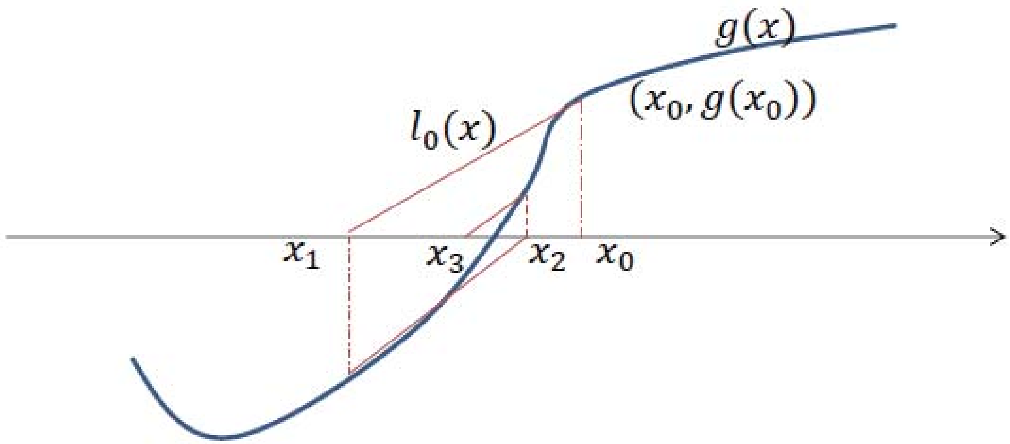
STAT3006: TUTORIAL2

1. Newton's method.
2. Expectation Maximization (EM) algorithm.

NEWTON'S METHOD

- Also called Newton-Raphson's method.
- Used to iteratively approximate zero points of the equation $\mathrm{g}(x) = 0$, where g(x) must be differentiable.



- $x_{n+1} = x_n - \frac{g(x_n)}{g'(x_n)}$

NEWTON'S METHOD

- We have a sequence (Newton sequence) of $\{x_n\}^{\infty}{}_{n=1}$ from the Newton's method.
  - From the lecture note2, Newton sequence is quadratic convergence.
  - Quadratic convergence implies that $\lim \frac{|x_{n+1} - x_n|}{|x_n - x_\infty|} = 1$, so we can use $|x_{n+1} - x_n| < \varepsilon$ as stopping rule.

- Usually, we would like to maximize f(x) instead of searching zero points of g(x).
  - In some cases (e.g. f is convex), maximizing f(x) is equivalent to searching zero points of f'(x).
  - Let g(x) be f'(x).
  - $x_{n+1} = x_n - \frac{f'(x_n)}{f''(x_n)}$.
  - Requirement for f(x), f must be twice differentiable, and calculating the $\left(f''(x_n)\right)^{-1}$ is computationally feasible.

NEWTON'S METHOD

- In the multivariate case, (maximize $f(x_1, x_2, \ldots, x_p)$, the range of f is in R).
- $\boldsymbol{x}^{(n+1)} = \boldsymbol{x}^{(n)} - \left(Hf\left(\boldsymbol{x}^{(n)}\right)\right)^{-1} Jf\left(\boldsymbol{x}^{(n)}\right)$
- $Hf\left(\boldsymbol{x}^{(n)}\right)$ is the Hessian matrix of $f(x_1^{(n)}, x_2^{(n)}, \ldots, x_p^{(n)})$.
- $Jf\left(\boldsymbol{x}^{(n)}\right)$ is the Jacobian vector of $f(x_1^{(n)}, x_2^{(n)}, \ldots, x_p^{(n)})$.

- Application: we maximize log likelihood function to obtain MLE, where f=logL.
  - In this case, $Jf\left(\boldsymbol{x}^{(n)}\right)$ is called the score function.
  - $-Hf\left(\boldsymbol{x}^{(n)}\right)$ is called the observed information matrix.

NEWTON'S METHOD

- Example (Weibull distribution): $p(x) = \boldsymbol{\alpha\beta} x^{\boldsymbol{\beta}-1} e^{-\boldsymbol{\alpha} x^{\boldsymbol{\beta}}}$ $(x > 0, \boldsymbol{\alpha} > 0, \boldsymbol{\beta} > 0)$.
- logL = $n log\boldsymbol{\alpha} + n log\boldsymbol{\beta} + (\boldsymbol{\beta} - 1)\sum log X_i - \boldsymbol{\alpha}\sum \mathrm{X}_{\mathrm{i}}^{\boldsymbol{\beta}}$
- Calculate the score function.
- Calculate the observed Information matrix.

EM ALGORITHM

- When we use Newton's method, we have to calculate the inverse of a matrix.
- It is usually computationally infeasible when the dimension of parameters is high.
- In some problems, the observed data is X, and the missing (unobserved) data is Z.
  - On the one hand, directly maximizing observed-data likelihood $L(\Theta|X)$ is very difficult.
  - On the other hand, complete-data likelihood $L(\Theta|X, Z)$ is more tractable.
- EM algorithm is a very useful tool to maximize $L(\Theta|X)$ by playing with $L(\Theta|X,Z)$.
  - Given current estimates for Θ, $\Theta^{(t)}$
  - E (Expectation) step: calculate the conditional expectation $Q\left(\Theta\middle|\Theta^{(t)}\right) = E[\log L(\Theta|X, Z)|\ X, \Theta^{(t)}]$.
  - M (Maximization) step: $\Theta^{(t+1)} = \text{argmax}\, Q\left(\Theta\middle|\Theta^{(t)}\right)$
  - Why EM algorithm works? It can be shown that $L(\Theta^{(t+1)}\ |\ X) \geq L(\Theta^{(t)}\ |\ X)$.
  - Sometimes, multiple initial values should be tried to avoid falling into a local mode.

EM ALGORITHM

- Problem:
  - There are two coins, coin A and coin B.
  - The probability of coin A's head up is θ_A; The probability of coin B's head up is θ_B.
  - We first randomly select a coin from coin A and coin B, and then toss the selected coin ten times. We repeat the preceding procedure five times.
  - Data:
    - HTTTT HTTHT
    - HHTTH THHTH
    - TTTHT THHTT
    - TTHTH TTTHT
    - THHTH HTHHT

EM ALGORITHM

- When $Z_i = 1$, coin A is selected. When $Z_i = 2$, coin B is selected.
  - $P(Z_i = 1) = P(Z_i = 2) = 1/2$.
- Denote the number of heads up in experiment I by X_i.
- Missing data is Z_i, the observed data is X_i.
- Observed-data likelihood function is too complicated to deal with.
- The complete-data likelihood function is

$$L(\theta_A,\theta_B|\mathbf{X},\mathbf{Z}) = \prod_{i=1}^{5}\left[\binom{10}{X_i}\theta_A^{X_i}(1-\theta_A)^{10-X_i}\right]^{I(Z_i=1)}\cdot\left[\binom{10}{X_i}\theta_B^{X_i}(1-\theta_B)^{10-X_i}\right]^{I(Z_i=2)},$$

EM ALGORITHM

- The log complete-data likelihood function is

$$
\begin{aligned}
l(\theta_A,&\theta_B|\mathbf{X},\mathbf{Z}) \\
&= \sum_{i=1}^{5} I(Z_i=1)\cdot\left[\log\binom{10}{X_i} + X_i\log\theta_A + (10-X_i)\log(1-\theta_A)\right] + \\
&\quad I(Z_i=2)\cdot\left[\log\binom{10}{X_i} + X_i\log\theta_B + (10-X_i)\log(1-\theta_B)\right],
\end{aligned}
$$

- The E step is

$$
\begin{aligned}
E(I(Z_i=1)|X_i,\theta_A^{(t)},\theta_B^{(t)}) &= P(Z_i=1|X_i,\theta_A^{(t)},\theta_B^{(t)}) \\
&= \frac{\binom{10}{X_i}(\theta_A^{(t)})^{X_i}(1-\theta_A^{(t)})^{10-X_i}}{\binom{10}{X_i}(\theta_A^{(t)})^{X_i}(1-\theta_A^{(t)})^{10-X_i}+\binom{10}{X_i}(\theta_B^{(t)})^{X_i}(1-\theta_B^{(t)})^{10-X_i}}, \\
E(I(Z_i=2)|X_i,\theta_A^{(t)},\theta_B^{(t)}) &= \frac{\binom{10}{X_i}(\theta_B^{(t)})^{X_i}(1-\theta_B^{(t)})^{10-X_i}}{\binom{10}{X_i}(\theta_A^{(t)})^{X_i}(1-\theta_A^{(t)})^{10-X_i}+\binom{10}{X_i}(\theta_B^{(t)})^{X_i}(1-\theta_B^{(t)})^{10-X_i}}.
\end{aligned}
$$

EM ALGORITHM

- The M step is

$$\theta_A^{(t+1)} = \frac{\sum_{i=1}^{5} E(I(Z_i = 1)|X_i, \theta_A^{(t)}, \theta_B^{(t)}) \cdot X_i}{\sum_{i=1}^{5} E(I(Z_i = 1)|X_i, \theta_A^{(t)}, \theta_B^{(t)}) \cdot 10}$$

$$\theta_B^{(t+1)} = \frac{\sum_{i=1}^{5} E(I(Z_i = 2)|X_i, \theta_A^{(t)}, \theta_B^{(t)}) \cdot X_i}{\sum_{i=1}^{5} E(I(Z_i = 2)|X_i, \theta_A^{(t)}, \theta_B^{(t)}) \cdot 10}$$