Mean Field Asymptotics in Statistical Learning.

Apr 7

Lecture 20. Derivation of AMP. I.

[Montanari, 2011]
Graphical model concepts in CS.

[Rangan, Schniter, Fletcher, 2016]
Vector AMP

[Donoho, Maleki, Montanari, 2009]
MP algorithm for CS: I. motivation and construction.

① The overall methodology.

Gibbs measure $\xrightarrow{\text{Factor graph}}$ Graphical model (Markov random field)

a standard mechanism.

Belief-propagation algorithm

exponential family approx of beliefs

Message-passing algorithm

Aggregating messages

Approximate message passing algorithm $\xrightarrow[\text{Rigorous.}]{\text{Probablistic analysis}}$ State evolution

② Markov random field.

Probability measure on $\Omega = \bigotimes_{i\in[N]} \mathcal{X}_i = \mathcal{X}^{\otimes N}$ $\quad \mu(x) \propto \exp\{-\beta H(x)\}$

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

$\partial a \subseteq V, \quad \forall a \in F$

$x_{\partial a} = (x_{i_1}, \cdots, x_{i_{|\partial a|}}) \in \mathcal{X}^{|\partial a|}$

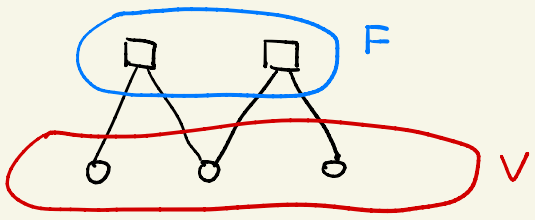
$$\psi_a : \mathcal{X}^{\otimes|\partial a|} \to \mathbb{R}_{\geq 0}, \quad a \in F$$

$$\psi_i : \mathcal{X} \to \mathbb{R}_{\geq 0}, \quad i \in V.$$

Factor graph visualization

$G = (V, F, E)$. $\quad |V| = N, \quad |F| = M,$

$E \subseteq \{(i,a) : i \in V, a \in F\}$.



Example 1: 1-D Ising model with 3 spins.

$$\Omega = \{\pm 1\}^3$$

$$\mu(x_1, x_2, x_3) \propto e^{-\beta(x_1x_2 + x_2x_3)} = \psi_{a_1}(x_1, x_2) \times \psi_{a_2}(x_2, x_3).$$

$$\partial a_1 = \{1, 2\}, \quad \partial a_2 = \{2, 3\}.$$

$$\psi_{a_1}(x_1, x_2) = e^{-\beta x_1 x_2}, \qquad \psi_{a_2}(x_2, x_3) = e^{-\beta x_2 x_3}$$

$$\psi_1(x_1) = \psi_2(x_1) = \psi_3(x_1) = 1.$$

Example 2: Bayes linear model.

$$y = Ax_0 + w \in \mathbb{R}^n$$

$x_0 \in \mathbb{R}^d$, $A \in \mathbb{R}^{n \times d}$, $w \in \mathbb{R}^n$,

$x_{0,i} \sim$ iid $p_0(x)$. $\quad w_i \sim$ iid $N(0, \sigma^2)$.

$$\mu(x) = p(x \mid A, y) \propto \exp\left\{-\frac{\|y - Ax\|_2^2}{2\sigma^2}\right\} \prod_{i=1}^{d} p_0(x_i)$$

$$= \prod_{a=1}^{n} \exp\left\{-\frac{(y_a - \langle A_a, x\rangle)^2}{2\sigma^2}\right\} \prod_{i=1}^{d} p_0(x_i)$$

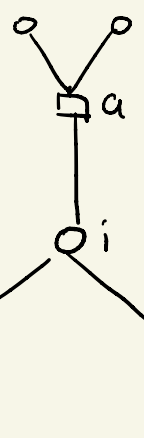
$$\psi_a(x) = \exp\left\{-\frac{(y_a - \langle A_a, x\rangle)^2}{2\sigma^2}\right\}$$

$$\psi_i(x_i) = p_0(x_i).$$

Remark: For $\mu \in \mathcal{P}(\Omega)$, the MRF representation is not unique.

Task: Compute the marginal distributions

$$\forall i \in [d], \quad \mu_i(x_i) \equiv \int \mu(x) \prod_{j \neq i} dx_j$$



③ Belief propagation algorithm on tree graphs

Let G be a tree factor graph.

include exclude

Denote $V_{a \to i}$ be all reachable vertices starting at "a" by blocking "i",

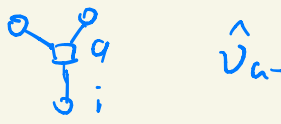
$V_{i \to a}$ - - - - - - - - "i" by blocking "a".

$F_{a \to i}$ - - - - factors starting at "a" by blocking "i"

$F_{i \to a}$ - - - - factors starting at "i" by blocking "a".

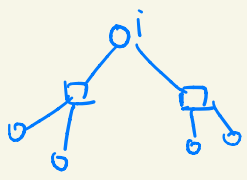
Define

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



$\hat{\nu}_{a \to i}(x)$

$$\nu_{i \to a}(x) \propto \prod_{b \in F_{i \to a}} \psi_b(x_{\partial b}) \prod_{j \in F_{i \to a}} \psi_j(x_j)$$



$\nu_{i \to a}(x)$

$$X_{\setminus i} = (x_1, x_2, \cdots x_{i-1}, x_{i+1}, \cdots x_d).$$

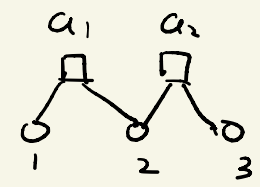
Define $\hat{\mu}_{a\to i}(x_i) \propto \sum_{x_{\setminus i}} \hat{\nu}_{a\to i}(x) \propto \sum_{x_{\setminus i}} \prod_{b\in F_{a\to i}} \psi_b(x_{\partial b}) \prod_{j\in F_{a\to i}} \psi_j(x_j) \in \mathcal{P}(\mathcal{X})$

$\mu_{i\to a}(x_i) \propto \sum_{x_{\setminus i}} \nu_{i\to a}(x) \propto \sum_{x_{\setminus i}} \prod_{b\in F_{i\to a}} \psi_b(x_{\partial b}) \prod_{j\in F_{i\to a}} \psi_j(x_j) \in \mathcal{P}(\mathcal{X})$

Claim :
$$\begin{cases} \mu_{i\to a}(x_i) \propto \psi_i(x_i) \prod_{b\in\partial i\setminus a} \hat{\mu}_{b\to i}(x_i) \\ \hat{\mu}_{a\to i}(x_i) \propto \sum_{x_{\partial a\setminus i}} \psi_a(x_{\partial a}) \prod_{j\in\partial a\setminus i} \mu_{j\to a}(x_j) \end{cases} \star$$

$$\mu_i(x_i) \propto \mu_{i\to a}(x_i)\hat{\mu}_{a\to i}(x_i) \propto \psi_i(x_i) \prod_{b\in\partial i} \hat{\mu}_{b\to i}(x_i)$$

Example : $\mu(x_1, x_2, x_3) \propto \psi_{a_1}(x_1, x_2) \cdot \psi_{a_2}(x_2, x_3)$



$\mu_{1\to a_1}(x_1) \propto 1.$

$\hat{\mu}_{a_1\to 2}(x_2) \propto \sum_{x_1} \psi_{a_1}(x_1, x_2)\, \mu_{1\to a_1}(x_1) = \sum_{x_1} \psi_{a_1}(x_1, x_2)$

$\mu_{2\to a_2}(x_2) \propto \hat{\mu}_{a_1\to 2}(x_2)\, \psi_2(x_2) = \hat{\mu}_{a_1\to 2}(x_2) = \sum_{x_1} \psi_{a_1}(x_1, x_2)$

$\hat{\mu}_{a_2\to 3}(x_3) \propto \sum_{x_2} \psi_{a_2}(x_2, x_3)\, \mu_{2\to a_2}(x_2) = \sum_{x_2} \psi_{a_2}(x_2, x_3) \sum_{x_1} \psi_{a_1}(x_1, x_2)$

$\mu_3(x_3) \propto \psi_3(x_3)\, \hat{\mu}_{a_2\to 3}(x_3) = \sum_{x_2} \psi_{a_2}(x_2, x_3) \sum_{x_1} \psi_{a_1}(x_1, x_2)$

$= \sum_{x_1, x_2} \psi_{a_1}(x_1, x_2)\, \psi_{a_2}(x_2, x_3).$

Def (BP on trees). (Make ★ iterative updates)

$\{\hat{\mu}^k_{a\to i}, \mu^k_{i\to a}\}_{a\in F, i\in V, k\geq 0} \subseteq \mathcal{P}(\mathcal{X})$ beliefs.

Update rule: Calculate $\{\hat{\mu}^{k+1}_{a\to i}, \mu^{k+1}_{i\to a}\}_{a,i}$ using $\{\hat{\mu}^k_{a\to i}, \mu^k_{i\to a}\}_{a,i}$.

$$\begin{cases} \mu^{k+1}_{i\to a}(x_i) \propto \psi_i(x_i) \prod_{b\in\partial i\setminus a} \hat{\mu}^k_{b\to i}(x_i) \\ \hat{\mu}^{k+1}_{a\to i}(x_i) \propto \sum_{x_{\partial a\setminus i}} \psi_a(x_{\partial a}) \prod_{j\in\partial a\setminus i} \mu^k_{j\to a}(x_j) \end{cases}$$

Extract marginals: $\mu^k_i(x_i) \propto \mu^k_{i\to a}(x_i)\, \hat{\mu}^k_{a\to i}(x_i) \propto \psi_i(x_i) \prod_{b\in\partial i} \hat{\mu}^k_{b\to i}(x_i).$

Initialization: $\hat{\mu}^0_{a\to i}(x_i)$, $\mu^0_{i\to a}(x_i)$: arbitrary, often uniform.

Thm: BP algorithm on trees will give exact results after $2K$ iterations.

where K is diameter of the tree (length of longest path)

$\mu^{2K}_i(x_i) = \mu_i(x_i)$ $\mu^{2K}_{i\to a}(x_i) = \mu_{i\to a}(x_i)$. $\hat{\mu}^{2K}_{a\to i}(x_i) = \hat{\mu}_{a\to i}(x_i)$.

④ Loopy belief propagation on general graphs.

Def (Loopy BP).

$$\{\hat{\mu}^k_{a\to i}, \mu^k_{i\to a}\}_{a\in F, i\in V, k\geq 0} \subseteq \mathcal{P}(X) \quad \text{beliefs.}$$

Update rule:

$$\begin{cases} \mu^{k+1}_{i\to a}(x_i) \propto \psi_i(x_i) \prod_{b\in \partial i\setminus a} \hat{\mu}^k_{b\to i}(x_i) \\ \hat{\mu}^{k+1}_{a\to i}(x_i) \propto \sum_{x_{\partial a\setminus i}} \psi_a(x_{\partial a}) \prod_{j\in \partial a\setminus i} \mu^k_{j\to a}(x_j) \end{cases}$$

Extract marginals: $\mu^k_i(x_i) \propto \mu^k_{i\to a}(x_i)\hat{\mu}^k_{a\to i}(x_i) \propto \psi_i(x_i) \prod_{b\in\partial i} \hat{\mu}^k_{b\to i}(x_i)$.

Example: LASSO w/ temperature β

$$\begin{cases} \mu^{k+1}_{i\to a}(x_i) \propto \prod_{b\in\partial i\setminus a} \hat{\mu}^k_{b\to i}(x_i) \times \exp\{-\beta\lambda|x_i|\} \\ \hat{\mu}^{k+1}_{a\to i}(x_i) \propto \int dx_{\sim i} \exp\left\{-\frac{\beta}{2}(y_a - \langle A_a, x\rangle)^2\right\} \prod_{j\in\partial a\setminus i} \mu^k_{j\to a}(x_j) \end{cases}$$

$$\mu^k_i(x_i) \propto \prod_{b\in\partial i} \hat{\mu}^k_{b\to i}(x_i) \times \exp\{-\beta\lambda|x_i|\}$$

Remark:

- ⊗ We hope $\mu^k_i(x_i) \to \mu_i(x_i)$ as $k\to\infty$.
- ⊗ For many models, especially mean field models, this happens in practice.
- ⊗ Practical when there is closed form update.
- ⊗ When ψ_a, ψ_i's are not simple, no closed form update, Not practical.

⑤ From BP to message passing algorithm.

Consider the case when $X = \mathbb{R}$.

Gaussian approximation of (exponential family approx).

Wrong intuition: the beliefs are approximately Gaussian.

Intuition: In the update rule, only the mean and var of input beliefs are important.

Input beliefs can be approximated by Gaussian dist. in the update rule.

Real belief is still non-Gaussian.

Def (Message passing algorithm)

mean and variance of beliefs.

$$\{ m_{i\to a}^k,\ v_{i\to a}^k,\ \hat{m}_{a\to i}^k,\ \hat{v}_{a\to i}^k \}_{a\in F,\ i\in V,\ k\geq 0} \subseteq \mathbb{R}.$$ Messages.

Update rule: Calculating $\{ m_{i\to a}^{k+1},\ v_{i\to a}^{k+1},\ \hat{m}_{a\to i}^{k+1},\ \hat{v}_{a\to i}^{k+1} \}$ using $\{ m_{i\to a}^k, v_{i\to a}^k, \hat{m}_{a\to i}^k, \hat{v}_{a\to i}^k \}$.

$$\rho_{i\to a}^k(x_i) = \frac{1}{\sqrt{2\pi v_{i\to a}^k}} \exp\left\{ -\frac{(x_i - m_{i\to a}^k)^2}{2 v_{i\to a}^k} \right\}.$$ density of $\mathcal{N}(m_{i\to a}^k, v_{i\to a}^k)$

$$\hat{\rho}_{a\to i}^k(x_i) = \frac{1}{\sqrt{2\pi \hat{v}_{a\to i}^k}} \exp\left\{ -\frac{(x_i - \hat{m}_{a\to i}^k)^2}{2 \hat{v}_{a\to i}^k} \right\}.$$ density of $\mathcal{N}(\hat{m}_{a\to i}^k, \hat{v}_{a\to i}^k)$

$$r_{i\to a}^{k+1}(x_i) \propto \psi_i(x_i) \prod_{b\in \partial i\setminus a} \hat{\rho}_{b\to i}^k(x_i)$$

$$\hat{r}_{a\to i}^{k+1}(x_i) \propto \sum_{x_{\partial a\setminus i}} \psi_a(x_{\partial a}) \prod_{j\in \partial a\setminus i} \rho_{j\to a}^k(x_j)$$

$$(m_{i\to a}^{k+1}, v_{i\to a}^{k+1}) = (\text{mean}, \text{variance}) \text{ under } r_{i\to a}^{k+1}(x_i)$$

$$(\hat{m}_{a\to i}^{k+1}, \hat{v}_{a\to i}^{k+1}) = (\text{mean}, \text{variance}) \text{ under } \hat{r}_{a\to i}^{k+1}(x_i)$$

Extract marginals:

$$r_i^k(x_i) \propto \psi_i(x_i) \prod_{b\in \partial i} \hat{\rho}_{b\to i}^{k-1}(x_i)$$

$\{ \rho_{i\to a}^k, \hat{\rho}_{a\to i}^k \}$ input beliefs, assumed to be Gaussian

$\{ r_{i\to a}^k, \hat{r}_{a\to i}^k \}$ output beliefs, possibly non-Gaussian

Example: LASSO w/ temperature β.

$$\otimes \quad r_{i\to a}^{k+1}(x_i) \propto \exp\{-\beta\lambda|x_i|\} \times \exp\left\{ -\sum_{b\neq a} \frac{(x_i - \hat{m}_{b\to i}^k)^2}{2\hat{v}_{b\to i}^k} \right\}$$

$$\propto \exp\left\{ -\beta \left(\frac{(x_i - \theta_{i\to a}^k)^2}{2\zeta_{i\to a}^k} + \lambda|x_i| \right) \right\}.$$

$$\begin{cases} \beta(\zeta_{i\to a}^k)^{-1} = \sum_{b\neq a} (\hat{v}_{b\to i}^k)^{-1} \\ \beta(\zeta_{i\to a}^k)^{-1} \theta_{i\to a}^k = \sum_{b\neq a} (\hat{v}_{b\to i}^k)^{-1} \hat{m}_{b\to i}^k \\ m_{i\to a}^{k+1} = \mathbb{E}_{X_i \sim \pi(\beta, \lambda, \theta_{i\to a}^k, \zeta_{i\to a}^k)}[X_i] \\ v_{i\to a}^{k+1} = \mathrm{Var}_{X_i \sim \pi(\beta, \lambda, \theta_{i\to a}^k, \zeta_{i\to a}^k)}[X_i] \end{cases}$$

where $\pi(\beta, \lambda, \theta, \zeta) \sim \exp\left\{ -\beta \left(\frac{(x-\theta)^2}{2\zeta} + \lambda|x| \right) \right\}$.

⊗ $$\hat{\nu}^{k+1}_{a\to i}(x_i) \propto \int_{\mathbb{R}^{d-1}} \exp\left\{-(\beta/2)\left(y_a - \langle A_a, x\rangle\right)^2\right\} \times \exp\left\{-\sum_{j\neq i} \frac{(x_j - m^k_{j\to a})^2}{2\nu^{k+1}_{j\to a}}\right\} dx_{\sim i}.$$

This is a Gaussian integration, which gives a Gaussian density.

$$\propto \exp\left\{-\frac{(x_i - \hat{m}^{k+1}_{a\to i})^2}{2\hat{\nu}^{k+1}_{a\to i}}\right\}.$$

$$\begin{cases} \hat{m}^{k+1}_{a\to i} = \dfrac{y_a - \sum_{j\neq i} A_{aj} m^k_{j\to a}}{A_{ai}} \\ \hat{\nu}^{k+1}_{a\to i} = \left(\sum_{j\neq i} A^2_{aj} \nu^k_{j\to a} + 1/\beta\right) / A^2_{ai}. \end{cases}$$

⊗ Simplification as $\beta \to \infty$.

$$\lim_{\beta\to\infty} \mathbb{E}_{x\sim\pi(\beta,\lambda,\theta,s)}[x] = \arg\min_x \left\{\frac{(x-\theta)^2}{2s} + \lambda|x|\right\}$$

$$= \eta(\theta; \lambda s).$$

$$\lim_{\beta\to\infty} \beta \cdot \mathrm{Var}_{x\sim\pi(\beta,\lambda,\theta,s)}[x] = \begin{cases} s, & \text{if } |\theta| > \lambda s \\ 0, & \text{if } |\theta| < \lambda s \end{cases}$$

$$= s \cdot \eta'(\theta; \lambda s).$$

We are short of alphabets. Abuse notation:

$$\hat{m}_{a\to i} \leftarrow \lim_{\beta\to\infty} \hat{m}_{a\to i}, \quad m_{i\to a} \leftarrow \lim_{\beta\to\infty} m_{i\to a}$$

$$\hat{\nu}_{a\to i} \leftarrow \lim_{\beta\to\infty} \beta\hat{\nu}_{a\to i}, \quad \nu_{i\to a} \leftarrow \lim_{\beta\to\infty} \nu_{i\to a}.$$

$$\begin{cases} (s^k_{i\to a})^{-1} = \sum_{b\neq a} (\hat{\nu}^k_{b\to i})^{-1} \\ (s^k_{i\to a})^{-1}\theta^k_{i\to a} = \sum_{b\neq a} (\hat{\nu}^k_{b\to i})^{-1} \hat{m}^k_{b\to i} \\ m^{k+1}_{i\to a} = \eta(\theta^k_{i\to a}; \lambda s^k_{i\to a}) \\ \nu^{k+1}_{i\to a} = s^k_{i\to a} \cdot \eta'(\theta^k_{i\to a}, \lambda s^k_{i\to a}). \\ \hat{m}^{k+1}_{a\to i} = \dfrac{y_a - \sum_{j\neq i} A_{aj} m^k_{j\to a}}{A_{ai}} \\ \hat{\nu}^{k+1}_{a\to i} = \left(\sum_{j\neq i} A^2_{aj} \nu^k_{j\to a} + 1\right) / A^2_{ai}. \end{cases}$$

Define $\quad \tau_{a\to i}^{k+1} = A_{ai}^2 \hat{v}_{a\to i}^{k+1} = \sum_{j\neq i} A_{aj}^2 v_{j\to a}^{k} + 1$

$$z_{a\to i}^{k+1} = A_{ai} \hat{m}_{a\to i}^{k+1} = y_a - \sum_{j\neq i} A_{aj} m_{j\to a}^{k} .$$

$$\Rightarrow \quad \theta_{i\to a}^{k} = \frac{\sum_{b\neq a} A_{bi} z_{b\to i}^{k} / \tau_{b\to i}^{k}}{\sum_{b\neq a} A_{bi}^2 / \tau_{b\to i}^{k}}$$

$$\zeta_{i\to a}^{k} = \frac{1}{\sum_{b\neq a} A_{bi}^2 / \tau_{b\to i}^{k}}$$