Department of Statistics and Data Science at SUSTech

MAT7035: Computational Statistics

Tutorial 4: Optimization (I): Newton's Method

A. Optimization

- Optimizing a function means maximizing or minimizing this function.
- A typical optimization problem in statistics is maximizing the log-likelihood function for calculating MLEs of parameters.

B. Newton's Method

B.1 Newton's method for root finding and optimization

(a) Root finding: For a given differentiable function $f(x)$, Newton's method is an iterative root finding technique to solve $f(x) = 0$, defined by

$$x^{(t+1)} = x^{(t)} - \frac{f(x^{(t)})}{f'(x^{(t)})},$$

where $x^{(0)}$ is an initial value.

(b) Optimization: For a twice differentiable function $g(x)$, under some conditions, an optimum $x^{(\infty)}$ satisfies $g'(x^{(\infty)}) = 0$. Then Newton's method for finding the maximizer or the minimizor of $g(x)$ is derived as

$$x^{(t+1)} = x^{(t)} - \frac{g'(x^{(t)})}{g''(x^{(t)})}.$$

B.2 Remarks

(a) Newton's method is highly sensitive to the initial value. Inappropriate initial values may lead to divergence or a local optimum.

(b) Besides, there is no assurance that all $x^{(t)}$ will locate in the support.

Example T4.1 (Maximizor of a function). Let

$$f(x) = \left(\frac{x}{2}\right)^{1/2} + 2\left(\frac{1-x}{3}\right)^{1/2}.$$

(a) Find the accurate x maximizing $f(x)$.

(b) Use Newton's method to calculate the numerical solution x^*. The initial value is set as $x^{(0)} = 0.1$. The stopping rule is: $|x^{(t+1)} - x^{(t)}| < 10^{-6}$.

Solution: (a) On the one hand, let

$$f'(x) = \frac{1}{4}\left(\frac{x}{2}\right)^{-1/2} - \frac{1}{3}\left(\frac{1-x}{3}\right)^{-1/2} = 0,$$

we obtain $x = 3/11$. On the other hand, since

$$\begin{aligned} f''(x) &= -\frac{1}{4}\left(\frac{1}{4}\right)\left(\frac{x}{2}\right)^{-3/2} - \left(\frac{1}{3}\right)\left(\frac{1}{2}\right)\left(\frac{1}{3}\right)\left(\frac{1-x}{3}\right)^{-3/2} \\ &= -\frac{1}{16}\left(\frac{x}{2}\right)^{-3/2} - \frac{1}{18}\left(\frac{1-x}{3}\right)^{-3/2}, \end{aligned}$$

we have $f''(3/11) = -1.7066 < 0$, indicating that $f(x)$ has the strictly local maximum at $x = 3/11 \approx 0.2727273$ with $f(3/11) = 1.3540064$.

(b) Let $x^{(0)} = 0.1$, Newton's method shows that

$$\begin{aligned} x^{(1)} &= x^{(0)} - \frac{f'(x^{(0)})}{f''(x^{(0)})} = 0.1859363, \\ x^{(2)} &= x^{(1)} - \frac{f'(x^{(1)})}{f''(x^{(1)})} = 0.2552335, \\ x^{(3)} &= x^{(2)} - \frac{f'(x^{(2)})}{f''(x^{(2)})} = 0.2721640, \end{aligned}$$

$$\begin{aligned} x^{(4)} &= x^{(3)} - \frac{f'(x^{(3)})}{f''(x^{(3)})} = 0.2727267, \\ x^{(5)} &= x^{(4)} - \frac{f'(x^{(4)})}{f''(x^{(4)})} = 0.2727273. \end{aligned}$$

Note that $|x^{(5)} - x^{(4)}| = 6 \times 10^{-7} < 10^{-6}$, thus the maximum of the $f(x)$ is gotten when $x = x^{(5)} = 0.2727273$ and $f(0.2727273) = 1.3540064$. ‖

Example T4.2 (Exponential distribution). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Exponential}(1/\theta)$ with pdf

$$f(x|\theta) = \frac{1}{\theta} \exp\left(-\frac{x}{\theta}\right), \quad x > 0, \quad \theta > 0.$$

(a) Derive the score vector, the observed information matrix and the expected information matrix.

(b) Using the Newton–Raphson algorithm and the Fisher scoring algorithm to find the MLE $\hat{\theta}$ and the estimated asymptotic covariance matrix of $\hat{\theta}$.

Solution: Let $Y_{\text{obs}} = \{x_i\}_{i=1}^n$. (a) The log-likelihood function of θ is

$$\begin{aligned} \ell(\theta|Y_{\text{obs}}) &= \log\left[\prod_{i=1}^n f(x_i|\theta)\right] = \log\left\{\prod_{i=1}^n \left[\frac{1}{\theta}\exp\left(-\frac{x_i}{\theta}\right)\right]\right\} \\ &= \log\left[\frac{1}{\theta^n}\exp\left(-\frac{\sum_{i=1}^n x_i}{\theta}\right)\right] = -n\log\theta - \frac{\sum_{i=1}^n x_i}{\theta}. \end{aligned}$$

The score vector is

$$\ell'(\theta|Y_{\text{obs}}) = -\frac{n}{\theta} + \frac{\sum_{i=1}^n x_i}{\theta^2}.$$

The observed information matrix is

$$I(\theta|Y_{\text{obs}}) = -\ell''(\theta|Y_{\text{obs}}) = -\frac{n}{\theta^2} + \frac{2\sum_{i=1}^n x_i}{\theta^3}.$$

The expected information matrix is

$$J(\theta) = E_{Y_{\text{obs}}}[I(\theta|Y_{\text{obs}})] = -\frac{n}{\theta^2} + \frac{2\sum_{i=1}^n E(X_i)}{\theta^3} = -\frac{n}{\theta^2} + \frac{2n\theta}{\theta^3} = \frac{n}{\theta^2}.$$

(b) The iteration of the Newton–Raphson algorithm is

$$\begin{aligned}
\theta^{(t+1)} &= \theta^{(t)} + \frac{-\dfrac{n}{\theta^{(t)}} + \dfrac{\sum_{i=1}^n x_i}{[\theta^{(t)}]^2}}{-\dfrac{n}{[\theta^{(t)}]^2} + \dfrac{2\sum_{i=1}^n x_i}{[\theta^{(t)}]^3}}, \\
\Longrightarrow \theta^{(t+1)} &= \theta^{(t)} + \frac{\theta^{(t)}\sum_{i=1}^n x_i - n[\theta^{(t)}]^2}{2\sum_{i=1}^n x_i - n\theta^{(t)}} = \theta^{(t)}\frac{3\sum_{i=1}^n x_i - 2n\theta^{(t)}}{2\sum_{i=1}^n x_i - n\theta^{(t)}}.
\end{aligned}$$

The iteration of the Fisher scoring algorithm is

$$\begin{aligned}
\theta^{(t+1)} &= \theta^{(t)} + \frac{-\dfrac{n}{\theta^{(t)}} + \dfrac{\sum_{i=1}^n x_i}{[\theta^{(t)}]^2}}{\dfrac{n}{[\theta^{(t)}]^2}}, \\
\Longrightarrow \theta^{(t+1)} &= \theta^{(t)} + \frac{\sum_{i=1}^n x_i - n\theta^{(t)}}{n} = \frac{\sum_{i=1}^n x_i}{n}.
\end{aligned}$$

The estimated asymptotic covariance matrix is

$$\widehat{\mathrm{Cov}}(\hat\theta) = J^{-1}(\hat\theta) = \frac{\hat\theta^2}{n}. \qquad \|$$

Remark: The Fisher scoring algorithm amazingly gives a non-iterative result, which coincides the accurate MLE easily derived from $\ell'(\theta|Y_{\rm obs}) = 0$. $\|$

B.3 High-dimensional case

(a) Let $Y_1, \dots, Y_n \overset{\rm iid}{\sim} f(y|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the parameter vector. Let $Y_{\rm obs} = \{y_i\}_{i=1}^n$, then

— the log-likelihood function is $\ell(\boldsymbol{\theta}|Y_{\rm obs}) = \sum_{i=1}^n \log f(y_i|\boldsymbol{\theta})$;

— the score vector is $\nabla\ell(\boldsymbol{\theta}|Y_{\rm obs})$;

— the observed information matrix is $\boldsymbol{I}(\boldsymbol{\theta}|Y_{\rm obs}) = -\nabla^2\ell(\boldsymbol{\theta}|Y_{\rm obs})$;

— the expected information matrix is $\boldsymbol{J}(\boldsymbol{\theta}) = E_{Y_{\rm obs}}[\boldsymbol{I}(\boldsymbol{\theta}|Y_{\rm obs})]$.

(b) The Newton–Raphson algorithm is defined as

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

(c) The Fisher scoring algorithm is defined as

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

(d) The MLE $\hat{\boldsymbol{\theta}}$ has the property:

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \xrightarrow{\text{D}} N(\mathbf{0}, \boldsymbol{J}^{-1}(\boldsymbol{\theta})).$$

(e) The inverse covariance of the asymptotic distribution, $\boldsymbol{J}^{-1}(\boldsymbol{\theta})$, could be estimated by $\boldsymbol{J}^{-1}(\hat{\boldsymbol{\theta}})$ and denoted by $\widehat{\text{Cov}}(\hat{\boldsymbol{\theta}})$.

C. Derivative of a vector/matrix

Let $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ and $\boldsymbol{a} = (a_1, \ldots, a_n)^\top$ be two $n \times 1$ vectors, $\boldsymbol{b} = (b_1, \ldots, b_m)^\top$ an $m \times 1$ vector,

$$\boldsymbol{A}_{m\times n} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \quad \text{and} \quad \boldsymbol{B}_{n\times n} = \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nn} \end{pmatrix}.$$

Define

$$\frac{\partial \boldsymbol{b}^\top}{\partial \boldsymbol{x}} = \left(\frac{\partial b_1}{\partial \boldsymbol{x}}, \ldots, \frac{\partial b_m}{\partial \boldsymbol{x}}\right) = \begin{pmatrix} \frac{\partial b_1}{\partial x_1} & \frac{\partial b_2}{\partial x_1} & \cdots & \frac{\partial b_m}{\partial x_1} \\ \frac{\partial b_1}{\partial x_2} & \frac{\partial b_2}{\partial x_2} & \cdots & \frac{\partial b_m}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial b_1}{\partial x_n} & \frac{\partial b_2}{\partial x_n} & \cdots & \frac{\partial b_m}{\partial x_n} \end{pmatrix}.$$

We have

(a) $\partial(\boldsymbol{a}^\top\boldsymbol{x})/\partial\boldsymbol{x} = \boldsymbol{a}$

(b) $\partial(\boldsymbol{A}\boldsymbol{x})/\partial\boldsymbol{x}^\top = \boldsymbol{A}$

(c) $\partial(\boldsymbol{A}\boldsymbol{x})^\top/\partial\boldsymbol{x} = \boldsymbol{A}^\top$

(d) $\partial(\boldsymbol{x}^\top\boldsymbol{B}\boldsymbol{x})/\partial\boldsymbol{x} = (\boldsymbol{B}+\boldsymbol{B}^\top)\boldsymbol{x}$

(e) $\partial^2(\boldsymbol{x}^\top\boldsymbol{B}\boldsymbol{x})/\partial\boldsymbol{x}\partial\boldsymbol{x}^\top = \boldsymbol{B}+\boldsymbol{B}^\top$.

Proof: (a) Since $\boldsymbol{a}^\top\boldsymbol{x} = a_1x_1 + a_2x_2 + \cdots + a_nx_n$, we have

$$\frac{\partial(\boldsymbol{a}^\top\boldsymbol{x})}{\partial\boldsymbol{x}} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = \boldsymbol{a}.$$

(b) Note that

$$\boldsymbol{A}\boldsymbol{x} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \end{pmatrix},$$

then

$$\begin{aligned} &\frac{\partial(\boldsymbol{A}\boldsymbol{X})}{\partial\boldsymbol{x}^\top} \\ &= \begin{pmatrix} \dfrac{\partial(a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n)}{\partial x_1} & \cdots & \dfrac{\partial(a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n)}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial(a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n)}{\partial x_1} & \cdots & \dfrac{\partial(a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n)}{\partial x_n} \end{pmatrix} \\ &= \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} = \boldsymbol{A}. \end{aligned}$$

(c) Since $(\boldsymbol{Ax})^\top = (a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n, \cdots, a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n)$, we obtain

$$
\begin{aligned}
&\frac{\partial(\boldsymbol{Ax})^\top}{\partial \boldsymbol{x}} \\
= &\begin{pmatrix} \dfrac{\partial(a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n)}{\partial x_1} & \cdots & \dfrac{\partial(a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_1)}{\partial x_1} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial(a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n)}{\partial x_n} & \cdots & \dfrac{\partial(a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n)}{\partial x_n} \end{pmatrix} \\
= &\begin{pmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{mn} \end{pmatrix} = \boldsymbol{A}^\top.
\end{aligned}
$$

(d) Since

$$
\begin{aligned}
&\boldsymbol{x}^\top \boldsymbol{B} \boldsymbol{x} \\
= &(x_1, x_2, \cdots, x_n) \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \\
= &(x_1 b_{11} + x_2 b_{21} + \cdots + x_n b_{n1}, \cdots, x_1 b_{1n} + x_2 b_{2n} + \cdots + x_n b_{nn}) \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \\
= &(x_1^2 b_{11} + x_1 x_2 b_{21} + \cdots + x_1 x_n b_{n1}) + \cdots + (x_n x_1 b_{1n} + x_n x_2 b_{2n} + \cdots + x_n^2 b_{nn}) \\
= &\sum_{i=1}^n \sum_{j=1}^n x_i b_{ij} x_j,
\end{aligned}
$$

we obtain

$$
\begin{aligned}
\frac{\partial(\boldsymbol{x}^{\top}\boldsymbol{B}\boldsymbol{x})}{\partial \boldsymbol{x}} &= \begin{pmatrix} 2b_{11}x_1 + (b_{12}+b_{21})x_2 + (b_{13}+b_{31})x_3 + \cdots + (b_{1n}+b_{n1})x_n \\ (b_{12}+b_{21})x_1 + 2b_{22}x_2 + (b_{23}+b_{32})x_3 + \cdots + (b_{2n}+b_{n2})x_n \\ \vdots \\ (b_{1n}+b_{n1})x_1 + (b_{2n}+b_{n2})x_2 + (b_{3n}+b_{n3})x_3 + \cdots + 2b_{nn}x_n \end{pmatrix} \\
&= \begin{pmatrix} b_{11}+b_{11} & b_{12}+b_{21} & \cdots & b_{1n}+b_{n1} \\ b_{21}+b_{12} & b_{22}+b_{22} & \cdots & b_{2n}+b_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1}+b_{1n} & b_{n2}+b_{2n} & \cdots & b_{nn}+b_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \\
&= \left(\begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nn} \end{pmatrix} + \begin{pmatrix} b_{11} & b_{21} & \cdots & b_{n1} \\ b_{12} & b_{22} & \cdots & b_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ b_{1n} & b_{2n} & \cdots & b_{nn} \end{pmatrix} \right) \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \\
&= (\boldsymbol{B} + \boldsymbol{B}^{\top})\boldsymbol{x}.
\end{aligned}
$$

(e)

$$
\frac{\partial^2(\boldsymbol{x}^{\top}\boldsymbol{B}\boldsymbol{x})}{\partial \boldsymbol{x}\partial \boldsymbol{x}^{\top}} = \begin{pmatrix} 2b_{11} & b_{12}+b_{21} & \cdots & b_{1n}+b_{n1} \\ b_{12}+b_{21} & 2b_{22} & \cdots & b_{2n}+b_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ b_{1n}+b_{n1} & b_{2n}+b_{n2} & \cdots & 2b_{nn} \end{pmatrix} = \boldsymbol{B} + \boldsymbol{B}^{\top}. \qquad \|
$$

Example T4.3 (Poisson regression). Let $Y_{\text{obs}} = \{y_i\}_{i=1}^{n}$ and consider the following Poisson regression

$$
Y_i \stackrel{\text{ind}}{\sim} \text{Poisson}(\lambda_i), \quad \log(\lambda_i) = \boldsymbol{x}_{(i)}^{\top}\boldsymbol{\theta}, \quad 1 \leqslant i \leqslant n,
$$

where $\boldsymbol{x}_{(i)}$ is the $q \times 1$ covariates vector, and $\boldsymbol{\theta}_{q\times 1}$ is the unknown parameter vector.

(a) Derive the score vector and the observed information matrix.

(b) Using the Newton-Raphson algorithm to find the MLE $\hat{\boldsymbol{\theta}}$ and the estimated asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$.

Solution: (a) The log-likelihood function of $\boldsymbol{\theta}$, the score vector and the observed information matrix are

$$\begin{aligned}
\ell(\boldsymbol{\theta}|Y_{\text{obs}}) &= \log\left[\prod_{i=1}^n \Pr(Y_i = y_i)\right] = \log\left[\prod_{i=1}^n \frac{\lambda_i^{y_i}}{y_i!}\exp(-\lambda_i)\right] \\
&= \sum_{i=1}^n y_i \log(\lambda_i) - \sum_{i=1}^n \log(y_i!) - \sum_{i=1}^n \lambda_i \\
&= \sum_{i=1}^n y_i \boldsymbol{x}_{(i)}^\top \boldsymbol{\theta} - \sum_{i=1}^n \log(y_i!) - \sum_{i=1}^n \exp(\boldsymbol{x}_{(i)}^\top \boldsymbol{\theta}) \\
\nabla\ell(\boldsymbol{\theta}|Y_{\text{obs}}) &= \sum_{i=1}^n y_i \boldsymbol{x}_{(i)} - \sum_{i=1}^n \exp(\boldsymbol{x}_{(i)}^\top \boldsymbol{\theta})\boldsymbol{x}_{(i)} \\
-\nabla^2\ell(\boldsymbol{\theta}|Y_{\text{obs}}) &= \sum_{i=1}^n \exp(\boldsymbol{x}_{(i)}^\top \boldsymbol{\theta}).\boldsymbol{x}_{(i)}\boldsymbol{x}_{(i)}^\top
\end{aligned}$$

(b) The iteration of the Newton–Raphson algorithm is

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \left[\sum_{i=1}^n \exp(\boldsymbol{x}_{(i)}^\top \boldsymbol{\theta}^{(t)})\boldsymbol{x}_{(i)}\boldsymbol{x}_{(i)}^\top\right]^{-1}\left[\sum_{i=1}^n y_i\boldsymbol{x}_{(i)} - \sum_{i=1}^n \exp(\boldsymbol{x}_{(i)}^\top \boldsymbol{\theta}^{(t)})\boldsymbol{x}_{(i)}\right].$$

The estimated asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$ is

$$\widehat{\text{Cov}}(\hat{\boldsymbol{\theta}}) = [-\nabla^2\ell(\hat{\boldsymbol{\theta}}|Y_{\text{obs}})]^{-1} = \left[\sum_{i=1}^n \exp(\boldsymbol{x}_{(i)}^\top \hat{\boldsymbol{\theta}})\boldsymbol{x}_{(i)}\boldsymbol{x}_{(i)}^\top\right]^{-1}.$$

Note that the observed information matrix does not depend on the observation data Y_{obs}, then the expected covariance matrix is also the observed one. ‖