EE769 Introduction to ML
Basic Mathematics for ML

Amit Sethi, EE, IITB

asethi, 7483

Objective

- Revise and gain comfort with the following:
  - Vectors and matrices
  - Calculus and convex optimization
  - Probability and statistics

Scalar-vector operations

$a \in \mathbb{R}$

$\bar{x} \in \mathbb{R}^{N \times 1}$

$$\bar{x} = \begin{bmatrix} 100 \\ -100 \\ 22 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \in \mathbb{R}^{3 \times 1}$$

$\mathbb{R}^3$

$$a\bar{x} = \begin{bmatrix} a x_1 \\ a x_2 \\ a x_3 \end{bmatrix}$$

$$\bar{x} + a = \begin{bmatrix} x_1 + a \\ x_2 + a \\ x_3 + a \end{bmatrix}$$

$$\bar{x} + a\mathbb{1} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + a \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

$N \times 1$ $N \times 1$

Scalar-vector operations

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \qquad x_1 \text{ scalar}$$

$$\mathbf{x} \in \mathbb{R}^3$$

$$a\,\mathbf{x} = \begin{bmatrix} a x_1 \\ a x_2 \\ a x_3 \end{bmatrix}$$

$$a \in \mathbb{R}$$

$\mathbf{x} + a$ — Not a math function

$$\equiv \begin{bmatrix} x_1 + a \\ x_2 + a \\ x_3 + a \end{bmatrix} = \mathbf{x} + a \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

$$= \mathbf{x} + a\,\mathbb{1}$$

(vector) (scalar) (vector)

Vector-vector operations

$$\underset{\in \mathbb{R}^N}{\bar{x}} + \underset{\in \mathbb{R}^N}{\bar{y}} = \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \end{bmatrix}$$

$$\bar{x} \cdot \bar{y} = \bar{x}^T \bar{y} = \langle x, y \rangle = \underbrace{x_1 y_1 + x_2 y_2}_{\text{Scalar}}$$

$$\underset{\in \mathbb{R}^{1\times 2}}{\begin{bmatrix} x_1 & x_2 \end{bmatrix}} \quad \underset{\in \mathbb{R}^{2\times 1}}{\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}}$$

$$\underset{N\times 1}{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}} \odot \underset{N\times 1}{\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}} = \underset{N\times 1}{\begin{bmatrix} x_1 y_1 \\ x_2 y_2 \end{bmatrix}}$$

Vector-vector operations

$$x + y = \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \\ x_3 + y_3 \end{bmatrix}$$

x & y have the same dimension

$$x \cdot y = x_1 y_1 + x_2 y_2 + x_3 y_3 = x^T y$$

$$= \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \langle x, y \rangle$$

$$x \odot y = \begin{bmatrix} x_1 y_1 \\ x_2 y_2 \\ x_3 y_3 \end{bmatrix} \in \mathbb{R}^3$$

Sub-spaces spanned by vectors

$\forall \bar{v}_1, \bar{v}_2 \in V$

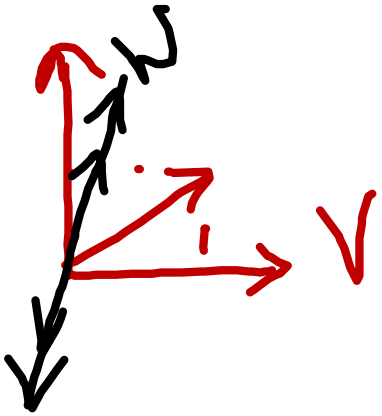
① $\bar{v}_1 + \bar{v}_2 \in V$

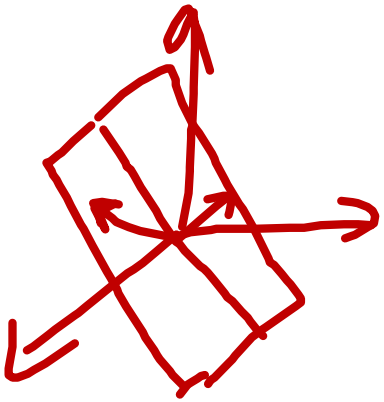
② $c\bar{v}_1 \in V$

$\forall \bar{w}_1, \bar{w}_2 \in W \subset V$

① $\bar{w}_1 + \bar{w}_2 \in W$

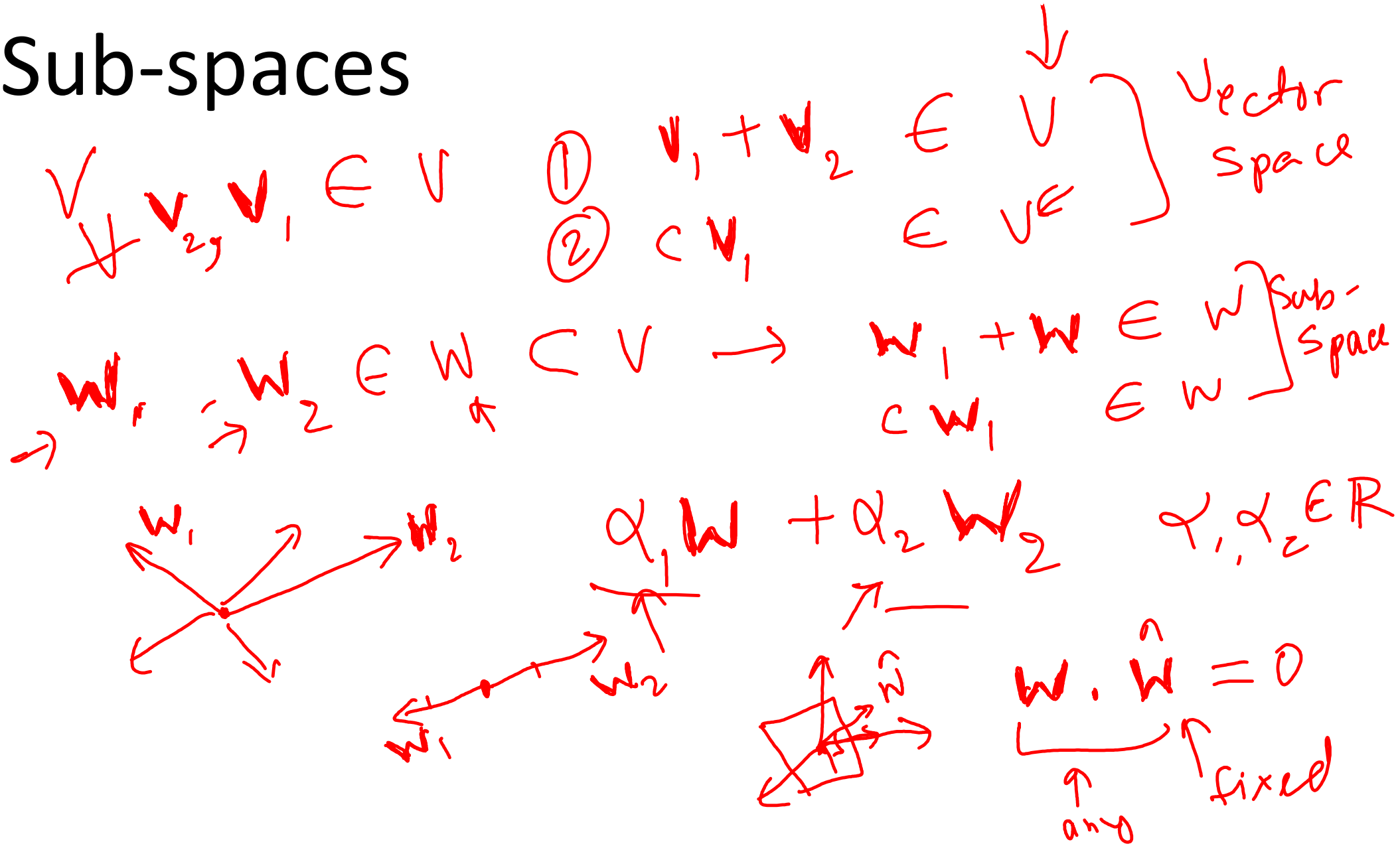
② $c\bar{w}_1 \in W$





$\alpha_1 \bar{w}_1 + \alpha_2 \bar{w}_2$

Sub-spaces


V
$\vec{V}_2, \vec{V}_1 \in V$
① $\vec{V}_1 + \vec{V}_2 \in V$
② $c\vec{V}_1 \in V$
Vector space
$\vec{W}_1, \vec{W}_2 \in W \subset V$
$\vec{W}_1 + \vec{W} \in W$
$c\vec{W}_1 \in W$
sub-space
$\vec{W}_1$
$\vec{W}_2$
$\alpha_1 \vec{W} + \alpha_2 \vec{W}_2$
$\alpha_1, \alpha_2 \in R$
$\vec{W}_1$
$\vec{W}_2$
$\hat{N}$
$\vec{W} \cdot \hat{N} = 0$
any
fixed

Matrix-matrix operations

$X \quad Y \in \mathbb{R}^{M \times N}$ ← Rows Cols

$X \in \mathbb{R}^{M \times N}$

$x \in \mathbb{R}^{N}$ $\quad x \in \mathbb{R}^{N \times 1}$

$$X + Y = \begin{bmatrix} (x_{11}+y_{11}) & \cdots & (x_{1N}+y_{1N}) \\ \vdots & & \\ (x_{M1}+y_{M1}) & \cdots & (x_{MN}+y_{MN}) \end{bmatrix}$$

$X \in \mathbb{R}^{M \times N}$ $\quad Z \in \mathbb{R}^{N \times P}$

$$XZ = \begin{bmatrix} (x_{11}z_{11} + \cdots x_{1N}z_{N1}) & \cdots & \\ \cdots & & (x_{M1}z_{1P} + \cdots x_{MN}z_{NP}) \end{bmatrix}$$

$$X \odot Y \quad \begin{bmatrix} (x_{11}y_{11}) & \cdots (x_{MN}y_{MN}) \end{bmatrix}$$

Transpose, determinant, and inverse of a matrix

$$X^T = \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{M1} \\ \vdots & \ddots & x_{ji} & \cdots \\ x_{1N} & \cdots & \cdots & x_{MN} \end{bmatrix} \in \mathbb{R}^{N \times M}$$

$$X \in \mathbb{R}^{2\times 2} \quad X = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \qquad \det(X) = ad - bc$$

$$X X^{-1} = I = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & 0 & \ddots & 1 \end{bmatrix}$$

$$X^{-1} = \frac{1}{|X|} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

$$3 \times 3^{-1} = 1$$

$$\begin{bmatrix} 1 & -2 \\ 3 & 4 \end{bmatrix} \times \begin{bmatrix} 0.4 & 0.2 \\ -0.3 & 0.1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Rank of a matrix

$$\bar{X} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \in \mathbb{R}^{2\times 3}$$

$\text{Rank}(\bar{X}) \leq \min(M, N)$

$$\bar{Y} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 8 & 12 \end{bmatrix}$$

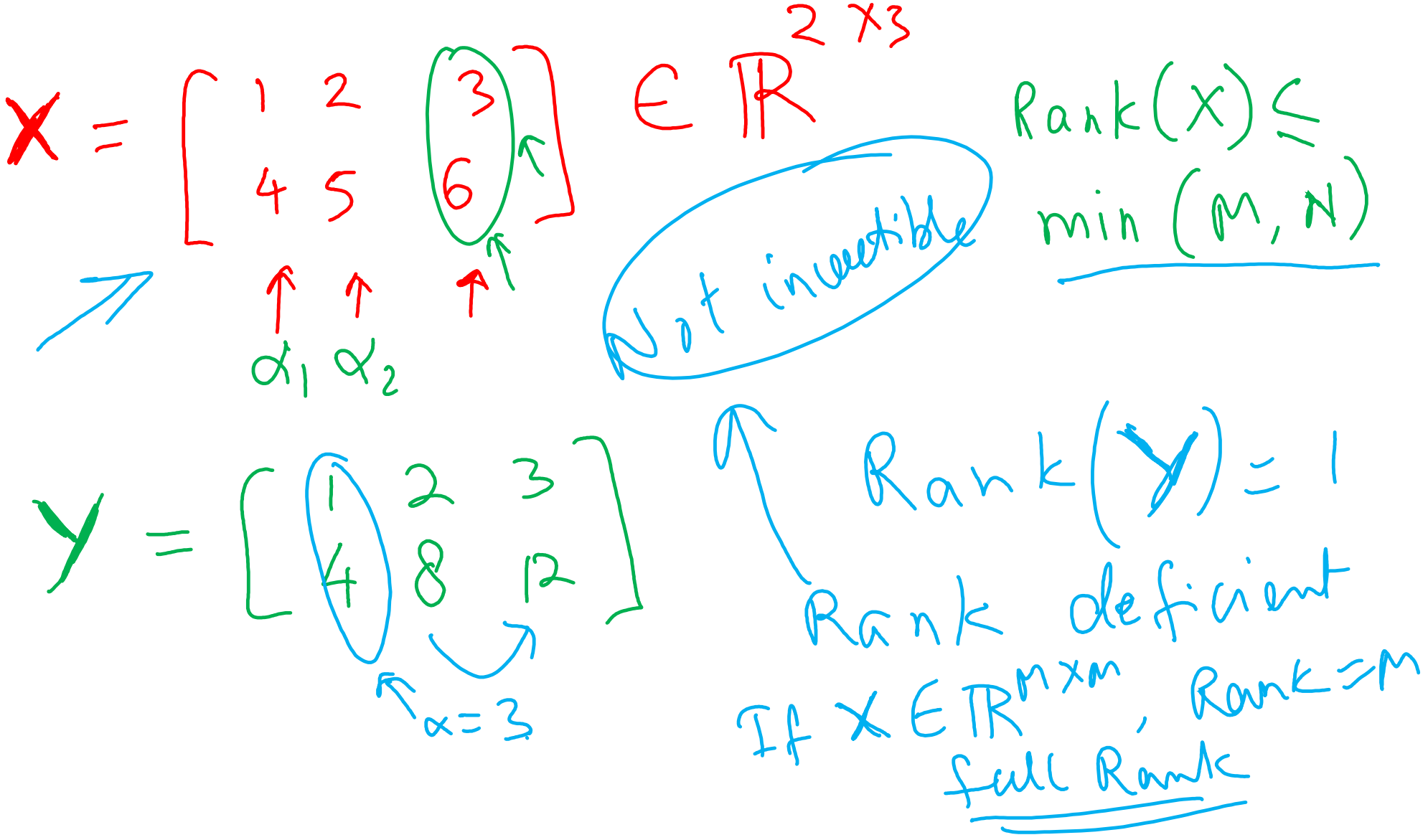
$\text{Rank}(\bar{Y}) = 1$

Rank deficient matrix

$$\bar{X} = \begin{bmatrix} 1 & 2 \\ 4 & 5 \end{bmatrix}$$ full Rank

$$\bar{Y} = \begin{bmatrix} 1 & 2 \\ 4 & 8 \end{bmatrix}$$ Rank deficient

Rank of a matrix



Pseudo-inverse of a matrix

$$\underset{M\times N}{\bar{\bar{X}}} \qquad \underset{N\times M}{\bar{\bar{X}}^{+}} = (\bar{\bar{X}}^{H}\bar{\bar{X}})^{-1}\bar{\bar{X}}^{H}$$

Conjugate transpose

$$\bar{\bar{X}}^{+}\bar{\bar{X}} = \underbrace{(\bar{\bar{X}}^{H}\bar{\bar{X}})^{-1}}\underbrace{\bar{\bar{X}}^{H}}\bar{\bar{X}} = I_{N\times N}$$

Pseudo-inverse of a matrix

X $\quad X^{+}$ $\quad (XX^{+})X = X$

$X^{+} = (X^{H}X)^{-1} X^{H}$

$M \times N$

$N \times M$

$N \times N$

Conjugate transpose

Eigen Decomposition

$\bar{A}\ \bar{v}_i = \lambda_i \bar{v}_i$ $\quad \bar{v}_i \rightarrow$ eigenvector of $\bar{A}$

$N \times N \quad N \times 1$ $\quad i \in \{1, \ldots, N\}$ $\quad \lambda_i \rightarrow$ eigenvalue of $\bar{A}$

$$\bar{A} = \bar{Q}\ \bar{\Lambda}\ Q^{-1}$$

$\bar{Q} = [\bar{v}_1 \cdots \bar{v}_N]$

$\bar{\Lambda} = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_N \end{bmatrix}$

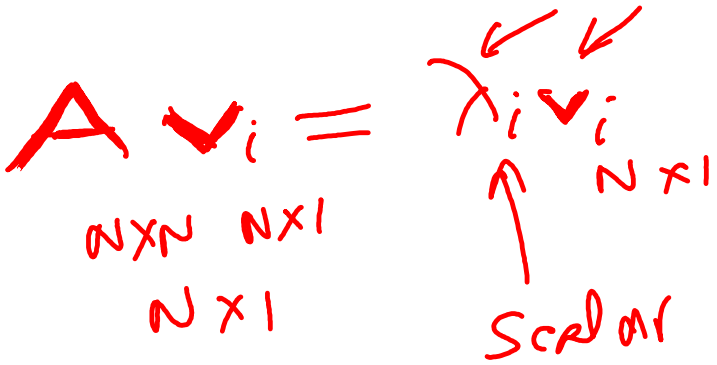
$\|v_i\|_2^2 = 1$

$v_i^T v_j = \delta_{ij}$

$\delta_{ij} = 0, \ i \neq j$

$= 1, \ i = j$

Eigen Decomposition



v is an eigen vector of A

λ " " " value of A

v_i is a unit vector

$$A = Q \Lambda Q^{-1}$$

$Q = [v_1 \; v_2 \; v_3 \cdots v_N]$

$\Lambda \rightarrow$ Diagonal matrix

$$\begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ 0 & 0 & \cdots & \lambda_N \end{bmatrix}$$

Tensors

$\underline{\alpha} \in \mathbb{R}$; $\underline{\underline{x}} \in \mathbb{R}^{N}$; $\underline{\underline{\underline{Y}}} \in \mathbb{R}^{M \times N}$

$\underline{\underline{\underline{\underline{T}}}} \in \mathbb{R}^{M \times N \times P \times \cdots Q}$

transpose(T, [3, 1, 2, 0])

0, 1, 2, 3

0, 1

1, 0

Tensors

$\alpha \in \mathbb{R}$ $\quad$ $\mathcal{X} \in \mathbb{R}^{N}$ $\quad$ $A \in \mathbb{R}^{M \times N}$

$\mathcal{X} \in \mathbb{R}^{\underset{0}{M} \times \underset{1}{N} \times \underset{2}{P} \times \underset{3}{Q}}$

$\mathcal{T} \in \mathbb{R}$

$R = \text{transpose}(T, [3,1,2,0])$

order of permutation of dimensions

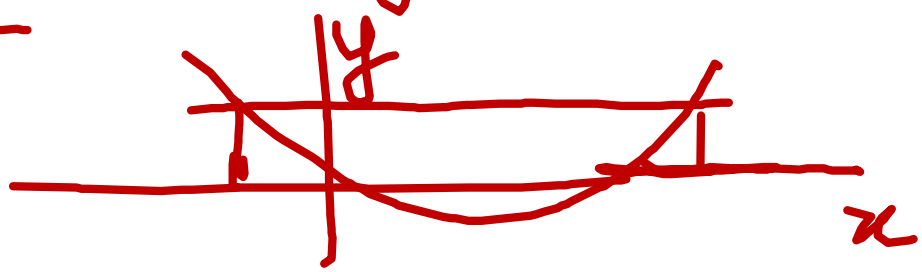
Functions

$f : X \to Y$ $\quad x \in X$ $\quad f(x) \in Y$

$y = f(x) \in Y$

many - to - one

$y = ax^2 + bx + c$



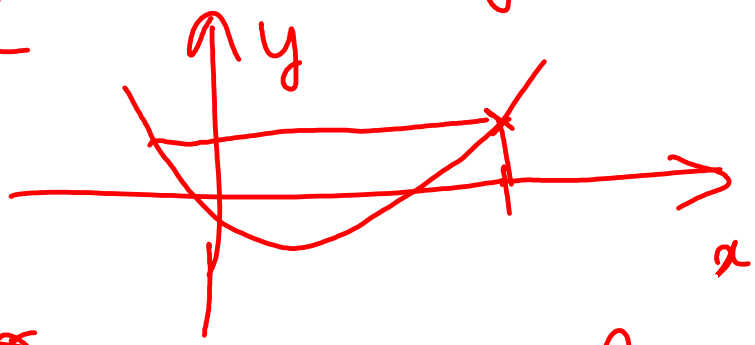
$x^2 + y^2 = 1 \Rightarrow y = \pm\sqrt{1-x^2}$ ✗ Not a function

Functions

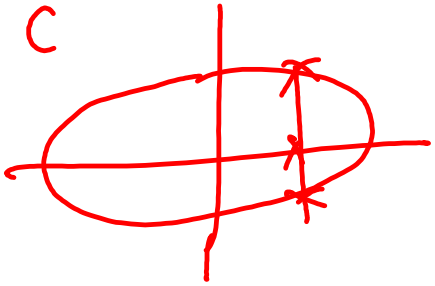
$f : X \to Y$ $\quad x \in X \quad f(x) \to y$

many to one

$f(x) = ax^2 + bx + c$



$ax^2 + by^2 = c^2$



Not a function

Continuity

$y = \sqrt{x}$ $y = \sin(x)$ $y = x^2$

x is continuous

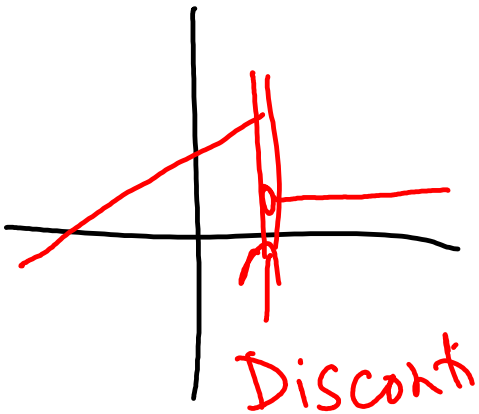
Lipschitz continuity

If $f(x + \Delta x) = f(x)$

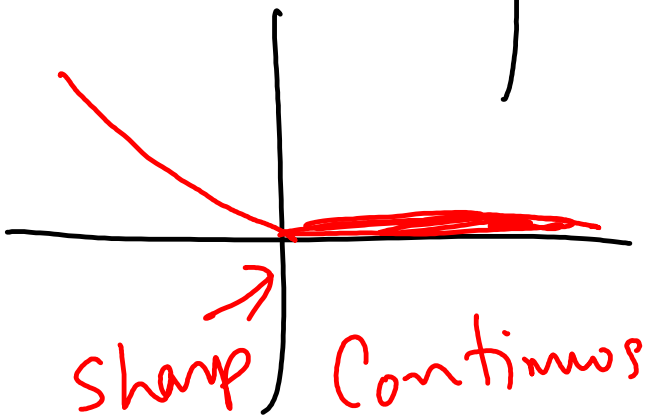
$\Delta x \to 0$

$|f(x_1) - f(x_2)| \leq K |x_1 - x_2|$

$+K$ $-K$

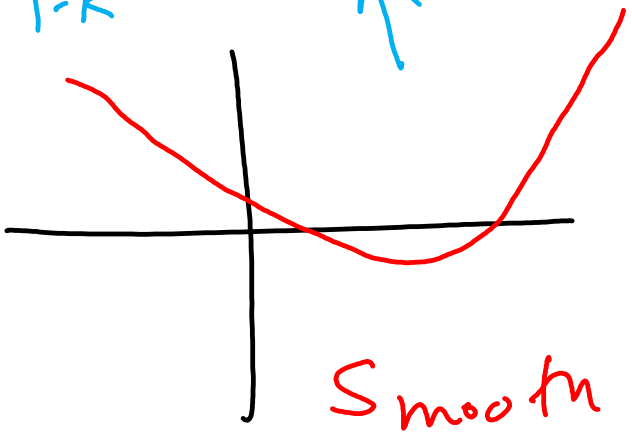


Discontinuous



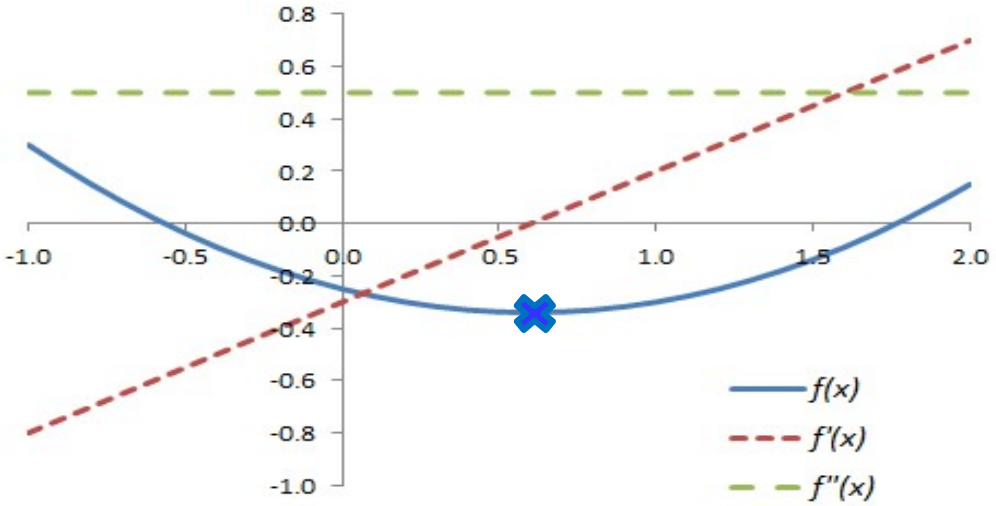
Sharp corner

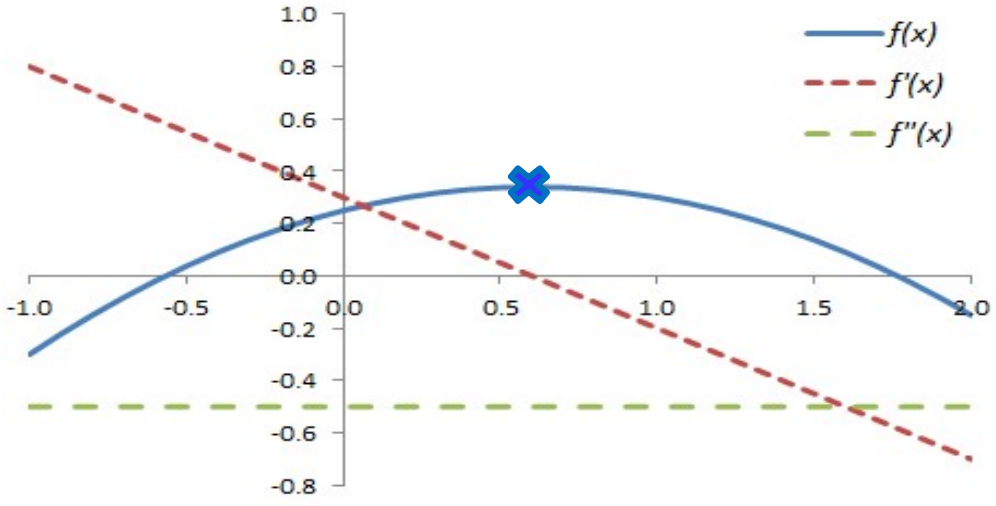
Continuous not smooth



Smooth & continuous

Derivative of a function of a scalar





E.g. $f(x) = ax^2 + bx + c, \quad f'(x) = 2ax + b, \quad f''(x) = 2a$

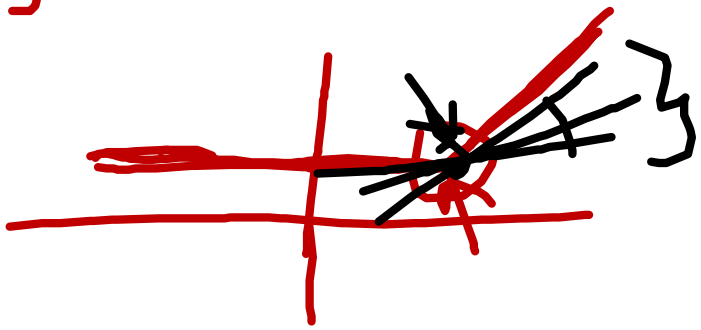
- Derivative $f'(x) = \frac{d\,f(x)}{d\,x}$ is the rate of change of $f(x)$ with x
- It is zero when then function is flat (horizontal), such as at the minimum or maximum of $f(x)$
- It is positive when $f(x)$ is sloping up, and negative when $f(x)$ is sloping down
- To move towards the maxima, taking a small step in a direction of the derivative

Derivative of a function

$$\lim_{\Delta x \to 0} f(x+\Delta x) = f(x) \quad \forall x$$

Continuous $f(x)$

$$f'(x) = \lim_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x}$$



Sub-tangent

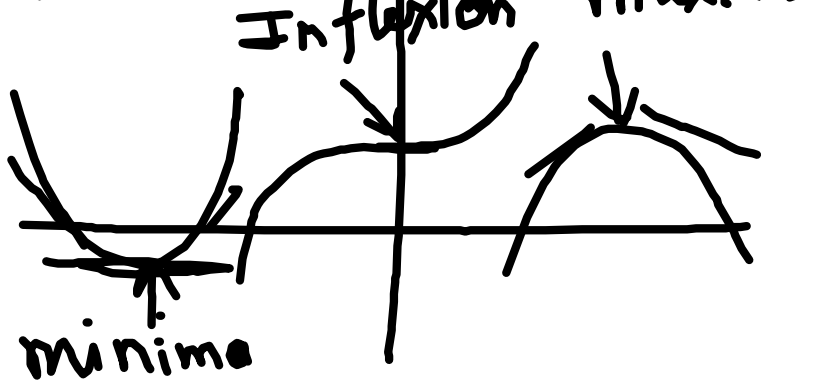
Sub-gradient

$f'(x) = 0 \rightarrow$ critical point

$f''(x) > 0 \rightarrow$ min

$< 0 \rightarrow$ max

$= 0 \rightarrow$ inflexion

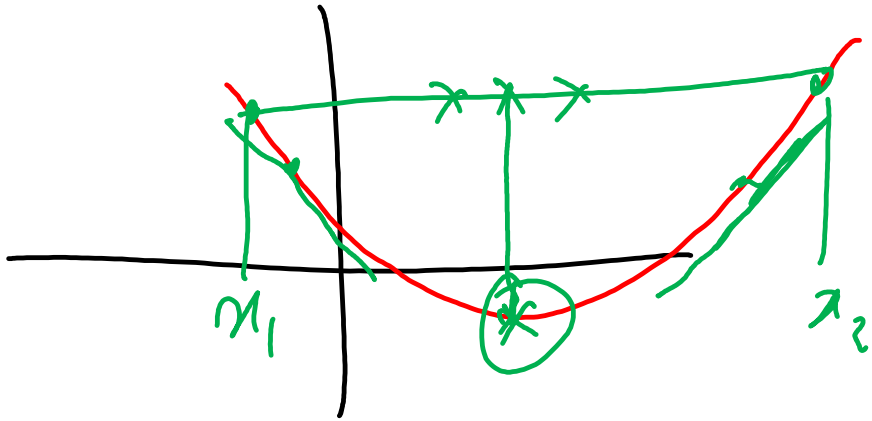


Derivative of a function

Gradient

$f(x)$

$f'(x) = 2ax + b$



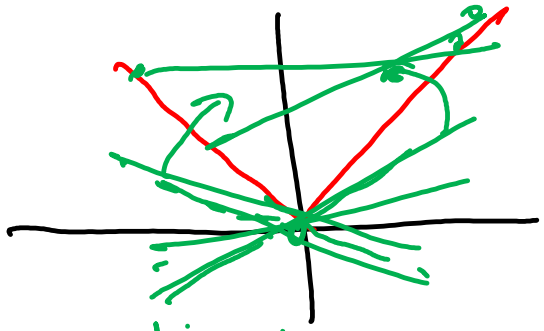
$\lambda f(x_1) + (1-\lambda) f(x_2)$

$\geq f(\lambda x_1 + (1-\lambda) x_2)$

$ax^2 + bx + c$

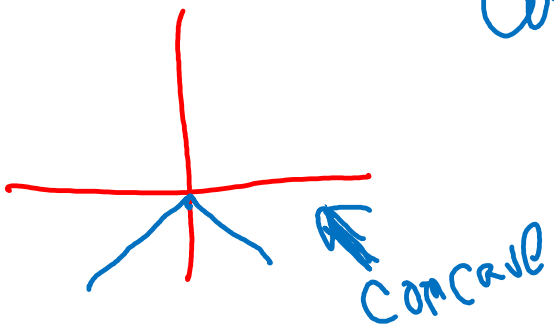
Convex function

Convex →



$f(x) = |x|$

$$f'(x) = \begin{cases} -1, & \text{if } x < 0 \\ +1, & \text{if } x > 0 \end{cases}$$



Subgradient

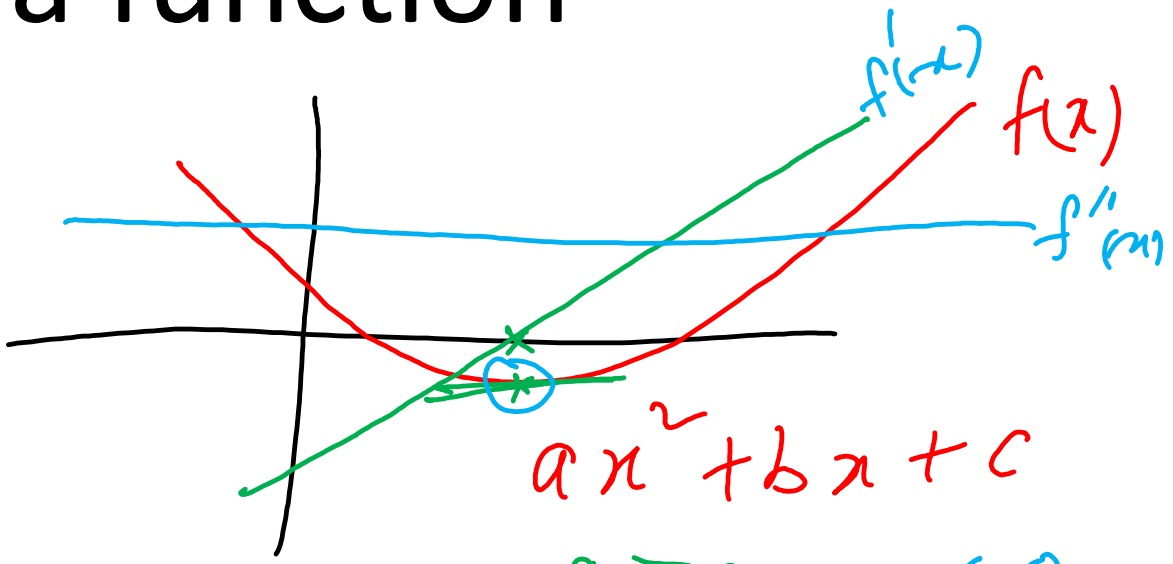
Sub-derivatives are slopes of subtangents

Critical points of a function

$f'(x) = 0$

$f'(x) = 2ax + b$

$x = \frac{-b}{2a}$



Maxima → if $a > 0$, $f''(x) < 0$

Minima → if $f''(x) > 0$

Inflection

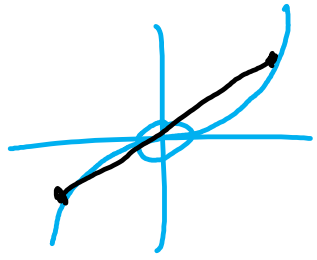
$f''(x) = 2a \geq 0$ ← Convex $\forall x$

Concave fn. $f''(x) \leq 0$

$f(x) = x^3$

$f'(x) = 3x^2$

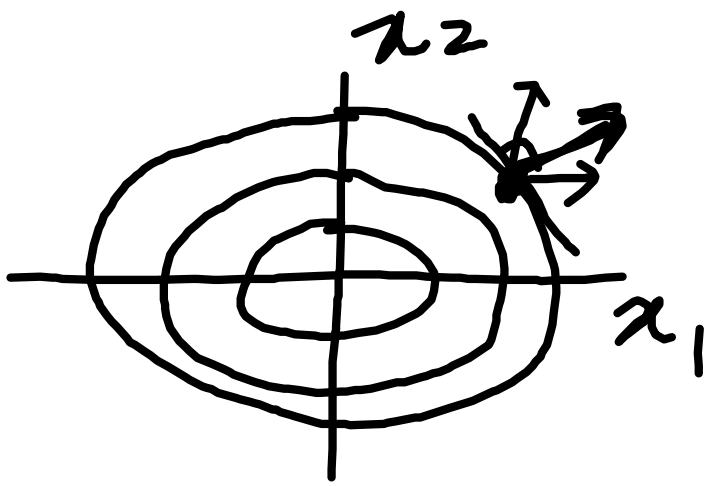
$f''(x) = 6x$



Multivariate functions

$$y = f(x_1, x_2) = a x_1^2 + b x_2^2 \Leftarrow$$

$$\nabla y = \begin{bmatrix} \partial y/\partial x_1 \\ \partial y/\partial x_2 \end{bmatrix} = \begin{bmatrix} 2a x_1 \\ 2b x_2 \end{bmatrix} \Leftarrow$$

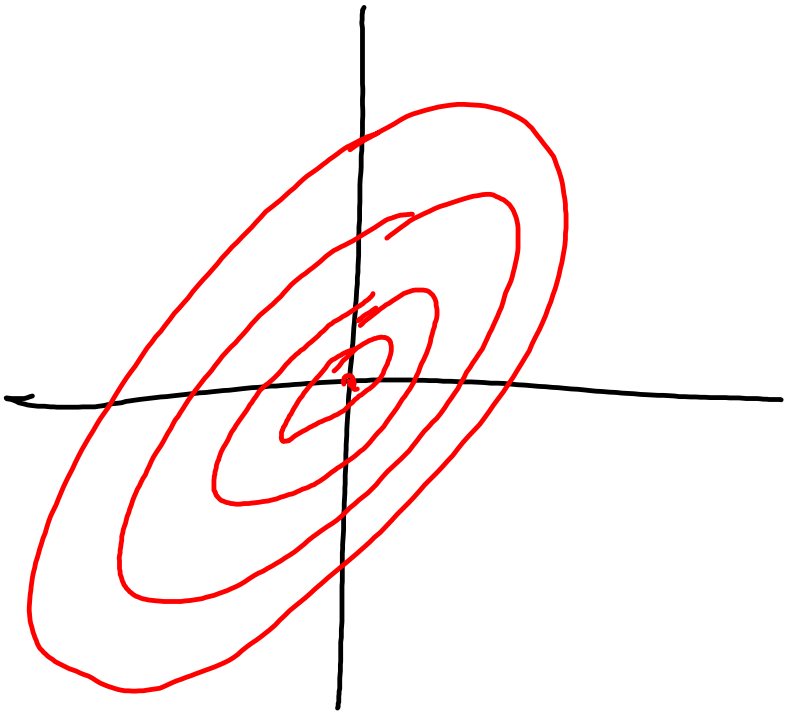


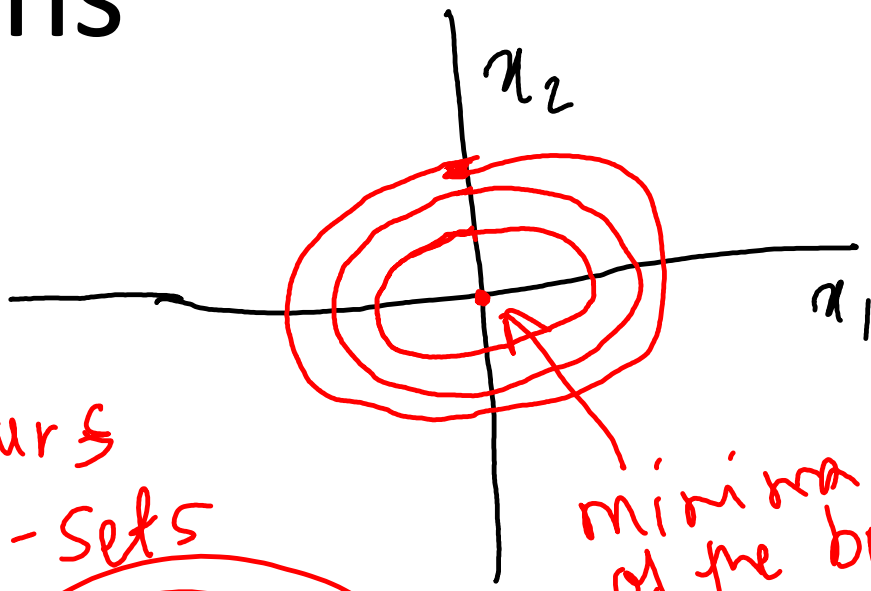
G. D. in 2-d

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \Leftarrow \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \eta \nabla y$$

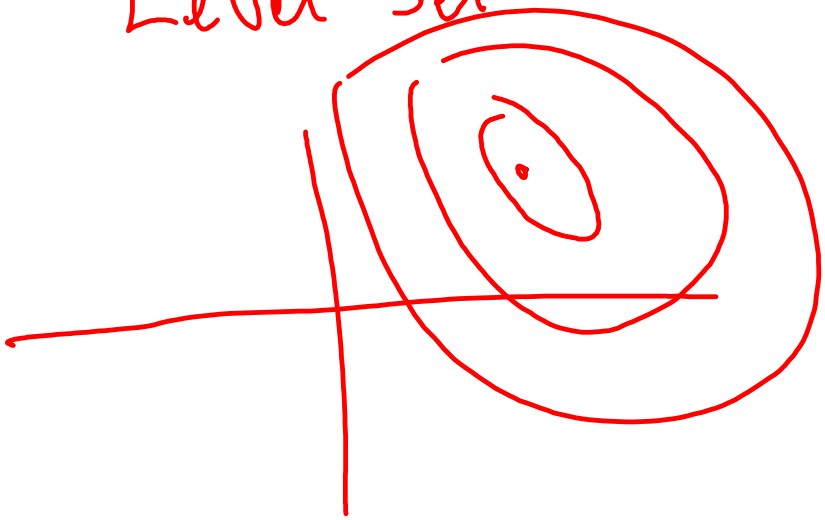
Multivariate functions

$$y = a x_1^2 + b x_2^2$$

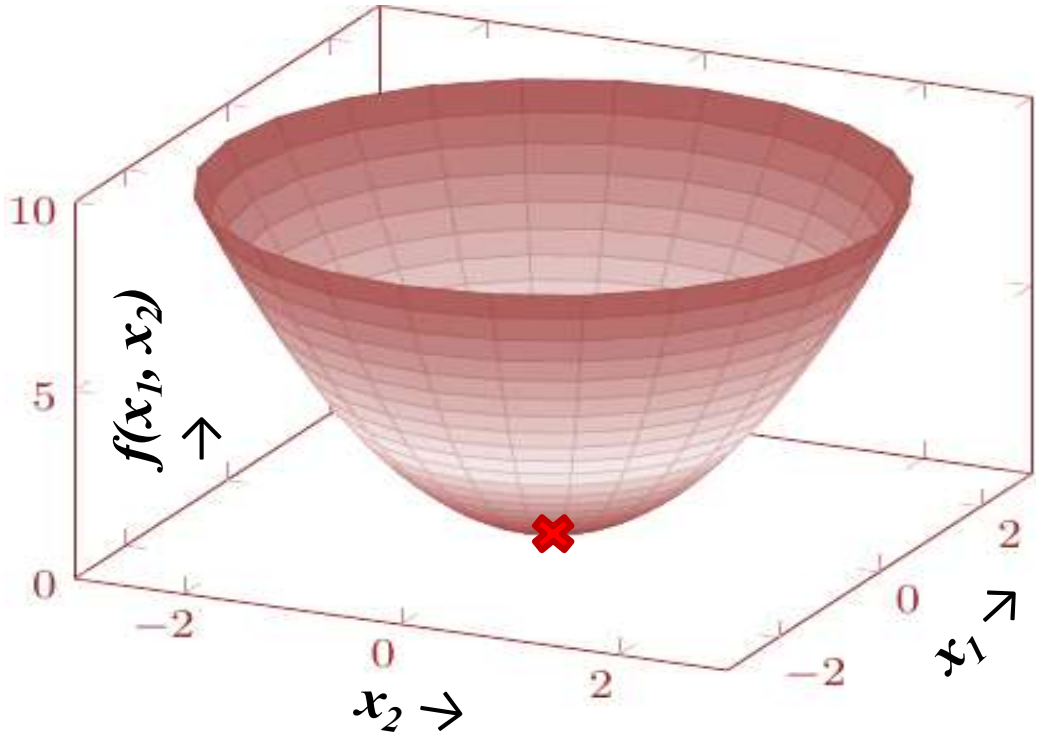




Contours
Level-sets



Gradient of a function of a vector

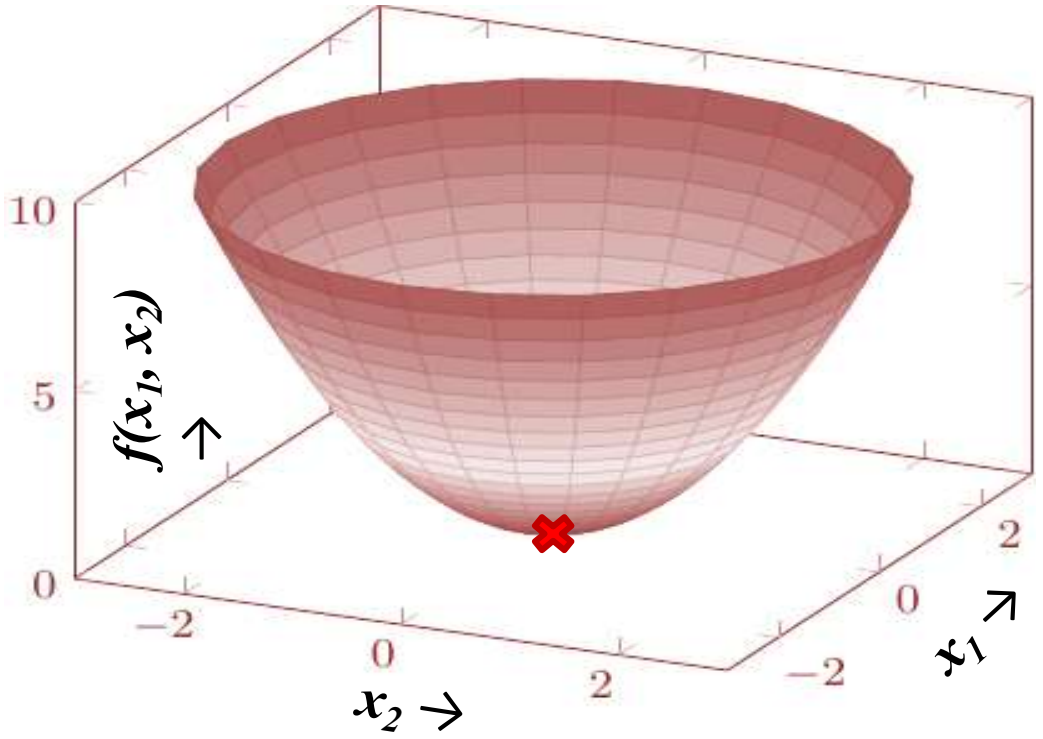


- Derivative with respect to each dimension, holding other dimensions constant
- $\nabla f(\boldsymbol{x}) = \nabla f(x_1, x_2) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{bmatrix}$
- At a minima or a maxima the gradient is a zero vector

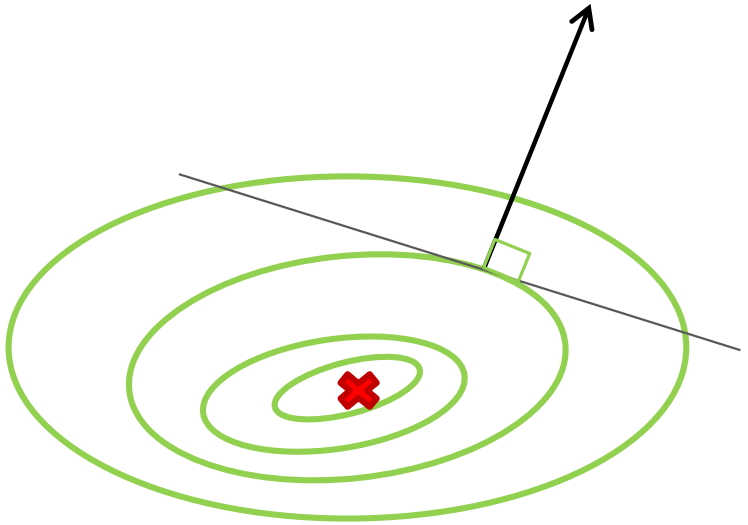
  The function is flat in every direction
- At a minima or a maxima the gradient is a zero vector

Original image source unknown

Gradient of a function of a vector



- Gradient gives a direction for moving towards the minima
- Take a small step towards negative of the gradient



Original image source unknown

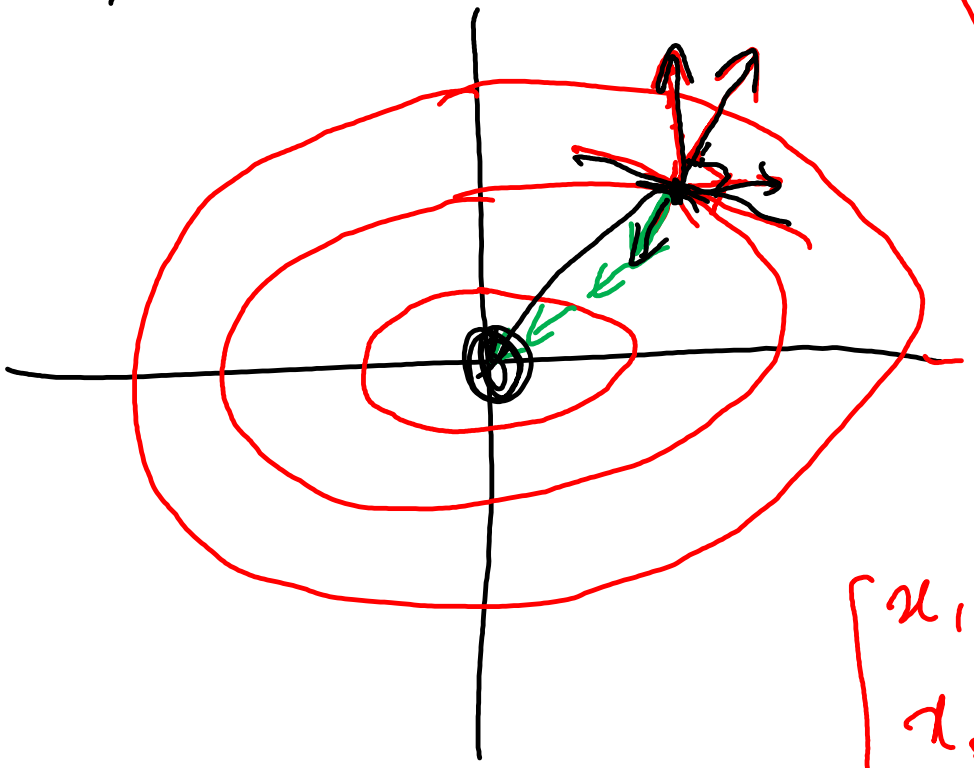
Example of gradient

- Let $f(\boldsymbol{x}) = f(x_1, x_2) = 5{x_1}^2 + 3{x_2}^2$
- Then $\nabla f(\boldsymbol{x}) = \nabla f(x_1, x_2) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 10x_1 \\ 6x_2 \end{bmatrix}$
- At a location $(2,1)$ a step in $\begin{bmatrix} 20 \\ 6 \end{bmatrix}$ or $\begin{bmatrix} 0.958 \\ 0.287 \end{bmatrix}$ direction will lead to maximal increase in the function

Gradients

$f(x_1, x_2) = y = a x_1^2 + b x_2^2$

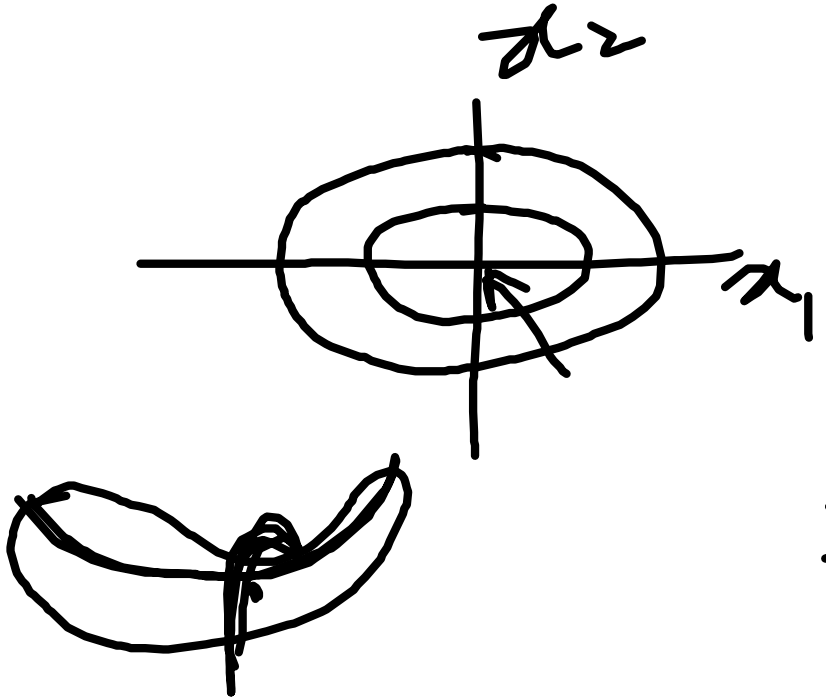
$$\nabla y = \begin{bmatrix} \frac{\partial y}{\partial x_1} \\ \frac{\partial y}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 2 a x_1 \\ 2 b x_2 \end{bmatrix}$$



Direction of steepest ascent

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \leftarrow \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \eta \begin{bmatrix} \partial y / \partial x_1 \\ \partial y / \partial x_2 \end{bmatrix}$$

Maxima and minima for multivariate functions



$$y = a x_1^2 + b x_2^2$$

$$\nabla y = 0 = \begin{bmatrix} 2ax_1 \\ 2bx_2 \end{bmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

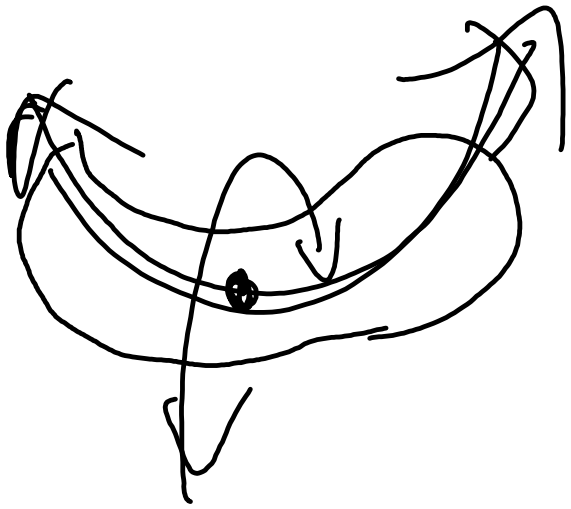
If $a > 0$; $b < 0$

$\Rightarrow (0,0)$ is max for x_2
min for x_1

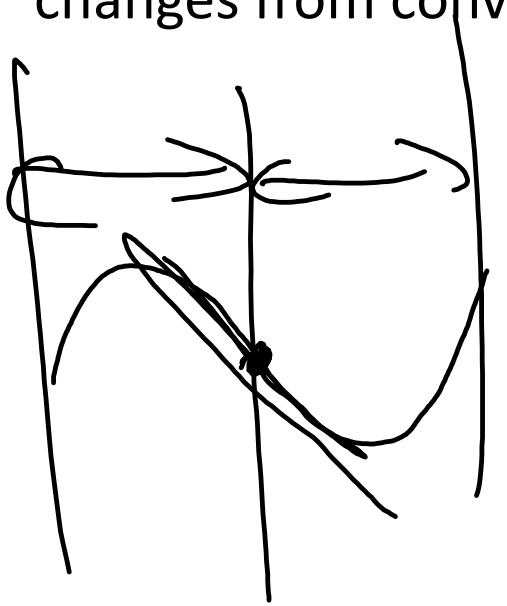
Saddle point

When functions are neither convex nor concave

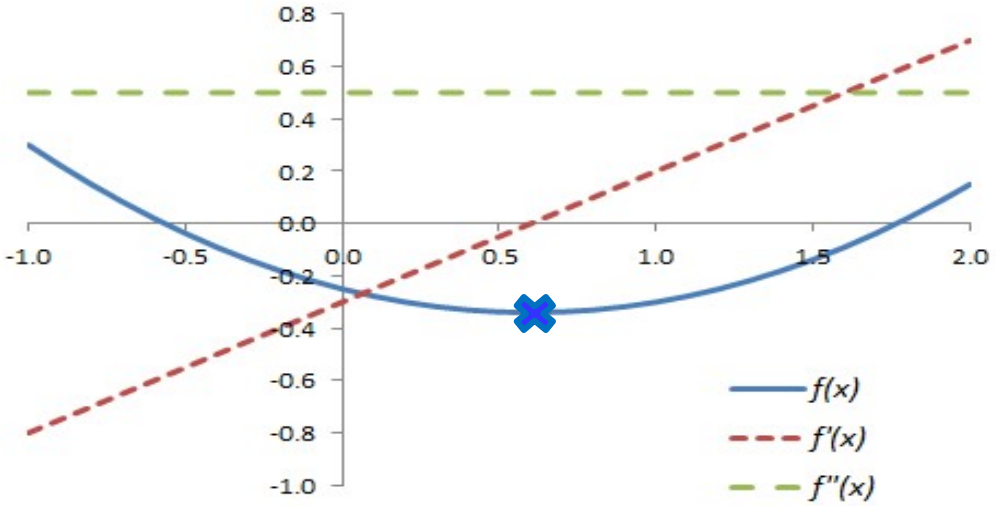
Saddle point
where derivative is zero
and one direction is convex,
and the other in concave

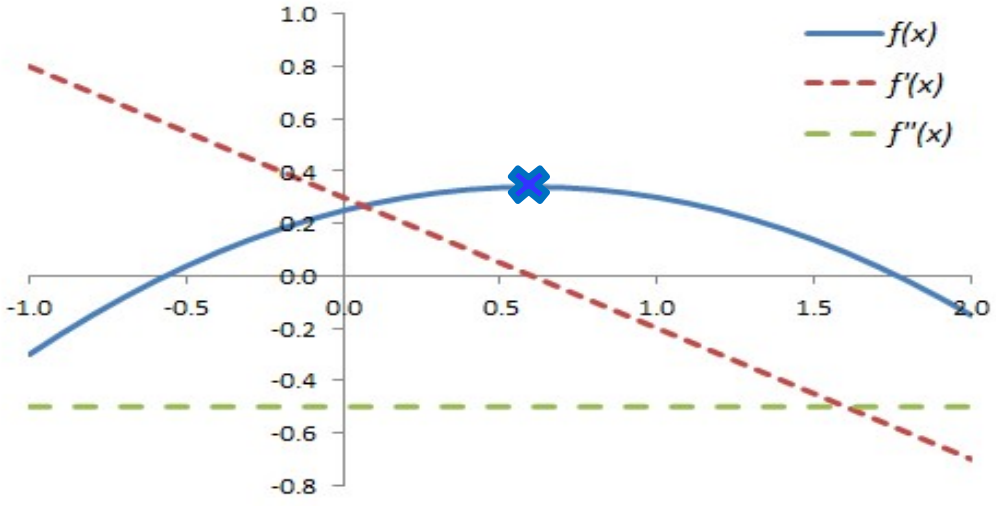


Inflection point
where second derivative
is zero and the function
changes from convex to concave



Double derivative

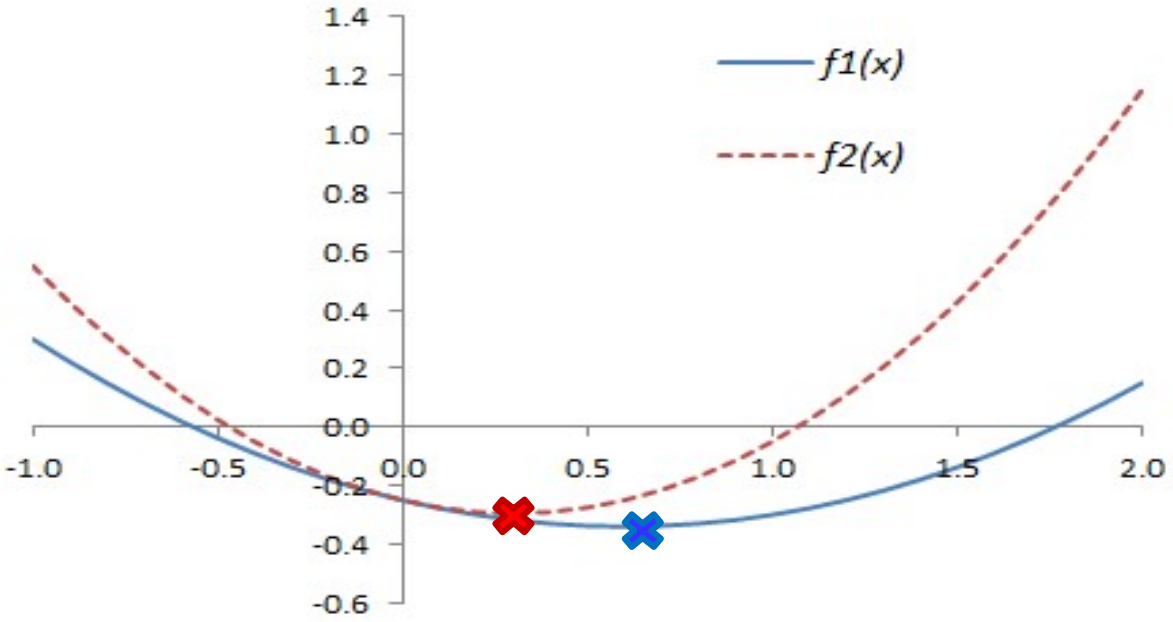




E.g. $f(x) = ax^2 + bx + c, \quad f'(x) = 2ax + b, \quad f''(x) = 2a$

- Double derivative $f''(x) = \frac{d^2 f(x)}{d x^2}$ is the derivative of derivative of $f(x)$
- Double derivative is positive for convex functions (have a single minima), and negative for concave functions (have a single maxima)

Double derivative



$$f(x) = ax^2 + bx + c,$$
$$f'(x) = 2ax + b,$$
$$f''(x) = 2a$$

- Double derivative tells how far the minima might be from a given point.
- From $x = 0$ the minima is closer for the red dashed curve than for the blue solid curve, because the former has a larger second derivative (its slope reverses faster)

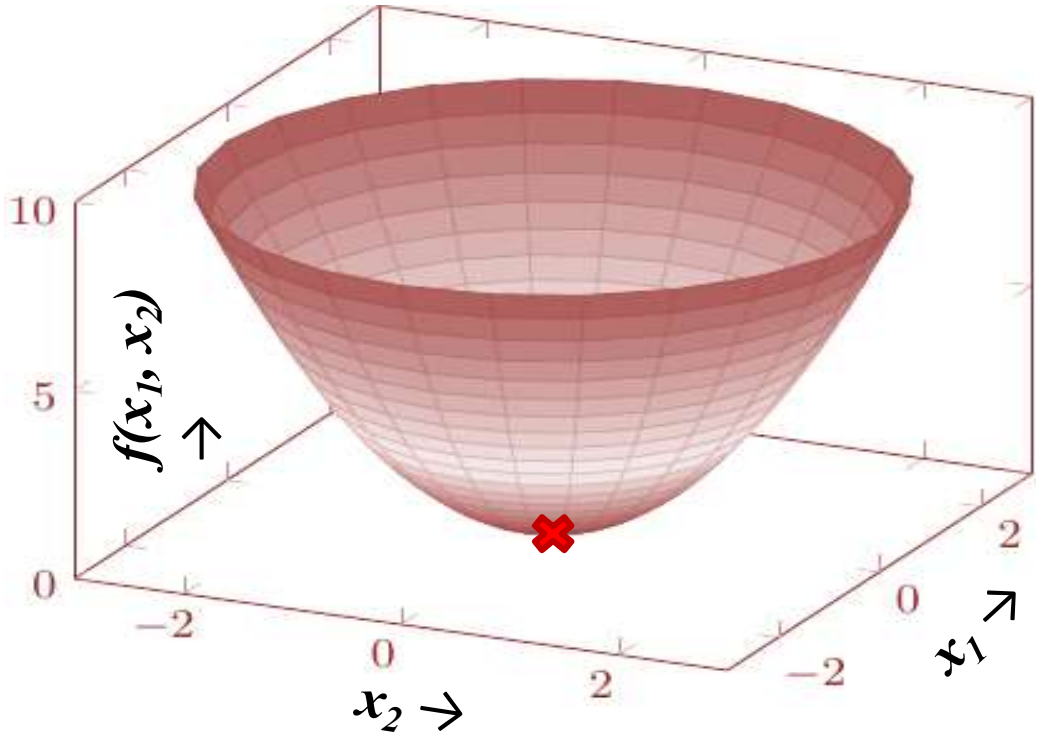
Perfect step size for a paraboloid

- Let $f(x) = ax^2 + bx + c$
- Assuming $a < 0$
- Minima is at: $x^* = -\frac{b}{2a}$
- For any x the perfect step would be:

$$-\frac{b}{2a} - x = -\frac{2ax+b}{2a} = -\frac{f'(x)}{f''(x)}$$

- So, the perfect learning rate is: $\eta^* = \frac{1}{f''(x)}$
- In multiple dimensions, $\boldsymbol{x} \leftarrow \boldsymbol{x} - H\big(f(\boldsymbol{x})\big)^{-1} \nabla(f(\boldsymbol{x}))$
- Practically, we do not want to compute the inverse of a Hessian matrix, so we approximate Hessian inverse

Hessian of a function of a vector



- Double derivative with respect to a pair of dimensions forms the Hessian matrix:
- If all eigenvalues of a Hessian matrix are positive, then the function is convex

Original image source unknown

Example of Hessian

- Let $f(\boldsymbol{x}) = f(x_1, x_2) = 5{x_1}^2 + 3{x_2}^2 + 4x_1x_2$
- Then $\nabla f(\boldsymbol{x}) = \nabla f(x_1, x_2) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 10x_1 + 4x_2 \\ 6x_2 + 4x_1 \end{bmatrix}$
- And, $H(f(\boldsymbol{x})) = \begin{bmatrix} \frac{\partial^2 f}{\partial {x_1}^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial {x_2}^2} \end{bmatrix} = \begin{bmatrix} 10 & 4 \\ 4 & 6 \end{bmatrix}$

Saddle points, Hessian and long local furrows

- Some variables may have reached a local minima while others have not
- Some weights may have almost zero gradient
- At least some eigenvalues may not be negative

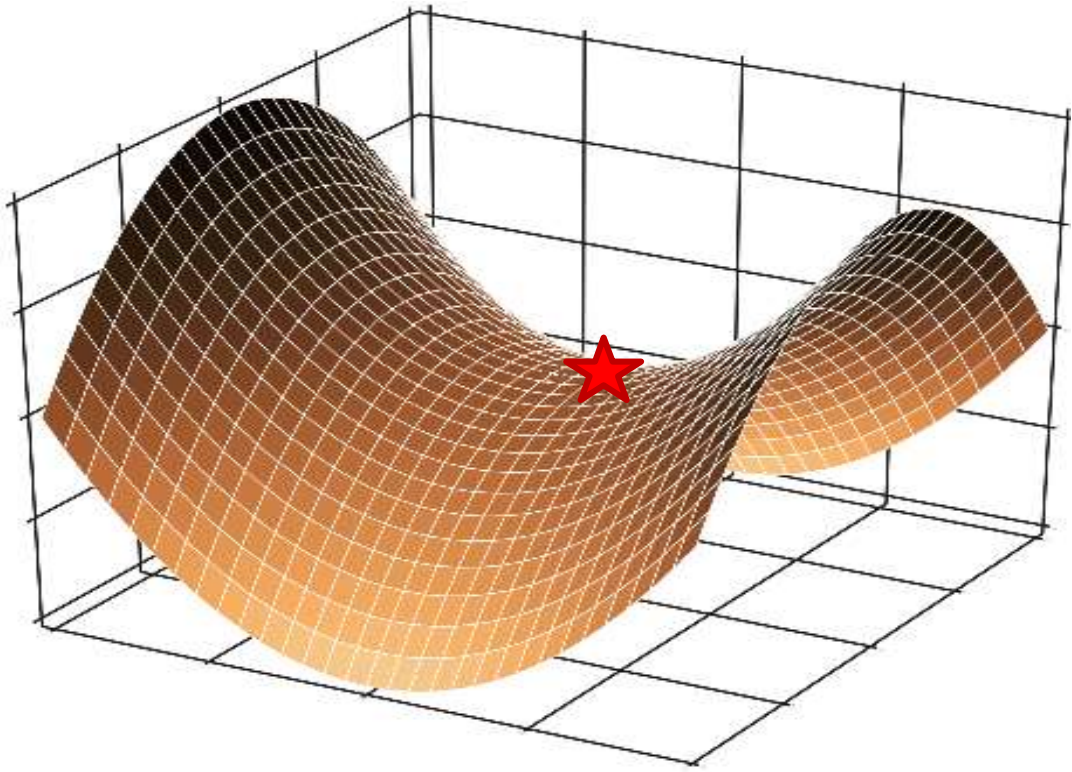


Image source: Wikipedia

A realistic picture

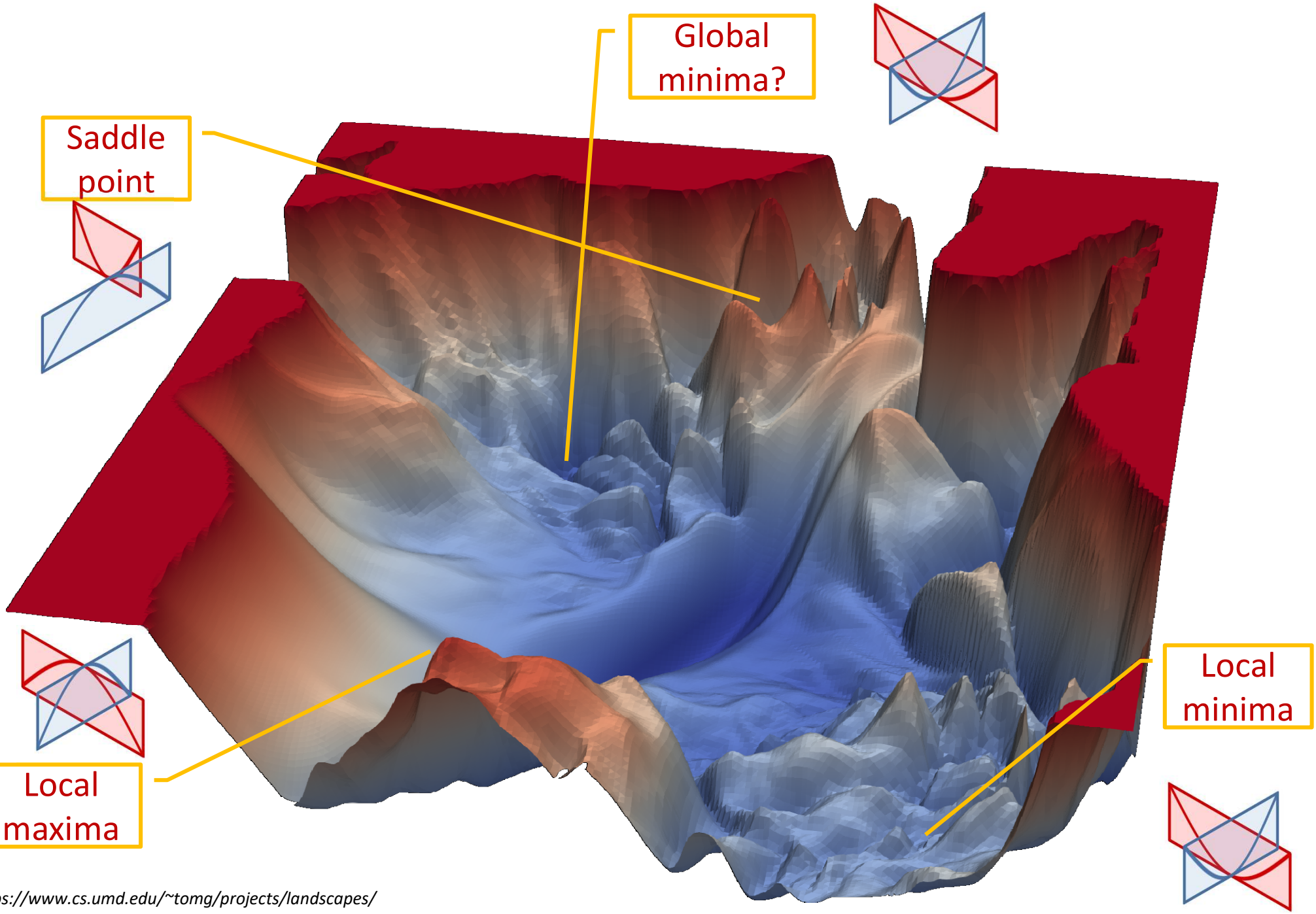
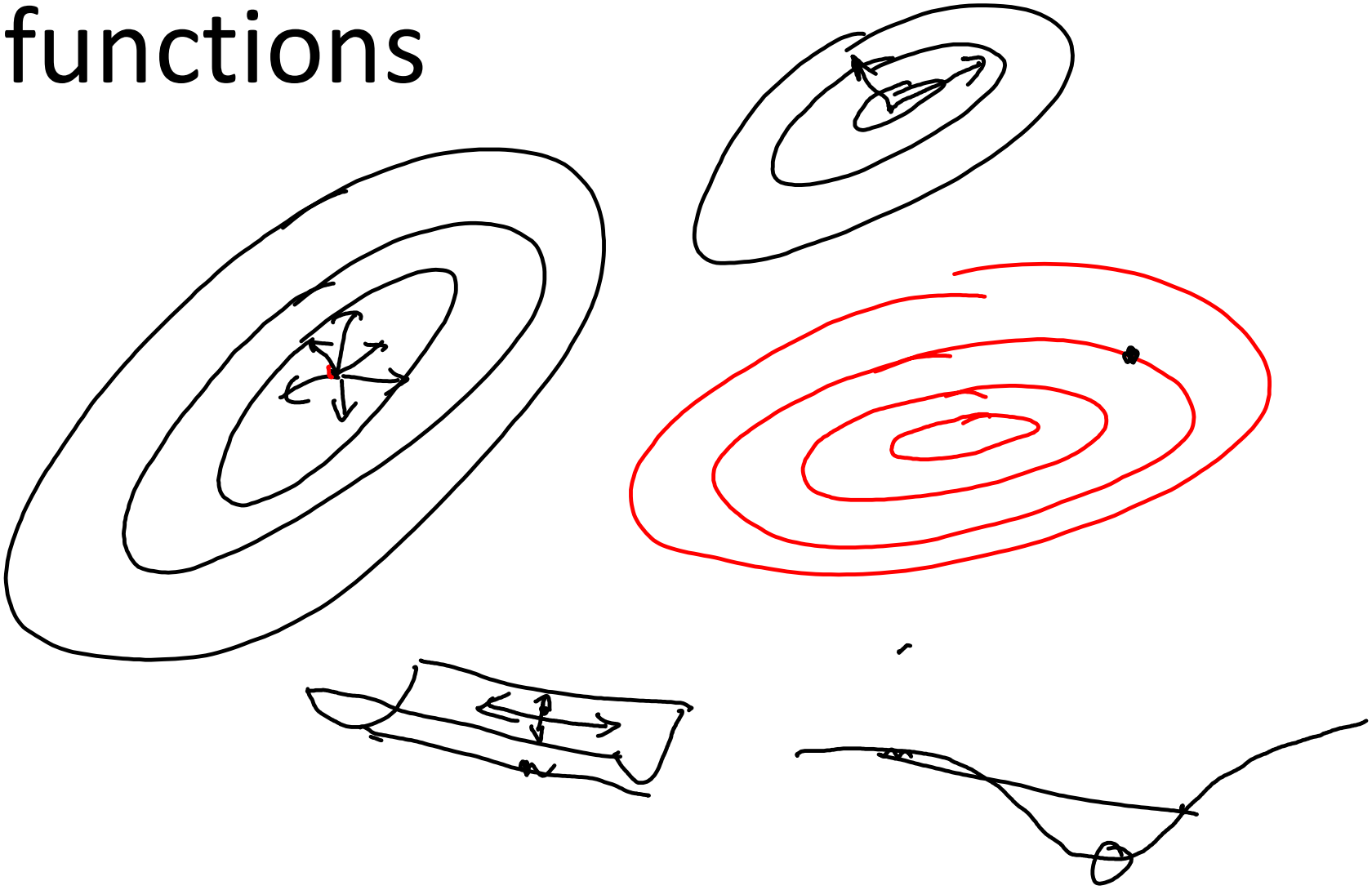
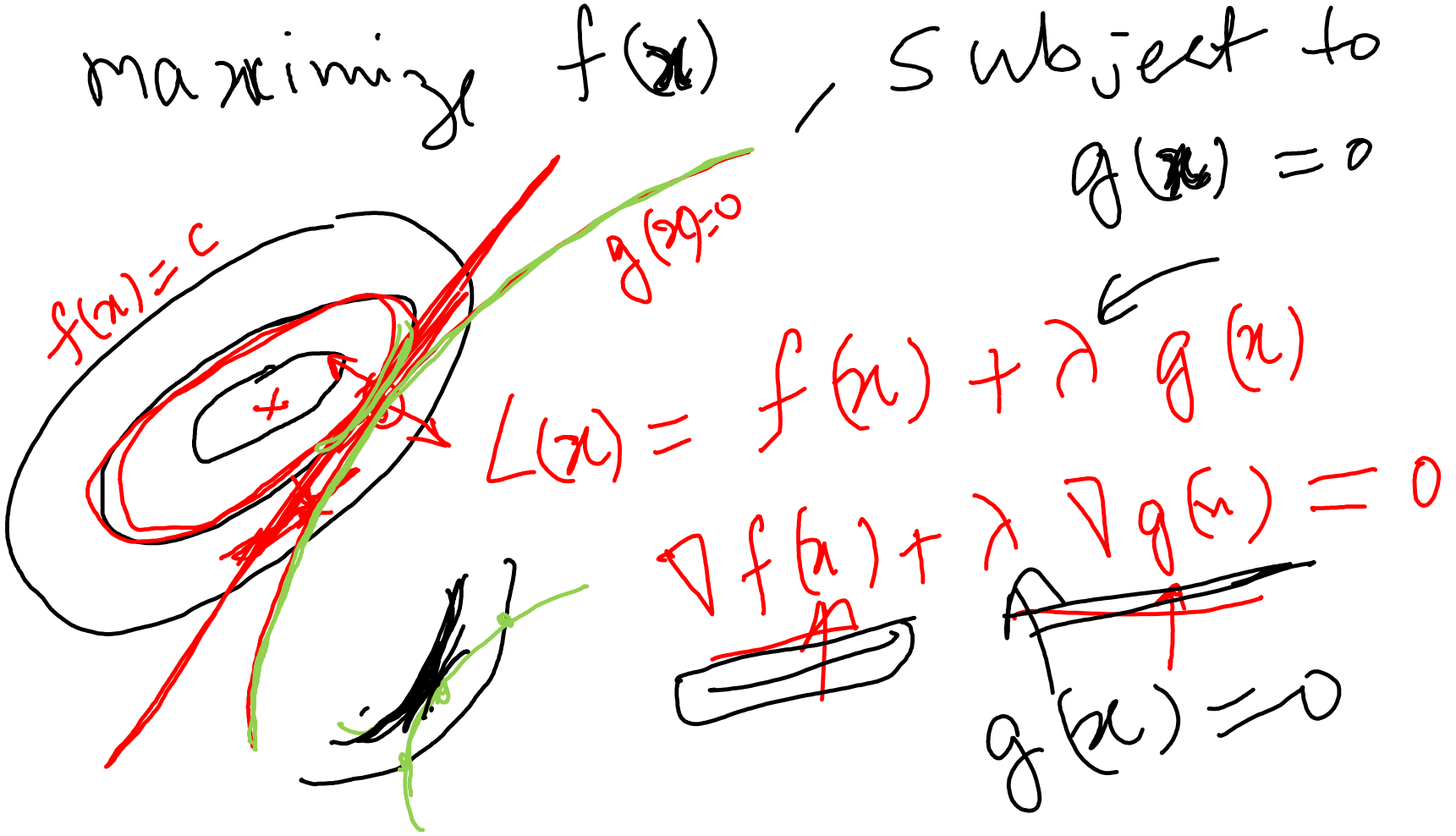


Image source: https://www.cs.umd.edu/~tomg/projects/landscapes/

Some functions



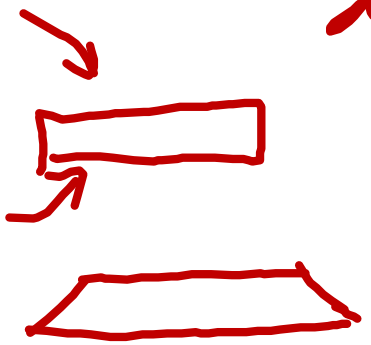
Constrained optimization using Lagrange multiplier



Random variable

$X = x$ $\{H, T\}$ M.E. C.E

$x \in \{0, 1\}$



Fair coin

$P(X=0) = 1 - P(X=1)$

$= 0.5$

$\forall x, P(X=x) \geq 0$

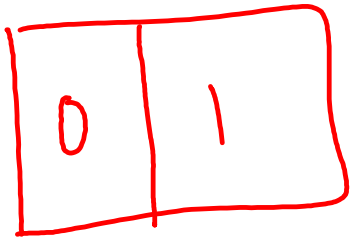
$\sum_x P(X=x) = 1$

Random variable

$X = x$ $\{H, T\}$ Event space

$X \in \{0, 1\}$ M.E.C.E.

Fair coin 



$P(X=0) = 1 - P(X=1) = 0.5$

Biased coin 

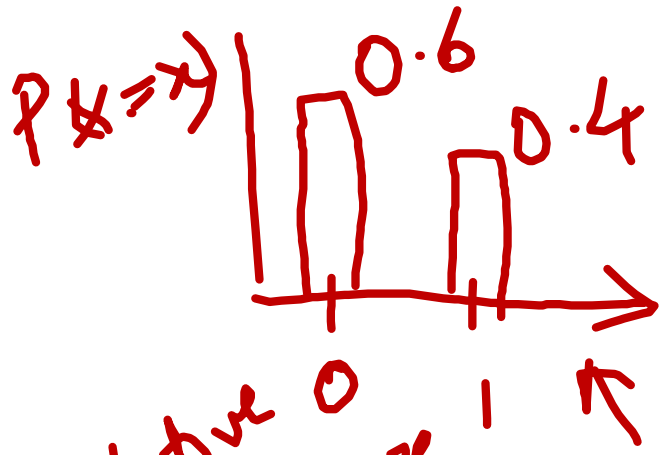
$P(X=0) = 0.6$

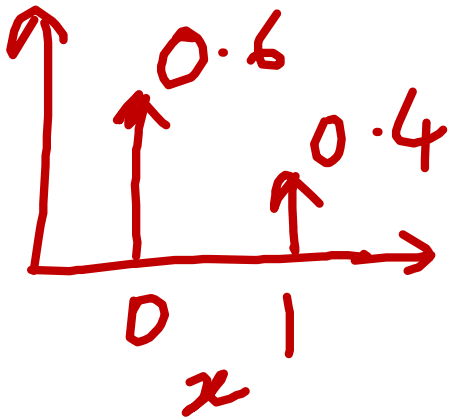
$P(X=1) = 0.4$

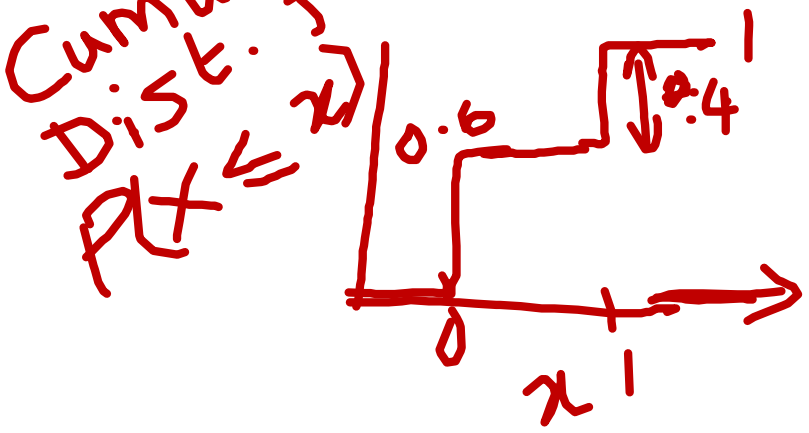
$\forall x, P(X=x) \geq 0$

$\sum_x P(X=x) = 1$

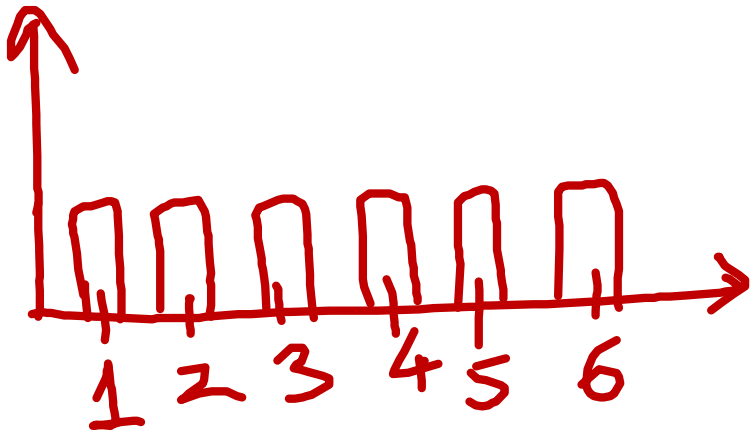
Probability mass function







6-faced die



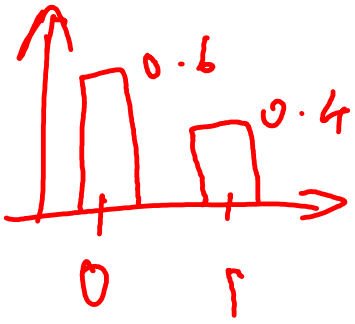
Probability mass function

$P(X = x)$

$P_X(x) : x \longrightarrow [0, 1]$

$\uparrow \quad \sum_x P_X(x) = 1$

Discreet R.V.



Some common PMFs

Bernoulli Dist.

$x \in \{0, 1\}$

$\mu = P(X = 1)$

$\text{Bern}(x \mid \mu) = \mu^x (1-\mu)^{1-x} = \begin{cases} \mu, & \text{if } x = 1 \\ 1-\mu, & \text{if } x = 0 \end{cases}$

$\mathbb{E}(x) = \sum_x x \, P(X = x) = \mu$

$\mathbb{E}[(x - \bar{x})^2] = \mu(1-\mu)$

Binomial Dist.

$\text{Bin}(y \mid \mu, N) = \binom{N}{y} \mu^y (1-\mu)^{N-y}$

$\mathbb{E}[f(x)] = \sum_x f(x) P(X = x)$

Some common PMFs

① Bernoulli $x \in \{0, 1\}$

$\mu = P(x=1)$ $\mu \in [0, 1]$

$\text{Bern}(x|\mu) = \mu^x (1-\mu)^{1-x}$

given $\begin{cases} \mu, \text{ if } x=1 \\ 1-\mu, \text{ if } x=0 \end{cases}$

$\mathbb{E}(x) = \sum_x x\, P(x) = 0 \cdot (1-\mu) + 1 \cdot \mu = \mu$

$\mathbb{E}f(x) = \sum_x f(x) P(x) = f(0)(1-\mu) + f(1)\mu$

$\mathbb{E}(x^2) = 0(1-\mu) + 1(\mu) = \mu$

Variance of x

$= \mathbb{E}(x - \mathbb{E}(x))^2$

$= \mathbb{E}(x-\mu)^2$

$= \mathbb{E}(x^2 - 2x\mu + \mu^2)$

$= \mathbb{E}(x^2) - 2\mu\mathbb{E}(x) + \mu^2$

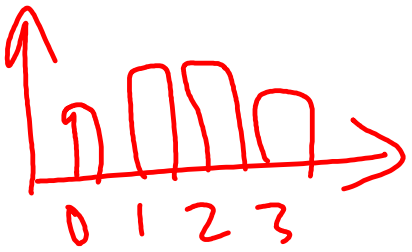
$= \mu(1-\mu)$

Binomial distribution

μ, N I.I.D.

$$Bin(y \mid \mu, N) = \binom{N}{y} \mu^{y} (1-\mu)^{N-y}$$

$$= {}^{N}C_{y}$$



0 0 0	0
0 0 1	1
0 1 0	1
0 1 1	2
1 0 0	1
1 0 1	2
1 1 0	2
1 1 1	3

$$E(y) = N\mu$$

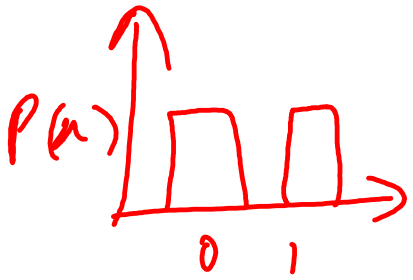
Entropy

Entropy of a random variable

$$H(x) = -E[\log(P(x))]$$

$$= -\sum_x \log[P(x)]\, P(x)$$

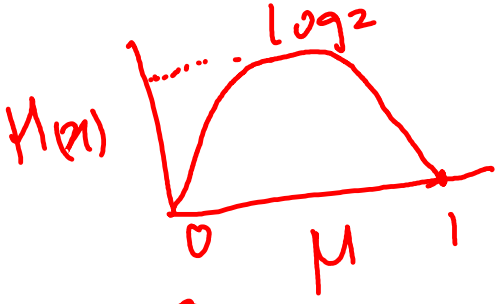
$$= -[(1-\mu)\log(1-\mu) + \mu \log \mu]$$



High $H(x)$

Low $H(x)$

$\mu = 0.5 \Rightarrow H(x) = -\log\left(\frac{1}{2}\right) = \log 2$



$\lim_{\mu \to 0}$ $H(x) = -(1-\mu)\log(1-\mu) - \mu \log \mu = 0$

Joint, conditional, and marginal probability

Dice: $x \in \{1,2,3,4,5,6\}$

$y \in \{0,1\}$, $0 \Rightarrow$ even

$z \in \{0,1\}$, $x < 3$

Joint

$P(y,z) \neq 0$

	$z=0$	$z=1$	$p(y)$
$y=0$	$1/6$	$2/6$	$1/2$
$y=1$	$1/6$	$2/6$	$1/2$
$p(z)$	$1/3$	$2/3$	

$\sum_y \sum_z p(y,z) = 1$

Marginal Dist

$p(z) = \sum_y p(y,z)$

$p(y) = \sum_z p(y,z)$

Conditional

$p(y|z) = \frac{p(y,z)}{p(z)}$

$p(y|z=0)$

$p(y=0|z=0) = \frac{1/6}{1/3} = \frac{1}{2}$

Joint, conditional, and marginal probability

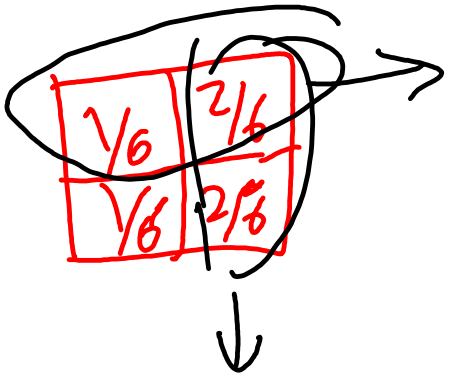
Dice $x \in \{1, 2, 3, 4, 5, 6\}$

$y \in \{0, 1\}$ 0 => Even, 1 means odd

$z \in \{0, 1\}$ $0 \Rightarrow x < 3$, $1 \Rightarrow x \geq 3$

$P(y, z)$

	$z=0$	$z=1$
$y=0$	$\{2\}$ 1/6	$\{4,6\}$ 2/6
$y=1$	$\{1\}$ 1/6	$\{3,5\}$ 2/6



$\sum_y \sum_z P(y, z) = 1$

$P(y, z) \in [0, 1]$

$P(y,z)$

	$z=0$	$z=1$
$y=0$	$1/6$	$2/6$
$y=1$	$1/6$	$2/6$

$P(y \mid z)$ ← given

Conditional

$$P(x \text{ is Even} \mid x \geq 3) = \frac{2/6}{2/6 + 2/6} = \frac{1}{2}$$

$$P(y \mid z) = \frac{P(y,z)}{P(z)} = \frac{2/6}{4/6} = \frac{1}{2}$$

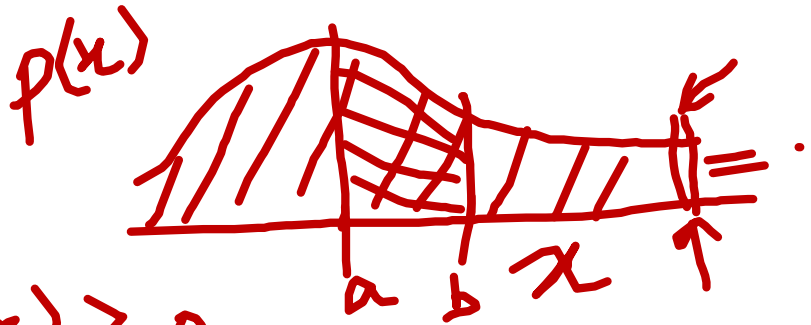
$$P(y,z) = P(y \mid z)\, P(z)$$

Marginal $P(z) = \sum_y P(y,z) = \frac{2}{6} + \frac{2}{6} = \frac{4}{6}$

Continuous random variable and PDF

$P(X = x) = 0$ $x = 170\,cm$

$p(x)$



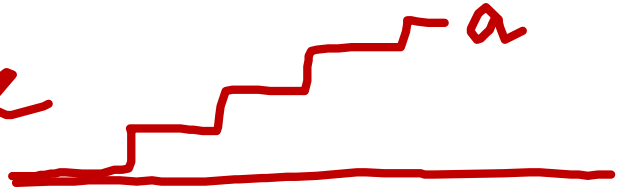
$$\int_{-\infty}^{\infty} p(x)\,dx = 1 \qquad p(x) \geq 0$$

$$P(a < x \leq b) = \int_{a}^{b} p(x)\,dx$$

Cumulative D.F.

$$\int_{-\infty}^{a} p(x)\,dx$$

Discrete



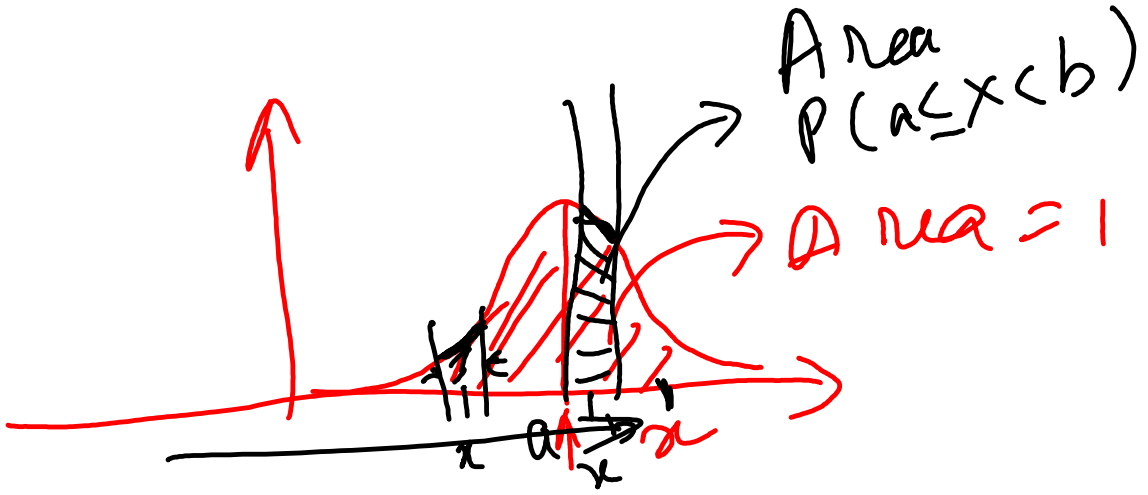
Continuous random variable and PDF

$P(x = x) = 0$

Prob. density fn.

$p(x) \geq 0$

$\int_{-\infty}^{\infty} p(x) = 1$



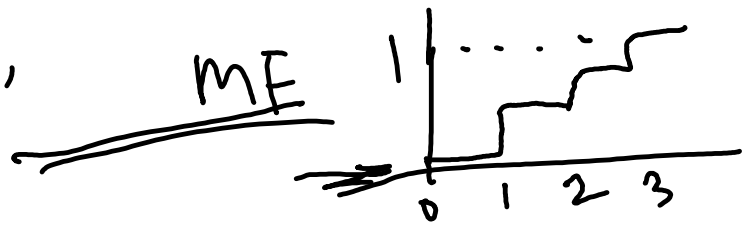
170 cm

169.99

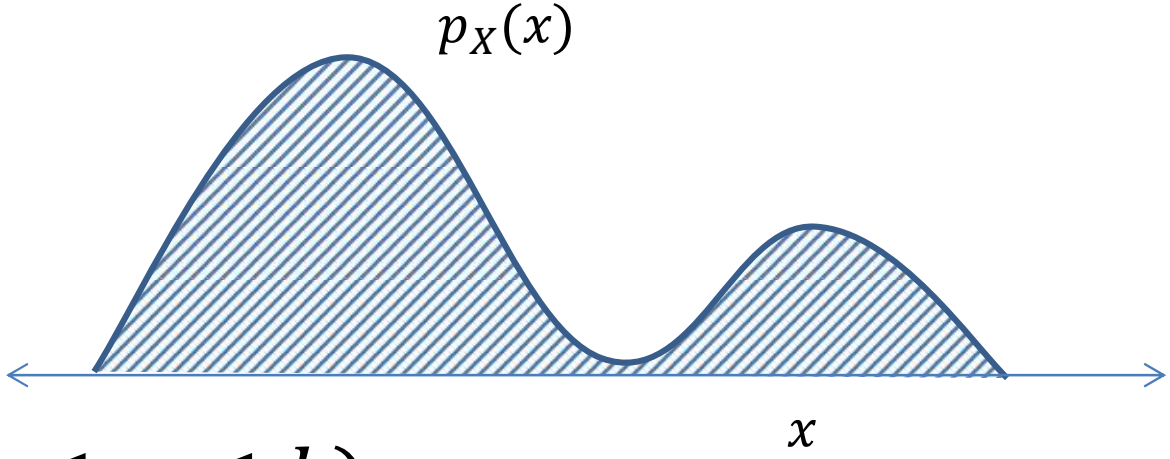
170.01

Cont. R.V. Cumulative DF $= \int_{-\infty}^{x} P_x(y)\,dy$

Discrete " " MF



Probability density function (PDF)

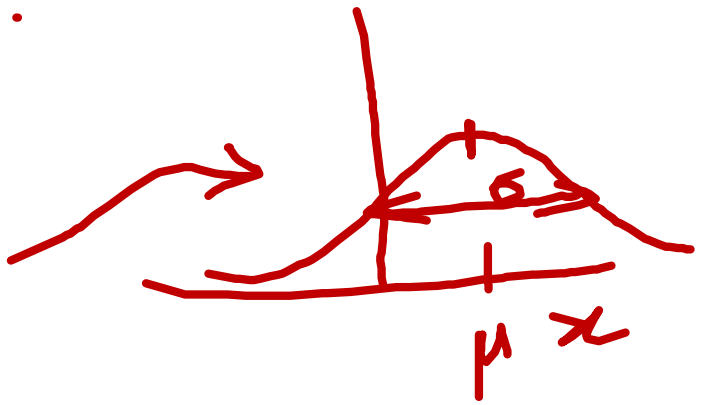


- PDF: $p_X(x) \geq 0$
- $\int_a^b p_X(\hat{x})d\hat{x} = Pr(a \leq x \leq b)$
- $\int_{-\infty}^{+\infty} p_X(\hat{x})d\hat{x} = 1$
- CDF: $P_X(x) = \int_{-\infty}^{x} p_X(\hat{x})d\hat{x}$
- $p_X(x) = \frac{dP_X(x)}{dx}$

Some common PDFs

Gaussian / Normal Dist.

$$\mathcal{N}(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



$$E(x) = \int_{-\infty}^{\infty} x\, p(x)\, dx = \mu$$

$$E[(x-\mu)^2] = \sigma^2$$

Unif$(x;a,b)$

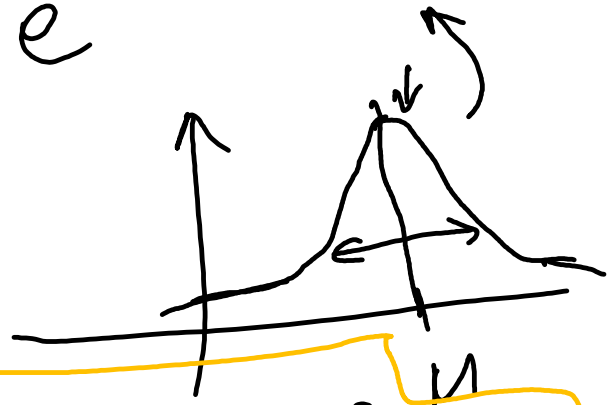
Beta dist. $x \in [0,1]$

$$\text{Beta}(x;a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$$

Some common PDFs

Gaussian / Normal

$$\mathcal{N}(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]}$$



$$H(x) = \int_{-\infty}^{\infty} [\log p(x)]\, p(x)\, dx$$

$$\mathbb{E}[f(x)] = \int_{-\infty}^{\infty} f(x)\, p(x)\, dx \qquad \mathbb{E}[x] = \int_{-\infty}^{\infty} x\, p(x)\, dx$$

Beta

Unif.

$\frac{1}{b-a}$

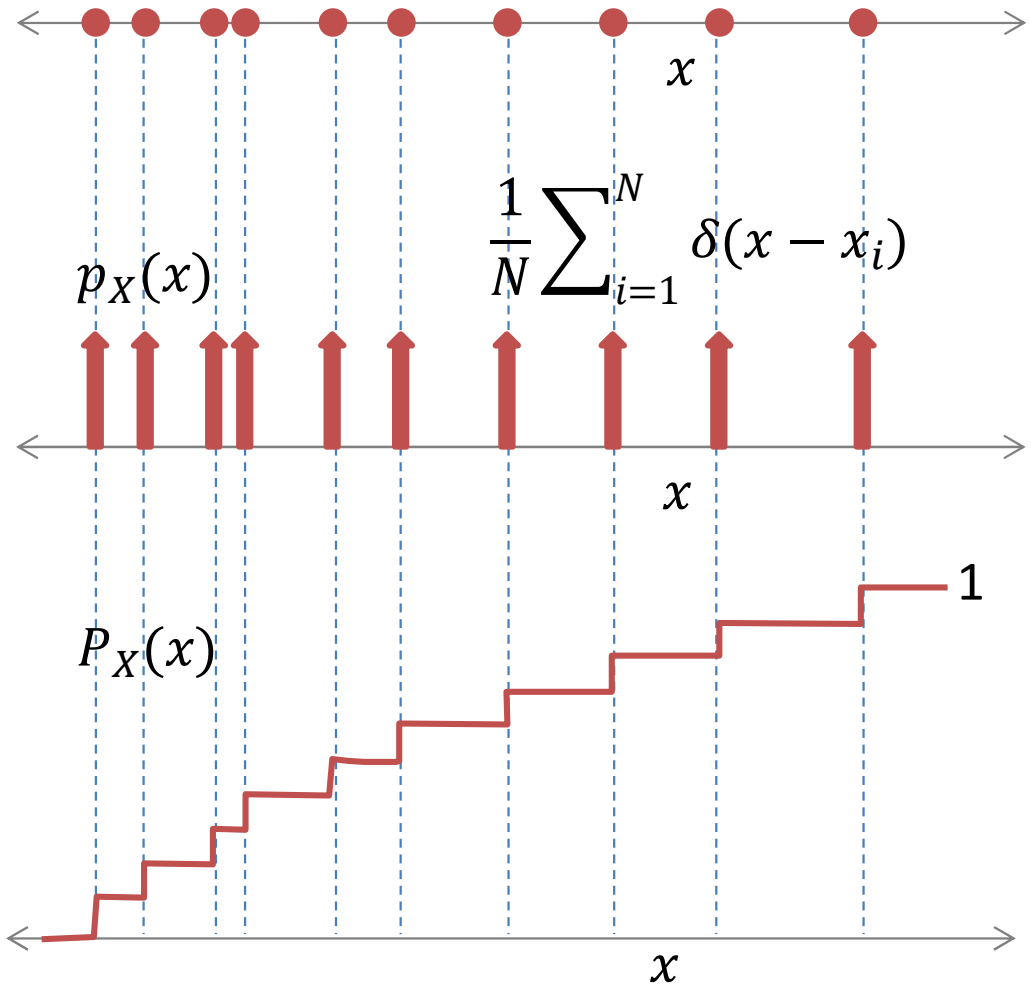
a, b $\quad \mathrm{Beta}(x|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$

for $0 \le x \le 1$

0 otherwise

Empirical distribution

- Given samples
- What is the “best” explanation of the data?
- Empirical PDF (train of Dirac delta functions scaled 1/N)
- CDF has N steps



Problem with empirical distribution

- Rote learning
- Does not generalize
- If we randomly split the sample into training and validation
- How likely is the validation subset based on the distribution learnt from the training subset?

Between two distributions, which is better?

- The one under which the data has higher likelihood
- If $p_A(x_i) > p_B(x_i)$ then p_A is a better distribution to explain the observation x_i
- This can be maximized by having a Dirac delta function $\delta(x - x_i)$
- What about for the entire data sample X?

The IID Assumption

- Independence of x_i and x_j for $i \neq j$.
- $p(x_i, x_j) = p(x_i)p(x_j)$

- All samples drawn from the same (identical) distributed
- $x_i, x_j \sim p_X$

Implication of the IID assumption

- Likelihood of the whole data factorizes:

- $p(X) = p(x_1, \ldots, x_N) = p_X(x_1) \times \ldots \times p_X(x_N)$
  $= \prod_{i=1}^{N} p_X(x_i)$

- $L(X) = \log p(X) = \sum_i \log p_X(x_i)$

MLE of parameterized distribution

- Between two distributions p_A is a better explanation than p_B of the entire data X if $\prod_i p_A(x_i) > \prod_i p_B(x_i)$

- By extension, if a family of distributions is parameterized by θ, then we are interested in

$$\begin{aligned} \arg\max_\theta \prod_i p_\theta(x_i) &= \arg\max_\theta \sum_i \log p_\theta(x_i) \\ &= \arg\max_\theta L_\theta(X) \end{aligned}$$

Example 1: Exponential distribution

- $p_\lambda(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0 & , \ x < 0 \end{cases}$
- Log likelihood $L_\lambda(X) = \sum_i \log p_\lambda(x_i)$
- Differntiating $\frac{\partial L_\lambda(X)}{\partial \lambda} = \sum_{x_{i \geq 0}} \left(\frac{\partial \log \lambda}{\partial \lambda} - \frac{\partial \lambda x_i}{\partial \lambda} \right) = 0$
- (assuming all samples are non-negative)
- Implies $\lambda = \frac{N}{\sum_i x_i}$ = inverse of the sample mean

Example 2: Uniform distribution

- $p_{a,b}(x) = \begin{cases} \frac{1}{b-a}, a \leq x \leq b \\ 0 \quad , \text{otherwise} \end{cases}$
- Log likelihood $L_{a,b}(X) = \sum_i \log p_{a,b}(x_i)$
- $= K \log(0) \; - (\mathrm{N} - \mathrm{K}) \log(b - a)$
- Reduce the K (terms outside $[a, b]$) to zero, and minimize $(b - a)$ by differentiating wrt a, b
- So, $a = \min x_i$, $b = \max x_i$

Example 3: Gaussian distribution

- $p_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
- MLE by differentiating log likelihood wrt μ, σ
- gives $\mu = \bar{x} = \frac{\sum_i x_i}{N}$; i.e., sample mean
- and, $\sigma = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{N}}$

Sufficient statistics

- Statistic is a function of the sample

$$T(X) = T(x_1, \ldots, x_N)$$

- For some distributions, computing a few statistics is sufficient for MLE estimate
- Gives complete information about the distribution
- Examples:
  - Sample mean and variance for Gaussian distribution
  - Sample mean for exponential distribution
  - Max and min for uniform distribution

Comparing two parametric distributions

- Let there be two candidate families of distributions $p_\theta(X)$ and $p_\phi(X)$ to explain the data
- Can we compare $\max L_\theta(X)$ and $\max L_\phi(X)$?
- Yes, we can, but we might overfit
- Narrow down the family of distributions based on domain knowledge (e.g. physical phenomenon)
- E.g. “Can the random variable take negative values?”

MLE vs. Bayesian estimate

- MLE finds θ that maximizes $L_\theta(X)$
- Bayesian estimate takes the expected value of θ w.r.t. $L_\theta(X)$
- Bayesian estimate: $\int \theta L_\theta(X) d\theta \ / \ \int L_\theta(X) d\theta$
- We can also incorporate a prior over θ
- $\int \theta p_\Theta(\theta) L_\theta(X) d\theta \ / \ \int p_\Theta(\theta) L_\theta(X) d\theta$

Multivariate PDF

$$p(x_1, x_2) \geq 0$$

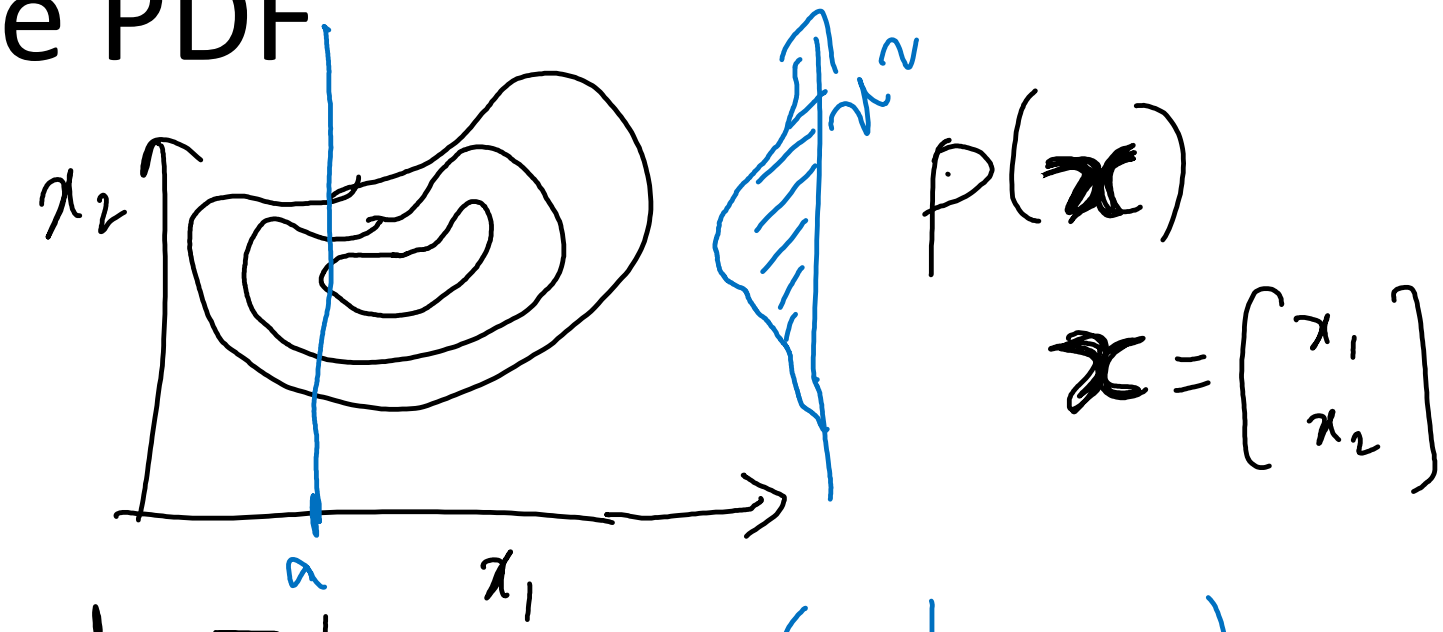
$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x_1, x_2)\, dx_1\, dx_2 = 1$$

$$\int_{-\infty}^{\infty} p(x_1, x_2)\, dx_1 = p(x_2)$$

$$p(x_2 \mid x_1 = a) = \frac{p(x_1 = a, x_2)}{p(x_1 = a)}$$

Multivariate PDF

$p(x_1, x_2)$



$p(x_1, x_2) \geq 0$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x_1, x_2)\, dx_1\, dx_2 = 1$$

$$p(x_1) = \int_{-\infty}^{\infty} p(x_1, x_2)\, dx_2$$

marginal

Conditional

$$p(x_2 \mid x_1 = a) = \frac{p(x_1 = a, x_2)}{p(x_1 = a)}$$

Multivariate Gaussian

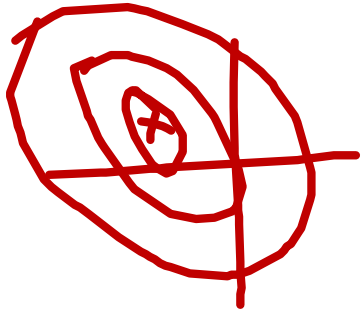
$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

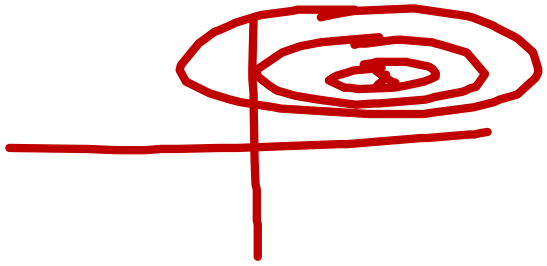
$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12} \\ \sigma_{21} & \sigma_{22}^2 \end{bmatrix}$$

$p(x_1, x_2)$

$$p(x) = \frac{1}{\sqrt{(2\pi)^k \, \det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

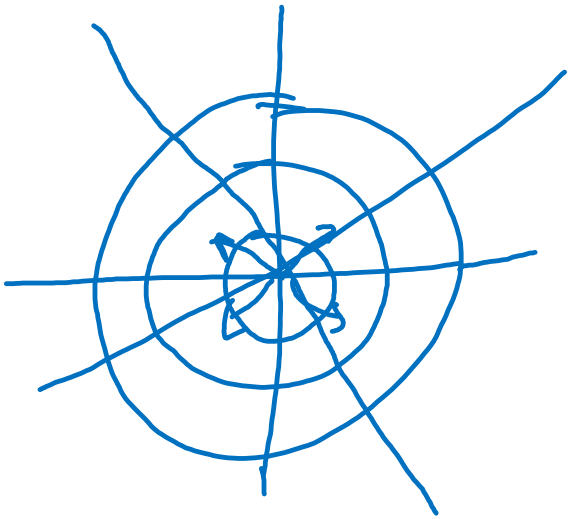




Multivariate Gaussian PDF

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12} \\ \sigma_{21} & \sigma_{22}^2 \end{bmatrix}$$



$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^k \det(\Sigma)}} \times e^{\left[-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu)\right]}$$

$1 \times k \quad k \times k \quad k \times 1 \quad 1 \times 1$