CSI 436/536
Introduction to Machine Learning

Review of multivariate calculus (1)

Professor Siwei Lyu
Computer Science
University at Albany, State University of New York

review of 1D optimization

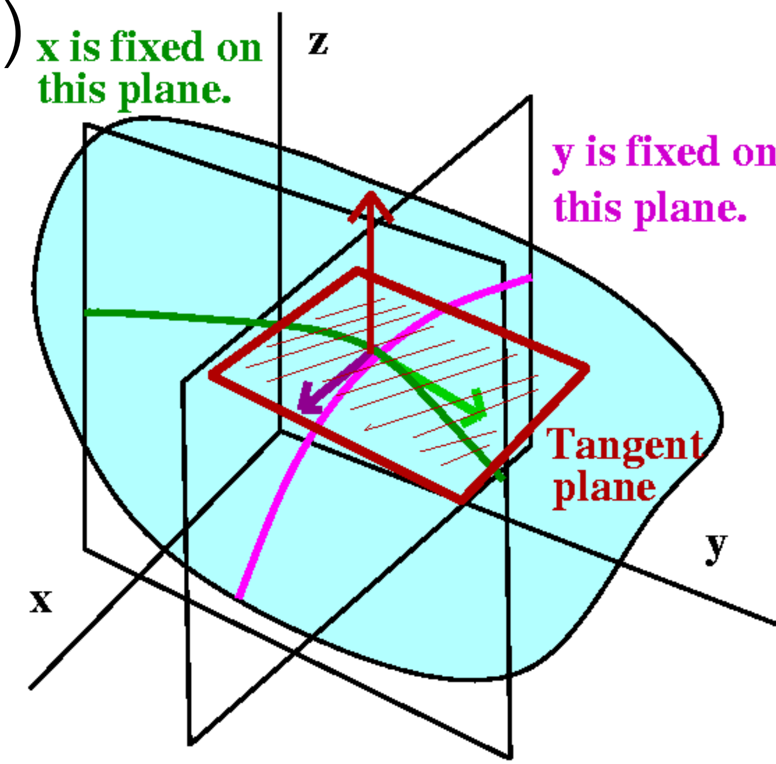
$$f(x) = x^3 + 3x^2 - 24x + 2$$

- f(x) > -∞ and f(x) second order differentiable
- first, solve f'(x) = 0, to get all solutions f'(x) = $3x^2$ + 6x -24 = 0, x=-4,x=2
- 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
- for all minimums, check if the solution is also global

vector functions

- we study function of vector input and scalar output
- Partial derivatives
  - fixing all other variables and take derivative of one variable as if it is a scalar function
  - Everything you know about differentiation still holds (chain rule, additivity, etc)
- gradient
  - vector formed by all partial derivatives

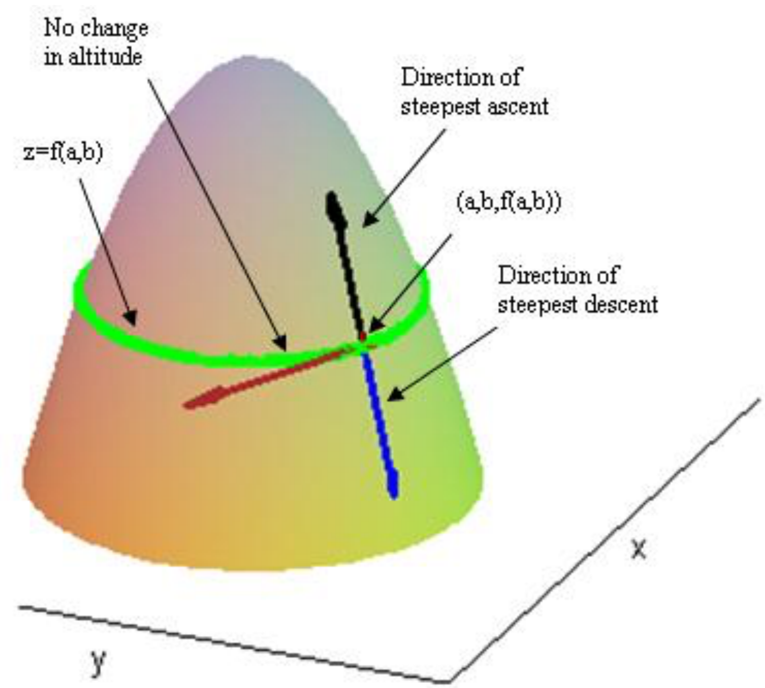
$$\nabla f(x) = \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{pmatrix}$$



gradient

- Geometric interpretation
- fastest descent (Taylor series) $f(\mathbf{x}+\mathbf{h}) \doteq f(\mathbf{x}) + \mathbf{h}^\top \nabla f(\mathbf{x})$ the maximum is reached using Cauchy-Schwartz inequality
  $\mathbf{h}^\top \nabla f(\mathbf{x}) \leq ||\mathbf{h}|| \times ||\nabla f(\mathbf{x})||$,

  so minimum reached for
  $\mathbf{h} = - \nabla f(\mathbf{x})$
  maximum reached for
  $\mathbf{h} = \nabla f(\mathbf{x})$

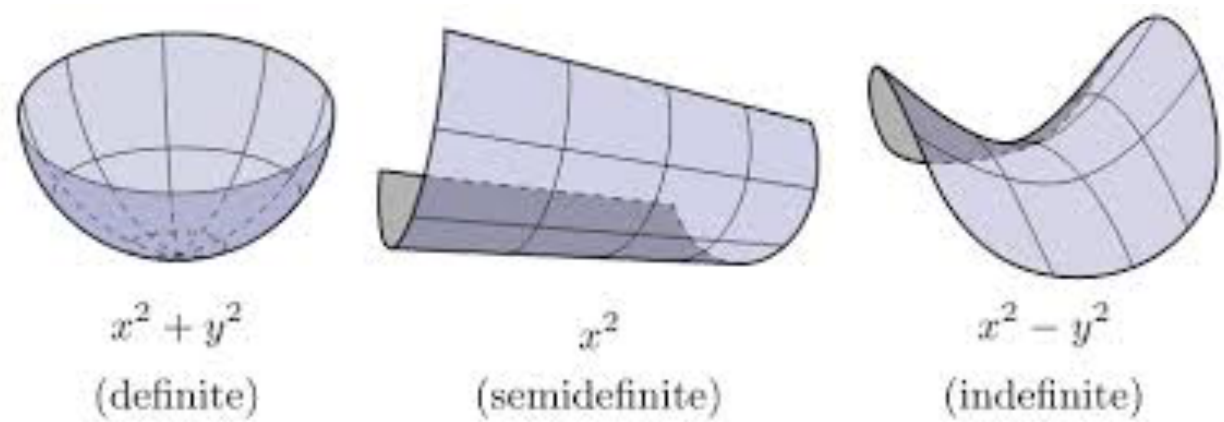


Hessian matrix

- symbolically, Hessian is outer product of gradient operator

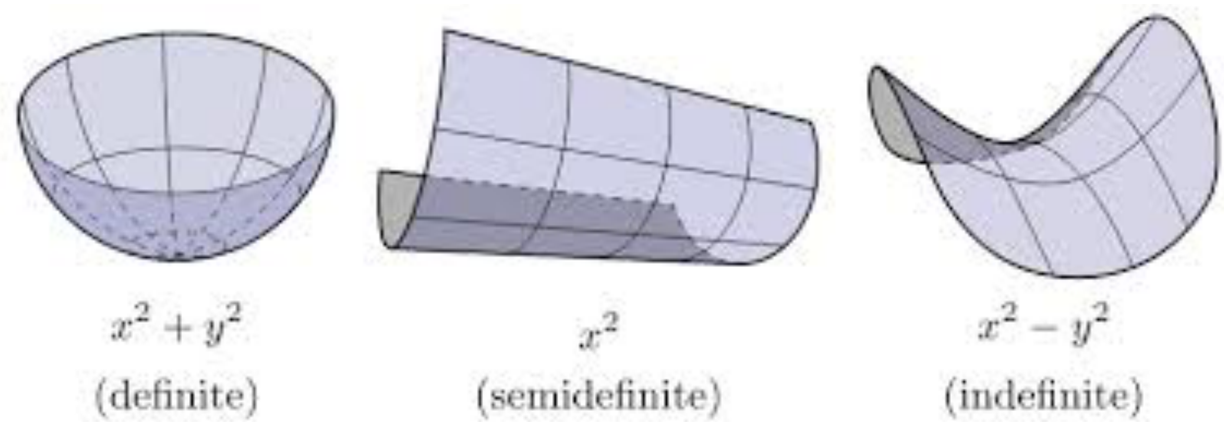
$$\nabla f = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix} \quad \text{and} \quad \nabla^2 f = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}$$

- intuition of Hessian matrix



Hessian matrix

- quadratic approximation of a function
  $f(\mathbf{x}+\mathbf{h}) \doteq f(\mathbf{x}) + \mathbf{h}^T \nabla f(\mathbf{x}) + \frac{1}{2}\, \mathbf{h}^T \nabla^2 f(\mathbf{x}) \mathbf{h}$



- Hessian matrix is symmetric
- Hessian matrix corresponds to the local curvature of the function
  - minimum: Hessian positive definite
  - maximum: Hessian negative definite
  - saddle point: : Hessian indefinite

Quadratic function

- Quadratic function (for A a symmetric matrix)

$$f(x) = \frac{1}{2}x^T Ax + b^T x + c$$

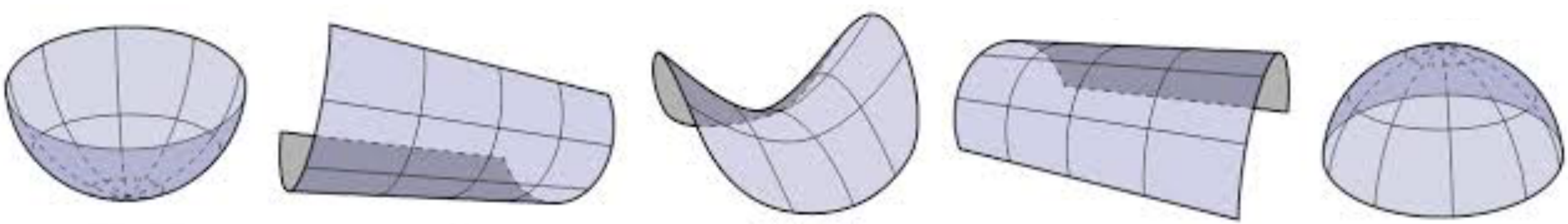
  - gradient $\nabla f(x) = Ax + b$
  - Hessian matrix $\nabla \nabla^T f(x) = A$
- Chain rule still works, e.g.,
  - Gaussian function $f(x) = e^{-\frac{1}{2}x^T A^{-1} x}$
  - Sigmoid function $g(x) = \dfrac{1}{1 + e^{-w^T x}}$

Quadratic programming

- minimize

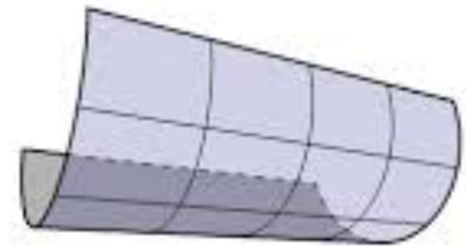
$$f(\mathbf{x}) = c + \mathbf{b}^T\mathbf{x} + ½\, \mathbf{x}^T A\mathbf{x}$$

- first, solve $\nabla f(\mathbf{x}) = \mathbf{b} + A\mathbf{x} = 0$
- check $\nabla^2 f(\mathbf{x})$:
  - $\nabla^2 f(\mathbf{x})$ is positive (semi)definite: minimum (local or global)
  - $\nabla^2 f(\mathbf{x})$ is negative (semi)definite: maximum (local or global) x = -4
  - $\nabla^2 f(\mathbf{x})$ is indefinite: undetermined, changing curvature
- semi-definiteness determines uniqueness of solution

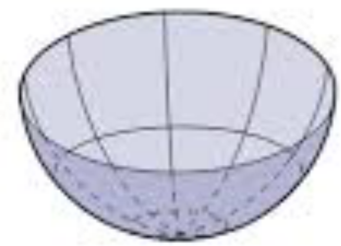


Convex function

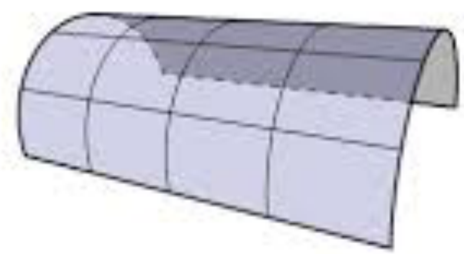
- Conditions on the Hessian matrix
  - $\nabla\nabla^T f(x) \succeq 0$ f(x) is convex
    - No local minimum



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



  - $-\nabla\nabla^T f(x) \succeq 0$ f(x) is concave
    - No local maximum



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

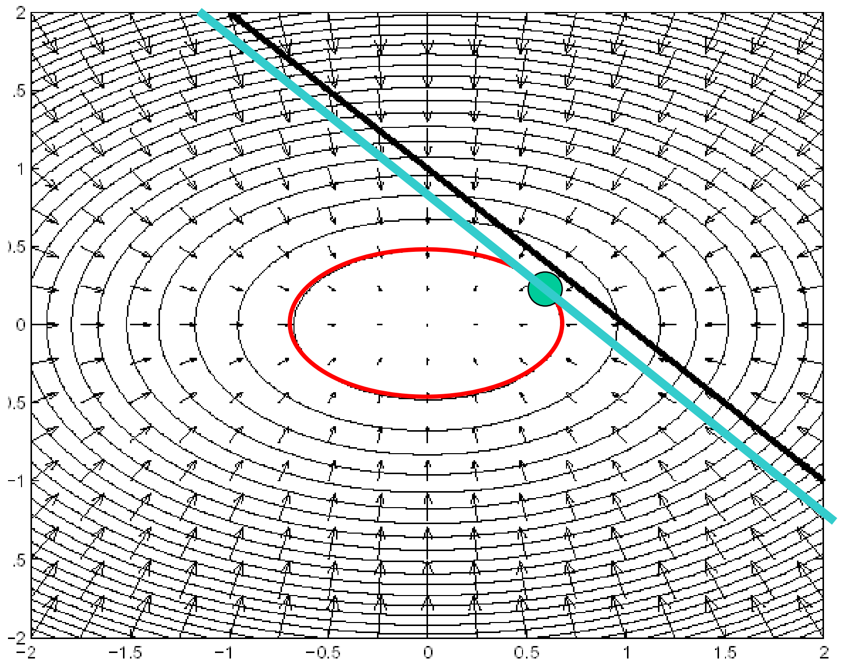
  - We will encounter many convex quadratic programing problems

Constrained optimization

- Example: maximize the area of a rectangular with fixed circumference $\max_{x,y} xy,$ s.t. $x + y = c$
- Approach 1: use y = c-x, and solve for x directly $\max_x x(c - x)$, $x = y = \dfrac{c}{2}$, symmetry and optimal
- Approach 2: use Lagrangian multipliers $L(x, y, \lambda) = xy - \lambda(x + y - c)$
  - Differentiate with regards to x and y, we have $x = y = \lambda$
  - Then using the constraint to get $x = y = \dfrac{c}{2}$
- Note xy is neither convex or concave, so only with constraint it has a solution

Equality constrained problem

- $min_{x,y} f(x, y) = x^2 + 2y^2 - 2$, s.t. $x + y = 1$



- objective gradient has to be perpendicular to the constraint, otherwise, we can still go done in the direction along the gradient

Equality constrained problem

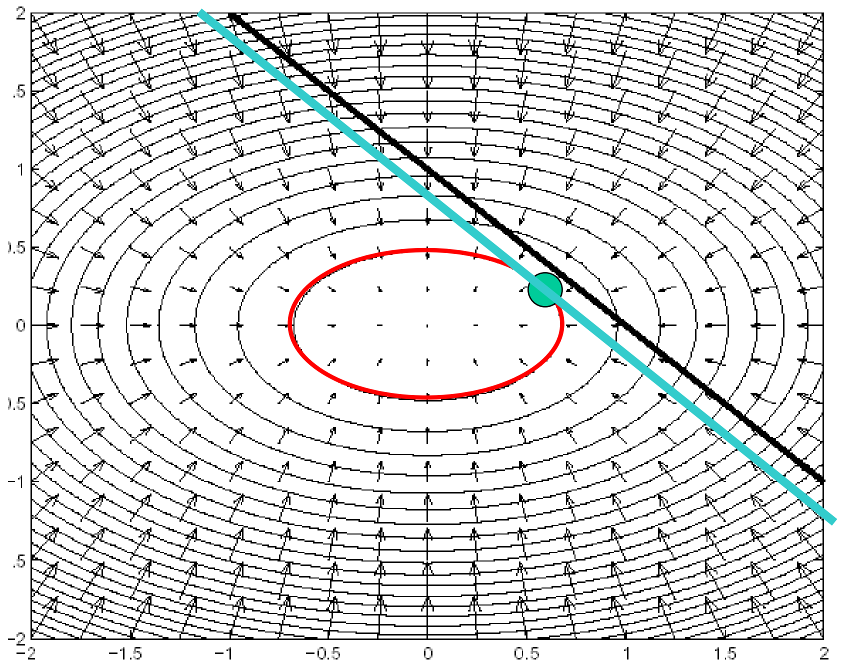
- Solve equality constrained
  $\min_x f(x) = \frac{1}{2}x^T A x + b^T x + c$, s.t. $Dx = e$
- introduce Lagrangian multiplier **v** and form Lagrangian $L(x, v) = f(x) - v^{\mathsf{T}}(Dx - e)$
- optimal solution given at the stationary point of L
  $$\frac{\partial L}{\partial x} = b + Ax - D^{\mathsf{T}}v = 0$$ (dual feasibility)
  $$\frac{\partial L}{\partial v} = Dx - e = 0$$ (primal feasibility)
- Solution: solving the KKT equation

$$\begin{pmatrix} A & -D^{\mathsf{T}} \\ D & 0 \end{pmatrix} \begin{pmatrix} x \\ v \end{pmatrix} = \begin{pmatrix} -b \\ e \end{pmatrix}$$

Previous example

- Rewrite the problem as

$$min_{x,y} f(x, y) = x^2 + 2y^2 - 2, \text{ s.t. } x + y = 1$$



- Solution given by

$$\begin{pmatrix} 2 & 0 & -1 \\ 0 & 4 & -1 \\ 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ v \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

constrained quadratic optimization

- minimize

  $$f(\mathbf{x}) = \tfrac{1}{2}\, \mathbf{x}^T A \mathbf{x}$$

  s.t., $\mathbf{x}^T\mathbf{x} - 1 = 0$

- introduce Lagrangian multiplier λ and form Lagrangian $L(\mathbf{x},\mathbf{v}) = f(\mathbf{x}) - \lambda(\mathbf{x}^T\mathbf{x} - 1)$
- optimal solution given at the stationary point of L $\partial L/\partial \mathbf{x} = A\mathbf{x} - \lambda\mathbf{x} = 0$, or $A\mathbf{x} = \lambda\mathbf{x}$
- so optimal solution is eigenvalue of A, objective function is λ
- to minimize, we should choose the one corresponding to the minimal eigenvalue (Ritz-Fisher theorem)