

Outline of the lecture

STK4051/STK9051 Computational Statistics

Riccardo De Bin

debin@math.uio.no

- Univariate Problems
 - Newton's method
 - Fisher scoring
 - Secant method
 - Fixed-point method
- Multivariate Problems
 - Newton's Method and Fisher Scoring
 - Iteratively Re-weighted Least Square

Univariate Problems: Newton's method

Newton's method is based on the **first-order Taylor series expansion**. Suppose $g'(x)$ continuously differentiable and $g''(x^*) \neq 0$ in a suitable interval around x^* . Then

$$0 = g'(x^*) \approx g'(x^{(t)}) + (x^* - x^{(t)})g''(x^{(t)})$$

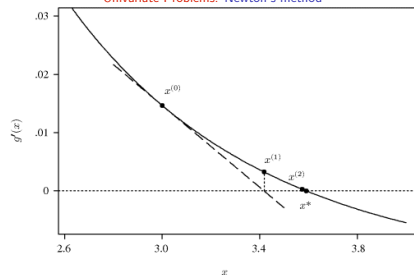
Solving for x^* ,

$$x^* = \underbrace{x^{(t)}}_{\text{current guess}} - \underbrace{\frac{g'(x^{(t)})}{g''(x^{(t)})}}_{\text{refinement}}.$$

This leads to the **updating equation** for Newton's method,

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

Univariate Problems: Newton's method



(Givens & Hoeting, 2013, figure 2.3)

Univariate Problems: Newton's method

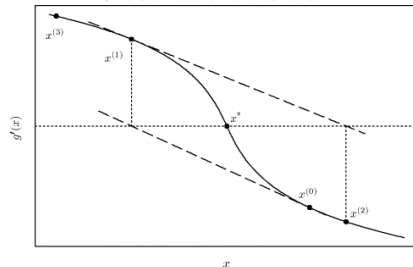
The step for this function ($g(x) = \log(x)/(1+x)$) is

$$x^{(t+1)} = x^{(t)} - \frac{(x^{(t)} + 1)(1 + 1/x^{(t)} - \log x^{(t)})}{-3 - 4/x^{(t)} - 1/(x^{(t)})^2 + 2 \log x^{(t)}}.$$

Starting from $x^{(0)} = 3.0$:

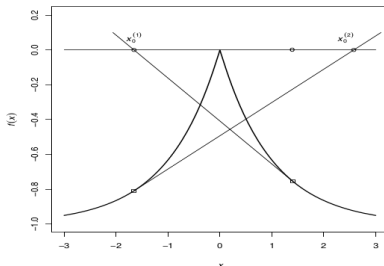
- we get $x^{(4)} \approx 3.59112$;
- the bisection methods takes 19 iterations to correctly determine the first 5 decimal;
- when converging, the order of **convergence is quadratic**:
 - when does it converge?
 - how to compute the order of convergence?

Univariate Problems: Newton's method



(Givens & Hoeting, 2013, figure 2.4)

Univariate Problems: Newton's method



(Gentle, 2009, figure 6.4)

Univariate Problems: Newton's method

Theorem: If g' has two continuous derivatives at x^* and $g''(x^*) \neq 0$, x^* being a simple root of g' , then $\exists$ a neighbourhood of x^* for which Newton's method **converges to x^*** when started from any x in that neighbourhood.

Proof: $g''(x^*) \neq 0$ and g'' is continuous at x^* , $\exists$ a neighbourhood of x^* within which $g''(x) \neq 0 \forall x$. Let us confine interest to this neighborhood, and define $\epsilon(t) = x(t) - x^*$.

A Taylor expansion yields

$$0 = g'(x^*) = g'(x^{(t)}) + (x^* - x^{(t)})g''(x^{(t)}) + (x^* - x^{(t)})^2 \frac{g'''(q)}{2}$$

for some q between $x^{(t)}$ and x^* .

Univariate Problems: Newton's method

Rearranging terms,

$$\underbrace{x^{(t)} - \frac{g'(x^{(t)})}{g''(x^{(t)})}}_{x^{(t+1)}} - x^* = \underbrace{(x^* - x^{(t)})^2}_{(\epsilon^{(t)})^2} \frac{g'''(q)}{2g''(x^{(t)})},$$

and since $x^{(t)} - x^* = \epsilon^{(t+1)}$,

$$\epsilon^{(t+1)} = \left(\epsilon^{(t)}\right)^2 \frac{g'''(q)}{2g''(x^{(t)})}.$$

Univariate Problems: Newton's method

Suppose that the starting value is not too bad, in the sense that $|x^{(0)} - x^*| \leq \delta$ (i.e., $\epsilon^{(0)} \leq \delta$), then

$$|\epsilon^{(t)}| = \frac{(c(\delta)\delta)^{2^t}}{c(\delta)},$$

which converges to zero as $t \rightarrow \infty$. Hence $x^{(t)} \rightarrow x^*$.

Univariate Problems: Newton's method

Define the neighbourhood of x^* , $\mathcal{N}_\delta(x^*) = [x^* - \delta, x^* + \delta]$, $\delta > 0$.

Let

$$c(\delta) = \max_{x_1, x_2 \in \mathcal{N}_\delta(x^*)} \left| \frac{g'''(x_1)}{2g''(x_2)} \right|$$

Since $c(\delta) \rightarrow \left| \frac{g'''(x^*)}{2g''(x^*)} \right|$ as $\delta \rightarrow 0$, then $\delta c(\delta) \rightarrow 0$ as $\delta \rightarrow 0$. If one chooses δ such that $\delta c(\delta) < 1$ and if $x^{(t)} \in \mathcal{N}_\delta(x^*)$, then

$$|c(\delta)\epsilon^{(t+1)}| = \left| \left(\epsilon^{(t)}\right)^2 \frac{g'''(q)}{2g''(x^{(t)})} c(\delta) \right| \leq \left(\epsilon^{(t)}\right)^2 c(\delta)^2.$$

Alternative check:

When g' is twice continuously differentiable, is **convex**, and has a root, then Newton's method converges from any starting point.

Another set of conditions is:

If

- $g''(x) \neq 0$ on $[a, b]$,
- $g'''(x)$ does not change sign on $[a, b]$,
- $g'(a)g'(b) < 0$,
- $g'(a)/g''(a) < b - a$ and $g'(b)/g''(b) < b - a$,

then Newton's method will converge from any $x^{(0)}$ in $[a, b]$.

Univariate Problems: Convergence order

The speed of a root-finding approach is typically measured by its **order of convergence**. A method has convergence of order β if $\lim_{t \rightarrow \infty} \epsilon(t) = 0$ and

$$\lim_{t \rightarrow \infty} \frac{|\epsilon(t+1)|}{|\epsilon(t)|^\beta} = c$$

for some constants $c \neq 0$ and $\beta > 0$. Note:

- **higher orders** of convergence means **faster** approximation to the true solution;
- faster algorithms are sometimes more prone **to not converge** to the right solution: **trade-off** between speed and robustness.

Univariate Problems: Newton's method

Newton's method in **algorithmic form**:

0. Set $t = 0$, and find an approximate solution $x^{(0)}$.
1. Set $x^{(t+1)} = x^{(t)} - \frac{g'(x^{(t)})}{g''(x^{(t)})}$
2. if $|x^{(t+1)} - x^{(t)}| < \epsilon$, **return** $x^{(t+1)}$,
otherwise, set $t = t + 1$ and go back to step 1.

Note:

- whether Newton's method converges depends on the **shape of g** and the **initial value $x^{(0)}$** .

Univariate Problems: Convergence order

For **Newton's method**,

$$\lim_{t \rightarrow \infty} \frac{|\epsilon(t+1)|}{|\epsilon(t)|^2} = \frac{g'''(x^*)}{2g''(x^*)}.$$

Therefore, if Newton's method converges,

- $c = g'''(x^*)/[2g''(x^*)]$;
- $\beta = 2$.

The **bisection method** does not satisfy the condition for determining order of convergence:

- there may be iteration in which $x^{(t)} - x^*$ **does not shrink**;
- it is consider having a **linear convergence** ($\beta = 1$) because:
 - the interval length is **halved** at each iteration;
 - $\lim_{t \rightarrow \infty} |\epsilon(t)| = 0$ if $\exists$ a root in the starting interval.

Univariate Problems: Fisher scoring

When we deal with **maximum likelihood estimation**,

$$\theta^{(t+1)} = \theta^{(t)} - \frac{\ell_{\theta}(\theta^{(t)})}{\ell_{\theta\theta}(\theta^{(t)})}$$

Since $I(\theta)$ is an alternative of $-\ell_{\theta\theta}(\theta^{(t)})$, it is reasonable to use

$$\theta^{(t+1)} = \theta^{(t)} + \frac{\ell_{\theta}(\theta^{(t)})}{I(\theta^{(t)})}.$$

This approach is called **Fisher scoring**.

- Fisher scoring and Newton's method have the **same asymptotic properties**;
- for **individual problems** one may be computationally or analytically easier than the other.

Univariate Problems: Secant method

The refining term of the Newton's method involves $g''(x^{(t)})$:

- it may be difficult to compute;
- it can be approximated by $[g'(x^{(t)}) - g'(x^{(t-1)})]/[x^{(t)} - x^{(t-1)}]$.

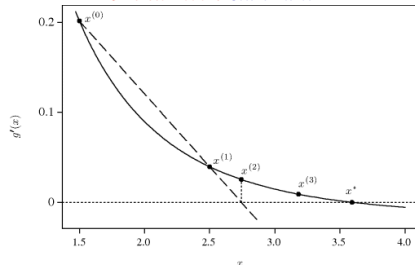
The result is the secant method, which has updating equation

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

Note:

- two starting points are needed ($x^{(0)}$ and $x^{(1)}$);
- local approximation of g' using the secant line between $x^{(0)}$ and $x^{(1)}$;
- the next estimation, $x^{(2)}$, is used together with $x^{(1)}$ to find the following approximation, and so on...

Univariate Problems: Secant method



(Givens & Hoeting, 2013, figure 2.5)

Univariate Problems: Secant method

The secant method:

- converges under similar assumptions to Newton's method;
- has order of convergence: $(1 + \sqrt{5})/2 \approx 1.62$.

To prove it, consider an interval $[a, b] : x^{(0)}, x^{(1)}, x^* \in [a, b]$ suitably small, with $g''(x) \neq 0$ and $g'''(x) \neq 0$. Let $\epsilon^{(t)} = x^{(t)} - x^*$. Then, from

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

one can find in a few steps

$$\epsilon^{(t+1)} = \left[\epsilon^{(t)} \epsilon^{(t-1)} \right] \underbrace{\left[\frac{x^{(t)} - x^{(t-1)}}{g'(x^{(t)}) - g'(x^{(t-1)})} \right]}_{A^{(t)}} \underbrace{\left[\frac{\frac{g'(x^{(t)}) - g'(x^{(t-1)})}{\epsilon^{(t)}} - \frac{g'(x^{(t-1)})}{\epsilon^{(t-1)}}}{x^{(t)} - x^{(t-1)}} \right]}_{B^{(t)}}.$$

Univariate Problems: Secant method

Then:

- $A^{(t)} \rightarrow 1/g''(x^*)$ as $x^{(t)} \rightarrow x^*$ (g'' is continuous);
- $B^{(t)} \approx g'''(x^*) \frac{\epsilon^{(t)} - \epsilon^{(t-1)}}{2(x^{(t)} - x^{(t-1)})} \rightarrow g'''(x^*)/2$ as $x^{(t)} \rightarrow x^*$.

Therefore

$$\epsilon^{(t+1)} \approx d^{(t)} \epsilon^{(t)} \epsilon^{(t-1)}$$

with

$$d^{(t)} \rightarrow \frac{g'''(x^*)}{2g''(x^*)} = d \text{ as } t \rightarrow \infty.$$

Univariate Problems: Secant method

Remember that finding the order of convergence means to find β such that

$$\lim_{t \rightarrow \infty} \frac{\epsilon^{(t+1)}}{|\epsilon^{(t)}|^\beta} = c$$

where c is a constant.

If this holds ($\epsilon^{(t+1)} \approx c|\epsilon^{(t)}|^\beta$), the equation $\epsilon^{(t+1)} \approx d^{(t)}\epsilon^{(t)}\epsilon^{(t-1)}$ becomes

$$c|\epsilon^{(t)}|^\beta \approx d^{(t)}|\epsilon^{(t)}||\epsilon^{(t-1)}|^{1/\beta}c^{(-1/\beta)}$$

i.e.,

$$\lim_{t \rightarrow \infty} |\epsilon^{(t)}|^{1-\beta+1/\beta} = \frac{c^{(1+1/\beta)}}{d}$$

As $c^{(1+1/\beta)}/d$ is a positive constant, we need $1 - \beta + 1/\beta = 0$, i.e. $\beta = (1 + \sqrt{5})/2 \approx 1.62$.

Univariate Problems: Fixed-point method

The fixed-point method: instead of finding a root of g' , find a fixed point of a function G such that $g'(x) = 0 \Leftrightarrow G(x) = x$

- a fixed point is a point where $f(x) = x$;
- then $x^{(t+1)} = G(x^{(t)})$;
- e.g., the obvious choice $G(x) = g'(x) + x$ leads to

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

Univariate Problems: Secant method

The secant method in algorithmic form:

- Set $t = 1$, and find two approximate solution $x^{(0)}$ and $x^{(1)}$;
- Set $x^{(t+1)} = x^{(t)} - g'(x^{(t)}) \frac{x^{(t)} - x^{(t-1)}}{g'(x^{(t)}) - g'(x^{(t-1)})}$
- if $|x^{(t+1)} - x^{(t)}| < \epsilon$, return $x^{(t+1)}$, otherwise, set $t = t + 1$ and go back to step 1.

Note:

- similar to the Newton's method, including conditions for convergence;
- it may be robustified by forcing the solution within an interval ("regula falsi method" or "Illinois method"), at the cost of reduced convergence rate.

Univariate Problems: Fixed-point method

The convergence depends on G . In particular if G satisfies

- $G(x) \in [a, b]$ if $x \in [a, b]$;
- $|G(x_1) - G(x_2)| \leq \lambda|x_1 - x_2|$, $\forall x_1, x_2 \in [a, b]$, $\lambda \in [0, 1)$;
- the function is Lipschitz continuous;
- λ is called the Lipschitz constant.

Then:

- $\exists!$ fixed point x^* in $[a, b]$;
- the algorithm converges to x^* from any $x^{(0)} \in [a, b]$;
- $|x^{(t)} - x^*| \leq \frac{\lambda^t}{1-\lambda}|x^{(1)} - x^{(0)}|$.

The (order of) convergence depends on λ :

- the Lipschitz condition holds if $|G'(x)| \leq \lambda < 1$, $\forall x \in [a, b]$;
- for $G(x) = g'(x) + x$ it means $|g''(x) + 1| < 1$ on $[a, b]$.

Univariate Problems: Fixed-point method

Rescaling:

- if g'' is bounded and does not change sign on $[a, b]$, we can rescale, $G(x) = \alpha g'(x) + x$, for $\alpha \neq 0$:
 - $\alpha g'(x) = 0 \Leftrightarrow g'(x) = 0$;
 - $\alpha : |\alpha g''(x) + 1| < 1$ on an interval including $x^{(0)}$;
- one can calculate a suitable α ;
- the book suggest a "try and fail" approach;
- the effectiveness of method depends on the choice of G ;
- For example, for $g'(x) = x + \log x$:
 - with $G(x) = (x + e^{-x})/2$ converges quickly;
 - with $G(x) = e^{-x}$ converges more slowly;
 - with $G(x) = -\log x$ fails to converge.

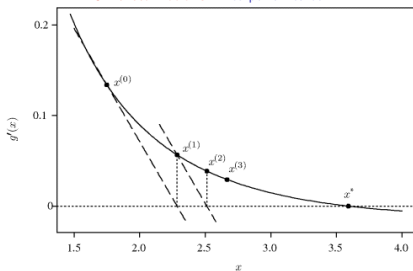
Note:

- both Newton's method and the secant method are special cases of fixed-point iteration;
- in general, all Newton-type methods.

Example:

- $g(x) = \log x/(1+x)$;
- $G(x) = g'(x) + x$;
- scaling with $\alpha = 4$.

Univariate Problems: Fixed-point method



(Givens & Hoeting, 2013, figure 2.6)

Multivariate Problems: Introduction

Multivariate optimization means to find the minimum (maximum) of a function g of a p -dimensional vector $\mathbf{x} = (x_1, \dots, x_p)^T$:

- nothing different from the univariate case, just extensions;
- still iterative methods;
- corresponds to find the solution of a system of equations;
- similar idea for the convergence criterion. Using a distance, e.g., $D(u, v) = \sum_{i=1}^p |u_i - v_i|$ or $D(u, v) = \sqrt{\sum_{i=1}^p (u_i - v_i)^2}$; we can define the absolute

$$D(x^{(t+1)}, x^{(t)}) < \epsilon$$

and relative convergence criteria

$$\frac{D(x^{(t+1)}, x^{(t)})}{D(x^{(t+1)}, 0)} < \epsilon \text{ or } \frac{D(x^{(t+1)}, x^{(t)})}{D(x^{(t+1)}, 0) + \epsilon} < \epsilon.$$

Multivariate Problems: Newton's Method and Fisher Scoring

Newton's method

As for the univariate case, from the first order Taylor expansion,

$$0 = g'(\mathbf{x}^*) \approx g'(\mathbf{x}^{(t)}) + (\mathbf{x}^* - \mathbf{x}^{(t)})^T g''(\mathbf{x}^{(t)})(\mathbf{x}^* - \mathbf{x}^{(t)})$$

we obtain

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

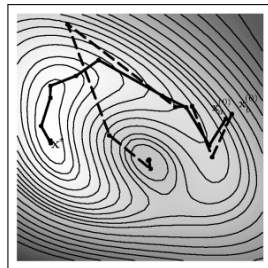
Fisher scoring

Similarly, substituting the second derivative with its expected value in a MLE problem,

$$\theta^{(t+1)} = \theta^{(t)} + I(\mathbf{x}\theta^{(t)})^{-1} \ell_{\theta}(\theta^{(t)}),$$

which is asymptotically equivalent to the Newton's method.

Multivariate Problems: Newton's Method and Fisher Scoring



(Givens & Hoeting, 2013, figure 2.7)

Multivariate Problems: Iteratively Re-weighted Least Square

Consider finding the MLE of a GLM (Generalised Linear Model):

- density from an exponential family

$$f(y|\theta) = \exp\{y\theta - b(\theta)\} / \phi + c(y, \phi),$$

where:

- θ is the canonical parameter;
- ϕ is the dispersion parameter;
- $E[Y] = b'(\theta)$ and $Var[Y] = b''(\theta)\phi$.

- given a vector of covariates $\mathbf{z}_i$ for Y_i , $i = 1, \dots, n$,

$$g(E[Y_i | \mathbf{z}_i]) = \mathbf{z}_i^T \beta,$$

where:

- g is the link function;
- β is a vector of unknown parameters.

Multivariate Problems: Iteratively Re-weighted Least Square

In particular, consider the logistic regression (Bernulli distribution),

- $Y_i | \mathbf{z}_i \sim \text{Bernulli}(\pi_i)$, i.i.d. $i = 1, \dots, n$;
- $\theta_i = \log(\pi_i / (1 - \pi_i))$;
- $b(\theta_i) = \log(1 + \exp\{\theta_i\}) = \log(1 + \mathbf{z}_i^T \beta)$;
- $\phi = 1$.

The log-likelihood function is

$$\ell(\beta) = \mathbf{y}^T \mathbf{Z} \beta - \mathbf{1}^T \mathbf{1}$$

where: $\mathbf{1} = (1 \dots 1)^T$, $\mathbf{y} = (y_1 \dots y_n)^T$, $\mathbf{b} = (b(\mathbf{z}_1 \beta) \dots b(\mathbf{z}_n \beta))^T$, and $\mathbf{Z}$ is the $n \times 2$ matrix with i -th row $(1 \ \mathbf{z}_i)$.

Multivariate Problems: Iteratively Re-weighted Least Square

To implement the Newton's method, we need the score function

$$\ell_{\beta}(\beta) = \mathbf{Z}^T (\mathbf{y} - \exp\{\mathbf{Z}^T \beta / (1 + \mathbf{Z}^T \beta)\}) = \mathbf{Z}^T (\mathbf{y} - \pi)$$

and the observed Fisher information

$$\ell_{\beta\beta}(\beta) = -\mathbf{Z}^T \mathbf{W} \mathbf{Z}$$

where $\mathbf{W} = (\pi_1(1 - \pi_1) \dots \pi_n(1 - \pi_n))\mathbf{I}$.

It result in the step

$$\begin{aligned} \beta^{(t+1)} &= \beta^{(t)} - \ell_{\beta\beta}(\beta^{(t)})^{-1} \ell_{\beta}(\beta^{(t)}) \\ &= \beta^{(t)} + (\mathbf{Z}^T \mathbf{W}^{(t)} \mathbf{Z})^{-1} \mathbf{Z}^T (\mathbf{y} - \pi). \end{aligned}$$

As $\mathbf{I}(\beta) = E[\mathbf{Z}^T \mathbf{W} \mathbf{Z}] = \mathbf{Z}^T \mathbf{W} \mathbf{Z} = -\ell_{\beta\beta}(\beta)$, there is no difference between Newton's method and Fisher scoring.

Multivariate Problems: Human Face Recognition

Consider the example on face recognition (Zhao et al., 1998):

- $n = 3215$ (1072 training, 2143 test);
- response $Y \in \{0, 1\}$, with 1 meaning success;
- covariate z : absolute difference in eye region difference;
- the model is that of the previous slide.

Using Newton's method:

(t)	0	1	2
$\beta^{(t)}$	$\begin{pmatrix} 0.95913 \\ 0.00000 \end{pmatrix}$	$\begin{pmatrix} 1.70694 \\ -14.20059 \end{pmatrix}$	$\begin{pmatrix} 1.73725 \\ -13.56988 \end{pmatrix}$
(t)	3	4	
$\beta^{(t)}$	$\begin{pmatrix} 1.73874 \\ -13.58839 \end{pmatrix}$	$\begin{pmatrix} 1.73874 \\ -13.58840 \end{pmatrix}$	

Multivariate Problems: Iteratively Re-weighted Least Square

Defining $\mathbf{e}^{(t)} = \mathbf{y} - \pi$, the step can be written as

$$\begin{aligned} \beta^{(t+1)} &= \beta^{(t)} + (\mathbf{Z}^T \mathbf{W}^{(t)} \mathbf{Z})^{-1} (\mathbf{Z}^T \mathbf{e}^{(t)}) \\ &= (\mathbf{Z}^T \mathbf{W}^{(t)} \mathbf{Z})^{-1} \left[(\mathbf{Z}^T \mathbf{W}^{(t)} \mathbf{Z}) \beta^{(t)} + \mathbf{Z}^T \mathbf{W}^{(t)} (\mathbf{W}^{(t)})^{-1} \mathbf{e}^{(t)} \right] \\ &= (\mathbf{Z}^T \mathbf{W}^{(t)} \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{W}^{(t)} \mathbf{x}^{(t)} \end{aligned}$$

with $\mathbf{x}^{(t)} = \mathbf{Z} \beta^{(t)} + (\mathbf{W}^{(t)})^{-1} \mathbf{e}^{(t)}$ being the working response.

Note:

- $\beta^{(t+1)}$ are coefficients computed from the weighted least squares regression of $\mathbf{x}^{(t)}$ on $\mathbf{Z}$;
- $\mathbf{x}$'s are computed at each steps and used to update the β 's.

References |

- GENTLE, J. E. (2009). *Computational Statistics*. Springer.
- GIVENS, G. H. & HOETING, J. A. (2013). *Computational Statistics*. Wiley.
- ZHAO, W., KRISHNASWAMY, A., CHELLAPPA, R., SWETS, D. & WENG, J. (1998). Discriminant analysis of principal components for face recognition. In *Face Recognition: from theory to applications*.