STAT 3006: Statistical Computing
Lecture 2*

15 January

2.2 Functional Iteration

When we search a maximum (or minimum) for a differentiable function $h(x)$, we usually turn to solving the equation $\frac{dh(x)}{dx} = 0$, i.e.

$$\frac{dh(x)}{dx} + x = x. \tag{2.1}$$

Let $f(x)$ be $\frac{dh(x)}{dx} + x$, the equation (2.1) becomes

$$f(x) = x. \tag{2.2}$$

All x^* solving equation (2.2) ($f(x^*) = x^*$) are called the *fixed points* of $f(x)$. Generally, our problem is that, for a function f which may be non-differentiable, we would like to find a *fixed point* of $f(x)$.

Algorithm: Fixed point finding algorithm.

Input: continuous and univariate function f, maximum number of iterations T, and tolerance ϵ; initial point $x^{(0)}$.
Output: $x^{(t)}$ in the last iteration.
1: $t \leftarrow 0$
2: **repeat**
3: let y be $x^{(t)}$;
4: calculate $x^{(t+1)} = f(y)$;
5: $t \leftarrow t + 1$;
6: **until** $|x^{(t)} - y| < \epsilon$ or $t \geq T$.

*If you have any question about the note, please send an email to xyluo@link.cuhk.edu.hk

Example: Given a positive number a, find $\sqrt{a}$.

Solution: notice that $\sqrt{a}$ is the solution of the equation $\frac{1}{2}(\frac{a}{x} - x) = 0$. Let $f(x) = \frac{1}{2}(\frac{a}{x} - x) + x = \frac{1}{2}(\frac{a}{x} + x)$, we implement the algorithm above by $x^{(t+1)} = \frac{1}{2}(\frac{a}{x^{(t)}} + x^{(t)})$.

Q Why don't we take $\tilde{f}(x) = (\frac{a}{x} - x) + x = \frac{a}{x}$? We have the following proposition.

Proposition 2.1. *Suppose $f : I \to \mathbb{R}$, where I is a closed interval such that*

(1) $f(x) \in I$ for $\forall x$.

(2) $|f(y) - f(x)| \leq \lambda |y - x|$ (Lipschitz continuous) for a constant λ (Lipschitz constant) and $\forall x, y \in I$.

If $\lambda \in [0, 1)$, then

(1) $f(x)$ has a unique fixed point $x_\infty \in I$.

(2) the sequence $x_n = f(x_{n-1})$ goes to x_∞, $\forall x_0 \in I$.

(3) $|x_n - x_\infty| \leq \frac{\lambda^n}{1-\lambda}|x_1 - x_0|$.

Proof.

$$\begin{aligned} |x_{k+1} - x_k| &= |f(x_k) - f(x_{k-1})| \\ &\leq \lambda |x_k - x_{k-1}| \leq \lambda^2 |x_{k-1} - x_{k-2}| \leq \ldots \leq \lambda^k |x_1 - x_0| \\ \forall m > n, \quad |x_m - x_n| &\leq \sum_{k=n}^{m-1} |x_{k+1} - x_k| \leq \sum_{k=n}^{m-1} \lambda^k |x_1 - x_0| \leq \frac{\lambda^n}{1-\lambda}|x_1 - x_0| \end{aligned}$$

The last inequality indicates that $\{x_n\}_{n=1}^{\infty}$ is a Cauchy sequence. In $\mathbb{R}$, Cauchy sequence implies the convergence of the sequence, so $\{x_n\}_{n=1}^{\infty}$ converges to a point x_∞. Moreover, $\{x_n\}_{n=1}^{\infty} \in I$ and I is closed, so $x_\infty \in I$. (2) is proved.

For the equation $x_n = f(x_{n-1})$, let n go to infinity and notice f is continuous, so we have $x_\infty = f(x_\infty)$. If there exists a $y \neq x_\infty$ s.t. $y = f(y)$, then

$$\begin{aligned} |y - x_\infty| &= |f(y) - f(x_\infty)| \\ &\leq \lambda |y - x_\infty| \\ &< |y - x_\infty|. \end{aligned}$$

The last inequality holds, because $\lambda \in [0, 1)$. It is contradictory that $|y - x_\infty| < |y - x_\infty|$, so x_∞ is the unique fixed point of f. We proved (1). (3) can be easily proved, so we omit it.

Example (continuing) $\tilde{f}(x) = \frac{a}{x}$, $x > 0$,

$$|\tilde{f}(y) - \tilde{f}(x)| = |\frac{a}{y} - \frac{a}{x}| = |\frac{a(x-y)}{xy}|$$
$$= |\frac{a}{xy}||y - x| = \frac{a}{xy}|y - x|.$$

We need to find $I = [c, d]$ such that

(1) $\tilde{f}(x) \in [c, d]$, $\forall x \in [c, d]$;

(2) $\frac{a}{xy} < 1$, $\forall x \in [c, d]$.

(2) implies that $\frac{a}{c^2} < 1$, so $c > \sqrt{a}$, $\sqrt{a} \notin I = [c, d]$. Therefore, we do not use $\tilde{f}(x)$ as the iteration operator to find $\sqrt{a}$.

As to $f(x) = \frac{1}{2}(\frac{a}{x} + x)$, $x > 0$,

$$|f(y) - f(x)| = |\frac{1}{2}(\frac{a}{y} + y) - \frac{1}{2}(\frac{a}{x} + x)|$$
$$= \frac{1}{2}|\frac{a}{y} - \frac{a}{x} + (y - x)|$$
$$= \frac{1}{2}|\frac{a}{xy}(x - y) + (y - x)| = \frac{1}{2}|1 - \frac{a}{xy}||y - x|.$$

Consider the interval $I = [\sqrt{\frac{2a}{3}}, \sqrt{2a}]$, $\sqrt{a} \in I$. For $\forall x \in I$, $f(x) \in I$. Additionally, for $\forall x, y \in I$, $|1 - \frac{a}{xy}| \leq \frac{1}{2}$, so $f(x)$ is Lipschitz continuous on I. Therefore, we can use the iterated operation $x_n = \frac{1}{2}(\frac{a}{x_{n-1}} + x_{n-1})$ to approximate $\sqrt{a}$.

For illustration, when $a = 2$ and $x_0 = 1.7$,

$$x_1 = \tilde{f}(x_0) = \frac{2}{1.7} = 1.176471$$
$$x_2 = \tilde{f}(x_1) = x_0 = 1.7$$
$$x_3 = x_1$$
$$x_4 = x_2 = x_0 \ldots$$

In contrast,

$$x_1 = f(x_0) = \frac{1}{2}(\frac{2}{1.7} + 1.7) = 1.438235$$
$$x_2 = f(x_1) = \frac{1}{2}(\frac{2}{1.438235} + 1.438235) = 1.414414$$
$$x_3 = 1.414214 \ldots$$

After three iterations, the result is very close to $\sqrt{2}$.

Q: How to verify f satisfies the two requirements of the proposition?

Lagrange's Mean Value Theorem: if f is continuous on the closed interval $[a, b]$ and differentiable on the open interval (a, b), then there exists a point ξ in (a, b) such that

$$f'(\xi) = \frac{f(b) - f(a)}{b - a}.$$

Solution(a sufficient condition): first, we have to find an interval $[a, b]$ s.t. f is continuous on $[a, b]$ and differentiable on (a, b), and $f(x) \in [a, b]$ when $x \in [a, b]$. Second, by the mean value theorem, if there exist a constant λ s.t. $1 > \lambda \geq \sup_{\xi \in (a,b)} |f'(\xi)|$, then $|f(x) - f(y)| \leq \lambda |x - y|$. When the two conditions hold, the corresponding f satisfies the two requirements of the proposition.

2.3 Newton's method

In the section, we provide another method called Newton's method to find the maximum (or minimum) for a function f. Assume function $f(x)$ is twice differentiable. Let $g(x)$ be $f'(x)$. In most cases, finding optimum of $f(x)$ is equivalent to finding the solution of the equation $g(x) = 0$. We will give two perspectives that motivates the Newton method.

1.(See Figure 1) Considering the equation $g(x) = 0$, from a starting point $x^{(0)}$, we draw a line that is tangent to $g(x)$ at point $(x_0, g(x_0))$. We can regard this line as an locally approximate curve to $g(x)$. After some simple algebra, this line $l_0(x)$ has the expression $l_0(x) = g(x_0) + g'(x_0)(x - x_0)$. As $l_0(x)$ is approximate to $g(x)$, the solution of $l_0(x) = 0$ is probably close to the solution of $g(x) = 0$. By solving $l_0(x) = 0$, we get the solution $x_1 = x_0 - \frac{g(x_0)}{g'(x_0)}$. Repeat the procedure, and then we have the general step $x_n = x_{n-1} - \frac{g(x_{n-1})}{g'(x_{n-1})}$ to find the solution of $g(x) = 0$.

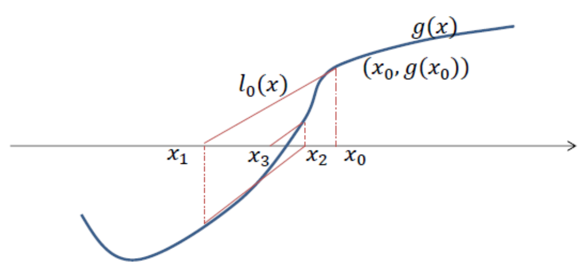

Figure 1: Figure demonstration for the Newton's method to solve $g(x) = 0$.

2. Notice that when we minimize (or maximize) a *convex* function $f(x)$, the problem is equivalent to finding the solution $g(x) = f'(x) = 0$. Plug $f'(x)$ into $x_n = x_{n-1} - \frac{g(x_{n-1})}{g'(x_{n-1})}$, we have $x_n = x_{n-1} - \frac{f'(x_{n-1})}{f''(x_{n-1})}$. What does that mean? (See Figure 2)when we minimize $f(x)$, given a starting point x_0, the Taylor expansion of $f(x)$ at x_0 (omit cubic term and terms with higher order) is $q_0(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2}(x - x_0)^2$. $q_0(x)$ can be regarded as an locally approximate curve to the function $f(x)$. Therefore, the point that minimizes $q_0(x)$ is probably close to the point that minimized $f(x)$. By minimizing $q_0(x)$, we get the point $x_1 = x_0 - \frac{f'(x_0)}{f''(x_0)}$. Repeat the procedure multiple times, we have the general step: $x_n = x_{n-1} - \frac{f'(x_{n-1})}{f''(x_{n-1})}$.

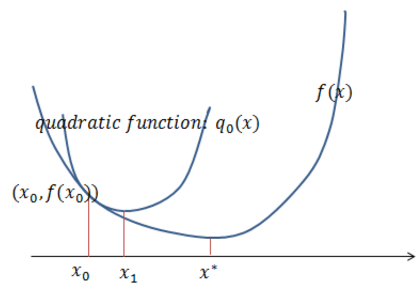

Figure 2: Figure demonstration for the Newton's method to minimize $f(x)$.

2.4 Rate of convergence

Definition 2.1. Assume $\{x_n\}_{n=0}^{\infty} \to x^*$. If $\exists p \geq 1$ and $\alpha > 0$ s.t. $\lim_{n\to\infty} \frac{\|x_{n+1} - x_\infty\|}{\|x_n - x_\infty\|^p} = \alpha$, then $\{x_n\}_{n=0}^{\infty}$ is p-order convergence.

- p = 1, linear convergence.
- p > 1, super-linear convergence.
- p = 2, quadratic convergence.

Theorem 2.2. *if* $\{x_n\}_{n=0}^{\infty}$ *super-linearly converges to* x_∞, *then when* $x_n \neq x_\infty$, $\lim_{n\to\infty} \frac{\|x_{n+1} - x_n\|}{\|x_n - x_\infty\|} = 1$.

Proof.

$$\begin{aligned}
\lim_{n\to\infty} \left| \frac{\|x_{n+1} - x_n\|}{\|x_n - x_\infty\|} - 1 \right| &= \lim_{n\to\infty} \left| \frac{\|x_{n+1} - x_n\| - \|x_n - x_\infty\|}{\|x_n - x_\infty\|} \right| \\
&\leq \frac{\|x_{n+1} - x_\infty\|}{\|x_n - x_\infty\|} = 0.
\end{aligned}$$

When a sequence is super-linear convergence, we can use $\|x_{n+1} - x_n\| < \epsilon$ as a stopping rule.

For Newton's method, let $M(x)$ be $x - \frac{g(x)}{g'(x)}$.

$$M'(x) = 1 - \frac{g'(x)}{g'(x)} + \frac{g(x)g''(x)}{g'(x)^2} = \frac{g(x)g''(x)}{g'(x)^2}$$
$$M'(x_\infty) = \frac{g(x_\infty)g''(x_\infty)}{g'(x_\infty)^2} = 0$$

The last equation holds since $g(x_\infty) = 0$.

$$\begin{aligned} x_n - x_\infty &= M(x_{n-1}) - M(x_\infty) \\ &= (Taylor \quad expansion) M'(x_\infty)(x_{n-1} - x_\infty) + \frac{1}{2}M''(z_n)(x_{n-1} - x_\infty)^2 \\ &= \frac{1}{2}M''(z_n)(x_{n-1} - x_\infty)^2. \end{aligned}$$

It follows that

$$\lim_{n\to\infty} \frac{\|x_n - x_\infty\|}{\|x_{n-1} - x_\infty\|^2} = \lim_{n\to\infty} \frac{1}{2}M''(z_n) = \frac{1}{2}M''(x_\infty).$$

Therefore, Newton sequence is quadratic convergence.

Example: Given a, we need to find $\frac{1}{a}$. Construct $g(x) = a - \frac{1}{x}$, then the Newton iteration is $x_{n+1} = x_n(2 - ax_n)$.

2.5 Multivariate case

So far we have talked about the application of Newton's method to the univariate function $f(x)$ (or $g(x)$). Next, we will discuss the Newton's method for a mapping $\vec{F}$ (e.g. $\vec{F} : \mathbb{R}^3 \to \mathbb{R}^3$). We consider the mapping $\vec{F}(\vec{x})$ from a $\mathbb{R}^m$ domain D to $\mathbb{R}^m$, where $\vec{x} = (x_1, x_2, \ldots, x_m)$ and $\vec{F}(\vec{x}) = (f_1(\vec{x}), f_2(\vec{x}), \ldots, f_m(\vec{x}))$. Our goal is to solve the equation system $\vec{F}(\vec{x}) = \vec{0}$.

Given current point $\vec{x}^{(n)}$, we carry out Taylor expansion for $f_i(\vec{x})$ $(i = 1, \ldots, m)$ at $\vec{x}^{(n)}$,

$$f_i(\vec{x}) \approx f_i(\vec{x}^{(n)}) + \frac{\partial f_i}{\partial x_1}(\vec{x}^{(n)})(x_1 - x_1^{(n)}) + \ldots + \frac{\partial f_i}{\partial x_m}(\vec{x}^{(n)})(x_m - x_m^{(n)}).$$

The equation above holds for $i = 1, \ldots, m$. We put these m equations together, which become

$$\vec{F}(\vec{x}) \approx \vec{F}(\vec{x}^{(n)}) + \vec{F}'(\vec{x}^{(n)})(\vec{x} - \vec{x}^{(n)}), \tag{2.3}$$

where the Jacobian matrix of $\vec{F}$ is

$$F'(\vec{x}_n) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(\vec{x}^{(n)}) & \cdots & \frac{\partial f_1}{\partial x_m}(\vec{x}^{(n)}) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(\vec{x}^{(n)}) & \cdots & \frac{\partial f_m}{\partial x_m}(\vec{x}^{(n)}) \end{pmatrix},$$

and

$$(\vec{x} - \vec{x}^{(n)}) = \begin{pmatrix} x_1 - x_1^{(n)} \\ \vdots \\ x_m - x_m^{(n)} \end{pmatrix}.$$

Let the left hand side of equation (2.3) be zero. It yields that

$$\vec{x}^{(n+1)} = \vec{x}^{(n)} - (\vec{F}'(\vec{x}^{(n)}))^{-1}\vec{F}(\vec{x}^{(n)}).$$

The equation above can be decomposed to two steps:

- solve $\vec{F}'(\vec{x}_n)\Delta x^{(n)} = -\vec{F}(x^{(n)})$;
- $x^{(n+1)} = x^{(n)} + \Delta x^{(n)}$.

Example(calculate MLE): $l(\Theta|x_1, \ldots, x_n) = logL(\Theta|x_1, \ldots, x_n)$. Under some regular conditions, $T\hat{h}eta$ solves the following equation,

$$\begin{pmatrix} \frac{\partial l}{\partial \theta_1} \\ \vdots \\ \frac{\partial l}{\partial \theta_m} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}.$$

By Newton's method, we iteratively update the $\Theta^{(n)}$ according to

$$\Theta^{(n+1)} = \Theta^{(n)} - \begin{pmatrix} \frac{\partial^2 l}{\partial \theta_1 \partial \theta_1}(\Theta^{(n)}) & \cdots & \frac{\partial^2 l}{\partial \theta_1 \partial \theta_m}(\Theta^{(n)}) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 l}{\partial \theta_m \partial \theta_1}(\Theta^{(n)}) & \cdots & \frac{\partial^2 l}{\partial \theta_m \partial \theta_m}(\Theta^{(n)}) \end{pmatrix}^{-1} \begin{pmatrix} \frac{\partial l}{\partial \theta_1}(\Theta^{(n)}) \\ \vdots \\ \frac{\partial l}{\partial \theta_m}(\Theta^{(n)}) \end{pmatrix}. \tag{2.4}$$

Example (MLE of Poisson distribution):

$$\begin{aligned} f(y_1, \ldots, y_n|\lambda) &= \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!} \\ l(\lambda|y_1, \ldots, y_n) &= \sum_{i=1}^{n}(y_i \log\lambda - \lambda - \log y_i!) \\ &= (\sum_{i=1}^{n} y_i)\log\lambda - n\lambda - \sum_{i=1}^{n}\log y_i! \\ \frac{dl}{d\lambda} &= \frac{\sum_{i=1}^{n} y_i}{\lambda} - n. \end{aligned}$$

- MLE direct derivation: $\hat{\lambda} = \frac{\sum_{i=1}^{n} y_i}{n}$.
- Newton's method to solve: $\lambda_{k+1} = \lambda_k + \frac{\lambda_k^2}{\sum_{i=1}^{n} y_i}(\frac{\sum_{i=1}^{n} y_i}{\lambda_k} - n)$.

Example (Poisson regression):
We have independent count data $\{y_1, \ldots, y_n\}$. For each Y_i, Y_i follows $Poi(\lambda_i)$, where $log(\lambda_i) =$

$\alpha + \beta x_i$, α and β are parameters and x_i is the fixed covariate. The p.d.f (probability density function) of y_i is $f(y_i|\alpha, \beta, x_i) = e^{-e^{(\alpha+\beta x_i)}} \frac{(e^{\alpha+\beta x_i})^{y_i}}{y_i!}$. It follows that the joint p.d.f. is

$$f(y_1, y_2, \ldots, y_n|\alpha, \beta) = \prod_{i=1}^{n} e^{-e^{(\alpha+\beta x_i)}} \frac{(e^{\alpha+\beta x_i})^{y_i}}{y_i!}.$$

$$\begin{aligned}
l(\alpha, \beta) = \log f(y_1, y_2, \ldots, y_n|\alpha, \beta) &= \sum_{i=1}^{n} [-e^{\alpha+\beta x_i} + y_i(\alpha + \beta x_i) - \log y_i!] \\
\frac{\partial l}{\partial \alpha} &= -\sum_{i=1}^{n} e^{\alpha+\beta x_i} + \sum_{i=1}^{n} y_i \\
\frac{\partial l}{\partial \beta} &= -\sum_{i=1}^{n} x_i e^{\alpha+\beta x_i} + \sum_{i=1}^{n} x_i y_i \\
\frac{\partial^2 l}{\partial \alpha^2} &= -\sum_{i=1}^{n} e^{\alpha+\beta x_i} \\
\frac{\partial^2 l}{\partial \alpha \partial \beta} &= -\sum_{i=1}^{n} x_i e^{\alpha+\beta x_i} \\
\frac{\partial^2 l}{\partial \beta^2} &= -\sum_{i=1}^{n} x_i^2 e^{\alpha+\beta x_i}.
\end{aligned}$$

The Newton step is

$$\begin{pmatrix} \alpha_{k+1} \\ \beta_{k+1} \end{pmatrix} = \begin{pmatrix} \alpha_k \\ \beta_k \end{pmatrix} - \begin{pmatrix} -\sum_{i=1}^{n} e^{\alpha_k+\beta_k x_i} & -\sum_{i=1}^{n} x_i e^{\alpha_k+\beta_k x_i} \\ -\sum_{i=1}^{n} x_i e^{\alpha_k+\beta_k x_i} & -\sum_{i=1}^{n} x_i^2 e^{\alpha_k+\beta_k x_i} \end{pmatrix}^{-1} \begin{pmatrix} -\sum_{i=1}^{n} e^{\alpha_k+\beta_k x_i} + \sum_{i=1}^{n} y_i \\ -\sum_{i=1}^{n} x_i e^{\alpha_k+\beta_k x_i} + \sum_{i=1}^{n} x_i y_i \end{pmatrix}.$$