Microeconometrics
Maximum Likelihood and Numerical Optimization

Pedro Portugal

NOVA School of Business and Economics

Spring 2025

The Likelihood Principle

- Choose as estimator of the parameter vector θ_0 that value of θ that maximizes the likelihood of observing the actual sample
  - Discrete case: this likelihood is the probability obtained from the probability mass function
  - Continuous case: this likelihood is the density
- The joint probability mass function or density $f(\mathbf{y}, \mathbf{X}|\theta)$ is viewed as a function of θ given the data ($\mathbf{y}$,$\mathbf{X}$)
- This is called the **likelihood function** and is denoted by $L_N(\theta|\mathbf{y}, \mathbf{X})$
- Maximizing $L_N(\theta)$ is equivalent to maximizing the **log-likelihood function**

$$\mathcal{L}_N(\theta) = \ln L_N(\theta)$$

Conditional Likelihood

- For cross-section data the observations $(y_i, \mathbf{x}_i)$ are independent over i with conditional density function $f(y_i|\mathbf{x}_i, \theta)$
- Then,

$$f(\mathbf{y}|\mathbf{X}, \theta) = \prod_{i=1}^{N} f(y_i|\mathbf{x}_i, \theta)$$

- Leading to the (conditional) log-likelihood function

$$\mathcal{Q}_N(\theta) = N^{-1}\mathcal{L}_N(\theta) = \frac{1}{N}\sum_{i=1}^{N} \ln f(y_i|\mathbf{x}_i, \theta)$$

  - where we divide by N so that the objective function is an average

Maximum Likelihood: Commonly Used Densities

Model	Range of y	Density $f(y)$	Common Parameterization
Normal	$(-\infty, \infty)$	$[2\pi\sigma^2]^{-1/2}e^{-(y-\mu)^2/2\sigma^2}$	$\mu = \mathbf{x}'\beta, \sigma^2 = \sigma^2$
Bernoulli	0 or 1	$p^y(1-p)^{1-y}$	Logit $p = e^{\mathbf{x}'\beta}/(1+e^{\mathbf{x}'\beta})$
Exponential	$(0, \infty)$	$\lambda e^{-\lambda y}$	$\lambda = e^{\mathbf{x}'\beta}$ or $1/\lambda = e^{\mathbf{x}'\beta}$
Poisson	$0, 1, 2, \ldots$	$e^{-\lambda}\lambda^y/y!$	$\lambda = e^{\mathbf{x}'\beta}$

Source: Cameron and Trivedi, 2005

Log-Likelihood Function

- Suppose we have a random sample of N observations of y and $\mathbf{x}$
- We can use these data to construct a series of probabilities corresponding to the sequence of observations on y

i	1	2	3	...	N
y_i	1	1	0	...	0
F_i	F($\mathbf{x}_1\beta$)	F($\mathbf{x}_2\beta$)	1-F($\mathbf{x}_3\beta$)	...	1-F($\mathbf{x}_N\beta$)

- The likelihood of each observation i will be

$$\ell_i(y_i|\mathbf{x}_i,\beta) = F(\mathbf{x}_i\beta)^{y_i}[1 - F(\mathbf{x}_i\beta)]^{1-y_i}$$

- And the log-likelihood function will be

$$\ln L_N(\beta) = \sum_{i=1}^{N}\{y_i \ln F(\mathbf{x}_i\beta) + (1 - y_i)\ln(1 - F(\mathbf{x}_i\beta))\}$$

Maximum Likelihood Estimator

- Maximizes the (conditional) log-likelihood function and is clearly an extremum estimator
- Usually the MLE is the local maximum that solves the first-order conditions

$$\frac{1}{N}\frac{\partial \mathcal{L}_N(\theta)}{\partial \theta} = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial \ln f(y_i|\mathbf{x}_i,\theta)}{\partial \theta} = \mathbf{0}$$

- This estimator is the **conditional MLE** as it is based on the conditional density of y given $\mathbf{x}$
- The gradient vector $\partial \mathcal{L}_N(\theta)/\partial \theta$ is called the **score vector** as it sums the first derivatives of the log density, and when evaluated at θ_0 it is called the **efficient score**

MLE (Cont.)

- Derive the foc for the optimum from the maximization of $\ln L_N$

$$\frac{\partial \ln L}{\partial \beta} = 0$$

  - and solve for $\widehat{\beta}$
- The sample analogue of this estimator is the solution of the following equation

$$\frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^{N} \frac{y_i - F(\mathbf{x}_i\beta)}{F(\mathbf{x}_i\beta)[1 - F(\mathbf{x}_i\beta)]} f(\mathbf{x}_i\beta)\mathbf{x}_i = 0$$

  - which is the sum of scores for each observation i
- For the *Logit* and *Probit*, the L_N function is concave (see Amemiya for a proof). Thus, the maximum is unique and easy to compute

Information Matrix Equality

ML Regularity Conditions

$$E_f\left[\frac{\partial \ln f(y|\mathbf{x},\theta)}{\partial \theta}\right] = \int \frac{\partial \ln f(y|\mathbf{x},\theta)}{\partial \theta} f(y|\mathbf{x},\theta) = \mathbf{0}$$

- and

$$-E_f\left[\frac{\partial^2 \ln f(y|\mathbf{x},\theta)}{\partial \theta \partial \theta'}\right] = E_f\left[\frac{\partial \ln f(y|\mathbf{x},\theta)}{\partial \theta}\frac{\partial \ln f(y|\mathbf{x},\theta)}{\partial \theta'}\right]$$

Information Matrix (Fisher Information)

- Is the expectation of the **outer product of the score vector**

$$\mathcal{I} = E\left[\frac{\partial \mathcal{L}_N(\theta)}{\partial \theta}\frac{\partial \mathcal{L}_N(\theta)}{\partial \theta'}\right]$$

Information Matrix Equality

- For log-likelihood function, the regularity condition implies that

$$-E_f\left[\left.\frac{\partial^2 \mathcal{L}_N(\theta)}{\partial\theta\partial\theta'}\right|_{\theta_0}\right] = E_f\left[\left.\frac{\partial \mathcal{L}_N(\theta)}{\partial\theta}\frac{\partial \mathcal{L}_N(\theta)}{\partial\theta'}\right|_{\theta_0}\right]$$

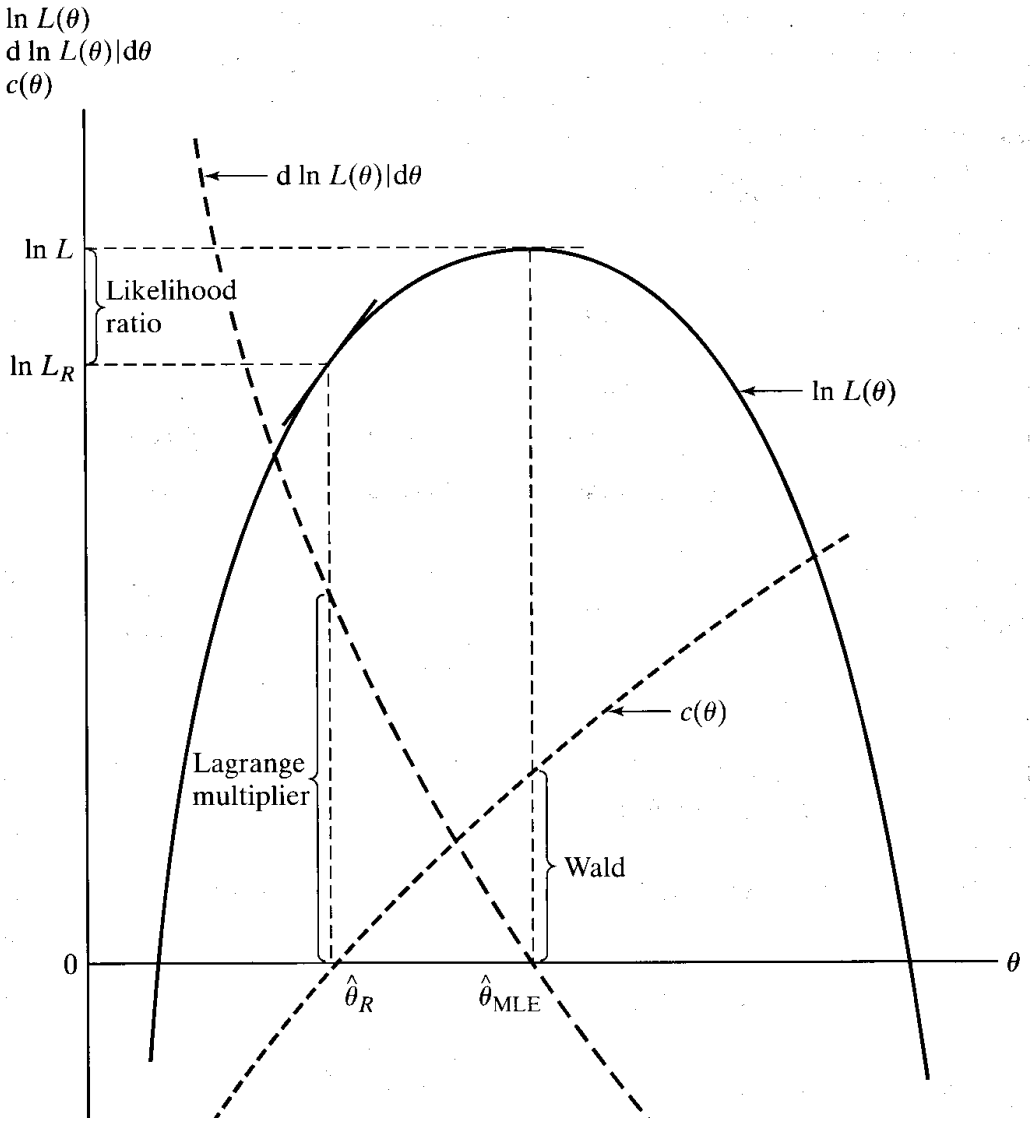
  - if the expectation is with respect to $f(y|\mathbf{x},\theta_0)$
- Implies that the information matrix also equals

$$\mathcal{I} = -E\left[\frac{\partial^2 \mathcal{L}_N(\theta)}{\partial\theta\partial\theta'}\right]$$

- The asymptotic distribution of the MLE is often expressed as

$$\widehat{\theta}_{ML} \overset{a}{\sim} \mathcal{N}\left[\theta, -\left(E\left[\frac{\partial^2 \mathcal{L}_N(\theta)}{\partial\theta\theta'}\right]\right)^{-1}\right]$$

Three Asymptotically Equivalent Testing Procedures

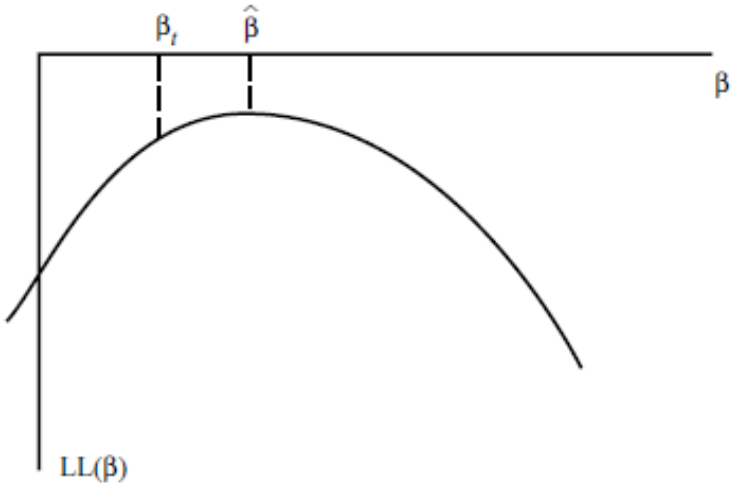


Three Asymptotically Equivalent Testing Procedures

Testing the hypothesis $H_0 : c(\theta) = 0$

- **Likelihood ratio test**: If the restriction $c(\theta) = 0$ is valid then imposing it should not lead to a large reduction in the log-likelihood function. The test is based on the difference $\ln L_U - \ln L_R$, where L_U and L_R are the values of the likelihood function at the unconstrained value of θ and at the restricted estimate, respectively.
- **Wald Test**: If the restriction is valid then $c(\hat{\theta}_{MLE})$ should be close to zero since the MLE is consistent. We reject the hypothesis if this value is significantly different from zero.
- **Lagrange multiplier test**: If the restriction is valid then the restricted estimator should be near the point that maximizes the log-likelihood. Therefore, the slope of the log-likelihood function should be near zero at the restricted estimator.

Maximum Likelihood



- Find the value of β that maximizes the $LL(\beta)$, ie, $\hat{\beta}$
- Note in this figure that LL is always negative, since the likelihood is a probability between 0 and 1 and the log of any number between 0 and 1 is negative
- The researcher specifies starting values β_0 and at each iteration moves to a new value of the parameters at which $LL(\beta)$ is higher than at the previous value

Maximum Likelihood

- The question is: what is the best step we can take next, that is, what is the best value for β_{t+1}?
- The gradient at β_t is the vector of first derivatives of $LL(\beta)$ evaluated at β_t

$$g_t = \left(\frac{\partial LL(\beta)}{\partial \beta} \right)_{\beta_t}$$

- This vector tells us how to move in order to go up the likelihood function. The Hessian is the matrix of second derivatives:

$$H_t = \left(\frac{\partial g_t}{\partial \beta'} \right)_{\beta_t} = \left(\frac{\partial^2 LL(\beta)}{\partial \beta \partial \beta'} \right)_{\beta_t}$$

Newton-Raphson

- To determine the best value of β_{t+1} take a second-order Taylor's approximation of $LL(\beta_{t+1})$ around $LL(\beta_t)$

$$LL(\beta_{t+1}) = LL(\beta_t) + (\beta_{t+1} - \beta_t)' g_t + \frac{1}{2}(\beta_{t+1} - \beta_t)' H_t (\beta_{t+1} - \beta_t)$$

- Find the value of β_{t+1} that maximizes this approximation to $LL(\beta_{t+1})$

$$\frac{\partial LL(\beta_{t+1})}{\partial \beta_{t+1}} = g_t + H_t(\beta_{t+1} - \beta_t) = 0$$

$$H_t(\beta_{t+1} - \beta_t) = -g_t$$

$$\beta_{t+1} - \beta_t = -H_t^{-1} g_t$$

$$\beta_{t+1} = \beta_t + (-H_t^{-1}) g_t$$

Figure: Direction of step follows the slope

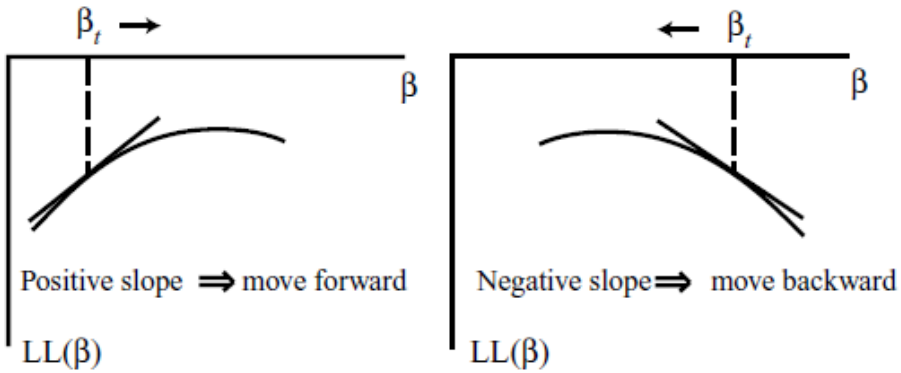
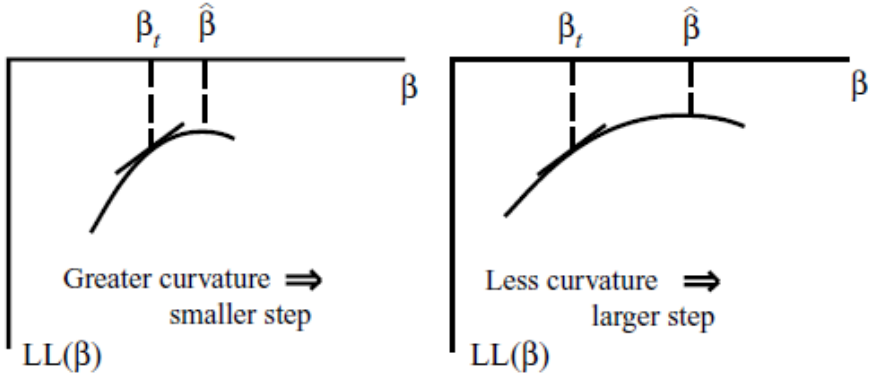
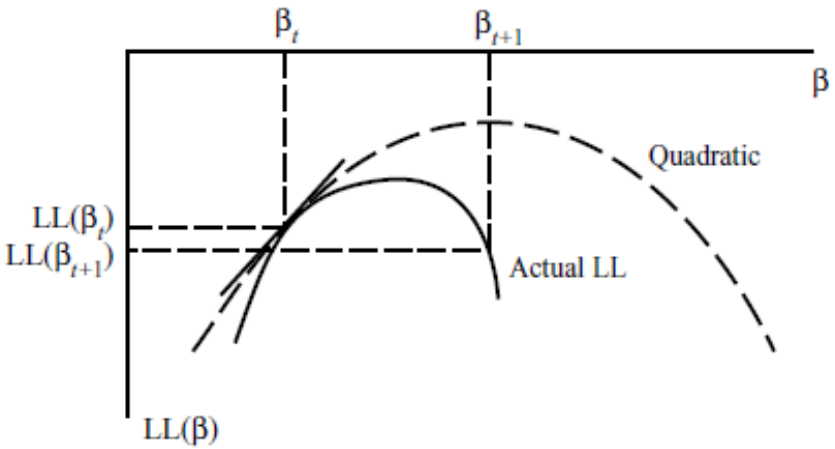


Figure: Step size is inversely related to curvature



NR (Cont.)

- It is possible for the NR procedure to step past the maximum and move to a lower $LL(\beta)$
- The actual LL is given by the solid line. The dashed line is a quadratic function that has the slope and curvature that LL has at the point β_t
- The NR procedure moves to the top of the quadratic, to β_{t+1}. However, $LL(\beta_{t+1})$ is lower than $LL(\beta_t)$ in this case

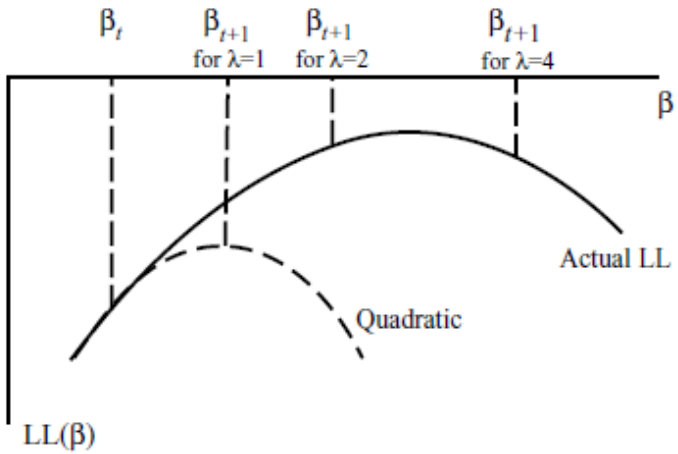


Step Size

- To allow for this possibility

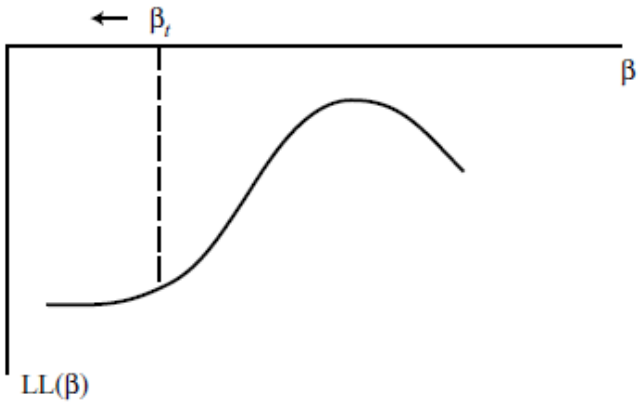
$$\beta_{t+1} = \beta_t + \lambda(-H_t^{-1})g_t$$

  - The vector $(-H_t^{-1})g_t$ is called the direction, and λ is called the step size
- The step size λ is reduced to assure that each step of the NR procedure provides an increase in $LL(\beta)$



Concavity

- Suppose the log-likelihood function has regions that are not concave. In these areas, the NR procedure can fail to find an increase
- If the function is convex at β_t, then the NR procedure moves in the opposite direction to the slope of the log-likelihood function
- The NR step with $K = 1$ is $LL'(\beta)/(-LL''(\beta))$. The second derivative is positive at β_t, since the slope is rising, and therefore $(-LL''(\beta))$ is negative and the step is in the opposite direction to the slope



Berndt, Hall, Hall, and Hausman (1974)

- Uses B_t in the optimization routine in place of $-H_t$
- Each iteration os defined by

$$\beta_{t+1} = \beta_t + \lambda B_t^{-1} g_t$$

- This step is the same as for NR except that B_t is used in place of $-H_t$ where B_t is the average outer product in the sample

$$B_t = \sum_n s_n(\beta_t) s_n(\beta_t)' / N$$

- The score of an observation is the derivative of that observation's LL with respect to the parameters

$$s_n(\beta_t) = \partial \ln P(\beta)/\partial\beta \text{ evaluated at } \beta_t$$

BHHH

- The gradient is the average score

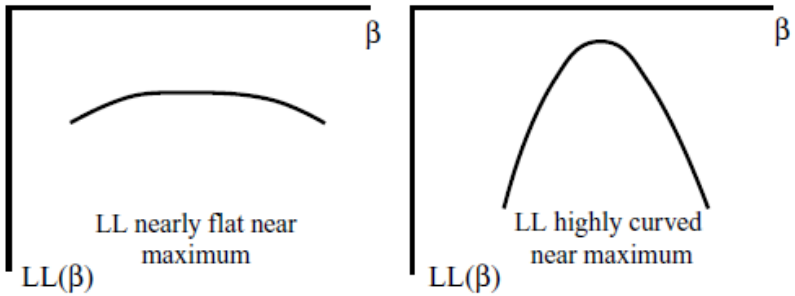
$$g_t = \sum_n s_n(\beta_t)/N$$

- The outer product of observation n's score is the $k \times K$ matrix

$$s_n(\beta_t)s_n(\beta_t)' = \begin{bmatrix} s_n^1 s_n^1 & s_n^1 s_n^2 & \dots & s_n^1 s_n^K \\ s_n^1 s_n^2 & s_n^2 s_n^2 & \dots & s_n^2 s_n^K \\ \dots & \dots & & \dots \\ s_n^1 s_n^K & s_n^2 s_n^K & \dots & s_n^K s_n^K \end{bmatrix}$$

  - where s_n^k is the kth element of $s_n(\beta_t)$

Shape of LL function near maximum



- If all individuals in the sample have similar scores, the LL function is fairly flat, given that different values of the parameters fit the data about the same
  - The curvature is small when the variance of the scores is small
- Scores differing greatly over observations mean that the observations are quite different, the LL function is highly peaked, given that the sample provides good information on the values of β
  - The curvature is great when the variance of the scores is high

BHHH-2

- Variant on the BHHH procedure obtained by subtracting out the mean score before taking the outer product
- For any level of the average score, the covariance of the scores over the sampled decision makers is

$$W_t = \sum_n \frac{(s_n(\beta_t) - g_t)(s_n(\beta_t) - g_t)'}{N}$$

  - where the gradient g_t is the average score
- W_t is the covariance of the scores around their mean, and B_t is the average outer product of the scores
- The maximization procedure can use W_t instead of B_t

$$\beta_{t+1} = \beta_t + \lambda W_t^{-1} g_t$$

Steepest Ascent

- This procedure is defined by the iteration formula

$$\beta_{t+1} = \beta_t + \lambda g_t$$

  - The defining matrix is the identity matrix I
- It provides the greatest possible increase in $LL(\beta)$ for the distance between β_t and β_{t+1}, at least for small enough distance

DFP and BFGS

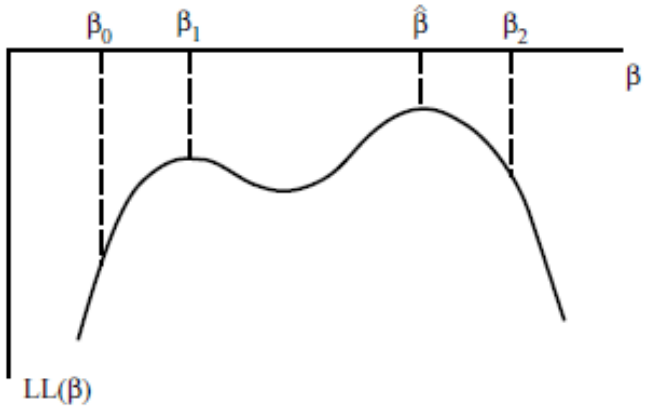
- Calculate the approximate Hessian using information at more than one point on the likelihood function
- NR uses the actual Hessian at β_t to determine the step to β_{t+1}, and BHHH and BHHH-2 use the scores at β_t to approximate the Hessian
- In contrast, the DFP and BFGS procedures use information at several points to obtain a sense of the curvature of the LL function
- The Hessian is the matrix of second derivatives and therefore it gives the amount by which the slope of the curve changes as one moves along the curve
- Since we are interested in making large steps is useful to understand how the slope changes for noninfinitesimal movements

arc Hessian

- Consider a function $f(x)$ with slope at $x = 3$ equal to 25 and at $x = 4$ equal to 19
- The change in slope for a one unit change in x is -6
- Therefore the arc Hessian is -6, representing the change in the slope as a step is taken from $x = 3$ to $x = 4$
- The DFP and BFGS procedures calculate the gradient at each step in the iteration process
  - The difference in the gradient between the various points that have been reached is used to calculate an arc Hessian over these points
  - The arc Hessian reflects the change in the gradient that occurs for actual movement on the curve, as opposed to the Hessian which reflects the change in slope for infinitesimally small steps around that point

Local versus Global Maximum

- All of the methods previously discussed are susceptible to converging at a local maximum that is not the global maximum



- Starting at β_0 will lead to convergence at β_1
- Unless other starting values were tried, the researcher would mistakenly believe that the maximum of $LL(\beta)$ had been achieved
- Starting at β_2, convergence is achieved at $\hat{\beta}$. Comparing $LL(\hat{\beta})$ with $LL(\beta_1)$ shows that β_1 is not a maximizing value