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 $\xrightarrow{\text{exponential family approx of beliefs}}$ Message-passing algorithm

Aggregating messages

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

② Markov random field.

$V = [N]$ $F = [M]$.

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

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

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

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

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

$a \in F$, ∂a

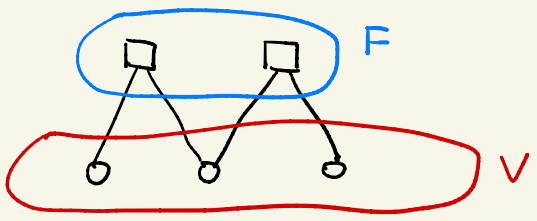
$i \in V$, ∂i

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

Factor graph visualization

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

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



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

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

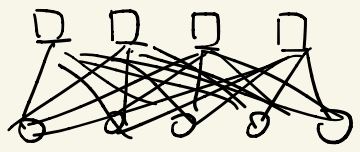
$$\mu(x_1, x_2, x_3) \propto e^{-\beta(x_1x_2 + x_2x_3)} = \psi_{a_1}(x_1, x_2) \times \psi_{a_2}(x_2, x_3).$$

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

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

$$\psi_1(x_1) = \psi_2(x_2) = \psi_3(x_3) = 1.$$

Example 2: Bayes linear model.



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

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

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

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

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

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

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

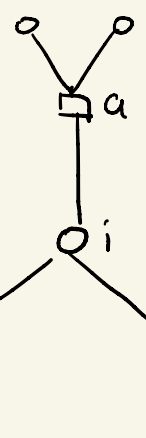
$\psi(x) = \mu(x)$

$\mu(x) \propto \psi(x)$

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",

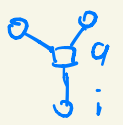
$V_{i \to a}$ - - - - - - - - "i" by blocking "a".

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

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

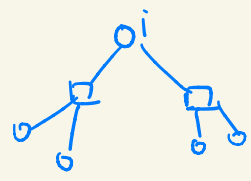
Define

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



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

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



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

$x_{\sim i} = (x_1, \cdots x_{i-1}, x_{i+1}, \cdots x_N)$

$$\hat{\mu}_{a\to i}(x_i) \propto \sum_{x_{\sim i}} \hat{\nu}_{a\to i}(x) \in \mathcal{P}(\mathcal{X})$$

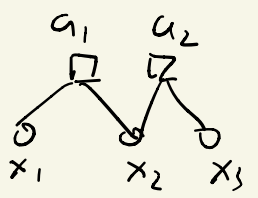
$$\mu_{i\to a}(x_i) \propto \sum_{x_{\sim i}} \nu_{i\to a}(x) \in \mathcal{P}(\mathcal{X}).$$

Claim:

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

$$\mu_i(x_i) \equiv \int \mu(x)\, dx_{\sim i} \propto \mu_{i\to a}(x_i)\, \hat{\mu}_{a\to i}(x_i) \propto \psi_i(x_i) \prod_{b\in\partial i} \hat{\mu}_{b\to i}(x_i)$$

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



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

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

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

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

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

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

Def (Belief propagation algorithm, on trees).

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

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

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

Extract marginal: $\mu^{k+1}_i(x_i) \propto \psi_i(x_i) \prod_{b\in\partial i} \hat{\mu}^k_{b\to i}(x_i)$

Thm: BP algorithm on trees will be exact after $2K$ iteration

K is the diameter of the tree.

$$\mu_i^{2K}(x_i) = \mu_i(x_i) \qquad \mu_{i\to a}^{2K}(x_i) = \mu_{i\to a}(x_i), \quad \hat{\mu}_{a\to i}^{2K}(x_i) = \hat{\mu}_{a\to i}(x_i)$$

④ Loopy BP algorithm on general graphs.

Def (Loopy BP).

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

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

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

Extract marginal: $\mu_i^{k+1}(x_i) \propto \psi_i(x_i) \prod_{b\in\partial i} \hat{\mu}_{b\to i}^k(x_i)$

Example: LASSO. w/ temperature β.

$$\mu_\beta(x) = \prod_{a=1}^n \exp\{-(\beta/2)(y_a - \langle A_a, x\rangle)^2\} \cdot \prod_{i=1}^d \exp\{-\beta\lambda|x_i|\}.$$

$$\begin{cases} \hat{\mu}_{a\to i}^{k+1}(x_i) \propto \int_{\mathbb{R}^{d-1}} \exp\{-(\beta/2)(y_a - \langle A_a, x\rangle)^2\} \prod_{j\in\partial a\setminus i} \mu_{j\to a}^k(x_j) \prod_{j\neq i} dx_j. \\ \mu_{i\to a}^{k+1}(x_i) \propto \exp\{-\beta\lambda|x_i|\} \times \prod_{b\in\partial i\setminus a} \hat{\mu}_{b\to i}^k(x_i) \end{cases}$$

$$\mu_i^{k+1}(x_i) \propto \prod_{b\in\partial i} \hat{\mu}_{b\to i}^k(x_i) \times \exp\{-\beta\lambda|x_i|\}.$$

Remark:

- We hope $\mu_i^k(x_i) \to \tilde{\mu}_i(x_i)$ as $k\to\infty$.
- This will happen in practice. in many case.
- Practical when ψ_a ψ_i, are simple.
- Not very practical for LASSO.

⑤ From BP to message passing. ($\mathcal{X} = \{\pm 1\}$, $\mathcal{X} = \mathbb{R}^k$)

Consider $\mathcal{X} = \mathbb{R}$

Gaussian approx.

Wrong intuition: the beliefs are approximately Gaussian.

Intuition: In the update rule, only mean and variance of incoming beliefs are important.

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

Real beliefs are still non-Gaussian.

Def (Message passing).

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

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