a \in F && (4.3.2) \\
\sum_{x_i} b_i(x_i) &= 1, \quad \forall i \in V && (4.3.3) \\
\sum_{\underline{x}_{\partial a \setminus i}} b_a(\underline{x}_{\partial a}) &= b_i(x_i), \quad \forall a \in F, i \in \partial a && (4.3.4) \\
\sum_{\underline{x}_{\partial a}} b_a(\underline{x}_{\partial a}) &= 1, \quad \forall a \in F. && (4.3.5)
\end{aligned}$$

Note that Eqn(4.3.3) and Eqn(4.3.4) imply Eqn(4.3.5), and hence we can remove that equation from the requirements of a valid marginal.

The set of marginals $\{b_i, b_a\}$ on a graph G which satisfy the above requirements is known as the set of *locally consistent marginals* on G, denoted by $LOC(G)$. As the above constraints are linear, it can be readily seen that the set $LOC(G)$ is a convex set. Also the dimension of $LOC(G)$ is bounded above by $|\mathcal{X}|^{\max|\partial a|}|F|$.

We will now contrast the set of locally consistent marginals $LOC(G)$ for the cases when G is a tree and when G is not.

- G is a Tree
  - When G is a tree, then for any $\{b_i, b_a\} \in LOC(G)$, there exists an unique measure $b_* \in Measures(\mathcal{X}^\nu)$ whose marginals are given by $\{b_i, b_a\}$.
  - The measure b_* is given by

$$b_*(\underline{x}) = \prod_{a \in F} \left(\frac{b_a(\underline{x}_{\partial a})}{\prod_{i \in \partial a} b_i(x_i)} \right) \prod_{i \in V} b_i(x_i)$$

  - The Gibbs free energy for b_* is given by $\mathbb{G}(b_*) = \mathbb{G}(b_a, b_i)$ and hence $\log Z = \max_{LOC(G)} \mathbb{G}(b_a, b_i)$
- G is not a Tree
  - In this case, there exists a set of marginals $\{b_i, b_a\} \in LOC(G)$ that are not marginals of any distribution b on the graph G.

The example of a case where the set $\{b_i, b_a\} \in LOC(G)$ fails to be the marginals for any distribution on G, can be obtained by considering the case of cyclic graph consisting of three variable nodes and three factor nodes as in Fig(**??**). Each variable x_i takes values in the set $\{0, 1\}$. The compatibility function for each factor node is specified in the matrix form as,

$$b_{12} = \begin{bmatrix} 0.49 & 0.01 \\ 0.01 & 0.49 \end{bmatrix} b_{23} = \begin{bmatrix} 0.49 & 0.01 \\ 0.01 & 0.49 \end{bmatrix} b_{13} = \begin{bmatrix} 0.01 & 0.49 \\ 0.49 & 0.01 \end{bmatrix}$$

From the compatibility function matrices, one observes that the factor node $(1, 2)$ prefers that variable nodes 1 and 2 be in the same state. Similarly factor node $(2, 3)$ prefers that variable nodes 2 and 3 be in the same state. On the other hand, factor node $(1, 3)$ prefers variable nodes 1 and 3 be in different states. It follows that not all of the compatibility functions can be satisfied simultaneously by a distribution. On the other hand, it can be readily checked that the marginals defined by $b_i(x_i) = (0.5, 0.5)$ and $b_{ij} = \psi_{ij}$ are locally consistent. The relation between the set of marginals of some distribution on G and the set $LOC(G)$ is summed up in Fig(**??**)

The following two results describe the relationship between the stationary points of Gibbs free energy on the set $LOC(G)$ and the fixed point of the Belief Propagation Algorithm on G and thus the relationship between the Gibbs free energy and Bethe free energy.

Proposition 4.3.3. *If the compatibility functions of a factor graph G are such that $\psi_a(\underline{x}_{\partial a}) > 0$ for all $\underline{x}_{\partial a}$ and $a \in F$, then there exists a stationary point of $\mathbb{G}(b_i, b_a)$ in the interior of $LOC(G)$. In this case, there exists a fixed point of the Belief Propagation Algorithm on G, which is given by the above stationary point.*

Claim 4.3.4. *The Bethe free energy is the Lagrangian dual of the Gibbs free energy for the locally consistent marginals .*

Proof. Before writing the Lagrangian dual for the Gibbs free energy on the set $LOC(G)$, note that the constraints are specified by Eqn(4.3.3) and Eqn(4.3.4). Thus, in the dual, there will be a variable λ_i for each variable node $i \in V$, and a variable $\lambda_{i \to a}(x_i)$ for all x_i and for each edge $(ia) \in G$. Thus the Lagrangian dual is given by

$$\mathcal{L}(\{b\}, \{\lambda\}) = \mathbb{G}(b) - \sum_i \lambda_i \left(\sum_{x_i} b_i(x_i) - 1 \right) - \sum_{(ia)} \sum_{x_i} \lambda_{i \to a}(x_i) \left(\sum_{\underline{x}_{\partial a \setminus i}} b_a(\underline{x}_{\partial a}) - b_i(x_i) \right)$$

Setting $\frac{\partial \mathcal{L}}{\partial b} = 0$ leads to an expression of b as a function of λ. Substituting this into the above expression, one gets $\mathcal{L}$ as a function of just λ, which after a variable change from λ to $\{\nu_{i\to a}, \hat{\nu}_{a\to i}\}$ leads to the expression for Bethe Free energy. □

Last lecture we saw that the Bethe free energy is the Lagrange dual of the Gibbs free energy for the locally consistent marginals. Given a set of messages $\{\hat{\nu}_{a\to i}, \nu_{i\to a}\}$ we can construct the marginals by setting $\frac{\partial \mathcal{L}}{\partial b} = 0$

$$b_i(x_i) \propto \prod_{a\in\partial i} \hat{\nu}_{a\to i}(x_i)$$
$$b_a(\underline{x}_{\partial a}) \propto \psi_a(\underline{x}_{\partial a}) \prod_{i\in\partial a} \nu_{i\to a}(x_i)$$

Now by imposing the locally consistency condition we get the belief propagation fixed points. Let

$$\nu_{i\to a}(x_i) \propto \prod_{b\in\partial i\setminus a} \hat{\nu}_{a\to i}(x_i)$$

Imposing the locally consistency condition $\sum_{\underline{x}_{\partial a\setminus i}} b_a(\underline{x}_{\partial a}) = b_i(x_i)$, we get

$$\prod_{b\in\partial i} \hat{\nu}_{b\to i}(x_i) \propto \sum_{\underline{x}_{\partial a\setminus i}} \psi(\underline{x}_{\partial a}) \prod_{j\in\partial a} \nu_{j\to a}(x_j)$$
$$\hat{\nu}_{a\to i}(x_i) \prod_{b\in\partial i\setminus a} \hat{\nu}_{b\to i}(x_i) \propto \sum_{\underline{x}_{\partial a\setminus i}} \psi(\underline{x}_{\partial a})\nu_{i\to a}(x_i) \prod_{j\in\partial a\setminus i} \nu_{j\to a}(x_j)$$
$$\hat{\nu}_{a\to i}(x_i) \propto \sum_{\underline{x}_{\partial a\setminus i}} \psi(\underline{x}_{\partial a}) \prod_{j\in\partial a\setminus i} \nu_{j\to a}(x_j)$$

4.3.3 Bethe free energy as a functional over messages

Recall the definition of the Bethe Free Energy $\mathbb{G}_{\mathrm{B}} : \{\nu, \hat{\nu}\} \to \mathrm{R}$ as

$$\begin{aligned}\mathbb{G}_{\mathrm{B}}(\nu, \hat{\nu}) \;=\; & -\sum_{i,a} \log\left(\sum_{x_i} \nu_{i\to a}(x_i)\hat{\nu}_{a\to i}(x_i)\right) + \sum_{a\in F} \log\left(\sum_{\underline{x}_{\partial a}} \psi_a(\underline{x}_{\partial a}) \prod_{i\in\partial a} \nu_{i\to a}(x_i)\right) \\ & + \sum_{i\in V} \log\left(\sum_{x_i} \prod_{a\in\partial i} \hat{\nu}_{a\to i}(x_i)\right).\end{aligned}$$

Note that there are three terms in the expression, one for each edge, one for each factor node, and one for each variable node in the factor graph. The following claim justifies the study of Bethe free energy.

Proposition 4.3.5. *For any factor graph, the stationary points of the Bethe free energy correspond to the fixed points of the Belief Propagation algorithm and vice versa. That is,*

$$\frac{\partial \mathbb{G}_{\mathrm{B}}}{\partial \nu} = 0$$

if and only if the messages are a fixed point of belief propagation.

Proof. Note first from the definition of the Bethe free energy, that it is independent of scale with respect to each of its arguments. That is, if we replace, $\nu_{i\to a}(x_i)$ by $\lambda\nu_{i\to a}(x_i)$ for some $\lambda > 0$, then the value of the Bethe free energy remains the same. Differentiating the expression for the Bethe Free energy with respect to $\nu_{i\to a}(x_i)$, we get

$$\frac{\partial \mathbb{G}_{\mathrm{B}}}{\partial \nu_{i\to a}(x_i)} = -\frac{\hat{\nu}_{a\to i}(x_i)}{\sum_{x_i}\nu_{i\to a}(x_i)\hat{\nu}_{a\to i}(x_i)} + \frac{\sum_{\underline{x}_{\partial a\setminus i}}\psi_a(\underline{x}_{\partial a})\prod_{j\in\partial a\setminus i}\nu_{j\to a}(x_j)}{\sum_{\underline{x}_{\partial a}}\psi_a(\underline{x}_{\partial a})\prod_{j\in\partial a}\nu_{j\to a}(x_j)}$$

Setting $\frac{\partial \mathbb{G}_{\mathrm{B}}}{\partial \nu_{i\to a}(x_i)} = 0$ and rearranging the expression we see that,

$$\begin{aligned}\hat{\nu}_{a\to i}(x_i) &= \left(\frac{\sum_{x_i}\nu_{i\to a}(x_i)\hat{\nu}_{a\to i}(x_i)}{\sum_{\underline{x}_{\partial a}}\psi_a(\underline{x}_{\partial a})\prod_{j\in\partial a}\nu_{j\to a}(x_j)}\right)\sum_{\underline{x}_{\partial a\setminus i}}\psi_a(\underline{x}_{\partial a})\prod_{j\in\partial a\setminus i}\nu_{j\to a}(x_j)\\ &\propto \sum_{\underline{x}_{\partial a\setminus i}}\psi_a(\underline{x}_{\partial a})\prod_{j\in\partial a\setminus i}\nu_{j\to a}(x_j)\end{aligned}$$

which is nothing but the equation for the fixed point of the Belief propagation algorithm (for factor message). The corresponding equation for the variable message can be obtained similarly by differentiating the expression for Bethe Free energy with respect to $\hat{\nu}_{a\to i}(x_i)$ and setting $\frac{\partial \mathbb{G}}{\partial \hat{\nu}_{a\to i}(x_i)} = 0$. Doing this we get,

$$\nu_{i\to a}(x_i) \propto \prod_{b\in\partial i\setminus a}\hat{\nu}_{b\to i}(x_i)$$

Thus the stationary points of the Bethe free energy are the fixed points for the Belief Propagation algorithm. The converse can be obtained by direct substitution. □

4.4 Region-Based Approximation of the Free Energy

The idea of this approximation is that we have a complicated system for which we cannot compute the free energy. Therefore, We decompose the system into subsystems and then approximate the free energy by combining the free energies of the subsystems.

4.4.1 Regions and Region-Based Free Energy

We define a *region* R of a factor graph $G = (V, F, E)$ to be (V_R, F_R, E_R) such that

$$\begin{aligned} a \in F_R &\Rightarrow \partial a \subseteq V_R\\ i \in V_R, a \in F_R, (i,a) \in E &\Rightarrow (i,a) \in E_R\end{aligned}$$

An example of a region is shown in Figure 4.3 (a). Given a region R, the Gibbs free energy of the region $\mathbb{G}_R : M(\mathcal{X}^{V_R}) \to \mathbb{R}$ is

$$\mathbb{G}_R[b_R] = H[b_R] + \sum_{\underline{x}_R}\sum_{a\in F_R} b_R(\underline{x}_R)\log(\psi_a(\underline{x}_{\partial a})) = H_R[b_R] - U_R[b_R]$$

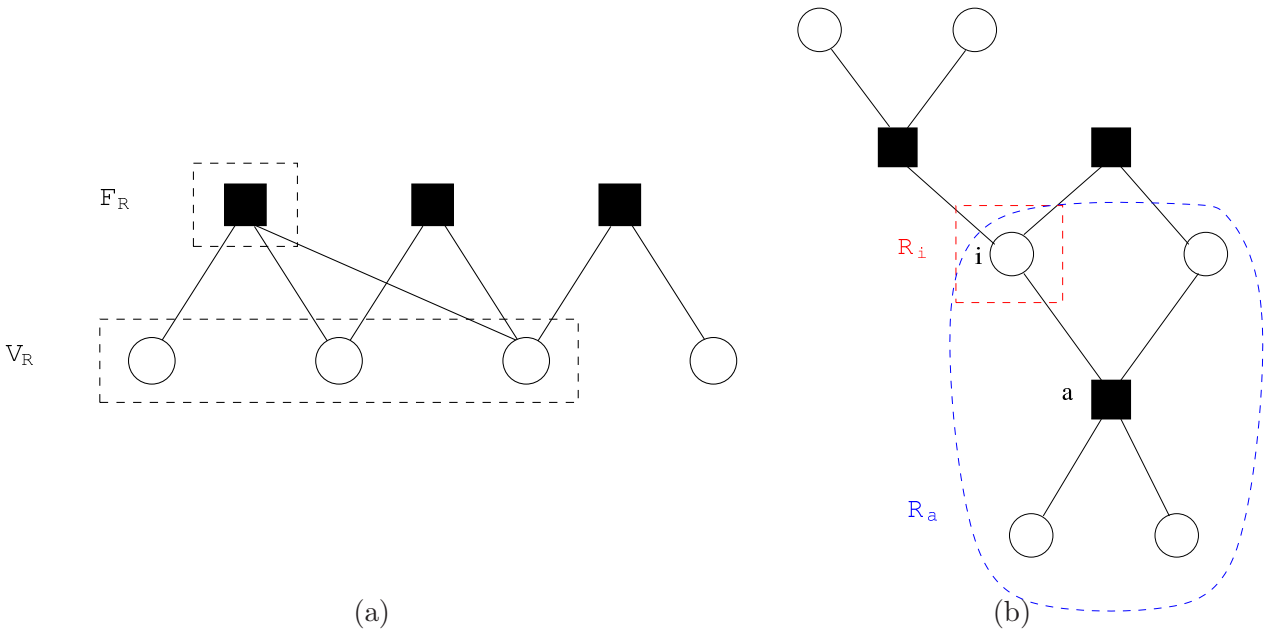


Figure 4.3: (a) An example of a region. (b) Regions of the Bethe approximation.

where $\underline{x}_R = \{x_i : i \in V_R\}$.
The size of the domain of $\mathbb{G}_R$ is exponential in the size of the region. Therefore for small regions we can compute the associated free energy.

For a family of regions $\mathcal{R} = \{R_1, R_2, ..., R_q\}$ and the associated set of coefficients $c_\mathcal{R} = \{c_R : R \in \mathcal{R}\}$, define the *region-based free energy* $\mathbb{G}_\mathcal{R} : M(\mathcal{X}^{V_{R_1}}) \times M(\mathcal{X}^{V_{R_2}}) \times ... \times M(\mathcal{X}^{V_{R_q}}) \rightarrow R$ as

$$\mathbb{G}_\mathcal{R}\{b_\mathcal{R}\} = \sum_{R \in \mathcal{R}} c_R \mathbb{G}_R[b_R]$$

where $b_\mathcal{R} = \{b_R(\underline{x}_R) : R \in \mathcal{R}\}$.

As a check, assume that we have two disjoint factor graphs G_1 and G_2 and let $G = G_1 \bigcup G_2$. Then clearly, the free energy of G is the summation of free energies of G_1 and G_2. Now let $\mathcal{R} = \{G_1, G_2\}$, and $c_1 = c_2 = 1$. Region-based free energy, $\mathbb{G}_\mathcal{R}$, gives the exact value of the free energy. When the regions overlap then $\mathbb{G}_\mathcal{R}$ will give an approximation of the free energy.

Example 1 Bethe Free Energy
Bethe approximation is an example of region-based approximation where the regions are (Figure 4.3 (b))

$$\begin{aligned} \mathcal{R} &= \{R_i, R_a : i \in V, a \in F\} \\ R_i &= (\{i\}, \emptyset, \emptyset), & \forall i \in V \\ R_a &= (\{\partial a\}, \{a\}, \{(i, a) : i \in \partial a\}), & \forall a \in F \end{aligned}$$

and the coefficients are

$$\begin{aligned} c_i &= 1 - |\partial i|, \quad \forall i \in V \\ c_a &= 1, \qquad\quad \forall a \in F \end{aligned}$$

The corresponding region-based free energy is

$$\mathbb{G}_{\mathcal{R}}\{b_a, b_i\} = \sum_{i \in V}(1 - |\partial i|)H[b_i] + \sum_{a \in F}\left(H[b_a] + \sum_{\underline{x}_{\partial a}} b_a(\underline{x}_{\partial a}) \log(\psi_a(\underline{x}_{\partial a}))\right)$$

4.4.2 Region-Based Approximation

We approximate the free energy by

$$F = \log(Z) \approx \max_{b_{\mathcal{R}}} \mathbb{G}_{\mathcal{R}}\{b_{\mathcal{R}}\}$$

where the marginals of the regions must be consistent, i.e.

$$\forall R \supset R' : \sum_{\underline{x}_{R \setminus R'}} b_R(\underline{x}_R) = b_{R'}(\underline{x}_{R'})$$

and the coefficients must satisfy the following rules

$$\sum_{R \in \mathcal{R}} c_R \mathbb{I}(i \in V_R) = 1, \quad \forall i \in V \tag{4.4.1}$$

$$\sum_{R \in \mathcal{R}} c_R \mathbb{I}(a \in F_R) = 1, \quad \forall a \in F \tag{4.4.2}$$

Constraining the coefficients to satisfy rules (4.4.1) and (4.4.2) has the following consequences

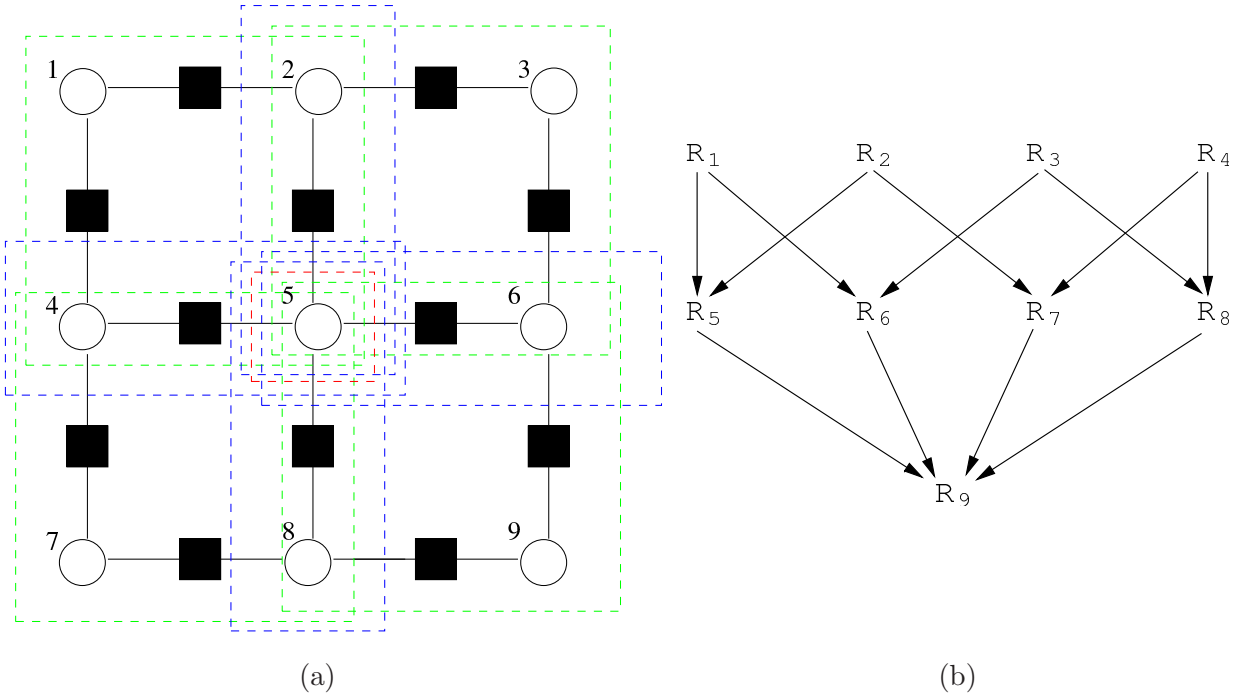
1. If (4.4.2) holds and the marginals $\{b_R : R \in \mathcal{R}\}$ are the real marginals then $U_{\mathcal{R}}(b_{\mathcal{R}})$ is equal to the energy term of the Gibbs free energy of the real distribution, i.e, $U[\mu] = -\mathbb{E}_\mu \log(\psi(\underline{x}))$.

$$\begin{aligned} U_{\mathcal{R}}(b_{\mathcal{R}}) &= -\sum_{R \in \mathcal{R}} c_R \sum_{\underline{x}_R} b_R(\underline{x}_R) \sum_{a \in F_R} \psi_a(\underline{x}_{\partial a}) \\ &= -\sum_{R \in \mathcal{R}} \sum_{a \in F} \mathbb{I}(a \in F_R) c_R \sum_{\underline{x}_{\partial a}} \mu_a(\underline{x}_{\partial a}) \psi_a(\underline{x}_{\partial a}) \\ &= -\sum_{a \in F} \left(\sum_{R \in \mathcal{R}} \mathbb{I}(a \in F_R) c_R\right) \sum_{\underline{x}_{\partial a}} \mu_a(\underline{x}_{\partial a}) \psi_a(\underline{x}_{\partial a}) \\ &= -\sum_{a \in F} \sum_{\underline{x}_{\partial a}} \mu_a(\underline{x}_{\partial a}) \psi_a(\underline{x}_{\partial a}) \\ &= U[\mu] \end{aligned}$$

2. If (4.4.1) holds, the marginals $\{b_R : R \in \mathcal{R}\}$ are the real marginals, and the real distribution is uniform then $H_{\mathcal{R}}(b_{\mathcal{R}})$ is equal to the entropy term of the Gibbs free energy of the real distribution.

Example 2 Regions with Short Loops
Given the factor graph shown in Figure 4.4 (a), we define the following set of regions (and the corresponding coefficients) where each region has at most one short loop. Regions are shown in Figure 4.4 (a) as dashed boxes. It is not hard to check that the coefficients satisfy rules (4.4.1) and (4.4.2).

$$
\begin{array}{llll}
V_{R_1} = \{1,2,4,5\}, & c_1 = 1 & V_{R_5} = \{2,5\}, & c_5 = -1 \\
V_{R_2} = \{2,3,5,6\}, & c_2 = 1 & V_{R_6} = \{4,5\}, & c_6 = -1 \\
V_{R_3} = \{4,5,7,8\}, & c_3 = 1 & V_{R_7} = \{5,6\}, & c_7 = -1 \\
V_{R_4} = \{5,6,8,9\}, & c_4 = 1 & V_{R_8} = \{5,8\}, & c_8 = -1 \\
V_{R_5} = \{5\}, & c_9 = 1 & &
\end{array}
$$



(a) (b)

Figure 4.4: (a) Factor graph and the regions. (b) Region graph corresponding to the regions.

4.4.3 Region Graph

Region graph $\mathcal{G}(\mathcal{R})$ of a set of regions $\mathcal{R}$ is a directed graph with vertices $R \in \mathcal{R}$ and directed edges where

$$R \to R' \Rightarrow R' \subseteq R$$

Figure 4.4 (b) shows a region graph corresponding to the regions of Example 2. Note that because of the condition on directed edges the graph must be a directed acyclic graph.

If there is a directed edge between R and R' then we say that R is a parent of R', $R \in P(R')$, and R' is a child of R, $R' \in C(R)$. If there is a directed path between R and R' then we say that R is an ancestor of R', $R \in A(R')$, and R' is a descendent of R, $R' \in D(R)$.

If a set of regions $\mathcal{R}$ is represented as a region graph (such representation is not necessarily unique) and the corresponding coefficients satisfy (4.4.1) and (4.4.2) and following condition

$$C_R = 1 - \sum_{R' \in A(R)} C_{R'} \quad \forall R \in \mathcal{R}$$

then the marginals of regions, $b_{\mathcal{R}}$, can be computed iteratively using Generalized Belief Propagation (GBP) algorithm.

4.5 Generalized Belief Propagation

To solve the following optimization problem

$$\begin{aligned} &maximaize \quad \mathbb{G}_{\mathcal{R}}\{b_{\mathcal{R}}\} \\ &subject\ to \quad \sum_{\underline{x}_{R \setminus R'}} b_R(\underline{x}_R) = b_{R'}(\underline{x}_{R'}), \quad \forall R \to R' \end{aligned}$$

We form the Lagrangian

$$\mathcal{L}(\{b_R\}, \{\lambda_{R \to R'}\}) = \mathbb{G}_{\mathcal{R}}\{b_{\mathcal{R}}\} - \sum_{R \to R'} \lambda_{R \to R'}(\underline{x}_{R'}) C_{R \to R'}(\underline{x}_{R'}))$$

where

$$C_{R \to R'}(\underline{x}_{R'}) = \sum_{\underline{x}_{R \setminus R'}} b_R(\underline{x}_R) - b_{R'}(\underline{x}_{R'})$$

Setting $\frac{\partial \mathcal{L}}{\partial b} = 0$ leads to an expression of marginals as functions of Lagrange multipliers. Now by imposing the consistency conditions, $C_{R \to R'}(\underline{x}_{R'}) = 0$, we get the update rules of the message passing algorithm.

After a variable change from $\lambda_{R \to R'}$ to $\nu_{R \to R'}$, the marginal of a region b_R is given by

$$b_R(\underline{x}_R) \propto \prod_{a \in F_R} \psi_a(\underline{x}_{\partial a}) \prod_{R_1 \in P(R)} \nu_{R_1 \to R}(\underline{x}_R) \prod_{R_2 \in D(R)} \prod_{R_3 \in P(R_2) \setminus R, D(R)} \nu_{R_3 \to R_2}(\underline{x}_{R_2})$$

The relationships among R, R_1, R_2, and R_3 are described in Figure 4.5.

4.6 Tree-based bounds

In this section we discuss a different aproach that aims at improving Bethe approximation to the free energy, and constructing message passing algorithms with better properties than belief propagation.

In particular, Bethe free energy has two problems that limit its applicability

(a) (b)

Figure 4.5: Relationships among R, R_1, R_2, and R_3

1. It is no is not a concave function of the beliefs, and hence it is dificult to optimize.

2. It does not yield a bound on the Helmotz free energy (log-partition function).

The general idea is to construct bounds for the log-partition function that are obtained by optimizing appropriate functionals of 'beliefs' (local marginals). However the emphasis is on obtaining tractable bounds (specifically, concave functionals).

We begin in Section 4.6.1 by recalling some useful properties of exponential families of probability distributions.

4.6.1 Exponential families

Given the finite space $\mathcal{X}^V$, and a collection of functions $T_1, \ldots, T_d : \mathcal{X}^V \to \mathbb{R}$, the corresponding *exponential family*[2] is a family of probability distributions $\{\mu_\theta\}$ on $\mathcal{X}^V$ indexed by $\theta = (\theta_1, \ldots, \theta_d) \in \mathbb{R}^d$. Given the parameters θ, the corresponding distribution is given by

$$\mu_\theta(x) = \frac{1}{Z(\theta)} \exp\big\{\langle \theta, T(x)\rangle\big\} = \frac{1}{Z(\theta)} \exp\Big\{\sum_{\ell=1}^{d} \theta_\ell T_\ell(x)\Big\}. \tag{4.6.1}$$

The partition function $Z(\theta) = e^{\Phi(\theta)}$ is as usual defined by the normalization condition

$$\Phi(\theta) = \log\Big\{\sum_{\underline{x}\in\mathcal{X}} e^{\langle \theta, T(x)\rangle}\Big\} \tag{4.6.2}$$

The following is an elementary, but useful fact.

Lemma 4.6.1. *The function $\theta \mapsto \Phi(\theta)$ is convex and, denoting by $\mathbb{E}_\theta$, Cov_θ expectation and covariance with respect to μ_θ, we have*

$$\nabla\Phi(\theta) = \mathbb{E}_\theta\{T(x)\}, \qquad \mathrm{Hess}\Phi(\theta) = \mathsf{Cov}_\theta(T(x); T(x)). \tag{4.6.3}$$

[2]Slightly more general definitions are possible, but we for our purposes the present one is sufficient.

Further $\theta \mapsto \Phi(\theta)$ is strictly convex if and only if the functions 1, T_1, $\dots T_d$ to be linearly independent on $\mathcal{X}^V$ (i.e. there are no numbers $c_0, c_1, \dots, c_d$ such that $c_1 + c_1 T_1(x) + \cdots + C_d T_d(x) = 0$ identically).

Given a set of functions $T = (T_1, \dots, T_d)$, we let $\mathsf{L}(T) \subseteq \mathbb{R}^d$ denote the polytope

$$\mathsf{L}(T) = \Big\{ \sum_{x \mathcal{X}^V} p(x) T(x) \,, \qquad p \in \mathsf{M}(\mathcal{X}^V) \Big\} \,. \tag{4.6.4}$$

This is the set of possible expectations of T, when the probability distribution over x is arbitrary. Equivalently, is the convex hull of the sets of points $\{T(x) \ : \ x \in \mathcal{X}^V\}$. Remarkably, exponential families allow to realize any point in the interior of $\mathsf{L}(T)$.

Proposition 4.6.2. *For any $t = (t_1, \dots, t_d)$ in the relative interior of $\mathsf{L}(T)$, there exists $\theta \in \mathbb{R}^d$ such that*

$$\mathbb{E}_\theta\{T(x)\} = t \,. \tag{4.6.5}$$

Proof. Without loss of generality, we can assume that the functions $1, T_1, \dots, T_d$ are linearly dependent (because otherwise we can work with a linearly independent subset), and hence that $\mathsf{L}(T)$ has full dimension. Let t be a point in the interior of $\mathsf{L}(T)$. This means that $t = \sum_x p(x) T(x)$ for some $p(x)$ strictly positive for all x. Consider the lagrangian

$$\mathcal{L}(\theta, t) = \Phi(\theta) - \langle t, \theta \rangle = \Phi(\theta) - \sum_x p(x) \langle T(x), \theta \rangle \tag{4.6.6}$$

The function $\theta \mapsto \mathcal{L}(\theta, t)$ is strictly convex. Further, for any $\theta \in \mathbb{R}^d$, as $b \to +\infty$

$$\mathcal{L}(b\theta, t) = b \max\{\langle \theta, T(x) \rangle \,:\, x \in \mathcal{X}^V\} - b \sum_x p(x) \langle T(x), \theta \rangle + O(1) \to \infty \tag{4.6.7}$$

Define $\theta_* = \arg\min_{\theta \in \mathbb{R}^d} \mathcal{L}(\theta, t)$ (this is unique because of strict convexity of Φ). By the above, θ_* is finite, and hence $\nabla \mathcal{L}(\theta_*, t) = 0$. Using Lemma 4.6.1, this immediately implies $\mathbb{E}_{\theta_*}\{T(x)\} = t$.

Uniqueness follows because otherwise there would be two points θ, θ' such that $\nabla \mathcal{L}(\theta, t) = \nabla \mathcal{L}(\theta', t) = 0$, i.e. two distinct stationary points of $\mathcal{L}$, which is impossible by strict convexity. □

Given the last proof, it is natural to consider the function $\mathbb{H} : \mathrm{relint}(\mathsf{L}(T)) \to \mathbb{R}$ defined by

$$\mathbb{H}(t) = \min\{\mathcal{L}(\theta, t) \ : \ \theta \in reals^d\} \,. \tag{4.6.8}$$

Proposition 4.6.3. *The function $t \mapsto \mathbb{H}(t)$ is concave (strictly concave if $1, T_1, \dots, T_d$ are linearly independent) and*

$$\mathbb{H}(t) = \max \big\{ H(p) \ : \ p \in \mathsf{M}(\mathcal{X}^V) \,, \mathbb{E}_p\{T(x)\} = t \big\} \,. \tag{4.6.9}$$

Further, the maximum is achieved when $p = \mu_\theta$ for a certain θ.

Proof. Concavity follows because $\mathbb{H}$ is the lagrange dual of Φ. To prove the characterization (4.6.9), call $\mathbb{H}'(t)$ the right hand side. Then $\mathbb{H}(t) \le \mathbb{H}'(t)$, because simple calculus shows that $\mathbb{H}(t) = H(\mu_\theta)$. On the other hand for any $\theta \in \mathbb{R}^d$,

$$\mathbb{H}(t) \le \max_p \big\{ H(p) - \mathbb{E}_p \langle \theta, T(x) \rangle + \langle \theta, t \rangle \big\} = \mathcal{L}(\theta, t) \,. \tag{4.6.10}$$

□

4.6.2 Concavity of Bethe Free Energy on Trees

How do we apply the formalism of exponential families to graphical models? Consider a factor graph model with graph $G = (V, F, E)$:

$$\mu(x) = \frac{1}{Z} \prod_{a \in F} \psi_a(\underline{x}_a) \,. \tag{4.6.11}$$

It is easy to construct an exponential family, of which μ is a special member. This is just the family that includes the following functions

$$\begin{aligned} T_{i,z}(x) &= \mathbb{I}(x_i = z) \,, \qquad \forall i \in V \; z \in \mathcal{X} \,, & (4.6.12) \\ T_{a,\underline{z}_{\partial a}}(x) &= \mathbb{I}(\underline{x}_{\partial a} = \underline{z}_{\partial a}) \,, \qquad \forall a \in F \; \underline{z}_{\partial a} \in \mathcal{X}^{\partial a} \,., & (4.6.13) \end{aligned}$$

In other words, this is the family of models that factorize according to the graph G. We shall denote the set of parameters of this family by $\theta_i(z)$ (for the first set of functions) and $\theta_a(\underline{z}_{\partial a})$ (for the second set of functions). An element of this family then takes the form

$$\mu_\theta(x) = \frac{1}{Z(\theta)} \exp\Big\{ \sum_{a \in F} \theta_a(\underline{x}_{\partial a}) + \sum_{i \in V} \theta_i(x_i) \Big\} \,. \tag{4.6.14}$$

Notice that in this case the functions $T(x)$ are not linearly independent, and therefore a given distribution μ in the exponential family admits, in general more than one parametrization of the form (4.6.14). The model (4.6.11) is embedded in this family by letting

$$\begin{aligned} \overline{\theta}_i(x_i) &= 0 \,, & (4.6.15) \\ \overline{\theta}_a(\underline{x}_{\partial a}) &= \log \psi_a(\underline{x}_{\partial a}) \,. & (4.6.16) \end{aligned}$$

The results of the previous section have immediate consequences for graphical models. Recall that the Bethe entropy is defined as

$$H_{\mathrm{B}}(\underline{b}) = \sum_{a \in F} H(b_a) - \sum_{i \in V} (1 - |\partial i|) H(b_i) \,,$$

where $\underline{b} = \{b_i, b_a \; : \; i \in V \,, \; a \in F\}$ is a vector of beliefs.

Proposition 4.6.4. *If G is a tree, then Bethe entropy $H_{\mathrm{B}} : \mathsf{LOC}(G) \to \mathbb{R}$ is concave.*

Proof. This is in fact a corollary of Proposition 4.6.3, by using the formulation given above, and the exactness of Bethe free energy for factor graph models, when the underlying graph is a tree. □

4.7 Upper bound

For graphs that are not trees, we can use the upper bound

$$\Phi(\overline{\theta}) \leq \sum_T \rho_T \Phi(\theta_T) \tag{4.7.1}$$

for any collection of $\{\theta_T\}$ of parameters and weights $\{\rho_T\}$ such that

$$\sum_T \rho_T = 1, \qquad \rho_T \geq 0 \tag{4.7.2}$$

$$\sum_T \theta_T \rho_T = \overline{\theta} \tag{4.7.3}$$

Idea :

- Choose θ_T such that $F(\theta_T)$ can be computed easily
- Optimize over $\{\rho_T\}$, $\{\theta_T\}$ under the constraints (3),(4).

For example, we could have $T = (V_T, F_T, E_T)$ a spanning tree, where $V_T = V, F_T \subseteq F$. A tree is a spanning tree if $V_T = V$ is connected. Let's assume ρ_T is fixed and optimize over θ_T.

$$\mathcal{L}(\{b\},\{\theta\}) = \sum_T \rho_T \Phi(\theta_T) - \sum_{a,\underline{x}_{\partial a}} b_a(\underline{x}_{\partial a})[\sum_T \rho_T \theta_a^T(\underline{x}_{\partial a}) - \overline{\theta}_a(\underline{x}_{\partial a})] - \sum_{i,x_i} b_i(x_i)[\sum_T \rho_T \theta_i^T(x_i) - \overline{\theta}_i(x_i)]\,.$$

Stationarity conditions with respect to θ^T gives

$$b_a(\underline{x}_{\partial a}) = \mu_a^{\theta_T}(\underline{x}_{\partial a}) \tag{4.7.4}$$

$$b_i(x_i) = \mu_i^{\theta_T}(x_i) \tag{4.7.5}$$

Further, since θ_T is non-vanishing on a tree, and using the stationarity conditions we get

$$\begin{aligned}\Phi(\theta_T) &= \sum_{a\in F} H(\mu_a^{\theta_T}) + \sum_{i\in V}(1-|\partial i|)H(\mu_i^{\theta_T}) + \sum_{i\in V}\sum_{x_i}\theta_i^T(x_i)\mu_i^{\theta_T}(x_i) + \sum_{a\in F}\sum_{\underline{x}_a}\theta_a^T(\underline{x}_{\partial a})\mu_a^{\theta_T}(\underline{x}_a)\\ &= \sum_{a\in F} H(b_a) + \sum_{i\in V}(1-|\partial i|)H(b_i) + \sum_{i\in V}\sum_{x_i}\theta_i^T(x_i)b_i(x_i) + \sum_{a\in F}\sum_{\underline{x}_a}\theta_a^T(\underline{x}_{\partial a})b_a(\underline{x}_a)\,.\end{aligned}$$

Using this expression and eliminating θ^T from the lagrangian, we get the upper bound

$$\Phi(\overline{\theta}) \leq \max_{\underline{b}} \mathbb{G}_{\mathrm{T}}(\underline{b}) \tag{4.7.6}$$

where

$$\mathbb{G}_{\mathrm{T}}\{b\} = \sum_{a,\underline{x}_{\partial a}} b_a(\underline{x}_{\partial a})\overline{\theta}_a(\underline{x}_{\partial a}) + \sum_{i,x_i} b_i(x_i)\overline{\theta}_i(x_i) + \sum_a \rho(a)[H(b_a) - \sum_{i\in\partial a} H(b_i)] + \sum_{i\in V} H(b_i)\,.$$

Here, the parameters $\rho(a)$, $a \in F$ are defined as follows

$$\rho(a) = \sum_a \sum_{T\ni a} \rho_T\,. \tag{4.7.7}$$

Notice that $\rho(a)$ can be interpreted as the probability that the factor node a belongs to the random tree T drawn from the probability distribution ρ_T.

The function $\mathbb{G}_{\mathrm{T}} : \mathsf{LOC}(G) \to \mathbb{R}$ is concave by construction (it is a linear combination of concave functions with positive coefficients). Hence the mazimization over $\underline{b}$ is tractable. In order to optimize

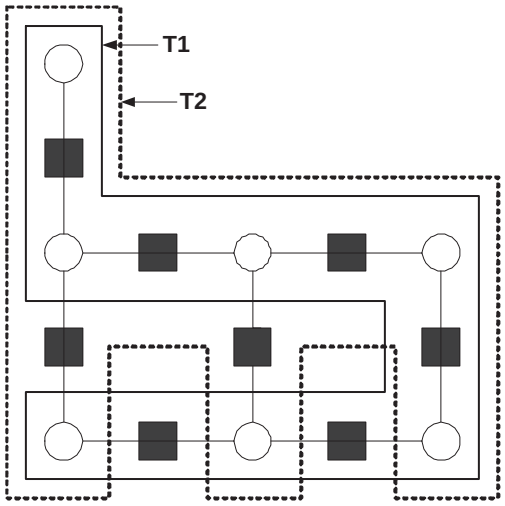


Figure 4.6: Example of choosing spanning trees on a graph

the upper bound, we need to minimize over the parameters $\{\rho_T\}$. Notice however, the bound depends on ρ only through the $O(n)$ quantities $\{\rho(a)\}$. These belong to the so-called spanning tree polytope.

$$\begin{aligned}
\rho_T = \frac{1}{|T|} \rightarrow dn\rho(a) = \sum_a \rho(a) &= \sum_{a,T} \rho_T \mathbb{I}(a \in T) \\
&= \sum_T \rho_T (n-1) \\
&= n-1 \\
\rightarrow \rho(a) = \frac{n-1}{dn}
\end{aligned}$$

when $\rho(a) = 0$, $\mathbb{G}_{\mathrm{T}}$ looks like Naive Mean Field free energy.

Chapter 5

Learning graphical models

In these notes we overview some basic facts about learning graphical models form data. For more information on these topics we refer to [AKN06, BMS08].

5.1 General setting and global approaches

As usual we a factor graph model of the form

$$\mu(\underline{x}) = \frac{1}{Z(G,\psi)} \prod_{a\in F} \psi_a(\underline{x}_{\partial a})\,, \tag{5.1.1}$$

with $G = (V = [n], F, E)$ the underlying factor graph. Notice that we made explicit the dependence of the partition function on the graph G and on the potentials $\psi = \{\psi_a\}_{a\in F}$. In learning we assume to be given s iid configurations $\underline{x}^{(1)}$, $\dots \underline{x}^{(s)}$ with distribution $\mu(\,\cdot\,)$ and want to reconstruct the underlying model from these samples.

It is important to stress that, given a distribution $\mu(\,\cdot\,)$ and a graph G, there are many possible sets of compatibility functions $\{\psi_a\}$ such that $\mu(\,\cdot\,)$ can be written in the form (5.1.1).

One large family of approaches consists in maximizing the likelihood $\prod_\ell \mu(\underline{x}^{(\ell)})$. It is more convenient in this case to introduce the exponential parametrization $\psi_a(\underline{x}_{\partial a}) \equiv e^{\theta_a(\underline{x}_{\partial a})}$ and rather consider the rescaled log-likelihood

$$L(\underline{\theta}; G) \equiv \frac{1}{s}\sum_{\ell=1}^{s} \log \mu(\underline{x}^{(\ell)}) \tag{5.1.2}$$

$$= \frac{1}{s}\sum_{a\in F}\sum_{\ell=1}^{s} \theta_a(\underline{x}_{\partial a}) - \log Z(G;\underline{\theta})\,. \tag{5.1.3}$$

One pleasing fact is that this function is concave in $\underline{\theta}$ (indeed $\log Z(G;\theta) = \log\{\sum_{\underline{x}} \prod_a e^{\theta_a(\underline{x}_{\partial a})}\}$ is convex). However, evaluating it requires to evaluate the log partition function, which is in general #-P hard. Further, if one is interested in learning the graph, the natural search space might be discrete.

There are several options to deal with these problems: one can use an approximate free-energy expression for $\log Z(G;\underline{\theta})$ (naive mean field, Bethe, etc.), and relax combinatorial constraints. We will not follow these interesting directions but rather focus on local approaches.

5.2 Local approaches: Parameter learning

The basic idea is to learn the model 'one piece at a time.' We start in this section by considering the (easier) subproblem of *parameter learning*. This means that we assume that the graph G is given and we are required to learn the compatibility functions $\psi_a(\,\cdot\,)$. In order to get some intuition of this problem, it is convenient to consider a couple of simple examples.

Example 1: Bonsider a model with only one variable $x \in \{0,1\}$ and only one factor node a connected to it. Of course in this case it is supefluous to speak of graphical models, and we can directly specify $\mu(\,\cdot\,)$. A simple parametrization is

$$\mu(x) = \begin{cases} \theta & \text{if } x = 1, \\ 1-\theta & \text{otherwise.} \end{cases} \tag{5.2.1}$$

The data consists therefore in s iid samples $x^{(1)}, \dots, x^{(s)}$ with bias θ. A straightforward procedure for estimating the parameter consists in using $\widehat{\theta} = \sum_{\ell=1}^{s} x^{(\ell)}/s$. We have the following simple guarantee.

Proposition 5.2.1. *For any $\varepsilon, \delta > 0$, let $s^*(\varepsilon,\delta)$ be the minimum number of samples such that $|\widehat{\theta} - \theta| \le \varepsilon$ with probability larger than $1-\delta$. Then*

$$s^*(\varepsilon,\delta) \le \frac{1}{2\varepsilon^2} \log\left(\frac{2}{\delta}\right). \tag{5.2.2}$$

Proof. The statement is an immediate consequence of the following large deviation bound (which follows for instance, from Chernoff bound [CT91])

$$\mathbb{P}\left\{ |\sum_{\ell=1}^{s} x^{(\ell)} - s\theta| \ge s\varepsilon \right\} \le 2\, e^{-2s\varepsilon^2}. \tag{5.2.3}$$

□

Example 2: As a second example, consider k variables $x_1, \dots, x_k \in \mathcal{X}$, connected by a single function node a. In other words $\partial a = \{1, \dots, k\}$ and

$$\mu(x_1, \dots, x_k) = \frac{1}{Z}\, \psi_a(x_1, \dots, x_k). \tag{5.2.4}$$

Of course we cannot pretend to learn $\psi_a(\cdots)$ because any compatibility function obtained by rescaling it leads to the same distribution. We shall therefore content ourselves with the objective of learning a 'canonical' representative $\widetilde{\psi}_a(x_1, \dots, x_k) \equiv \psi_a(x_1, \dots, x_k)/Z$ (well, in this case the choice of such a representative is easy, but in general it is the core of the question).

A natural way of estimating it consists in setting $\widehat{\psi}(x_1, \dots, x_k)$ equal to the fraction of samples ℓ such that $x_1^{(\ell)} = x_1$, …, $x_k^{(\ell)} = x_k$.

Proposition 5.2.2. *Assume $\psi_a(x_1, \dots, x_k) > 0$ for all $x_1, \dots, x_k \in \mathcal{X}$. For any $\varepsilon, \delta > 0$, let $s^*(\varepsilon,\delta)$ be the minimum number of samples such that $|\widehat{\psi}(x_1, \dots, x_k) - \widetilde{\psi}(x_1, \dots, x_k)| \le \varepsilon$ for all $x_1, \dots, x_k \in \mathcal{X}$, with probability larger than $1-\delta$. Then*

$$s^*(\varepsilon,\delta) \le \frac{1}{2\varepsilon^2} \log\left(\frac{2|\mathcal{X}|^k}{\delta}\right). \tag{5.2.5}$$

Proof. The proof is an immediate generalization of the one above. Indeed we are trying to estimate the probabilities of $|\mathcal{X}|^k$ events $(x_1, \dots, x_k)$. The factor $|\mathcal{X}|^k$ comes in because we apply union bound to these events. □

Notice in passing that, as long as ψ_a is strictly positive as in this statement, the same procedure allows to estimate conditional probabilities of the form $\mu(\underline{x}_A|\underline{x}_B)$ for $A, B \in [k]$.

General case. Let us now go back to the general graphical model in Eq. (5.1.1). The above examples show that we can use the samples $\underline{x}^{(1)}, \dots, \underline{x}^{(s)}$ to efficiently learn marginals of the form $\mu_U(\underline{x}_U)$, as long as U has bounded size. On the other hand, for U large, the probability $\mu_U(\underline{x}_U)$ decreases exponentially in $|U|$ and the additive error ε becomes much larger than the quantity to estimate itself. In other words, to learn the probability $\mu_U(\underline{x}_U)$, we need at least one sample such that $\underline{x}_U^{(\ell)} = \underline{x}_U$ (and indeed more than one).

In order to learn the compatibility functions we need then to solve two problems: (I) Define a set of 'canonical' compatibility functions; (II) Express them in terms of 'local' functions. The following result solves both problems.

Theorem 5.2.3 (Hammersley, Clifford)**.** *Assume $\mu(\,\cdot\,)$ to be a factor graph model Eq. (5.1.1) with strictly positive factors. If $\underline{x}^* \in \mathcal{X}^V$ is a configuration, then*

$$\mu(\underline{x}) = \mu(\underline{x}^*) \prod_{a \in F} \widetilde{\psi}_a(\underline{x}_{\partial a})\,, \tag{5.2.6}$$

where

$$\widetilde{\psi}_a(\underline{x}_{\partial a}) \equiv \prod_{U \subseteq \partial a} \left[\frac{\mu_{U,V\setminus U}(\underline{x}_U, \underline{x}^*_{V\setminus U})}{\mu_{U,V\setminus U}(\underline{x}^*_U, \underline{x}^*_{V\setminus U})} \right]^{(-1)^{|\partial a \setminus U|}} . \tag{5.2.7}$$

Proof. The proof makes use of the following well known identity. For any finite set A:

$$\sum_{B \subseteq A} (-1)^{|B|} = \sum_{B \subseteq A} (-1)^{|A\setminus B|} = \begin{cases} 1 & \text{if } A = \emptyset, \\ 0 & \text{otherwise,} \end{cases}$$

where the sums include all distinct subsets of A, including the empty set.

We define, for each nonempty subset of vertices $S \subseteq V$,

$$\widetilde{\psi}_S(\underline{x}_S) \equiv \prod_{U \subseteq S} \mu(\underline{x}_U, \underline{x}^*_{V\setminus U})^{(-1)^{|S\setminus U|}} . \tag{5.2.8}$$

The proof reduces to showing that the following claims hold

1. If $S = \partial a$ for some factor node a, then the latter definition (5.2.8) coincides with the one in the statement, cf. Eq. (5.2.7).
2. The factor graph model obeys $\mu(\underline{x}) = \mu(\underline{x}^*) \prod_{S \subseteq V} \widetilde{\psi}_S(\underline{x}_S)$ for all $\underline{x} \in \mathcal{X}^V$.
3. If $S \neq \partial a$ for all factor nodes $a \in F$, then $\widetilde{\psi}_S(\underline{x}_S) = 1$.

We now turn to proving these three claims.

Proof of Claim 1. Consider the definition (5.2.7). The denominator is the same for each term of the product and coincides with $\mu(\underline{x}^*)$. By taking the product, this denominator is raised to the power $\sum_{U\subseteq S}(-1)^{|S\setminus U|} = 0$. Hence the overall denominator is $\mu(\underline{x}^*)^0 = 1$.

Proof of Claim 2. Reordering terms we have

$$\prod_{S\subseteq V}\widetilde{\psi}_S(\underline{x}_S) = \prod_{S\subseteq V}\Big\{\prod_{U\subseteq S}\mu(\underline{x}_U, \underline{x}^*_{V\setminus U})^{(-1)^{|S\setminus U|}}\Big\} = \prod_{U\subseteq V}\mu(\underline{x}_U, \underline{x}^*_{V\setminus U})^{\kappa_U}\,,$$

where

$$\kappa_U = \sum_{S\neq\emptyset, U\subset S\subset V}(-1)^{|S\setminus U|}\,.$$

Now it is easy to compute

$$\kappa_\emptyset = \sum_{\emptyset\neq S\subset V}(-1)^{|S|} = \sum_{S\subset V}(-1)^{|S|} - 1 = -1\,,$$
$$\kappa_V = 1\,,$$

and, for $U\neq\emptyset, V$, by letting $R = S\setminus U$,

$$\kappa_U = \sum_{R\subset V\setminus U}(-1)^{|R|} = 0\,.$$

We thus obtained

$$\prod_{S\subseteq C}\widetilde{\psi}_S(\underline{x}_S) = \frac{\mu(\underline{x})}{\mu(\underline{x}^*)}\,,$$

which proves the claim.

Proof of Claim 3. onsider a set $S\subseteq V$ that does not coincide with any neighborhood of a factor node. By inspection $\widetilde{\psi}_S(\underline{x}^*_S) = 1$. It is therefore sufficient to show that $\widetilde{\psi}_S(\underline{x}_S)$ does not depend on $\underline{x}_S$. Let $F_S\subseteq F$ denote the set of factor nodes a such that $\partial a\cap S\neq\emptyset$. The joint probability $\mu(\underline{x}_U, \underline{x}^*_{V\setminus U})$ depends on $\underline{x}_U$ only through the factors in F_S, whence

$$\widetilde{\psi}_S(\underline{x}_S) \cong \prod_{U\subseteq S}\Big\{\prod_{a\in F_S}\psi_a(\underline{x}_{\partial a\cap U}, \underline{x}^*_{\partial a\setminus U}\Big\}^{(-1)^{|S\setminus U|}}\,.$$

Since the variables outside S are always fixed to $\underline{x}^*$ in the above expression, we can redefine the factors by letting $\partial a\cap S\to\partial a$. Reordering the terms in the above product and letting $U_a = U\cap\partial a$, $U' = U\setminus\partial a$, we get

$$\begin{aligned}\widetilde{\psi}_S(\underline{x}_S) &\cong \prod_{a\in F_S}\prod_{U_a\subseteq\partial a}\prod_{U'\subseteq S\setminus\partial a}\psi_a(\underline{x}_{U_a}, \underline{x}^*_{\partial a\setminus U_a})^{(-1)^{|S\setminus\partial a|-|U'|}(-1)^{|\partial a|-|U_a|}}\\ &\cong \prod_{a\in F_S}\prod_{U_a\subseteq\partial a}\psi_a(\underline{x}_{U_a}, \underline{x}^*_{\partial a\setminus U_a})^{\kappa(a)(-1)^{|\partial a|-|U_a|}}\,,\end{aligned}$$

where

$$\kappa(a) \equiv \sum_{U' \subseteq S \setminus \partial a} (-1)^{|S \setminus \partial a| - |U'|} = 0 \, .$$

The last equality here holds unless $S = \partial a$ for some factor node a. □

The Hammersley-Clifford theorem clearly solves problem (I) above. To see why it solves (II) as well, thefine MB(a) (the 'Markov blanket' of a) to be the set of variable nodes j that ara at distance 1 from some $i \in \partial a$, but are not in ∂a. It is then easy to show that

$$\frac{\mu_{U,V\setminus U}(\underline{x}_U, \underline{x}^*_{V\setminus U})}{\mu_{U,V\setminus U}(\underline{x}^*_U, \underline{x}^*_{V\setminus U})} = \frac{\mu(\underline{x}_U, \underline{x}^*_{\partial a \setminus U} | \underline{x}^*_{\mathrm{MB}(a)\setminus \partial a})}{\mu(\underline{x}_U, \underline{x}^*_{\partial a \setminus U} | \underline{x}^*_{\mathrm{MB}(a)\setminus \partial a})} \, . \tag{5.2.9}$$

As long as MB(a) is bounded, we can estimate efficiently the above probabilities and hence the functions $\widetilde{\psi}_a(\,\cdot\,)$.

Proposition 5.2.4. *Consider a graphical model with degree bounded by Δ, and $\psi_a(\,\cdot\,) \geq \xi > 0$ for all $a \in F$. For any $\varepsilon, \delta > 0$, let $s^*(\varepsilon, \delta)$ be the minimum number of samples that allow to estimate the compatibility functions with precision ε, with probability larger than $1 - \delta$. Then there exist constants $A, B > 0$ such that*

$$s^*(\varepsilon, \delta) \leq \frac{A}{\varepsilon^2} \log\left(\frac{Bn}{\delta}\right) . \tag{5.2.10}$$

5.3 Local approaches: Structural learning

In this section we consider, for simplicity, a pairwise graphical model (a Markov Random Field) on the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$:

$$\mu(\underline{x}) = \frac{1}{Z} \prod_{(i,j)\in\mathcal{E}} \psi_{ij}(x_i, x_j) \, . \tag{5.3.1}$$

and assume $\mathcal{G}$ to have degree bounded by Δ. The structural learning problem requires to reconstruct the graph $\mathcal{G}$ (i.e. its edge set) from the i.i.d. samples $\underline{x}^{(1)}, \dots, \underline{x}^{(s)}$.

As for parameter learning, we can use these samples to estimate conditional probabilities $\mu(\underline{x}_A | \underline{x}_B)$ for bounded sets $A, B \in \mathcal{V}$. Let us denote by $\widehat{\mu}(\underline{x}_A | \underline{x}_B)$ such an estimate. One can use these estimates to construct a test of the type: *Does U contain the neighborhood of i?* Suppose for a moment that such a test is available, and it fails with sufficiently small probability. One can then use it to reconstruct $\mathcal{G}$ by reconstructing the neighborhood ∂i of each node $i \in \mathcal{V}$. Indeed it is sufficient to apply the test to each of the n^Δ set of vertices $U \subseteq \mathcal{V}$ with $|U| \leq \Delta$. If the test is positive on more than one subset, we return the smallest subset (under inclusion).

One such test is easily described. It depends on a parameter $\delta_0 > 0$ and takes as input i and $U \subseteq \mathcal{V} \setminus \{i\}$. For each $j \in \mathcal{V} \setminus \{i\} \cup U$, each $\underline{x}_U \in \mathcal{X}^U$, x_j, $x'_j \in \mathcal{X}$ it computes

$$\delta(j, \underline{x}_U, x_j, x'_j) \equiv \sum_{x_i \in \mathcal{X}} \left| \widehat{\mu}(x_i | \underline{x}_U, x_j) - \widehat{\mu}(x_i | \underline{x}_U, x'_j) \right| . \tag{5.3.2}$$

If $\delta \geq \delta_0$ for some choice of $j, \underline{x}_U, x_j, x'_j$ then U is rejected: it does not include the neighborhood ∂i. Otherwisr, if $\delta < \delta_0$ for all $j, \underline{x}_U, x_j, x'_j$, then U is accepted.

Bibliography

[AKN06] P. Abbeel, D. Koller, and A. Y. Ng, *Learning Factor Graphs in Polynomial Time and Sample Complexity*, Jour. of Mach. Learn. Res. **7** (2006), 1743–1788.

[AM00] S.M. Aji and R.J. McEliece, *The generalized distributive law*, IEEE Trans. on Inform. Theory **46** (2000), 325–343.

[BMS08] G. Bresler, E. Mossel, and A. Sly, *Reconstruction of Markov random fields from samples: Some easy observations and algorithms*, Random (Cambridge, MA), 2008.

[CT91] T. M. Cover and J. A. Thomas, *Elements of information theory*, Wiley Interscience, New York, 1991.

[JS89] M. Jerrum and A. Sinclair, *Approximate counting, uniform generation and rapidly mixing Markov chains*, Inform. and Comput. **82** (1989), 93–133.

[JVV86] M. Jerrum, L. Valiant, and V. Vazirani, *Random generation of combinatorial structures from a uniform distribution*, Theoret. Comput. Sci. **43** (1986), 169–188.

[WJ08] M. J. Wainwright and M. I. Jordan, *Graphical models, exponential families, and variational inference*, Foundations and Trends in Machine Learning (2008), no. 1.

[WJW05] M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky, *A New Class of Upper Bounds on the Log Partition Function*, IEEE Trans. on Inform. Theory **51** (2005), no. 7, 2313–2335.

[YFW05] J. S. Yedidia, W. T. Freeman, and Y. Weiss, *Constructing free energy approximations and generalized belief propagation algorithms*, IEEE Trans. on Inform. Theory **51** (2005), 2282–2313.