Lecture 2: Numerical Optimization

Professor: Mauricio Sarrias

Universidad de Talca

2020

Reading

Reading (Mandatory):

- (Ruud) - Chapter 16
- (Train) - Chapter 8

Goals

- Understand the main algorithms to estimate nonlinear models.
- Understand the advantages and disadvantages of each procedure.

Motivation

- Most estimation procedure involves maximization of some function (likelihood function, simulated likelihood function, or squared moment condition)
- We will review numerical procedures that are used to maximize a likelihood function.
- The procedures for maximization are fairly straightforward and easy to implement.

We begin with a general discussion on how to search for **a solution to a nonlinear optimization problem** and describe some specific commonly used methods.

Motivation

- ML estimates are obtained by setting the gradient of the log likelihood to 0 and solving for the parameters using algebra.
- Algebraic solutions are rarely possible with nonlinear models.
- Numerical methods start with a ***guess*** of the values of the parameters and **iterate** to improve that guess.

Notation

The (average) log-likelihood takes the form:

$$\log L(\boldsymbol{\beta}) = \frac{\sum_{i=1}^{n} \ln P_i(\boldsymbol{\beta})}{n}$$

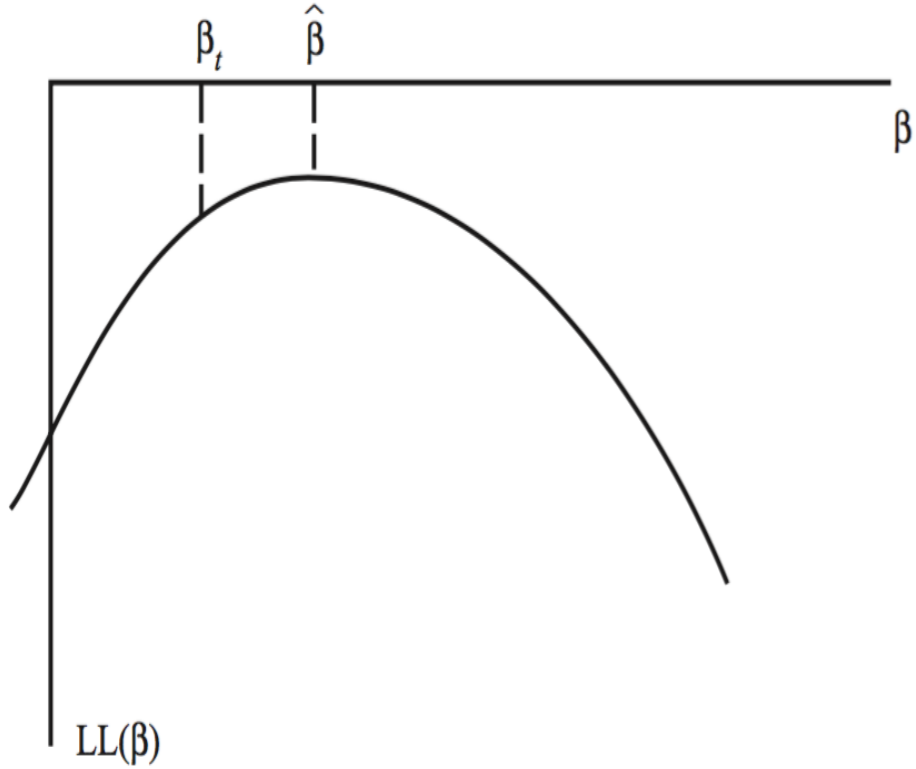
where:

- $P_i(\boldsymbol{\beta})$ is the probability of the observed outcome for decision maker i,
- n is the sample size,
- $\boldsymbol{\beta}$ is a $K \times 1$ vector of parameter.

Including n facilities interpretation, however all the procedures operate the same whether or not the log-likelihood is divided by n.

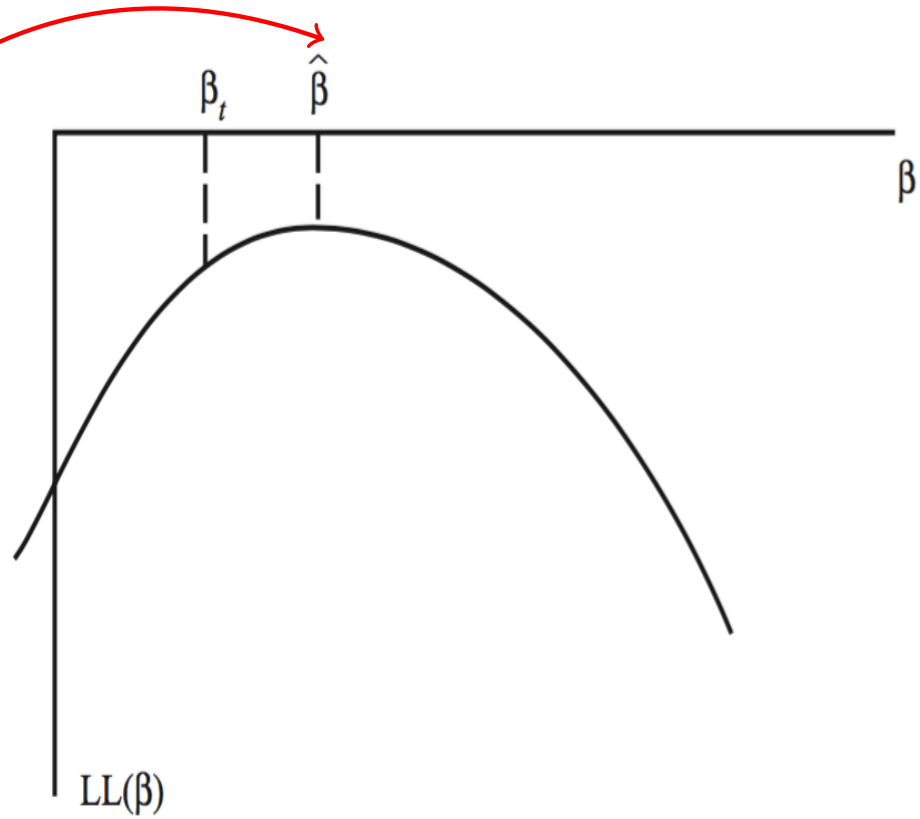
The goal: To find β

- The goal is to locate $\widehat{\beta}$
- Why the LL is always negative?
- The max can be found by 'walking up' the likelihood function until no further increase can be found
- Set starting values β_0 and iterate.
- β_t is the $\widehat{\beta}$ in iteration t.



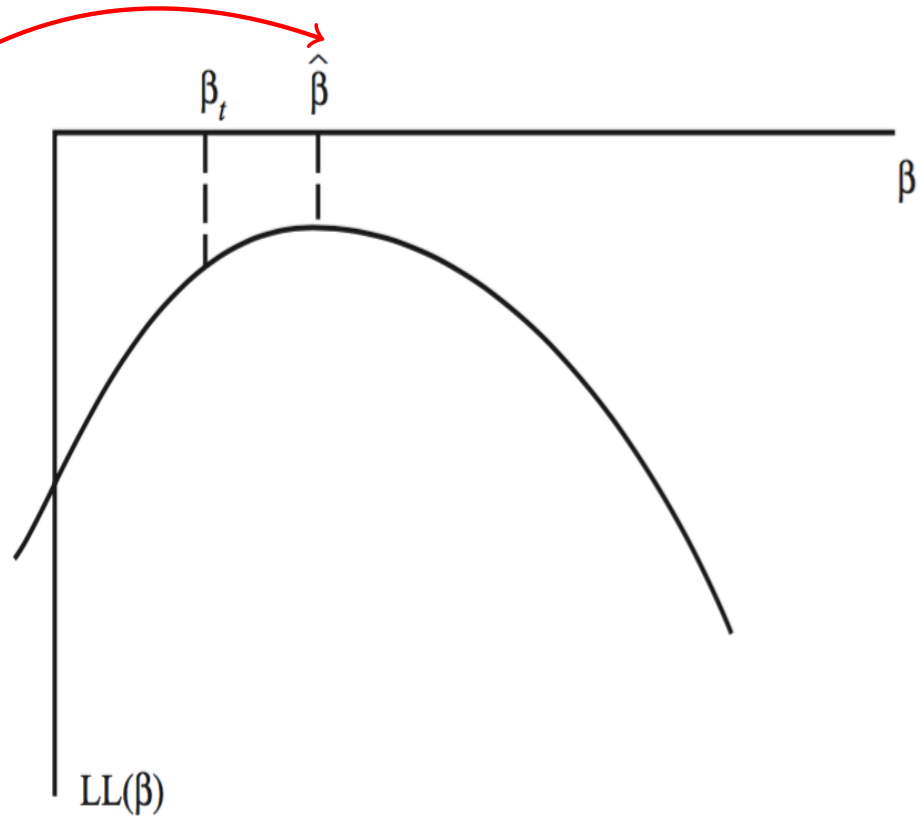
The goal: To find β

- The goal is to locate $\widehat{\beta}$
- Why the LL is always negative?
- The max can be found by 'walking up' the likelihood function until no further increase can be found
- Set starting values β_0 and iterate.
- β_t is the $\widehat{\beta}$ in iteration t.



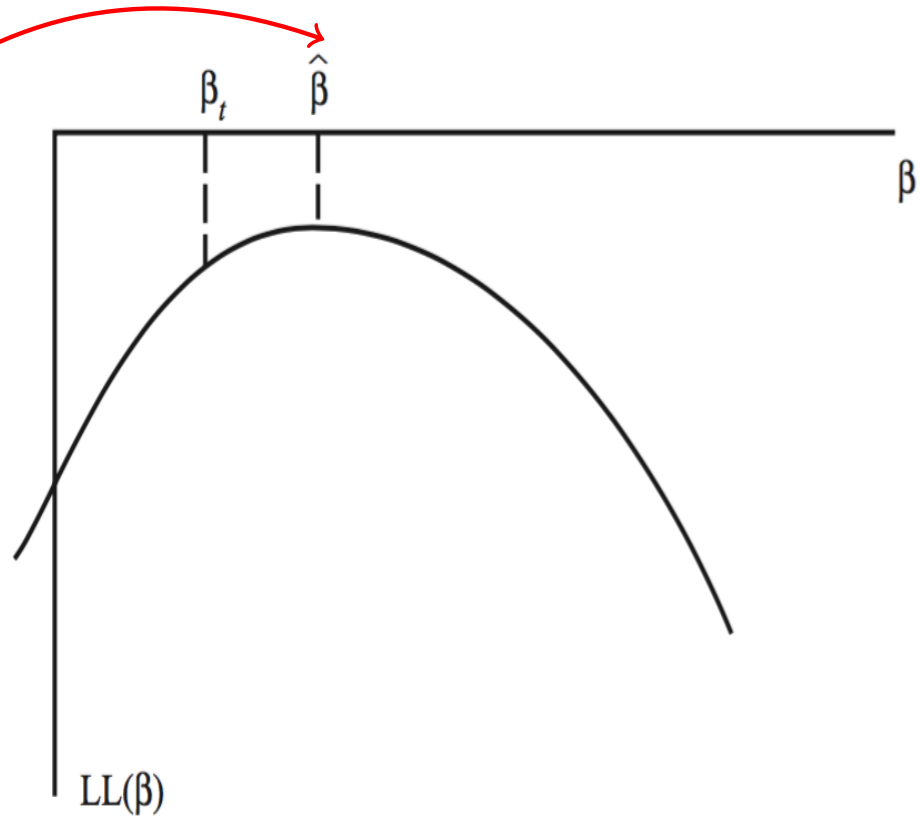
The goal: To find β

- The goal is to locate $\widehat{\beta}$
- Why the LL is always negative?
- The max can be found by 'walking up' the likelihood function until no further increase can be found
- Set starting values β_0 and iterate.
- β_t is the $\widehat{\beta}$ in iteration t.



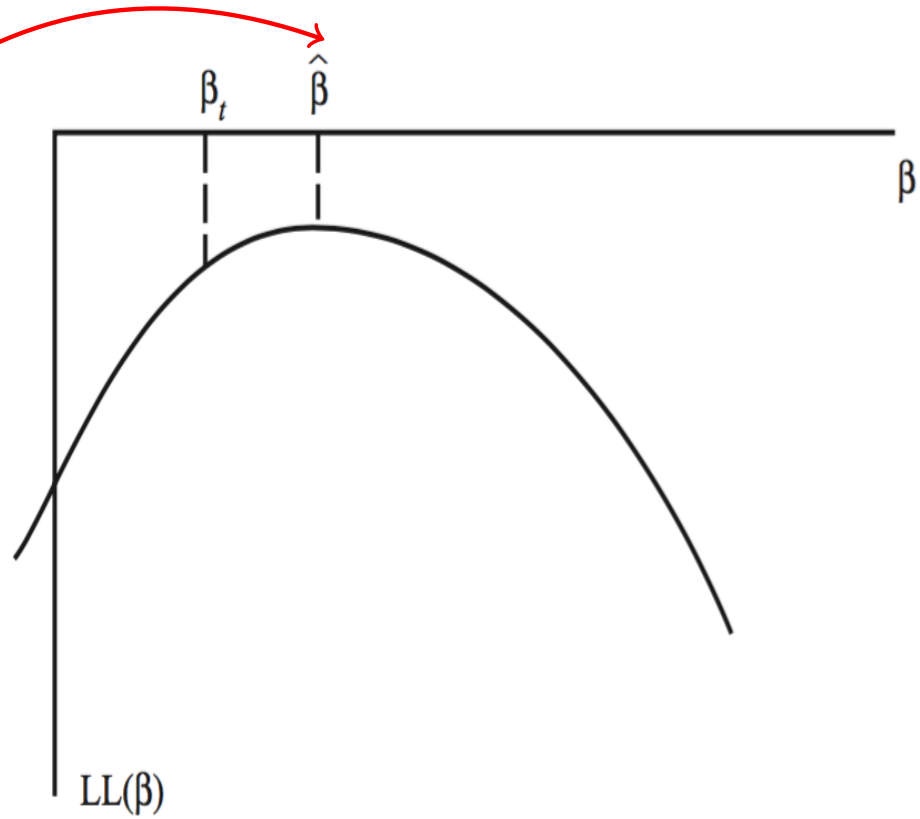
The goal: To find β

- The goal is to locate $\widehat{\beta}$
- Why the LL is always negative?
- The max can be found by 'walking up' the likelihood function until no further increase can be found
- Set starting values β_0 and iterate.
- β_t is the $\widehat{\beta}$ in iteration t.



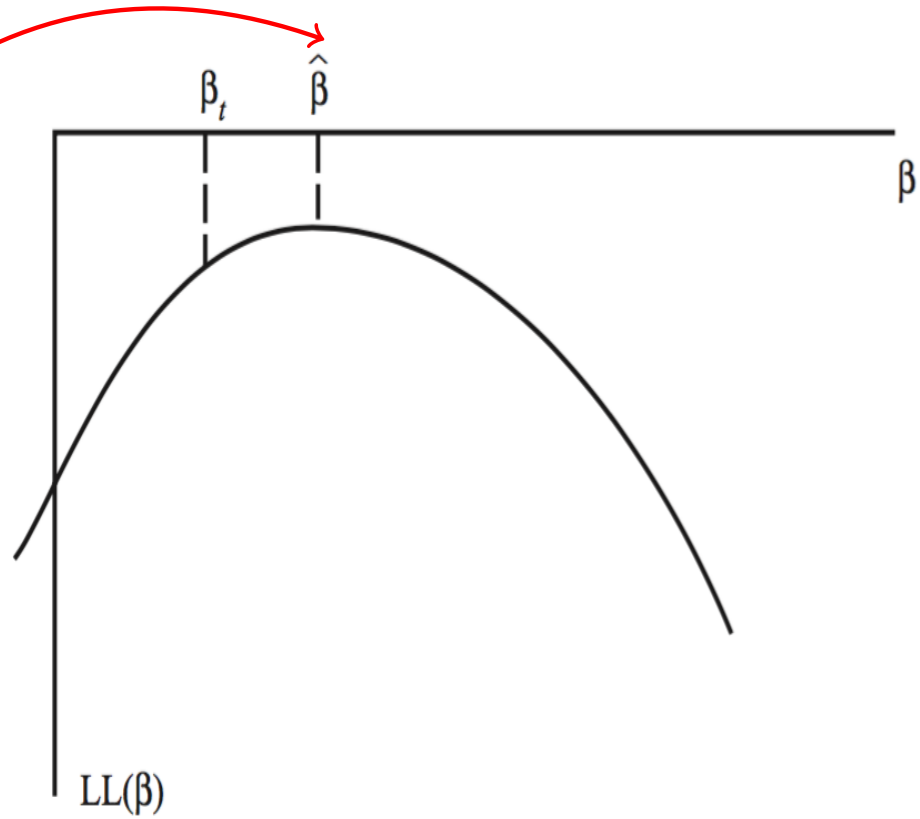
The goal: To find β

- The goal is to locate $\widehat{\beta}$
- Why the LL is always negative?
- The max can be found by 'walking up' the likelihood function until no further increase can be found
- Set starting values β_0 and iterate.
- β_t is the $\widehat{\beta}$ in iteration t.



The goal: To find β

- The goal is to locate $\widehat{\beta}$
- Why the LL is always negative?
- The max can be found by 'walking up' the likelihood function until no further increase can be found
- Set starting values β_0 and iterate.
- β_t is the $\widehat{\beta}$ in iteration t.



Notation

- Each iteration (step) moves to a new value of the parameters at which $\log L(\boldsymbol{\beta})$ is higher than at the previous values.
- Let $\boldsymbol{\beta}_t$ denotes the current value of $\boldsymbol{\beta}$ attained after t steps from the starting values.

Question

What is the best step we can take next, that is, what is the best value of $\boldsymbol{\beta}_{t+1}$?

Notation: Gradient

The gradient $\boldsymbol{\beta}_t$ is the vector of first derivatives of $\log L(\boldsymbol{\beta})$ evaluated at $\boldsymbol{\beta}_t$:

$$\mathbf{g}_t = \left. \frac{\partial \log L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right|_{\beta_t}$$

This vector tell us **which way to step** in order to go up the likelihood function.

But ... **How much to move?**

Notation: Hessian

Using the Hessian the matrix of second derivatives evaluated at $\boldsymbol{\beta}_t$:

$$\mathbf{H}_t = \left.\frac{\partial \mathbf{g}_t}{\partial \boldsymbol{\beta}'}\right|_{\boldsymbol{\beta}_t} = \left.\frac{\partial^2 \log L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}\right|_{\boldsymbol{\beta}_t}$$

The gradient has dimension $K \times 1$, and the Hessian is $K \times K$. As we will see, the Hessian can help up to know **how far to step**, given that the gradient tell us **in which direction** to step.

Iterative Solutions

We start with an initial guess $\boldsymbol{\beta}_0$ (called **starting values**) and attempt to improve on this guess by adding a vector $\boldsymbol{\zeta}_0$ of adjustments:

$$\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0 + \boldsymbol{\zeta}_0$$

We proceed by updating the previous iteration according to the equation:

$$\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t + \boldsymbol{\zeta}_t$$

Iterations continue until there is **convergence**.

1. Roughly, convergence occurs when the gradient of the log-likelihood is close to 0 or the estimates do not change form one step to the next.
2. Convergence must occur to obtain the ML estimator $\widehat{\boldsymbol{\beta}}$

Iterative Solutions

The problem is to find a $\boldsymbol{\zeta}_t$ that moves the process rapidly toward a solution.

- Think of $\boldsymbol{\zeta}_t$ consisting of two parts:

$$\boldsymbol{\zeta}_t = \mathbf{D}_t \boldsymbol{\gamma}_t$$

  - $\boldsymbol{\gamma}_t = \partial \log L / \partial \boldsymbol{\beta}_t$, which indicates the direction of change.
  - $\mathbf{D}_t$ is a direction matrix that reflects the curvature of the log likelihood function (**How rapidly the gradient is changing!**)

The Method of Steepest Ascent

The method of steepest ascent lets $\mathbf{D} = \mathbf{I}$:

$$\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t + \lambda \left. \frac{\partial \log L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right|_{\boldsymbol{\beta}_t}$$

- An estimate increases if the gradient is positive, and it decreases if the gradient is negative.
- Iterations stop when the derivative becomes nearly 0.
- It is called 'steepest ascent" because it provides the greatest possible increase in $\log L(\boldsymbol{\beta})$ for the distance between $\boldsymbol{\beta}_t$ and $\boldsymbol{\beta}_{t+1}$, at least for small enough distance.
- The problem: it considers the slope of $\log L$, but no how quickly the slope is changing.
- You should move more gradually for the function that is changing quickly, in order to avoid moving too far.

The Method of Steepest Ascent

The method stops if $\|\mathbf{g}_t\| < \epsilon$, where ϵ =small number.

Recall that $\|\mathbf{x}\| = \sqrt{\sum_{k=1}^{K} x_k^2}$

The Method of Steepest Ascent

- The next commonly used methods address this problem by adding a direction matrix that assesses how quickly the log likelihood function is changing.
- They differ in the choice of $\mathbf{D}$.
- In all cases, it takes longer to compute the direction matrix.
- Usually, the additional computational cost are made up by the fewer iterations that are required to reach convergence.

The Newton-Raphson Method

The rate of change in the slope of $\log L$ is indicated by the second derivatives (Hessian Matrix). If $\boldsymbol{\theta} = (\alpha, \beta)'$. Then:

$$\mathbf{D} = \mathbf{H} = \frac{\partial^2 \log L}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \begin{pmatrix} \frac{\partial^2 \ln L}{\partial \alpha \partial \alpha} & \frac{\partial^2 \ln L}{\partial \alpha \partial \beta} \\ \frac{\partial^2 \ln L}{\partial \beta \partial \alpha} & \frac{\partial^2 \ln L}{\partial \beta \partial \beta} \end{pmatrix}$$

- If $\partial^2 \ln L / \partial \alpha \alpha$ is large relative to $\partial^2 \ln L / \partial \beta \partial \beta$, the gradient is changing more rapidly as α changes than β changes.
- Thus small adjustments to the estimate of α would be indicated

So:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \left(\frac{\partial^2 \log L}{\partial \boldsymbol{\theta}_t \partial \boldsymbol{\theta}_t'} \right)^{-1} \frac{\partial \ln L}{\partial \boldsymbol{\theta}_t}$$

The Newton-Raphson Method

Proof NR.

Take a second-order Taylor's approximation of $\ln L(\boldsymbol{\theta}_{t+1})$ around $\ln L(\boldsymbol{\theta}_t)$:

$$\ln L(\boldsymbol{\theta}_{t+1}) = \ln L(\boldsymbol{\theta}_t) + (\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t)'\mathbf{g}_t + \frac{1}{2}(\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t)'\mathbf{H}_t(\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t)$$

FOC gives:

$$\begin{aligned}
\frac{\partial \ln L(\boldsymbol{\theta}_{t+1})}{\partial \boldsymbol{\theta}_{t+1}} = \mathbf{g} + \mathbf{H}_t(\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t) &= \mathbf{0} \\
\mathbf{H}_t(\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t) &= -\mathbf{g}_t \\
\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t &= -\mathbf{H}_t^{-1}\mathbf{g}_t \\
\boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t - \mathbf{H}^{-1}\mathbf{g}_t.
\end{aligned}$$

□

The Newton-Raphson Method

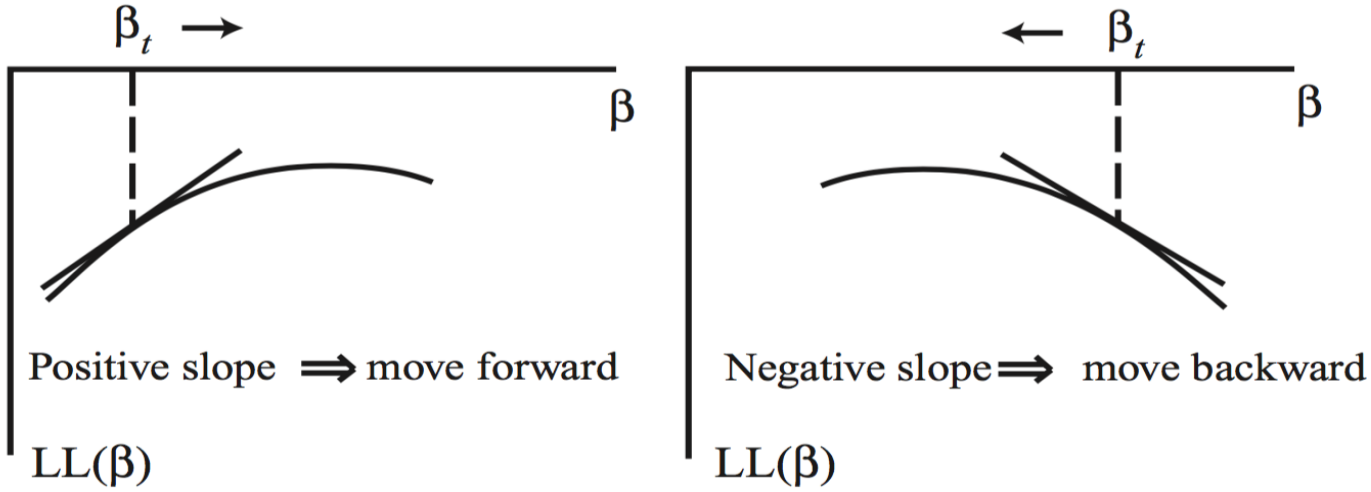


Figure 8.2. Direction of step follows the slope.

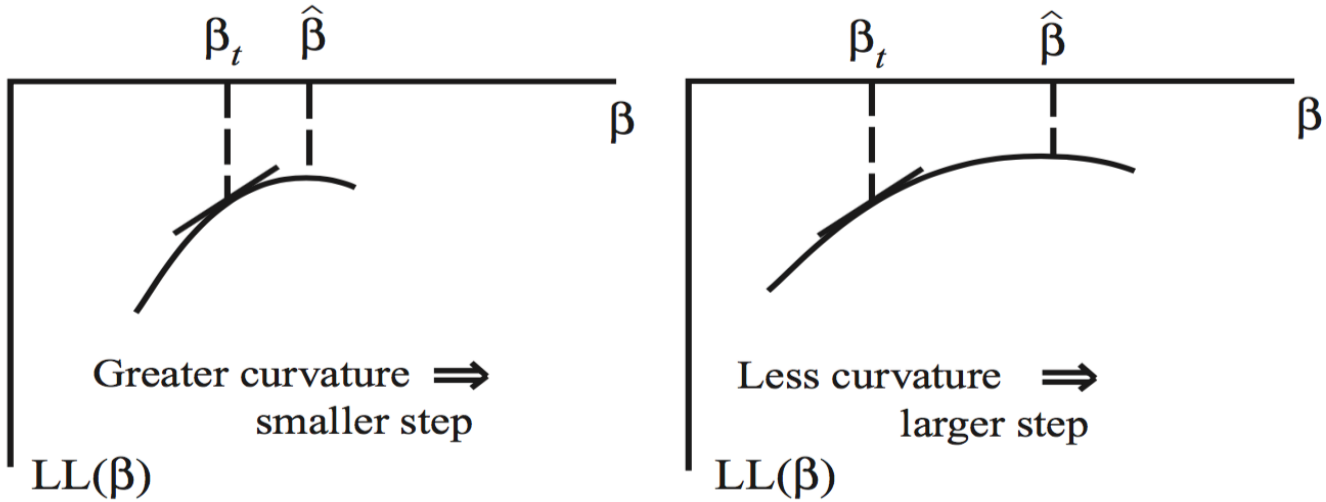


Figure 8.3. Step size is inversely related to curvature.

Step Size

- It is possible to step past the max and move to a lower $\ln L(\boldsymbol{\beta}_{t+1})$;
- The NR procedure moves to the top of the quadratic, to $\boldsymbol{\beta}_{t+1}$.
- But, $\ln L(\boldsymbol{\beta}_{t+1})$ is lower than $\ln L(\boldsymbol{\beta}_t)$

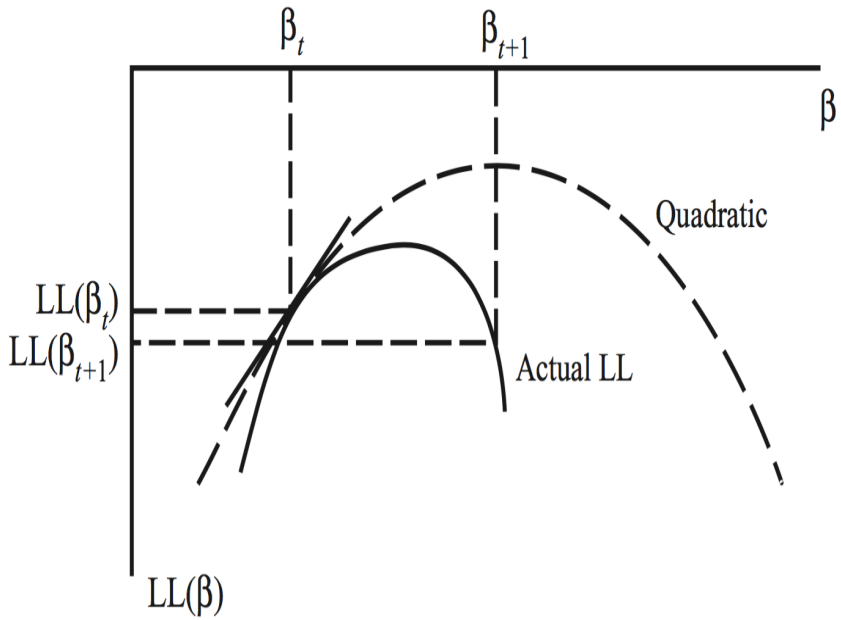


Figure 8.4. Step may go beyond maximum to lower LL.

Step Size

To allow for this possibility, the step is multiplied by a scalar λ in the NR formula:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \lambda(-\mathbf{H}_t)^{-1}\mathbf{g}_t$$

- $(-\mathbf{H}_t)^{-1}\mathbf{g}_t$ is called the direction,
- λ is called the step size.
- The **step size** λ is reduced to assure that each step of the NR procedure provides and increase in $\log L(\boldsymbol{\theta})$

Step Size

The adjustment is performed separately in each iteration, as follows.

1. Start with $\lambda = 1$. If $\ln L(\boldsymbol{\theta}_{t+1}) > \ln L(\boldsymbol{\theta}_t)$, move to $\boldsymbol{\theta}_{t+1}$ and start a new iterations.
2. If $\ln L(\boldsymbol{\theta}_{t+1}) < \ln L(\boldsymbol{\theta}_t)$, then set $\lambda = 1/2$ and try again.
3. If, with $\lambda = 1/2$, then set $\lambda = 1/4$ and try again.
4. Continue this process until a λ is found for which $\ln L(\boldsymbol{\theta}_{t+1}) > \ln L(\boldsymbol{\theta}_t)$
5. If this process results in a tiny λ, then little progress is made in finding the maximum.
6. This can be taken as a signal to the researcher that a different procedure may be needed.

Concavity

- If $\ln L$ function is globally concave (GC), then the NR procedure is guaranteed to provide an increase in the LL at each iteration.
- If $\ln L$ is GC $\implies$ $\mathbf{H}_n$ is negative definite at all values of $\boldsymbol{\beta}$
- IF $\mathbf{H}$ is negative definite, then $\mathbf{H}^{-1}$ is also negative definite, and $-\mathbf{H}^{-1}$ is positive definite

Definition (Positive definite Matrix)

A symmetric matrix $\mathbf{M}$ is positive definite if $\mathbf{x}'\mathbf{M}\mathbf{x} > 0$ for any $\mathbf{x} \neq \mathbf{0}$.

Concavity

Consider a first-order Taylor's approximation of $\ln L(\boldsymbol{\theta}_{t+1})$ around $\ln L(\boldsymbol{\theta}_{t+1})$:

$$\ln L(\boldsymbol{\theta}_{t+1}) = \ln L(\boldsymbol{\theta}_t) + (\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t)'\mathbf{g}_t$$

Under NR, $\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t = \lambda(-\mathbf{H}_t)^{-1}\mathbf{g}_t$. Substituting gives:

$$\begin{aligned}\ln L(\boldsymbol{\theta}_{t+1}) &= \ln L(\boldsymbol{\theta}_t) + \left[\lambda(-\mathbf{H}_t)^{-1}\mathbf{g}_t\right]'\mathbf{g}_t \\ &= \ln L(\boldsymbol{\theta}_t) + \lambda\mathbf{g}_t'(-\mathbf{H}_t)^{-1}\mathbf{g}_t\end{aligned}$$

- Since $-\mathbf{H}^{-1}$ is positive definite, we have $\mathbf{g}_t'(-\mathbf{H}_t)^{-1}\mathbf{g}_t > 0$ and $\ln L(\boldsymbol{\theta}_{t+1}) - \ln L(\boldsymbol{\theta}_t) > 0$.
- Since this comparison is based on a first-order approximation, and increase in $\ln L(\boldsymbol{\theta})$ may only be obtained in a small neighborhood of $\boldsymbol{\theta}$.
- This implies a low λ.
- However, an increase is indeed guaranteed at each iteration if $\ln L(\boldsymbol{\theta})$ is GC.

What if the LL is not concave?

- NR with one parameter is $LL'(\beta)/(-LL''(\beta))$.
- $LL''(\beta) > 0 \implies -LL''(\beta) < 0$: the step is in the opposite direction to the slope.

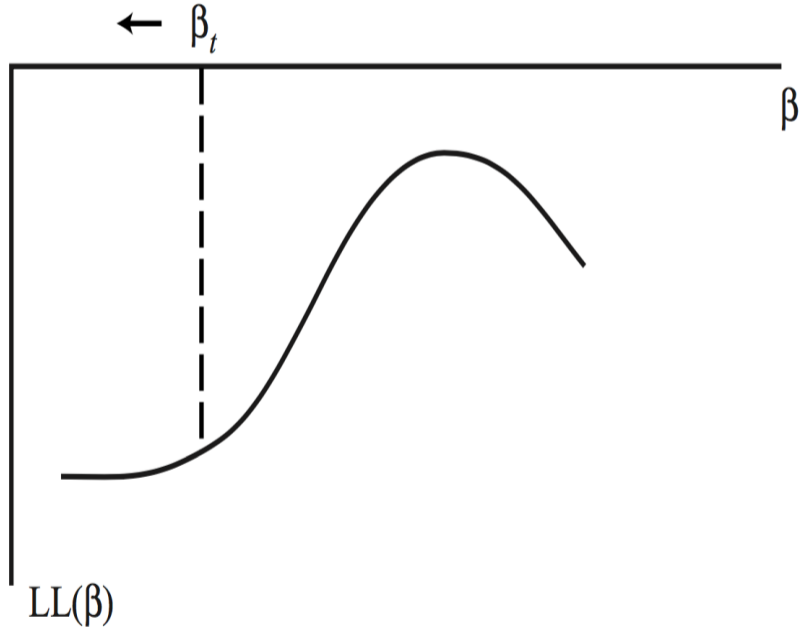


Figure 8.6. NR in the convex portion of LL.

NR

Disadvantages

- Calculation the Hessian is usually computation-intensive:
  - Procedures that avoid calculating the Hessian at every iteration can be much faster.
- NR does not guarantee an increase in each step if the log-likelihood function is not globally concave.
  - When $-\mathbf{H}_t^{-1}$ is not positive definite, an increase is not guaranteed.

BHHH

Berndt et al. (1974) propose using an outer product of the gradient approximation to the information matrix:

$$\mathbf{D} = \mathbf{H} = \sum_{i=1}^{N} \left(\frac{\partial \ln L_i}{\partial \boldsymbol{\theta}_t} \right) \left(\frac{\partial \ln L_i}{\partial \boldsymbol{\theta}_t} \right)'$$

where $\ln L_i$ is the value of the likelihood function evaluated for the ith observation. This approximation is often simpler to compute since only the gradient needs to be evaluated.

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \left(\sum_{i=1}^{N} \left(\frac{\partial \ln L_i}{\partial \boldsymbol{\theta}_t} \right) \left(\frac{\partial \ln L_i}{\partial \boldsymbol{\theta}_t} \right)' \right)^{-1} \frac{\partial \ln L}{\partial \boldsymbol{\theta}_t}$$

which is known as the BHHH algorithm or the modified method of scoring.

BHHH

Recall the score function

$$\mathbf{g}_t = \sum_i \mathbf{s}_i(\boldsymbol{\beta}_t)/n$$

where the score function:

$$\mathbf{s}_i(\boldsymbol{\beta}_t) = \frac{\partial \ln P_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$$

evaluated at $\boldsymbol{\beta}_t$

- The average outer product in the sample is related to the covariance matrix:
  - if the average score were zero, then $\mathbf{B}_t = \sum_i \mathbf{s}_i(\boldsymbol{\beta}_t)\mathbf{s}_i(\boldsymbol{\beta})'/n$ would be the covariance matrix of scores in the sample.
- Suppose that all the people in the sample have similar scores. Then the sample contains very little information. Then $\log L$ function is fairly flat in this situation, reflecting the fact that different values of the parameters fit the data about the same.

BHHH

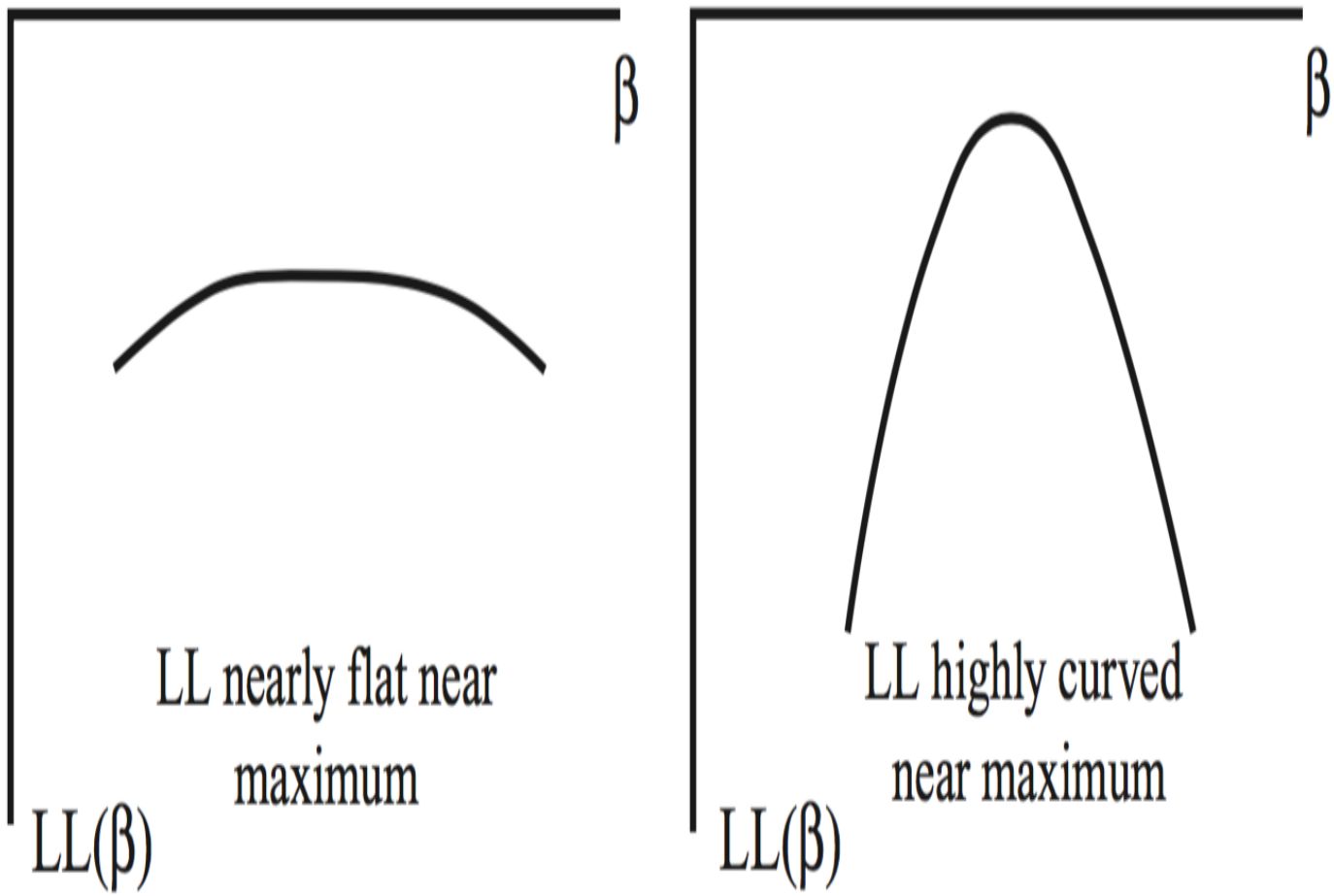


Figure 8.7. Shape of log-likelihood function near maximum.

BHHH

There are tow advantages to the BHHH procedure over NR:

- $\mathbf{B}_t$ is far faster to calculate that $\mathbf{H}_t$
- $\mathbf{B}_t$ is necessarily positive definite.

Drawbacks:

- The procedure can give small steps that raise $\log L(\boldsymbol{\beta})$ very little, especially when the iterative process is far from the maximum.
- This behaviour can arise because $\mathbf{B}_t$ is not a good approximation to $-\mathbf{H}_t$ far from the true value, or because $\log L(\boldsymbol{\beta})$ is highly non-quadratic in the are the problem is occurring.

DFP and BFGS

- The Davidon-Fletcher-Powerll (DFP) and Broyden-Fletcher-Goldfarb-Shanno (BFGS) methods calculate the approximate Hessian in a way that uses information at more than one point on the likelihood function.
- They update the arc Hessian.
- The two procedures differ in how the updating is performed.
- Both methods are extremely effective—usually far more efficient that NR, BHHH, or steepest ascent.
- BFGS is the default algorithm in the optimization routines of many commercial software packages.

Numerical Derivatives

Consider $K = 1$. The gradient is approximated by computing the slope of the change in $\log L$ when θ changes by a small amount. If Δ is a small number relative to θ,

$$\frac{\partial \log L}{\partial \theta} \approx \frac{\ln L(\theta + \Delta) - \ln L(\theta)}{\Delta}$$

- Numerical estimates can greatly increase the time and number of iterations needed, and results can be sensitive to the choice of Δ.
- Different start values can result in different estimates of the Hessian at convergence (different standard errors).
- Programs that use numerical methods for computing derivatives should only be used if no alternatives are available.

Convergence Criterion

- Stop if $\mathbf{g}_t = \mathbf{0}$ will never occur.
- When are we sufficiently close to the maximum to justify stopping the iterative process?
  - The statistic $m_t = \mathbf{g}_t'(-\mathbf{H}_t)^{-1}\mathbf{g}_t$ is often used to evaluate convergence.
  - Set $\breve{m} = 0.0001$, for example, and determines in each iteration if $\mathbf{g}_t'(-\mathbf{H}_t)^{-1}\mathbf{g}_t < \breve{m}$.
  - $(\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t)'(\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t) < \breve{m}$.
  - $(\log L(\boldsymbol{\theta}_{t+1}) - \log L(\boldsymbol{\theta}_t)) < \breve{m}$

Local versus Global Maximum

- Convergence to a local maximum that is not the global maximum;
- Globally concave: just one maximum;
- Try different starting values;

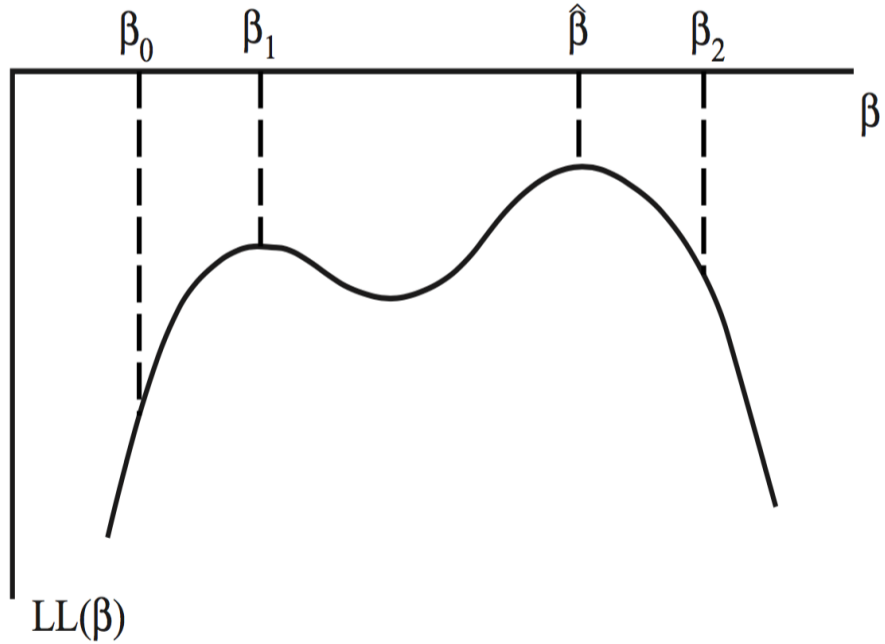


Figure 8.8. Local versus global maximum.

Problems with numerical Methods

- "Convergence not obtained after 250 iterations"
- Hessian is nearly flat:
  - "Singularity encountered"
  - "Hessian could not be inverted"
  - "Hessian was not full rank"
- Wrong solution: Local maximum or saddle point

Problems with numerical Methods

Potential solutions:

1. Incorrect Variables: Check summary statistics.
2. Number of observations: Convergence generally occurs more rapidly when there are more observations, and when the ratio of the number of observations to the number of variables is larger.
3. Scaling variables: The larger the ratio between the largest standard deviation and the smallest standard deviation, the more problem you will have with numerical methods. Experience suggests that problems are much more likely when the ratio between the largest and smallest standard deviation exceeds 10.
4. Distribution of the outcome: To many 1s. (or 0s).