Department of Statistics and Data Science at SUSTech

MAT7035: Computational Statistics

Tutorial 7: Optimization (IV): Two-class Special MM Algorithms

I. Construction of minorizing functions via Jensen's inequality

I.1 Discrete version of Jensen's inequality

(a) Jensen's inequality:

Let $\varphi(\cdot)$ be a concave function. If X is an r.v. taking values in the domain of $\varphi(\cdot)$, then

$$\varphi[E(X)] \geqslant E[\varphi(X)], \tag{7.1}$$

provided that both expectations $E(X)$ and $E[\varphi(X)]$ exist.

(b) Discrete version:

For any concave function $f(\cdot)$ and a discrete random variable X with pmf $\Pr(X = x_i) = \alpha_i$, Jensen's inequality (7.1) states that

$$f\left(\sum_{i=1}^{n} \alpha_i x_i\right) \geqslant \sum_{i=1}^{n} \alpha_i f(x_i), \tag{7.2}$$

where $\alpha_i > 0, i = 1, \ldots, n$, and $\sum_{i=1}^{n} \alpha_i = 1$.

(c) Two important facts:

- $\sum_{i=1}^{n} \alpha_i f(x_i)$ is a convex combination of the same separable function with different arguments, resulting in a diagonal Hessian matrix of the form

$$\frac{\partial^2[\sum_{i=1}^{n} \alpha_i f(x_i)]}{\partial \boldsymbol{x} \partial \boldsymbol{x}^{\top}} = \mathrm{diag}\Big\{\alpha_1 f''(x_1), \ldots, \alpha_n f''(x_n)\Big\}.$$

- On both sides of the inequality (7.2), the function family does not change.

I.2 Linear combination of parameters: $\boldsymbol{a}^\top\boldsymbol{\theta}$

(a) Suppose that we can decompose

$$\underbrace{\ell(\boldsymbol{\theta}|Y_{\rm obs})}_{\text{concave}} = \underbrace{\ell_1(\boldsymbol{\theta})}_{\text{effective concave}} + \underbrace{\ell_2(\boldsymbol{a}^\top\boldsymbol{\theta})}_{\text{concave of one dimension}}, \tag{7.3}$$

- $\ell_1(\boldsymbol{\theta})$ is an "effective" concave function in the sense that there exist explicit solutions to the partial score equations $\nabla\ell_1(\boldsymbol{\theta}) = \frac{\partial\ell_1(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = \mathbf{0}_q$;
- $\ell_2(\cdot)$ depends on $\boldsymbol{\theta}$ only through $\boldsymbol{a}^\top\boldsymbol{\theta}$;
- $\boldsymbol{a} = (a_1, \ldots, a_q)^\top$ is a constant vector.

(b) We construct a minorizing function of the form

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \ell_1(\boldsymbol{\theta}) + Q_2(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}),$$

where $Q_2(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ minorizes $\ell_2(\boldsymbol{a}^\top\boldsymbol{\theta})$; i.e., Q_2 satisfies

$$Q_2(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \leqslant \ell_2(\boldsymbol{a}^\top\boldsymbol{\theta}), \quad \forall\, \boldsymbol{\theta}, \boldsymbol{\theta}^{(t)} \in \boldsymbol{\Theta}, \quad \textbf{and} \tag{7.4}$$

$$Q_2(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) = \ell_2(\boldsymbol{a}^\top\boldsymbol{\theta}^{(t)}). \tag{7.5}$$

It's easy to prove that $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ minorizes $\ell(\boldsymbol{\theta}|Y_{\rm obs})$ at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$.

Proof: We have

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \ell_1(\boldsymbol{\theta}) + Q_2(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \overset{(7.4)}{\leqslant} \ell_1(\boldsymbol{\theta}) + \ell_2(\boldsymbol{a}^\top\boldsymbol{\theta}) \overset{(7.3)}{=} \ell(\boldsymbol{\theta}|Y_{\rm obs}), \quad \forall\, \boldsymbol{\theta} \in \boldsymbol{\Theta},$$

$$Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) = \ell_1(\boldsymbol{\theta}^{(t)}) + Q_2(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) \overset{(7.5)}{=} \ell_1(\boldsymbol{\theta}^{(t)}) + \ell_2(\boldsymbol{a}^\top\boldsymbol{\theta}^{(t)}) = \ell(\boldsymbol{\theta}^{(t)}|Y_{\rm obs}).$$

□

(c) Construction of a separable minorizing function:

- Given the constant vector $\boldsymbol{a} = (a_1, \ldots, a_q)^\top$, we can appropriately select a vector $\boldsymbol{w} = (w_1, \ldots, w_q)^\top$ such that $a_j w_j \geqslant 0$ for all $j = 1, \ldots, q$.
- We have

$$\begin{aligned}
\ell_2(\boldsymbol{a}^\top \boldsymbol{\theta}) &= \ell_2\left(\sum_{j=1}^q a_j \theta_j\right) = \ell_2\left(\sum_{j=1}^q \frac{a_j w_j}{\boldsymbol{a}^\top \boldsymbol{w}} \cdot \frac{\boldsymbol{a}^\top \boldsymbol{w}}{w_j}\theta_j\right) \\
&= \ell_2\left(\sum_{j=1}^q \frac{a_j w_j}{\boldsymbol{a}^\top \boldsymbol{w}} x_j\right) \qquad \left[\text{where } \ x_j \hat{=} \frac{\boldsymbol{a}^\top \boldsymbol{w}}{w_j}\theta_j\right] \\
&\overset{(7.2)}{\geqslant} \sum_{j=1}^q \frac{a_j w_j}{\boldsymbol{a}^\top \boldsymbol{w}}\, \ell_2\left(\frac{\boldsymbol{a}^\top \boldsymbol{w}}{w_j}\theta_j\right).
\end{aligned}$$

- If we could choose $\boldsymbol{w} = \boldsymbol{\theta}^{(t)}$, provided that $a_j \theta_j^{(t)} \geqslant 0$ for all $j = 1, \ldots, q$, then we can set

$$Q_2(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \sum_{j=1}^q \frac{a_j \theta_j^{(t)}}{\boldsymbol{a}^\top \boldsymbol{\theta}^{(t)}}\, \ell_2\left(\frac{\boldsymbol{a}^\top \boldsymbol{\theta}^{(t)}}{\theta_j^{(t)}}\theta_j\right). \tag{7.6}$$

It is easy to verify that this $Q_2(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ satisfies (7.4) and (7.5).

Example T7.1 (Genetic linkage model / Example 2.6 in Lecture Notes). The offspring of an $AB/ab \times AB/ab$ mating fall into the four categories AB, Ab, aB and ab with cell probabilities

$$\frac{\theta+2}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}, \quad 0 \leqslant \theta \leqslant 1,$$

where $\theta = (1-r)^2$. Observed frequencies $Y_{\text{obs}} = (y_1, y_2, y_3, y_4)^\top$ follow a multinomial distribution with above cell probabilities, i.e.,

$$Y_{\text{obs}} \sim \text{Multinomial}\left(n; \frac{\theta+2}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}\right).$$

Design an MM algorithm to find the maximum likelihood estimator of θ.

Hint: (a) Log-beta function family: A function $g(\theta)$ is said to be a member of the family of log-beta functions, denoted by $g(\theta) \in \mathrm{LB}(\theta)$, if

$$g(\theta) = c + a\log(\theta) + b\log(1-\theta), \quad \theta \in [0,1],$$

where $c \in \mathbb{R}$ is a constant not depending on θ and $a, b \geqslant 0$. We call $\log(\theta)$ and $\log(1-\theta)$ two complemental *base functions* (or *assemblies*) of the log-beta function family.

(b) Effective concavity: $g(\theta)$ is effectively concave iff $a > 0$ and $b > 0$. It has mode

$$\hat{\theta} = \arg\max_{\theta \in [0,1]} g(\theta) = \frac{a}{a+b}. \tag{7.7}$$

Solution: The observed-data likelihood function of θ is given by

$$\begin{aligned} L(\theta|Y_{\mathrm{obs}}) &= \binom{N}{y_1, \ldots, y_4} \left(\frac{\theta+2}{4}\right)^{y_1} \left(\frac{1-\theta}{4}\right)^{y_2} \left(\frac{1-\theta}{4}\right)^{y_3} \left(\frac{\theta}{4}\right)^{y_4} \\ &\propto \left(\frac{\theta+2}{4}\right)^{y_1} \left(\frac{1-\theta}{4}\right)^{y_2+y_3} \left(\frac{\theta}{4}\right)^{y_4}. \end{aligned}$$

Therefore, we can divide the log-likelihood function into two parts:

$$\ell(\theta|Y_{\mathrm{obs}}) = \underbrace{c_1 + y_4\log(\theta) + (y_2+y_3)\log(1-\theta)}_{\ell_1(\theta)} + \underbrace{y_1\log(\theta+2)}_{\ell_2(\cdot)},$$

where c_1 is a constant independent of the parameter θ.

(a) The first MM algorithm: Obviously, we have $\ell_1(\theta) \in \mathrm{LB}(\theta)$, and both y_4 and $y_2 + y_3$ are positive. Therefore, $\ell_1(\theta)$ is effectively concave.

Let $\ell_2(u) = y_1\log(u)$, where $u > 0$ and $0 < y_1 \leqslant N$. It is clear that $\ell_2(u)$ is a strictly concave function defined in $\mathbb{R}_+$. Hence, from (7.6), we can define

$$\begin{aligned} Q_2(\theta|\theta^{(t)}) &= y_1\left[\frac{\theta^{(t)}}{\theta^{(t)}+2}\log\left(\frac{\theta^{(t)}+2}{\theta^{(t)}}\cdot\theta\right) + \frac{2}{\theta^{(t)}+2}\log\left(\frac{\theta^{(t)}+2}{1}\cdot 1\right)\right] \\ &= c_2 + \frac{y_1\theta^{(t)}}{\theta^{(t)}+2}\log(\theta) \in \mathrm{LB}(\theta), \end{aligned}$$

where c_2 is another constant free of the parameter θ. Then we can construct the Q function as follows:

$$\begin{aligned} Q(\theta|\theta^{(t)}) &= \ell_1(\theta) + Q_2(\theta|\theta^{(t)}) \\ &= c_1 + c_2 + \left(y_4 + \frac{y_1\theta^{(t)}}{\theta^{(t)}+2} \right) \log(\theta) + (y_2+y_3)\log(1-\theta) \in \mathrm{LB}(\theta). \end{aligned}$$

Therefore, from (7.7), we obtain the first MM iteration:

$$\theta^{(t+1)} = \frac{y_4 + y_1\theta^{(t)}/(\theta^{(t)}+2)}{y_4 + y_1\theta^{(t)}/(\theta^{(t)}+2) + y_2 + y_3}.$$

(b) <u>The second MM algorithm</u>: Note that we can write

$$\theta + 2 = 3\theta + 2(1-\theta) = (3,2)\begin{pmatrix} \theta \\ 1-\theta \end{pmatrix}.$$

Hence, let $\boldsymbol{a} = (a_1, a_2)^\top = (3,2)^\top$, $\boldsymbol{\theta} = (\theta_1, \theta_2)^\top = (\theta, 1-\theta)^\top \in \mathbb{T}_2$, and $\boldsymbol{\theta}^{(t)} = (\theta_1^{(t)}, \theta_2^{(t)})^\top = (\theta^{(t)}, 1-\theta^{(t)})^\top$, we have $a_1\theta_1^{(t)} = 3\theta^{(t)} \geqslant 0$ and $a_2\theta_2^{(t)} = 2(1-\theta^{(t)}) \geqslant 0$. The corresponding

$$\begin{aligned} Q_2^*(\theta|\theta^{(t)}) &= \frac{3y_1\theta^{(t)}}{\theta^{(t)}+2} \log\left(\frac{\theta^{(t)}+2}{\theta^{(t)}} \cdot \theta \right) + \frac{2y_1(1-\theta^{(t)})}{\theta^{(t)}+2} \log\left[\frac{\theta^{(t)}+2}{1-\theta^{(t)}}(1-\theta) \right] \\ &= c_2^* + \frac{3y_1\theta^{(t)}}{\theta^{(t)}+2}\log(\theta) + \frac{2y_1(1-\theta^{(t)})}{\theta^{(t)}+2}\log(1-\theta) \in \mathrm{LB}(\theta) \end{aligned}$$

and

$$\begin{aligned} Q^*(\theta|\theta^{(t)}) =& c_1 + c_2^* + \left(y_4 + \frac{3y_1\theta^{(t)}}{\theta^{(t)}+2} \right) \log(\theta) \\ &+ \left[y_2 + y_3 + \frac{2y_1(1-\theta^{(t)})}{\theta^{(t)}+2} \right] \log(1-\theta) \in \mathrm{LB}(\theta). \end{aligned}$$

The resultant second MM algorithm is

$$\theta^{(t+1)} = \frac{y_4 + 3y_1\theta^{(t)}/(\theta^{(t)}+2)}{y_1+y_2+y_3+y_4}. \qquad \|$$

I.3 Linear combinations of parameters: $\{\boldsymbol{a}_i^{\top}\boldsymbol{\theta}\}_{i=1}^n$

(a) Suppose that we can decompose

$$\underbrace{\ell(\boldsymbol{\theta}|Y_{\rm obs})}_{\text{concave}} = \underbrace{\ell_0(\boldsymbol{\theta})}_{\text{concave}} + \underbrace{\ell_1(\boldsymbol{\theta})}_{\text{effective concave}} + \sum_{i=1}^n \underbrace{\ell_{2i}(\boldsymbol{a}_i^{\top}\boldsymbol{\theta})}_{\text{concave of one dimension}}, \tag{7.8}$$

— $\ell_0(\boldsymbol{\theta})$ is concave but not necessarily effectively concave;

— $\ell_1(\boldsymbol{\theta})$ is effectively concave;

— $\ell_{2i}(\cdot)$ is a one-dimensional concave function in $\mathbb{R}$ or in a subset of $\mathbb{R}$;

— ℓ_{2i} depends on $\boldsymbol{\theta}$ only through the linear combination $\boldsymbol{a}_i^{\top}\boldsymbol{\theta}$;

— $\boldsymbol{a}_i = (a_{i1}, \dots, a_{iq})^{\top}$ is a constant vector.

(b) Provided that for any $i \in \{1, \dots, n\}$, $a_{ij}\theta_j^{(t)} \geqslant 0$ for all $j = 1, \dots, q$, a minorizing function can be constructed as

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \ell_0(\boldsymbol{\theta}) + \ell_1(\boldsymbol{\theta}) + \sum_{i=1}^n Q_{2i}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}), \tag{7.9}$$

where

$$Q_{2i}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \sum_{j=1}^q \frac{a_{ij}\theta_j^{(t)}}{\boldsymbol{a}_i^{\top}\boldsymbol{\theta}^{(t)}}\, \ell_{2i}\left(\frac{\boldsymbol{a}_i^{\top}\boldsymbol{\theta}^{(t)}}{\theta_j^{(t)}}\theta_j\right). \tag{7.10}$$

Example T7.2 (Poisson additive model). Let $Y_i \stackrel{\text{ind}}{\sim} \text{Poisson}(\boldsymbol{a}_{\rm i}^{\top}\boldsymbol{\theta})$, $1 \leqslant i \leqslant n$, where $\boldsymbol{a}_i = (a_{i1}, \dots, a_{ip})^{\top}$ is a known vector and each element is nonnegative. Design an MM algorithm to find the maximum likelihood estimators of $\boldsymbol{\theta}$.

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

$$L(\boldsymbol{\theta}|Y_{\rm obs}) = \prod_{i=1}^n \frac{(\boldsymbol{a}_i^{\top}\boldsymbol{\theta})^{y_i}\exp(-\boldsymbol{a}_i^{\top}\boldsymbol{\theta})}{y_i!}.$$

We rewrite the log-likelihood function of $\boldsymbol{\theta} \in \mathbb{R}_+^q$ as two (or three) parts:

$$\ell(\boldsymbol{\theta}|Y_{\rm obs}) \;=\; c_1 + \sum_{i=1}^n [y_i \log(\boldsymbol{a}_i^{\top}\boldsymbol{\theta}) - \boldsymbol{a}_i^{\top}\boldsymbol{\theta}]$$

$$= \underbrace{c_1 - \textstyle\sum_{i=1}^n \boldsymbol{a}_i^\top \boldsymbol{\theta}}_{\ell_0(\boldsymbol{\theta})} + \underbrace{0}_{\ell_1(\boldsymbol{\theta})} + \sum_{i \notin \mathbb{I}_0} \underbrace{y_i \log(\boldsymbol{a}_i^\top \boldsymbol{\theta})}_{\ell_{2i}(\cdot)},$$

where $\mathbb{I}_0 \hat{=} \{i\colon\ y_i = 0,\ 1 \leqslant i \leqslant n\}$ and $\ell_{2i}(u) = y_i \log(u)$ for $y_i \geqslant 1$. From (7.9) and (7.10), we have

$$\begin{aligned}
Q_{2i}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= y_i \sum_{j=1}^q \frac{a_{ij}\theta_j^{(t)}}{\boldsymbol{a}_i^\top \boldsymbol{\theta}^{(t)}} \log\left(\frac{\boldsymbol{a}_i^\top \boldsymbol{\theta}^{(t)}}{\theta_j^{(t)}}\theta_j\right) \\
&= c_{2i} + y_i \sum_{j=1}^q \frac{a_{ij}\theta_j^{(t)}}{\boldsymbol{a}_i^\top \boldsymbol{\theta}^{(t)}} \log(\theta_j) \quad \text{and} \\
Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= \ell_0(\boldsymbol{\theta}) + \sum_{i \notin \mathbb{I}_0} Q_{2i}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \\
&= c_1 + c_2 - \sum_{i=1}^n \boldsymbol{a}_i^\top \boldsymbol{\theta} + \sum_{i=1}^n y_i \sum_{j=1}^q \frac{a_{ij}\theta_j^{(t)}}{\boldsymbol{a}_i^\top \boldsymbol{\theta}^{(t)}} \log(\theta_j) \\
&= c + \sum_{i=1}^n \left[\sum_{j=1}^q \frac{y_i \cdot a_{ij}\theta_j^{(t)}}{\boldsymbol{a}_i^\top \boldsymbol{\theta}^{(t)}} \log(\theta_j) + \sum_{j=1}^q a_{ij}(-\theta_j)\right] \\
&= c + \sum_{j=1}^q \left[\left(\sum_{i=1}^n \frac{y_i \cdot a_{ij}\theta_j^{(t)}}{\boldsymbol{a}_i^\top \boldsymbol{\theta}^{(t)}}\right) \log(\theta_j) + \left(\sum_{i=1}^n a_{ij}\right)(-\theta_j)\right] \\
&\hat{=} c + \sum_{j=1}^q \Big[a_j \log(\theta_j) + b_j(-\theta_j)\Big],
\end{aligned}$$

where all parameters $\{\theta_j\}_{j=1}^q$ are separated. So we obtain the following MM iterations:

$$\theta_j^{(t+1)} = \frac{a_j}{b_j} = \theta_j^{(t)} \frac{\sum_{i=1}^n [y_i a_{ij}/\boldsymbol{a}_i^\top \boldsymbol{\theta}^{(t)}]}{\sum_{i=1}^n a_{ij}}, \quad 1 \leqslant j \leqslant q. \qquad \|$$

J. De Pierro's (DP) algorithm

J.1 A special member of the family of MM algorithms

(a) In examples such as Poisson regression, QLB algorithm fails because of the absence of a positive definite B satisfying conditions.

(b) **The main idea of the DP algorithm**: Transferring the optimization of a high-dimensional function $\ell(\boldsymbol{\theta}|Y_{\rm obs})$ to the optimization of a low dimensional surrogate function $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$.

(c) $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ is a sum of convex combinations of a series of one-dimensional concave functions. Maximizing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ can be implemented via the one-step Newton–Raphson method.

J.2 Summary of the DP Algorithm

(a) Let the log-likelihood function be of the form

$$\ell(\boldsymbol{\theta}|Y_{\rm obs}) = \sum_{i=1}^{m} f_i(\boldsymbol{x}_{(i)}^\top \boldsymbol{\theta}),$$

— $\boldsymbol{X}_{m\times q} = (\boldsymbol{x}_{(1)}, \ldots, \boldsymbol{x}_{(m)})^\top$ is the covariates matrix;

— $\boldsymbol{x}_{(i)} = (x_{i1}, \ldots, x_{iq})^\top$ denotes the i-th row vector of $\boldsymbol{X}$;

— $\boldsymbol{\theta}$ is the parameter of interest;

— $\{f_i\}_{i=1}^m$ are twice continuously differentiable and strictly concave functions defined in one-dimensional real space $\mathbb{R}$.

(b) The score and the observed information are given by

$$\begin{aligned} \nabla\ell(\boldsymbol{\theta}|Y_{\rm obs}) &= \sum_{i=1}^{m} f_i'(\boldsymbol{x}_{(i)}^\top\boldsymbol{\theta})\boldsymbol{x}_{(i)}, \quad \text{and} \\ -\nabla^2\ell(\boldsymbol{\theta}|Y_{\rm obs}) &= \sum_{i=1}^{m} \left\{-f_i''(\boldsymbol{x}_{(i)}^\top\boldsymbol{\theta})\right\}\boldsymbol{x}_{(i)}\boldsymbol{x}_{(i)}^\top. \end{aligned}$$

Thus, $\ell(\boldsymbol{\theta}|Y_{\rm obs})$ is strictly concave provided that at least one $f_i''(\cdot) < 0$.

(c) We first define two index sets:

$$\begin{aligned} \mathbb{J}_i &= \{j\colon\, x_{ij} \neq 0\}, \quad 1 \leqslant i \leqslant m, \\ \mathbb{I}_j &= \{i\colon\, x_{ij} \neq 0\}, \quad 1 \leqslant j \leqslant q, \end{aligned}$$

and probability weights: for a fixed i,

$$\lambda_{ij} = \frac{|x_{ij}|}{\sum_{j' \in \mathbb{J}_i} |x_{ij'}|} > 0, \quad j \in \mathbb{J}_i, \quad \text{and} \quad \sum_{j \in \mathbb{J}_i} \lambda_{ij} = 1.$$

(d) Then we can construct a surrogate function for a given $\boldsymbol{\theta}^{(t)} \in \boldsymbol{\Theta}$ as follows:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \sum_{i=1}^{m} \sum_{j \in \mathbb{J}_i} \lambda_{ij} f_i \left(\lambda_{ij}^{-1} x_{ij} (\theta_j - \theta_j^{(t)}) + \boldsymbol{x}_{(i)}^{\top} \boldsymbol{\theta}^{(t)} \right), \quad \boldsymbol{\theta} \in \boldsymbol{\Theta}.$$

(e) The DP algorithm is defined by

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

(f) $\theta^{(t+1)}$ can be obtained as the solution to the system of equations

$$\sum_{i \in \mathbb{I}_j} f_i' \left(\lambda_{ij}^{-1} x_{ij} (\theta_j - \theta_j^{(t)}) + \boldsymbol{x}_{(i)}^{\top} \boldsymbol{\theta}^{(t)} \right) x_{ij} = 0, \qquad 1 \leqslant j \leqslant q.$$

(g) If it cannot be solved explicitly, the one step NR algorithm yields

$$\theta_j^{(t+1)} = \theta_j^{(t)} + \tau_j^2(\boldsymbol{\theta}^{(t)}) \sum_{i \in \mathbb{I}_j} f_i'(\boldsymbol{x}_{(i)}^{\top} \boldsymbol{\theta}^{(t)}) x_{ij}, \tag{7.11}$$

where

$$\tau_j^2(\boldsymbol{\theta}^{(t)}) = \left\{ \sum_{i \in \mathbb{I}_j} \{-f_i''(\boldsymbol{x}_{(i)}^{\top} \boldsymbol{\theta}^{(t)})\} x_{ij}^2 / \lambda_{ij}, \right\}^{-1}, \quad 1 \leqslant j \leqslant q.$$

Example T7.3 (Logistic regression models/Example T6.1 in Tutorial 6). Let $Y_{\text{obs}} = \{y_i\}_{i=1}^m$ and consider the following logistic regression

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

where y_i is 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 DP algorithm to find the MLEs of $\boldsymbol{\theta}$.

Solution: From Example T6.1 in Tutorial 6, we have the log-likelihood function of $\boldsymbol{\theta}$

$$\ell(\boldsymbol{\theta}|Y_{\rm obs}) = c + \sum_{i=1}^{m}\{y_i(\boldsymbol{x}_{(i)}^{\top}\boldsymbol{\theta}) - n_i \log[1+\exp(\boldsymbol{x}_{(i)}^{\top}\boldsymbol{\theta})]\} = c + \sum_{i=1}^{m} f_i(\boldsymbol{x}_{(i)}^{\top}\boldsymbol{\theta}),$$

where

$$f_i(u) = y_i u - n_i \log(1+\mathrm{e}^u).$$

Noting that

$$f_i'(u) = y_i - n_i\frac{\mathrm{e}^u}{1+\mathrm{e}^u} \quad \text{and} \quad -f_i''(u) = n_i\frac{\mathrm{e}^u}{(1+\mathrm{e}^u)^2}.$$

From (7.11), we obtain the following DP iterations:

$$\theta_j^{(t+1)} = \theta_j^{(t)} + \frac{\sum_{i\in\mathbb{I}_j}(y_i - n_i p_i^{(t)})x_{ij}}{\sum_{i\in\mathbb{I}_j} n_i p_i^{(t)}(1-p_i^{(t)})x_{ij}^2/\lambda_{ij}}, \quad 1 \leqslant j \leqslant q,$$

where

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