CS5321 Numerical Optimization

01 Introduction

Class information

- Class webpage: http://www.cs.nthu.edu.tw/~cherung/cs5321
- Text and reference books:
  - Numerical optimization, Jorge Nocedal and Stephen J. Wright (http://www.mcs.anl.gov/otc/Guide)
  - Linear and Nonlinear Programming, Stephen G. Nash and Ariela Sofer (1996, 2005)
  - Numerical Methods for Unconstrained Optimization and Nonlinear Equations, J. Dennis and R. Schnabel
- TA: 王治權 pponywong@gmail.com

Grading

- Class notes (50%)
  - Using latex to write one class note.
  - Peer review system
    - You must review others' notes and give comments
    - The grade is given on both works (30%-20%)
- Project (40%)
  - Applications, software survey, implementations
  - Proposal (10%), presentation(10%), and report (20%)
- Class attendance (10%)

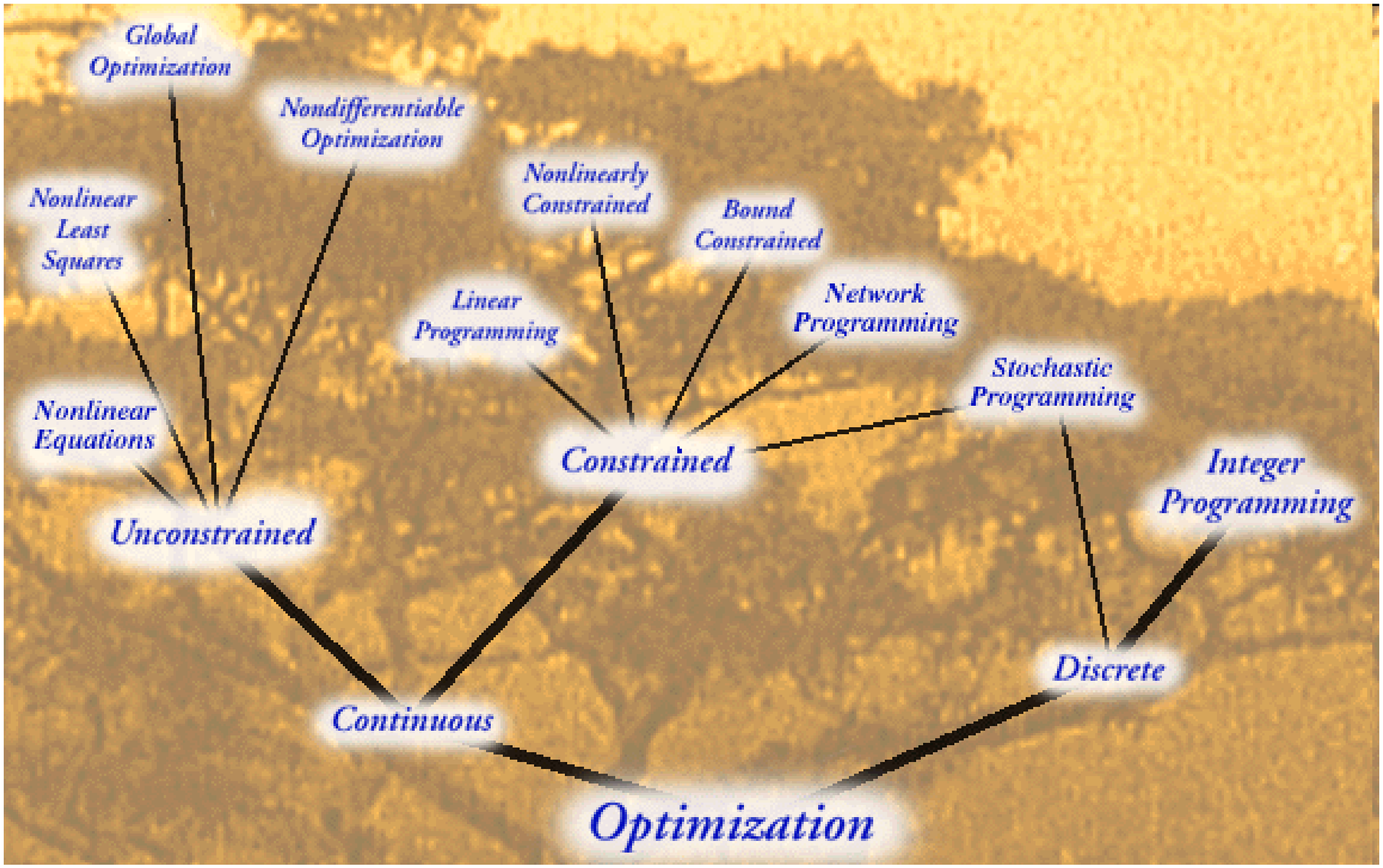
Outline of class

- Introduction
  - Background
- Unconstrained optimization
  - Fundamental of unconstrained optimization
  - Line search methods
  - Trust region methods
  - Conjugate gradient methods
  - Quasi-Newton methods
  - Inexact Newton methods

Outline of class-continue

- Linear programming
  - The simplex method
  - The interior point method
- Constrained optimization
  - Optimality conditions
  - Quadratic programming
  - Penalty and augmented Largrangian methods
  - Active set methods
  - Interior point methods

Optimization Tree



Classification

- Minimization or maximization
- Continuous vs. discrete optimization
- Constrained and unconstrained optimization
- Linear and nonlinear programming
- Convex and non-convex problems
- Global and local solution

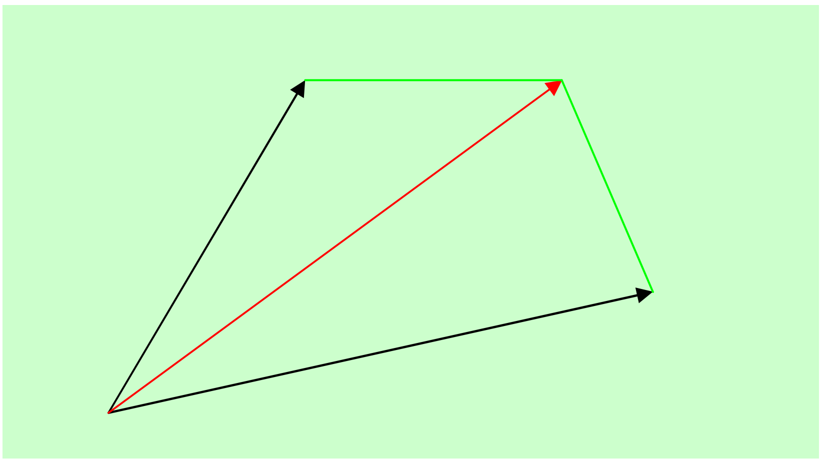
Affine, convex, and cones

$$x = \sum_{i=1}^{p} \lambda_i x_i = \lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_p x_p$$

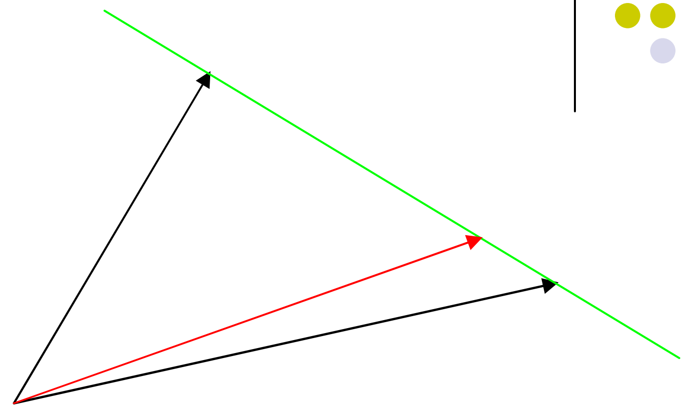
where $x_i \in \mathbb{R}^n$ and $\lambda_i \in \mathbb{R}$.

- x is a *linear combination* of $\{x_i\}$
- x is an *affine combination* of $\{x_i\}$ if $\sum_{i=1}^{p} \lambda_i = 1$
- x is a *conical combination* of $\{x_i\}$ if $\lambda_i \geq 0$
- x is a *convex combination* of $\{x_i\}$ if $\sum_{i=1}^{p} \lambda_i = 1, \lambda_i \geq 0$

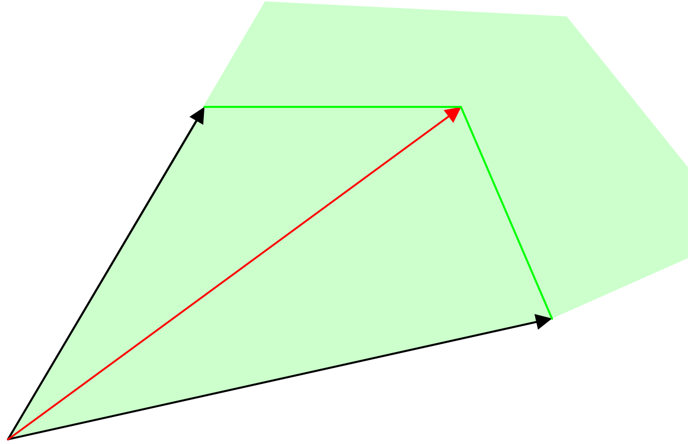
- *linear combination*



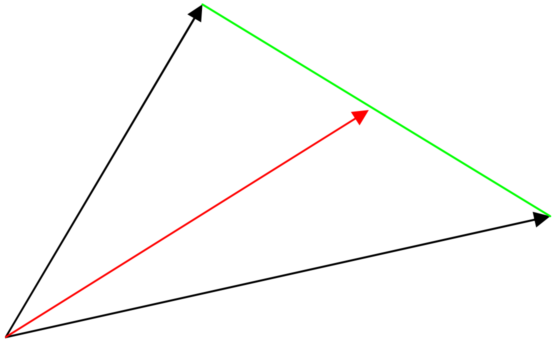
- *affine combination*



- *conical combination*



- *convex combination*

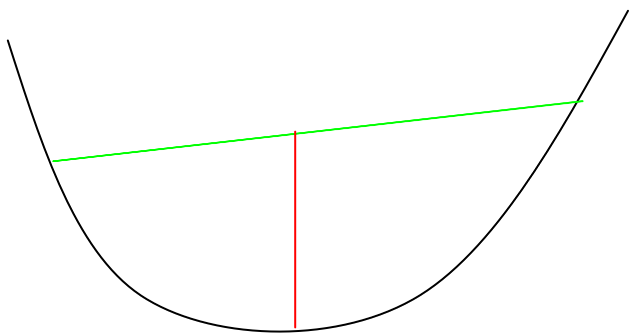


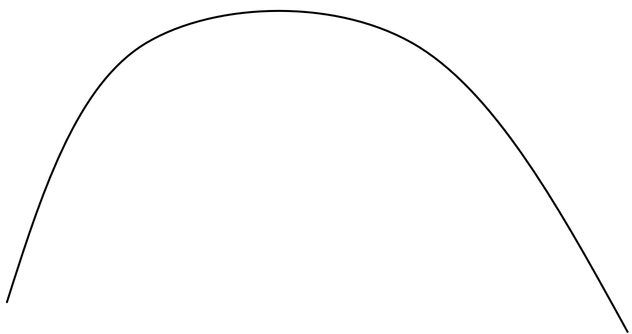
Convex function

- A function f is convex if for $\alpha \in [0,1]$

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y)$$

- A function f is concave if $-f$ is convex





Some calculus

- Rate of convergence
- Lipschitz continuity
- Single valued function
  - Derivatives and Hessian
  - Mean value theorem and Taylor's theorem
- Vector valued function
  - Derivatives and Jacobian

Rate of convergence

- Let $\{x_k | x_k \in \mathbb{R}^n\}$ be a sequence converge to x^*
  - The convergence is Q-linear if for a constant $r \in (0,1)$

$$\frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} \leq r$$

  - The convergence is Q-superlinear if

$$\lim_{k \to \infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = 0$$

  - The convergence is p Q-order if for a constant M

$$\frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|^p} \leq M$$

Lipschitz continuity

- A function f is said to be Lipschitz continuous on some set $\mathcal{N}$ if there is a constant $L>0$ such that

$$\| f(x) - f(y)\| \leq L \, \|x - y\| \text{ for all } x,y \in \mathcal{N}$$

- If function f and g are Lipschitz continuous on $\mathcal{N}$, $f+g$ and fg are Lipschitz continuous on $\mathcal{N}$.

Derivatives

- For a single valued function $f(x_1,\ldots,x_n): \mathbb{R}^n \rightarrow \mathbb{R}$
  - Partial derivative of x_i: $$\frac{\partial f}{\partial x_i} = \lim_{h \rightarrow 0} \frac{f(x + he_i) - f(x)}{h}$$
  - Gradient of f is $$\nabla f(x) = \begin{pmatrix} \partial f / \partial x_1 \\ \partial f / \partial x_2 \\ \vdots \\ \partial f / \partial x_n \end{pmatrix}$$
  - Directional derivatives: for $\|p\|=1$

$$D(f(x), p) = \lim_{h \leftarrow 0} \frac{f(x + hp) - f(x)}{h} = \nabla f(x)^T p.$$

Hessian

- In some sense of the second derivative of f

$$\nabla^2 f(x) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1}(x) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(x) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(x) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(x) & \frac{\partial^2 f}{\partial x_2 \partial x_2}(x) & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n}(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(x) & \frac{\partial^2 f}{\partial x_n \partial x_2}(x) & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n}(x) \end{pmatrix}$$

- If f is twice continuously differentiable, Hessian is symmetric.

Taylor's theorem

- Mean value theorem: for $\alpha \in (0,1)$

$$f(x+p) = f(x) + \nabla f(x+\alpha p)^T p$$

- Taylor's theorem: for $\alpha \in (0,1)$

$$f(x+p) = f(x) + \nabla f(x)^T p + \frac{1}{2} p^T \nabla^2 f(x+\alpha p) p$$

Vector valued function

- For a vector valued function $r: \mathbb{R}^n \rightarrow \mathbb{R}^m$

$$r(x) = \begin{pmatrix} f_1(x_1, x_2, \cdots, x_n) \\ f_2(x_1, x_2, \cdots, x_n) \\ \vdots \\ f_m(x_1, x_2, \cdots, x_n) \end{pmatrix}$$

- The Jacobian of r at x is

$$J(x) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(x) & \frac{\partial f_1}{\partial x_2}(x) & \cdots & \frac{\partial f_1}{\partial x_n}(x) \\ \frac{\partial f_2}{\partial x_1}(x) & \frac{\partial f_2}{\partial x_2}(x) & \cdots & \frac{\partial f_2}{\partial x_n}(x) \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1}(x) & \frac{\partial f_m}{\partial x_2}(x) & \cdots & \frac{\partial f_m}{\partial x_n}(x) \end{pmatrix} = \begin{pmatrix} \nabla f_1(x)^T \\ \nabla f_2(x)^T \\ \vdots \\ \nabla f_m(x)^T \end{pmatrix}$$

Some linear algebra

- Vectors and matrix
- Eigenvalue and eigenvector
- Singular value decomposition
- LU decomposition and Cholesky decomposition
- Subspaces and QR decomposition
- Sherman-Morrison-Woodbury formula

Vector

- A column vector $x \in R^n$ is denoted as $x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$
- The transpose of x is $x^T = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix}$
- The inner product of $x, y \in R^n$ is $x^T y = \sum_{i=1}^{n} x_i y_i$
- Vector norm
  - 1-norm $\|x\|_1 = \sum_{i=1}^{n} |x_i|$
  - 2-norm $\|x\|_2 = \sqrt{x^T x} = \sqrt{\sum_{i=1}^{n} x_i^2}$
  - ∞-norm $\|x\|_\infty = \max_{i=1\ldots n} |x_i|$

Matrix

- A matrix $A \in \mathbb{R}^{m \times n}$ is $A = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{pmatrix}$
- The transpose of A is $A^T = \begin{pmatrix} A_{11} & A_{21} & \cdots & A_{m1} \\ A_{12} & A_{22} & \cdots & A_{m2} \\ \vdots & \vdots & \cdots & \vdots \\ A_{1n} & A_{2n} & \cdots & A_{mn} \end{pmatrix}$
- Matrix A is symmetric if $A^{\mathrm{T}}=A$
- Matrix norm: $\|A\|_p = \max\|Ax\|_p$ for $\|x\|_p=1$, $p=1,2,\infty$

Eigenvalue and eigenvector

- A scalar λ is an eigenvalue of an $n\times n$ matrix A if there is a nonzero vector x such that $Ax = \lambda x$.
  - Vector x is called an eigenvector.
- Matrix A is symmetric positive definite (SPD) if $A^T = A$ and all its eigenvalues are positive.
- If A has n linearly independent eigenvectors, A can have the eigen-decomposition: $A = X\Lambda X^{-1}$.
  - Λ is diagonal with eigenvalues as its diagonal elements
  - Column vectors of X are corresponding eigenvectors

Spectral decomposition

- If A is real and symmetric, all its eigenvalues are real, and there are n orthogonal eigenvectors.
- The spectral decomposition of a symmetric matrix A is $A=Q\Lambda Q^{\mathrm{T}}$.
  - Λ is diagonal with eigenvalues as its diagonal elements
  - Q is orthogonal, i.e. $Q^{\mathrm{T}}Q = QQ^{\mathrm{T}} = I$.
  - Column vectors of Q are corresponding eigenvectors.

Singular value

- The singular values of an $m \times n$ A are the square roots of the eigenvalues of $A^T A$.
- Any matrix A can have the singular value decomposition (SVD): $A = U \Sigma V^T$.
  - Σ is diagonal with singular values as its elements.
  - U and V are orthogonal matrices.
  - The column vectors of U are called left singular vectors of A; the column vectors of V is called the right singular vector of A.

LU decomposition

- The LU decomposition with pivoting of matrix A is $PA=LU$
  - P is a permutation matrix
  - L is lower triangular; U is upper triangular.
- The linear system $Ax=b$ can be solved by
  1. Perform LU decompose $PA=LU$
  2. Solve $Ly=Pb$
  3. Solve $Ux=y$

Cholesky decomposition

- For a SPD matrix A, there exists the Cholesky decomposition $P^{\mathrm{T}}AP = LL^{\mathrm{T}}$
  - P is a permutation matrix
  - L is a lower triangular matrix
- If A is not SPD, the LBL decomposition can be used: $P^{\mathrm{T}}AP = LBL^{\mathrm{T}}$
  - B is a block diagonal matrix with blocks of dimension 1 or 2.

Subspaces, QR decomposition

- The null space of an $m \times n$ matrix A is

$$\mathbf{Null}(A)=\{w|Aw=0, w\neq 0\}$$

- The range of A is $\mathbf{Range}(A)=\{w|w=Av, \forall v\}$.
- Fundamental of linear algebra:

$$\mathbf{Null}(A) \oplus \mathbf{Range}(A^{\mathrm{T}}) = \mathbb{R}^n$$

- Matrix A has the QR decomposition $AP = QR$
  - P is permutation matrix; Q is an orthogonal matrix; R is an upper triangular matrix.

Singularity and ill-conditioned

- An $n \times n$ matrix A is singular (noninvertible) iff
  - A has 0 eigenvalues
  - A has 0 singular values
  - The null space of A is not empty
  - The determinant of A is zero
- The condition number of A is $\kappa(A) = \|A\| \|A^{-1}\|$
  - A is ill-conditioned if it has a large condition number.

Sherman-Morrison-Woodbury formula

- For a nonsinular matrix $A \in \mathbb{R}^{n\times n}$, if a rank-one update of $\hat{A} = A + uv^{T}$ is also nonsingular,

$$\hat{A}^{-1} = A^{-1} - \frac{A^{-1}uv^{T}A^{-1}}{1 + v^{T}A^{-1}u}.$$

- For matrix $U, V \in \mathbb{R}^{n\times p}$, $1 \le p \le n$, if $\hat{A} = A + UV^{T}$ is nonsingular,

$$\hat{A}^{-1} = A^{-1} - A^{-1}U(I + V^{T}A^{-1}U)^{-1}V^{T}A^{-1}$$