Numerical Optimization

Miguel Sarzosa

Department of Economics
University of Maryland

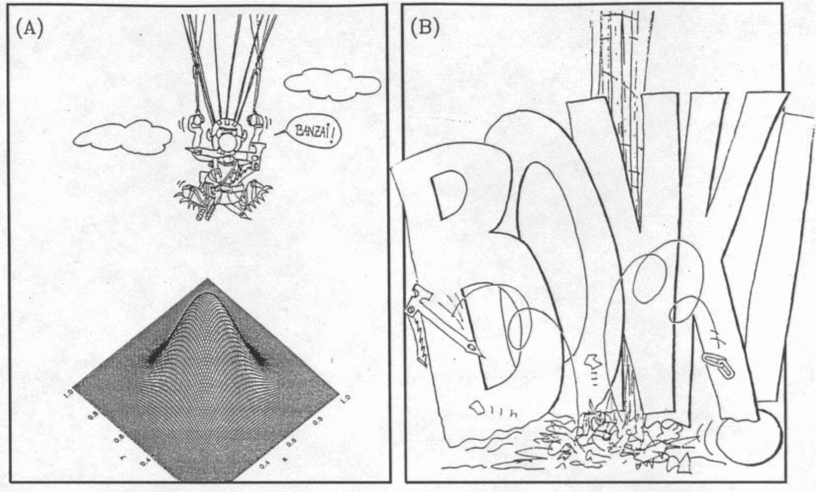
Econ626: Empirical Microeconomics, 2012

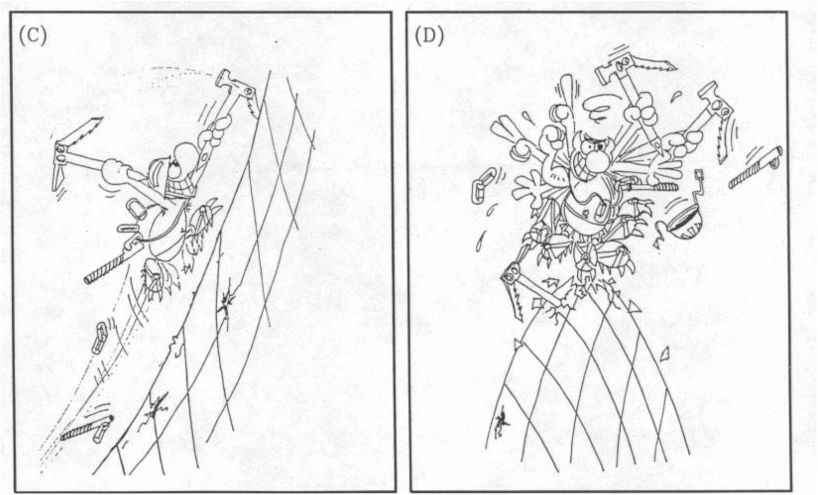
The Problem of the hill climber

Hill Climbing Method: Finding the highest altitude in a 2D landscape

1. Choose a starting location (Choose initial parameters)
2. Determine the steepest uphill direction
3. Move a certain distance in that direction
4. Go on until all surrounding directions are downhill

Numerical optimization methods differ in how they take on steps 1 to 3.


(A)
BANZAI!
(B)
BONK!



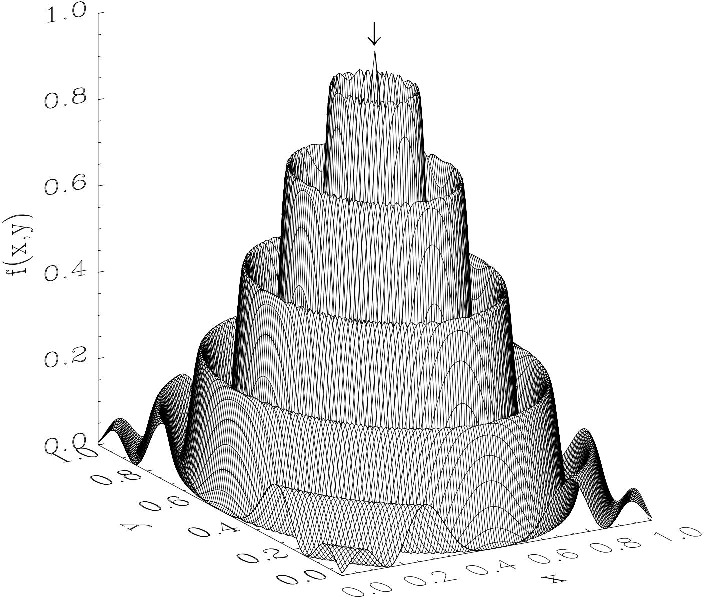
Taken from Charbonneau (2002)

Complications

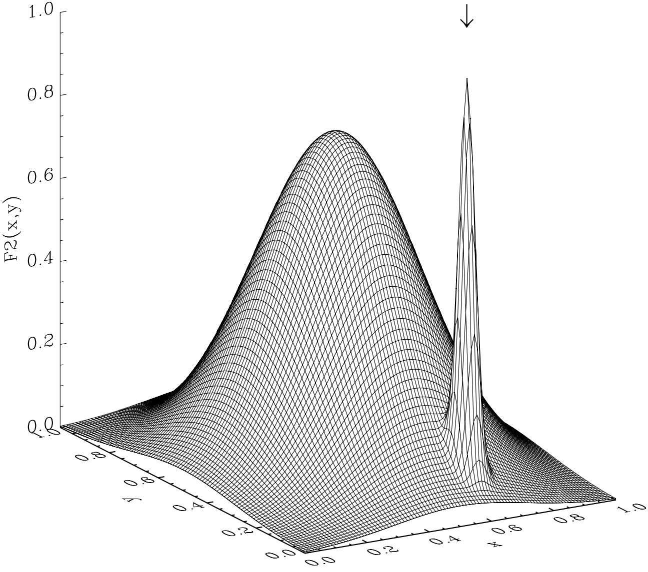
Again our method is:

1. Choose a starting location (Choose initial parameters)
2. Determine the steepest uphill direction
3. Move a certain distance in that direction
4. Go on until all surrounding directions are downhill

It is easy to see that many things can go wrong in our recipe of climbing the mountain



Taken from Charbonneau (2002)



Taken from Charbonneau (2002)

What climbing the hill looks like in math

- Our problem is always find $\hat{\theta} = \arg\max Q_N(\theta)$ where $Q_N(\cdot)$ is a given objective function.
- Usually $Q_N^{'}(\theta) = 0$ has an analytical solution.
- In nonlinear applications, this is not the case. Then we need a way to implement the sequence 2-3 explained above $\Rightarrow$**Iterative Methods**
- In Iterative Methods you have an step s and there is a rule that yields where to find $\hat{\theta}_{s+1}$, where ideally $Q_N\left(\hat{\theta}_{s+1}\right) > Q_N\left(\hat{\theta}_s\right)$

Gradient Methods

- Most iterative methods are gradient methods.
- The derivative tells them where to go

$$\hat{\theta}_{s+1} = \hat{\theta}_s + \mathbf{A}_s \mathbf{g}_s \tag{1}$$

where $\mathbf{A}_s = A\left(\hat{\theta}_s\right)$ and $\mathbf{g}_s = \frac{\partial Q_N(\theta)}{\partial \theta}\Big|_{\hat{\theta}_s}$

- Different methods use different $\mathbf{A}_s$
- What is a natural $\mathbf{A}_s$?
  - Answer: The Hessian (Newton-Raphson)

Simple Example

Consider the exponential regression

$$Q_N(\theta) = -(2N)^{-1} \sum_{i=1}^{N} \left(y_i - e^{\theta}\right)^2$$

It is easy to see that the gradient becomes

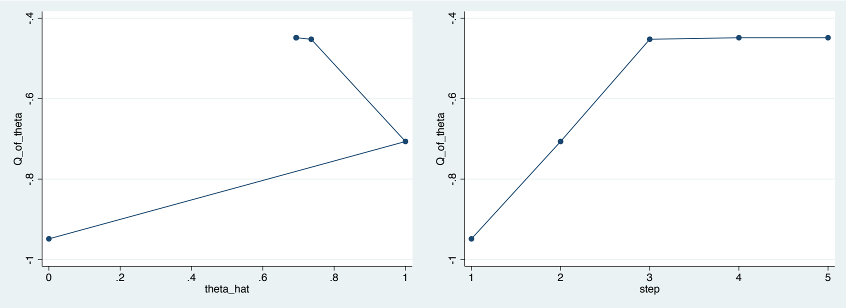
$$g = N^{-1} \sum_{i=1}^{N} \left(y_i - e^{\theta}\right) e^{\theta} = \left(\bar{y} - e^{\theta}\right) e^{\theta}$$

Suppose $A_s = e^{-2\theta}$. Then following (1)

$$\hat{\theta}_{s+1} = \hat{\theta}_s + e^{2\theta} \left(\bar{y} - e^{\theta}\right) e^{\theta} = \hat{\theta}_s + \left(\bar{y} - e^{\theta}\right) e^{-\theta}$$

Suppose $\bar{y} = 2$ and $\hat{\theta}_1 = 0$. Then,
$\hat{\theta}_2 = 1$ and $g_1 = 1 \to \hat{\theta}_3 = 1 + (2 - e)\, e^{-1}$ and $g_2 = (2 - e)\, e^{-1}$. And so on....

Figure: $Q(\theta)$ and $\hat{\theta}_s$; $Q(\theta)$ and the Iteration s



Convergence

Iterations will continue forever if we do not define some criteria

1. A small change in $Q_N\left(\hat{\theta}_s\right)$
2. A small change in $\mathbf{g}_s$ relative to the Hessian
3. A small change in parameter estimates $\hat{\theta}_s$

Convergence is often 10^{-6}

Initial Values (1. Starting Location)

- Iterations needed to reach hill top reduce if initial values are chosen to be close to θ^*
- A poor initial values choice can lead to failure
- Stata chooses 20 places in the parameter space at random

Derivatives (2. Determine the steepest uphill direction)

$$\frac{\Delta Q_N\left(\hat{\theta}_s\right)}{\Delta \theta_j} = \frac{Q_N\left(\hat{\theta}_s + h\mathbf{e}_j\right) + Q_N\left(\hat{\theta}_s - h\mathbf{e}_j\right)}{2h}$$

where $\mathbf{e}_j = (0, 0, \ldots, 0, 1, 0, \ldots, 0)^{'}$ and h should be very small

- Computer calculates them
- Drawback is that they can be computationally burdensome as the number of parameters increases (parameter space dimensions).
- Advantage: no coding
- Alternative: Analytical derivatives provided by user
  - Analytical derivatives reduce the computational work and make easier the computation of the second derivatives (Hessian).
  - Users can also provide second analytical derivatives

Newton-Raphson Method

$$\hat{\theta}_{s+1} = \hat{\theta}_s + \mathbf{H}_s^{-1}\mathbf{g}_s$$

where $\mathbf{H}_s = \left.\frac{\partial^2 Q_N(\theta)}{\partial\theta\partial\theta'}\right|_{\hat{\theta}_s}$ is of dimension $q \times q$

Motivation: From the Taylor approximation around $\hat{\theta}_s$

$$Q_N^*(\theta) = Q_N\left(\hat{\theta}\right) + \mathbf{g}_s^{'}\left(\theta - \hat{\theta}_s\right) + {}^1\!/_2\left(\theta - \hat{\theta}_s\right)^{'}\mathbf{H}_s\left(\theta - \hat{\theta}_s\right)$$

To find the optimal θ^* of this Taylor expansion we calculate the derivative with respect to θ which yields

$$\mathbf{g}_s + \mathbf{H}_s\left(\theta - \hat{\theta}_s\right) = 0$$

Solving for $\theta \Rightarrow \theta = \hat{\theta} - \mathbf{H}_s^{-1}\mathbf{g}_s$

Note that $\mathbf{H}_s$ has to be non-singular

BHHH and DFP

$$\mathbf{H}_{BHHH,s} = -\sum_{i=1}^{N} \frac{\partial q_i(\theta)}{\partial \theta} \frac{\partial q_i(\theta)}{\partial \theta'} \bigg|_{\hat{\theta}_s}$$

Note that we only require to calculate the first derivatives. Less burdensome.

$$\mathbf{A} = \mathbf{A}_{s-1} + \frac{\delta_{s-1}\delta_{s-1}^{'}}{\delta_{s-1}\gamma_{s-1}} + \frac{\mathbf{A}_{s-1}\gamma_{s-1}\gamma_{s-1}^{'}\mathbf{A}_{s-1}}{\gamma_{s-1}^{'}\mathbf{A}_{s-1}\gamma_{s-1}}$$

where $\delta_{s-1} = \mathbf{A}_{s-1}\mathbf{g}_{s-1}$ and $\gamma_{s-1} = \mathbf{g}_s - \mathbf{g}_{s-1}$

Prepare for the Example (Poisson Model)

A Poisson model optimizes the following objective function

$$Q(\theta) = \sum_{i=1}^{N} \left[-e^{\mathbf{x}_i^{'}\theta} + y_i \mathbf{x}_i^{'}\theta - \ln y_i! \right]$$

It is easy to see that the gradient and the Hessian are

$$\mathbf{g}(\theta) = \sum_{i=1}^{N} \left[y_i - e^{\mathbf{x}_i^{'}\theta} \right] \mathbf{x}_i$$

$$\mathbf{H}(\theta) = \sum_{i=1}^{N} -e^{\mathbf{x}_i^{'}\theta} \mathbf{x}_i \mathbf{x}_i^{'}$$

Using the Newton-Raphson method

$$\hat{\theta}_{s+1} = \hat{\theta}_s + \left[\sum_{i=1}^{N} e^{\mathbf{x}_i^{'}\theta} \mathbf{x}_i \mathbf{x}_i^{'} \right]^{-1} \sum_{i=1}^{N} \left[y_i - e^{\mathbf{x}_i^{'}\theta} \right] \mathbf{x}_i$$

Now we go to Stata!