GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Deep Learning

Lecture 2: Math basics

Alexander Ecker
Institut für Informatik, Uni Göttingen

alxecker

https://alexanderecker.wordpress.com

– Credit: slides largely based on Ian Goodfellow's and Chris Bishop's slides –

Recommended literature

Primary book:

- **Goodfellow, Bengio, Courville: Deep Learning**
  https://www.deeplearningbook.org

Good general machine learning book:

- Christopher Bishop: Pattern Recognition and Machine Learning
  https://www.microsoft.com/en-us/research/people/cmbishop/prml-book/

Good resource for Deep Reinforcement Learning (last week only):

- OpenAI: Spinning Up in Deep RL
  https://openai.com/blog/spinning-up-in-deep-rl/

Homework assignments

First homework assignment on Thursday

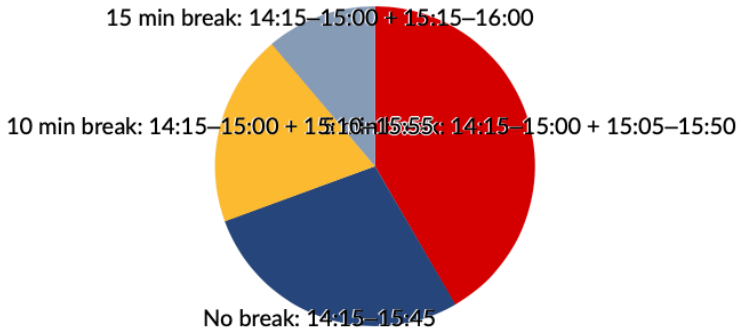
Form teams of 2–3 students

- No individuals please
- You'll present your solution to the tutors in tutorial (5 min)
- It's ok to split work, but everybody needs to be able to explain everything

Details on submission on Thursday

To break or not to break?

Should we take a break after 45 minutes or do 90 min straight?



No break: 14:15–15:45	(28% | 10/36)
5 min break: 14:15–15:00 + 15:05–15:50	(42% | 15/36)
10 min break: 14:15–15:00 + 15:10–15:55	(19% | 7/36)
15 min break: 14:15–15:00 + 15:15–16:00	(11% | 4/36)

$\mathbb{E}[\text{break length}] = 5 \text{ min } 39 \text{ sec}$ → **5 min break**

Topics today

Linear algebra

Probability theory

Information theory

Linear algebra

Notation

Scalar: a single number

$$a, n, x \in \mathbb{R}$$

0d tensor 

Vector: 1d array of numbers

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n$$

1d tensor 

Matrix: 2d array of numbers

$$\boldsymbol{A} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \in \mathbb{R}^{m \times n}$$

2d tensor 

Tensor: *N*-dim array of numbers $\boldsymbol{W} = \left(W_{i_1 i_2 \cdots i_N}\right)$

3d tensor 

Matrix transpose

$$(\boldsymbol{A}^T)_{ij} = A_{ji}$$

$$\boldsymbol{A} = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \end{bmatrix} \Rightarrow \boldsymbol{A}^T = \begin{bmatrix} A_{11} & A_{21} \\ A_{12} & A_{22} \\ A_{13} & A_{23} \end{bmatrix}$$

The matrix transpose is a mirror image across the diagonal

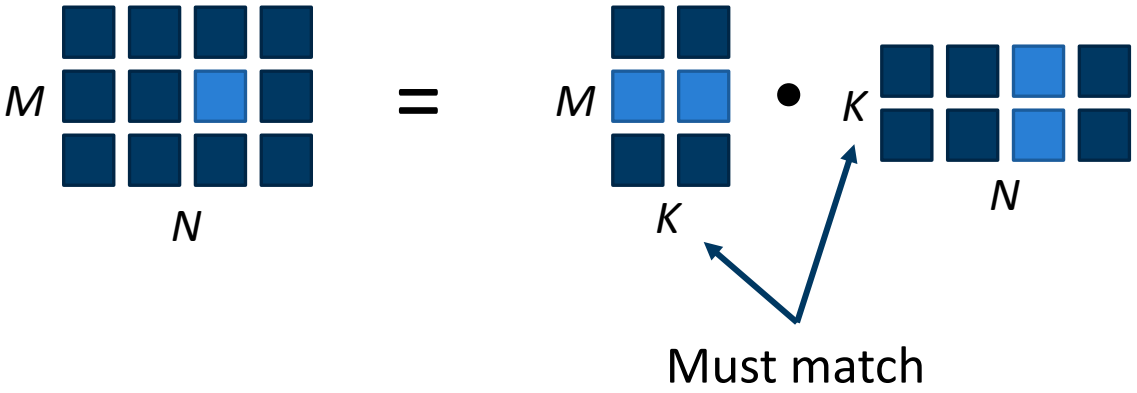
$$(\boldsymbol{AB})^T = \boldsymbol{B}^T \boldsymbol{A}^T$$

Dot product

(a.k.a. inner product, matrix product)

$$\boldsymbol{C} = \boldsymbol{AB}$$

$$C_{ij} = \sum_{k=1}^{K} A_{ik} B_{kj}$$



Identity matrix

$$\boldsymbol{I}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\boldsymbol{Ix} = \boldsymbol{x}$$

$$\boldsymbol{IA} = \boldsymbol{A} = \boldsymbol{AI}$$

Linear system of equations

$$\boldsymbol{Ax} = \boldsymbol{b}$$

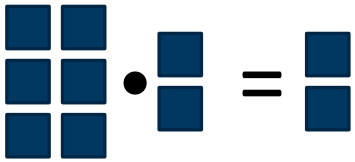
expands to

$$\begin{aligned} A_{1:}\boldsymbol{x} &= b_1 \\ A_{2:}\boldsymbol{x} &= b_2 \\ &\dots \\ A_{m:}\boldsymbol{x} &= b_m \end{aligned}$$

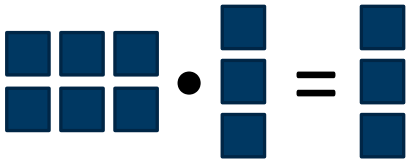
Solving linear systems of equations

A linear system of equations $\boldsymbol{Ax} = \boldsymbol{b}$ can have…

- No (exact) solution
  $\boldsymbol{A}$ is *tall*



- Many solutions
  $\boldsymbol{A}$ is *fat*



- Exactly one solution
  $\boldsymbol{A}$ is *square* and *invertible*

Matrix inversion

Matrix inverse

$$\boldsymbol{A^{-1}A = I = AA^{-1}}$$

Solving linear system via matrix inverse

$$\begin{aligned} \boldsymbol{Ax} &= \boldsymbol{b} \\ \boldsymbol{A^{-1}Ax} &= \boldsymbol{A^{-1}b} \\ \boldsymbol{Ix} &= \boldsymbol{A^{-1}b} \\ \boldsymbol{x} &= \boldsymbol{A^{-1}b} \end{aligned}$$

Note: don’t do it like this in practice!

Numerically unstable, but useful for analytical analysis

Norms

Measures the "length" of a vector

If $\|\cdot\|$ represents a norm, it must hold:

(1) $\|\boldsymbol{x}\| = 0 \quad \Rightarrow \quad \boldsymbol{x} = 0$

(2) $\|\boldsymbol{x} + \boldsymbol{y}\| < \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$ (triangle inequality)

(3) $\|a \cdot \boldsymbol{x}\| = |a| \cdot \|\boldsymbol{x}\|$

L_p norm (most popular: L_2):

$$\|\boldsymbol{x}\|_p = \left(\sum_i |x_i|^p\right)^{1/p}$$

L_1 norm: $\|\boldsymbol{x}\|_1 = \sum|x_i|$

Maximum (L_∞) norm: $\|\boldsymbol{x}\|_\infty = \max_i |x_i|$

Special matrices and vectors

Unit vector:

$$\|\boldsymbol{x}\| = 1$$

Symmetric matrix:

$$\boldsymbol{A}^T = \boldsymbol{A}$$

Orthogonal matrix:

$$\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{I}$$
$$\boldsymbol{A}^T = \boldsymbol{A}^{-1}$$

Eigenvalue decomposition

Eigenvector $\boldsymbol{v}$ and eigenvalue λ:

$$\boldsymbol{A}\boldsymbol{v} = \lambda\boldsymbol{v}$$

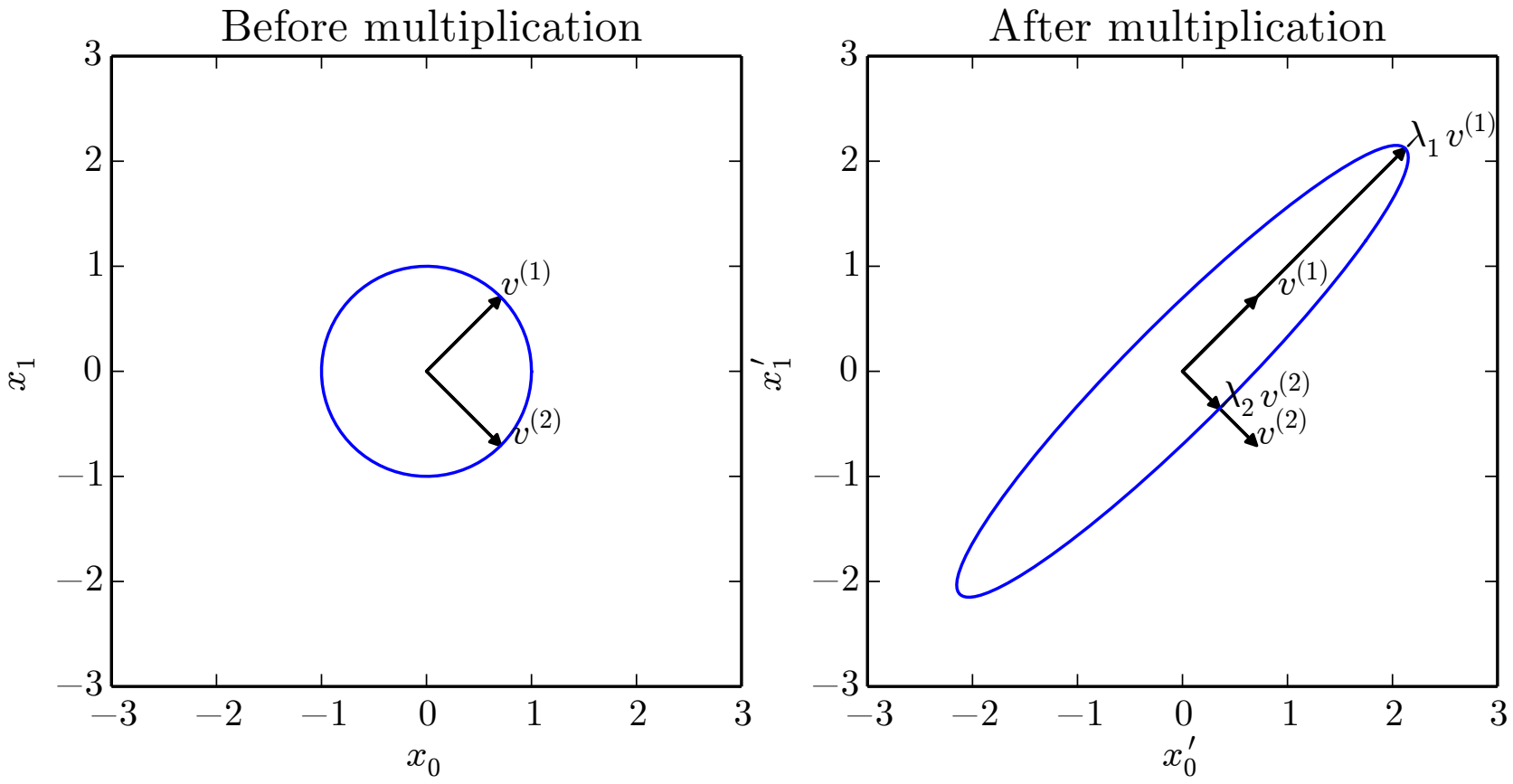
Eigenvalue decomposition of a diagonalizable matrix $\boldsymbol{A}$

$$\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T$$

$$\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I} \qquad \boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix}$$

Every real symmetric matrix is diagonalizable

Eigenvalue decomposition graphically



Singular value decomposition

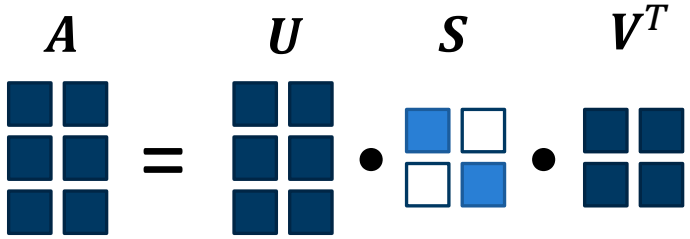
Similar to eigenvalue decomposition

More general: matrix need not be square

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T$$

Singular values

$$\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I} \qquad \boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I} \qquad \boldsymbol{S} = \begin{bmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_n \end{bmatrix}$$



Matrix pseudoinverse

(a.k.a. Moore-Penrose inverse)

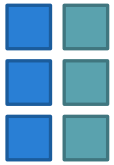
$$\boldsymbol{A}\boldsymbol{A}^{\dagger}\boldsymbol{A} = \boldsymbol{A} \qquad \boldsymbol{A}^{\dagger}\boldsymbol{A}\boldsymbol{A}^{\dagger} = \boldsymbol{A}^{\dagger}$$

Symmetric

$$\left(\boldsymbol{A}^{\dagger}\boldsymbol{A}\right)^{T} = \boldsymbol{A}^{\dagger}\boldsymbol{A} \qquad \left(\boldsymbol{A}\boldsymbol{A}^{\dagger}\right)^{T} = \boldsymbol{A}\boldsymbol{A}^{\dagger}$$

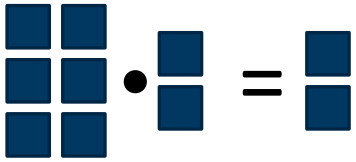
For linearly independent columns of $\boldsymbol{A}$:

$$\boldsymbol{A}^{\dagger} = (\boldsymbol{A}^{T}\boldsymbol{A})^{-1}\boldsymbol{A}^{T}$$
$$\boldsymbol{A}^{\dagger}\boldsymbol{A} = \boldsymbol{I}$$



Least-squares solution to overdetermined system of linear equations:

$$\boldsymbol{x} = \boldsymbol{A}^{\dagger}\boldsymbol{b}$$



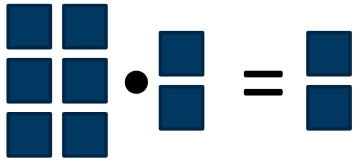
Matrix pseudoinverse

System of linear equations:

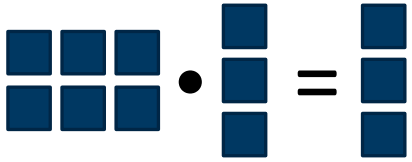
$$\boldsymbol{x} = \boldsymbol{A}^{\dagger}\boldsymbol{b}$$

1. System has exactly one solution:
   same as $\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b}$

2. System has no solution:
   $\boldsymbol{x}$ with the smallest quadratic error $\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2$



3. System has many solutions:
   $\boldsymbol{x}$ with the smallest L_2 norm $\|\boldsymbol{x}\|_2$



Trace

Sum of the diagonal elements of a matrix

$$\mathrm{Tr}(\boldsymbol{A}) = \sum_{\boldsymbol{i}} A_{ii}$$

Trace permutes cyclically

$$\mathrm{Tr}(\boldsymbol{ABC}) = \mathrm{Tr}(\boldsymbol{CAB}) = \mathrm{Tr}(\boldsymbol{BCA})$$

Determinant

For square matrices:

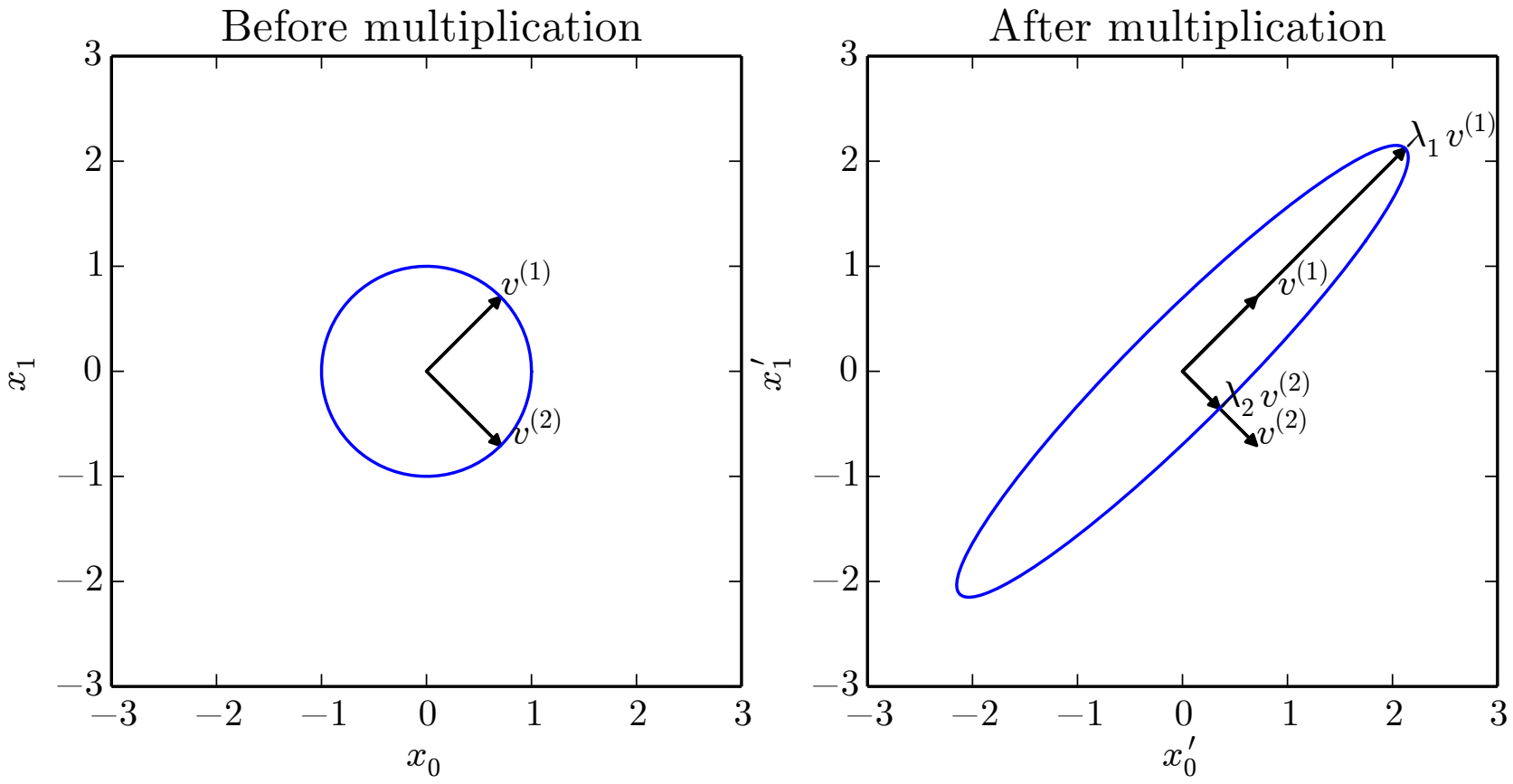
$$\det(\boldsymbol{A}) = \prod_k \lambda_k$$ ← Eigenvalues

Absolute value of the determinant measures how much multiplication by a matrix expands or contracts the space

Matrix $\boldsymbol{A}$ is invertible if $\det(\boldsymbol{A}) \neq 0$

Volume-conserving transformations have $\det(\boldsymbol{A}) = 1$

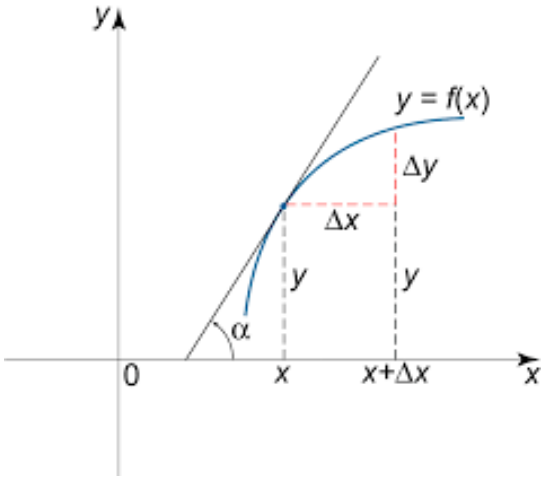
Eigenvalue decomposition graphically

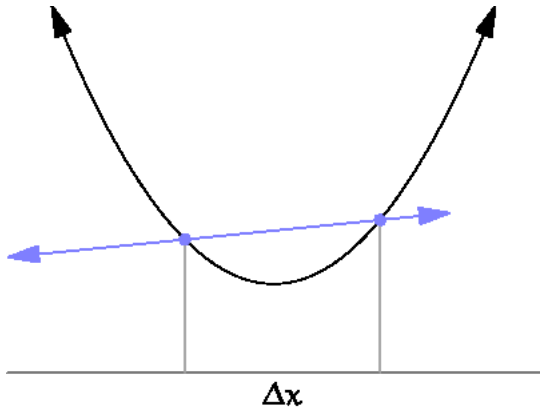


Derivative

The derivate $f'(x)$ of a function $f(x)$ is its slope at x

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

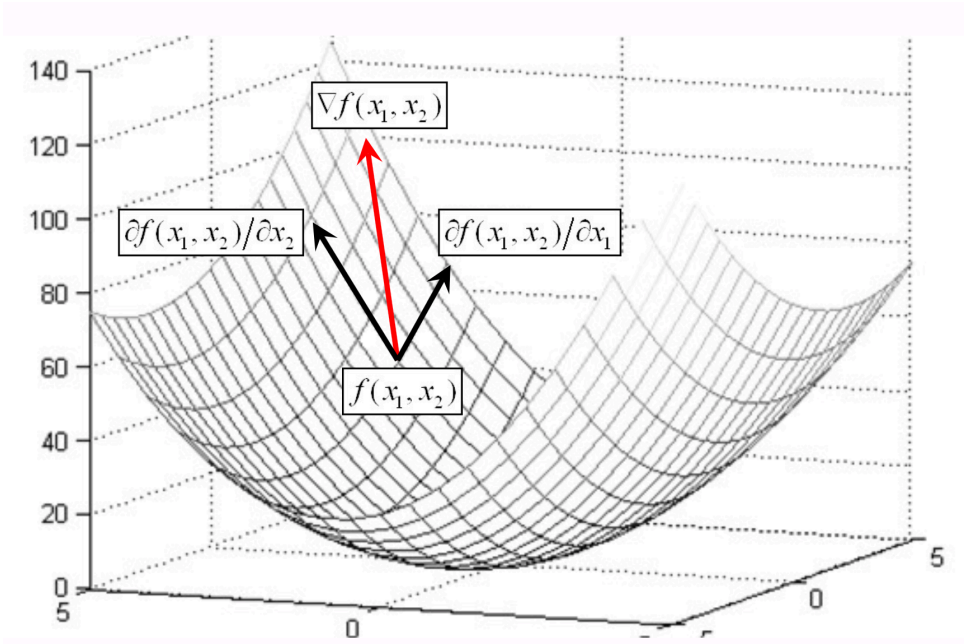




Gradient

"Derivative" of a function $f(\boldsymbol{x})$ with vector-valued inputs $\boldsymbol{x}$
Points into direction of steepest ascent

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = \frac{\partial f}{\partial \boldsymbol{x}}(\boldsymbol{x}) = [f'(x_1), \ldots, f'(x_n)]$$



Jacobian

Gradient matrix of function $\boldsymbol{f}(\boldsymbol{x})$ with vector-valued input and output

$$J_{\boldsymbol{x}} = \begin{bmatrix} \nabla_{\boldsymbol{x}} f_1(\boldsymbol{x}) \\ \vdots \\ \nabla_{\boldsymbol{x}} f_m(\boldsymbol{x}) \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{x_1} & \cdots & \frac{\partial f_1}{x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{x_1} & \cdots & \frac{\partial f_m}{x_n} \end{bmatrix}$$

Chain rule

Derivate of compositions

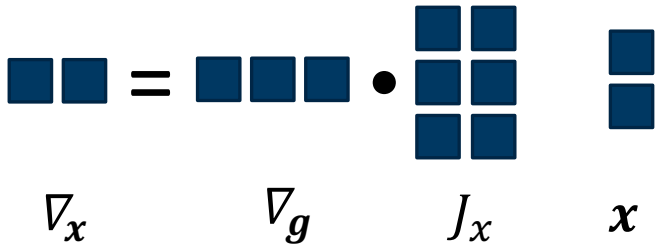
$$f'\big(g(x)\big) = \frac{\partial f}{\partial g}\frac{\partial g}{\partial x}(x)$$

Example $f(x) = \sin(x^2)$

$$f'(x) = \frac{\partial \sin(g)}{\partial g} \cdot \frac{\partial x^2}{\partial x} = \cos(x^2) \cdot 2x$$

Vector-valued functions:

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{g}(\boldsymbol{x})) = \frac{\partial f}{\partial \boldsymbol{g}} \cdot \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}}(\boldsymbol{x})$$



Linear algebra cheat sheet

The Matrix Cookbook
http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=3274

Contains lots of useful linear algebra identities

Probability theory

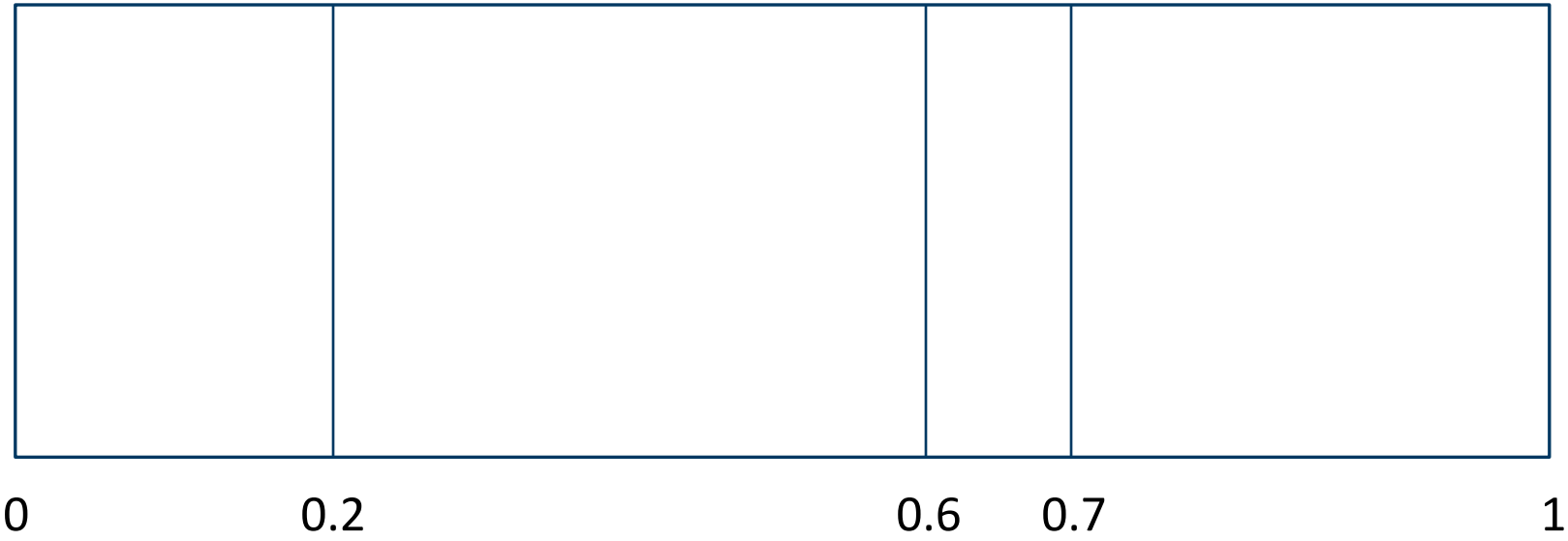
Why probability theory?

Reasoning about the future state of the world requires dealing with uncertainty

- Inherent stochasticity in the system
- Observation noise
- Incomplete observability
- Incomplete modeling

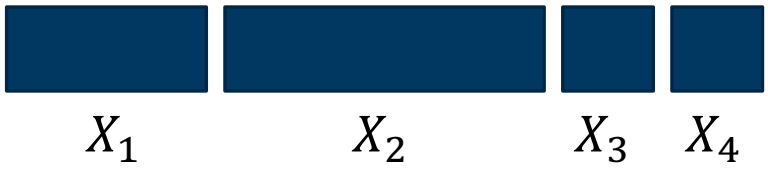
Events

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$



Discrete vs. continuous random variables

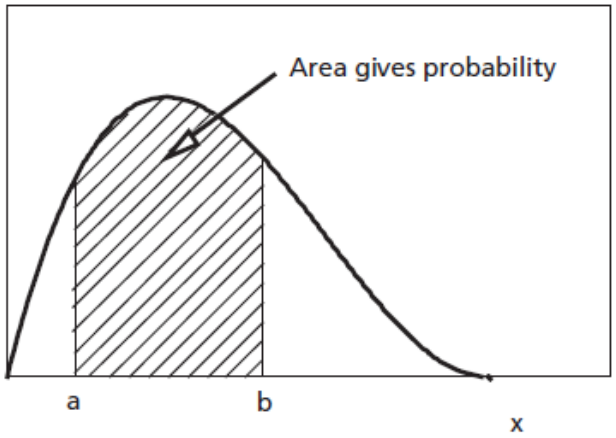
Discrete



Probability mass function

$$P(X = X_1)$$

Continuous

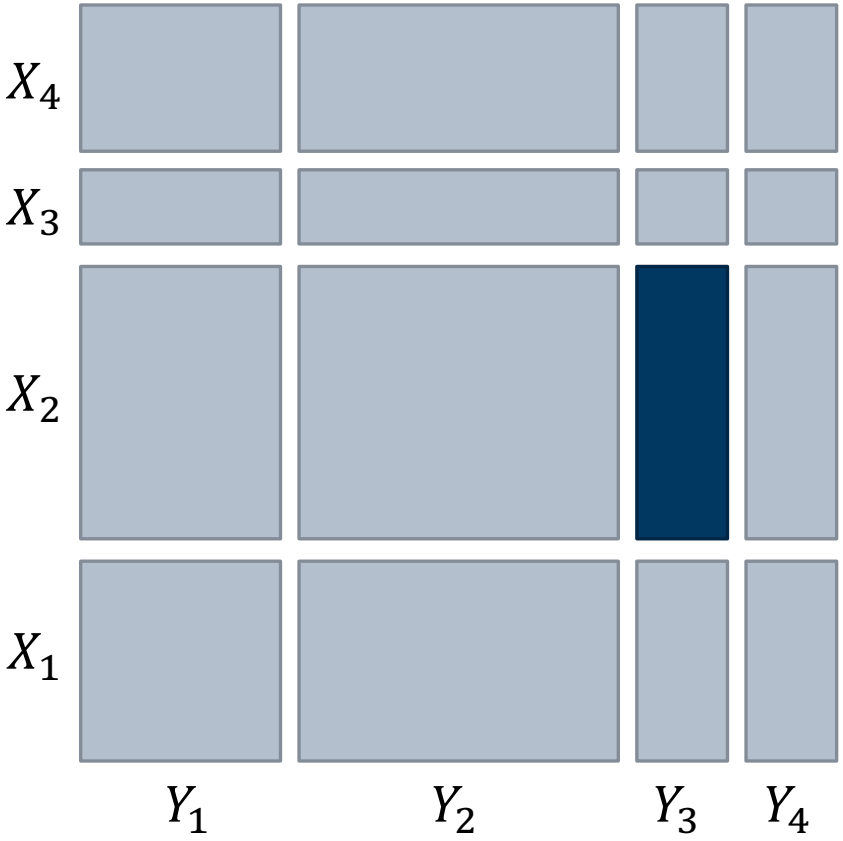


Probability density function $p(x)$

$$P(X \in [a, b]) = \int_a^b p(x)dx$$

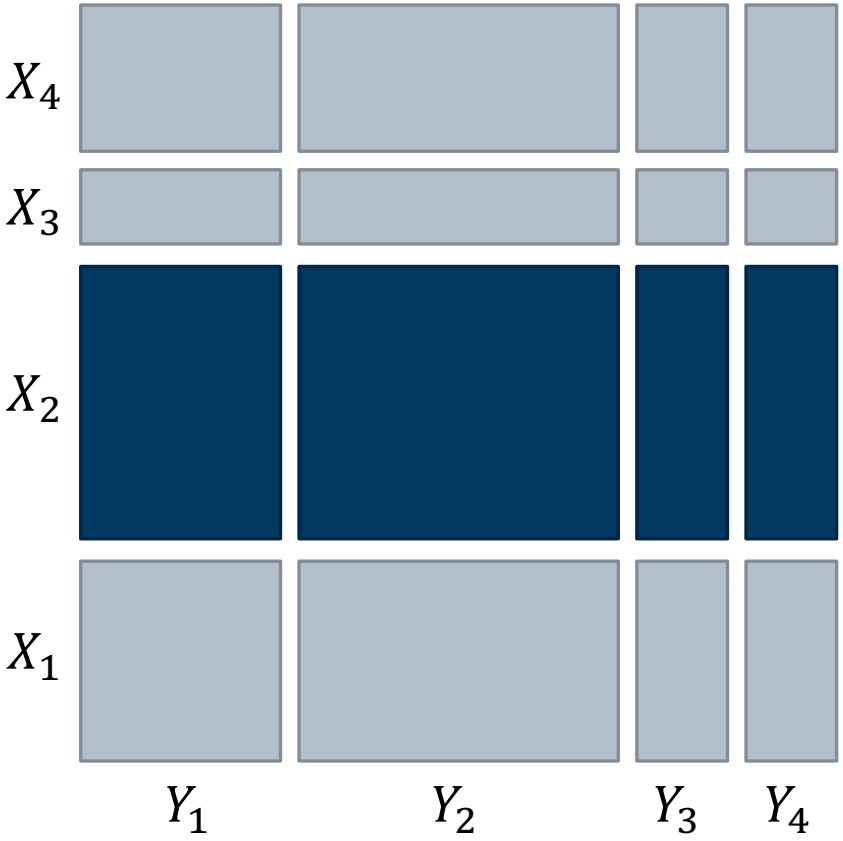
https://intellipaat.com/blog/tutorial/statistics-and-probability-tutorial/probability-distributions-of-continuous-variables/

Joint probability



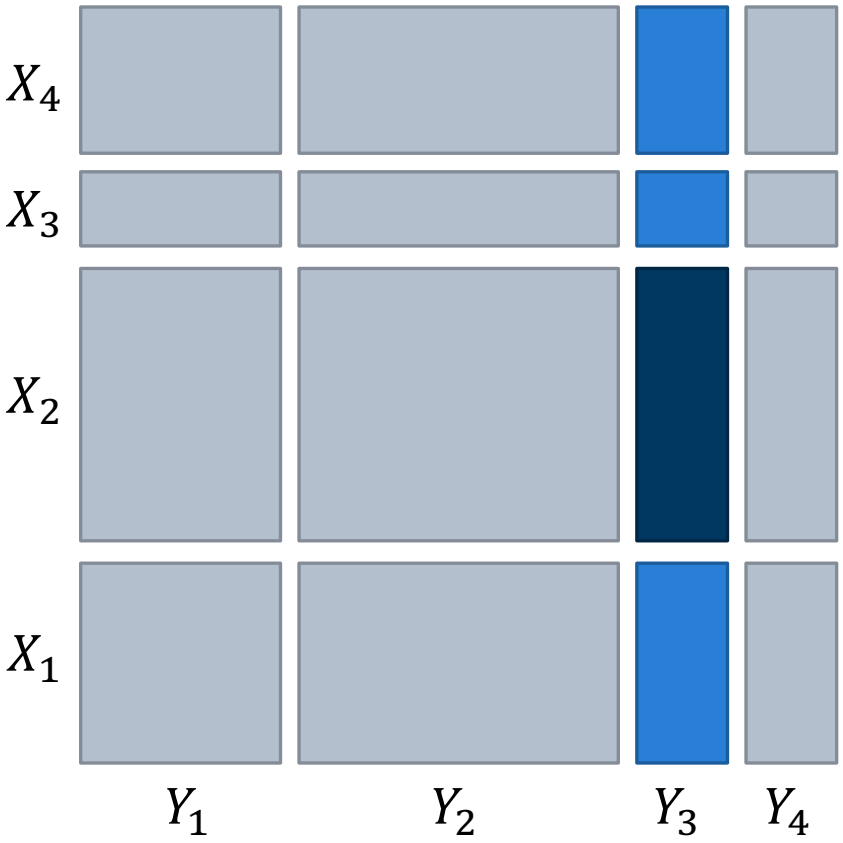
$P(X_2, Y_3)$

Marginal probability



$$P(X_2) = \sum_k P(X_3, Y_k)$$

Conditional probability



$$P(X_2|Y_3) = \frac{P(X_2, Y_3)}{P(Y_3)}$$

Bayes' rule

$$= \begin{cases} P(X,Y) = P(X|Y)P(Y) \\ P(Y,X) = P(Y|X)P(X) \end{cases}$$

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Posterior $\propto$ Likelihood $\times$ Prior

Example: Bayes' rule

Morse code (long "–" or short "•") transmitted through a noisy line

$$P(S = \text{“–”}) = \frac{4}{7} \qquad P(S = \text{“•”}) = \frac{3}{7} \qquad P(\text{flip}) = \frac{1}{8}$$

A "•" was received. What is the probability that a "•" was sent?

$$\begin{aligned} P(R = \text{“•”}) &= P(R = \text{“•”} | S = \text{“•”}) P(S = \text{“•”}) + \\ &\quad P(R = \text{“•”} | S = \text{“–”}) P(S = \text{“–”}) \\ &= \frac{7}{8} \cdot \frac{3}{7} + \frac{1}{8} \cdot \frac{4}{7} = \frac{25}{56} \end{aligned}$$

$$P(S = \text{“•”} | R = \text{“•”}) = \frac{P(R = \text{“•”} | S = \text{“•”}) P(S = \text{“•”})}{P(R = \text{“•”})} = \frac{\frac{7}{8} \cdot \frac{3}{7}}{\frac{25}{56}} = \frac{21}{25} = 84\%$$

http://www.sci.utah.edu/~gerig/CS6640-F2010/prob-tut.pdf

Independence

X and Y are ***independent*** if and only if

$$P(X, Y) = P(X)P(Y) \qquad \text{for all } X, Y$$

X and Y are ***conditionally*** independent if and only if

$$P(X, Y|Z) = P(X|Z)P(Y|Z) \qquad \text{for all } X, Y, Z$$

Expectation

Discrete random variables:

$$\mathbb{E}[X] = \sum_k P(X = x_k) x_k$$

Continuous random variables

$$\mathbb{E}[X] = \int_{\mathcal{X}} p(x) x \, dx$$

Expectation is a linear operation:

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$$

Variance and covariance

Variance measures the spread of a distribution
(average squared deviation from the mean)

$$\mathrm{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

Standard deviation: $\mathrm{SD}[X] = \sqrt{\mathrm{Var}[X]}$

Covariance measures the co-fluctuations of two random variables

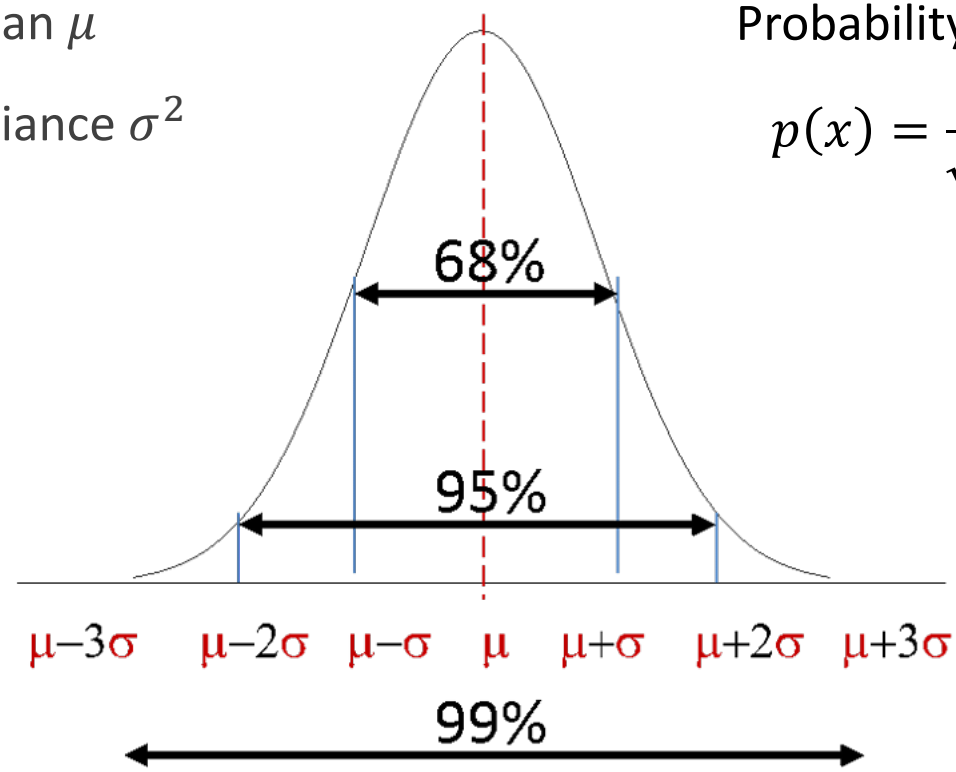
$$\begin{aligned}\mathrm{Cov}[X, Y] &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]\end{aligned}$$

Covariance matrix $\boldsymbol{C}_{ij} = \mathrm{Cov}\left[X_i, X_j\right]$

Gaussian distribution

Mean μ

Variance σ^2



Probability density function

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

http://sphweb.bumc.bu.edu/otlt/MPH-Modules/PH717-QuantCore/PH717-Module6-RandomError/PH717-Module6-RandomError5.html

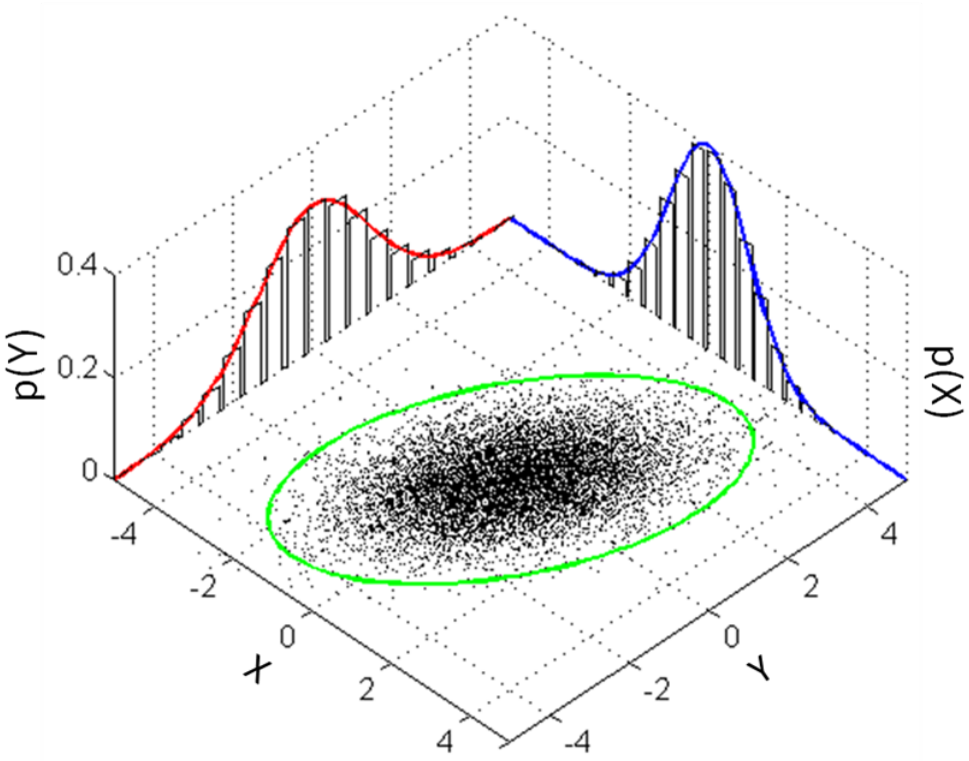
Multivariate normal distribution

Probability density function

$$p(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{C}|}} \exp(-(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{C}^{-1} (\boldsymbol{x}-\boldsymbol{\mu}))$$

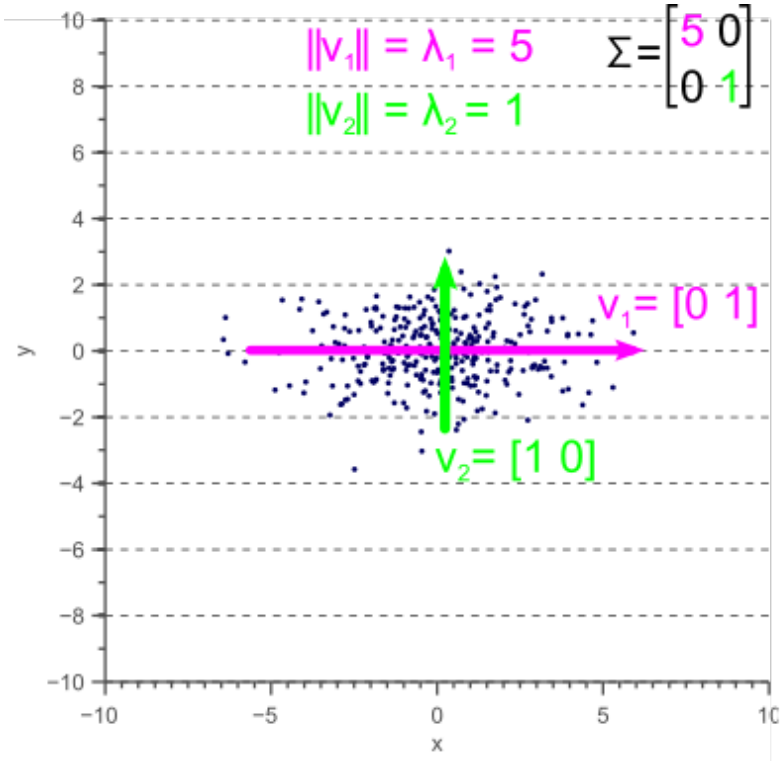
Mean $\boldsymbol{\mu}$

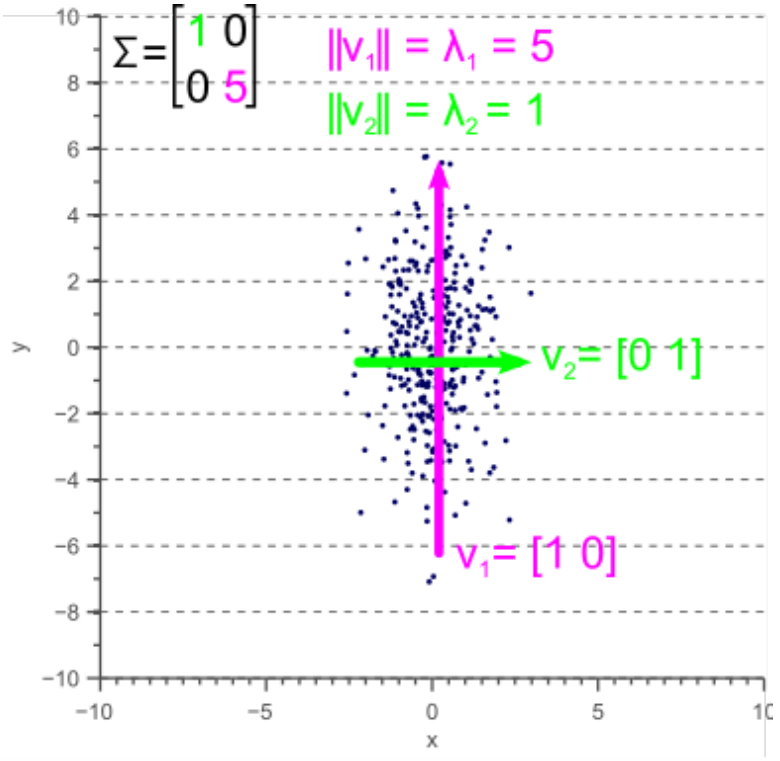
Covariance matrix $\boldsymbol{C}$



https://www.wikiwand.com/en/Multivariate_normal_distribution

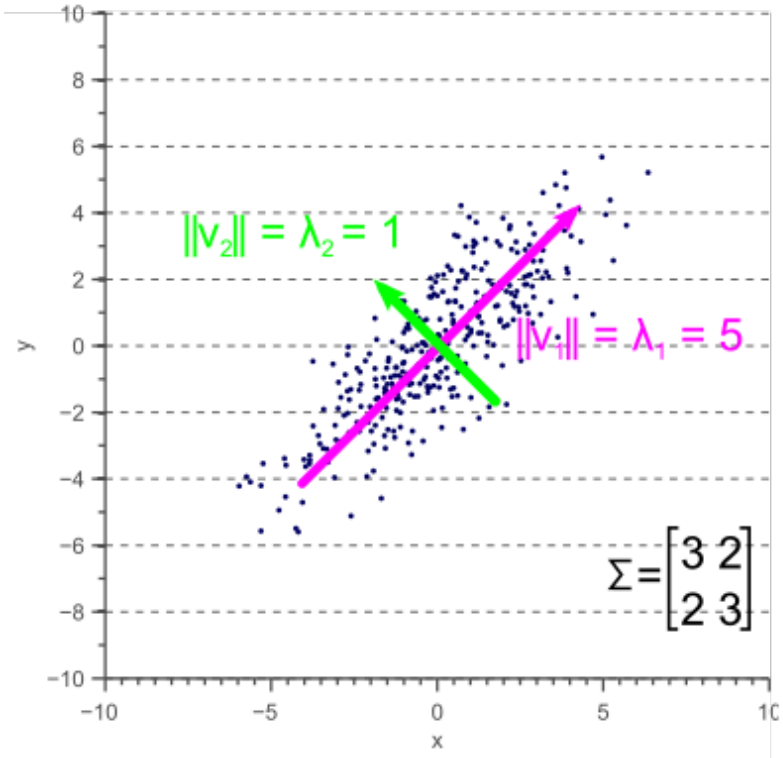
Uncorrelated 2d Gaussian

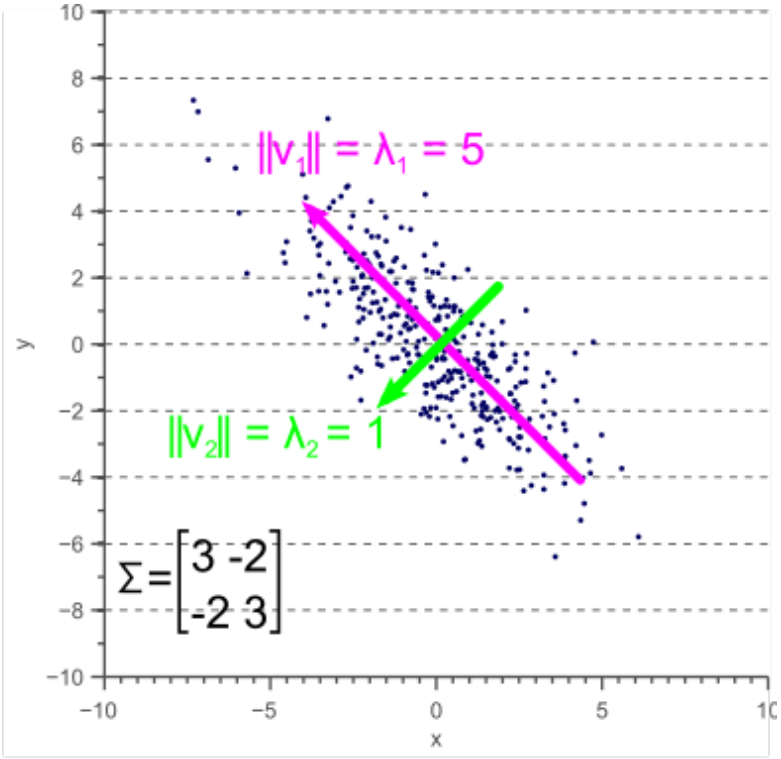




https://www.visiondummy.com/2014/04/geometric-interpretation-covariance-matrix/

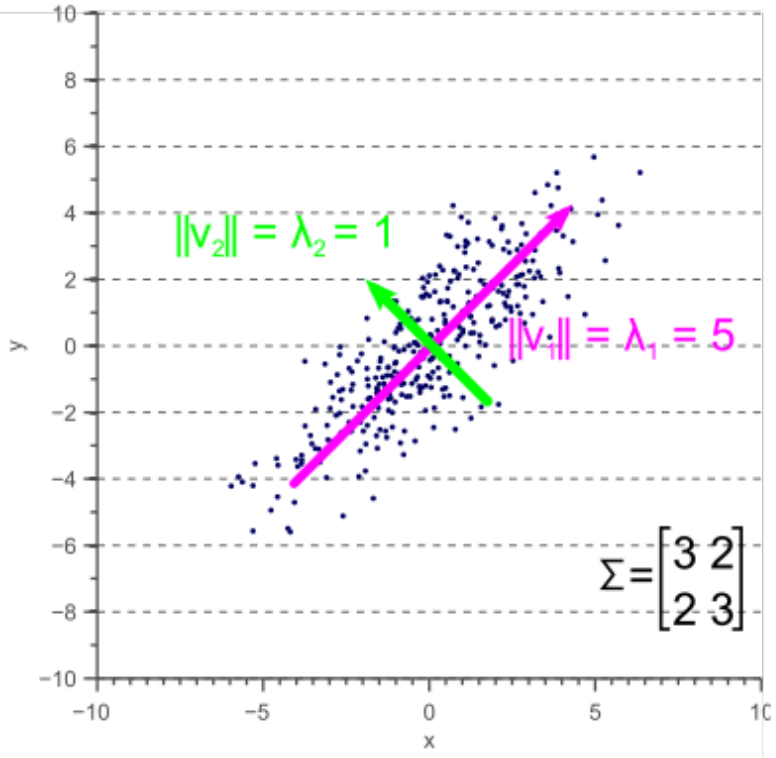
Correlated 2d Gaussian





https://www.visiondummy.com/2014/04/geometric-interpretation-covariance-matrix/

Correlated 2d Gaussian



Eigenvalue decomposition of covariance matrix $\boldsymbol{C}$

$$\boldsymbol{C} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T$$

$$\boldsymbol{V} = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

$$\boldsymbol{\Lambda} = \begin{bmatrix} 5 & \\ & 1 \end{bmatrix}$$

→ Principal Component Analysis (PCA)

https://www.visiondummy.com/2014/04/geometric-interpretation-covariance-matrix/

Bernoulli + multinomial distribution

Bernoulli distribution models binary random variable

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

$$P(X = x) = p^x(1 - p)^{1-x}$$

Multinomial distribution ($N = 1$) models categorical random variable

$$P(X = \boldsymbol{x}) = \prod_k p_k^{x_k} \quad \text{where } \boldsymbol{p} \in [0,1]^K \quad \text{and } p_K = 1 - \sum_{k=1}^{K} p_k$$

Often used in classification problems

Information theory

Entropy

$$\mathrm{H}[x] = -\sum_{x} p(x) \log_2 p(x)$$

Important quantity in

- coding theory
- statistical physics
- machine learning

Entropy

Coding theory: discrete x with 8 possible states; how many bits to transmit the state of x?

All states equally likely

$$\mathrm{H}[x] = -8 \times \frac{1}{8} \log_2 \frac{1}{8} = 3 \text{ bits.}$$

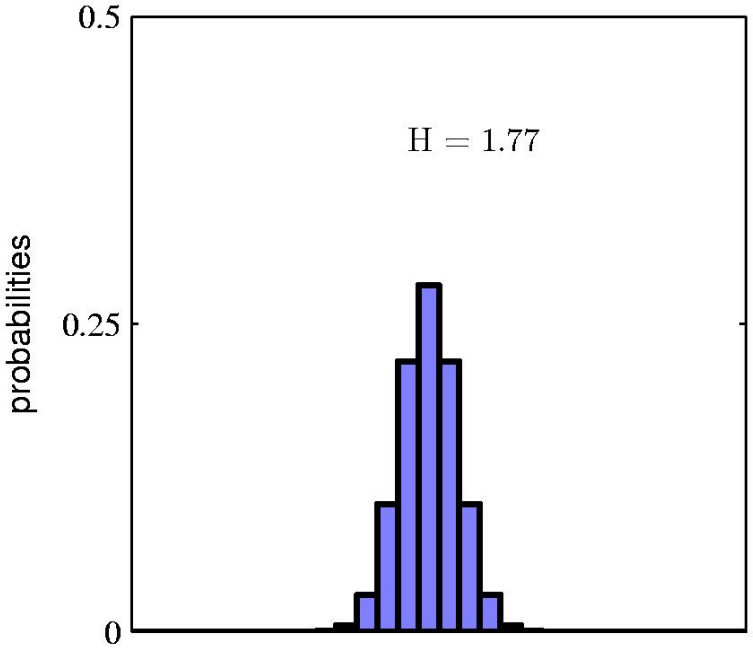
Entropy

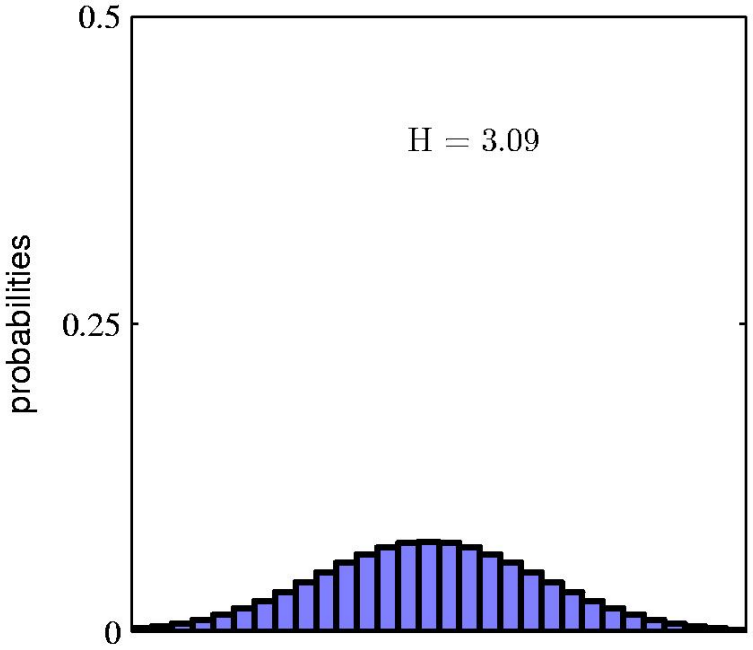
x	a	b	c	d	e	f	g	h
$p(x)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{64}$	$\frac{1}{64}$	$\frac{1}{64}$	$\frac{1}{64}$
code	0	10	110	1110	111100	111101	111110	111111

$$\begin{aligned} \mathrm{H}[x] &= -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{8}\log_2\frac{1}{8} - \frac{1}{16}\log_2\frac{1}{16} - \frac{4}{64}\log_2\frac{1}{64} \\ &= 2 \text{ bits} \end{aligned}$$

$$\begin{aligned} \text{average code length} &= \frac{1}{2}\times 1 + \frac{1}{4}\times 2 + \frac{1}{8}\times 3 + \frac{1}{16}\times 4 + 4\times\frac{1}{64}\times 6 \\ &= 2 \text{ bits} \end{aligned}$$

Entropy

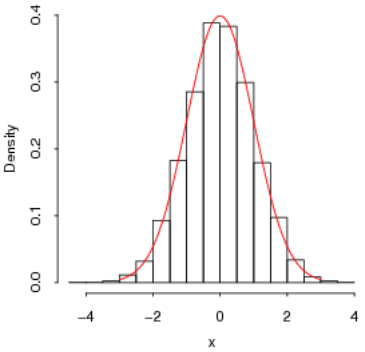




Differential Entropy

Put bins of width Δ along the real line

$$\lim_{\Delta \to 0} \left\{ -\sum_i p(x_i) \Delta \ln p(x_i) \right\} = -\int p(x) \ln p(x)$$



Differential entropy maximized (for fixed σ^2) when

$$p(x) = \mathcal{N}(x|\mu, \sigma^2)$$

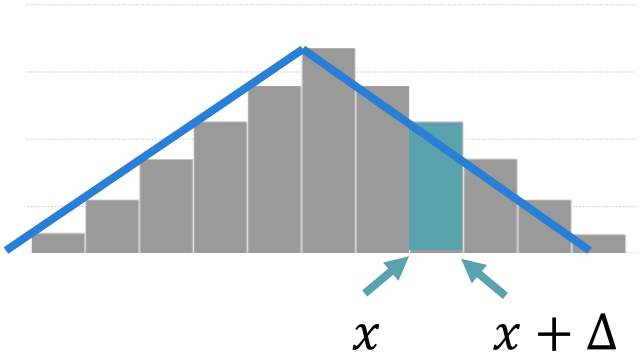
in which case $\mathrm{H}[x] = \frac{1}{2}\left\{1 + \ln(2\pi\sigma^2)\right\}.$

Differential Entropy

Put bins of width Δ along the real line

$$H_\Delta = -\sum_i p(x_i)\Delta \ln(p(x_i)\Delta)$$

$$= -\sum_i p(x_i)\Delta \ln p(x_i) - \ln \Delta$$



$$H[x] = \lim_{\Delta \to 0} \left\{ -\sum_i p(x_i)\Delta \ln p(x_i) \right\} = -\int p(x) \ln p(x)\, dx$$

Conditional Entropy

Suppose we have a joint distribution p(x, y) from which we draw pairs of values of x and y

If a value of x is already known, then the additional information needed to specify the corresponding value of y is given by −ln p(y|x).

$$\mathrm{H}[\mathbf{y}|\mathbf{x}] = -\iint p(\mathbf{y},\mathbf{x}) \ln p(\mathbf{y}|\mathbf{x})\, \mathrm{d}\mathbf{y}\, \mathrm{d}\mathbf{x}$$

$$\mathrm{H}[\mathbf{x},\mathbf{y}] = \mathrm{H}[\mathbf{y}|\mathbf{x}] + \mathrm{H}[\mathbf{x}]$$

The Kullback-Leibler Divergence

Suppose we have modeled some unknown distribution p(x) using an approximating distribution q(x)

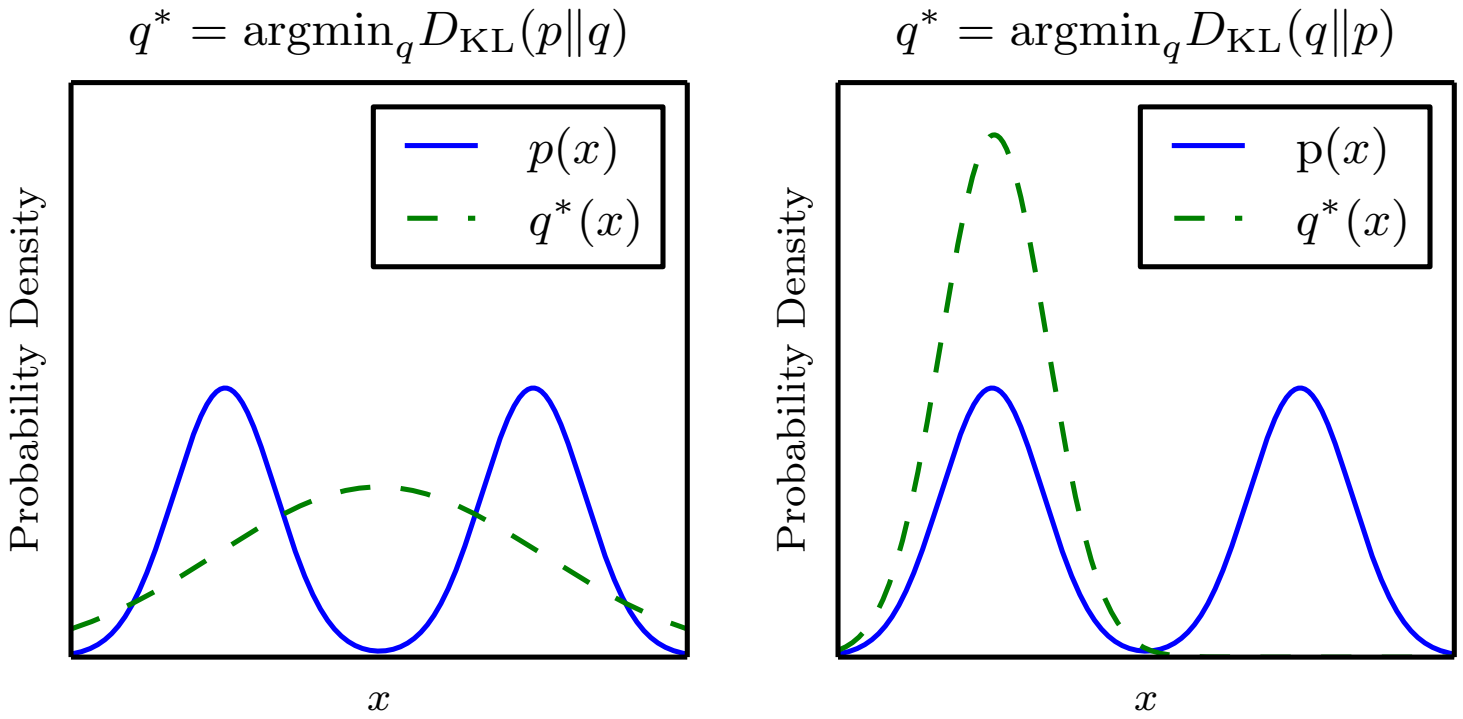
The KL divergence quantifies additional amount of information required to specify the value of x as a result of using q(x) instead of p(x)

$$\begin{aligned}\mathrm{KL}(p\|q) &= -\int p(\mathbf{x})\ln q(\mathbf{x})\,\mathrm{d}\mathbf{x} - \left(-\int p(\mathbf{x})\ln p(\mathbf{x})\,\mathrm{d}\mathbf{x}\right) \\ &= -\int p(\mathbf{x})\ln\left\{\frac{q(\mathbf{x})}{p(\mathbf{x})}\right\}\mathrm{d}\mathbf{x}\end{aligned}$$

$$\mathrm{KL}(p\|q) \geqslant 0$$

$$\mathrm{KL}(p\|q) \not\equiv \mathrm{KL}(q\|p)$$

KL divergence is asymmetric!



$$\text{KL}(p\|q) \neq \text{KL}(q\|p)$$

Mutual Information

Quantifies how much information one variable carries about another

$$\begin{aligned} \mathrm{I}[\mathbf{x},\mathbf{y}] &\equiv \mathrm{KL}(p(\mathbf{x},\mathbf{y})\|p(\mathbf{x})p(\mathbf{y})) \\ &= -\iint p(\mathbf{x},\mathbf{y}) \ln\left(\frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x},\mathbf{y})}\right) \mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} \end{aligned}$$

Symmetry: $\mathrm{I}[\mathbf{x},\mathbf{y}] = \mathrm{H}[\mathbf{x}] - \mathrm{H}[\mathbf{x}|\mathbf{y}] = \mathrm{H}[\mathbf{y}] - \mathrm{H}[\mathbf{y}|\mathbf{x}]$

Independent random variables: $p(x,y) = p(x)p(y) \;\Rightarrow\; I(x,y) = 0$

Questions?