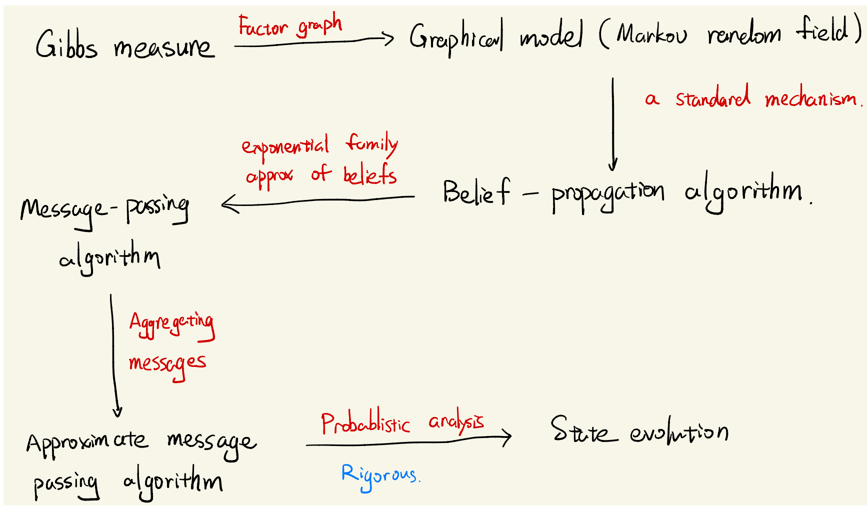
STAT260 Mean Field Asymptotics in Statistical Learning	Lecture 20 - 04/07/2021

Lecture 20: Derivation of AMP I

Lecturer: Song Mei *Scriber: Yihong Wu* *Proof reader: Alexander Tsigler*

The overall methodology of deriving the AMP and related algorithms is summarized as follows:



1 Markov random field

A factor graph $G = (V, F, E)$ is a bipartite graph where $|V| = N, |F| = M$ and $E \subset V \times F$. Here V and F are the sets of variable nodes and factor nodes, respectively.

Given G, a Markov random field is a probability measure μ on the configuration space $\Omega = \mathcal{X}^{\otimes N}$ which admits the following form

$$\mu(x) = \frac{1}{Z} \prod_{a \in F} \psi_a(x_{\partial a}) \prod_{i \in V} \psi_i(x_i) \tag{1}$$

Here for each factor $a \in F$, $\partial a \subset V$ are its neighbors (variables), and we denote $x_{\partial a} = (x_i : i \in \partial a)$.

Many Gibbs measure of the form $\mu(x) \propto \exp(-\beta H(x))$ can be rewritten in the form (1) with appropriately chosen factor graph. This representation, however, is not unique in general.

Example 1 (1-D Ising model with 3 spins)**.** Consider $\Omega = \{\pm 1\}^3$,

$$\mu(x_1, x_2, x_3) \propto e^{-\beta(x_1 x_2 + x_2 x_3)} = \psi_{a_1}(x_1, x_2)\psi_{a_2}(x_2, x_3), \tag{2}$$

where $\partial a_1 = \{x_1, x_2\}$, $\partial a_2 = \{x_2, x_3\}$, $\psi_{a_1}(x_1, x_2) = e^{-\beta x_1 x_2}, \psi_{a_2}(x_2, x_3) = e^{-\beta x_2 x_3}$, $\psi_1(x_1) = \psi_2(x_2) = \psi_3(x_3) = 1$. See

Example 2 (Bayes linear model)**.** Consider the setting of linear regression $y = Ax_0 + w$, where $x_0 \in \mathbb{R}^d, A \in \mathbb{R}^{n \times d}, w \in \mathbb{R}^n$. Assume that $x_{0i} \overset{iid}{\sim} P_0$ and $w_i \overset{iid}{\sim} N(0, \sigma^2)$. Then the posterior of x_0 is

$$\mu(x) = P(x|A, y) \propto \exp\left\{-\frac{\|y - Ax\|_2^2}{2\sigma^2}\right\} \prod_{i=1}^d P_0(x_i) = \prod_{a=1}^n \underbrace{\exp\left\{-\frac{(y_a - \langle A_a, x\rangle)_2^2}{2\sigma^2}\right\}}_{\psi_a(x_{\partial a})} \prod_{i=1}^d \underbrace{P_0(x_i)}_{\psi_i(x_i)}.$$

In this case, the factor graph is fully connected.

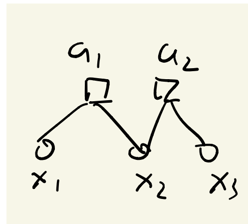


Figure 1: 1-D Ising model with 3 spins

Our main task is to compute the marginal distribution for all x_i

$$\mu_i(x_i) \equiv \int \mu(x) \prod_{j \neq i} \mathrm{d}x_j.$$

To this end, we consider a number of algorithms.

2 Belief propagation algorithm on trees

Let G be a tree factor graph. Let

- $V_{a\to i}$ be the all reachable variables starting from a by blocking i;
- $V_{i\to a}$ be the all reachable variables starting from i by blocking a;
- $F_{a\to i}$ be the all reachable factors starting from a by blocking i;
- $F_{i\to a}$ be the all reachable factors starting from i by blocking a.

Define

$$\hat{\nu}_{a\to i}(x) \propto \prod_{b\in F_{a\to i}} \psi_b(x_{\partial b}) \prod_{j\in V_{a\to i}} \psi_j(x_j) \in \mathcal{P}(\Omega),$$
$$\nu_{i\to a}(x) \propto \prod_{b\in F_{i\to a}} \psi_b(x_{\partial b}) \prod_{j\in V_{i\to a}} \psi_j(x_j) \in \mathcal{P}(\Omega),$$

and their respective marginals

$$\hat{\mu}_{a\to i}(x_i) = \sum_{x_{\setminus i}} \hat{\nu}_{a\to i}(x) \in \mathcal{P}(\mathcal{X}),$$
$$\mu_{i\to a}(x_i) = \sum_{x_{\setminus i}} \nu_{i\to a}(x) \in \mathcal{P}(\mathcal{X}).$$

Here we denote $x_{\setminus i} = (x_j : j \neq i)$.

We claim that $\{\hat{\mu}_{a\to i}, \mu_{i\to a}\}$ satisfies the following relations:

$$\hat{\mu}_{a\to i}(x_i) \propto \sum_{x_{\partial a\setminus i}} \psi_a(x_{\partial a}) \prod_{j\in\partial a\setminus i} \mu_{j\to a}(x_j), \tag{3}$$
$$\mu_{i\to a}(x_i) \propto \psi_i(x_i) \prod_{b\in\partial i\setminus a} \hat{\mu}_{b\to i}(x_i), \tag{4}$$

and the true marginals of μ can be computed as follows

$$\mu_i(x_i) \propto \mu_{i\to a}(x_i)\hat{\mu}_{a\to i}(x_i) \overset{(4)}{=} \psi_i(x_i) \prod_{b\in\partial i} \hat{\mu}_{b\to i}(x_i). \tag{5}$$

The claim can be proved by induction. Next we revisit Example 1, where μ is given in (2). The factor graph in Figure 1 is a tree. Suppose we want to compute the marginal of x_3. We can do this starting from x_1 as follows. Since x_1 is a degree-one node, both $V_{1\to a_1}$ and $F_{1\to a_1}$ are empty. So $\nu_{1\to a_1}$ is uniform and $\mu_{1\to a_1}(x_1) \propto 1$. Continuing this using (3)–(4), we get

$$\begin{aligned}
\hat{\mu}_{a_1\to 2}(x_2) &\propto \sum_{x_1} \psi_{a_1}(x_1 x_2)\mu_{1\to a_1}(x_1) \propto \sum_{x_1} \psi_{a_1}(x_1 x_2) \\
\mu_{2\to a_2}(x_2) &\propto \psi_2(x_2)\hat{\mu}_{a_1\to 2}(x_2) = \hat{\mu}_{a_1\to 2}(x_2) \\
\hat{\mu}_{a_2\to 3}(x_3) &\propto \sum_{x_2} \psi_{a_2}(x_2 x_3)\mu_{2\to a_2}(x_2) \propto \sum_{x_2} \psi_{a_2}(x_2 x_3)\psi_{a_1}(x_1 x_2).
\end{aligned}$$

Finally, using (5),

$$\mu_3(x_3) \propto \psi_3(x_3)\hat{\mu}_{a_2\to 3}(x_3) = \hat{\mu}_{a_2\to 3}(x_3)$$

In general, we can use (3) and (4) as recursions then extract the marginals using (5), resulting in the following BP algorithms for trees:

Definition 1 (BP on trees)**.** *For each time k, $\{\hat{\mu}^k_{a\to i}, \mu^k_{i\to a} : i \in V, a \in F\}$ are called "beliefs", which are probability measures on $\mathcal{X}$. Given some initialization $\{\mu^0_{i\to a} : i \in V, a \in F\}$, we update $\{\hat{\mu}^0_{i\to a}, \mu^1_{i\to a}, \ldots\}$ in succession according to the following rule*[1]

$$\hat{\mu}^k_{a\to i}(x_i) \propto \sum_{x_{\partial a\setminus i}} \psi_a(x_{\partial a}) \prod_{j\in\partial a\setminus i} \mu^k_{j\to a}(x_j), \tag{6}$$

$$\mu^{k+1}_{i\to a}(x_i) \propto \psi_i(x_i) \prod_{b\in\partial i\setminus a} \hat{\mu}^k_{b\to i}(x_i), \tag{7}$$

and extract the marginal by

$$\mu^{k+1}_i(x_i) \propto \psi_i(x_i) \prod_{b\in\partial i} \hat{\mu}^k_{b\to i}(x_i). \tag{8}$$

Theorem 2. *For trees, BP algorithm converges to the true marginals after $2K$ iterations, where K is the diameter of the tree (length of the longest path). In other words, $\mu^k_i(x_i) = \mu_i(x_i)$ for all $i \in V$ and all $k \geq 2K$.*

3 Loopy BP on general graphs

Definition 1 and Theorem 2 hold for trees. Nevertheless, for general graphs, one can still consider Definition 1, known as loopy BP. Next let's look at an example in the context of linear regression. Just like Example 2, the factor graph is complete and not a tree.

Example 3 (LASSO with temperature β)**.** Consider

$$\mu_\beta(x) = \prod_{a=1}^{n} \underbrace{\exp\left\{-\frac{(y_a - \langle A_a, x\rangle)_2^2}{2}\right\}}_{\psi_a(x_{\partial a})} \prod_{i=1}^{d} \underbrace{\exp(-\beta\lambda|x_i|)}_{\psi_i(x_i)}.$$

[1] For continuous space, $\sum_{x_{\partial a\setminus i}}$ in (6) is replaced by $\int dx_{\partial a\setminus i}$.

Here $V = [d]$ and $F = [n]$. The BP update rule is given by

$$\hat{\mu}^k_{a\to i}(x_i) \propto \int_{\mathbb{R}^{d-1}} \prod_{j\neq i} dx_j \exp\left\{-\frac{(y_a - \langle A_a, x\rangle)_2^2}{2}\right\} \prod_{j\neq i} \mu^k_{j\to a}(x_j),$$
$$\mu^{k+1}_{i\to a}(x_i) \propto \exp(-\beta\lambda|x_i|) \prod_{b\neq a} \hat{\mu}^k_{b\to i}(x_i),$$

and the extracted marginal is

$$\mu_i^{k+1}(x_i) \propto \exp(-\beta\lambda|x_i|) \prod_{b\in[n]} \hat{\mu}^k_{b\to i}(x_i).$$

Remark 3.

- *Although there is no general theorem like Theorem 2, the hope that as $k \to \infty$, μ_i^{k+1} converges to some $\tilde{\mu}_i$. This limit however is in general not μ_i.*
- *In many cases, μ_i^{k+1} does converge empirically.*
- *This results in a practical algorithm if ψ_a and ψ_i are "simple". But this does not hold for LASSO (cannot integrate in close form).*

4 From BP to message passing algorithms

Note that in general each belief being updated in the BP algorithm is a probability distribution on $\mathcal{X}$ (such as a density). It will be more convenient to operate on the basis of real-valued messages. The idea of message passing algorithm is to approximate each belief by parametric distributions such as exponential family, then update the parameters. Consider $\mathcal{X} = \mathbb{R}$ and Gaussian approximation

Definition 4 (Message passing algorithm)**.** *For each k, $\{m^k_{i\to a}, v^k_{i\to a}, \hat{m}^k_{a\to i}, \hat{v}^k_{a\to i}\}$ are called "beliefs", which are real values. Define $\rho^k_{i\to a}(x_i)$ and $\hat{\rho}^k_{a\to i}(x_i)$ as the densities of $N(m^k_{i\to a}, v^k_{i\to a})$ and $N(\hat{m}^k_{a\to i}, \hat{v}^k_{a\to i})$ respectively, i.e.,*

$$\rho^k_{i\to a}(x_i) = \frac{1}{\sqrt{2\pi v^k_{i\to a}}} \exp\left\{-\frac{(x_i - m^k_{i\to a})^2}{2v^k_{i\to a}}\right\},$$
$$\hat{\rho}^k_{a\to i}(x_i) = \frac{1}{\sqrt{2\pi \hat{v}^k_{i\to a}}} \exp\left\{-\frac{(x_i - \hat{m}^k_{i\to a})^2}{2\hat{v}^k_{i\to a}}\right\}.$$

Given initialization $\{m^0_{i\to a}, v^0_{i\to a}, \hat{m}^0_{a\to i}, \hat{v}^0_{a\to i}\}$, compute

$$\hat{\gamma}^k_{a\to i}(x_i) \propto \int \psi_a(x_{\partial a}) \prod_{j\in\partial a\setminus i} \rho^k_{j\to a}(x_j) dx_{\partial a\setminus i}, \tag{9}$$
$$\gamma^{k+1}_{i\to a}(x_i) \propto \psi_i(x_i) \prod_{b\in\partial i\setminus a} \hat{\rho}^k_{b\to i}(x_i), \tag{10}$$

and update the messages as

$$(\hat{m}^k_{a\to i}, \hat{v}^k_{a\to i}) = \text{ mean and variance of } \hat{\gamma}^k_{a\to i}(x_i), \tag{11}$$
$$(m^k_{i\to a}, v^k_{i\to a}) = \text{ mean and variance of } \gamma^k_{i\to a}(x_i), \tag{12}$$

Finally, we extract the marginal as

$$\gamma_i^{k+1}(x_i) \propto \psi_i(x_i) \prod_{b\in\partial i} \hat{\rho}^k_{b\to i}(x_i). \tag{13}$$

Remark 5. *Why "Gaussian approximation"?*

- *The wrong intuition is that beliefs are approximately Gaussian. For example, for LASSO, this is due to non-Gaussian terms* $\psi_x(x_i) = \exp(-\beta\lambda|x_i|)$.
- *The correct intuition is that in the update rule, only means and variances of incoming beliefs are "important", so we can approximate the input beliefs by Gaussians (*ρ *and* $\hat{\rho}$*). But the output beliefs (*γ *and* $\hat{\gamma}$*) are non-Gaussian. For example, in* (13)*, the product of* $\hat{\rho}$*'s are Gaussian, but* $\psi_x(x_i)$ *is not.*