Numerical Optimization

Instructor : Sung Chan Jun

Week #6 : October 7 – 11, 2019

BioComputing

Announcements

- Midterm Exam
  - Date : October 16 (Wednesday), 2019
  - Time : 10:30 AM – Noon
  - Scope : Week #1 – Week #7
- No Class
  - Date : October 23 (Wednesday), 2019
- Makeup Class
  - Date : October 21 (Monday), 2019
  - Time : 7:00 PM – 8:15 PM
  - No attendance check

BioComputing

Course Syllabus (tentative)

1st week	Sept. 2, 4	Introduction of optimization	
2nd week	Sept. 9, 11	Univariate Optimization	
3rd week	Sept. 16, 18	Univariate Optimization	
4th week	Sept. 23, 25	Unconstrained Multivariate Optimization	
5th week	Sept. 30, Oct. 2	Unconstrained Multivariate Optimization	
6th week	Oct. 7, 9	Unconstrained Multivariate Optimization	National Holiday (Oct. 9)
7th week	Oct. 14, 16	Unconstrained Multivariate Optimization	Midterm (Oct. 16)
8th week	Oct. 21, 23	Constrained Multivariate Optimization	

Course Syllabus (tentative)

9th week	Oct. 28, 30	Constrained Multivariate Optimization	
10th week	Nov. 4, 6	Constrained Multivariate Optimization	
11th week	Nov. 11, 13	Constrained Multivariate Optimization	
12th week	Nov. 18, 20	Global Optimization	
13th week	Nov. 25, 27	Global Optimization	
14th week	Dec. 2, 4	Global Optimization, Wrap-up	
15th week	Dec. 9	Final Exam	Final Exam (Dec. 9)

Recall Last Week

- Multivariate Optimization: Methods for Smooth Functions

$$\boxed{\text{Minimize } f(\mathbf{x}) \text{ on } \mathbf{x} = (x_1, x_2, \ldots, x_n)^T \in R^n}$$

  - Typical Algorithm (model algorithm)

    S1. [Test for convergence]

    If termination condition is satisfied, the algorithm terminates with $\mathbf{x}_k$ as the solution.

    S2. [Compute (or determine) a search direction]

    Compute a non-zero n-vector $\mathbf{p}_k$ (direction of search).

    S3. [Compute (or determine) a step length]

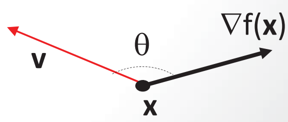
    Compute α_k (step length) such that $f(\mathbf{x}_k + \alpha_k \mathbf{p}_k) < f(\mathbf{x}_k)$.

    S4. [Update the estimate of the minimum]

    Set $\mathbf{x}_{k+1} := \mathbf{x}_k + \alpha_k \mathbf{p}_k$ and $k := k + 1$, and go back to S1.

Recall Last Week

- Multivariate Optimization: Methods for Smooth Functions
  - Descent Methods
    - Method for which a descent condition $f_{k+1} < f_k$ for all $k \geq 0$, that is, function values are strictly decreasing.
  - Descending direction **p** at **x**
    - When $\mathbf{p} \cdot \nabla f(\mathbf{x}) < 0$, i.e the angle between vectors **p** and $\nabla f(\mathbf{x})$ is $> \pi/2$.
    - Estimate a slope of $f(\mathbf{x})$ along unit **v** direction at **x**
      - $df(\mathbf{x} + t\mathbf{v})/dt \,|_{t=0} = \nabla f(\mathbf{x}) \cdot \mathbf{v} = |\nabla f(\mathbf{x})| \cdot |\mathbf{v}| \cos(\theta) = |\nabla f(\mathbf{x})| \cos(\theta)$
      - $\nabla f(\mathbf{x}) \cdot \mathbf{v}$ yields the biggest slope when $\theta = 0$, that is, $\mathbf{v} = \nabla f(\mathbf{x}) / |\nabla f(\mathbf{x})|$.



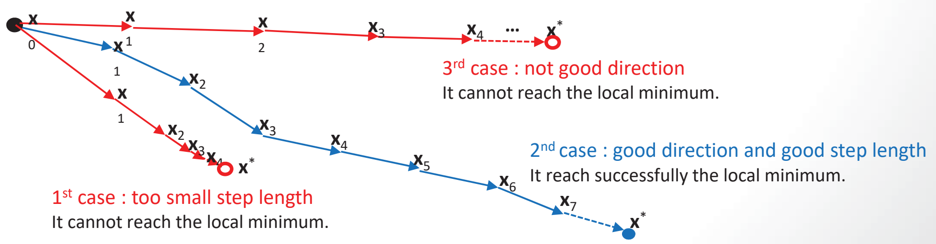
Recall Last Week

- Multivariate Optimization: Methods for Smooth Functions
  - Existence of a reasonable step length at descending direction
    - (Theorem) Let **p** be a descending direction at **x**.

      $\exists\, \alpha_0 > 0$ such that $f(\mathbf{x} + \alpha \mathbf{p}) < f(\mathbf{x})$, $\forall\, 0 \le \alpha \le \alpha_0$. (by Taylor expansion).
  - Does the descent condition ($f_{k+1} < f_k$ for all $k \ge 0$) imply that the sequence $\{\mathbf{x}_k\}$ always converges to a local minimum point **x**?
    - No.
    - This case happens when
      - Step lengths α_k are chosen so that the reduction in function values gets far smaller at each iteration.
      - Search direction $\mathbf{p}_k$ is almost parallel to the contour line, i.e, almost orthogonal to $\nabla f(\mathbf{x})$.

Recall Last Week

- Multivariate Optimization: Methods for Smooth Functions
  - Descent condition ($f_{k+1} < f_k$ for all $k \ge 0$) doesn't imply that the sequence $\{\mathbf{x}_k\}$ always converges to a local minimum point **x**.
  - How to overcome when these cases happen?
    - (To overcome 3st case) Step lengths α_k are chosen so that the reduction in function values gets far smaller at each iteration.
      - Wolfe conditions or Armijo-Goldstein conditions
    - (To overcome 1st case) Search direction $\mathbf{p}_k$ is almost orthogonal to $\nabla f(\mathbf{x})$.
      - Direction $\mathbf{p}_k$ keeps away from the orthogonality to $\nabla f(\mathbf{x})$.
        - Consider some condition such as $|\mathbf{p} \cdot \nabla f(\mathbf{x})| > \delta > 0$ for a small δ



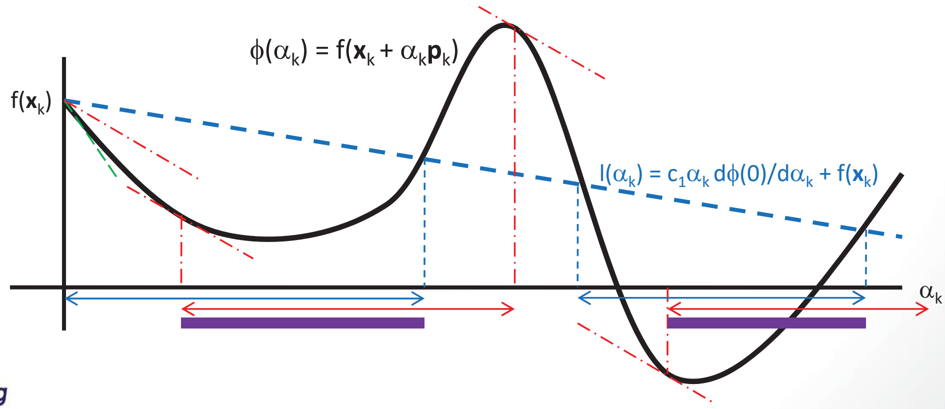
Recall Last Week

- Multivariate Optimization: Methods for Smooth Functions
  - Smart ways to choose step lengths α_k?
    - Wolfe Conditions

$$f(x_k + \alpha_k p_k) \le c_1 \alpha_k \nabla f(x_k) \cdot p_k + f(x_k)$$
$$\nabla f(x_k + \alpha_k p_k) \cdot p_k \ge c_2 \nabla f(x_k) \cdot p_k, \quad 0 < c_1 < c_2 < 1$$

Letting $\phi(\alpha_k) = f(\mathbf{x}_k + \alpha_k \mathbf{p}_k)$

$$\phi(\alpha_k) \le c_1 \alpha_k \, d\phi(0)/d\alpha_k + f(\mathbf{x}_k)$$
$$d\phi(\alpha_k)/d\alpha_k \ge c_2 d\phi(0)/d\alpha_k$$



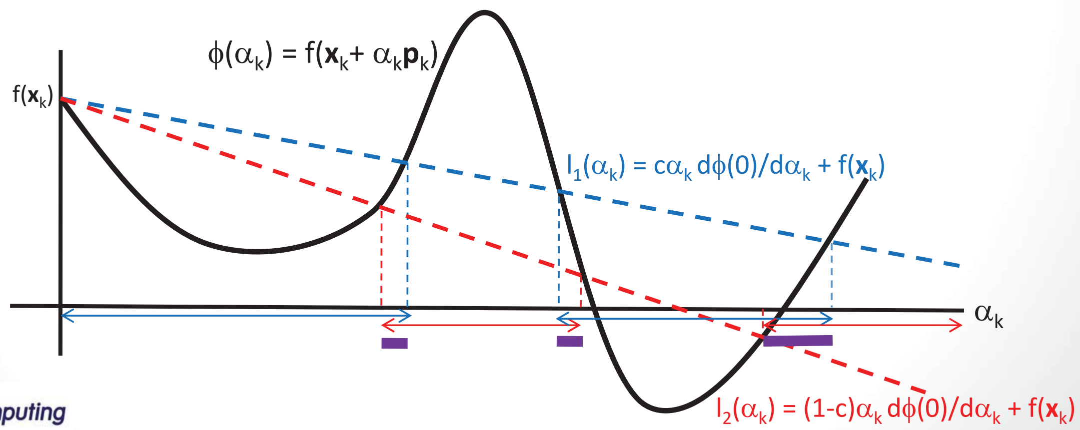
Recall Last Week

- Multivariate Optimization: Methods for Smooth Functions
  - Smart ways to choose step lengths α_k?
    - Goldstein Conditions

$$(1-c)\alpha_k \nabla f(\mathbf{x}_k) \cdot \mathbf{p}_k + f(\mathbf{x}_k) \le f(\mathbf{x}_k + \alpha_k \mathbf{p}_k) \le c\alpha_k \nabla f(\mathbf{x}_k) \cdot \mathbf{p}_k + f(\mathbf{x}_k), 0 < c < 1/2$$

$$(1-c)\alpha_k \, d\phi(0)/d\alpha_k + f(\mathbf{x}_k) \le \phi(\alpha_k) \le c\alpha_k \, d\phi(0)/d\alpha_k + f(\mathbf{x}_k)$$

Letting $\phi(\alpha_k) = f(\mathbf{x}_k + \alpha_k \mathbf{p}_k)$



Recall Last Week

- Multivariate Optimization: Methods for Smooth Functions
  - (Theorem) Existence of α satisfying Wolfe Conditions
    - Assume f($\mathbf{x}$) is continuously differentiable and f($\mathbf{x}$) > M (some number) on the ray $\{\mathbf{x}_k + \alpha \mathbf{p}_k : \alpha > 0\}$. Then $\exists$ interval of α satisfying Wolfe Conditions
  - (Theorem) Existence of α satisfying Goldstein Conditions
    - Assume f($\mathbf{x}$) is continuously differentiable and f($\mathbf{x}$) > M (some number) on the ray $\{\mathbf{x}_k + \alpha \mathbf{p}_k : \alpha > 0\}$. Then $\exists$ interval of α satisfying Goldstein Conditions.

Recall Last Week

- Multivariate Optimization: Methods for Smooth Functions

Iteration formula : $\mathbf{x}_k = \mathbf{x}_{k-1} + \alpha_k \mathbf{p}_k$

Assumptions

1. Let $\mathbf{p}_k$ be a descent direction away from orthogonality to $\nabla f(\mathbf{x}_k)$.
2. Let α_k satisfy Wolfe conditions.
3. Let f($\mathbf{x}$) > M (some number), continuously differentiable in a set D = $\{\mathbf{x}: f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$, and ∇f is Lipschitz continuous on D.

Then $\mathbf{x}_k$ converges to a stationary point, i.e, $\lim_{k\to\infty} \|\nabla f(x_k)\| = 0$

Recall Last Week

- Line search : Finding step length

Minimize $\phi(\alpha) = f(\mathbf{x}_k + \alpha\mathbf{p}_k)$ for the given search direction $\mathbf{p}_k$ and $\alpha > 0$

- Easy thinking
  - Find a local minimizer (exact line search). It may be too expensive.
- Smart thinking
  - Instead finding a local minimizer, choose α to give a substantial reduction in f(x) in a cheaper way (inexact line search).
  - Inexact line search
    - Backtracking line search
      - Choose $\alpha_0 > 0$, $\rho \in (0,1)$, $c \in (0,1)$
      - Set $\alpha := \alpha_0$
      - Repeat until $f(\mathbf{x}_k + \alpha\mathbf{p}_k) \leq \alpha c \nabla f_k \cdot \mathbf{p}_k + f(\mathbf{x}_k)$

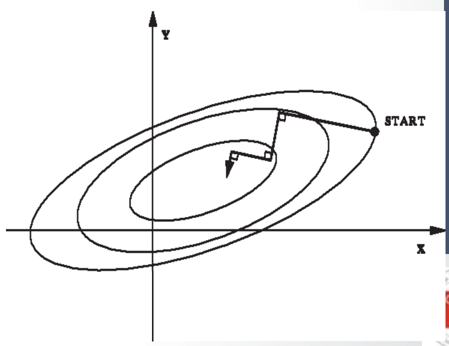
        Set $\alpha := \alpha \cdot \rho$
      - Terminate with $\alpha_k = \alpha$.

Recall Last Week

- The method of steepest descent (Cauchy's method)
  - Directional derivative at **x** along direction **p**

$$\lim_{\alpha \to 0} \frac{f(\mathbf{x} + \alpha\mathbf{p}) - f(\mathbf{x})}{\alpha} = \mathbf{p} \cdot \nabla f(\mathbf{x})$$

  - Steepest descent unit direction **p**
    - the greatest negative value of $\mathbf{p} \cdot \nabla f(\mathbf{x})$ is $\mathbf{p} = -\nabla f(\mathbf{x}) / |\nabla f(\mathbf{x})|$.
  - Using steepest descent direction $-\nabla f(\mathbf{x})$ yields
    - $\mathbf{x}_{k+1} := \mathbf{x}_k + \alpha_k \mathbf{p}_k \Rightarrow \boxed{\mathbf{x}_{k+1} := \mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k)}$



Recall Last Week

- The method of steepest descent
  - Pros
    - Reliable for any starting point ("global convergence")
    - Easy to implement. Used as a starting method for other methods
  - Cons
    - Slow convergence near the minimum point
  - Evaluation of the gradient (first derivative approximation)
    - When it is not practical, finite difference approximation is used.

$$\left.\frac{\partial f}{\partial x_i}\right|_{\mathbf{x}} \approx \frac{f(\mathbf{x} + h_i\mathbf{e}_i) - f(\mathbf{x})}{h_i}, \text{ 'forward difference formula'}$$

$$\left.\frac{\partial f}{\partial x_i}\right|_{\mathbf{x}} \approx \frac{f(\mathbf{x}) - f(\mathbf{x} - h_i\mathbf{e}_i)}{h_i}, \text{ 'backward difference formula'}$$

$$\left.\frac{\partial f}{\partial x_i}\right|_{\mathbf{x}} \approx \frac{f(\mathbf{x} + h_i\mathbf{e}_i) - f(\mathbf{x} - h_i\mathbf{e}_i)}{2h_i}, \text{ 'central difference formula'}$$

Multivariate Optimization: Method of Steepest Descent

- Convergence
  - Convex quadratic function $f(\mathbf{x}) = 1/2\mathbf{x}^T\mathbf{Q}\mathbf{x} - \mathbf{b}^T\mathbf{x}$ where $\mathbf{Q}$ is positive definite.
    - Steepest descent method with exact line search (step length) converges linearly. That is, it satisfies the following:

$$\left\|\mathbf{x}_{k+1} - \mathbf{x}^*\right\|_Q^2 \le \left(\frac{1-r}{1+r}\right)^2 \left\|\mathbf{x}_k - \mathbf{x}^*\right\|_Q^2, \quad r = \lambda_{min}/\lambda_{max} \iff \left\|\mathbf{x}_k - \mathbf{x}^*\right\|_Q \le \left(\frac{1-r}{1+r}\right)^k \left\|\mathbf{x}_0 - \mathbf{x}^*\right\|_Q$$

Here λ is eigenvalue of $\mathbf{Q}$.

  - General smooth function $f(\mathbf{x})$ (twice continuously differentiable)
    - Assume steepest descent method with exact line search converges to a point $\mathbf{x}^*$, where Hessian $\nabla^2 f(\mathbf{x}^*)$ is positive definite. Then

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \le \left(\frac{1-r}{1+r}\right)^2 [f(\mathbf{x}_k) - f(\mathbf{x}^*)], \quad r = \lambda_{min}/\lambda_{max}$$

Here λ is eigenvalue of $\nabla^2 f(\mathbf{x}^*)$.

Multivariate Optimization:
Second Derivative methods

- Newton's Method
  - By Taylor's expansion for multivariate function at current point x_k,

$$f(\mathbf{x}_k + \mathbf{p}_k) \approx f(\mathbf{x}_k) + \mathbf{p}_k \cdot \nabla f(\mathbf{x}_k) + \frac{1}{2}\mathbf{p}_k^T H(\mathbf{x}_k)\mathbf{p}_k$$

Looking for direction p_k to yield a minimum of the right hand side is

$$H(\mathbf{x}_k)\mathbf{p}_k = -\nabla f(\mathbf{x}_k) \quad \therefore \mathbf{p}_k = -H(\mathbf{x}_k)^{-1}\nabla f(\mathbf{x}_k).$$

So, Newton's iteration formula is $\boxed{\mathbf{x}_{k+1} = \mathbf{x}_k - H(\mathbf{x}_k)^{-1}\nabla f(\mathbf{x}_k).}$

When a step length procedure is included,

$$\boxed{\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k H(\mathbf{x}_k)^{-1}\nabla f(\mathbf{x}_k).}$$

However, step length $\alpha_k = 1$ is commonly used.

Multivariate Optimization:
Second Derivative methods

- Recall : Newton's method in Univariate Optimization
  - $f \approx$ quadratic interpolation function $f^\wedge$. By Taylor's expansion, with $f(x_k)$, $f'(x_k)$ and $f''(x_k)$

$$f^\wedge(x) = f(x_k) + f'(x_k)(x - x_k) + \frac{1}{2}f''(x_k)(x - x_k)^2$$

  - Find its minimum and call it x_{k+1} , then

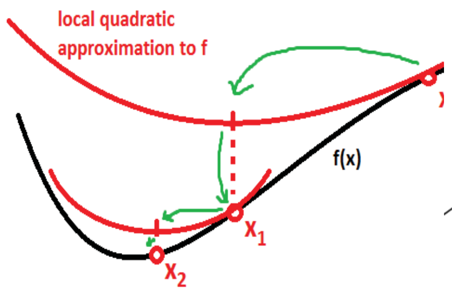
$$\boxed{x_{k+1} = x_k - f'(x_k)/f''(x_k)}$$

- Newton's Method (in Multivariate Optimization)

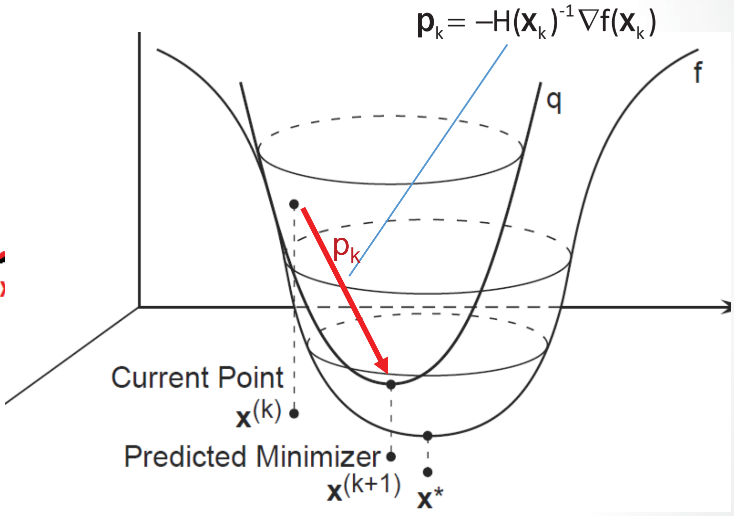
$$\boxed{\mathbf{x}_{k+1} = \mathbf{x}_k - H(\mathbf{x}_k)^{-1}\nabla f(\mathbf{x}_k).}$$

Multivariate Optimization: Second Derivative methods

- Geometrical view of Newton's methods



Univariate (1-dim) problem

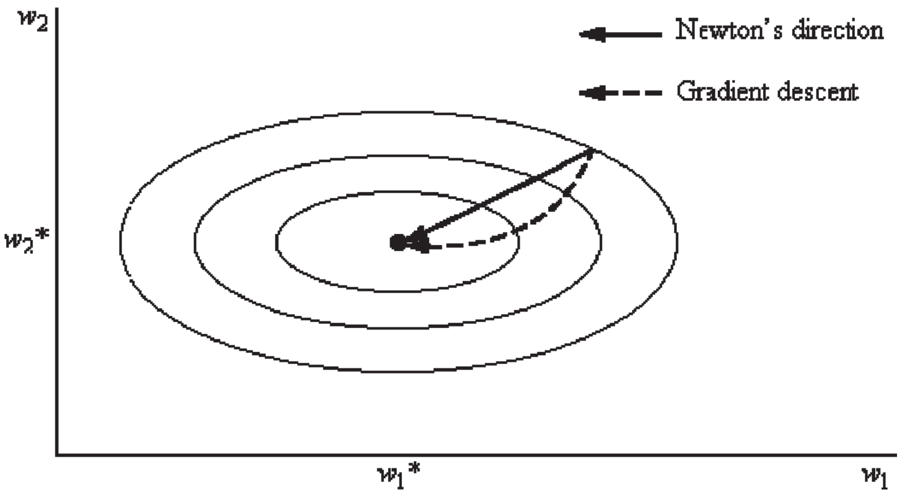


Multivariate (2-dim or higher) problem

Multivariate Optimization: Newton's Method

- (Theorem) Convergence of Newton's Method
  - We assume that
    - $\mathbf{x}_{k+1} = \mathbf{x}_k - H(\mathbf{x}_k)^{-1}\nabla f(\mathbf{x}_k)$
    - $f(\mathbf{x})$ is twice differentiable and $\nabla^2 f(\mathbf{x})$ is Lipschitz continuous around neighborhood of a local minimum $\mathbf{x}^*$, where $\nabla f(\mathbf{x}^*) = 0$ and $\nabla^2 f(\mathbf{x}^*)$ is positive definite.
  - Then
    - Iterates sequences $\{\mathbf{x}_k\} \rightarrow \mathbf{x}^*$ and $\{\nabla f(\mathbf{x}_k)\} \rightarrow 0$ when $\mathbf{x}_0$ is sufficiently close to $\mathbf{x}^*$.
    - Convergence rate of $\{\mathbf{x}_k\}$ and gradient $\{\nabla f(\mathbf{x}_k)\}$ are quadratic.

Multivariate Optimization: Newton's Method



- Newton's direction : more likely pointing to a local minimum
- Gradient direction : pointing to maximum direction of change

Multivariate Optimization: Newton's Method

- When all Hessians $H(\mathbf{x}_k)$ are positive definite and step length is reasonable, then Newton's ($\mathbf{p}_k = -H(\mathbf{x}_k)^{-1}\nabla f(\mathbf{x}_k)$) is a descent method and converges quadratically.

$$\text{For } \mathbf{p}_k = -H(\mathbf{x}_k)^{-1}\nabla f(\mathbf{x}_k) \text{ and } \nabla f(\mathbf{x}_k) \neq 0,$$
$$\nabla f(\mathbf{x}_k)\cdot\mathbf{p}_k = -\nabla f(\mathbf{x}_k)^T H(\mathbf{x}_k)^{-1}\nabla f(\mathbf{x}_k) < 0.$$

- In general, convergence is dependent on the accuracy of Taylor expansion and on the distance between initial point and minimum point.

Multivariate Optimization: Newton's Method

- Pros
  - Rapid convergence (quadratic)
- Cons
  - It need very expensive computation (evaluation of Hessian + its inversion) every iteration.
  - Convergence depends on initial point (starting point).
  - Positive definiteness of Hessian is required.

Multivariate Optimization: Modified Newton's methods

- When Hessian $H(\mathbf{x}_k)$ is indefinite, i.e, $H(\mathbf{x}_k)$ has both negative and positive eigenvalues
  - Strategy 1. Find a matrix **M** such that $H(\mathbf{x}_k) + \mathbf{M}$ is positive definite.
    - For example, choose $\mathbf{M} = \tau\mathbf{I}$ such $H(\mathbf{x}_k) + \mathbf{M}$ is sufficiently positive definite.
  - Strategy 2. Modify $H(\mathbf{x}_k)$ into positive definite matrix accordingly or approximate it by positive definite matrix.
- When Hessian $H(\mathbf{x}_k)$ is singular (noninvertible) and $\nabla f(\mathbf{x}_k) \neq 0$
  - Newton method is not applicable.
  - Choose the method of steepest descent.