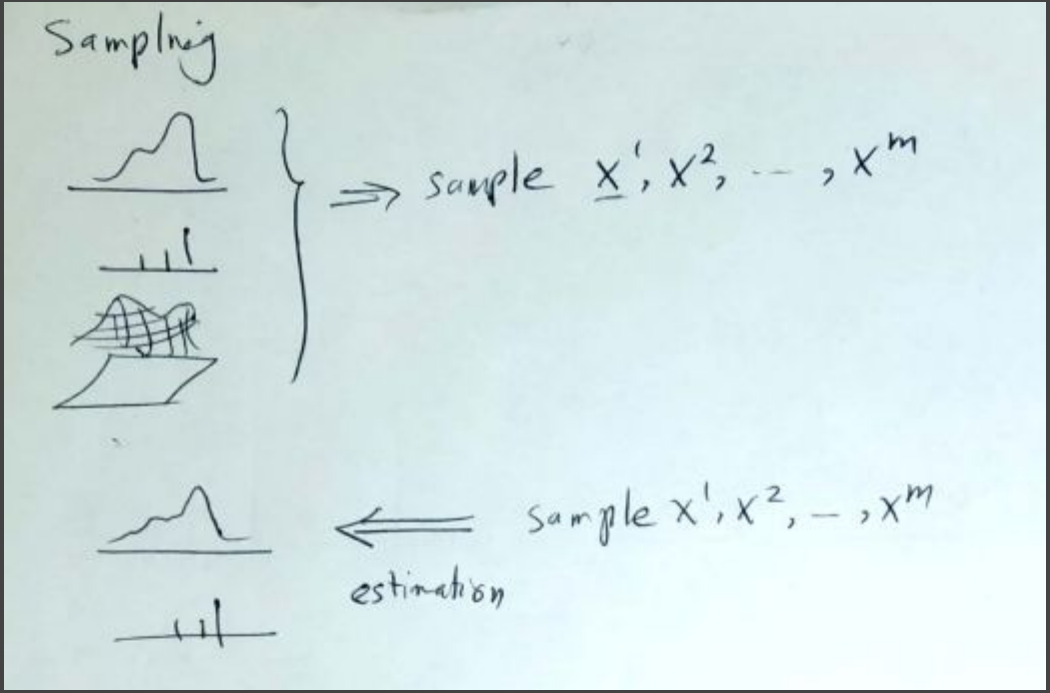
Mathematics for AI

Lecture 27

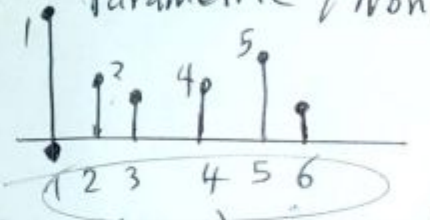
Statistical Estimation, Maximum Likelihood Solution
Introduction to Optimization

Statistical Estimation



Example: Rolling a dice

Estimation Parametric / Nonparametric



$Pr(X=1) = P_1 \quad P(X=2) = P_2 \quad \cdots \quad P_6$

$P_1 + P_2 + \cdots + P_6 = 1$

$\theta = \begin{bmatrix} P_1 \\ P_2 \\ \vdots \\ P_6 \end{bmatrix}$

$X^1 \; X^2 \; X^3 \cdots X^m$

$2 \quad 1 \quad 2 \quad 3 \quad \quad 6$

$p_\theta(x) = Pr(X = x) = P_x \quad x \in \{1, 2, \ldots, 6\}$

$x^1, x^2, \ldots, x^m$ i.i.d independent

$x^i \sim P$

The likelihood function

K. N. Toosi
University of Technology

Fix $\theta = [p_1, p_2, \dots, p_6]$ find Prob. $x^1, x^2, \dots x^m$

$$\mathcal{L}(\theta) = \underbrace{\text{Prob}(x^1, x^2, \dots, x^m \mid \theta)}_{\text{likelihood}} = \text{Prob}(x^1 \mid \theta)\, \text{Prob}(x^2 \mid \theta) \cdots \text{Prob}(x^m \mid \theta)$$

$$= \prod_{i=1}^{m} \text{Prob}(x^i \mid \theta) = \prod_{i=1}^{m} p_\theta(x^i) = \prod_{i=1}^{m} p_{x^i}$$

$$= p_1^{m_1}\, p_2^{m_2} \cdots p_6^{m_6} = \mathcal{L}(\theta) = \mathcal{L}(p_1, p_2, \dots, p_6)$$

$m_1 = \#(x^i = 1)$

$m_2 = \#(x^i = 2)$

$m_6 = \#(x^i = 6)$

Maximum Likelihood (ML) Solution

MA 27 II

$$L(\theta) = L(p_1, p_2, \dots, p_6) = p_1^{m_1}\, p_2^{m_2} \cdots p_6^{m_6} = L(\theta) = L(\theta; D)$$

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\; \underbrace{L(\theta, D)}_{\text{likelihood}}$$

θ^* → Maximum Likelihood solution

$$L(\theta; x^1, \dots, x^m)$$

$$= Pr(D \mid \theta)$$

$$= Pr(x^1, \dots, x^m \mid \theta)$$

Example: Rolling a dice

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\ L(\theta; D) \quad \text{subject to} \quad \boxed{p_1, p_2, \dots, p_6 \geq 0}$$

$$p_1 + p_2 + \dots + p_6 = 1$$

$$p_6 = 1 - p_1 - p_2 - \dots - p_5$$

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\ p_1^{m_1}\, p_2^{m_2} \cdots p_5^{m_5}\, (1 - p_1 - p_2 \cdots - p_5)^{m_6}$$

$$= \underset{\theta}{\operatorname{argmax}}\ \log \left(p_1^{m_1}\, p_2^{m_2} \cdots p_5^{m_5} \left(1 - \sum_{i=1}^{5} p_i\right)^{m_6} \right)$$

$$= \underset{\theta}{\operatorname{argmax}}\ \underbrace{m_1 \log p_1 + m_2 \log p_2 + \dots + m_5 \log p_5 + m_6 \log \left(1 - \sum_{i=1}^{5} p_i\right)}_{\text{log-likelihood}}$$

$$\theta = [p_1\ p_2 \cdots p_5]$$

Example: Rolling a dice

$$= \underset{\theta}{\arg\max} \; \cancel{\log} \; \underbrace{m_1 \log p_1 + m_2 \log p_2 + \dots + m_5 \log p_5 + m_6 \log\Big(1-\sum_{i=1}^{5} p_i\Big)}_{\text{log-likelihood } LL(\theta)}$$

$$\theta = [p_1 \; p_2 \; \dots \; p_5]$$

$$\frac{\partial}{\partial p_1} LL(\theta) = \frac{m_1}{p_1} + m_6 \frac{-1}{1-\sum_{i=1}^{5} p_i} = 0$$

$$\frac{\partial}{\partial p_i} LL(\theta) = \boxed{\frac{m_i}{p_i}} + m_6 \frac{-1}{1-\sum p_i} = 0 \qquad i = 1, \dots, 5$$

$$\Longrightarrow \frac{m_1}{p_1} = \frac{m_2}{p_2} = \frac{m_3}{p_3} = \dots = \frac{m_5}{p_5} = \frac{m_6}{1-\sum_{i=1}^{5} p_i} = \frac{m_6}{p_6} = \frac{1}{\alpha}$$

Example: Rolling a dice

K. N. Toosi
University of Technology

$$\frac{\partial}{\partial p_1} LL(\theta) = \frac{m_1}{p_1} + m_6 \frac{-1}{1-\sum_{i=1}^{5} p_i} = 0 \qquad LL(\theta)$$

$$\frac{\partial}{\partial p_i} LL(\theta) = \boxed{\frac{m_i}{p_i}} + m_6 \frac{-1}{1-\sum p_i} = 0 \qquad i = 1, \dots, 5$$

$$\Rightarrow \frac{m_1}{p_1} = \frac{m_2}{p_2} = \frac{m_3}{p_3} = \dots = \frac{m_5}{p_5} = \frac{m_6}{1-\sum_{i=1}^{5} p_i} = \frac{m_6}{p_6} = 1/\alpha$$

$$p_1 = \alpha m_1 \qquad p_2 = \alpha m_2 \dots \qquad p_6 = \alpha m_6$$

$$\sum p_i = 1 \Rightarrow \alpha(\underbrace{m_1 + m_2 + \dots + m_6}_{m}) = 1 = \alpha m = 1 \Rightarrow \alpha = \frac{1}{m}$$

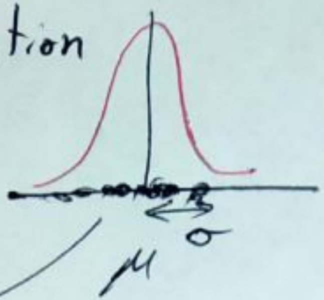
$$p_1 = \alpha m_1 = \frac{m_1}{m} \qquad p_2 = \frac{m_2}{m} \quad \dots \quad p_6 = \frac{m_6}{m}$$

maximum-likelihood solution

Example: Normal Distribution

Example: Normal Distribution MA27 (III)

$$p_\theta(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$



$\theta = [\mu, \sigma]$ samples $x^1, x^2, \dots, x^n$

Likelihood $L(\theta) = L(\mu, \sigma) = P(x^1 - x^m) = \prod_{i=1}^{m} p_\theta(x^i)$

$$= \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\frac{(x^i-\mu)^2}{\sigma^2}}$$

Example: Normal Distribution Likelihood and log-likelihood

Likelihood $L(\theta) = L(\mu, \sigma) = P(x^1 — x^m) = \prod_{i=1}^{m} P_\theta(x^i)$

$$= \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\frac{(x^i-\mu)^2}{\sigma^2}}$$

Log-likelihood $ll(\theta) = \log_e L(\mu, \sigma) =$

$$= \sum_{i=1}^{m} \left(-\log\sqrt{2\pi} - \log\sigma - \frac{1}{2}\frac{(x^i-\mu)^2}{\sigma^2} \right)$$

$$\Rightarrow ll(\theta) = -m\log\sqrt{2\pi} - m\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{m}(x^i-\mu)^2$$

Example: Normal Distribution - ML Solution

$$\Rightarrow ll(\theta) = -m\log\sqrt{2\pi} - m\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{m}(x^i-\mu)^2$$

$$\frac{\partial ll(\theta)}{\partial\mu} = -\frac{1}{2\sigma^2}\sum_{i=1}^{m}(-2)(x^i-\mu) = 0 \Rightarrow \sum_{i=1}^{m}(x^i-\mu) = 0$$

$$\Rightarrow 0 = \sum_{i=1}^{m} x^i - m\mu \Rightarrow \boxed{\hat{\mu}^* = \frac{1}{m}\sum_{i=1}^{m} x^i}$$

$$\frac{\partial ll(\theta)}{\partial\sigma} = \frac{-m}{\sigma} - \frac{1\,(-2)}{2\sigma^3}\sum(x^i-\mu)^2 = 0$$

$$\Rightarrow -m + \frac{1}{\sigma^2}\sum_{i=1}^{m}(x^i-\mu)^2 = 0$$

$$\sigma^{*2} = \frac{1}{m}\sum_{i=1}^{m}(x^i-\mu^*)^2$$

Exercise

Exercise: find ML solution for

uniform dist. $U[a,b] = p(x) = \begin{cases} \frac{1}{b-a} & x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$

exponential dist. $p(x) = \cancel{\lambda e^{-\lambda}} \; \lambda e^{-\lambda}$

$x \geq 0$

Exercise

Exercise: $\mathcal{N}(\mu, \Sigma)$ IV

multivariable Gaussian

$$P(x) = \frac{1}{(2\pi)^{d/2} \det(\Sigma)} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

$x \in \mathbb{R}^d$

$\ell\ell(\vec{\mu}, \Sigma)$

$x^1, x^2, \dots, x^m$

~~maximum~~

ML Solution: $\mu^* = \frac{1}{m} \sum_{i=1}^{m} x^i$

$$\Sigma^* = \frac{1}{m} \sum_{i=1}^{m} (x^i - \mu^*)(x^i - \mu^*)^T$$

Maximum Likelihood General Approach

Find $L(\theta)$ or $ll(\theta)$

$$\nabla_L(\theta) = 0 \Longrightarrow \theta = \checkmark$$

$$\theta = (\theta_1, \theta_2, \ldots, \theta_n)$$

$$\nabla_L(\theta) = \begin{bmatrix} \dfrac{\partial L}{\partial \theta_1} \\ \dfrac{\partial L}{\partial \theta_2} \\ \vdots \\ \dfrac{\partial L}{\partial \theta_n} \end{bmatrix} = \vec{0} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \left\{ \begin{array}{l} n\text{-equations} \\ n\text{-unknown} \end{array} \right.$$

Maximum Likelihood General Approach

K. N. Toosi
University of Technology

$$\theta = (\theta_1, \theta_2, \ldots, \theta_n)$$

$$\nabla_L(\theta) = \begin{bmatrix} \frac{\partial L}{\partial \theta_1} \\ \frac{\partial L}{\partial \theta_2} \\ \vdots \\ \frac{\partial L}{\partial \theta_n} \end{bmatrix} = \vec{0} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \left\{ \begin{array}{l} n\text{-equations} \\ n\text{-unknown} \end{array} \right.$$

A (usually non-linear) system of equations.

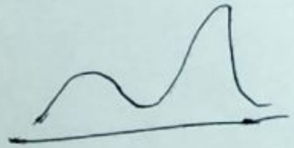
In many cases it is very hard to solve (impossible to solve!)

Example: Mixture of Gaussians

K. N. Toosi
University of Technology

Example: $p(x) = \frac{1}{3} \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2} \frac{(x-\mu_1)^2}{\sigma_1^2}}$

$+ \frac{2}{3} \frac{1}{\sqrt{2\pi}\sigma_2} e^{-1/2 \frac{(x-\mu_2)}{\sigma_2^2}}$



mixture of Gaussians.

$\theta = [\mu_1, \sigma_1, \mu_2, \sigma_2]$

$\Rightarrow x^* = \underset{\theta}{\operatorname{argmax}} L(\theta)$

$\hookrightarrow$ Hard to find global optimum

Remember: Learning from data

$$x \in R^m \longrightarrow \boxed{f_\theta} \longrightarrow y = f(\theta, x)$$

- Training data $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$
  - choose θ such that $f(\theta, x_i)$ is close to y_i

K. N. Toosi
University of Technology

Remember: Learning from data: Cost function

$$x \in R^m \longrightarrow \boxed{f_\theta} \longrightarrow y = f(\theta, x)$$

- Training data $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$
  - choose θ such that $f(\theta, x_i)$ is close to y_i

Cost function

K. N. Toosi
University of Technology

Remember: Learning from data: Cost function

$$x \in R^m \longrightarrow \boxed{f_\theta} \longrightarrow y = f(\theta, x)$$

- Training data $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$
  - choose θ such that $f(\theta, x_i)$ is close to y_i
  - cost function:

$$C(\theta) = \Sigma_{i=1..N}\, d(\, f(\theta, x_i), y_i\,)$$

Remember: Learning from data: Cost function

K. N. Toosi
University of Technology

$$x \in R^m \longrightarrow \boxed{f_\theta} \longrightarrow y = f(\theta, x)$$

- Training data $(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$
  - choose θ such that $f(\theta, x_i)$ is close to y_i
  - cost function:

$$C(\theta) = \Sigma_{i=1..n}\, d(\, f(\theta, x_i), y_i \,)$$

data output

Remember: Learning from data: Cost function

$$x \in R^m \longrightarrow \boxed{f_\theta} \longrightarrow y = f(\theta, x)$$

- Training data $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$
  - choose θ such that $f(\theta, x_i)$ is close to y_i
  - cost function:

$$C(\theta) = \Sigma_{i=1..n}\, d(\, f(\theta, x_i), y_i\,)$$

model output given x_i

Remember: Learning from data: Cost function

$$x \in R^m \longrightarrow \boxed{f_\theta} \longrightarrow y = f(\theta, x)$$

- Training data $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$
  - choose θ such that $f(\theta, x_i)$ is close to y_i
  - cost function:

$$C(\theta) = \Sigma_{i=1..n}\ d(\ f(\theta, x_i), y_i\ )$$

distance

Remember: Learning from data: Cost function

K. N. Toosi
University of Technology

$$x \in R^m \longrightarrow \boxed{f_\theta} \longrightarrow y = f(\theta, x)$$

- Training data $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$
  - choose θ such that $f(\theta, x_i)$ is close to y_i
  - cost function:

$$C(\theta) = \Sigma_{i=1..n} \| f(\theta, x_i) - y_i \|^2$$

distance

Remember: Learning from data: Cost function

K. N. Toosi
University of Technology

$$x \in R^m \rightarrow \boxed{f_\theta} \rightarrow y = f(\theta, x)$$

- Training data $(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$
  - choose θ such that $f(\theta, x_i)$ is close to y_i
  - cost function:

$$C(\theta) = \Sigma_{i=1..n}\, d(\, f(\theta, x_i), y_i\,)$$

Remember: Learning from data: Cost function

K. N. Toosi
University of Technology

$$x \in R^m \rightarrow \boxed{f_\theta} \rightarrow y = f(\theta, x)$$

- Training data $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$
  - choose θ such that $f(\theta, x_i)$ is close to y_i
  - cost function:

$$C(\theta) = \Sigma_{i=1..n}\, d(\, f(\theta, x_i), y_i\,)$$

choose θ such that $C(\theta)$ is small

Remember: Learning from data: Cost function

$$x \in R^m \longrightarrow \boxed{f_\theta} \longrightarrow y = f(\theta, x)$$

- Training data $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$
  - choose θ such that $f(\theta, x_i)$ is close to y_i
  - cost function:

$$C(\theta) = \Sigma_{i=1..N}\, d(\, f(\theta, x_i), y_i\,)$$

$$\theta^* = \operatorname{argmin}_\theta C(\theta)$$

Optimization: Continuous vs Discrete

K. N. Toosi
University of Technology

Optimization

$f(x)$

$f: \mathbb{R}^d \rightarrow \mathbb{R}$

$f: S \rightarrow \mathbb{R}$

$S \subseteq \mathbb{R}^d$

$$\boxed{x^* = \underset{x}{\operatorname{argmin}}\ f(x)}$$

$$x^* = \underset{x}{\operatorname{argmax}}\ f(x)$$

Continuous Optimization

Discrete Optimization

(Combinatorial)

$f(\theta, x) = Ax + b$

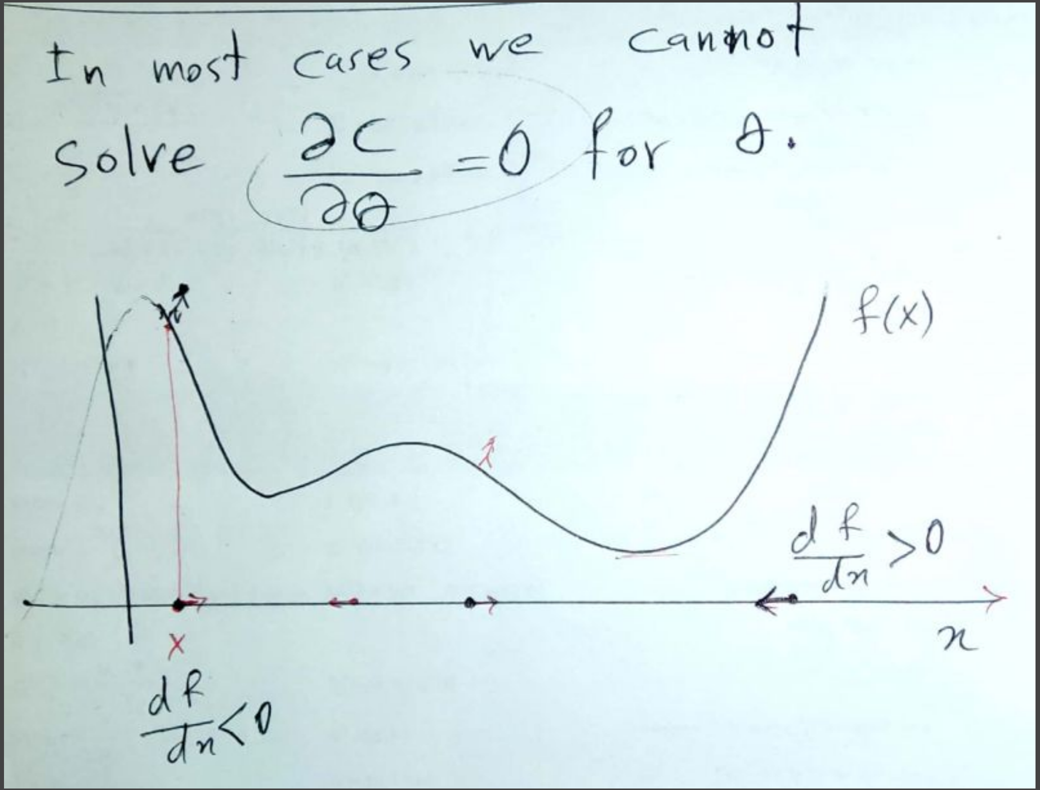
Remember: Linear Regression

$$C(\theta) = C(A, b) = \sum_{i=1}^{m'} \|Ax^i - b - y^i\|^2$$

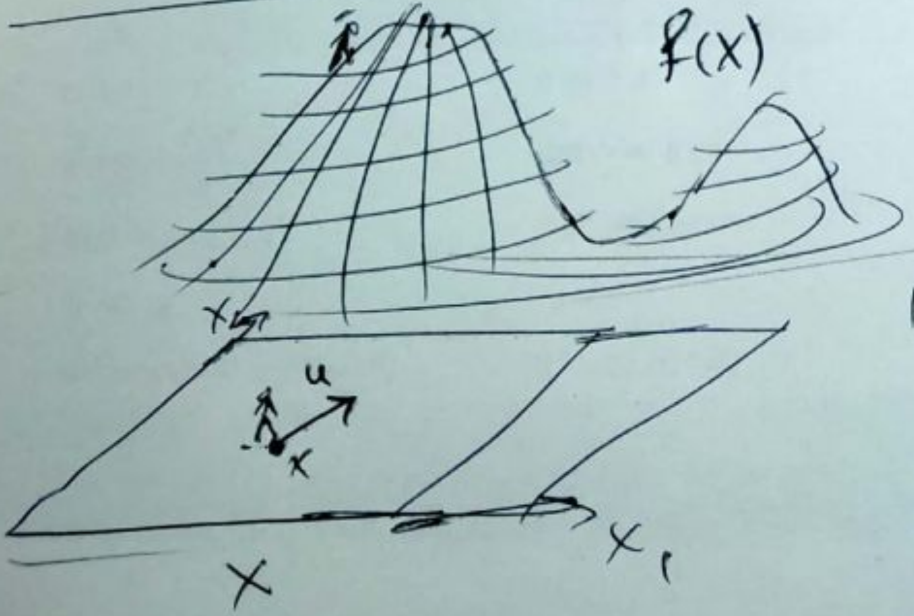
VI

$$\left.\begin{aligned} \frac{\partial C}{\partial A} &= 0_{m \times n} \\ \frac{\partial C}{\partial b} &= \vec{0} \end{aligned}\right\} \Rightarrow \text{closed form solution}$$

Iterative Optimization Algorithms



Steepest Descent



$\|u\|=1$ $\left\{ D[u]f(x) = \nabla_f^T u \right.$

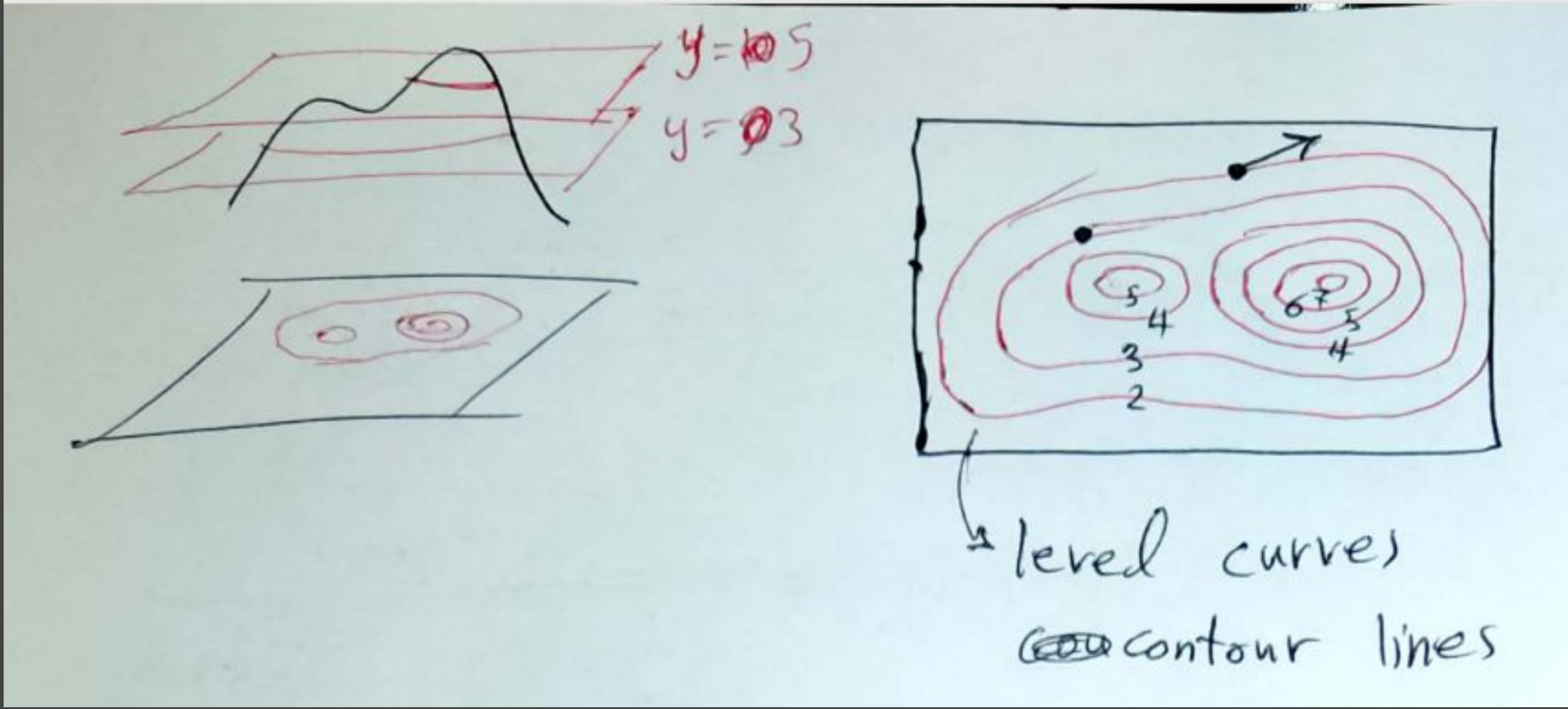
$$u = \frac{\nabla_f}{\|\nabla_f\|} = \operatorname{argmax} \nabla_f^T u \quad \text{s.t.} \quad \|u\|=1$$

steepest ascen~~d~~t

$$u = -\nabla_f / \|\nabla_f\| = \operatorname*{argmin}_{\|u\|=1} \nabla_f^T u$$

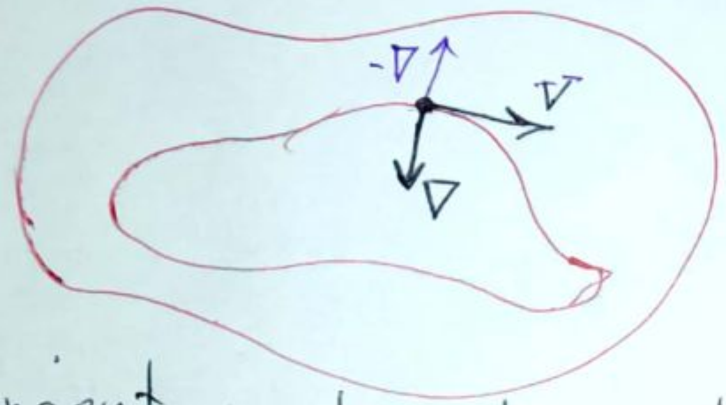
steepest descen~~d~~t

Level Curves



Level Curves and the Gradient Vector

K. N. Toosi
University of Technology



$v \perp \nabla$

$D[v]f = 0$

$\Rightarrow v^T \nabla = 0$

any tangent vector to contour curves is perpendicular to the gradient of f at that point.