Department of Statistics and Data Science at SUSTech

MAT7035: Computational Statistics

Tutorial 6: Optimization (III): General MM Algorithms

F. The MM Algorithms

F.1 Definition

(b) Assume that $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ minorizes $\ell(\boldsymbol{\theta}|Y_{\text{obs}})$ at $\boldsymbol{\theta}^{(t)}$, i.e.,

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &\leqslant \ell(\boldsymbol{\theta}|Y_{\text{obs}}), \quad \forall \boldsymbol{\theta},\ \boldsymbol{\theta}^{(t)} \in \boldsymbol{\Theta} \quad \textbf{and} \\ Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) &= \ell(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}}). \end{aligned}$$

(b) If we could find such a real value function $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ depending on $\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)} \in \boldsymbol{\Theta}$, where $\boldsymbol{\theta}^{(t)}$ denotes the t-th approximation of the MLE $\hat{\boldsymbol{\theta}}$,

(c) then by maximizing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ instead of the target log-likelihood function $\ell(\boldsymbol{\theta}|Y_{\text{obs}})$, we obtain the maximizer of $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ as

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}). \tag{6.1}$$

F.2 The ascent property of the MM algorithm

(a) Let $\boldsymbol{\theta}^{(t+1)}$ be defined in (6.1), then we have

$$\ell(\boldsymbol{\theta}^{(t+1)}|Y_{\text{obs}}) \geqslant Q(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) \geqslant Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) = \ell(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}}).$$

(b) An increase in $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ forces an increase in $\ell(\boldsymbol{\theta}|Y_{\text{obs}})$.

(c) This ascent property guarantees a monotone convergence of an MM algorithm.

G. The Quadratic Lower-Bound (QLB) Algorithm

G.1 Definition

(a) The QLB algorithm is a special case of MM algorithms and can be used to find the MLE $\hat{\boldsymbol{\theta}}$.

(b) The key for the QLB algorithm is to find a positive definite matrix $\boldsymbol{B} > 0$ not depending on $\boldsymbol{\theta}$ such that

$$\nabla^2 \ell(\boldsymbol{\theta}|Y_{\text{obs}}) + \boldsymbol{B} \geqslant 0 \quad \forall \boldsymbol{\theta} \in \boldsymbol{\Theta}.$$

(c) The minorizing function is defined by

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \ell(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}}) + (\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)})^\top \nabla \ell(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}}) - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)})^\top \boldsymbol{B}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)}).$$

(d) Let

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}).$$

(e) To maximize $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$, we let

$$\begin{aligned}
\nabla Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= \nabla\Big[\ell(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}}) + (\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)})^\top \nabla \ell(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}}) \\
&\quad -\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)})^\top \boldsymbol{B}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)})\Big] \\
&= \nabla\big[\ell(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}})\big] + \nabla\big[(\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)})^\top \nabla \ell(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}})\big] \\
&\quad -\frac{1}{2}\nabla\big[(\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)})^\top \boldsymbol{B}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)})\big] \\
&= \mathbf{0} + \nabla \ell(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}}) - \frac{1}{2}\big[2\boldsymbol{B}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)})\big] \\
&= \nabla \ell(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}}) - \boldsymbol{B}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)}) = \mathbf{0},
\end{aligned}$$

and obtain

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \boldsymbol{B}^{-1}\nabla \ell(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}}).$$

G.2 Proving that $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ minorizes $\ell(\boldsymbol{\theta}|Y_{\rm obs})$ at $\boldsymbol{\theta}^{(t)}$

We only need to prove that

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &\leqslant \ell(\boldsymbol{\theta}|Y_{\rm obs}), \quad \forall \boldsymbol{\theta},\ \boldsymbol{\theta}^{(t)} \in \boldsymbol{\Theta} \quad \textbf{and} \\ Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) &= \ell(\boldsymbol{\theta}^{(t)}|Y_{\rm obs}). \end{aligned}$$

Proof: By the second-order Taylor's expansion of $\ell(\boldsymbol{\theta}|Y_{\rm obs})$ in a neighborhood of $\boldsymbol{\theta}^{(t)}$, we have

$$\begin{aligned} \ell(\boldsymbol{\theta}|Y_{\rm obs}) &= \ell(\boldsymbol{\theta}^{(t)}|Y_{\rm obs}) + (\boldsymbol{\theta}-\boldsymbol{\theta}^{(t)})^{\top}\nabla\ell(\boldsymbol{\theta}^{(t)}|Y_{\rm obs}) \\ &\quad + \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}^{(t)})^{\top}\nabla^2\ell(\boldsymbol{\theta}^{*}|Y_{\rm obs})(\boldsymbol{\theta}-\boldsymbol{\theta}^{(t)}) \\ &\geqslant \ell(\boldsymbol{\theta}^{(t)}|Y_{\rm obs}) + (\boldsymbol{\theta}-\boldsymbol{\theta}^{(t)})^{\top}\nabla\ell(\boldsymbol{\theta}^{(t)}|Y_{\rm obs}) + \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}^{(t)})^{\top}(-\boldsymbol{B})(\boldsymbol{\theta}-\boldsymbol{\theta}^{(t)}) \\ &= Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}), \end{aligned}$$

for all $\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)} \in \boldsymbol{\Theta}$ and some point $\boldsymbol{\theta}^{*}$ between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^{(t)}$. Let $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$, we obtain $Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) = \ell(\boldsymbol{\theta}^{(t)}|Y_{\rm obs})$. Therefore, $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ minorizes $\ell(\boldsymbol{\theta}|Y_{\rm obs})$ at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$. ‖

H. EM Algorithm is a Special Case of MM Algorithms

For any EM algorithm, let

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \int_{\mathbb{Z}} \ell(\boldsymbol{\theta}|Y_{\rm obs}, \boldsymbol{z}) \times f(\boldsymbol{z}|Y_{\rm obs}, \boldsymbol{\theta}^{(t)})\, {\rm d}\boldsymbol{z}.$$

Define

$$Q^{*}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) + \ell(\boldsymbol{\theta}^{(t)}|Y_{\rm obs}) - Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)})$$

as the surrogate function of an MM algorithm. We can prove

(a) $Q^{*}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ minorizes $\ell(\boldsymbol{\theta}|Y_{\rm obs})$ at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$.

(b) Maxmizing $Q^*(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ with respect to $\boldsymbol{\theta}$ is equivalent to maxmizing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$.

Proof: (a) When proving the ascent property of an EM algorithm, we obtain the result for all $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^{(t)}$,

$$\ell(\boldsymbol{\theta}^{(t)}|Y_{\rm obs}) - Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) \leqslant \ell(\boldsymbol{\theta}|Y_{\rm obs}) - Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$$

and $\ell(\boldsymbol{\theta}|Y_{\rm obs}) - Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ attains its minimum at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)} \in \boldsymbol{\Theta}$. Then

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) + \ell(\boldsymbol{\theta}^{(t)}|Y_{\rm obs}) - Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) \leqslant \ell(\boldsymbol{\theta}|Y_{\rm obs}),$$

i.e., $Q^*(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \leqslant \ell(\boldsymbol{\theta}|Y_{\rm obs})$ for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and they are equal when $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$.

(b) Note that $\ell(\boldsymbol{\theta}^{(t)}|Y_{\rm obs})$ and $Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)})$ are independent of $\boldsymbol{\theta}$. ‖

Example T6.1 (Logistic regression). Let $Y_{\rm obs} = \{y_i\}_{i=1}^m$ and consider the following logistic regression

$$\begin{aligned} y_i &\stackrel{\rm ind}{\sim} \text{Binomial}(n_i, p_i), \\ \text{logit}(p_i) &= \log\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{x}_{(i)}^{\top}\boldsymbol{\theta}, \quad 1 \leqslant i \leqslant m, \end{aligned}$$

where y_i denotes the number of subjects with positive response in the i-th group with n_i trials, p_i the probability of a subject in the i-th group with positive response, $\boldsymbol{x}_{(i)}$ covariates vector, and $\boldsymbol{\theta}_{q\times 1}$ unknown parameters. Use the QLB algorithm to find the MLE of $\boldsymbol{\theta}$.

Hint: Define a positive definite matrix $\boldsymbol{B} > 0$ and set $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \boldsymbol{B}^{-1}\nabla\ell(\boldsymbol{\theta}^{(t)}|Y_{\rm obs})$.

Solution: The log-likelihood function of $\boldsymbol{\theta}$ is

$$\ell(\boldsymbol{\theta}|Y_{\rm obs}) = \sum_{i=1}^m \log\binom{n_i}{y_i} + \sum_{i=1}^m [y_i \log(p_i)] + \sum_{i=1}^m [(n_i - y_i)\log(1-p_i)],$$

where

$$p_i = \frac{\exp[\boldsymbol{x}_{(i)}^\top \boldsymbol{\theta}]}{1 + \exp[\boldsymbol{x}_{(i)}^\top \boldsymbol{\theta}]}.$$

Then the score vector is

$$\nabla \ell(\boldsymbol{\theta}|Y_{\text{obs}}) = \sum_{i=1}^{m} (y_i - n_i p_i)\boldsymbol{x}_{(i)}$$

and the observed information matrix $\boldsymbol{I}(\boldsymbol{\theta}|Y_{\text{obs}})$ is

$$-\nabla^2 \ell(\boldsymbol{\theta}|Y_{\text{obs}}) = \sum_{i=1}^{m} n_i p_i (1 - p_i)\boldsymbol{x}_{(i)}\boldsymbol{x}_{(i)}^\top.$$

Let

$$\begin{aligned}
\boldsymbol{X} &= (\boldsymbol{x}_{(1)}, \ldots, \boldsymbol{x}_{(m)})^\top, \\
\boldsymbol{y} &= (y_1, \ldots, y_m)^\top, \\
\boldsymbol{N} &= \text{diag}(n_1, \ldots, n_m), \\
\boldsymbol{p} &= (p_1, \ldots, p_m)^\top, \quad p_i = \frac{\exp[\boldsymbol{x}_{(i)}^\top \boldsymbol{\theta}]}{1 + \exp[\boldsymbol{x}_{(i)}^\top \boldsymbol{\theta}]} \\
\boldsymbol{P} &= \text{diag}(p_1(1-p_1), \ldots, p_m(1-p_m)).
\end{aligned}$$

Then

$$\nabla \ell(\boldsymbol{\theta}|Y_{\text{obs}}) = \boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{N}\boldsymbol{p}) \quad \text{and} \quad -\nabla^2 \ell(\boldsymbol{\theta}|Y_{\text{obs}}) = \boldsymbol{X}^\top \boldsymbol{N}\boldsymbol{P}\boldsymbol{X}.$$

Note that $p_i(1-p_i) = -(p_i - \frac{1}{2})^2 + \frac{1}{4} \leqslant \frac{1}{4}$, then

$$-\nabla^2 \ell(\boldsymbol{\theta}|Y_{\text{obs}}) = \sum_{i=1}^{m} n_i p_i (1-p_i)\boldsymbol{x}_{(i)}\boldsymbol{x}_{(i)}^\top \leqslant \frac{1}{4}\sum_{i=1}^{m} n_i \boldsymbol{x}_{(i)}\boldsymbol{x}_{(i)}^\top.$$

Define $\boldsymbol{B} = \frac{1}{4}\sum_{i=1}^{m} n_i \boldsymbol{x}_{(i)}\boldsymbol{x}_{(i)}^\top = \frac{1}{4}\boldsymbol{X}^\top \boldsymbol{N}\boldsymbol{X}$ so that $\nabla^2 \ell(\boldsymbol{\theta}|Y_{\text{obs}}) + \boldsymbol{B} \geqslant 0, \forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$. Therefore, we obtain

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + 4(\boldsymbol{X}^\top \boldsymbol{N}\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{N}\boldsymbol{p}^{(t)}),$$

where

$$p_i^{(t)} = \frac{\exp[\boldsymbol{x}_{(i)}^\top \boldsymbol{\theta}^{(t)}]}{1 + \exp[\boldsymbol{x}_{(i)}^\top \boldsymbol{\theta}^{(t)}]}$$

is the i-th component of $\boldsymbol{p}^{(t)}$ and $1 \leqslant i \leqslant m$. $\|$