UNIVERSITY AT ALBANY
STATE UNIVERSITY OF NEW YORK

CSI 436/536 (Spring 2025)

Machine Learning

Lecture 3: Review of Calculus and Optimization

Chong Liu

Department of Computer Science

Jan 29, 2025

Announcement

- Enroll in Gradescope ASAP if you haven't done yet
- Participation points
  - Come to me to claim 1 point after each lecture, if
    - You asked a question, or
    - You showed/explained your solutions to in-class exercise problems
  - Maximum 4 points this semester
- Study group registration due next Monday!
- HW 1 will be released next Monday

Recap: linear algebra review

- Vector:
  - Norm (one vector):
    - l_p norm (l_1, l_2, l_∞)
  - Distance and angle (two vectors)
  - Linear (in)dependence
  - Orthogonality: $x^{\mathrm{T}}y = 0$
- Matrix:
  - Matrix-vector multiplication, matrix-matrix multiplication

Properties of a matrix

- General matrix
  - Rank: max number of independent column vectors / row vectors
  - Transpose: switch rows and columns

$$A \in \mathbb{R}^{m \times n} \qquad A^T \in \mathbb{R}^{n \times m}$$

- Square matrix
  - Trace: Sum of diagonal elements
  - Determinant:

$$tr\left(\begin{bmatrix} 5 & 3 & 5 \\ 4 & -1 & 2 \\ -3 & 8 & 7 \end{bmatrix}\right) = 5 - 1 + 7 = 11.$$

$$\det\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$$

Invertible matrix

$$A^{-1}A = I$$

Orthogonal matrix

$$A^{-1} = A^T$$

Symmetric matrix

$$A^T = A$$

Eigenvalues and eigenvectors of a (square) matrix

Let A be a $n \times n$ matrix. The vector $v \neq 0$ that satisfies

$$Av = \lambda v$$

for some scalar λ is called the eigenvector of A and λ is the eigenvalue corresponding to the eigenvector v.

1. A is symmetric, then $\lambda \in \mathbb{R}$.
2. A is symmetric and positive semi-definite, then $\lambda \geq 0$
3. A is symmetric and positive definite, then $\lambda > 0$

Positive (semi)-definite matrix

Very important property for optimization, kernel methods

- A symmetric matrix $A \in \mathbb{R}^{n \times n}$ is positive semi-definite, if and only if $x^T A x \geq 0$, for any $x \in \mathbb{R}^n$.
  - All eigenvalues of A are non-negative.
  - $X^T A X$ for any $X \in \mathbb{R}^{n \times m}$ is positive semi-definite.
- A symmetric matrix $A \in \mathbb{R}^{n \times n}$ is positive definite, if and only if $x^T A x > 0$, for any $0 \neq x \in \mathbb{R}^n$.
  - All eigenvalues of A are positive.
  - All diagonal entries of A are positive.

In class exercise: prove $A = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$ is a positive semi-definite matrix

- Solution 1: prove $x^T A x \geq 0$ for any vector x.

- Solution 2: prove all eigenvalues of A are all non-negative.
  - Hint: solve $\det(A - \lambda I) = 0$ to find eigenvalues.

Today's agenda

- Multi-variate calculus
  - Partial derivative and gradient
  - Chain rule
  - Multiple integrals
  - Jacobian matrix and Hessian matrix
- Optimization
  - Convex set and convex function
  - Optimization problem formulation
  - Properties of convex optimization
  - Lagrange Multipliers

Multi-variate function

- Definition:
  - A function of two or more variables takes multiple inputs and produces a single output.
  - Examples: $f(x, y) = e^{x+y} + e^{3xy} + e^{y^4}$
- Domain:
  - Set of all possible inputs
- Range:
  - Set of possible output values.

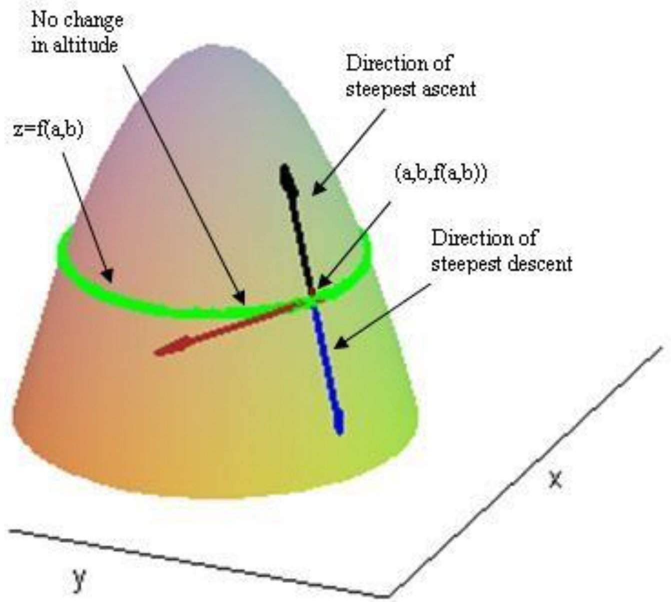
Partial derivative

- Definition:
  - The rate of change of a function with respect to one variable, holding other variables constant.
- Notations:
  - $\frac{\partial f}{\partial x}$ or $\nabla_x f(x, y)$
- Example:
  - $f(x, y) = e^{x+y} + e^{3xy} + e^{y^4}$
    - $\frac{\partial f}{\partial x} = e^{x+y} + 3ye^{3xy}$
    - $\frac{\partial f}{\partial y} = e^{x+y} + 3xe^{3xy} + 4y^3e^{y^4}$

Gradient

- Definition:
  - A vector that points in the direction of the steepest change.
  - Consist of multiple partial derivatives

- Example of $f(x, y)$:
  - $\nabla f(x, y) = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right)$



Chain rule

- To compute derivative of a composite function
- Example:
  - $z = f\big(g(t)\big)$
  - $\frac{\mathrm{d}z}{\mathrm{d}t} = \frac{\mathrm{d}f}{dg}\frac{\mathrm{d}g}{\mathrm{d}t}$
- In-class exercise:
  - $f(x) = e^{2x}, g(x) = \sin(x)$. Find $\nabla f(g(x))$.
  - $\frac{\mathrm{d}f}{dg} = 2e^{2g(x)} = 2e^{2\sin(x)}$
  - $\frac{\mathrm{d}z}{\mathrm{d}t} = \frac{\mathrm{d}f}{dg}\frac{\mathrm{d}g}{\mathrm{d}t} = 2e^{2\sin(x)}\cos(x)$

Multiple Integrals

- Double integral: compute the volume under a surface in two dimensions.
- Example: a function $f(x, y)$ over a region R
  - $\iint_R f(x, y)\, dx\, dy$
- In-class exercise: find double integral of the function $f(x, y) = x^2 + y^2$ over $0 \leq x \leq 2$ and $1 \leq y \leq 3$.
  - $\int_0^2 x^2\, dx = 8/3$
  - $\int_0^2 y^2\, dx = 2y^2$
  - $\int_1^3 \left(\frac{8}{3} + 2y^2\right) dy = 16/3 + 52/3 = \ 68/3$

Jacobian matrix – first order

$$\mathbf{J}_{ij} = \frac{\partial f_i}{\partial x_j}$$

$$\mathbf{J} = \begin{bmatrix} \frac{\partial \mathbf{f}}{\partial x_1} & \cdots & \frac{\partial \mathbf{f}}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \nabla^{\mathrm{T}} f_1 \\ \vdots \\ \nabla^{\mathrm{T}} f_m \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

- In-class exercise:
  - $f(x, y) = (f_1, f_2, f_3)$
  - $f_1 = x^2 y, f_2 = y^3, f_3 = 4xy + 5$

$$J_{3x2} = \begin{bmatrix} \frac{\partial f_1}{\partial x} & \frac{\partial f_1}{\partial y} \\ \frac{\partial f_2}{\partial x} & \frac{\partial f_2}{\partial y} \\ \frac{\partial f_3}{\partial x} & \frac{\partial f_3}{\partial y} \end{bmatrix} = \begin{bmatrix} 2xy & x^2 \\ 0 & 3y^2 \\ 4y & 4x \end{bmatrix}$$

Hessian matrix – second order

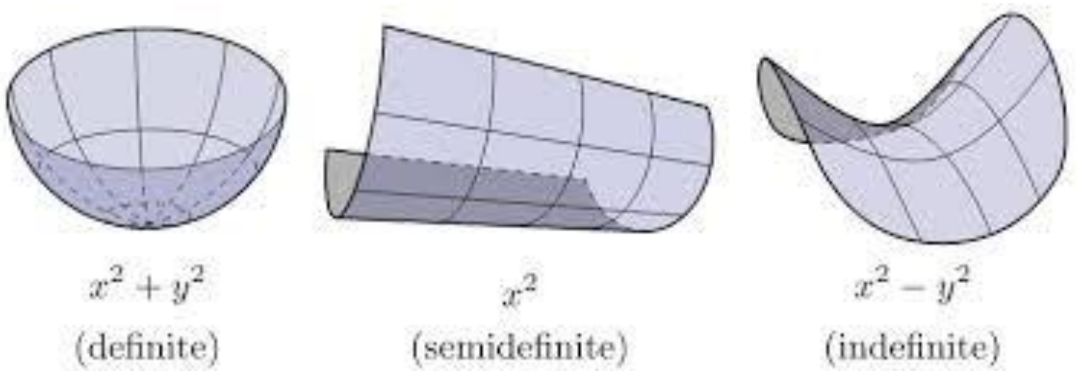
$$(\mathbf{H}_f)_{i,j} = \frac{\partial^2 f}{\partial x_i \, \partial x_j}$$

$$\mathbf{H}_f = \begin{bmatrix} \dfrac{\partial^2 f}{\partial x_1^2} & \dfrac{\partial^2 f}{\partial x_1 \, \partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_1 \, \partial x_n} \\[2.2ex] \dfrac{\partial^2 f}{\partial x_2 \, \partial x_1} & \dfrac{\partial^2 f}{\partial x_2^2} & \cdots & \dfrac{\partial^2 f}{\partial x_2 \, \partial x_n} \\[2.2ex] \vdots & \vdots & \ddots & \vdots \\[2.2ex] \dfrac{\partial^2 f}{\partial x_n \, \partial x_1} & \dfrac{\partial^2 f}{\partial x_n \, \partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

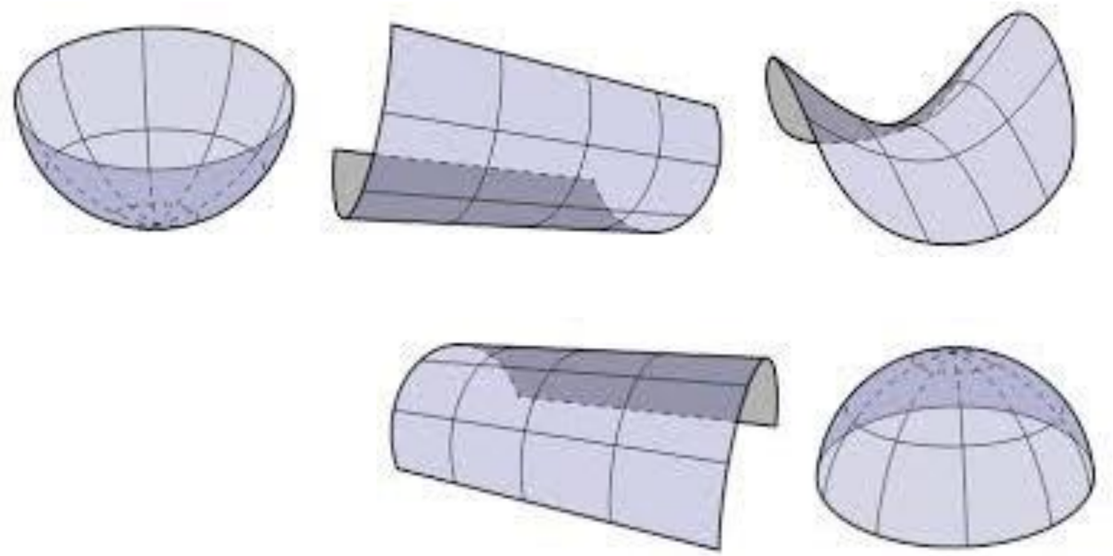
- Quadratic approximation of a function
  - $f(x+y) = f(x) + y^T \nabla f(x) + \frac{1}{2} y^T \nabla^2 f(x) y$

Hessian matrix – second order

- Hessian matrix is symmetric
- Hessian matrix and local curvature of the function
  - Minimum: Hessian is positive definite
  - Maximum: Hessian is negative definite
  - Saddle point: Hessian is indefinite (not positive/negative definite)



Quadratic Function



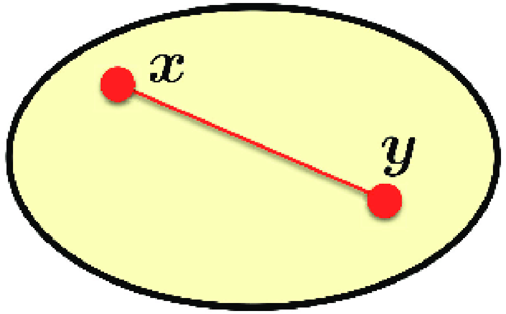
- $f(x) = \frac{1}{2}x^T A x + b^T x + c$
  - Gradient: $\nabla f(x) = Ax + \mathrm{b}$
  - Hessian: $\nabla^2 f(x) = A$
- Quadratic programming:
  - $\min f(x) = \frac{1}{2}x^T A x + b^T x + c$
  - Key: check Hessian matrix!
    - Hessian is positive (semi)definite: minimum (local or global)
    - Hessian is negative (semi)definite: maximum (local or global)
    - Hessian is indefinite: undetermined, changing curvature

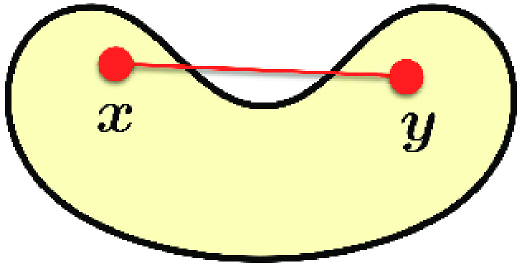
Today's agenda

- Multi-variate calculus
  - Partial derivative and gradient
  - Chain rule
  - Multiple integrals
  - Jacobian matrix and Hessian matrix
- Optimization
  - Convex set and convex function
  - Optimization problem formulation
  - Properties of convex optimization
  - Lagrange Multipliers

Convex Sets

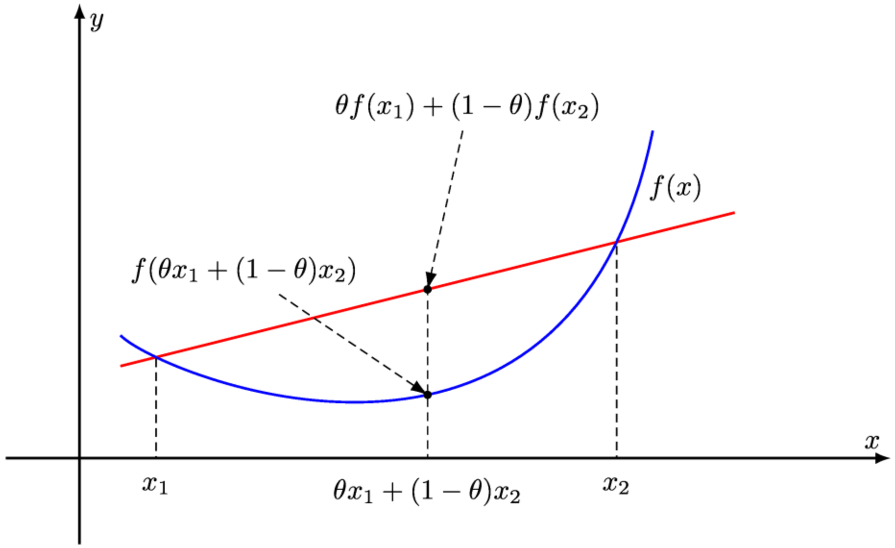
- Definition:
  - A set $C \subseteq R^n$ is convex if for any two points $x_1, x_2 \in C, \theta x_1 + (1-\theta)x_2 \in C$ for all $\theta \in [0,1]$.
- Interpretation:
  - A set $C \subseteq R^n$ is convex if, for any two points $x_1, x_2 \in C$, the line segment connecting them is also entirely within C.
- Discussion: are they convex sets?
  - (1) [0,1]
  - (2-3)





Convex functions

- Definition:
  - A function $f: C \to R$ is convex if C is a convex set and for all $x_1, x_2 \in C$ and $\theta \in [0,1]$:
  - $f(\theta x_1 + (1-\theta)x_2) \leq \theta f(x_1) + (1-\theta) f(x_2)$
- Interpretation:
  - A convex function lies below the line segment connecting any two points on its graph.
- Discussion: propose some convex functions
- Example: linear functions, quadratic functions, exponential functions.



Convex optimization problem formulation

- $\min f(x)$,
- $\text{s.t. } \mathrm{g}(x) \leq 0, h(x) = 0.$

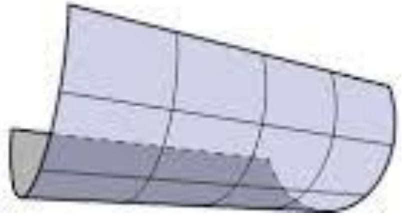
- $f(x)$ is the convex objective function
- $g(x)$ is convex inequality constraint
- h(x) is equality constraint

Review of 1-dimensional optimization

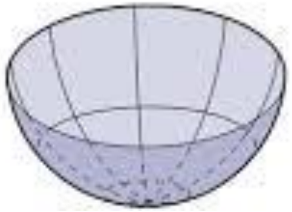
- $f(x) = x^3 + 3x^2 - 24x + 2$
  - First, solve $f'(x) = 0$ to get all solutions $f'(x) = 3x^2 + 6x - 24 = 0, x_1 = -4, x_2 = 2.$
  - Second, for each solution, check $f''(x)$: $f''(x) = 6x + 6$
    - $f''(x) > 0$: minimum (local or global) $x = 2$
    - $f''(x) < 0$: maximum (local or global) $x = -4$
    - $f''(x) = 0$: undetermined, changing curvature

Hessian matrix and convex function

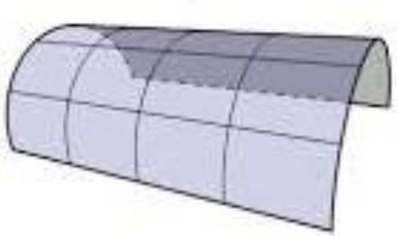
- $\nabla^2 f(x) \succcurlyeq 0$, then f(x) is convex
  - No local minimum



- $\nabla^2 f(x) \succ 0$, then f(x) is strongly convex
  - Unique global minimum



- $-\nabla^2 f(x) \succcurlyeq 0$, then f(x) is concave
  - No local maximum



- $-\nabla^2 f(x) \succ 0$, then f(x) is strongly concave
  - Unique global maximum

Lagrange multipliers to handle constraints

- The Lagrangian function combines the objective function with the constraints using multipliers.
- Example: $\max xy, \text{s.t.}\ x + y = c$
  - Solution 1: use $y = c - x$, then objective problem is $\max x(c - x)$, so $x = y = c/2$ is the optimal solution.
  - Solution 2 (Lagrange multiplier):
    - $L(x, y, \lambda) = xy - \lambda(x + y - c)$
    - Differentiate with regards to x and y, we have $x = y = \lambda$