STAT260 Mean Field Asymptotics in Statistical Learning Lecture 21 - 04/12/2021

Lecture 21: Derivation of AMP II

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

1 MP for LASSO

Previously we have defined the following Message Passing (MP) algorithm for a general Markov random field:

Definition 1 (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.,*

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

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

$$\hat{\gamma}^k_{a\to i}(x_i) \propto \int \psi_a(x_{\partial a}) \prod_{j\in\partial a\setminus i} \rho^k_{j\to a}(x_j) dx_{\partial a\setminus i}, \tag{1}$$
$$\gamma^{k+1}_{i\to a}(x_i) \propto \psi_i(x_i) \prod_{b\in\partial i\setminus a} \hat{\rho}^k_{b\to i}(x_i), \tag{2}$$

and update the messages as

$$(\hat{m}^k_{a\to i}, \hat{v}^k_{a\to i}) = \text{ mean and variance of } \hat{\gamma}^k_{a\to i}(x_i), \tag{3}$$
$$(m^k_{i\to a}, v^k_{i\to a}) = \text{ mean and variance of } \gamma^k_{i\to a}(x_i). \tag{4}$$

Finally, we extract the marginal as

$$\gamma^{k+1}_i(x_i) \propto \psi_i(x_i) \prod_{b\in\partial i} \hat{\rho}^k_{b\to i}(x_i). \tag{5}$$

A very important property of this algorithm is that it is "non-backtracking", meaning that the message that a sends to i does not depend on the message that i sent to a in the previous iteration and vice versa. Note, however, that the expression for the final output includes the product over the whole ∂i.

1.1 LASSO with finite temperature

In this lecture we apply that procedure to the LASSO problem:

Example 1 (LASSO with temperature β)**.** Consider

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

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

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

and the extracted marginal is

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

We need to understand what are the means and variances of distributions $\gamma^k_{a\to i}$ and $\hat{\gamma}^k_{i\to a}$. Let's start with $\gamma^k_{a\to i}$. Plugging in the expressions for LASSO in the MP we get

$$\begin{aligned}
\gamma^{k+1}_{i\to a}(x_i) &\propto \psi_i(x_i) \prod_{b\in\partial i\setminus a} \hat{\rho}^k_{b\to i}(x_i) \\
&\propto \exp(-\beta\lambda|x_i|) \times \exp\Big\{-\sum_{b\neq a} \frac{(x_i - \hat{m}^k_{b\to i})^2}{2\hat{v}^k_{b\to i}}\Big\} \\
&\propto \exp\Big\{-\beta\Big(\frac{(x_i - \theta^k_{i\to a})^2}{2\zeta^k_{i\to i}} + \lambda|x_i|\Big)\Big\},
\end{aligned}$$

where $\theta^k_{i\to a}$ and $\zeta^k_{i\to i}$ are mean and variance of the Gaussian distribution with density, proportional to

$$\exp\Big\{-\sum_{b\neq a} \frac{(x_i - \hat{m}^k_{b\to i})^2}{2\beta\hat{v}^k_{b\to i}}\Big\}.$$

The exact expressions are the following

$$\beta(\zeta^k_{i\to a})^{-1} = \sum_{b\neq a} (\hat{v}^k_{b\to i})^{-1}, \tag{6}$$
$$\beta(\zeta^k_{i\to a})^{-1}\theta^k_{i\to a} = \sum_{b\neq a} (\hat{v}^k_{b\to i})^{-1}\hat{m}^k_{b\to i}, \tag{7}$$
$$m^{k+1}_{i\to a} = \mathbb{E}_{x_i\sim\pi(\beta,\lambda,\theta^k_{i\to a},\zeta^k_{i\to a})}[x_i], \tag{8}$$
$$v^{k+1}_{i\to a} = \mathrm{Var}_{x_i\sim\pi(\beta,\lambda,\theta^k_{i\to a},\zeta^k_{i\to a})}[x_i], \tag{9}$$
$$\tag{10}$$

where

$$\pi(\beta,\lambda,\theta,\zeta) \propto \exp\Big\{-\beta\Big(\frac{(x-\theta)^2}{2\zeta} + \lambda|x|\Big)\Big\}. \tag{11}$$

The expression for $\hat{\gamma}^k_{a\to i}$ turns out to be a bit simpler:

$$\hat{\gamma}^k_{a\to i}(x) \propto \int_{\mathbb{R}^{d-1}} \underbrace{\exp\Big\{\frac{\beta}{2}(y_a - \langle A_a, x\rangle)^2\Big\}}_{\psi_a(x_{\partial a})} \times \underbrace{\exp\Big\{-\sum_{j\neq i} \frac{x_j - m^k_{j\to a})^2}{2v^k_{j\to a}}\Big\}}_{\prod_{j\in\partial a\setminus i}\rho^k_{j\to a}(x_j)} \mathrm{d}x_{\setminus i}.$$

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

$$\hat{\gamma}^k_{a\to i}(x) \propto \exp\Big\{-\sum_{j\neq i} \frac{(x_i - \hat{m}^{k+1}_{a\to i})^2}{2\hat{v}^{k+1}_{a\to i}}\Big\}$$

Because of that the result is explicit: mean and variance are affine functions of means and variances of incoming messages correspondingly.

$$A_{ai} \cdot \hat{m}^k_{a\to i} = y_a - \sum_{j\neq i} A_{aj} m^k_{j\to a}, \tag{12}$$

$$A^2_{ai} \cdot \hat{v}^k_{a\to i} = \sum_{j\neq i} A^2_{aj} v^k_{j\to a} + \frac{1}{\beta}. \tag{13}$$

1.2 LASSO with infinite temperature

The formulas above give the AMP algorithm with finite β. This is sufficient in Bayesian setting with the corresponding prior. However, for the vanilla LASSO we need to send β to infinity. Recall the definition of π in Equation 11. Let's write out the following for convenience:

$$\frac{d}{dx}\Big(\frac{(x-\theta)^2}{2\zeta} + \lambda|x|\Big) = \begin{cases} \zeta^{-1}(x-\theta-\lambda\zeta), & x<0, \\ \zeta^{-1}(x-\theta+\lambda\zeta), & x>0. \end{cases}$$

By Laplace approximation we have for expectation

$$\begin{aligned} \lim_{\beta\to\infty} \mathbb{E}_{X\sim\pi(\beta,\lambda,\theta,\zeta)}[X] = &\ \arg\min_x \Big\{ \frac{(x-\theta)^2}{2\zeta} + \lambda|x| \Big\} \\ = &\begin{cases} \theta+\lambda\zeta, & \text{if } \theta < -\lambda\zeta, \\ \theta-\lambda\zeta, & \text{if } \theta > \lambda\zeta, \\ 0, & \text{else;} \end{cases} \\ = &\ \eta(\theta;\lambda\zeta) \text{ — soft-threshold.} \end{aligned}$$

When it comes to variance, we look at the second derivative at the minimum:

$$\frac{d^2}{dx^2}\Big(\frac{(x-\theta)^2}{2\zeta} + \lambda|x|\Big)_{x=\eta(\theta;\lambda\zeta)}.$$

We see that that second derivative is ζ^{-1} if $|\theta| > \lambda\zeta$, but the function is not twice differentiable at the minimum for $|\theta| \le \lambda\zeta$, i.e. the second derivative is infinite. Thus we obtain

$$\begin{aligned} \lim_{\beta\to\infty} \mathrm{Var}_{X\sim\pi(\beta,\lambda,\theta,\zeta)}[X] \times \beta = &\begin{cases} \zeta, & \text{if } |\theta| > \lambda\zeta, \\ 0, & \text{if } |\theta| < \lambda\zeta; \end{cases} \\ = &\ \zeta \cdot \eta'(\theta;\lambda\zeta). \end{aligned}$$

In view of these asymptotic relations, we **change notation:**

$$m^k_{i\to a} \leftarrow \lim_{\beta\to\infty} m^k_{i\to a}, \quad \hat{m}^k_{i\to a} \leftarrow \lim_{\beta\to\infty} \hat{m}^k_{a\to i}$$
$$v^k_{i\to a} \leftarrow \lim_{\beta\to\infty} \beta v^k_{i\to a}, \quad \hat{v}^k_{i\to a} \leftarrow \lim_{\beta\to\infty} \beta \hat{v}^k_{a\to i}$$

Plugging this into equations 6–9, 12, 13 gives

$$
\begin{aligned}
(\zeta_{i\to a}^k)^{-1} &= \sum_{b\neq a}(\hat{v}_{b\to i}^k)^{-1},\\
(\zeta_{i\to a}^k)^{-1}\theta_{i\to a}^k &= \sum_{b\neq a}(\hat{v}_{b\to i}^k)^{-1}\hat{m}_{b\to i}^k,\\
m_{i\to a}^{k+1} &= \eta(\theta_{i\to a}^k;\lambda\zeta_{i\to a}^k),\\
v_{i\to a}^{k+1} &= \zeta_{i\to a}^k\cdot\eta'(\theta_{i\to a}^k;\lambda\zeta_{i\to a}^k),\\
A_{ai}\cdot\hat{m}_{a\to i}^k &= y_a - \sum_{j\neq i}A_{aj}m_{j\to a}^k,\\
A_{ai}^2\cdot\hat{v}_{a\to i}^k &= \sum_{j\neq i}A_{aj}^2 v_{j\to a}^k + 1.
\end{aligned}
$$

So far we treated matrix $\boldsymbol{A}$ as a given design matrix. To further analyze the AMP for LASSO we will need to impose some assumptions on it. Eventually we would like to have $A_{ai}\sim_{i.i.d.}\mathcal{N}(0,1/n)$. However, for now we assume $A_{ai}\sim_{i.i.d.}\mathrm{Unif}(\{\pm 1/\sqrt{n}\})$, so we can replace A_{ai}^2 by $1/n$. The intuition for this is that by universality the distribution of A_{ai} doesn't matter as long we match the first and the second moments. An alternative explanation is that the objects of interest (such as $\sum_{j\neq i}A_{aj}^2 v_{j\to a}^k$) should concentrate around their expectations, and therefore only the moments of A_{ai} matter.

Let's make one more slight change of variables:

$$
\begin{aligned}
z_{a\to i}^k &\equiv A_{ai}\cdot\hat{m}_{a\to i}^k = y_a - \sum_{j\neq i}A_{aj}m_{j\to a}^k,\\
\tau_{a\to i}^k &\equiv A_{ai}^2\hat{v}_{a\to i}^k = \sum_{j\neq i}A_{aj}^2 v_{j\to a}^k + 1.
\end{aligned}
$$

Our MP equations become:

$$
\theta_{i\to a}^k = \frac{\sum_{b\neq a}A_{bi}z_{b\to i}^k/\tau_{b\to i}^k}{\frac{1}{n}\sum_{b\neq a}1/\tau_{b\to i}^k}, \tag{14}
$$

$$
m_{i\to a}^{k+1} = \eta(\theta_{i\to a}^k;\lambda\zeta_{i\to a}^k), \tag{15}
$$

$$
z_{a\to i}^k = y_a - \sum_{j\neq i}A_{aj}m_{j\to a}^k, \tag{16}
$$

$$
\zeta_{i\to a}^k = \Big(\frac{1}{n}\sum_{b\neq a}1/\tau_{b\to i}^k\Big)^{-1}, \tag{17}
$$

$$
v_{i\to a}^{k+1} = \zeta_{i\to a}^k\cdot\eta'(\theta_{i\to a}^k;\lambda\zeta_{i\to a}^k), \tag{18}
$$

$$
\tau_{a\to i}^k = \frac{1}{n}\sum_{j\neq i}v_{j\to a}^k + 1. \tag{19}
$$

These give the Message Passig algorithm for LASSO. Note that since we have $i\in V$, $a\in F$, $|V|=n$ and $|F|=d$, the number of messages is of order dn.

2 From message passing to approximate message passing.

Our goal in this section is to simplify MP ($2\times n\times d$ # of messages). After simplification there are going to be only d messages. First we will introduce the main idea that leads to such simplification in Section 2.1. However, the straightforward application of this idea does not lead to a good approximation. We will then show how to relax it and make it work in Section 2.2

2.1 Crude approximation of MP

Observe that the source of such a large number of messages is a non-backtracking property. For example, we only need to track the dependence of $\theta^k_{i\to a}$ on a because we don't want to account for the message from a in its update rule. A simple idea that comes to mind is to give up on this property and to hope that adding one extra term to each summation will not influence much. This leads to the following crude approximation:

$$\begin{aligned}
\theta_i^k =& \frac{\sum_b A_{bi} z_b^k/\tau_b^k}{\frac{1}{n}\sum_b 1/\tau_b^k},\\
m_i^{k+1} =& \eta(\theta_i^k; \lambda\zeta_i^k),\\
z_a^k =& y_a - \sum_j A_{aj} m_j^k,\\
\zeta_i^k =& \Big(\frac{1}{n}\sum_b 1/\tau_b^k\Big)^{-1},\\
v_i^{k+1} =& \zeta_i^k \cdot \eta'(\theta_i^k, \lambda\zeta_i^k),\\
\tau_a^k =& \frac{1}{n}\sum_j v_j^k + 1,
\end{aligned}$$

From the last equation we see that $\tau_a^k =: \tau^k$ — does not depend on a. This propagates to the equation for ζ_i^k and gives $\zeta_i^k = \tau^k =: \zeta^k$ — does not depend on i. τ^k also cancels in the equation for θ_i^k. After these simplifications we only get 3 equations:

$$\boldsymbol{m}^{k+1} = \eta(\boldsymbol{m}^k + \boldsymbol{A}^\mathsf{T}\boldsymbol{z}^k; \lambda\zeta^k) \qquad \in \mathbb{R}^d, \tag{20}$$

$$\boldsymbol{z}^k = \boldsymbol{y} - \boldsymbol{A}\boldsymbol{m}^k \text{(without Onsager term)} \qquad \in \mathbb{R}^n, \tag{21}$$

$$\zeta^{k+1} = \zeta^k \times \frac{1}{n}\sum_{i\in[n]} \eta'(\boldsymbol{m}^k + \boldsymbol{A}\boldsymbol{z}^k; \lambda\zeta^k)_i + 1 \qquad \in \mathbb{R}. \tag{22}$$

It turns out that the approximation above is too crude, so it needs a correction. The equations for ζ, v and τ (the variables that correspond to variances of the messages) are fine. The equations for θ, m, z (variables that correspond to means) are off by $O(1)$. We cannot drop non-backtracking property in equations for those variables.

2.2 Derivation of AMP from MP

As we mentioned before, we only want to drop the non-backtracking property in equations for ζ, v and τ. When we replace $\tau^k_{i\to a}$ by τ^k, $\zeta^k_{i\to a}$ by ζ^k and v_i^k by v_k, the equations 17, 18 and 19 can be combined into a single update rule

$$\zeta^{k+1} = \zeta^k \times \frac{1}{n}\sum_{i\in[n]} \eta'(m^k + \boldsymbol{A}\boldsymbol{z}^k; \lambda\zeta^k)_i + 1.$$

So far, the equations 14, 15 and 16 become

$$\begin{aligned}
\theta^k_{i\to a} &= \sum_{b\neq a} A_{bi} z^k_{b\to i},\\
m^k_{i\to a} &= \eta(\theta^k_{i\to a}; \lambda\zeta^k),\\
z^k_{a\to i} &= y_a - \sum_{j\neq i} A_{aj} m^k_{j\to a},
\end{aligned}$$

(equations for m and z remain unchanged yet, in the equation for θ we cancelled τ^k and plugged in $(n-1)/n \approx 1$ in the denominator).

Now we would like to deal with the variables θ, m, z. The idea to completely eliminate non-backtracking property does not work for these variables, so let's try to relax it a little bit. For example, instead of fully dropping the dependence of $\theta_{i\to a}$ on a, let's say that this dependence results in a small correction to the update rule. More precisely, we would like to introduce the following decomposition:

$$\theta^k_{i\to a} = \theta^k_i + \delta\theta^k_{i\to a}, \tag{23}$$
$$m^k_{i\to a} = m^k_i + \delta m^k_{i\to a}, \tag{24}$$
$$z^k_{a\to i} = z^k_a + \delta z^k_{a\to i}, \tag{25}$$
$$\tag{26}$$

where $\delta\theta$, δm and δz are terms of smaller order of magnitude than θ, m and z correspondingly.

Defining $\delta\theta^k_{i\to a}$ is rather straightforward:

$$\theta^k_{i\to a} = \underbrace{\sum_b A_{bi} z^k_{b\to i}}_{\theta^k_i} - \underbrace{A_{ai} z^k_{a\to i}}_{\delta\theta^k_{i\to a}}.$$

Note that because of CLT we can think of θ^k_i as of something of order $O(1)$, while $\delta\theta^k_{i\to a}$ is $O(1/\sqrt{n})$. So it is indeed a small correction. The update rules are as follows:

$$\begin{cases} \theta^k_i = \sum_b A_{bi} z^k_{b\to i} = \sum_b A_{bi} z^k_b + \sum_b A_{bi}\delta z^k_{b\to i}, \\ \delta\theta^k_{i\to a} = -A_{ai} z^k_{a\to i} = -A_{ai}(z^k_i + \delta z^k_{a\to i}). \end{cases}$$

The derivation for m is a bit more involved because we need to push small correction $\delta\theta^k_{i\to a}$ through non-linear function η:

$$m^{k+1}_{i\to a} = \underbrace{\eta(\theta^k_i;\lambda\zeta^k)}_{m^{k+1}_i} + \underbrace{\eta'(\theta^k_i;\lambda\zeta^k)\cdot\delta\theta^k_{i\to a}}_{\delta m^{k+1}_i},$$

$$\begin{cases} m^{k+1}_i = \eta(\theta^k_i;\lambda\zeta^k), \\ \delta m^{k+1}_{i\to a} = \eta'(\theta^k_i,\lambda\zeta^k)\cdot\delta\theta^k_{i\to a}. \end{cases}$$

Finally, we derive the decomposition for z. It is quite similar to that for θ:

$$z^k_{a\to i} = \underbrace{y_a - \sum_b A_{bj} m^k_{j\to a}}_{z^k_a} + \underbrace{A_{ai} m^k_{i\to a}}_{\delta z^k_{a\to i}}$$

$$\begin{cases} z^k_a = y_a - \sum_j A_{aj} m^k_{j\to a} = y_a - \sum_j A_{aj} m^k_j - \sum_j A_{aj}\delta m^k_{j\to a}, \\ \delta z^k_{a\to i} = A_{ai} m^k_{i\to a} = A_{ai}(m^k_i + \delta m^k_{i\to a}). \end{cases}$$

So far we didn't actually change much: by introducing the small correction terms we didn't decrease the number of updates, but only emphasized which quantities in those updates may be negligible. Now we are in a good shape to make some simplifying approximations. We have already stated that simply neglecting $\delta\theta^k_{i\to a}$, $\delta m^{k+1}_{i\to a}$ and $\delta z^k_{a\to i}$ in the updates of θ^k_i, m^{k+1}_i and z^k_a is too crude. So let's try to neglect those quantities in their own update rules, i.e. let's take

$$\begin{aligned} \delta\theta^k_{i\to a} &= -A_{ai} z^k_i, \\ \delta m^{k+1}_{i\to a} &= \eta'(\theta^k_i,\lambda\zeta^k)\cdot\delta\theta^k_{i\to a}, \\ \delta z^k_{a\to i} &= A_{ai} m^k_i. \end{aligned}$$

We get

$$\begin{aligned} m_i^{k+1} &= \eta(\theta_i^k; \lambda\zeta^k), \\ \theta_i^k &= \sum_b A_{bi} z_b^k + \big(\sum_b A_{bi}^2\big) m_i^k \\ &\approx m_i^k + \sum_b A_{bi} z_b^k, \\ z_a^k &= y_a - \sum_j A_{aj} m_j^k + \sum_j A_{aj}^2 \partial\eta(\theta_j^k; \lambda\zeta^k) z_a^{k-1} \\ &\approx y_a - \sum_j A_{aj} m_j^k + \frac{1}{n} \sum_j \partial\eta(\theta_j^k; \lambda\zeta^k) z_a^{k-1}, \end{aligned}$$

where we also substituted $A_{ai}^2 \to 1/n$ for every a, i.

Our final system of update rules is

$$\begin{aligned} \boldsymbol{m}^{k+1} &= \eta(\boldsymbol{m}^k + \boldsymbol{A}^\mathsf{T} \boldsymbol{z}^k; \lambda\zeta^k) && \in \mathbb{R}^d, \\ \boldsymbol{z}^k &= \boldsymbol{y} - \boldsymbol{A}\boldsymbol{m}^k + \underbrace{\big[\frac{1}{n} \sum_j \partial\eta(\theta_j^k; \lambda\zeta^k)\big] \boldsymbol{z}^{k-1}}_{\text{Onsager term}} && \in \mathbb{R}^n, \\ \zeta^{k+1} &= \zeta^k \times \frac{1}{n} \sum_{i \in [n]} \eta'(\boldsymbol{m}^k + \boldsymbol{A}\boldsymbol{z}^k; \lambda\zeta^k)_i + 1 && \in \mathbb{R}. \end{aligned}$$

It is interesting to compare it to our crude system (Equations 20, 21, 22). The only correction appears in the update rule for $\boldsymbol{z}^k$. This correction term is called "Onsager term".