# Missing Data and EM

## MAST90083 Computational Statistics and Data Mining

Karim Seghouane

School of Mathematics & Statistics
The University of Melbourne

# Outline

§5.1 Introduction

§5.2 Motivation

§5.3 Expectation-Maximization

§5.4 Derivation of the EM

§5.5 Newton-Raphson and Fisher Scoring

# Introduction

- Assume a set of observations $\mathbf{y} = \{y_1, \ldots, y_N\}$ representing i.i.d. samples from a random variable $y$
- We aim to model this data set by specifying a parametric probability density model

$$y \sim g\left(y; \boldsymbol{\theta}\right)$$

- The vector $\boldsymbol{\theta}$ represents one or more unknown parameters that governs the distribution of the random variable $y$

# Example

If we assume that $y$ has a normal distribution with mean $\mu$ and variance $\sigma^2$ than

$$\boldsymbol{\theta} = \left(\mu, \sigma^2\right)$$

and

$$g\left(y; \boldsymbol{\theta}\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

Given the sample **y**, we aim to find the parameter vector that is most likely the "true" parameter vector of the DGP that generated the sample set **y**

# Maximum Likelihood

- The probability density function of the set of observations under the model $g\left(y;\boldsymbol{\theta}\right)$ is

$$L\left(\mathbf{y};\boldsymbol{\theta}\right) = g\left(\mathbf{y};\boldsymbol{\theta}\right) = g\left(y_1,...,y_N;\boldsymbol{\theta}\right) = \prod_{i=1}^{N} g\left(y_i;\boldsymbol{\theta}\right)$$

- $L\left(\mathbf{y};\boldsymbol{\theta}\right)$ defines the likelihood function. It is a function of the $\boldsymbol{\theta}$ (unknown) with the set of observations $\mathbf{y} = \{y_1,\ldots,y_N\}$ fixed.
- The maximum likelihood method is the most popular technique of parameter estimation. It consists in finding the most likely estimate $\hat{\boldsymbol{\theta}}$ by maximizing $L\left(\mathbf{y};\boldsymbol{\theta}\right)$

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} L\left(\mathbf{y};\boldsymbol{\theta}\right)$$

# Maximum Likelihood

- The log-likelihood corresponds to the logarithm of $L(\mathbf{y};\boldsymbol{\theta})$

$$\ell(\mathbf{y};\boldsymbol{\theta}) = \sum_{i=1}^{N} \ell(y_i;\boldsymbol{\theta}) = \sum_{i=1}^{N} \log g(y_i;\boldsymbol{\theta})$$

- and $\ell(y_i;\boldsymbol{\theta}) = \log g(y_i;\boldsymbol{\theta})$ is called log-likelihood component
- The maximum likelihood method is generally obtained by maximizing $\ell(\mathbf{y};\boldsymbol{\theta})$
- The likelihood function is also used to assess the precision of $\hat{\boldsymbol{\theta}}$

# Maximum Likelihood

- The score function is defined by

$$\dot{\ell}\left(\mathbf{y};\boldsymbol{\theta}\right)=\sum_{i=1}^{N}\dot{\ell}\left(y_i;\boldsymbol{\theta}\right)$$

- where

$$\dot{\ell}\left(\mathbf{y};\boldsymbol{\theta}\right)=\frac{\partial\ell\left(\mathbf{y};\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}}$$

- At the maximum in the parameter space

$$\dot{\ell}\left(\mathbf{y};\boldsymbol{\theta}\right)=0$$

# Properties of Maximum Likelihood

- The information matrix is defined by

$$I(\boldsymbol{\theta}) = -\sum_{i=1}^{N} \frac{\partial^2 \ell(y_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}$$

- $I(\boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ is the observed information and $\mathrm{Var}\left(\hat{\boldsymbol{\theta}}\right) = I\left(\hat{\boldsymbol{\theta}}\right)^{-1}$
- The Fisher information or expected information is

$$i(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\left[I(\boldsymbol{\theta})\right]$$

- Assume $\boldsymbol{\theta}_0$ denotes the trues value of $\boldsymbol{\theta}$

# Properties of Maximum Likelihood

- The sampling distribution of the maximum likelihood estimator is a normal distribution

$$\hat{\boldsymbol{\theta}} \to N\left(\boldsymbol{\theta}_0, i\left(\boldsymbol{\theta}_0\right)^{-1}\right)$$

- The samples are independently obtained from $g\left(y, \boldsymbol{\theta}_0\right)$
- This suggests that the sampling distribution of $\hat{\boldsymbol{\theta}}$ may be approximated by $N\left(\hat{\boldsymbol{\theta}}, I\left(\hat{\boldsymbol{\theta}}\right)^{-1}\right)$
- The corresponding estimates for the standard errors of $\hat{\theta}_j$ are obtained from $\sqrt{I\left(\hat{\boldsymbol{\theta}}\right)^{-1}_{jj}}$

# Local likelihood

- Any parametric model can be made local if the fitting method accommodates observation weights
- Local likelihood allows a relation from a globally parametric model to one that is local

$$\ell(\mathbf{z}, \boldsymbol{\theta}(z_0)) = \sum_{i=1}^{N} K_h(z_0, z_i)\ell(z_i, \boldsymbol{\theta}(z_0))$$

- For example $\ell(\mathbf{z}, \boldsymbol{\theta}) = \left(y - \mathbf{x}^\top \boldsymbol{\theta}\right)^2$. This fits a linear varying coefficient model $\boldsymbol{\theta}(z)$ by maximizing the local likelihood

# Likelihood and Kullback-Leibler Divergence

- The maximum likelihood estimate is obtained by

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ell\left(\mathbf{y}, \boldsymbol{\theta}\right) = \arg\max_{\boldsymbol{\theta}} \frac{1}{N} \sum_{i=1}^{N} \log g\left(y_i, \boldsymbol{\theta}\right)$$

- Using the empirical density

$$g_N\left(\mathbf{y}\right) = \frac{1}{N} \sum_{i=1}^{N} \delta\left(y - y_i\right)$$

- which puts a mass $1/N$ at $y_i'$s we have

$$\frac{1}{N} \sum_{i=1}^{N} \log g\left(y_i, \boldsymbol{\theta}\right) = \int \log g\left(y, \boldsymbol{\theta}\right) g_N\left(\mathbf{y}\right) dy$$

# Likelihood and Kullback-Leibler Divergence

- We have

$$\int \log g\left(y, \boldsymbol{\theta}\right) g_N\left(\mathbf{y}\right) dy = \int \log g\left(y, \boldsymbol{\theta}\right) dG_N\left(\mathbf{y}\right)$$

- The maximization can be replaced by

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left[ \int \log g\left(y, \boldsymbol{\theta}_0\right) dG_N\left(\mathbf{y}\right) - \int \log g\left(y, \boldsymbol{\theta}\right) dG_N\left(\mathbf{y}\right) \right]$$

# Likelihood and Kullback-Leibler Divergence

- Using the law of large numbers

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left[ \int \log g\left(y, \boldsymbol{\theta}_0\right) dG\left(\mathbf{y}, \boldsymbol{\theta}_0\right) - \int \log g\left(y, \boldsymbol{\theta}\right) dG\left(\mathbf{y}, \boldsymbol{\theta}_0\right) \right]$$

# Likelihood and Kullback-Leibler Divergence

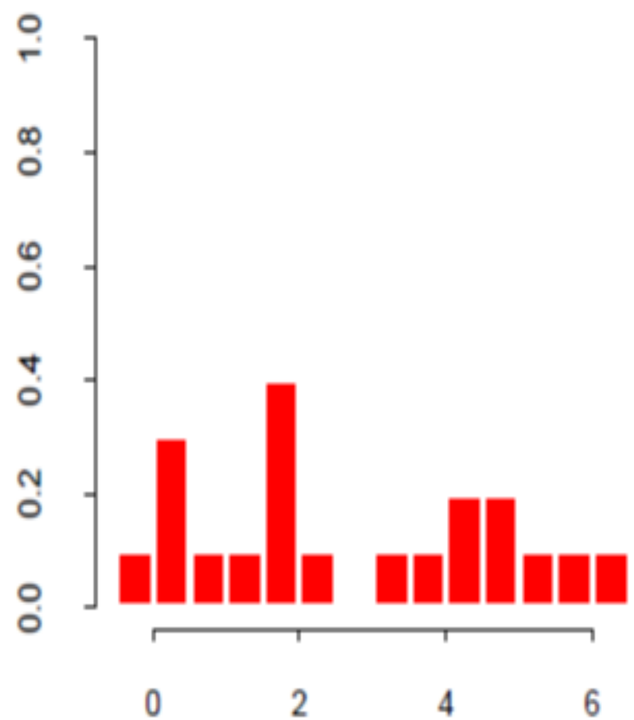
- and we have

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \mathrm{KL}\left[g\left(y, \boldsymbol{\theta}_0\right), g\left(y, \boldsymbol{\theta}\right)\right]$$

- Therefore the ML estimate is also the one that minimizes the KLD between a family of parametrized distributions and the true distribution.

# Motivation

Simple mixture model for density estimation using maximum likelihood




A Gaussian density would not be appropriate → because there are two regimes

# Motivation

We model $y$ as a mixture of two model densities

$$y_1 \sim N\left(\mu_1, \sigma_1^2\right) \quad y_2 \sim N\left(\mu_2, \sigma_2^2\right)$$

$$y \sim (1-z)y_1 + zy_2$$

where $z \in \{0,1\}$ with $p(z=1)=\pi$ is the mixing coefficient

# Motivation

The generative representation can be seen as

- Generate a $z \in \{0, 1\}$ with probability $\pi$
- Depending on the outcome deliver $y_1$ or $y_2$

Let $\phi(y)$ denote the normal density with parameters $\boldsymbol{\theta} = (\mu, \sigma^2)$. Then the density of $y$ is

$$p(y) = (1 - \pi)\phi_{\boldsymbol{\theta}_1}(y) + \pi\phi_{\boldsymbol{\theta}_2}(y) = \sum_{i=1}^{2} m_i \phi_{\boldsymbol{\theta}_i}(y)$$

where $m_1 = 1 - \pi$, $m_2 = \pi$ and $\sum_{i=1}^{2} m_i = 1$.

# Motivation

Now suppose we are given a data set of size $N$ and we want to fit this model using maximum likelihood to estimate

$$\boldsymbol{\theta} = (\pi, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = (\pi, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$$

The likelihood is

$$\ell(\mathbf{y}, \boldsymbol{\theta}) = \sum_{i=1}^{N} \log\left[(1-\pi)\phi_{\boldsymbol{\theta}_1}(y_i) + \pi\phi_{\boldsymbol{\theta}_2}(y_i)\right]$$

Direct work with $\ell(\mathbf{y}, \boldsymbol{\theta})$ is difficult instead we make use of $z$

# Illustration

- Suppose one of the component mixture say $\phi_{\boldsymbol{\theta}_2}$ has its mean $\mu_2$ exactly equal to one of the observation so that $\mu_2 = y_i$
- This data point will contribute a term in the likelihood function of the form $\frac{1}{\sqrt{2\pi}\sigma_2}$
- If we consider the limit $\sigma_2 \to 0$, then this term goes to infinity and as is the likelihood function
- Thus maximizing of the log-likelihood function is not a well posed problem because such singularities will always be present when one Gaussian is identified to an observation

# Motivation

- It is preferable to work with the joint density $p(y, z)$
- The marginal density of $z$ is specified in terms of the mixing coefficient $\pi$, $p(z = 1) = \pi$ and

$$p(z) = \pi^z (1 - \pi)^{1-z}$$

- Similarly, the conditional

$$p(y/z = 1) = \phi_{\boldsymbol{\theta}_2}(y)$$

- which can also be written as

$$p(y/z) = \phi_{\boldsymbol{\theta}_2}(y)^z \phi_{\boldsymbol{\theta}_1}(y)^{1-z}$$

# Motivation

- The joint density is given by $p(y/z)p(z)$ and the marginal of $y$ is obtained by summing the joint density over all possible states of $z$ to give

$$p(y) = \sum_z p(z)p(y/z) = \pi\phi_{\boldsymbol{\theta}_2}(y) + (1-\pi)\phi_{\boldsymbol{\theta}_1}(y)$$

- Thus the marginal density of $y$ is the Gaussian mixture
- If we have several observations $\mathbf{y} = \{y_1, \ldots, y_N\}$ and because we have represented the marginal distribution in the form $p(y) = \sum_z p(y, z)$, it follows that for every observed data $y_i$ there is a corresponding $z_i$
- Therefore there is an equivalent formulation of the Gaussian mixture involving an explicit latent variable.

# Justification for the EM

- We are now able to work with the joint density $p(y, z)$ instead of the marginal $p(y)$
- This leads to the introduction of the expectation maximization (EM) algorithm
- Another important quantity is the conditional density of $z$ given $y$.
- We use $\gamma(z)$ to denote $p(z = 1/y)$

# Justification for the EM

- The value of this conditional can be obtained using Bayes theorem

$p(z=1|y)=\gamma(z)$

$$\gamma(z) = p(z=1/y) = \frac{p(z=1)p(y/z=1)}{p(y)} =$$

$$\frac{p(z=1)p(y/z=1)}{\pi\phi_{\boldsymbol{\theta}_2}(y) + (1-\pi)\phi_{\boldsymbol{\theta}_1}(y)}$$

- $\pi$ is the probability of $z=1$ while $\gamma(z)$ is the corresponding probability once we have observed $y$
- $\gamma(z)$ can be seen as the responsibility that $\phi_{\boldsymbol{\theta}_2}$ takes for explaining the observation $y$

# Justification for the EM

For the two component mixture, start with initial estimates for the parameters (part 1)

- Expectation step:

$\gamma_1$''

$\gamma_2$

$$\gamma_i = \frac{\hat{\pi}\phi_{\hat{\boldsymbol{\theta}}_2}(y_i)}{\hat{\pi}\phi_{\hat{\boldsymbol{\theta}}_2}(y_i) + (1-\hat{\pi})\phi_{\hat{\boldsymbol{\theta}}_1}(y_i)} \qquad i = 1, \ldots, N$$

- Maximization step:

$$\hat{\mu}_1 = \frac{\sum_{i=1}^N (1-\hat{\gamma}_i) y_i}{\sum_{i=1}^N (1-\hat{\gamma}_i)} \qquad \hat{\sigma}_1^2 = \frac{\sum_{i=1}^N (1-\hat{\gamma}_i)(y_i - \hat{\mu}_1)^2}{\sum_{i=1}^N (1-\hat{\gamma}_i)}$$

$$\hat{\mu}_2 = \frac{\sum_{i=1}^N \hat{\gamma}_i y_i}{\sum_{i=1}^N \hat{\gamma}_i} \qquad \hat{\sigma}_2^2 = \frac{\sum_{i=1}^N \hat{\gamma}_i (y_i - \hat{\mu}_2)^2}{\sum_{i=1}^N \hat{\gamma}_i}$$

# Justification for the EM

For the two component mixture, start with initial estimates for the parameters (part 2)

- Maximization step:

$$\hat{\pi} = \sum_{i=1}^{N} \hat{\gamma}_i / N$$

- Iterate these two steps until convergence

$$\ell(\mathbf{y}, \mathbf{z}, \boldsymbol{\theta}) = \sum_{i=1}^{N} [(1 - z_i) \log \phi_{\boldsymbol{\theta}_1}(y_i) + z_i \log \phi_{\boldsymbol{\theta}_2}(y_i)]$$
$$+ \sum_{i=1}^{N} [(1 - z_i) \log(1 - \pi) + z_i \log \pi]$$

# When to use the EM

The EM algorithm is very useful when:

- We have missing values due to the observation process, including unknown clusters.
- Assuming hidden (latent) parameter for problem simplification.

# EM - how does it work?

- We observed $\mathbf{Y}$ (**observed data**) with the pdf $g(\mathbf{y}|\theta)$.
  $\mathbf{Y}$ can be either a number, a vector, a matrix, or of a more general form.
- We assume that some hidden parameter $\mathbf{Z}$ exists.
  Let $(\mathbf{Y}, \mathbf{Z})$ be the **complete data** having the pdf $f(\mathbf{y}, \mathbf{z}|\theta)$.
- We also assume/specify the joint pdf of $(\mathbf{Y}, \mathbf{Z})$

$$f(\mathbf{Y}, \mathbf{Z}|\theta)$$

# Reminder

- The objective is to maximize $\ln g(\mathbf{y}|\theta)$ w.r.t. $\theta$ in order to find the MLE of $\theta$
- If it is easy to do maximization of $\ln g(\mathbf{y}|\theta)$ directly, then there is no need to use the EM.
- So suppose maximizing $\ln g(\mathbf{y}|\theta)$ is difficult but maximizing $\ln f(\mathbf{y}, \mathbf{z}|\theta)$ is relatively easy provided that $(\mathbf{Y}, \mathbf{Z})$ were completely observed.

# EM - how does it work?

Now let's define a new likelihood function

$$\mathcal{L}(\theta|\mathbf{y},\mathbf{z}) = f(\mathbf{y},\mathbf{z}|\theta)$$

We call it the **complete-data likelihood**
What is constant and what is random in this function?

- $\mathbf{y}$ - the set of the observed data, known and fixed
- $\theta$ - parameter/s of the DGP, fixed but unknown
- $\mathbf{z}$ - latent variables ,unknown random variables

Therefore, $\mathcal{L}(\theta|\mathbf{y},\mathbf{z}) = h_{\theta,\mathbf{y}}(\mathbf{z})$

We need a tool to solve the optimization of the complete-data likelihood.

# EM - how does it work?

The expected value maximizes the likelihood!

So, what is the expected value of the complete-data likelihood?

$$(1)\;\; Q(\theta, \theta^{i-1}) \;=\; E[\ln f(\mathbf{y}, \mathbf{z}|\theta) \,|\, \mathbf{Y}, \theta^{i-1}]$$

Recall, the expectation of conditional density is
$E[h(y)|X = x] = \int_y h(y)\dot{f}(y|x)\, dy$

$$(2)\;\; E[\ln f(\mathbf{y}, \mathbf{z}|\theta) \,|\, \mathbf{Y}, \theta^{i-1}] \;=\; \int_y \ln f(\mathbf{y}, \mathbf{z}|\theta)\dot{k}(\mathbf{z}|\mathbf{y}, \theta^{i-1})\, d\mathbf{z}$$

Note, $k(\mathbf{z}|\mathbf{y}, \theta^{i-1})$ is a conditional distribution of the unobserved data. It depends on the current value of $\theta^{i-1}$ & on the observed data $\mathbf{y}$.

# EM - how does it work?

For a given value of $\theta^{i-1}$ & the observed dataset $\mathbf{y}$ we can evaluate $Q(\theta, \theta^{i-1})$.

This is **the** $1^{st}$ **step of the EM algorithm** - the **E-step** .

But, $Q(\theta, \theta^{i-1})$ will remain a function of $\theta$!

Now we can maximize it with respect to $\theta$

$$\theta^{i} = \underset{\theta}{\operatorname{argmax}}\, Q(\theta, \theta^{i-1}) \,.$$

and update $\theta^{i-1}$ by $\theta^{i}$.

This is **the** $2^{nd}$ **step of the EM algorithm** - the **M-step** .

# EM - how does it work?

In the EM algorithm the two steps are repeated many times.

**Each iteration is guaranteed to increase the $\log \mathcal{L}$. Moreover, EM algorithm is guaranteed to increase the observed-data $log - liklihood$.**

Why?

# Deriving the EM

**EM algorithm:**

- Start from an appropriate initial value $\theta^{(0)}$.
- Given the current iterate $\theta^{(r)}$, $r = 1, 2, \cdots$,

E-step Compute $Q(\theta, \theta^{(r)}) = E\left[\ln f(\mathbf{y}, \mathbf{z}|\theta)|\mathbf{y}, \theta^{(r)}\right]$.

M-step Maximize $Q(\theta, \theta^{(r)})$ as a function of $\theta$ to obtain $\theta^{(r+1)}$.

- The iteration continues until $||\theta^{(r+1)} - \theta^{(r)}||$ or $|Q(\theta^{(r+1)}, \theta^{(r)}) - Q(\theta^{(r)}, \theta^{(r)})|$ is smaller than a prescribed $\varepsilon > 0$ (e.g. $\varepsilon = 10^{-6}$).

*Remark*: If the M-step is replaced with

**M'-step**: Find $\theta^{(r+1)}$ so that $\quad Q(\theta^{(r+1)}, \theta^{(r)}) > Q(\theta^{(r)}, \theta^{(r)})$,

the resultant algorithm will be called the **GEM** (generalized EM).

# Properties

Note that the conditional pdf of $\mathbf{Z}$ given $\mathbf{Y} = \mathbf{y}$, $k(\mathbf{z}|\mathbf{y}, \theta)$, can be written as $k(\mathbf{z}|\mathbf{y}, \theta) = \frac{f(\mathbf{y},\mathbf{z}|\theta)}{g(\mathbf{y}|\theta)}$.

Then we have

$$\ln f(\mathbf{y}, \mathbf{z}|\theta) = \ln g(\mathbf{y}|\theta) + \ln k(\mathbf{z}|\mathbf{y}, \theta). \tag{1}$$

Namely, complete-data log-likelihood equals the sum of observed-data log-likelihood and conditional log-likelihood.

**Note 1**

# Properties

Taking expectation on both sides of (1) w.r.t. the conditional pdf of $\mathbf{Z}$ given $\mathbf{Y} = \mathbf{y}$ and some value $\theta'$ of $\theta$, we have

$$Q(\theta, \theta') = \ln g(\mathbf{y}|\theta) + H(\theta, \theta'), \quad \text{where} \tag{2}$$

$$\begin{aligned} Q(\theta, \theta') &= E_Z\left[\ln f(\mathbf{y}, \mathbf{Z}|\theta)|\mathbf{y}, \theta'\right] = \int k(\mathbf{z}|\mathbf{y}, \theta') \ln f(\mathbf{y}, \mathbf{z}|\theta) d\mathbf{z}, \quad \text{and} \\ H(\theta, \theta') &= E_Z\left[\ln k(\mathbf{Z}|\mathbf{y}, \theta)|\mathbf{y}, \theta'\right] = \int k(\mathbf{z}|\mathbf{y}, \theta') \ln k(\mathbf{z}|\mathbf{y}, \theta) d\mathbf{z}. \end{aligned}$$

Given a value $\theta'$, if we can find $\theta''$ such that $Q(\theta'', \theta') = \max_\theta Q(\theta, \theta')$ we know

$$\ln g(\mathbf{y}|\theta'') \geq \ln g(\mathbf{y}|\theta').$$

# Deriving the EM

**Theorem (Jensen's inequality)**

$$E[g(X)] \leq g(E[X]) \quad \textit{if } g(\cdot) \textit{ is concave.}$$

By Jensen's inequality,

$$\begin{aligned} E\left(\ln \frac{k(\mathbf{Z}|\mathbf{y},\theta)}{k(\mathbf{Z}|\mathbf{y},\theta')} \mid \mathbf{y},\theta'\right) &\leq \ln E\left(\frac{k(\mathbf{Z}|\mathbf{y},\theta)}{k(\mathbf{Z}|\mathbf{y},\theta')} \mid \mathbf{y},\theta'\right) \\ &= \ln \int \frac{k(\mathbf{z}|\mathbf{y},\theta)}{k(\mathbf{z}|\mathbf{y},\theta')} k(\mathbf{z}|\mathbf{y},\theta')d\mathbf{z} = \ln \int k(\mathbf{z}|\mathbf{y},\theta)d\mathbf{z} = \ln 1 = 0. \end{aligned}$$

$$\Rightarrow \quad E\left[\ln k(\mathbf{Z}|\mathbf{y},\theta)|\mathbf{y},\theta'\right] - E\left[\ln k(\mathbf{Z}|\mathbf{y},\theta')|\mathbf{y},\theta'\right] \leq 0$$

This implies that

$$H(\theta,\theta') \leq H(\theta',\theta'). \tag{3}$$

## Deriving the EM

Given a value $\theta'$, if we can find $\theta''$ such that $Q(\theta'', \theta') = \max_{\theta} Q(\theta, \theta')$, then by (2) and (3) we know

$$\ln g(\mathbf{y}|\theta'') \geq \ln g(\mathbf{y}|\theta').$$

(Note that $Q(\theta'', \theta') \geq Q(\theta', \theta')$ and $H(\theta'', \theta') \leq H(\theta', \theta')$.)

This suggests the following algorithm for calculating the MLE of $\theta$ which maximizes the observed-data log-likelihood $\ln g(\mathbf{y}|\theta)$.

# Deriving the EM

**Theorem (Dempster, Laird and Rubin, 1977)**

*Every EM or GEM algorithm increases the observed-data log-likelihood* $\ln g(\mathbf{y}|\theta)$ *at each iteration, i.e.*

$$\ln g(\mathbf{y}|\theta^{(r+1)}) \geq \ln g(\mathbf{y}|\theta^{(r)}),$$

*with the equality holding iff* $Q(\theta^{(r+1)}, \theta^{(r)}) = Q(\theta^{(r)}, \theta^{(r)})$.

**Proof**:
$$\begin{aligned} \ln g(\mathbf{y}|\theta^{(r+1)}) &= Q(\theta^{(r+1)}, \theta^{(r)}) - H(\theta^{(r+1)}, \theta^{(r)}) \quad \text{and} \\ \ln g(\mathbf{y}|\theta^{(r)}) &= Q(\theta^{(r)}, \theta^{(r)}) - H(\theta^{(r)}, \theta^{(r)}). \end{aligned}$$

Hence $\ln g(\mathbf{y}|\theta^{(r+1)}) \geq \ln g(\mathbf{y}|\theta^{(r)})$ because
$Q(\theta^{(r+1)}, \theta^{(r)}) \geq Q(\theta^{(r)}, \theta^{(r)})$ and $H(\theta^{(r+1)}, \theta^{(r)}) \leq H(\theta^{(r)}, \theta^{(r)})$. $\square$

# Deriving the EM

## Theorem (Wu(1983) and Little and Rubin (1987))

*Suppose a sequence of EM iterates $\theta^{(r)}$ satisfies*

1. $\dfrac{\partial Q(\theta, \theta^{(r)})}{\partial \theta} \Big|_{\theta=\theta^{(r+1)}} = 0;$
2. $\theta^{(r)}$ *converges to some value* $\theta_0$ *as* $r \to \infty$, *and* $k(\mathbf{z}|\mathbf{y}, \theta)$ *is "sufficiently smooth".*

*Then* $\dfrac{\partial \ln g(\mathbf{y}|\theta)}{\partial \theta} \Big|_{\theta=\theta_0} = 0.$

This theorem implies that, if $\theta^{(r)}$ converges, it will converge to a stationary point of $\ln g(\mathbf{y}|\theta)$, which is the MLE if there is only one such stationary point. If there are multiple such stationary points, the EM may not converge to the global maximum.

## Introduction

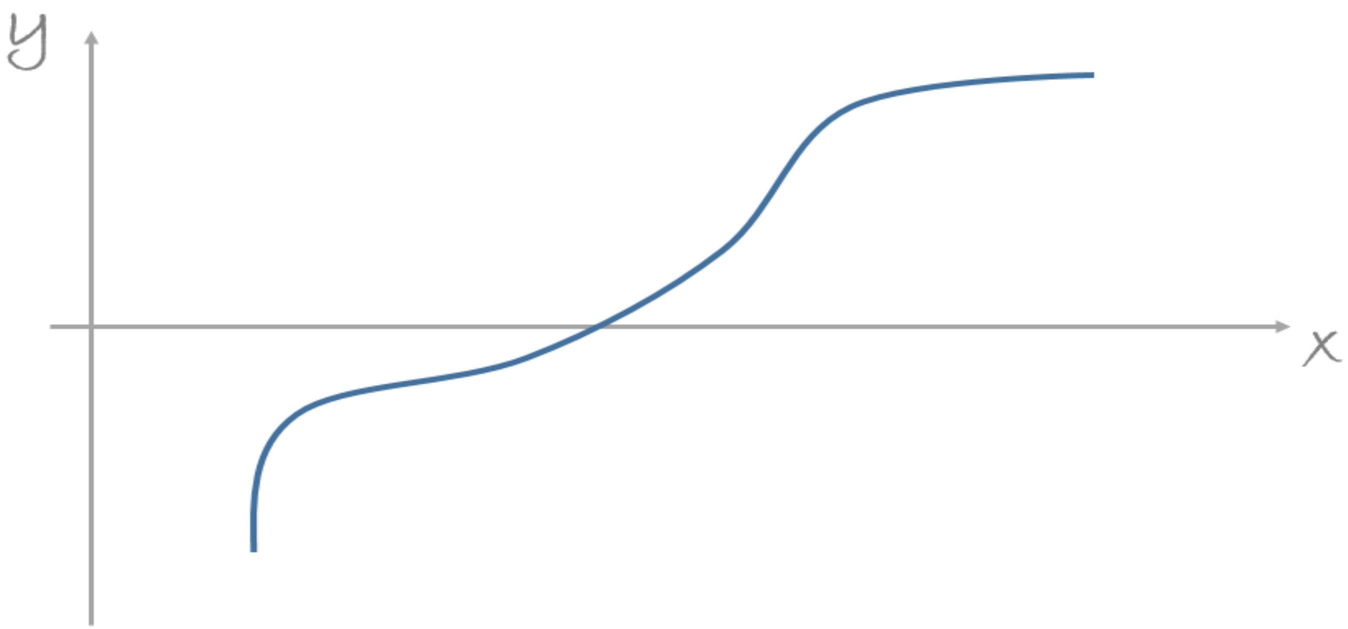
**Newton-Raphson is a method for finding to the roots (or zeroes) of a real-valued function.**

- Newton-Raphson is a more general optimisation algorithm which can also be used to find the MLE.
- NR can be faster than EM but often less stable numerically and less tractable analytically.

# Intuition and graphical explanation

## Graphical explanation in 2D

- The algorithm starts from some guess for the root, $x_0$.
- Then calculate the **tangent line** at this point.
- Compute the **x-intercept** of this tangent line, $x_1$
- Calculate a new tangent line and the new x-intercept.
- Repeat this untill convergency.



# Intuition and graphical explanation

Calculating the tangent line:

$$y = a + f'(x)x$$

To find the intercepr let's substitute the coordinate of our guess point ($x_0$ & $f(x_0)$):

$$f(x_0) = a + f'(x_0)x_0 \quad => \quad a = f(x_0) - f'(x_0)x_0$$

So the tangent line is

$$y = f(x_0) - f'(x_0)x_0 + f'(x_0)x$$

## Intuition and graphical explanation

To find the x-intercept of the tangent line we need to solve:

$$0 = f(x_0) - f'(x_0)x_0 + f'(x_0)x$$

$$0 = f(x_0) + f'(x_0)(x - x_0)$$

an $x$ that solves this equation is

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

# Notations and definitions

## Calculation in higher dimention

- $\mathbf{x}_n = (x_1, \cdots, x_n)^\top$: a random sample of $n$ observations from pdf $f(X|\theta)$.
- $\theta = (\theta_1, \cdots, \theta_q)^\top$: $q \times 1$ parameter vector.
- **Log-likelihood function**: $\ell(\theta) = \ln L(\theta) = \sum_{i=1}^n \ln f(x_i|\theta)$.
- **Score function**: $U(\theta) = \frac{\partial \ln L(\theta)}{\partial \theta} = \sum_{i=1}^n \frac{\partial \ln f(x_i|\theta)}{\partial \theta}$, is a $q \times 1$ vector.
- **Hessian function**: $H(\theta) = \frac{\partial U(\theta)}{\partial \theta^\top} = \frac{\partial^2 \ln L(\theta)}{\partial \theta \partial \theta^\top} = \sum_{i=1}^n \frac{\partial^2 \ln f(x_i|\theta)}{\partial \theta \partial \theta^\top}$, is a $q \times q$ matrix.
- **Observed information function**: $J(\theta) = -H(\theta)$.

# Newton-Raphson algorithm

The objective of the N-R algorithm is to solve $U(\theta) = 0$.

- Start with an appropriate initial value $\theta^{(0)}$.
- Compute $\theta^{(k)}$, $k = 0, 1, 2, \cdots$, successively with

$$\theta^{(k+1)} = \theta^{(k)} - \left[H(\theta^{(k)})\right]^{-1} U(\theta^{(k)}) \overset{\text{or}}{=} \theta^{(k)} + \left[J(\theta^{(k)})\right]^{-1} U(\theta^{(k)}).$$

- Continue until $\{\theta^{(k)}\}$ converges, i.e. until $|\theta^{(k+1)} - \theta^{(k)}|$ or $|U(\theta^{(k+1)})|$ or $|\ell(\theta^{(k+1)}) - \ell(\theta^{(k)})|$ is smaller than a small tolerance number (e.g. $10^{-6}$) computationally.

# Newton-Raphson algorithm

It is called the **Fisher-scoring algorithm** if computing via

$$\theta^{(k+1)} = \theta^{(k)} + \left[I(\theta^{(k)})\right]^{-1} U(\theta^{(k)}),$$

where, **Fisher information function**: $I(\theta) = E[J(\theta)] = -E[H(\theta)]$.

Fisher-scoring may be analytically more involving but is statistically more stable.

# Explaining N-R algorithm

- Suppose via Taylor expansion $\ln L(\theta)$ around the MLE $\hat{\theta}$ can be well approximated by a quadratic function

$$
\begin{aligned}
F(\theta) = \ln L(\theta^{(k)}) + \frac{\partial \ln L(\theta)}{\partial \theta}\Big|_{\theta=\theta^{(k)}}(\theta - \theta^{(k)}) \\
+\frac{1}{2}(\theta - \theta^{(k)})^\top \frac{\partial^2 \ln L(\theta)}{\partial \theta \partial \theta^\top}\Big|_{\theta=\theta^{(k)}}(\theta - \theta^{(k)}) \\
= \ln L(\theta^{(k)}) + U(\theta^{(k)})(\theta - \theta^{(k)}) + \frac{1}{2}(\theta - \theta^{(k)})^\top H(\theta^{(k)})(\theta - \theta^{(k)})
\end{aligned}
$$

- Solving $\frac{\partial F(\theta)}{\partial \theta} = 0 \Rightarrow U(\theta^{(k)}) + H(\theta^{(k)})(\theta - \theta^{(k)}) = 0$, we have

$$
\theta^{(k+1)} = \theta^{(k)} - \left[H(\theta^{(k)})\right]^{-1} U(\theta^{(k)}).
$$

- It implies $\hat{\theta} \approx \theta^{(k+1)} = \arg\max_\theta F(\theta)$.