Numerical optimization I

Woojoo Lee

Our problem is to find

$$x^* = \arg\max F(x)(\quad \text{or} \quad \arg\min F(x))$$

for a smooth function $F : R^n \to R$.

Example

Find x minimizing

$$F(x) = 2x^2 + 3x + 1$$

Find x minimizing

$$F(x) = (1/2)x^T Px + q^T x + r$$

where P is positive definite, $q \in R^n$ and $r \in R$.

Optimization problem $\rightarrow$ solving linear or nonlinear equations

For example,
a quadratic optimization problem $\rightarrow$ solving a linear equation

We will try to find

$$x^* \in R^n \text{ satisfying } f(x^*) = 0.$$

where $f = \nabla F$. x^* is called a zero of $f(\cdot)$.

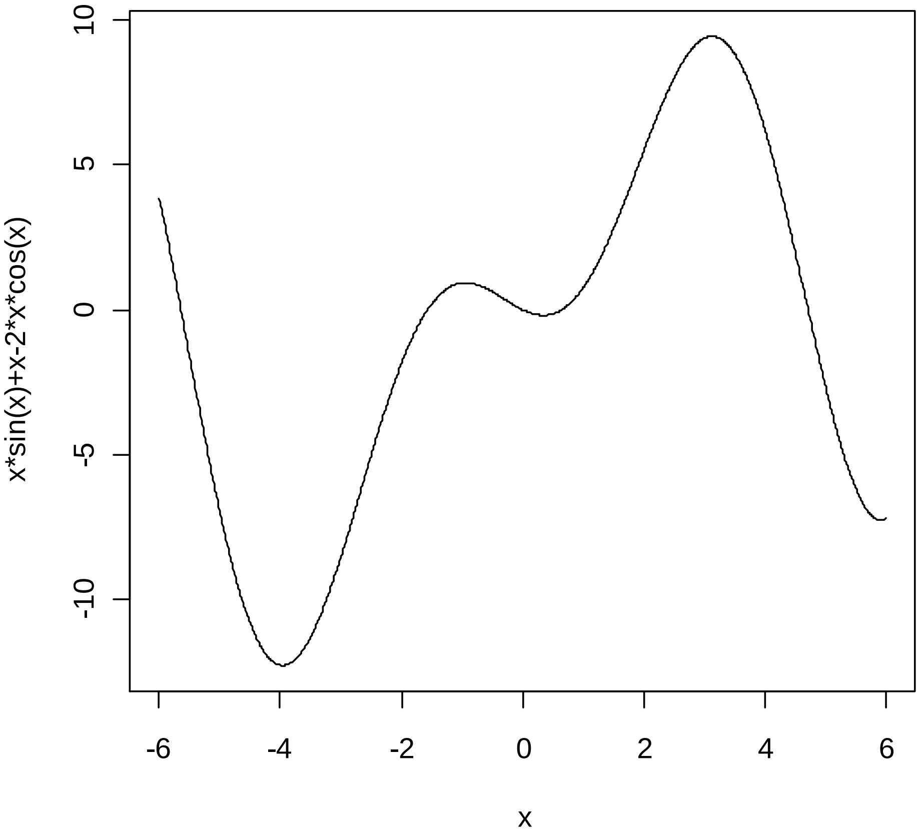
Example)

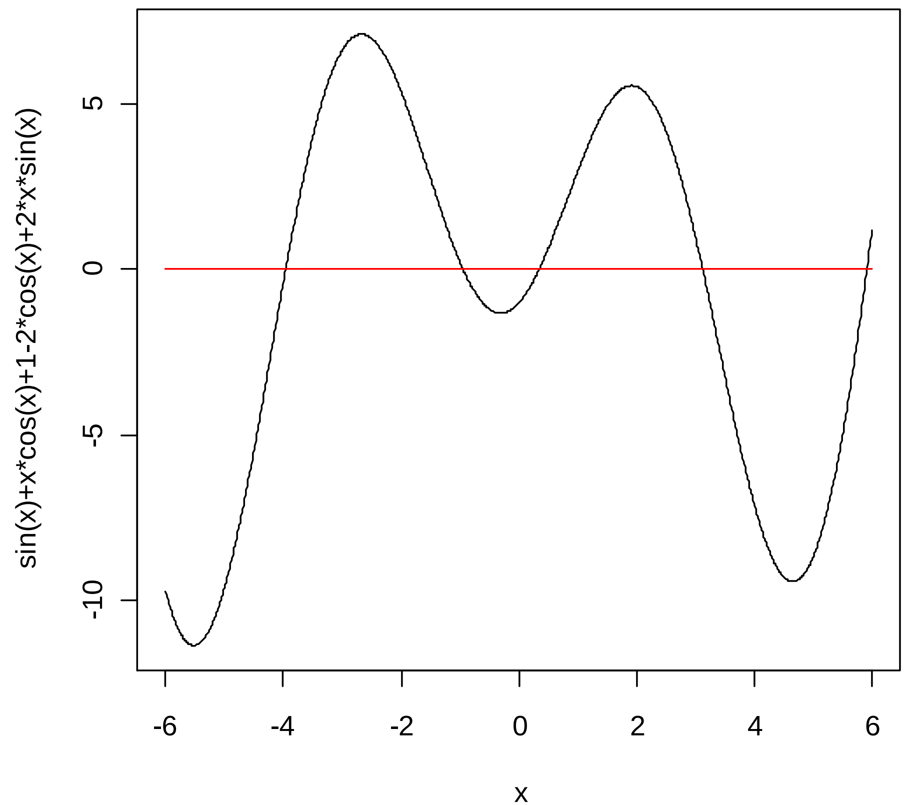
$$F(x) = x\sin(x) + x - 2x\cos(x)$$

where $x \in (-6, 6)$.

Generally, it is difficult to obtain an analytic solution.
$\rightarrow$ Numerical approach is required. In particular, iterative methods are popular.
$\rightarrow$ Every algorithm has its own pros and cons.

Example (continued)





We will study some numerical iteration methods for solving nonlinear equations.

Generate $k+1$-th iterate x_{k+1} from k-th iterate x_k. The iteration will stop if $|f(x_k)|$ is less than a specified threshold.

Remark) Note that your equation can have multiple solutions !

Some required mathematical knowledge to understand the following methods

- Basic calculus
- Differentiation
- Taylor expansion

Bisection method

Theorem: If $f(\cdot)$ is continuous on $[a, b]$, and $f(a)f(b) < 0$, then there exists at least one $x^* \in [a, b]$ for which $f(x^*) = 0$.

The bisection method shrinks the interval from $[a_0, b_0]$ to $[a_1, b_1] \supset \cdots \supset [a_n, b_n]$.

1. $x^0 = \frac{a_0+b_0}{2}$

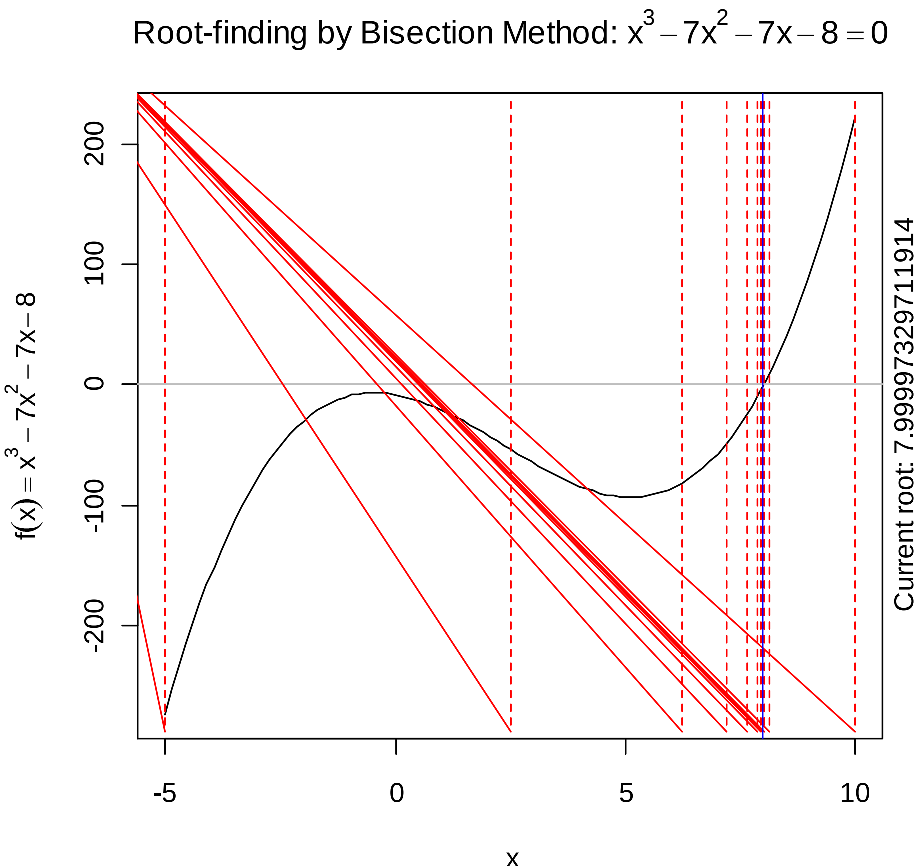
2.
$$[a_{t+1}, b_{t+1}] = \begin{array}{l} [a_t, x^{(t)}] \text{ if } f(a_t)f(x^{(t)}) < 0 \\ [x^{(t)}, b_t] \text{ if } f(a_t)f(x^{(t)}) > 0 \end{array}$$

3. $x^{t+1} = \frac{a_{t+1}+b_{t+1}}{2}$

Remark) $\lim_{t\to\infty} f(a_t)f(b_t) \leq 0$

Some animation for Bisection method

Example) $x^3 - 7x^2 - 7x - 8 = 0$ (with animation)



- Advantage: intuitive, simple, very weak assumption on f
- Disadvantage: often slow, limited when $f(a)f(b) > 0$, difficult in high-dimension

Fixed-point iteration method

Key idea: transform the problem of finding a root of $f(\cdot)$ into a problem of finiding a fixed point of $\phi(\cdot)$. ($\phi(\cdot)$ is determined by $f(\cdot)$.)

Fixed point: x^* satisfying $x^* = \phi(x^*)$ (easy to understand graphically)

$\rightarrow$ The simplest way to find a fixed point is to use:

$$\rightarrow x_{k+1} = \phi(x_k)$$

Consider $f(x) = x^2 - 10x + 21 = 0$. We know that roots are 3 and 7.

- $\phi(x) = x^2 - 9x + 21 = x$
- $\phi(x) = 10 - 21/x = x$
- $\phi(x) = (x^2 + 21)/10 = x$

Q) Do all $\phi(x)$ give the same solution ?

```
> ## x^2-9*x+21= x
> x<-rep(2.5,10)
> for (i in 1:10){
+ x[i+1]<-x[i]^2-9*x[i]+21
+ print(x[i+1])
+ }
[1] 4.75
[1] 0.8125
[1] 14.34766
[1] 97.72633
[1] 8691.899
[1] 75470907
[1] 5.695857e+15
[1] 3.244279e+31
[1] 1.052535e+63
[1] 1.107829e+126
```

```
> ## x= 10-21/x
> x<-rep(2.5,30)
> for (i in 1:30){
+ x[i+1]<-10-21/x[i]
+ print(x[i+1])
+ }
[1] 1.6
[1] -3.125
[1] 16.72
[1] 8.744019
[1] 7.598358
[1] 7.236245
[1] 7.097942
[1] 7.041396
[1] 7.017637
[1] 7.00754
[1] 7.003228
[1] 7.001383
[1] 7.000592
[1] 7.000254
[1] 7.000109
[1] 7.000047
[1] 7.00002
[1] 7.000009
[1] 7.000004
[1] 7.000002
[1] 7.000001
[1] 7
[1] 7
```

```
> ## x= (x^2+21)/10
> x<-rep(2.5,30)
> for (i in 1:30){
+ x[i+1]<-(x[i]^2+21)/10
+ print(x[i+1])
+ }
[1] 2.725
[1] 2.842563
[1] 2.908016
[1] 2.945656
[1] 2.967689
[1] 2.980718
[1] 2.988468
[1] 2.993094
[1] 2.995861
[1] 2.997518
[1] 2.998512
[1] 2.999107
[1] 2.999464
[1] 2.999679
[1] 2.999807
[1] 2.999884
[1] 2.999931
[1] 2.999958
[1] 2.999975
[1] 2.999985
[1] 2.999991
[1] 2.999995
[1] 2.999997
[1] 2.999998
[1] 2.999999
[1] 2.999999
[1] 3
```

The convergence of fixed-point iteration depends on whether ϕ is contractive or not. We say ϕ is contractive on $[a, b]$ if

$$\phi(x) \in [a, b]$$

whenever $x \in [a, b]$ and

$$|\phi(x) - \phi(y)| \leq L|x - y| \quad \text{(Lipschitz condition)}$$

for all $x, y \in [a, b]$ and some $L \in (0, 1)$.

$\rightarrow$ If ϕ is contractive on $[a, b]$, then there exists a unique fixed point x^* in $[a, b]$ and the convergence does not depend on the choice of a starting point.

Existence:

Uniqueness:

Convergence:

Local Convergence theorem: Let x^* be the fixed point of $\phi(x)$. Assume that $\phi(\cdot)$ is continuously differentiable near x^* and $|\phi'(x^*)| < 1$. Then,if the initial point is sufficiently close to x^*, the sequence $\{x_k\}$ converges to x^*.

Case 1) $|\phi'(x^*)| > 1$

Case 2) $|\phi'(x^*)| < 1$

Newton-Raphson method

Understanding Newton-Raphson method graphically

1. Consider a linear approximation of $f = \nabla F$.
2. Finding the point crossing the x-axis.
3. Repeat the the above steps until convergence.

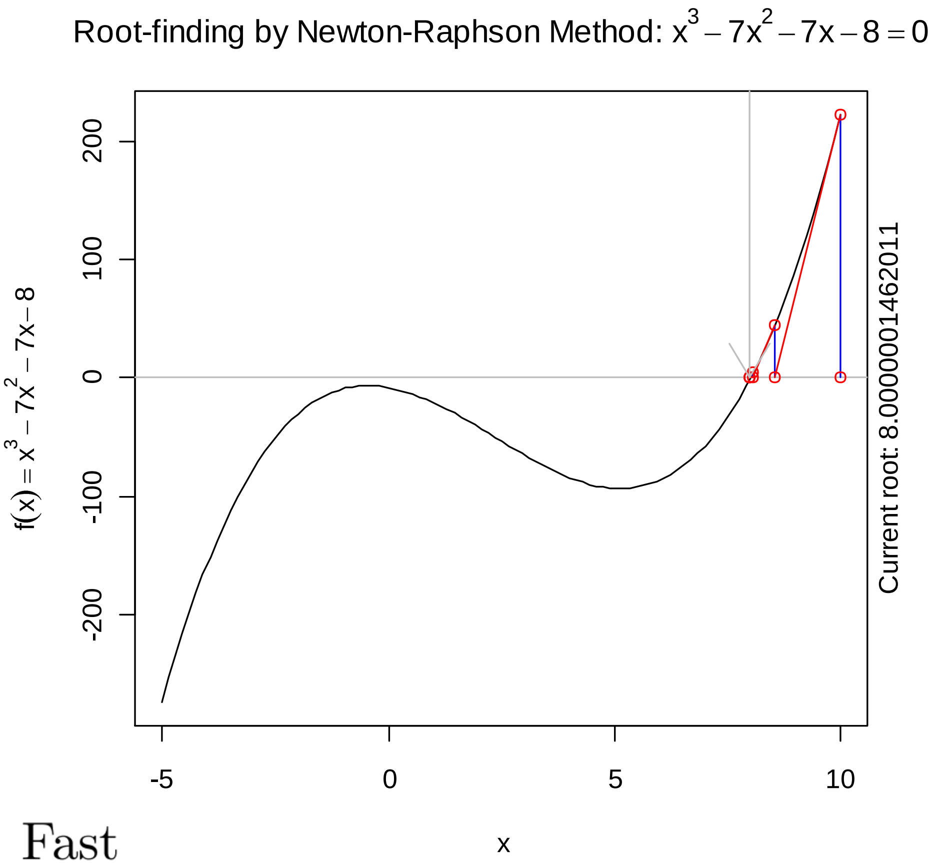
$\rightarrow$

$$x_{k+1} = x_k - f(x_k)/f'(x_k)$$

where $f'(x_k)$ is not zero.

Some animation for NR method

Example) $x^3 - 7x^2 - 7x - 8 = 0$ (with animation)



- Advantage: Fast
- Disadvantage: dependence on initial values, difficult to compute Hessian matrix

Secant method - think the difficulty of computing Hessian !

NR algorithm for general cases

Consider n equations in n variables:

$$\begin{aligned} f_1(x_1, \ldots, x_n) &= 0 \\ f_2(x_1, \ldots, x_n) &= 0 \\ &\vdots \\ f_n(x_1, \ldots, x_n) &= 0 \end{aligned}$$

This is simply denoted by $f(x) = 0$ where $x \in R^n$ and $f : R^n \to R^n$. Here, we assume that the functions $f_i(\cdot)$ are differentiable.

The derivative matrix $(Df(x))$of f at x is given by

$$Df(x) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_n} \end{pmatrix}$$

where the partial derivative functions are evaluated at x.

First-order Taylor expansion at $x = x_0$ (linearization)

$$f(x) \approx f(x_0) + Df(x_0)(x - x_0)$$

Q) Find the linear approximation for $f(x)$ at $x = 0$:

$$f(x) = \begin{pmatrix} \exp(2x_1 + x_2) - x_1 \\ x_1^2 - x_2 \end{pmatrix}$$

Numerical iteration

$$x_{k+1} = x_k - (Df(x_k))^{-1} f(x_k)$$

until $f(x_k)$ becomes sufficiently small.

Convergence rate (or convergence speed)

Let the limit of $\{x_k\}$ be x^* satisfying $f(x^*) = 0$, and let $e_k = x_k - x^*$. We say that a method has convergence of order r if

$$\lim_{k\to\infty} ||e_k|| \to 0$$

and

$$\lim_{k\to\infty} \frac{||e_{k+1}||}{||e_k||^r} = C$$

for some constant $C \neq 0$.

- $r = 1$: linear rate of convergence
- $r > 1$: superlinear rate of convergence
- $r = 2$: second order convergence

Remark) "Convergence of order r" can be defined in various ways. For example, in some other books, quadratice convergence is defined as

$$||e_{k+1}|| \leq C||e_k||^r$$

for some constant $C > 0$ and sufficiently large k.

What is the convergence rate of the Bisection method ?

Remark) Rigorously speaking, the bisection method does not formally meet the definition for determining order of convergence.

What is the convergence rate of Newton-Raphson method ?

Q) Does the secant method have the same order of convergence as Newton-Raphson method ?