LECTURE 7: SOLUTION METHODS FOR UNCONSTRAINED OPTIMIZATION

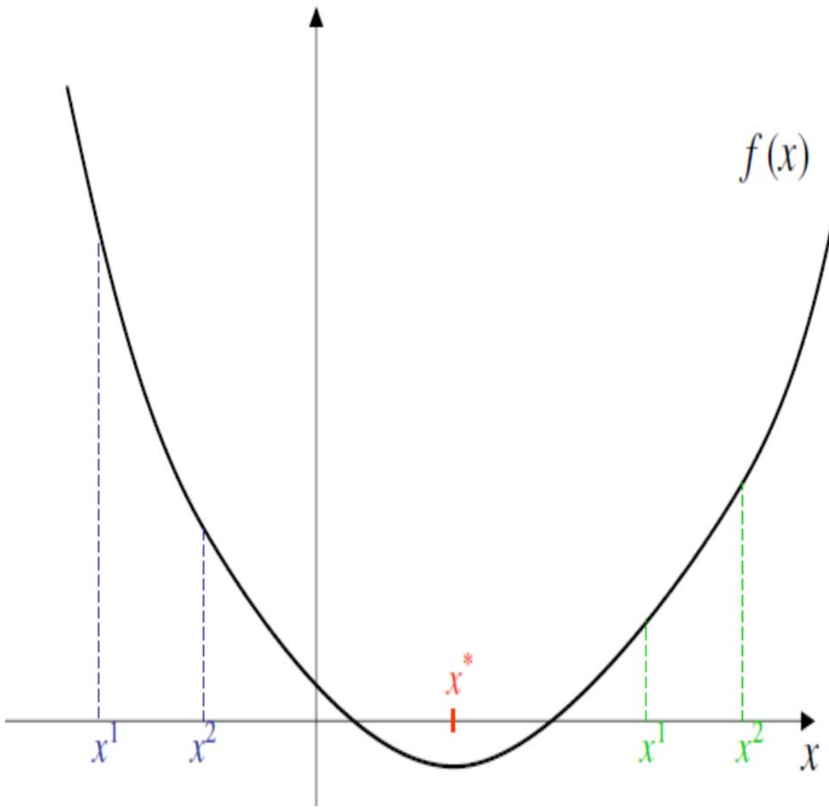
1. Line search – one dimensional
2. Curve fitting - one dimensional Newton’s method
3. Descent method - multidimensional

One dimensional search

- One dimensional search = Line search

- Definition:

  A <u>unimodal</u> function is a function that has only one valley (local minimum) in a given interval.



What's a line search?

- Given a unimodal function

$$f(x) : [x_s, x_f] \rightarrow R$$

with a minimum x^* .

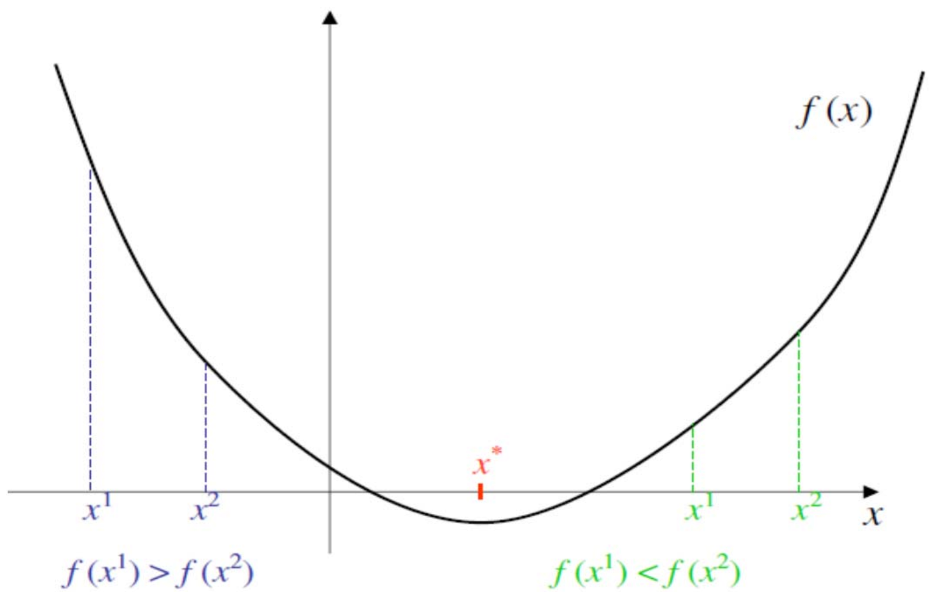
- A line search is a computational procedure with

<u>**Input**</u> : $f(\cdot)$ and $[x_s, x_f]$.

<u>**Output**</u> : a final interval $[x', x'']$ with $x^* \in [x', x'']$.

Observations

- We don't require continuity, differentiability, or any “analytic expression" of the function *f.*
- Evaluation of *f(.)* at a point *x* could be costly.



Performance of a line search

- **Challenge:**

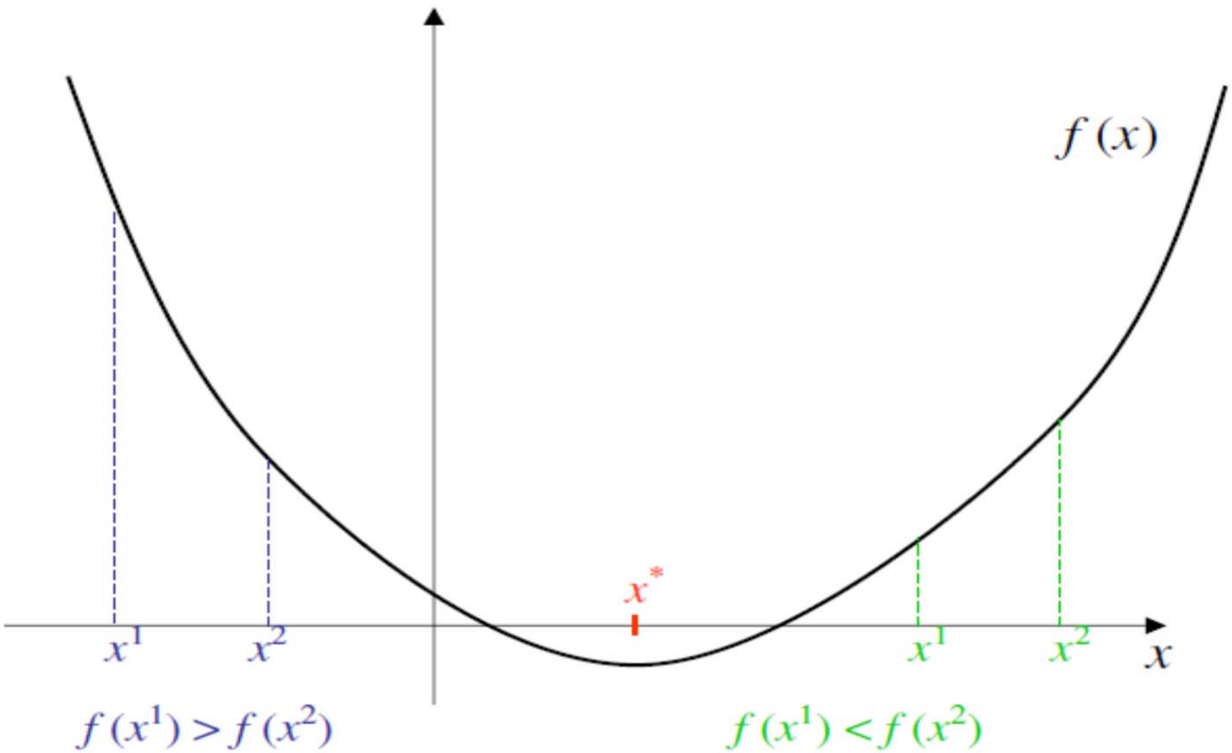
  Come up with a final interval $[x', x'']$, as small as possible, by evaluating $f(\cdot)$ for as few times as possible.

- **Criteria:**

  (1) Same number of evaluations, a smaller interval is preferred.

  (2) Same size of intervals, fewer number of evaluations is preferred.

What will you do?



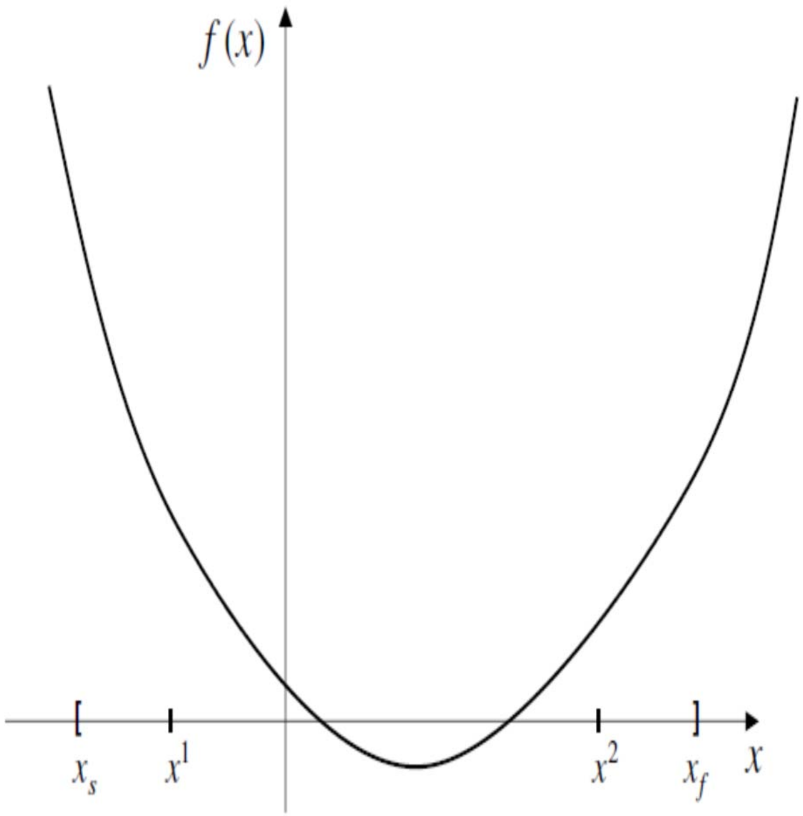
Line search by elimination

- Take two evaluations at x^1, x^2 with $x_s < x^1 < x^2 < x_f$.

(1) If $f(x^1) < f(x^2)$, then $[x^2, x_f]$ can be discarded.

(2) If $f(x^1) > f(x^2)$, then $[x_s, x^1]$ can be discarded.

(3) If $f(x^1) = f(x^2)$, then both $[x_s, x^1]$ and $[x^2, x_f]$ can be discarded.



Question

Let $f(\cdot)$ be a unimodal function on a given interval $[x_s, x_f]$. If we are willing to make n evaluations of $f(\cdot)$, where should we place them in order to get a final interval $[x', x'']$ as small as possible?

Observation:

WLOG, we may consider $n = 2k,\ k = 1, 2, \cdots$.

Exhaustive search

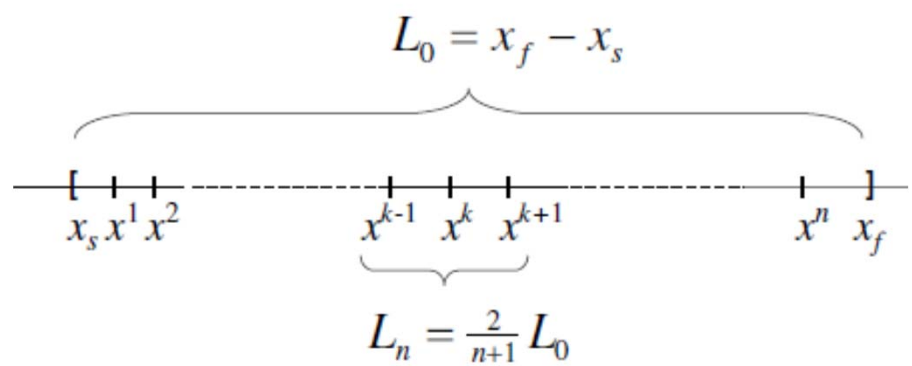
Evaluating the objective function at a predetermined n equally spaced points in the interval (x_s, x_f), $x_s < x^1 < \ldots < x^n < x_f$.

1. If the minimum value among the n function values is x^k, then the final interval of uncertainty is $[x^{k-1}, x^{k+1}]$ with length of

$$L_n = x^{k+1} - x^{k-1} = \frac{2}{n+1} L_0$$

where $L_0 = x_f - x_s$.

2. This is a simultaneous search method and is relatively inefficient.



n points / $n+1$ subintervals

$$f(x^k) = \min_{i=1,2,\ldots,n} \{f(x^i)\}$$

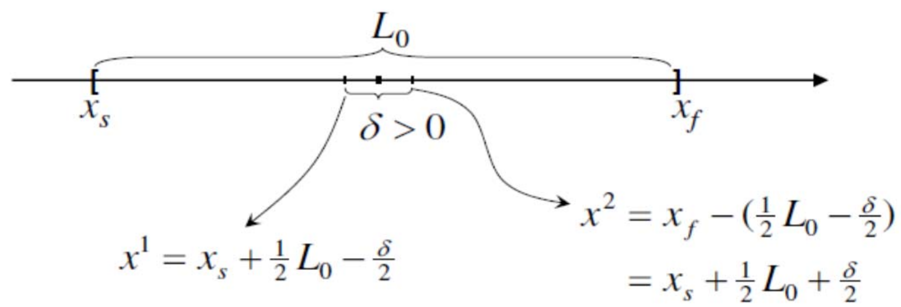
$$[x', x''] = [x^{k-1}, x^{k+1}]$$

Dichotomous (Bi-section) search

1. Sequentially evaluating $f(\cdot)$ at n $(= 2k)$ points in (x_s, x_f).

2. Let the interval of uncertainty after n evaluations $(n = 0, 2, \ldots)$, be $[x_s^{[n]}, x_f^{[n]}]$ with length $L_n = x_f^{[n]} - x_s^{[n]}$, where $x_s^{[0]} = x_s$ and $x_f^{[0]} = x_f$.

3. Evaluate function values at two new positions, $x_s^{[n]} + \frac{L_n}{2}$ - $\frac{\delta}{2}$ and $x_s^{[n]} + \frac{L_n}{2} + \frac{\delta}{2}$, where δ is a small positive number chosen such that the two evaluations give different results.



4. Based on the relative values of the objective function at the two points, using the unimodality assumption to eliminate almost half of the interval of uncertainty.

5. The final interval of uncertainty after conducting n experiments is

$$L_n = \frac{L_0}{2^{n/2}} + \delta(1 - \frac{1}{2^{n/2}}).$$

Observations

(1) Each time with 2 evaluations at x^1 and x^2, we cut at least $\frac{1}{2}L_0 - \frac{\delta}{2}$. The residue is at most $\frac{1}{2}L_0 + \frac{\delta}{2}$.

(2) To speed up the geometric reduction, smaller $\delta > 0$ is preferred.

(3) When δ is too small, $x^1 \approx x^2$. This may cause some precision problems in false evaluation.

(4) Can we separate x^1 and x^2 apart safely but not sacrificing much reduction?

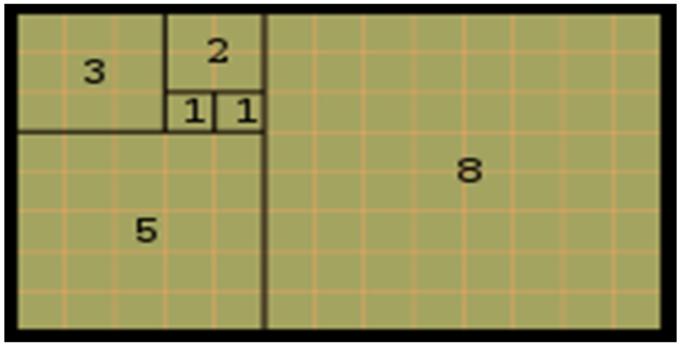
(5) Can the information on x^1 and x^2 be reused?

Have you heard of Fibonacci sequence?

- A sequence { 1, 1, 2, 3, 5, 8, X, XX, XXX, …}
- Questions: X = ?,

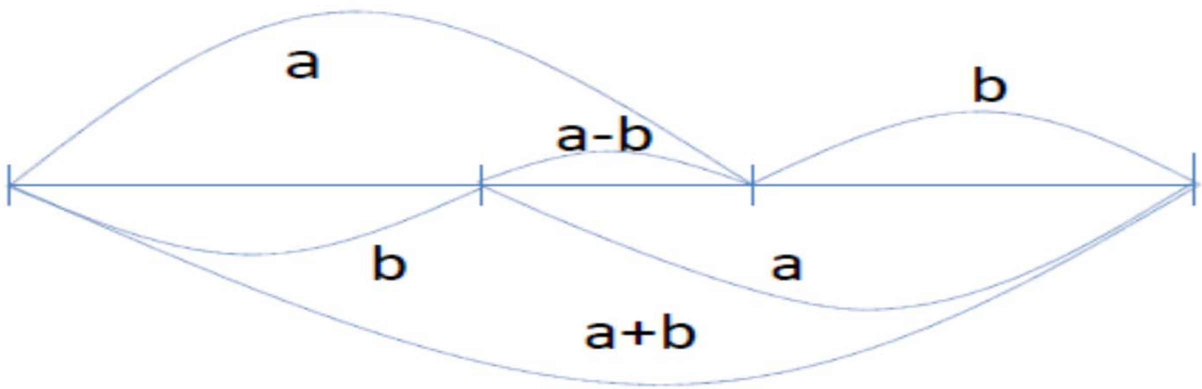
  XX = ?,

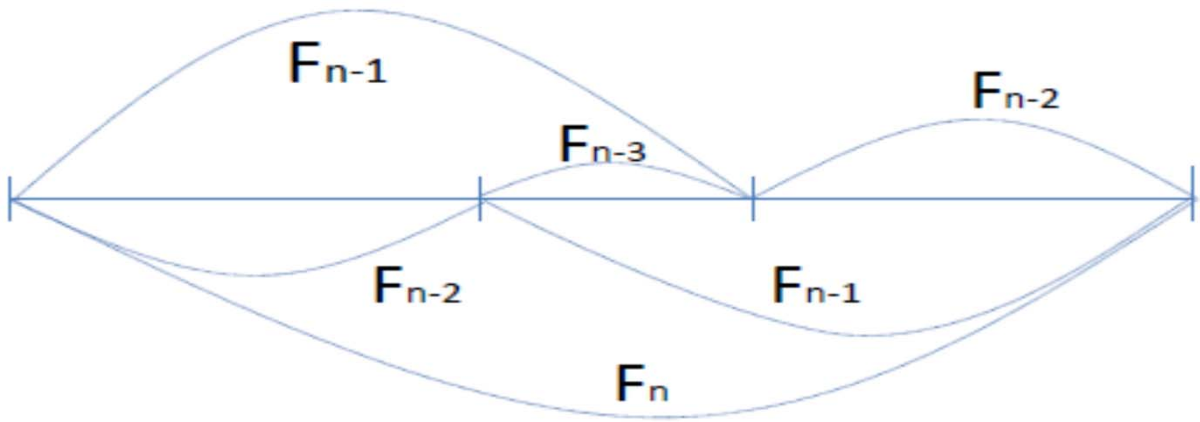
  XXX = ?



- The Fibonacci sequence is named after Leonardo of Pisa, who was known as Fibonacci. Fibonacci's 1202 book *Liber Abaci* introduced the sequence to Western European mathematics, although the sequence was independently described in Indian mathematics and it is disputed which came first.

Fibonacci sequence for search





$F_n = F_{n-1} + F_{n-2}$ $F_{n-1} = F_{n-2} + F_{n-3}$

Fibonacci method

1. The sequence of Fibonacci numbers, $\{F_n\}$,

$$\begin{aligned} F_0 &= F_1 = 1 \\ F_n &= F_{n-1} + F_{n-2}, \quad n = 2, 3, 4, \ldots \end{aligned}$$

$$\Rightarrow \{F_n\} = \{1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, \ldots\}$$

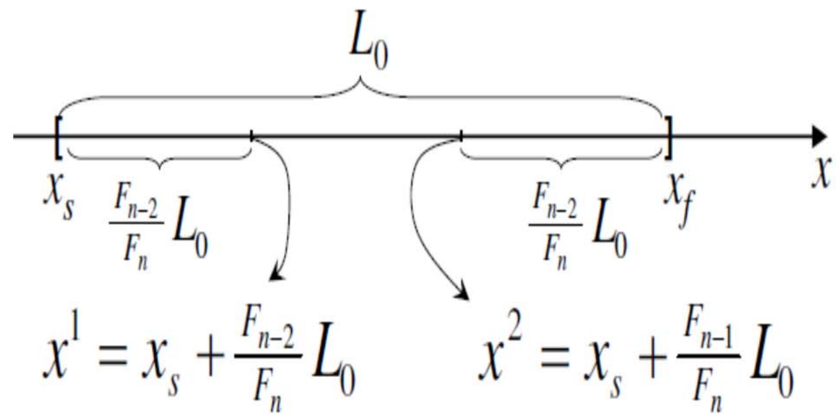
2. Let the initial interval of uncertainty be $[x_s, x_f]$ with length $L_0 = x_f$ - x_s. Let n be the predetermined total number of evaluations to be conducted.

3. Define

$$L_2^* = \frac{F_{n-2}}{F_n} L_0$$

Evaluate objective function values at x^1 and x^2 with

$$\begin{aligned} x^1 &= x_s + L_2^* = x_s + \frac{F_{n-2}}{F_n} L_0 \\ x^2 &= x_f - L_2^* = x_f - \frac{F_{n-2}}{F_n} L_0 = x_s + \frac{F_{n-1}}{F_n} L_0 \end{aligned}$$



4. Discard part of the interval by using the unimodality assumption. The new interval of uncertainty becomes

$$L_2 = L_0 - L_2^* = L_0(1 - \frac{F_{n-2}}{F_n}) = \frac{F_{n-1}}{F_n} L_0$$

5. Either x_1 or x_2 will be left in the new interval of uncertainty L_2 with a distance of

$$L_2^* = \frac{F_{n-2}}{F_n} L_0 = \frac{F_{n-2}}{F_{n-1}} L_2$$

from one end and a distance of

$$L_2 - L_2^* = \frac{F_{n-3}}{F_n} L_0 = \frac{F_{n-3}}{F_{n-1}} L_2$$

from the other end.

Observations

6. Perform one new evaluation at x_3 such that the two evaluations inside L_2 are located at a distance of

$$L_3^* = \frac{F_{n-3}}{F_n} L_0 = \frac{F_{n-3}}{F_{n-1}} L_2$$

from each end of the interval L_2 (symmetric placement).

7. The length of the new resulting interval of uncertainty is

$$L_3 = L_2 - L_3^* = L_2 - \frac{F_{n-3}}{F_{n-1}} L_2 = \frac{F_{n-2}}{F_{n-1}} L_2 = \frac{F_{n-2}}{F_n} L_0$$

The process repeats.

8. In general, we have,

$$L_j^* = \frac{F_{n-j}}{F_{n-(j-2)}} L_{j-1}$$

$$L_j = \frac{F_{n-(j-1)}}{F_n} L_0$$

9. The position of the nth evaluation is always the same as the $(n-1)$th one.

10. The final interval of uncertainty is given by

$$L_n = \frac{F_1}{F_n} L_0 = \frac{1}{F_n} L_0$$

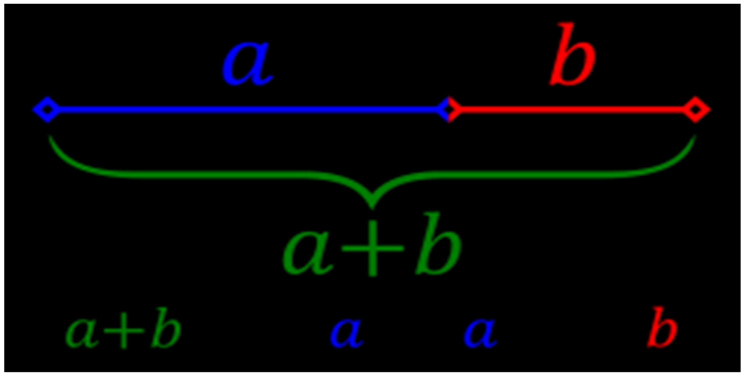
which permits us to determine n before the real process of evaluation.

Question

- Can we simplify this process a little without changing much performance?

- In particular, if the number “*n*" is not predetermined.

Have you heard of golden ratio?



- (a+b) /a = a/b = phi → phi = 1.618033…
- Ancient Greek mathematicians studied the golden ratio because of its frequent appearance in geometry The Greeks usually attributed discovery of this concept to Pythagoras or his followers.

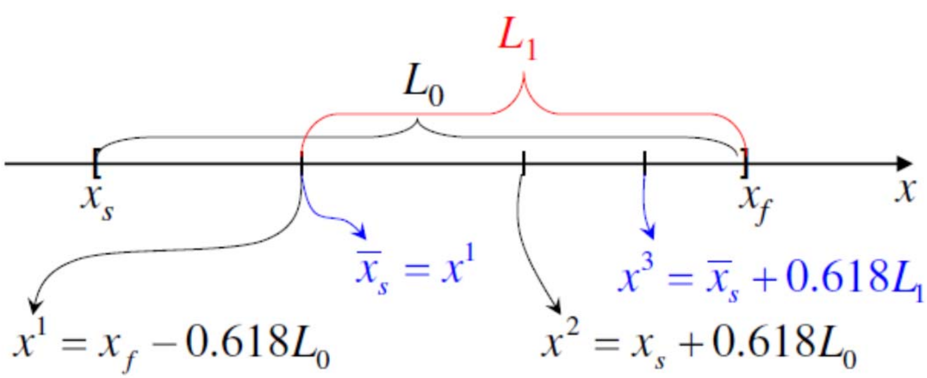
Golden section method

1. The solution to the Fibonacci difference equation
$$F_n = F_{n-1} + F_{n-2}$$
is of the form
$$F_n = A\tau_1^n + B\tau_2^n$$
where
$$\tau_1 = \frac{1+\sqrt{5}}{2} \approx 1.618 \quad \text{and} \quad \tau_2 = \frac{1-\sqrt{5}}{2} \approx -0.618$$
are roots of the characteristic equation
$$\tau^2 = \tau + 1$$

2. When n is large,
$$\frac{F_{n-1}}{F_n} = \frac{1}{\tau_1} \approx 0.618$$

3. We set in the golden section method,
$$L_k = 0.618 L_{k-1}, \qquad k = 1, 2, \ldots$$
with a symmetric placement of the function evaluation.

4. The final interval of uncertainty is
$$L_n = 0.618^{n-1} L_0$$



$$L_1 = 0.618 L_0$$

Comparison of various line search methods

Method	$\frac{1}{2}\frac{L_n}{L_0} \le 0.1$	$\frac{1}{2}\frac{L_n}{L_0} \le 0.01$
Exhaustive search	$n \ge 9$	$n \ge 99$
Dichotomous search	$n \ge 6$	$n \ge 14$
$(\delta = 0.01,\ L_0 = 1)$		
Fibonacci	$n \ge 4$	$n \ge 9$
Golden Section	$n \ge 5$	$n \ge 10$

– Fibonacci method seems to be the best.

– Dichotomous search:

Adding two points to reduce the interval by $\frac{1}{2}$.

– Golden Section:

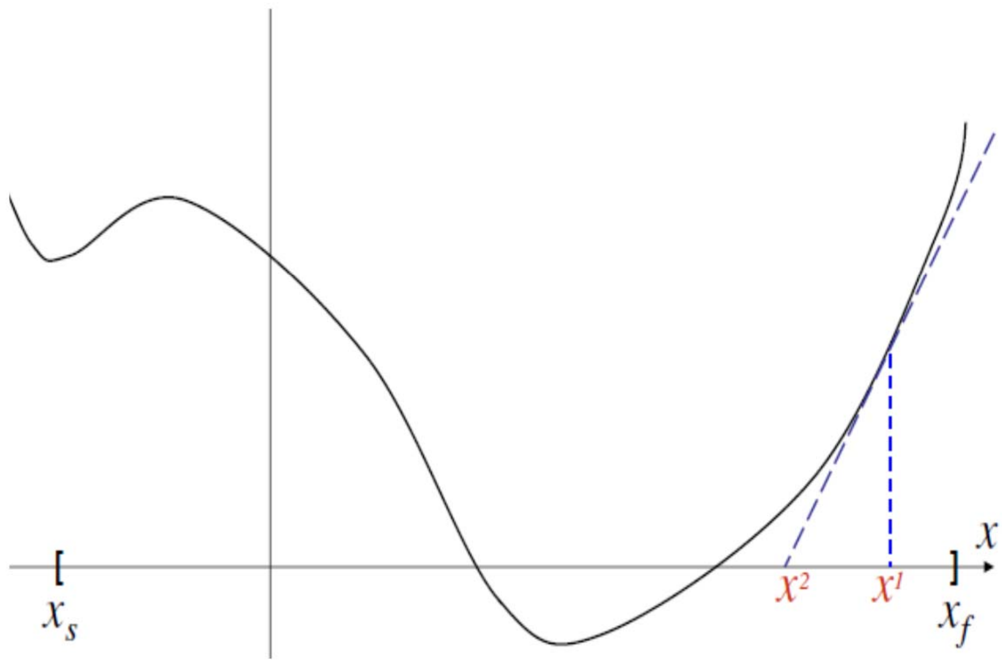
Adding one point to reduce the interval by $(1 - 0.618) = 0.382$.

– Fibonacci:

Adding one point to reduce the interval by $\frac{F_{n-2}}{F_n} = 1 - \frac{F_{n-1}}{F_n} \approx 0.382$.

Line search by curve fitting

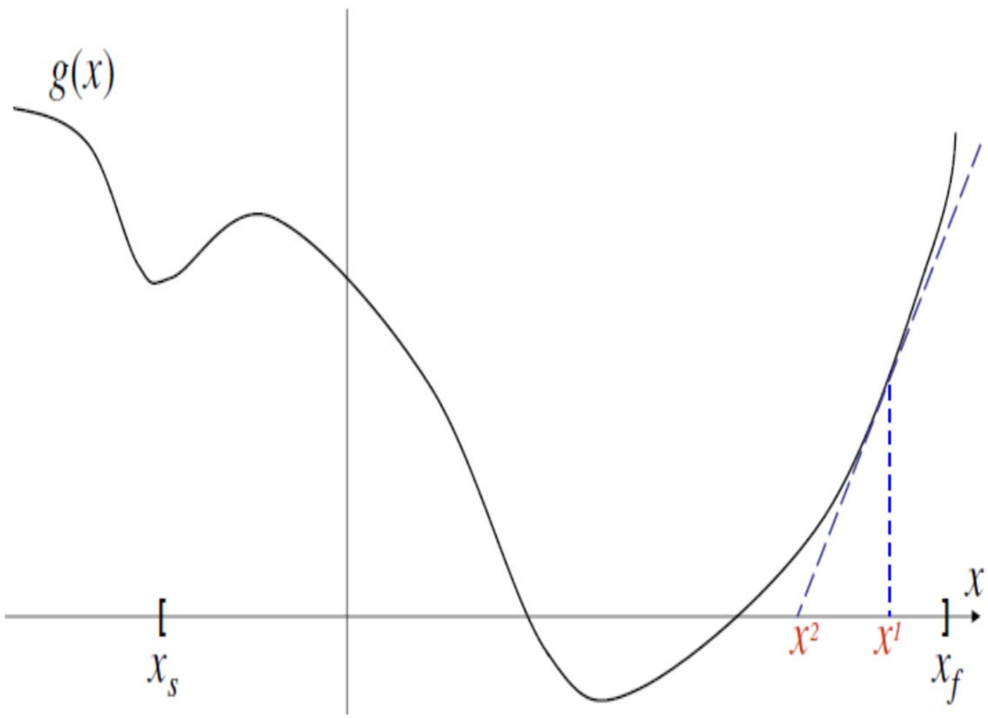
- In addition to the (local) unimodal property, the function *f* is assumed to possess certain degree of smoothness, such as C^2-smooth .



Newton's method – one dimensional case

- Motivation: Given a nonlinear function *g: R → R*,

  Newton's method is to find a solution to *g(x) = 0*.



$$\frac{g(x^1) - \cancelto{0}{g(x^2)}}{x^1 - x^2} \approx g'(x^1)$$

$$x^2 = x^1 - \frac{g(x^1)}{g'(x^1)}$$

Observations

- One dimensional case

  To minimize a C^2-smooth function $f(x)$,
  we consider solving $g(x) \triangleq f'(x) = 0$.
  Hence
  $$x^{k+1} = x^k - \frac{f'(x^k)}{f''(x^k)}.$$

- (It is a “conditional negative gradient” direction with unit step-length.)
- Multi-dimensional case (to be further discussed later)

$$x^{k+1} = x^k - \left[F(x^k)\right]^{-1} \nabla f(x^k)^T.$$

- Newton's method converges locally with an order of convergence p (greater than or equal to 2).

Main theorem: 1-d Newton's method

Theorem:

Let the function g have a continuous second derivative, and let x^* satisfy $g(x^*) = 0$, $g'(x^*) \neq 0$. Then, provided x^0 is sufficiently close to x^*, the sequence $\{x^k\}_{k=0}^{\infty}$ generated by Newton's method converges to x^* with an order of convergence at least two.

Proof

Since $g \in C^2$ and $g'(x^*) \neq 0$, within a sufficiently small neighborhood $\bar{N}(x^*)$, $\exists\ q_1,\ q_2 > 0$ such that $|g''(x)| < q_1$ and $|g'(x)| > q_2$, $\forall x \in \bar{N}(x^*)$.

Notice that $g(x^*) = 0$, we have

$$\begin{aligned} x^{k+1} - x^* &= x^k - x^* - \tfrac{g(x^k) - g(x^*)}{g'(x^k)} \\ &= -\tfrac{\left[g(x^k) - g(x^*) + g'(x^k)(x^k - x^*)\right]}{g'(x^k)}. \end{aligned}$$

By Taylor's theorem, $\exists\ \xi \in L^i(x^k, x^*)$ s.t.

$$x^{k+1} - x^* = -\frac{1}{2}\frac{g''(\xi)}{g'(x^k)}(x^k - x^*)^2.$$

Thus, $|x^{k+1} - x^*| \leq \frac{q_1}{2q_2}|x^k - x^*|^2.$

Also, as $\frac{q_1}{2q_2}|x^k - x^*| < 1$, then

$$|x^{k+1} - x^*| < |x^k - x^*|.$$

Therefore, $\{x^k\} \to x^*$ with an order of convergence $p \geq 2$.

False position method – one dimensional case

- Idea: Use the 1st order information only but at more points.
- Consider the approximation

$$f''(x^k) \approx \frac{f'(x) - f'(x^k)}{x - x^k}.$$

- Utilize known information at x^{k-1},

$$x^{k+1} = x^k - f'(x^k)\left[\frac{x^{k-1} - x^k}{f'(x^{k-1}) - f'(x^k)}\right].$$

- The method of false position converges locally with an order of convergence $\tau \approx 1.618.$

Main result: 1-d false position method

Theorem:

Let the function g have a continuous second derivative, and let x^* satisfy $g(x^*) = 0$, $g'(x^*) \neq 0$. Then, provided x^0 is sufficiently close to x^*, the sequence $\{x^k\}_{k=0}^{\infty}$ generated by the method of false position converges to x^* with order $\tau \approx 1.618$.

Descent methods for unconstrained optimization – multidimensional case

- Key ideas:

1. Reduce objective function value in each iteration.

2. Use the information embedded in the gradient function.

3. Consider the 2nd order information when necessary.

4. Taylor Theorem helps.

Recall - Taylor theorem

Let $f : E^n \to R$, $\quad S \subset E^n$ be open, $\quad f \in C^m(S)$,
$x^1, x^2 \in S$, $\quad x^1 \neq x^2$ and $L(x^1, x^2) \subset S$.

Then $\exists \ \bar{x} = \theta x^1 + (1-\theta)x^2 \in L^i(x^1, x^2)$ s.t.

$$f(x^2) = f(x^1) + \sum_{k=1}^{m-1} \frac{1}{k!} d^k f(x^1, x^2 - x^1)$$

$$+ \frac{1}{m!} d^m f(\bar{x}; x^2 - x^1)$$

where $d^k f(x;t)$ is the k-th order differential of function f along t.

Known facts

- $f \in C^1$

$$f(x^2) = f(x^1) + \nabla f(\bar{x})(x^2 - x^1)$$

- $f \in C^2$

$$f(x^2) = f(x^1) + \nabla f(x^1)(x^2 - x^1)$$
$$+\frac{1}{2}(x^2 - x^1)^T F(\bar{x})(x^2 - x^1)$$

When $x \approx x^1$

$$f(x) \approx f(x^1) + \sum_{k=1}^{m-1} \frac{1}{k!} d^k f(x^1; x - x^1)$$

Take $m = 2$

$$f(x) \approx f(x^1) + \nabla f(x^1)(x - x^1)$$

Assume $\nabla f(x^1) \neq 0$.

- Take $x - x^1 = \nabla f(x^1)$, i.e., moving from x^1 in the gradient direction at x^1

$$f(x) \approx f(x^1) + \|\nabla f(x^1)\| > f(x^1)$$

- For $x - x^1 = -[\nabla f(x^1)]$, i.e., moving from x^1 in the negative gradient direction

$$f(x) \approx f(x^1) - \|\nabla f(x^1)\| < f(x^1)$$

- For any $d \triangleq x - x^1$

$$\nabla f(x^1)(x - x^1) = \|d\| \underbrace{\|\nabla f(x^1)\| \cos\theta}_{\text{projection of } \nabla f(x^1) \text{ onto } d}$$

Gradient method

- Step 0. Select a very small $\epsilon > 0$ for being used in the stopping criterion. Start at an arbitrary initial point x^0 and set $k = 0$.
- Step 1. Optimality check. If

$$\|\nabla f(x^k)\| \leq \epsilon$$

  stop and $x^* = x^k$; otherwise go to Step 2.
- Step 2. Updating procedure.

$$x^{k+1} = x^k - \alpha_k g_k$$

  where the n-dimensional column vector $g_k = \nabla f(x^k)^T$ and α_k is a nonnegative scalar minimizing $f(x^k - \alpha g_k)$. Set $k = k + 1$. Go back to Step 1.

Observations

1. Gradient method is sometimes called the method of steepest descent.
2. At a current point x^k, the method searches along the direction of the negative gradient g_k to find a minimum point on this line for x^{k+1}.
3. The "global convergence theorem" applies here!

Minimizing a convex quadratic function

- Minimize

$$f(x) = \frac{1}{2} x^T Q x - x^T b$$

where Q is a symmetric positive definite $n \times n$ matrix.

Gradient method

1. Setting $\nabla f(x) = 0$, $f(x)$ has a unique minimum solution x^* such that

$$Qx^* = b.$$

2. Define

$$g_k \triangleq \nabla f(x^k)^T = Qx^k - b.$$

3. $f(x^k - \alpha g_k) = \frac{1}{2}(x^k - \alpha g_k)^T Q(x^k - \alpha g_k) - (x^k - \alpha g_k)^T b$

has a minimizer (in α)

$$\alpha_k = \frac{g_k^T g_k}{g_k^T Q g_k}.$$

4. Gradient method takes

$$x^{k+1} = x^k - \frac{g_k^T g_k}{g_k^T Q g_k} g_k$$

where $g_k = Qx^k - b$.

5. Define

$$E(x) \triangleq \frac{1}{2}(x - x^*)^T Q(x - x^*).$$

Then,

$$E(x) = f(x) + \frac{1}{2}(x^*)^T Q x^*.$$

Hence

$$\min_x E(x) \approx \min_x f(x).$$

Performance

- **Lemma**

  The gradient method in the quadratic case satisfies

$$E(x^{k+1}) = \{1 - \frac{(g_k^T g_k)^2}{(g_k^T Q g_k)(g_k^T Q^{-1} g_k)}\} E(x^k).$$

- **Kantorovich Inequality**

  Let Q be a positive definite symmetric $n \times n$ matrix. The following holds for any vector x,

$$\frac{(x^T x)^2}{(x^T Q x)(x^T Q^{-1} x)} \geq \frac{4aA}{(a+A)^2}$$

  where a and A are the smallest and largest eigenvalues of Q.

Main result

Theorem (Steepest descent - quadratic case)

For any $x^0 \in E^n$ the gradient method converges to the unique minimum point x^* with the following relation, for every step k,

$$E(x^{k+1}) \leq (\frac{A-a}{A+a})^2 E(x^k)$$

Observations

- The above theorem states that with respect to the error function E (or equivalently f) the gradient method converges linearly with a ratio no greater than $(\frac{A-a}{A+a})^2$.
- If the contour of f is a circle, how many steps does the gradient method need to REACH the minimum point?
- Roughly speaking, the convergence rate of the gradient method is slowed as the contours of f become more eccentric.
- See Figure 7.9 of Luenberger's book.

- Hirotugu Akaike (1959) showed that, if the ratio $(\frac{A-a}{A+a})^2$ is large, the process is very likely to converge at a rate close to the bound.

- Define $r \triangleq \frac{A}{a}$ (condition number).

Then, $(\frac{A-a}{A+a})^2 = (\frac{r-1}{r+1})^2$.

Larger condition number $\Rightarrow$ slower convergence.

Minimizing a non-quadratic function

- Main result:

Assume that $f \in C^2$ has a relative minimum at x^*. Assume further that $F(x^*)$ has a smallest eigenvalue $a > 0$ and a largest eigenvalue $A > 0$.

If $\{x^k\}$ is a sequence generated by the method of the gradient method that converges to x^*, then the sequence of objective values $\{f(x^k)\}$ converges to $f(x^*)$ linearly with a convergence ratio no great than $(\frac{A-a}{A+a})^2$.

Concept of scaling

- Scaling may substantially alter the convergence characteristics. Consider

$$\min f(x) = x_1^2 - 5x_1x_2 + x_2^4 - 25x_1 - 8x_2$$

At solution $x^* = (20, 3)^T$, we have

$$F(x^*) = \begin{bmatrix} 2 & -5 \\ -5 & 12x_2^2 \end{bmatrix}_{x^*} = \begin{bmatrix} 2 & -5 \\ -5 & 108 \end{bmatrix}$$

with eigenvalues of 108.23 and 1.765.

Since $(\frac{A-a}{A+a})^2 = 0.9368$, the convergence is very slow.

Now consider the following scaling:

$$\hat{x}_1 = x_1, \ \hat{x}_2 = 7x_2$$

$\Rightarrow$ Equivalent problem:

$$\min f(x) = \hat{f}(\hat{x}) = \hat{x}_1^2 - \frac{5}{7}\hat{x}_1\hat{x}_2 + \frac{\hat{x}_2^4}{7^4} - 25\hat{x}_1 - \frac{8}{7}\hat{x}_2$$

At solution $\hat{x}^* = (20, 21)^T$, we have

$$F(\hat{x}^*) = \begin{bmatrix} 2 & -5/7 \\ -5/7 & 12/7^4\hat{x}_2^2 \end{bmatrix}_{\hat{x}^*} = \begin{bmatrix} 2 & -5/7 \\ -5/7 & 2.2 \end{bmatrix}$$

with eigenvalues of 2.82 and 1.38.

Since $(\frac{A-a}{A+a})^2 = 0.118$, the convergence is much faster.

Question

- Scaling looks helpful for fast convergence. Should we do it first all the time?

- **Analysis:**

Consider $\min_{x \in E^n} f(x)$.

Define $Ty = x$ with T being an $n \times n$ invertible matrix.

Then

$$\min_{x \in E^n} f(x) \approx \min_{y \in E^n} f(Ty) \triangleq \min_{y \in E^n} h(y)$$

$$\begin{tikzcd}
\min\ f(x) \arrow[r, bend left, "y = T^{-1}x"] & \min\ h(y) \arrow[l, bend left, "x = Ty"]
\end{tikzcd}$$

$$\begin{aligned} d_x &= -T\nabla h(y)^T \\ &= -\underbrace{TT^T}_{\text{effect of scaling variables}} \nabla f(x)^T \end{aligned}$$

$$d_y = -\nabla h(y)^T$$

$$\nabla h(y) = \nabla f(x) T$$

Moreover, if $x^* = Ty^*$ is a solution, then

$$H(y^*) = T^T F(x^*) T$$

For better convergence ratio, we may consider diagonal T that evens out the eigenvalues of $H(y^*)$.

Not a simple problem

- Finding eigenvalues of a given matrix takes extra computational efforts. There is no polynomial-time algorithm for finding exact eigenvalues, although there exist $O(n^3)$ algorithms for finding approximate eigenvalues.