Unconstrained Optimization

Numerical Optimization

Outline

- Introduction to UO
- Solutions to the UO problems
  - Understanding “Solutions”
  - How to look for solutions and Taylor’s theorem
  - Non smooth functions
- Algorithms of Optimization
  - General Idea
  - Line search methods
  - Trust region methods
  - Comparison
  - Scaling

Introduction

- Minimize objective functions with real variables and no constrains:

$$\min_{x} \; f(x)$$

where $x \in \mathbb{R}^n$

$f : \mathbb{R}^n \to \mathbb{R}$ is a smooth function

Usually, we lack a global perspective on the function f.

All we know are the value of f and maybe some of its derivatives at a set of points.

Example of UO problem

- Given that we have measurements y_i taken at random time t_i, $i = 1, \ldots, m$.
- Assume we know y can be modelled by function

$$\phi(t; x) = x_1 + x_2 e^{-(x_3 - t)^2 / x_4} + x_5 \cos(x_6 t)$$

where $x_1 \ldots x_6$ are parameters to be fixed.

- We want to find $x_1 \ldots x_6$ so that Φ fit the observed data y_j as closely as possible

Example of UO problem

- First, group the parameters x_i into a vector of unknowns, $x = (x_1 , x_2 , . . . , x_6)^T$
- The discrepancy between observation and model is:

$$r_j(x) = y_j - \phi(t_j; x), \qquad j = 1, 2, \ldots, m$$

- To let the model fit the observation as close as possible, we minimize (optimize) :

$$\min_{x \in \mathbb{R}^6} f(x) = r_1^2(x) + r_2^2(x) + \cdots + r_m^2(x)$$

Objective function

Understanding "Solutions"

<u>Global minimizers:</u>

A point x^* is a *global minimizer* if $f(x^*) \leq f(x)$ for all x.

It is very good if we can find global minimizers.

Difficult to find compare to local minimizers.

Why? We usually do not have a good picture of the overall shape of *f*.

Example: $\phi(t;x) = x_1 + x_2 e^{-(x_3-t)^2/x_4} + x_5 \cos(x_6 t)$

Understanding "Solutions"

Local minimizers:

A point x^* is a *local minimizer* if there is a neighborhood $\mathcal{N}$ of x^* such that $f(x^*) \leq f(x)$ for all $x \in \mathcal{N}$.

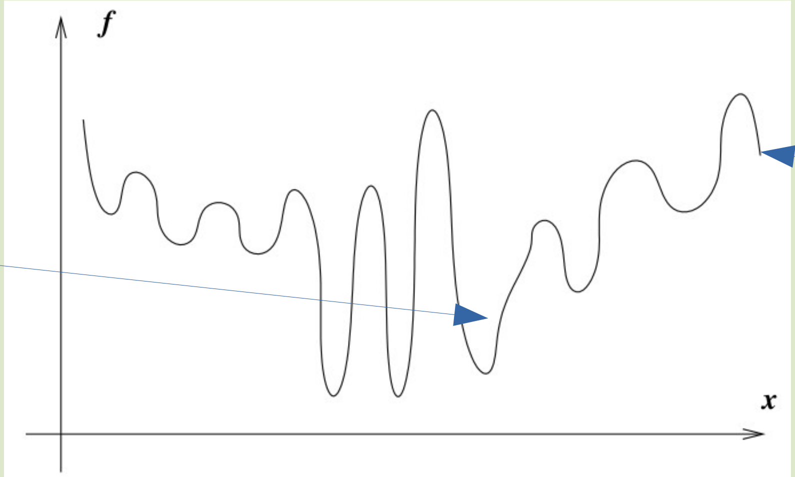
- Most algorithms are able to find only a local minimizer.
- Some are called strong local minimizer, but some are not.

A point x^* is a *strict local minimizer* (also called a *strong local minimizer*) if there is a neighborhood $\mathcal{N}$ of x^* such that $f(x^*) < f(x)$ for all $x \in \mathcal{N}$ with $x \neq x^*$.

Understanding “Solutions”

- Easier to find global minimizer if we have global knowledge about the function, otherwise might be “trapped” into local minimizers.
- Convex functions: every local minimizer is a global minimizer

Easy to be “trapped”



Not easy to find global minimizer

How to look for solution?

- If we assume x* is a local minimizer, examine all the points in its immediate vicinity and make sure that none of them has a smaller function value.
- If we can find gradient $\nabla f(x)$ and Hessian $\nabla^2 f(x^*)$ of f, we can use Taylor's Theorem.

Taylor's Theorem

Tool to study minimizers of smooth functions

If a function f is differentiable through order $n + 1$ in an interval I containing c, then, for each x in I, there exists z between x and c such that

$$f(x) = f(c) + f'(c)(x - c) + \frac{f''(c)}{2!}(x - c)^2 + \cdots + \frac{f^{(n)}(c)}{n!}(x - c)^n + R_n(x)$$

where

$$R_n(x) = \frac{f^{(n+1)}(z)}{(n + 1)!}(x - c)^{n+1}.$$

Taylor's Theorem

- With Taylor's theorem:

  $f : \mathbb{R}^n \to \mathbb{R}$ and is continuously differentiable. Given p, an n dimension real number and $0 < t < 1$:

$$f(x+p) = f(x) + \nabla f(x+tp)^T p \qquad \text{(eq1)}$$

Also,

$$f(x+p) = f(x) + \nabla f(x)^T p + \frac{1}{2} p^T \nabla^2 f(x+tp) p \qquad \text{(eq2)}$$

Studying Local Minimizers Using Taylor's Theorem

First-Order Necessary Conditions

In the case of f is a continuously differentiable in an open neighbourhood of x^*:
If x^* is a local minimizer, then $\nabla f(x^*) = 0$

Necessary condition: fulfilling the condition doesn't promise the conclusion, but failing the condition means failing the conclusion.

We call x^* a stationary point if $\nabla f(x^*) = 0$

Any local minimizer must be a stationary point

Studying Local Minimizers Using Taylor's Theorem

First-Order Necessary Conditions

In the case of *f* is a continuously differentiable in an open neighbourhood of x^*:
If x^* is a local minimizer, then $\nabla f(x^*) = 0$

Assume $f(x) = x^2$

$\nabla f(x) = 2x$

$\nabla f(0) = 0$

0 is a minimizer

But, assume $g(x) = -x^2$

$\nabla g(x) = -2x$

$\nabla g(0) = 0$

0 is not a minimizer

Studying Local Minimizers Using Taylor's Theorem

Second-Order Necessary Conditions

In the case of $\nabla^2 f$ exist and is continuous in an open neighbourhood of x^*:
If x^* is a local minimizer of f,
then $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive semidefinite.

positive semidefinite: for a symmetric square matrix A with real entries, if $x^T A x \geq 0$, it is positive semidefinite.
(x is any non-zero real column vector)

Studying Local Minimizers Using Taylor's Theorem

Second-Order Necessary Conditions

In the case of $\nabla^2 f$ exist and is continuous in an open neighbourhood of x^*:
If x^* is a local minimizer of f,
then $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive semidefinite.

Assume $f(x) = x^2$

$\nabla f(x) = 2x$

$\nabla^2 f(0) = 0$

0 is a minimizer

But, assume $g(x) = x^3$

$\nabla g(x) = 3x^2$

$\nabla^2 g(0) = 0$

0 is not a minimizer

Studying Local Minimizers Using Taylor's Theorem

Second-Order Sufficient Conditions

In the case of $\nabla^2 f$ is continuous in an open neighborhood of x^*:
If $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive definite, then x^* is a **strict** local minimizer of f

sufficient condition: fulfilling the condition guarantee the conclusion

But not fulfilling the condition doesn't means the conclusion is false

Studying Local Minimizers Using Taylor's Theorem

Second-Order Sufficient Conditions

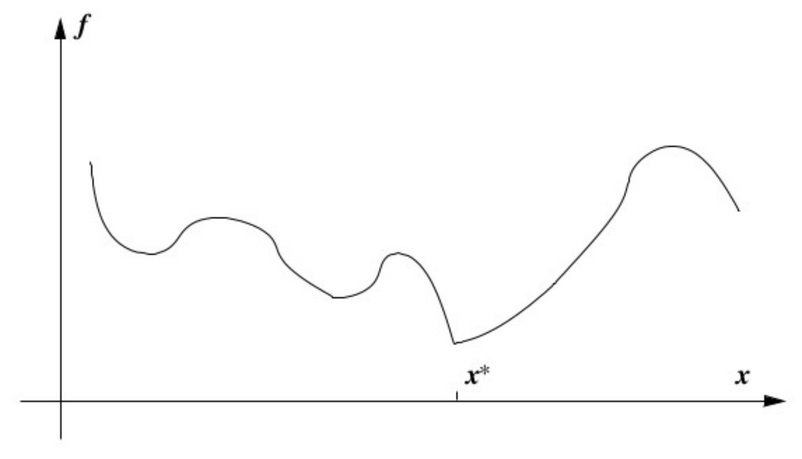
In the case of $\nabla^2 f$ is continuous in an open neighborhood of x^*:
If $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive definite,
then x^* is a **strict** local minimizer of f

This is "sufficient but not necessary" condition :

- *Some strict local minimizer may fail to satisfy this condition. For example: x^4 ($x^4 = 0$ is strict local minimizer, but fail this).*

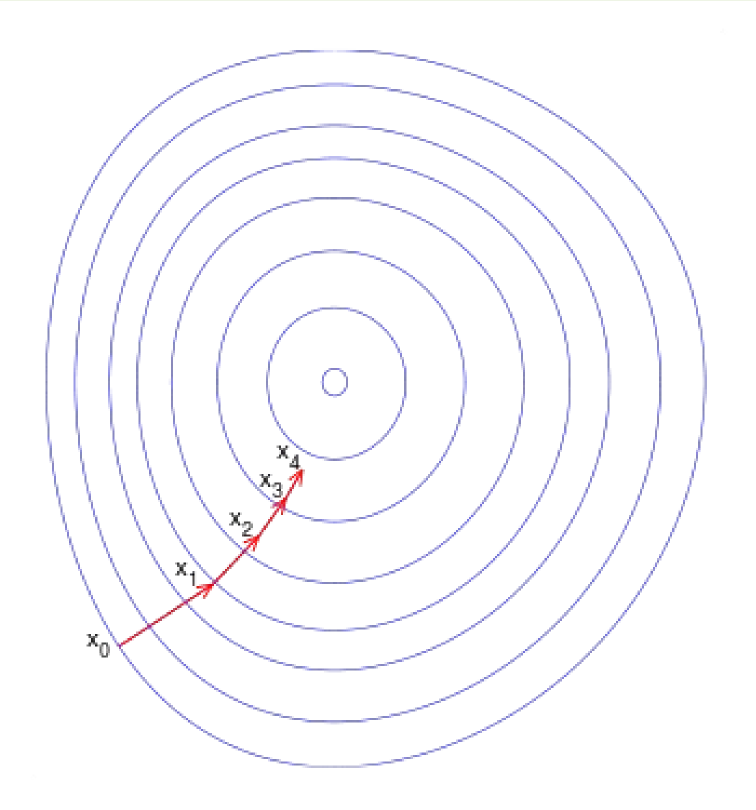
Local Minimizer for non smooth function

- No general solution for non smooth functions.
- However, if the function is a piecewise function and each piece is smooth, we can minimize each piece and compare to find the minimizer.

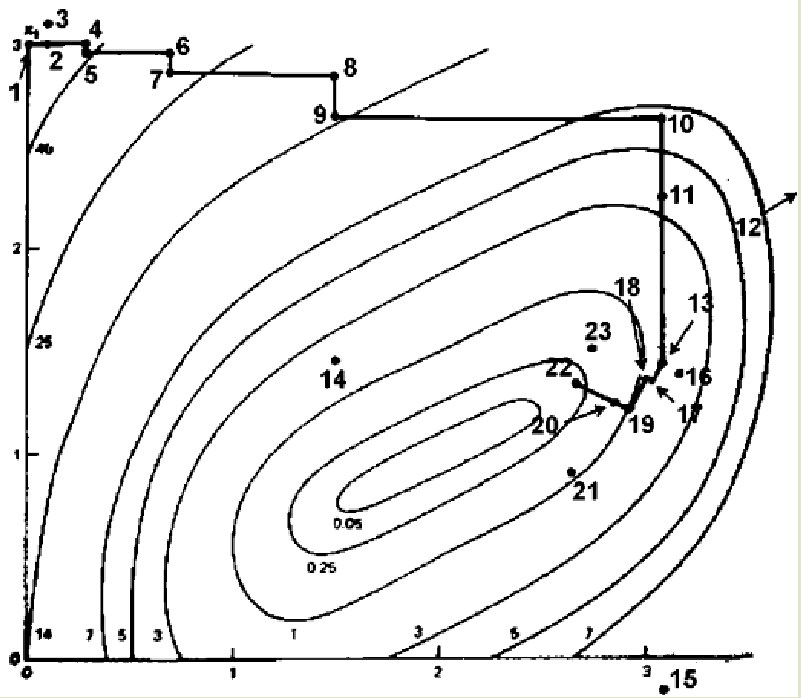


Algorithms of Optimization

- Always start with a starting point, x_0.
  - It will be good if we know where is a good starting point based on our understanding about the data
  - Or we may have some systematic or random guess.
- A sequence of iteration is performed by the algorithm to refine the x_k , until a stopping criteria is matched.
- Stopping criteria : no more progress can be made, or high accuracy solution is obtained.
- Two fundamental strategies: Line search and trust region



x_0
x_1
x_2
x_3
x_4



1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
0.05
0.25

LimCK

Line Search

- Many methods under this class of strategy, i.e. steepest descent, Newton, Quasi-Newton, conjugate gradient, ...
- Generally, these methods start from current point x_k, and search for a direction p_k to achieve next point x_{k+1}.
- $f(x_{k+1})$ must bring lower value compared to $f(x_k)$
- Once direction p_k is determined, solve the following to determine the step length, α – how far to move along p_k:

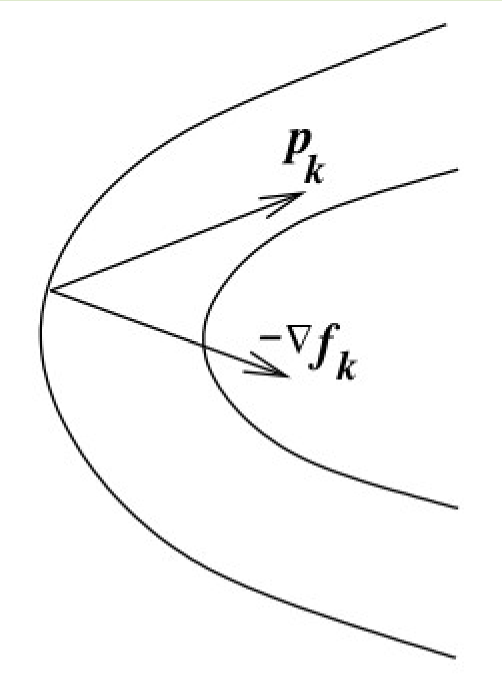
$$\min_{\alpha>0} f(x_k + \alpha p_k) \qquad \text{(eq3)}$$

Line Search

- 2 steps in line search methods:
  - Step 1: Determine p_k (discuss later)
  - Step 2: Determine α using eq3.
- However, to find the best α is costly.
- Instead of finding best α: generates a limited number of trial step lengths until it finds one that loosely approximates the minimum

Search Directions

- Any descent direction (makes an angle of strictly less than $\pi/2$ with $-\nabla f_k$) is guaranteed to produce a decrease in f
- Many possible directions:
  - Steepest descent
  - Newton
  - Quasi-Newton
  - Conjugate gradient



Steepest Descent Method

- Search along the direction which f decreases most rapidly:

$$p = -\nabla f_k \qquad \text{(eq4)}$$

- Advantage: it requires calculation of the gradient ∇f_k but not of second derivatives.
- Greedy algorithm – can be slow in actual, especially on difficult problems.

Newton Method

- One of the most important search direction.
- Derived from the second-order Taylor series approximation to $f(x_k + p)$

$$f(x_k + p) \approx f_k + p^T \nabla f_k + \frac{1}{2} p^T \nabla^2 f_k p \qquad \text{(eq5)}$$

- We want to obtain x_{k+1} which brings $f(x_k+p)$ to its minimum.
- If $\nabla^2 f_k$ is positive definite (i.e. $½p^T \nabla^2 f_k p > 0$), eq5 is a convex quadratic equation of p, in the form of:

$$a_{k+1} = a_k + bp + ½cp^2$$

Newton Method

- Since it is a quadratic equation, we can find the '*p*' that brings *f* to its minimum by:

$$\frac{df\left(x_k + p\right)}{dp} = 0$$

Therefore, $\nabla f_k + \nabla^2 f_k p = 0$

Newton direction, $p = -\left(\nabla^2 f_k\right)^{-1} \nabla f_k$ (eq6)

Newton Method

- Because eq5 gives only an approximation (compared to eq2), the third term is the only difference. $\nabla^2 f(x_k+tp) \leftrightarrow \nabla^2 f_k$
- If $\nabla^2 f$ is sufficiently smooth, the difference is small.
- In the other word, the direction is accurate if $||p||$ is small.

Newton Method

- Disadvantage – need to compute Hessian $\nabla^2 f$
- Disadvantage – This method works only if $\nabla^2 f_k$ is positive definite.
- Advantage - Methods that use the Newton direction have a fast rate of local convergence.

Quasi-Newton Method

- Instead of using Hessian $\nabla^2 f_k$, an approximation B_k is used.

$$\nabla^2 f_k \approx B_k$$

- B_k is updated after each step to take account of the additional knowledge gained during the step.

Quasi Newton direction $$p_k = -B_k^{-1} \nabla f_k \qquad \text{(eq7)}$$

Quasi-Newton Method

B_k is given by secant equation:

$$B_{k+1} s_k = y_k \qquad \text{(eq8)}$$

Where:

$$s_k = x_{k+1} - x_k \qquad y_k = \nabla f_{k+1} - \nabla f_k$$

Typically, at the first iteration, identity matrix is used as B_0

B_k is updated for the following iterations.

Quasi-Newton Method

Two common methods to update B_k

1) *symmetric-rank-one* (SR1) formula

$$B_{k+1} = B_k + \frac{(y_k - B_k s_k)(y_k - B_k s_k)^T}{(y_k - B_k s_k)^T s_k}$$

2) BFGS formula, named after its inventors

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k}$$

Questions?

Exercise

2.1 Compute the gradient $\nabla f(x)$ and Hessian $\nabla^2 f(x)$ of the Rosenbrock function

$$f(x) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2.$$

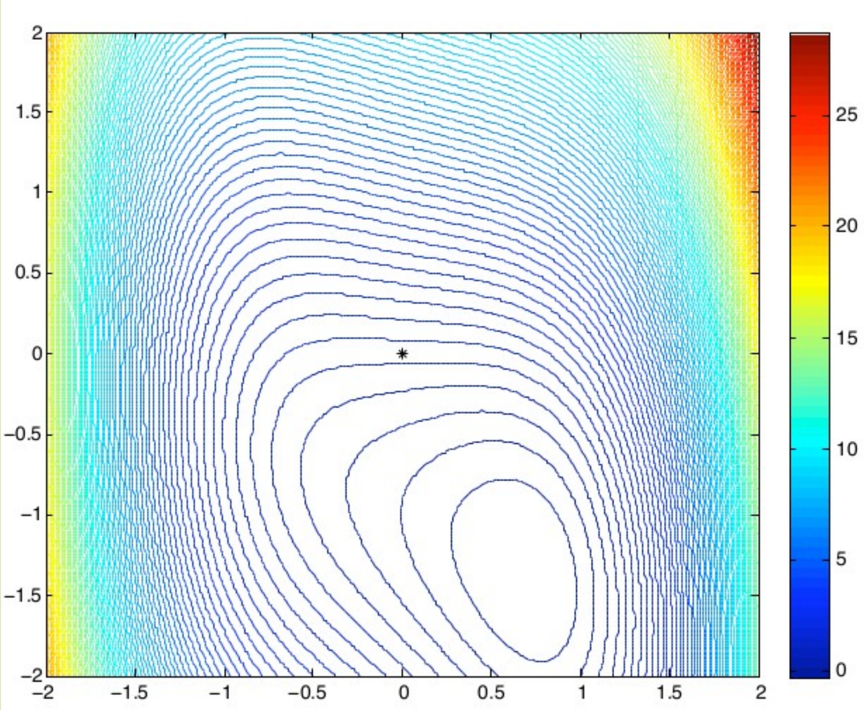
Show that $x^* = (1, 1)^T$ is the only local minimizer of this function, and that the Hessian matrix at that point is positive definite.

2.3 Let a be a given n-vector, and A be a given $n \times n$ symmetric matrix. Compute the gradient and Hessian of $f_1(x) = a^T x$ and $f_2(x) = x^T A x$.

Exercise

Q3. (Modified from Practical Methods of Optimization, Fletcher)
Obtain expressions for all first and second derivatives of the function of two variables:

$$f(x) = x_1^4 + x_1 x_2 + (1 + x_2)^2$$



(a) Argue why the point (0, 0) marked with an asterisk on the contour diagram cannot be a local minimizer.

(b) Show that the Hessian $\nabla^2 f(0)$ does not satisfy the property

$$p^T \nabla^2 f(0) p > 0 \text{ for all } p \neq 0.$$

(c) A local minimizer of f is $x^* = (0.6959, -1.3479)$. Verify that the first order necessary conditions for optimality are satisfied at x^*.