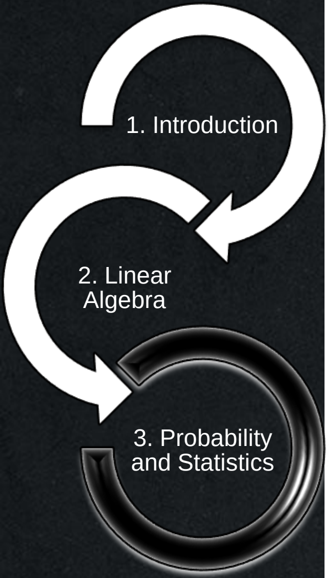
Scientific Machine and Deep Learning for Design and Construction in Civil Engineering

@ ETH Zürich 2021

Agenda



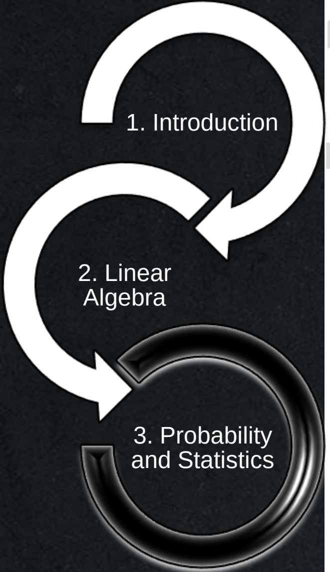
1 **Introduction**

2 **Linear Algebra**

- Scalars, Vectors, Matrices
- Matrix Operations
- Norms, Determinant
- Eigenvalues and Eigenvectors

3 **Probability and Statistics**

- Random Variables
- Expectation operator
- Bayes
- Important distributions
- Monte Carlo



01. Introduction
Lectures

1

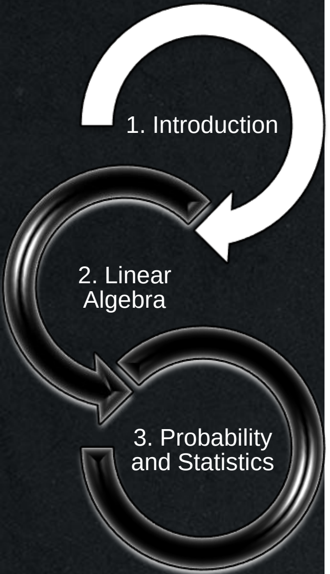
Dr. Michael A. Kraus
- PhD with honors 2019 @ Bundeswehr University Munich
- Post-Doc @ Stanford University
- Post-Doc @ ETH Zürich

2

Dr. Danielle Griego
- PhD 2020 @ ETH Zürich
- Post-Doc @ ETH Zürich
- Managing Director of Design++

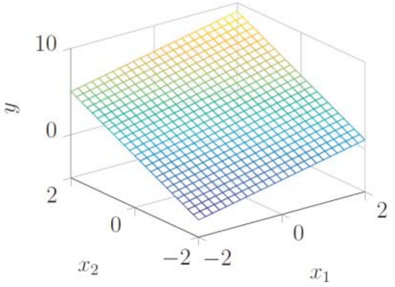
3

Sophia Kuhn, M.Sc.
- M.Sc. Civil Engineering 2021 @ ETH Zürich
- PhD candidate @ ETH Zürich



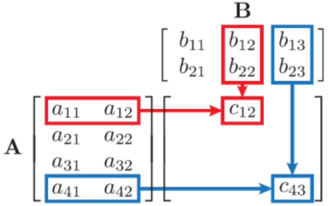
01. Introduction
Content of this Lecture

Linear Algebra



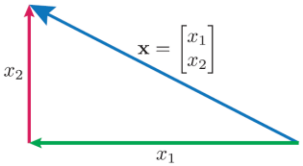
$$\begin{aligned} y &= x_1 + 2x_2 + 3 \\ &= \mathbf{a}^\intercal \mathbf{x} + b \end{aligned}$$

$$\mathbf{a} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad b = 3$$



$$\boxed{c_{12}} = a_{11}b_{12} + a_{12}b_{22}$$

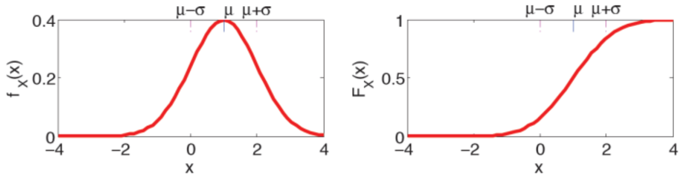
$$\boxed{c_{43}} = a_{41}b_{13} + a_{42}b_{23}$$



$$\begin{aligned} ||\mathbf{x}||_2 &= \sqrt{x_1^2 + x_2^2} \\ ||\mathbf{x}||_1 &= x_1 + x_2 \\ ||\mathbf{x}||_\infty &= \max|x_1, x_2| = x_1 \end{aligned}$$

Probability / Statistics

$$\Pr(E_1|A) = \frac{\Pr(A|E_1)\Pr(E_1)}{\Pr(A)}$$



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

Moment of order m: $\mathbb{E}[X^m]$

$$\mathbb{E}[X^m] = \int_{-\infty}^{\infty} x^m f_X(x) dx$$

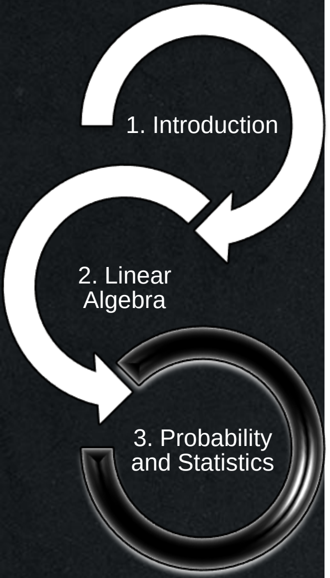
02. Linear Algebra
Basic Definitions

Notation

we assume, that there is at least some basic linear algebra background prior to this lecture

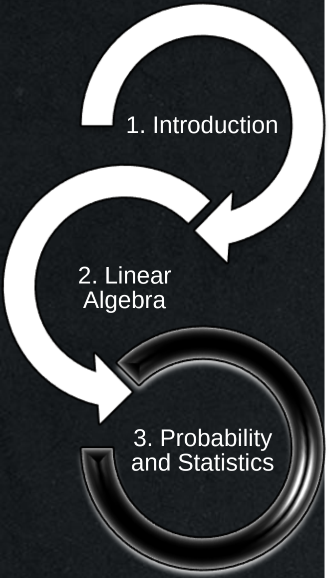
Relevant Elements from Linear Algebra

- scalars
- vectors
- matrices
- tensors



02. Linear Algebra

Basic Definitions



Notation

Scalar → single number

$$x \in \mathbb{R} = (-\infty; \infty)$$

$$\in \mathbb{R}^{+} = (0; \infty)$$

$$\in \mathbb{Z} = (-\infty; \ldots - 1, 0; 1, .., \infty)$$

$$\in (0; 1)$$

$$\in (0; 1]$$

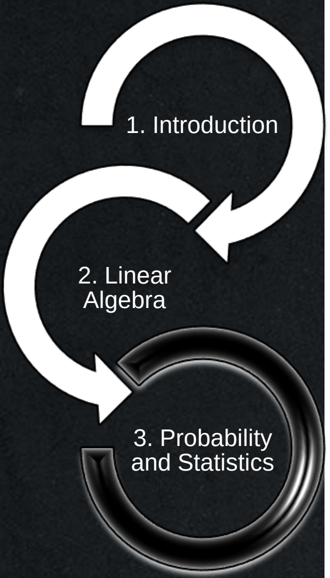
02. Linear Algebra
Basic Definitions

Notation

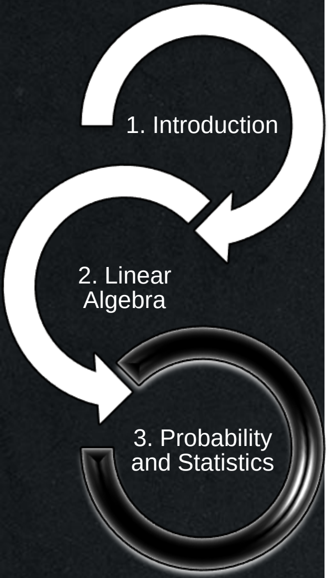
Vector → 1-D array containing numbers or scalars

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

If e.g. each $[\boldsymbol{x}]_i \in \mathbb{R}, \forall i = \{1{:}n\} \rightarrow \boldsymbol{x} \in \mathbb{R}^n$



02. Linear Algebra
Basic Definitions

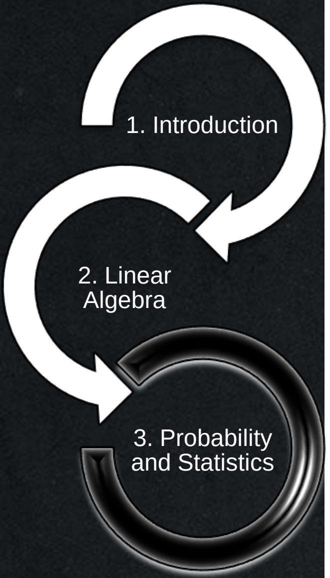


Notation

Matrix → 2-D array containing numbers or scalars

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & & x_{1n} \\ \vdots & \vdots & \dots & \vdots \\ x_{m1} & x_{m2} & & x_{mn} \end{bmatrix}$$

If e.g. each $[\boldsymbol{X}]_{ij} \in \mathbb{R}, \forall i = \{1{:}\,m\}, j = \{1{:}\,n\}, \rightarrow \boldsymbol{X} \in \mathbb{R}^{m \times n}$



02. Linear Algebra

Basic Definitions

Notation

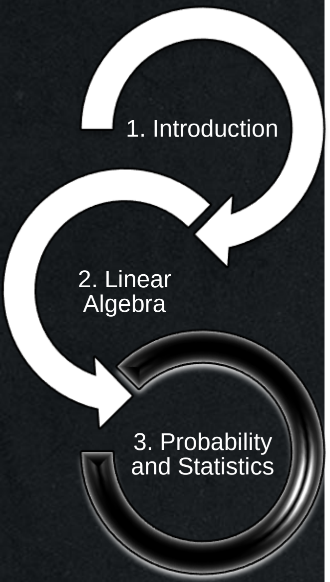
Matrix → 2-D array containing numbers or scalars

Special types:

- *square matrices:* $\boldsymbol{X} \in \mathbb{R}^{n \times n}$

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nn} \end{bmatrix}$$

02. Linear Algebra
Basic Definitions



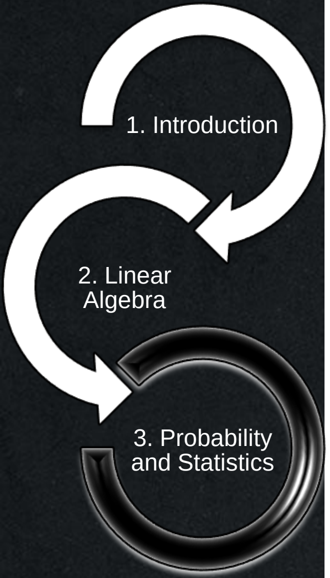
Notation

Matrix → 2-D array containing numbers or scalars

Special types:

- *square matrices:* $\boldsymbol{X} \in \mathbb{R}^{n \times n}$
- *diagonal matrices:* special square matrices with elements just on diagonal: $\boldsymbol{Y} = \mathbf{diag}(\boldsymbol{x})$

$$\mathbf{Y} = \mathrm{diag}(\mathbf{x}) = \begin{bmatrix} x_1 & 0 & \cdots & 0 \\ 0 & x_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & x_n \end{bmatrix}_{n \times n}$$



02. Linear Algebra

Basic Definitions

Notation

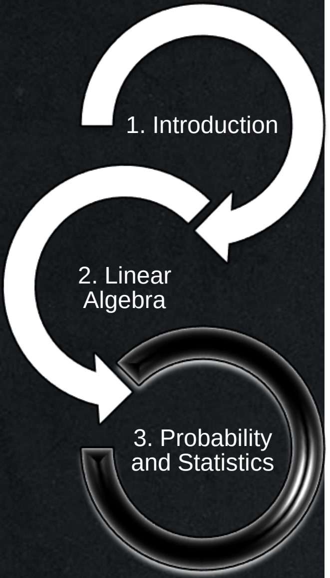
Matrix → 2-D array containing numbers or scalars

Special types:

- *square matrices:* $\boldsymbol{X} \in \mathbb{R}^{n \times n}$
- *diagonal matrices:* special square matrices with elements just on diagonal: $\boldsymbol{Y} = \mathbf{diag}(\boldsymbol{x})$
- Identity matrix: special diagonal matrix with just 1 on diagonal

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}_{n \times n}$$

02. Linear Algebra
Basic Definitions



Notation

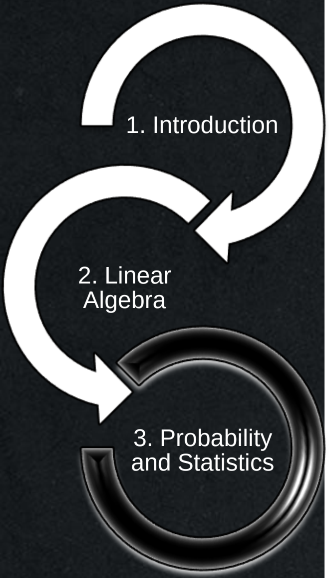
Matrix $\rightarrow$ 2-D array containing numbers or scalars

Special types:

- *square matrices:* $\boldsymbol{X} \in \mathbb{R}^{n \times n}$
- *diagonal matrices:* special square matrices with elements just on diagonal: $\boldsymbol{Y} = \mathbf{diag}(\boldsymbol{x})$
- Identity matrix: special diagonal matrix with just 1 on diagonal
- Block diagonal matrix: concatenates several matrices on diagonal

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \end{bmatrix} \qquad \mathsf{blkdiag}(\mathbf{A}, \mathbf{B}) = \begin{bmatrix} 1 & 2 & 0 & 0 & 0 \\ 3 & 4 & 0 & 0 & 0 \\ 0 & 0 & 4 & 5 & 6 \\ 0 & 0 & 7 & 8 & 9 \\ 0 & 0 & 10 & 11 & 12 \end{bmatrix}$$

02. Linear Algebra
Basic Definitions



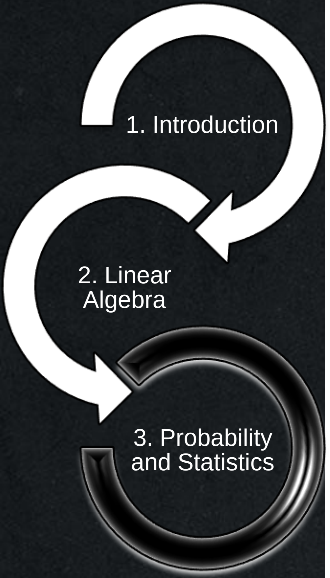
Linear Algebra Operations

- Transposition

- Matrix multiplication

- Matrix inversion

02. Linear Algebra
Basic Definitions



Linear Algebra Operations

- Transposition

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \end{bmatrix} \rightarrow \mathbf{X}^\top = \begin{bmatrix} x_{11} & x_{21} \\ x_{12} & x_{22} \\ x_{13} & x_{23} \end{bmatrix}$$

$$[\mathbf{X}^\top]_{ij} = [\mathbf{X}]_{ji}$$

- Matrix multiplication

- Matrix inversion

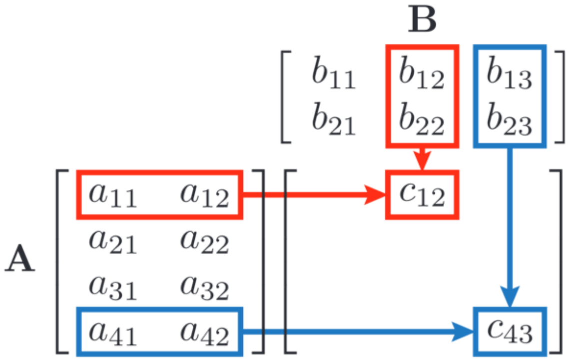
02. Linear Algebra

Basic Definitions

Linear Algebra Operations

- Transposition
- Matrix multiplication
- Matrix inversion

$$
\begin{aligned}
\mathbf{C} &= \mathbf{AB} \\
&= \mathbf{A} \times \mathbf{B} \\
[\mathbf{C}]_{ij} &= \textstyle\sum_k [\mathbf{A}]_{ik} \cdot [\mathbf{B}]_{kj} \\
\mathbf{C}_{m\times n} &= \mathbf{A}_{m\times k}\mathbf{B}_{k\times n}
\end{aligned}
$$



$$\boxed{c_{12}} = a_{11}b_{12} + a_{12}b_{22}$$
$$\boxed{c_{43}} = a_{41}b_{13} + a_{42}b_{23}$$

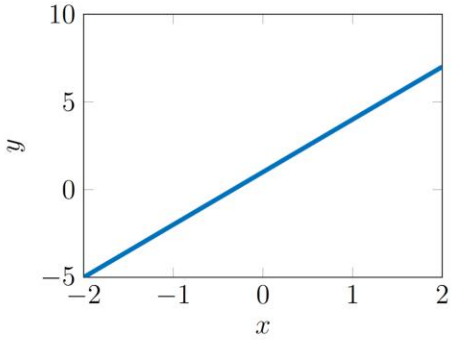
$$
\begin{aligned}
\mathbf{A}(\mathbf{B}+\mathbf{C}) &= \mathbf{AB}+\mathbf{AC} && \text{(Distributivity)} \\
\mathbf{A}(\mathbf{BC}) &= (\mathbf{AB})\mathbf{C} && \text{(Associativity)} \\
\mathbf{AB} &\neq \mathbf{BA} && \text{(Not commutative)} \\
(\mathbf{AB})^\top &= \mathbf{B}^\top\mathbf{A}^\top && \text{(Conjugate transposability)} \\
\mathbf{x}^\top\mathbf{y} &= \mathbf{x}\cdot\mathbf{y} && \text{(Inner product)}
\end{aligned}
$$

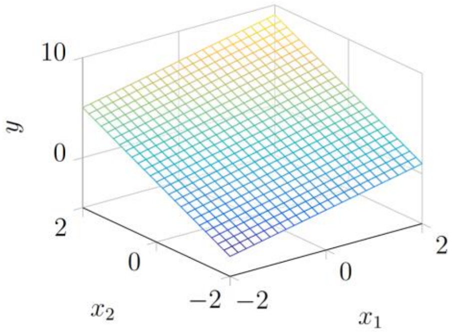
02. Linear Algebra
Basic Definitions

Linear Algebra Operations

- Transposition
- Matrix multiplication
- Matrix inversion

Example: Linear Systems





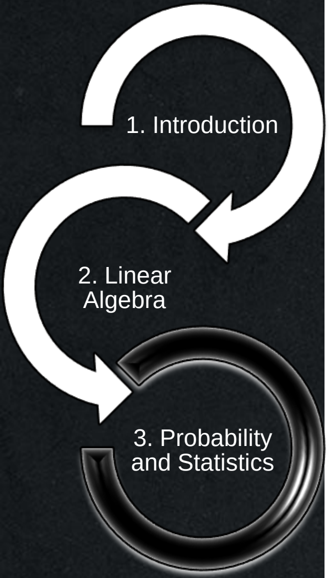
$$\begin{aligned} y &= 3x + 1 \\ &= ax + b \end{aligned}$$

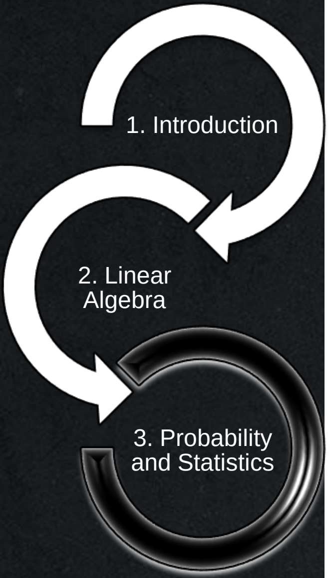
$$a = 3, \quad b = 1$$

$$\begin{aligned} y &= x_1 + 2x_2 + 3 \\ &= \mathbf{a}^\intercal \mathbf{x} + b \end{aligned}$$

$$\mathbf{a} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad b = 3$$

from [2]





02. Linear Algebra
Basic Definitions

Linear Algebra Operations

- Transposition
- Matrix multiplication
- Matrix inversion

Hadamar product
element-wise product

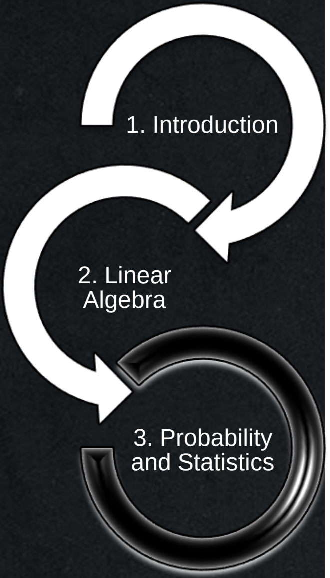
$$\mathbf{C} = \mathbf{A} \odot \mathbf{B}$$

$$[\mathbf{C}]_{ij} = [\mathbf{A}]_{ij} \cdot [\mathbf{B}]_{ij}$$

$$\mathbf{C}_{m \times n} = \mathbf{A}_{m \times n} \odot \mathbf{B}_{m \times n}$$

02. Linear Algebra

Basic Definitions



Linear Algebra Operations

- Transposition
- Matrix multiplication
- Matrix inversion

To be invertible, a matrix:

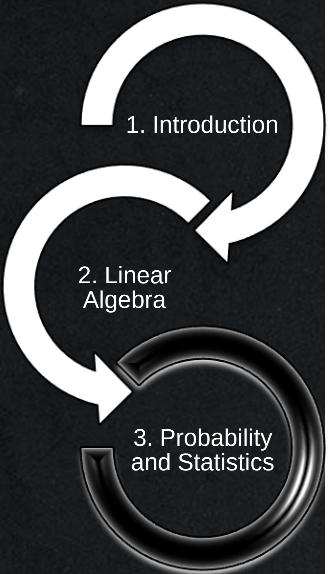
- must be square
- must not have linearly dependent rows or columns

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

Systems of linear equations

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$$
$$\boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b}$$
$$\boldsymbol{I}\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b}$$
$$\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b}$$

02. Linear Algebra
Basic Definitions



Norms of Vectors and Matrices

Measure how large a vector / matrix is.

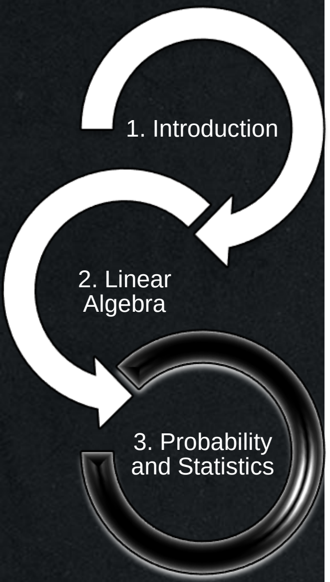
$\boldsymbol{L^p}$ norm is defined as

$$||\mathbf{x}||_p = \left(\sum_i |[\mathbf{x}]_i|^p\right)^{\frac{1}{p}} \qquad (L^p\text{-norm})$$

$$||\mathbf{x}||_2 = \sqrt{\sum_i [\mathbf{x}]_i^2} \equiv \sqrt{\mathbf{x}^\top \mathbf{x}} \qquad (\text{Euclidian norm})$$

$$||\mathbf{x}||_1 = \sum_i |[\mathbf{x}]_i| \qquad (\text{Manhattan norm})$$

$$||\mathbf{x}||_\infty = \max_i |[\mathbf{x}]_i| \qquad (\text{Max norm})$$

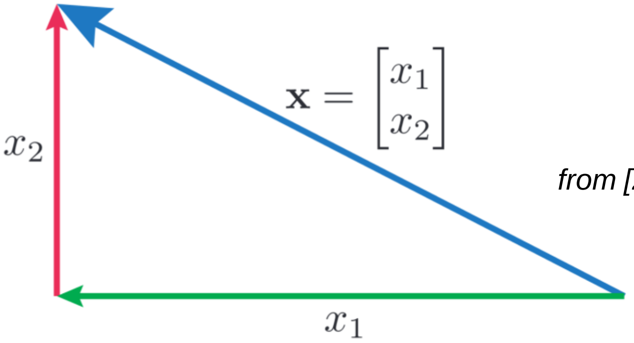


02. Linear Algebra

Basic Definitions

Norms of Vectors and Matrices

L^2 norm is the most common norm



from [2]

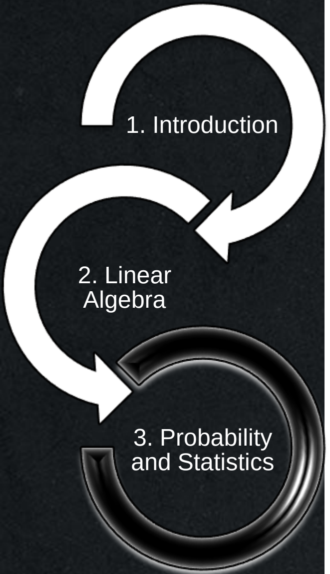
$$
\begin{aligned}
||\mathbf{x}||_2 &= \sqrt{x_1^2 + x_2^2} \\
||\mathbf{x}||_1 &= |x_1| + |x_2| \\
||\mathbf{x}||_\infty &= \max|x_1, x_2| = x_1
\end{aligned}
$$

Task (2 minutes)

draw lines for $||\boldsymbol{x}||_p = 1$

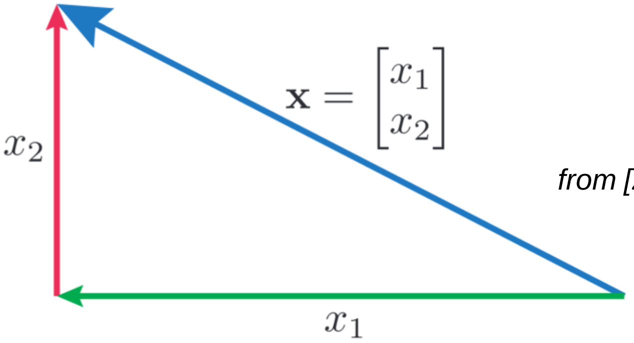
for p = 0.5; 1; 2; 20; ∞

02. Linear Algebra
Basic Definitions



Norms of Vectors and Matrices

L^2 norm is the most common norm



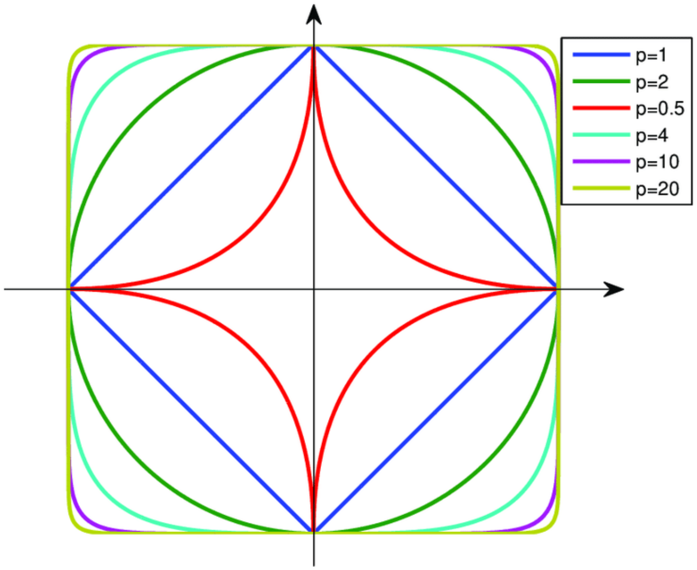
from [2]

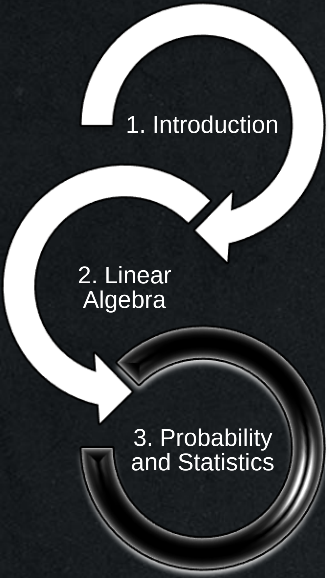
$$||\mathbf{x}||_2 = \sqrt{x_1^2 + x_2^2}$$
$$||\mathbf{x}||_1 = |x_1| + |x_2|$$
$$||\mathbf{x}||_\infty = \max |x_1, x_2| = x_1$$

Task (2 minutes)

draw lines for $||\boldsymbol{x}||_p = 1$

for p = 0.5; 1; 2; 20; ∞





02. Linear Algebra
Basic Definitions

Norms of Vectors and Matrices

Determinant $\mathbf{det}(\boldsymbol{A})$

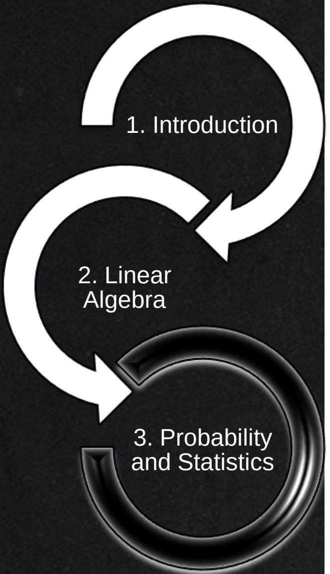
For a square matrix $[\boldsymbol{A}]_{n\times n}, \det(\boldsymbol{A}) \to \mathbb{R}$

The determinant measures how much a matrix contracts or expands in space

$\det(\boldsymbol{A}) = 1$: preservers space / volume

$\det(\boldsymbol{A}) = 0$: contracts the space / volume along 1-D

The determinant is the product of the eigenvalues of a matrix



02. Linear Algebra
Basic Definitions

Norms of Vectors and Matrices

Eigen Decomposition

A square matrix $[\boldsymbol{A}]_{n\times n}$ can be decomposed in eigenvectors $\{\boldsymbol{v_1}, \dots, \boldsymbol{v_n}\}$ and eigenvalues $\{\lambda_1, \dots, \lambda_n\}$. This can be written in matrix form:

$$\boldsymbol{V} = [\boldsymbol{v_1}, \dots, \boldsymbol{v_n}] \quad , \qquad \boldsymbol{\lambda} = [\lambda_1, \dots, \lambda_n]^{\boldsymbol{T}}$$

$$\boldsymbol{A} = \boldsymbol{V}\,\mathrm{diag}(\boldsymbol{\lambda})\boldsymbol{V}^{-\boldsymbol{1}}$$

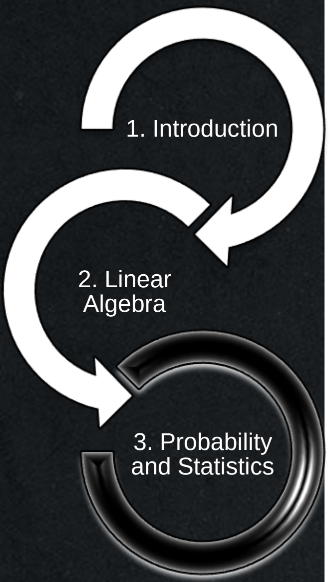
A matrix is positive definite if all eigenvalues > 0

A matrix is positive semi-definite if all eigenvalues ≥ 0

For positive semidefinite matrices: $\forall \boldsymbol{x}, \boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{A}\boldsymbol{x} \geq 0$

02. Linear Algebra

Summary



Matrix transposition

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \end{bmatrix} \rightarrow \mathbf{X}^\mathsf{T} = \begin{bmatrix} x_{11} & x_{21} \\ x_{12} & x_{22} \\ x_{13} & x_{23} \end{bmatrix}$$

Matrix multiplication

$$\begin{aligned} \mathbf{C} &= \mathbf{AB} \\ &= \mathbf{A} \times \mathbf{B} \\ [\mathbf{C}]_{ij} &= \textstyle\sum_k [\mathbf{A}]_{ik} \cdot [\mathbf{B}]_{kj} \\ \mathbf{C}_{m\times n} &= \mathbf{A}_{m\times k}\mathbf{B}_{k\times n} \end{aligned}$$

Element-wise product

$$\begin{aligned} \mathbf{C} &= \mathbf{A} \odot \mathbf{B} \\ [\mathbf{C}]_{ij} &= [\mathbf{A}]_{ij} \cdot [\mathbf{B}]_{ij} \\ \mathbf{C}_{m\times n} &= \mathbf{A}_{m\times n} \odot \mathbf{B}_{m\times n} \end{aligned}$$

Matrix inversion

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

Norm of a vector

$$||\mathbf{x}||_2 = \sqrt{\sum_i [\mathbf{x}]_i^2} \equiv \sqrt{\mathbf{x}^\mathsf{T}\mathbf{x}}$$

Determinant

$\det(\mathbf{A}) : \mathbb{R}^{n\times n} \rightarrow \mathbb{R}$ measures how much the matrix **contracts or expands the space**

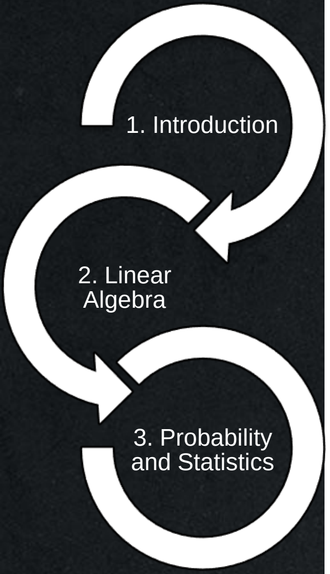
Eigendecomposition

A matrix is **positive semidefinite** if all eigenvalues ≥ 0. For positive semidefinite (PSD) matrices

$$\boxed{\forall \mathbf{x}, \mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} \geq 0}$$

Q & A

Break



03. Probability / Statistics

Basic Definitions of Probability

Probability mass and density function – univariate random variables

we assume, that there is at least some basic understanding of probability prior to this lecture

Random Phenomena; Experiments

- study of random phenomena
- Experiments have different outcomes ξ_i (random variable / RV)
- A set of all outcomes is $\Omega = \{\xi_1, \dots, \xi_K\}$
- An event is a subset of Ω: $E_i \subset \Omega$
- Outcomes have certain underlying patterns about them
- Experiments are conducted under repeatable conditions
- Probability of an event $\quad 1 \geq \Pr(E_i) \geq 0$
- Frequentist / Bayesian definition of probability

03. Probability / Statistics

Basic Definitions of Probability

Probability mass and density function – univariate random variables

Description of discrete RV: Probability Mass Function (PMF)

$$\Pr(X = x) = \Pr(x) = p_X(x) = p(x)$$

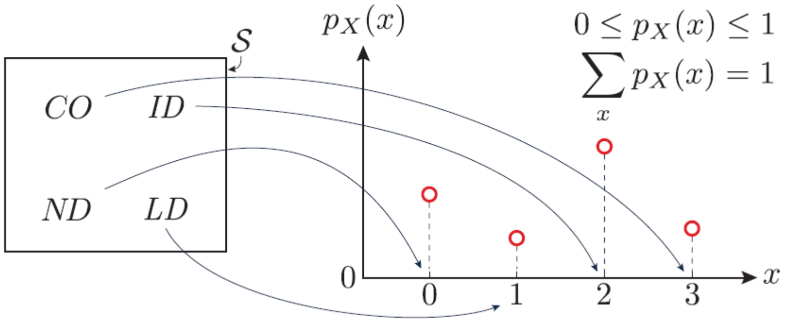
$$\mathcal{S} = \left\{ \begin{array}{r} \text{no damage (ND)} \\ \text{light damage (LD)} \\ \text{important damage (ID)} \\ \text{collapse (CO)} \end{array} \right\}$$

from [2]

Properties

$$0 \leq p_X(x) \leq 1$$

$$\sum_x p_X(x) = 1$$



Light or important damage: $\{1 \leq x \leq 2\}$

03. Probability / Statistics

Basic Definitions of Probability

Probability mass and density function – univariate random variables

Description of continuous RV: Probability Density Function (PDF)

$f_X(x) = f(x)$

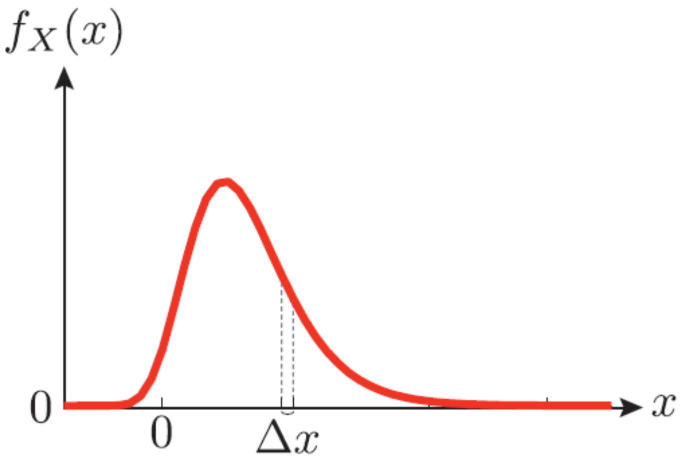
$\Pr(x \leq X \leq x + \Delta x) = f_X(x)\Delta x$

Properties

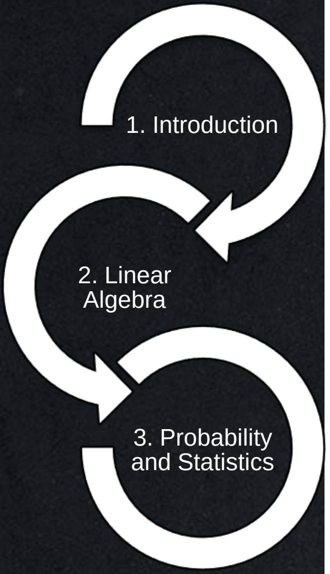
$\Pr(\mathrm{X} = \mathrm{x}) = 0$

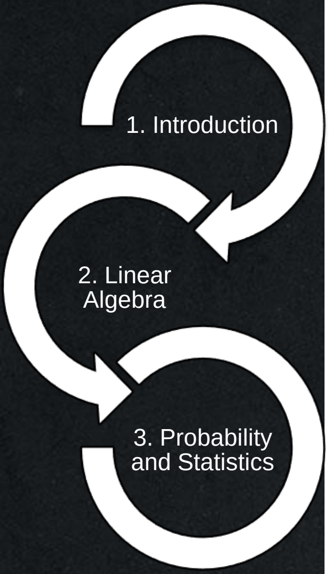
$f_X(x) \geq 0$

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



from [2]





03. Probability / Statistics

Basic Definitions of Probability

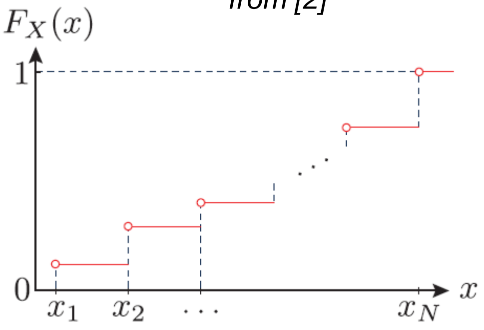
Probability mass and density function – univariate random variables

Description of RV: Cumulative Distribution Function (CDF)

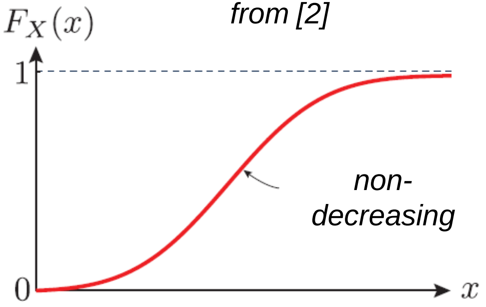
continous $$F_X(x) = \int_{-\infty}^{x'} f_X(x')dx \Leftrightarrow f_X(x) = \frac{dF_X(x)}{dx}$$

discrete $$F_X(x) = \Pr(X \le x) = \sum_{x'<x} p_X(x')$$

from [2]



from [2]



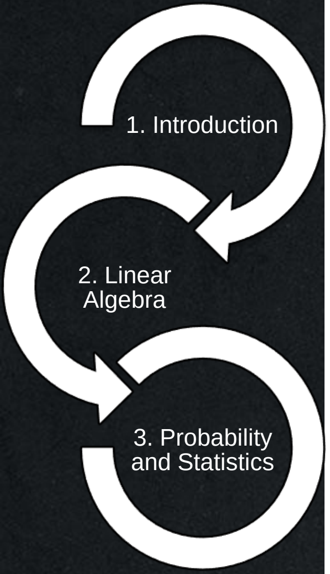
Properties

$$F_X(-\infty) = 0$$

$$F_X(\infty) = 1$$

03. Probability / Statistics

Basic Definitions of Probability



Probability mass and density function – multivariate random variables

Description of multivariate RV:

$$\text{Given } \mathbf{X} = [X_1, X_2, \cdots, X_n]^{\intercal} \left\{ \begin{array}{l} \text{vector (column) of} \\ \text{random variables} \end{array} \right.$$

$$\text{Given } \mathbf{x} = [x_1, x_2, \cdots, x_n]^{\intercal} \left\{ \begin{array}{l} \text{vector describing} \\ \text{the outcomes a} \\ \text{random variable } \mathbf{X} \end{array} \right.$$

X describes the simultaneous realization of several phenomena

03. Probability / Statistics

Basic Definitions of Probability

Probability mass and density function – multivariate random variables

Description of discrete RV:

Definition: $p_X(\boldsymbol{x}) = \Pr(X_1 = x_1 \cap X_2 = x_2 \cap \cdots \cap X_n = x_n)$

$0 \le p_{\boldsymbol{X}}(\boldsymbol{x}) \le 1$

Marginalization

$$\sum_{x_n} p_{X_1 \dots X_n}(x_1, \dots, x_n) = p_{X_1 \dots X_{n-1}}(x_1, \dots, x_{n-1})$$

$p_{X_i}(x_i)$: Marginal PMF

$$\sum_{x_1} \dots \sum_{x_n} p_{\boldsymbol{X}}(\boldsymbol{x}) = 1$$

1. Introduction
2. Linear Algebra
3. Probability and Statistics

03. Probability / Statistics

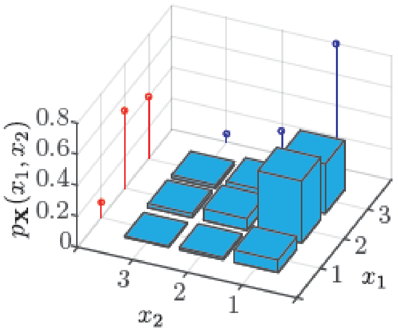
Basic Definitions of Probability

Probability mass and density function – multivariate random variables

Given two discrete R.V. $X_1 \perp X_2$

$$p_{X_1}(x_1) \begin{cases} p_{X_1}(1) = 0.1 \\ p_{X_1}(2) = 0.5 \\ p_{X_1}(3) = 0.4 \end{cases}$$

$$p_{X_2}(x_2) \begin{cases} p_{X_2}(1) = 0.8 \\ p_{X_2}(2) = 0.15 \\ p_{X_2}(3) = 0.05 \end{cases}$$



$$p_{X_1 X_2}(x_1, x_2) = p_{X_1}(x_1) \cdot p_{X_2}(x_2)$$

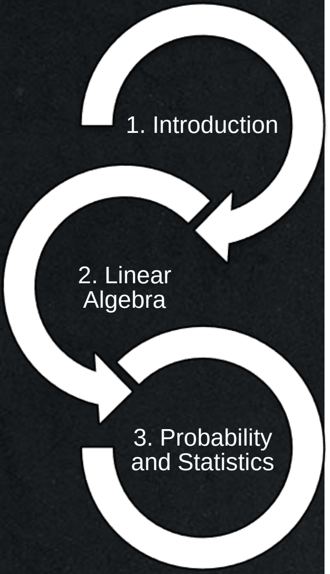
$p_{\mathbf{X}}(x_1, x_2) =$

	$x_2 = 1$	$x_2 = 2$	$x_2 = 3$	$\sum_{i=1}^{n=3} p_{\mathbf{X}}(x_1, i)$
$x_1 = 1$	0.08	0.015	0.005	0.1
$x_1 = 2$	0.4	0.075	0.025	0.5
$x_1 = 3$	0.32	0.06	0.02	0.4

from [2]

03. Probability / Statistics

Basic Definitions of Probability



Probability mass and density function – multivariate random variables

Description of continous RV:

Definition: $f_X(\boldsymbol{x})\Delta\boldsymbol{x} = \Pr(x_1 < X_1 < x_1 + \Delta x_1 \cap \cdots \cap x_1 < X_n < x_n + \Delta x_n)$

$0 \le f_X(\boldsymbol{x})$ *(may be larger than 1!!)*

Marginalization

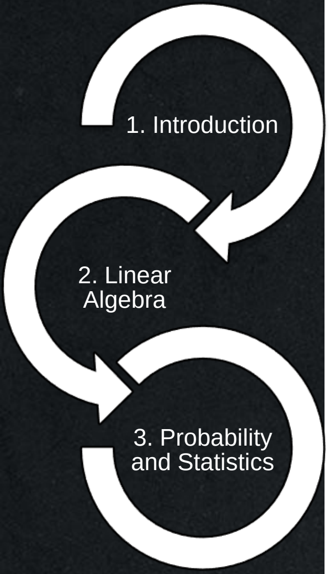
$$\int_{-\infty}^{\infty} f_{X_1 \ldots X_n}(x_1, \ldots, x_n)\, dx_n = f_{X_1 \ldots X_n}(x_1, \ldots, x_n)$$

$$\sum_{x_1} \ldots \sum_{x_n} p_X(\boldsymbol{x}) = 1$$

$p_{X_i}(x_i)$: Marginal PMF

03. Probability / Statistics

Basic Definitions of Probability



Expectation and variance operator

$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \cdot f_x(x)dx$ (continuous R.V.)

$\mathbb{E}[X] = \sum x \cdot p_x(x)dx$ (discrete R.V.)

Moment of order *m:* $\mathbb{E}[X^m]$ $\mathbb{E}[X^m] = \int_{-\infty}^{\infty} x^m f_X(x)dx$

For *m=1:* $\mathbb{E}[X] = \mu_X$ (expectation / center of gravity)

For *m=2:* $\mathbb{E}\left[X^2\right]$ (expectation of the squares)

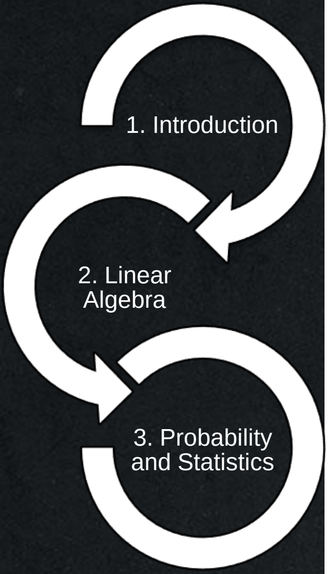
Centered moments of order *m:* $\mathbb{E}[(X-\mu_X)^m]$ $\mathbb{E}[(X-\mu_X)^m] = \int_{-\infty}^{\infty} (X-\mu_X)^m f_x(x)dx$

For *m=1:* $\mathbb{E}\left[(X-\mu_X)^1\right] = 0$

For *m=2:* $\mathbb{E}\left[(X-\mu_X)^2\right] = \sigma_X^2 = \mathrm{var}[X]$ (variance / inertia)

03. Probability / Statistics

Basic Definitions of Probability



Expectation and variance operator

$$Variance - \mathbb{E}\left[(X-\mu_X)^2\right]$$

$$\mathbb{E}\left[(X-\mu_X)^2\right] = \sigma_X^2 = var[X] = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2$$

$\mathrm{var}[X]$: The variance measures the dispersion of the probability density function. Thi s concept is analogue to the inertia of a cross-section.

σ_X: Standard deviation

$\delta_X = \frac{\sigma_X}{\mu_X}$: coefficient of variation (C.O.V.)

adimensional dispersion metric (⚠ $\mu_X \neq 0$)

03. Probability / Statistics

Basic Definitions of Probability

Expectation and variance operator

$Covariance - \mathbb{E}[(X-\mu_X)-(Y-\mu_Y)]$

Given two random variables X, Y

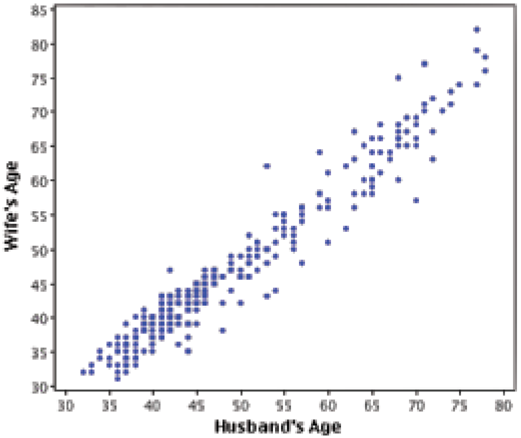
$$cov[XY] = \mathbb{E}[(X-\mu_X)-(Y-\mu_Y)] = \color{red}{\mathbb{E}[XY]-\mathbb{E}[X]\mathbb{E}[Y]}$$

$\color{red}{\rho_{XY}}$: *correlation coefficient*

(quantifies the **linear dependence** between X and Y)

$$\rho_{ij} = \frac{cov[XY]}{\sigma_i\sigma_j}, \quad -1 \le \rho_{XY} \le +1$$

from [2]



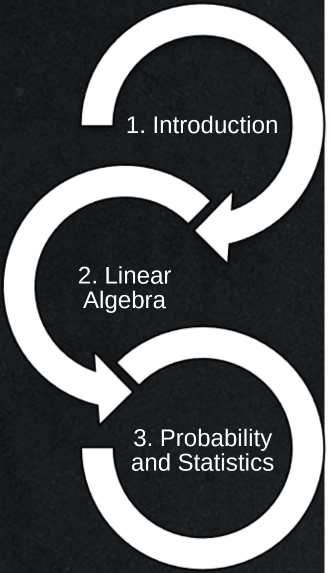
$\rightarrow X \perp\!\!\!\perp Y \Rightarrow \rho_{XY} = 0$

$\rightarrow \rho_{XY} = 0 \nRightarrow X \perp\!\!\!\perp Y$ ⚠

$\rightarrow correlation \nLeftrightarrow causality$

⚠

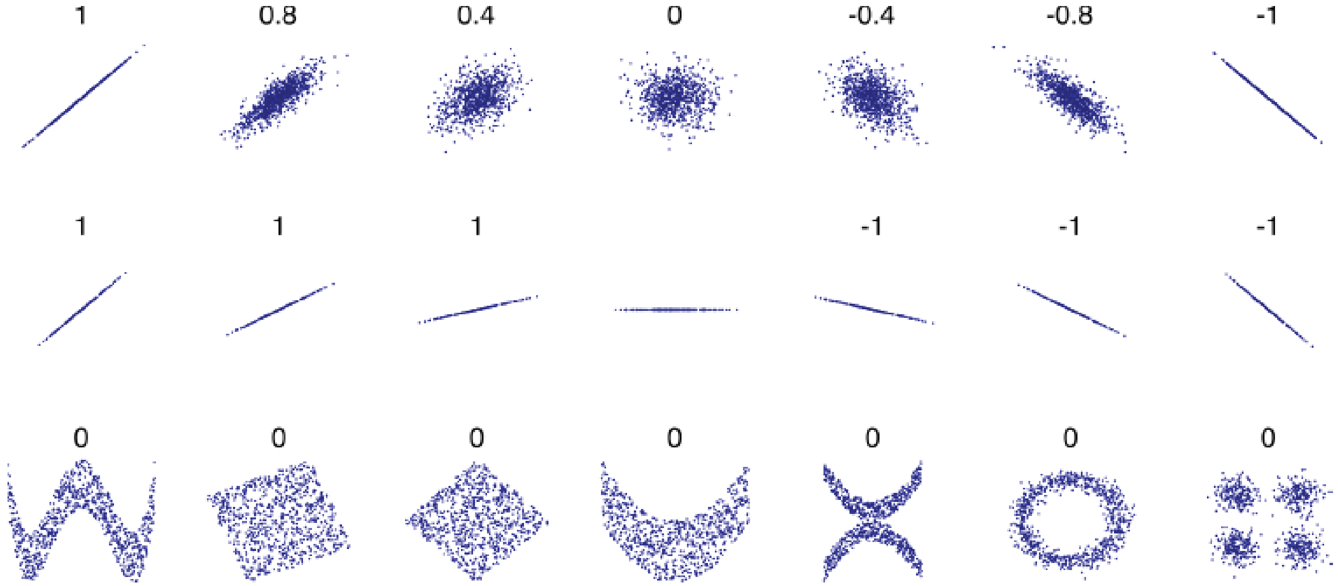
03. Probability / Statistics
Basic Definitions of Probability



Variance / Covariance

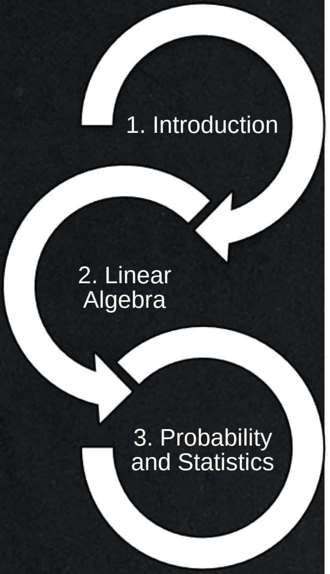
from [2]

- quantify <u>*linear*</u> dependence



03. Probability / Statistics

Basic Definitions of Probability



Distributions

- <u>Parametric Distributions</u>

  basic building block: $p(\boldsymbol{x}|\boldsymbol{\theta})$ **defined by parameters $\boldsymbol{\theta}$**

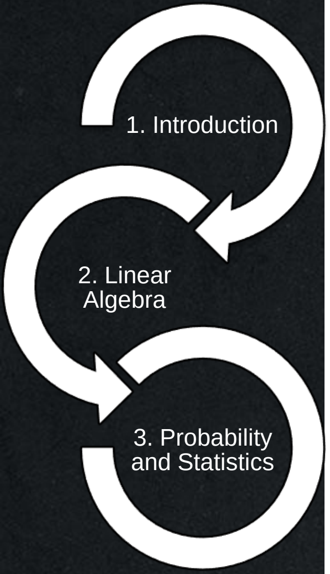
  need to determine θ given a sample $\{\boldsymbol{x_1}, \dots, \boldsymbol{x_N}\}$

- <u>Non-Parametric Distributions</u>

  are not restricted to specific functional forms

  make few assumptions about the shape of the distribution being modelled

03. Probability / Statistics
Important Distributions



Bernoulli Distribution

Coin flipping: heads=1, tails=0

$$p(x = 1|\mu) = \mu$$

Bernoulli Distribution

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

$$\mathbb{E}[x] = \mu$$

$$\mathrm{var}[x] = \mu(1-\mu)$$

03. Probability / Statistics
Important Distributions

Binomial Distribution

N coin flips: what is the probability of seeing *m* heads

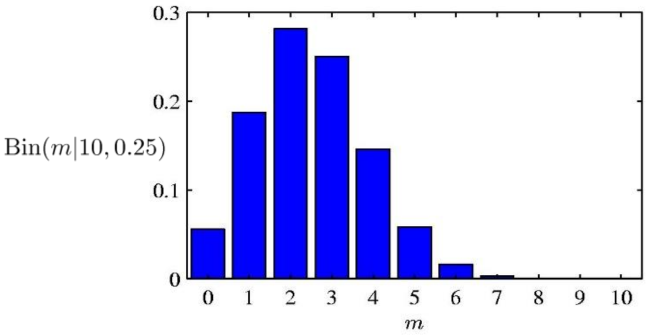
$$p(m\ \text{heads}|N,\mu)$$

Binomial Distribution

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

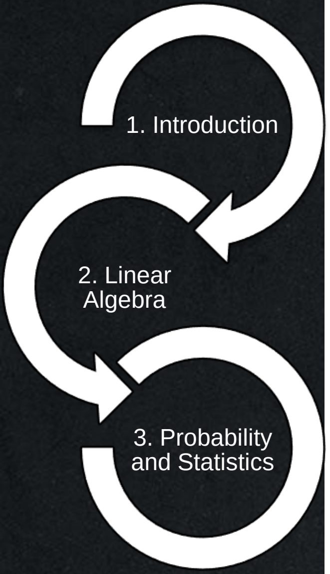
$$\mathbb{E}[m] \equiv \sum_{m=0}^{N} m\text{Bin}(m|N,\mu) = N\mu$$

$$\text{var}[m] \equiv \sum_{m=0}^{N} (m-\mathbb{E}[m])^2\text{Bin}(m|N,\mu) = N\mu(1-\mu)$$



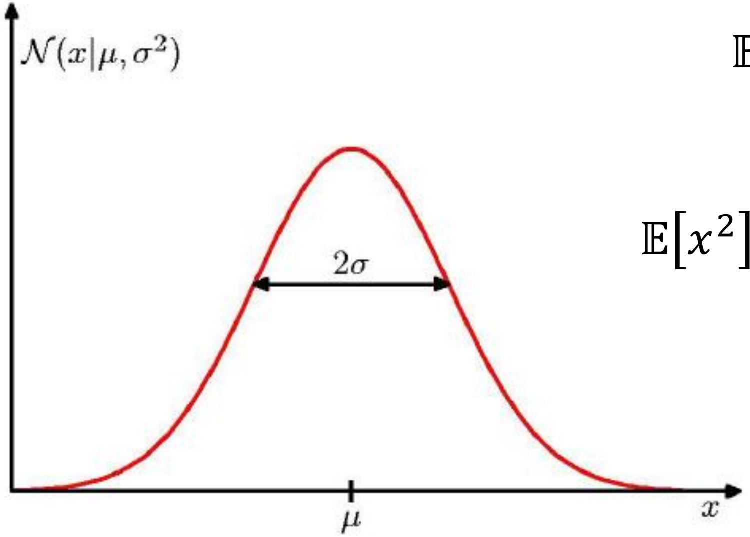
from [3]

03. Probability / Statistics
Important Distributions



Normal or Gaussian Distribution (univariate case)

$$\mathcal{N}\left(x|\mu,\sigma^2\right) = \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$



from [3]

$$\mathbb{E}[x] = \int_{-\infty}^{\infty}\mathcal{N}\left(x|\mu,\sigma^2\right)x\,dx = \mu$$

$$\mathbb{E}\left[x^2\right] = \int_{-\infty}^{\infty}\mathcal{N}\left(x|\mu,\sigma^2\right)x^2\,dx = \mu^2+\sigma^2$$

$$\mathrm{var}[x] = \mathbb{E}\left[x^2\right] - \mathbb{E}[x]^2 = \sigma^2$$

03. Probability / Statistics

Important Distributions

Normal or Gaussian Distribution (multivariate case)

5 Parameters: $\mu_{X_1}, \sigma_{X_1}, \mu_{X_2}, \sigma_{X_2}, \rho$

$$X \sim \mathcal{N}(x; \mu, \Sigma)$$

$$\mu_{X_1} = 0$$

$$\sigma_{X_1} = 2$$

$$\mu_{X_2} = 0$$

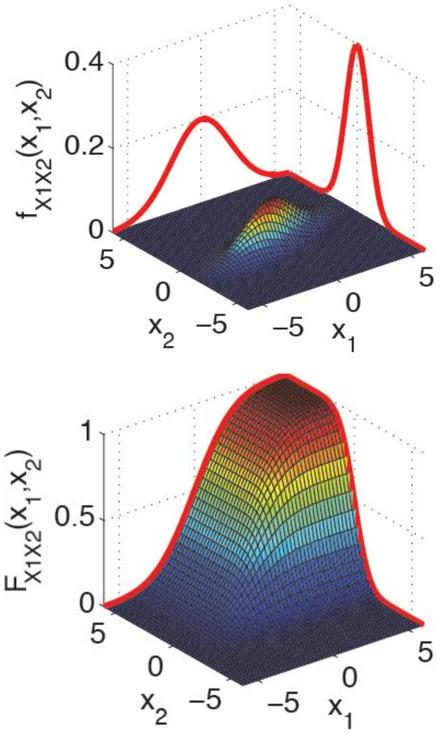
$$\sigma_{X_2} = 1$$

$$\rho = 0.6$$

$$\mu = \begin{bmatrix} \mu_{X_1} \\ \mu_{X_2} \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} \sigma_{X_1}^2 & \rho\sigma_{X_1}\sigma_{X_2} \\ \rho\sigma_{X_1}\sigma_{X_2} & \sigma_{X_2}^2 \end{bmatrix}$$

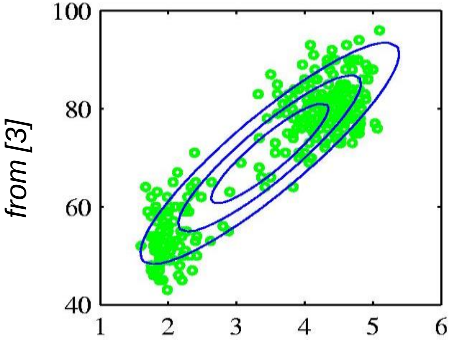
$$F_{X_1,X_2}(X_1, X_2) = \int_{-\infty}^{X_1} \int_{-\infty}^{X_2} f_{X_1,X_2}(X_1, X_2)\,\partial x\,\partial y$$

from [2]

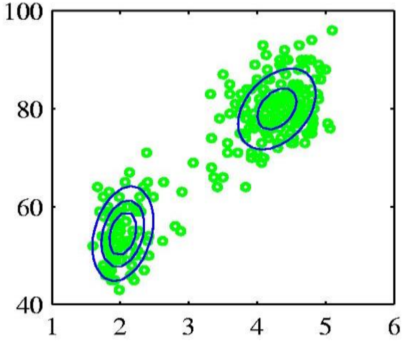


03. Probability / Statistics
Important Distributions

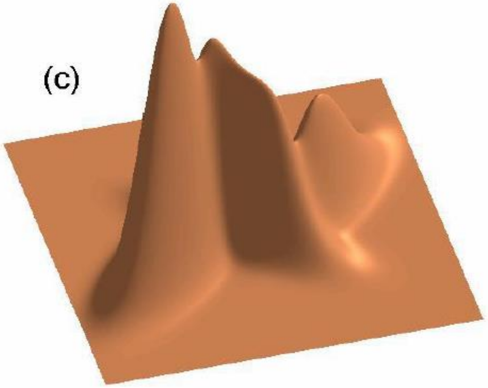
Mixture distributions



Single Gaussian



Mixture of two Gaussians



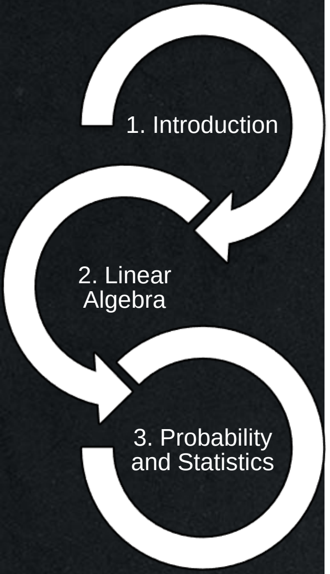
Mixture of three Gaussians

$$P(x) = \sum_i P(c = i)\, P(x|c = i)$$

$$p(x) = \sum_{k=1}^{K} \pi_k \underbrace{\mathcal{N}(x|\mu_k \Sigma_k)}_{\text{Component}}$$

Component

Mixing coefficient



03. Probability / Statistics
Important Distributions

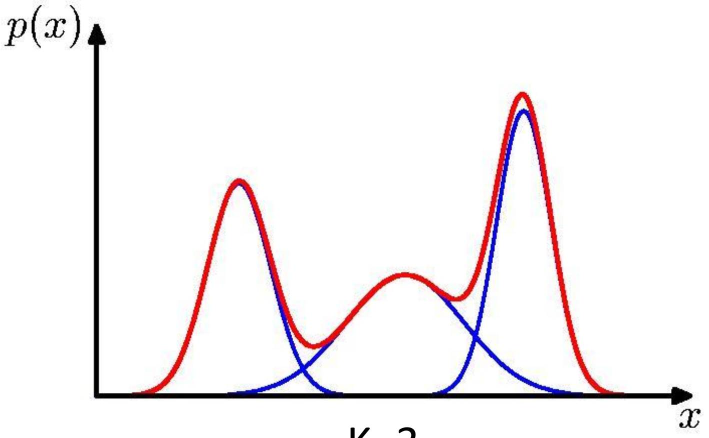
Mixture distributions

Combine simple models into a complex model:

$$p(x) = \sum_{k=1}^{K} \pi_k \underbrace{\mathcal{N}(x|\mu_k \Sigma_k)}_{\text{Component}}$$

Mixing coefficient

$$\forall k: \pi_k \geq 0 \qquad \sum_{k=1}^{K} \pi_k = 1$$

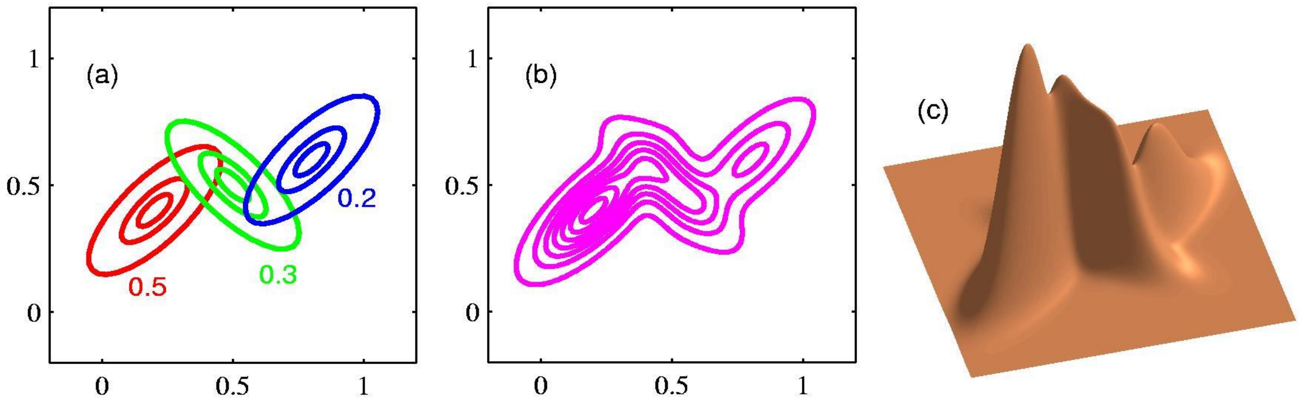


K=3

from [3]

03. Probability / Statistics
Important Distributions

Mixture distributions – Example: Mixutre of 3 Gaussians



from [3]

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k \Sigma_k) \qquad \forall k: \pi_k \geq 0 \qquad \sum_{k=1}^{K} \pi_k = 1$$

03. Probability / Statistics
Important Distributions

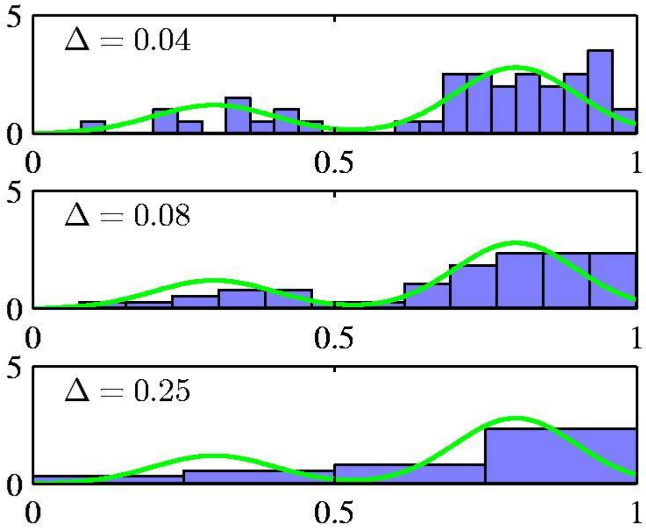
Nonparametric Methods

Histogram methods partition the data space in to distinct bins with widths Δ_i and count the number of observations, n_i, in each bin.

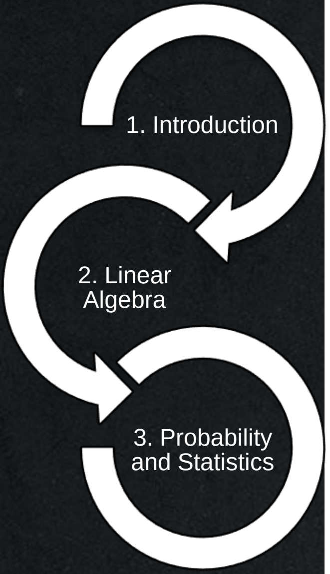
$$p_i = \frac{n_i}{N\Delta_i}$$

- often, the same width is used for all bins, $\Delta_i = \Delta$
- Δ acts as a smoothing parameter.

from [3]



03. Probability / Statistics
Important Distributions



Nonparametric Methods

Kernel Density Estimation: fix a volume V an d estimate the number of data points *K* within a certain region *R* from the data.

Let *R* be a hypercube centred on *x* and define the kernel function (Parzen window)

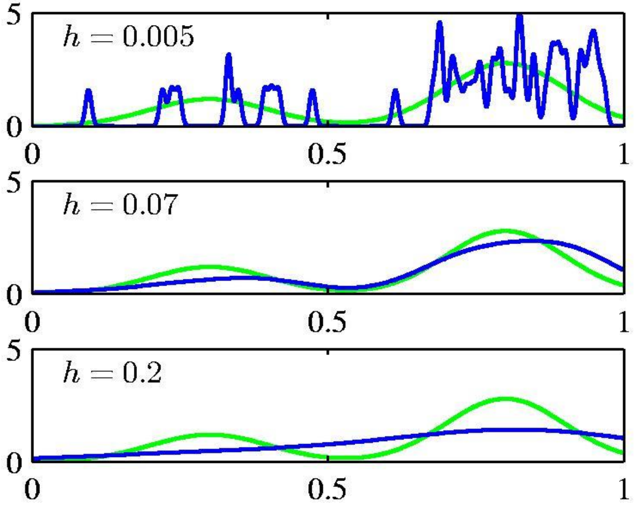
$$k\big((x-x_n)/h\big)=\begin{cases}1, & |(x-x_n)/h|\le 1/2, \quad i=1,\dots,D\\ 0, & \text{otherwise}\end{cases}$$

$$p(x)=\frac{1}{N}\sum_{n=1}^{N}\frac{1}{h^D}\,k\left(\frac{x-x_n}{h}\right)$$

To avoid discontinuities in $p(x)$, use a smooth kernel, e.g. a Gaussian

$$p(x)=\frac{1}{N}\sum_{n=1}^{N}\frac{1}{(2\pi h^2)^{D/2}}\exp\left\{-\frac{\|x-x_n\|^2}{2h^2}\right\}$$

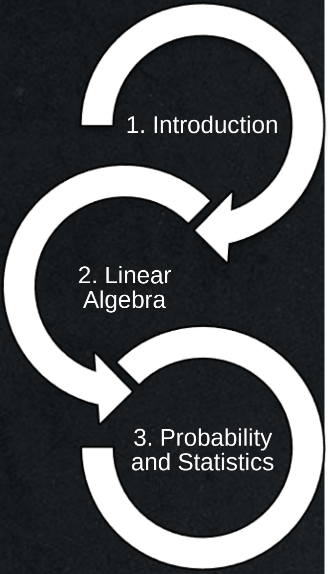
from [3]



h acts as a smoother.

03. Probability / Statistics

Functions of RV



Functions of random variables

Given $\boldsymbol{X} = [X_1, X_2, \dots, X_n]^T$ defined from their joint PDF $f_{\boldsymbol{X}}(\boldsymbol{x})$, and given $\boldsymbol{Y} = [Y_1, Y_2, \dots, Y_m]^T$ obtained from a function:

$$\mathbf{Y} = \boldsymbol{g}(\boldsymbol{X}) = \begin{bmatrix} g_1(\boldsymbol{X}) \\ g_2(\boldsymbol{X}) \\ \vdots \\ g_m(\boldsymbol{X}) \end{bmatrix}$$

3 cases:

1. $m = n = 1$
2. $m = n > 1$
3. $m = 1, n > 1$

Analytical expressions just for linear functions;

Taylor expansion or Monte-Carlo

03. Probability / Statistics

Bayes Theorem

Bayes Theorem

$p(\text{unknown}|\text{known})$

y: Observation

x: Constant

X: Random variable

$\boldsymbol{\theta}$: $X \sim f(x;\boldsymbol{\theta})$

Given $\boldsymbol{X} = [X_1, X_2, \cdots, X_X]^T$ a vector of random variables so that $\boldsymbol{X} \sim f(x)$

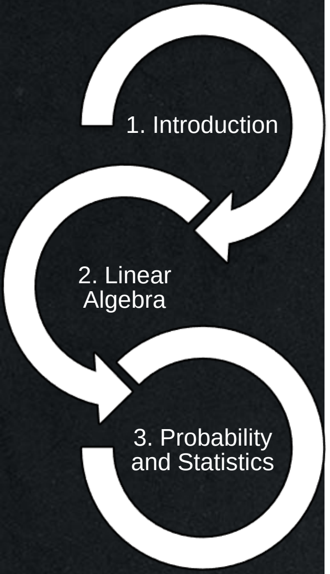
and given $\mathcal{D} = \{y_1, y_2, \cdots, y_D\}$ *a set of observations*

corresponding to realizations of $\boldsymbol{Y} = [Y_1, Y_2, \cdots, Y_D]^T$

so that $\boldsymbol{Y} \sim f(\boldsymbol{y})$

$$f(\boldsymbol{x}|\boldsymbol{y} = \mathcal{D}) = \frac{f(\boldsymbol{y} = \mathcal{D}|\boldsymbol{x}) \cdot f(\boldsymbol{x})}{f(\boldsymbol{y} = \mathcal{D})}$$

$$\underbrace{f(\boldsymbol{x}|\mathcal{D})}_{\text{posterior}} = \frac{\overbrace{f(\mathcal{D}|\boldsymbol{x})}^{\text{likelihod}} \cdot \overbrace{f(\boldsymbol{x})}^{\text{prior}}}{\underbrace{f(\mathcal{D})}_{\text{normalization cte.}}}$$



03. Probability / Statistics

Bayes Theorem

Bayes Theorem – highly topical example

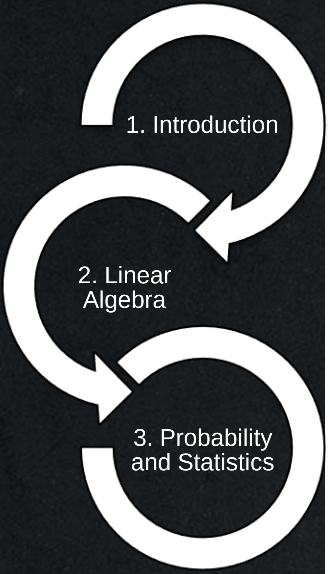
Given a deadly disease so rare that only one human on Earth has it.

We have a screening test so that

$$\text{test}+ \mapsto \begin{cases} \Pr(\text{test} + |\text{desease}) = 0.999 \\ \Pr(\text{test} + |\text{no desease}) = 0.001 \end{cases}$$

If you test positive, should you be worried?

03. Probability / Statistics
Bayes Theorem



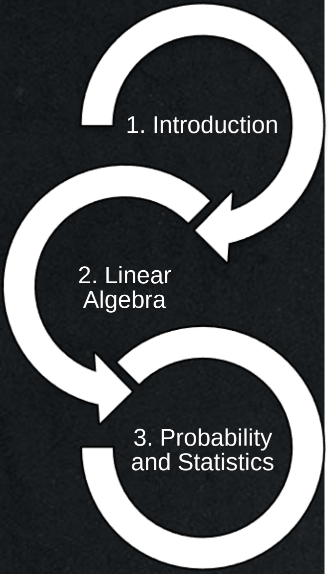
Bayes Theorem – highly topical example

from [2]

expect ≈ 0.001 x 8 x 10^9
= 8 x 10^6 false diagnoses

$$\Pr(\text{desease}|\text{test}+) = \frac{1}{8 \cdot 10^6} \approx 2 \cdot \Pr\left(\text{649}\right)$$

If you want to properly extract information from data, you must consider the prior probability of the phenomenon you are interested in.

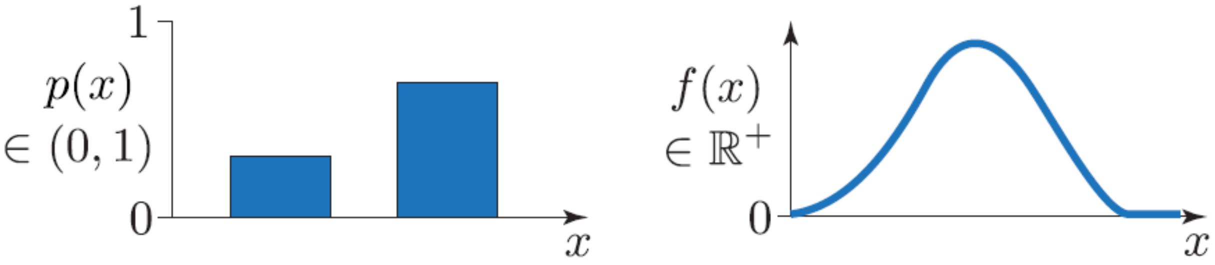


03. Probability / Statistics
Bayes Theorem

Bayes Theorem – Prior knowledge

$f(\boldsymbol{x})$ describes prior knowledge for the values that values $\boldsymbol{x}$ can take.

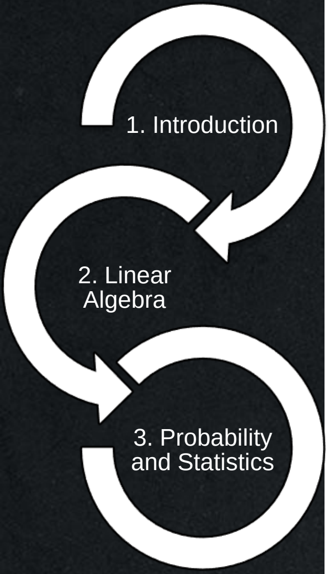
Prior knowledge can be based on:
- Engineering heuristics (expert knowledge)
- The posterior PDF obtained from previous data
- Non-informative prior (i.e. absence of prior knowledge)



from [2]

03. Probability / Statistics

Bayes Theorem



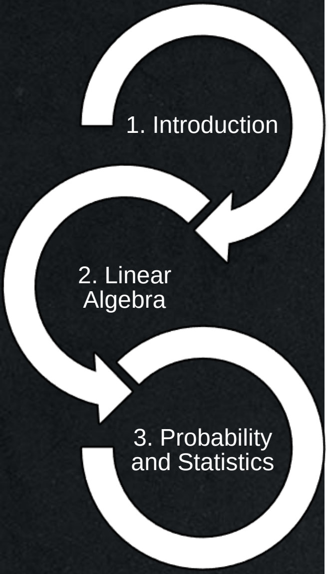
Bayes Theorem – Likelihood $\boldsymbol{f(\mathcal{D}|x)}$

$f(\mathcal{D}|\boldsymbol{x})$ describes the conditional probability of a set of observations D given the values that $\boldsymbol{x}$ can take.

"how the data is produced"

Examples:

- Hooke's law: $F = k \cdot w$
- Parabola: $y = a \cdot x^2 + b \cdot x + c$
- Tossing a dice: $f(\mathcal{D}|x) = 1$



Bayes Theorem – Evidence $\boldsymbol{f}(\boldsymbol{\mathcal{D}})$

$f(\mathcal{D})$ is called the evidence or the normalization constant.

The posterior integral must be equal to 1

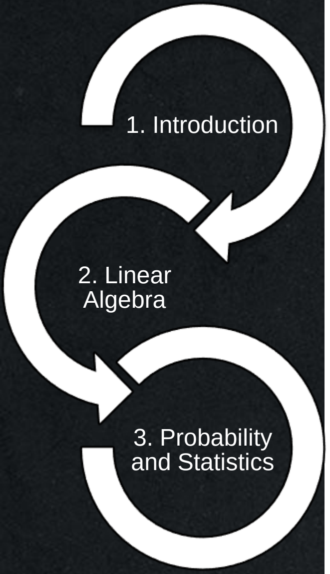
$$\underbrace{\sum_{x} p(\boldsymbol{x}|\mathcal{D})}_{\text{discrete case}} \equiv \underbrace{\int f(\boldsymbol{x}|\mathcal{D}) d\boldsymbol{x}}_{\text{continous case}} = 1$$

so that

$$p(\mathcal{D}) \underbrace{\sum_{x} p(\boldsymbol{y}|\boldsymbol{x}) \cdot p(\boldsymbol{x})}_{\text{discrete case}}, \qquad f(\mathcal{D}) \underbrace{\int f(\boldsymbol{y}|\boldsymbol{x}) \cdot f(\boldsymbol{x}) d\boldsymbol{x}}_{\text{continous case}} = 1$$

03. Probability / Statistics

Monte Carlo Methods



Monte Carlo Methods – Short Introduction

How to estimate integrals (expectations; evidence) for high-dimensional problems?

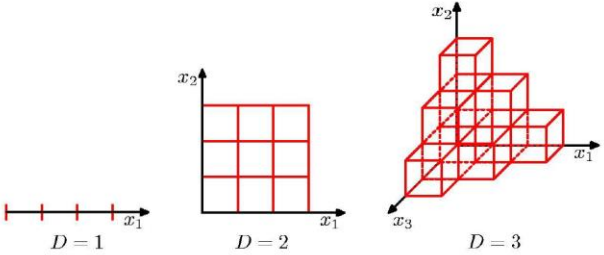
$$\underbrace{f(\boldsymbol{\theta}|\mathcal{D})}_{\text{posterior}} = \frac{\overbrace{f(\mathcal{D}|\boldsymbol{\theta})f(\boldsymbol{\theta})}^{\text{unnormalized posterior}}}{\underbrace{f(\mathcal{D})}_{\text{unknown normalization cte.}}}$$

$$f(\mathcal{D}) = \int f(\mathcal{D}|\boldsymbol{x}) \cdot f(\boldsymbol{x})dx$$

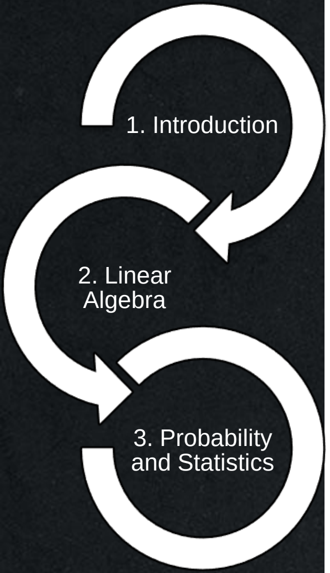
$$\approx \sum_{i=1}^{N} f(\mathcal{D}|\boldsymbol{x}_i) \cdot f(\boldsymbol{x}_i)\,\Delta\boldsymbol{x}_i$$

- standard (numerical) integration techniques are just efficient for few dimension
- # of parameters of a distribution is decisive for complexity of a problem
- Curse of Dimensionality

Solution: Monte Carlo Sampling



from [1]

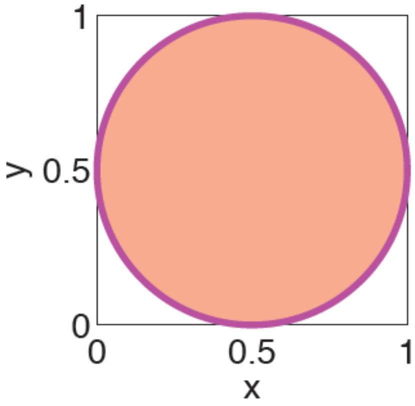


Monte Carlo Methods – Example: Estimate area content of a circle

Consider a circle with diameter *D* = 1: $(x-0.5)^2+(y-0.5)^2=r^2$

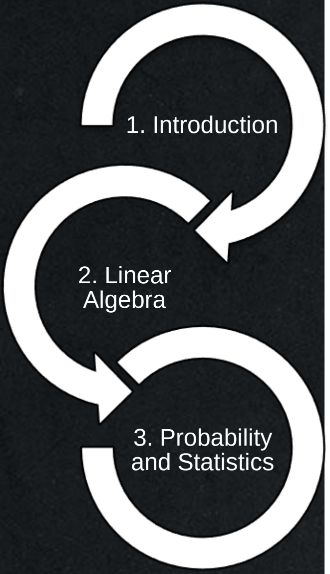
and *indicator function* $I(x,y)=\begin{cases}1 & \text{if } (x-0.5)^2+(y-0.5)^2\le r^2\\ 0 & \text{else}\end{cases}$

$$\boxed{\begin{aligned}\underbrace{a}_{\text{area}} &= \iint_{y\,x} I(x,y) f_{XY}(x,y)\,dx\,dy\\ &= \mathbb{E}[I(X,Y)]\end{aligned}}$$



from [2]

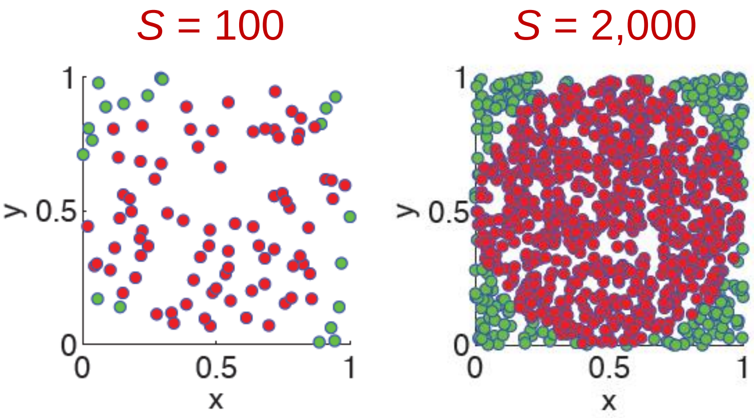
03. Probability / Statistics
Monte Carlo Methods



Monte Carlo Methods – Example: Estimate area content of a circle

$$a = \mathbb{E}[I(X,Y)] = \lim_{S\to\infty} \frac{1}{S} \sum_{s=1}^{S} I(x_s, y_s)$$

$$a \cong \mathbb{E}[I\widehat{(X,Y)}] = \frac{1}{S} \sum_{s=1}^{S} I(x_s, y_s)$$

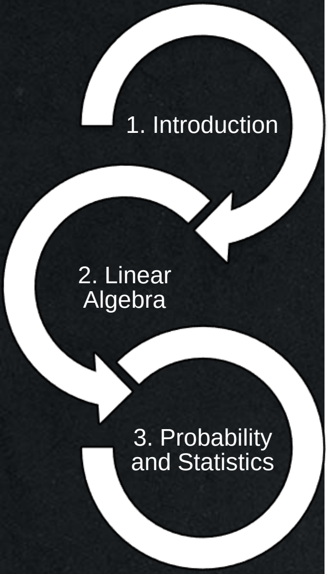


$\pi r^2 = 0.785$

from [2]

Estimation quality:

- $\mathbb{E}[I\widehat{(X,Y)}]$ depends on the number of samples
- Independent of the number of dimensions



Monte Carlo Methods – Metropolis Algorithm

Metropolis algorithm (Metropolis, 1953) developed during WWII within the Manhattan project (atomic bomb) in Los Alamos.

oldest MCMC algorithm

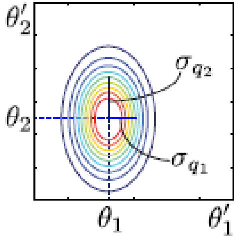
further MCMC samplers:
- Metropolis-Hasting
- Gibbs Sampling
- Slice Sampling

Algorithm 1: Metropolis

initialize $\boldsymbol{\theta}_0$;
for $s = 0, 1, 2, \cdots$ **do**
  define $\boldsymbol{\theta} = \boldsymbol{\theta}_s$;
  sample $\boldsymbol{\theta}' \sim q(\boldsymbol{\theta}'|\boldsymbol{\theta})$;
  compute $\alpha = \frac{\tilde{f}(\boldsymbol{\theta}')}{\tilde{f}(\boldsymbol{\theta})}$;
  compute $r = \min(1, \alpha)$;
  sample $u \sim \mathcal{U}(0, 1)$;
  if $u < r$ **then**
    $\boldsymbol{\theta}_{s+1} = \boldsymbol{\theta}'$;
  else
    $\boldsymbol{\theta}_{s+1} = \boldsymbol{\theta}_s$;

Proposal Distribution

$$q(\theta'|\theta) = \mathcal{N}\left(\theta'; \theta, \begin{bmatrix} \sigma_{q_1}^2 & 0 \\ 0 & \sigma_{q_2}^2 \end{bmatrix}\right)$$

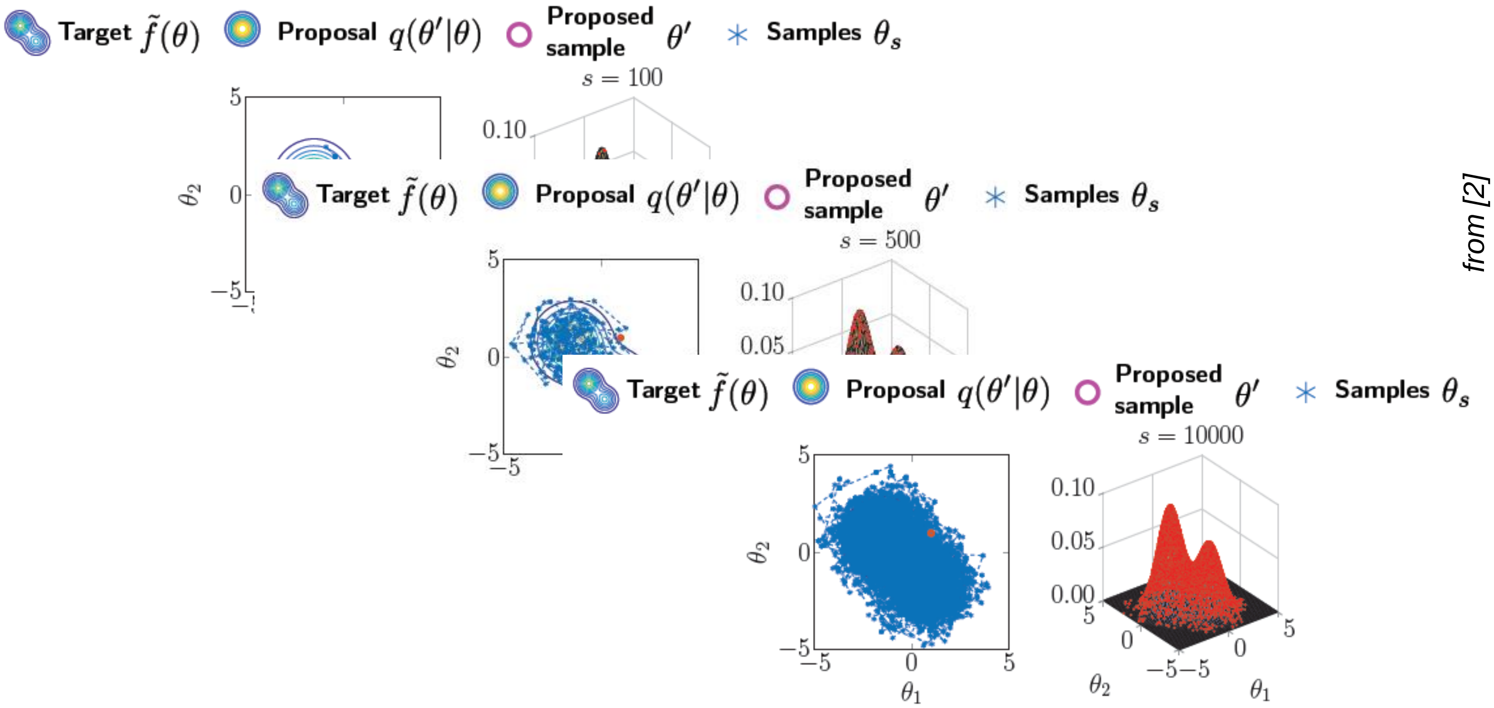


from [2]

random acceptance or rejection

03. Probability / Statistics
Monte Carlo Methods

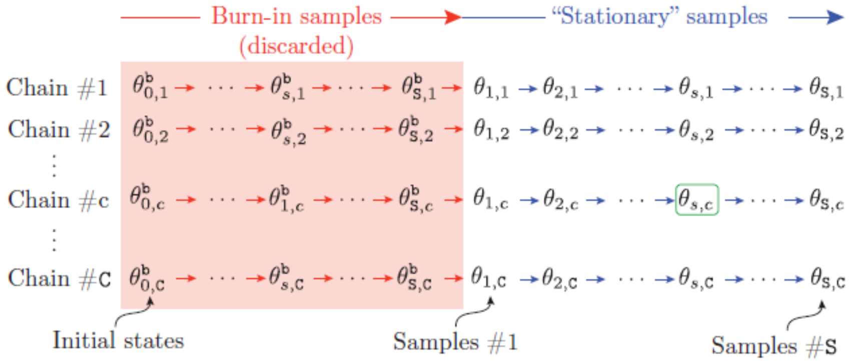
Monte Carlo Methods – Metropolis Algorithm



from [2]

03. Probability / Statistics
Monte Carlo Methods

Monte Carlo Methods – Convergence Monitoring



from [2]

03. Probability / Statistics
Monte Carlo Methods

Monte Carlo Methods – Convergence Monitoring

Within-chains

Mean:

$$\bar{\theta}_{\cdot c} = \frac{1}{S}\sum_{s=1}^{S}\theta_{s,c}$$

Variance:

$$\underbrace{W = \frac{1}{C}\sum_{c=1}^{C}\left[\frac{1}{S-1}\sum_{s=1}^{S}(\theta_{s,c} - \bar{\theta}_{\cdot c})^2\right]}_{\text{Underestimates } \mathrm{Var}[\theta_{s,c}]}$$

Between-chains

Mean:

$$\bar{\theta}_{\cdot\cdot} = \frac{1}{C}\sum_{c=1}^{C}\bar{\theta}_{\cdot c}$$

Variance:

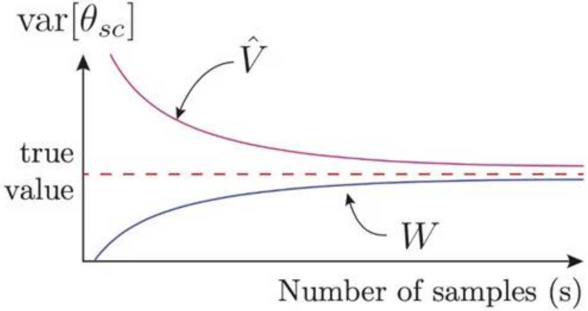
$$B = \frac{1}{C-1}\sum_{c=1}^{C}(\bar{\theta}_{\cdot c} - \bar{\theta}_{\cdot\cdot})^2$$

$$\overbrace{\hat{V} = \frac{S-1}{S}W + B}^{\text{Overestimates } \mathrm{Var}[\theta_{s,c}]}$$

Convergence $\hat{R} = \sqrt{\frac{\hat{V}}{W}}$

If $\hat{R} \approx 1 \rightarrow$ ✓

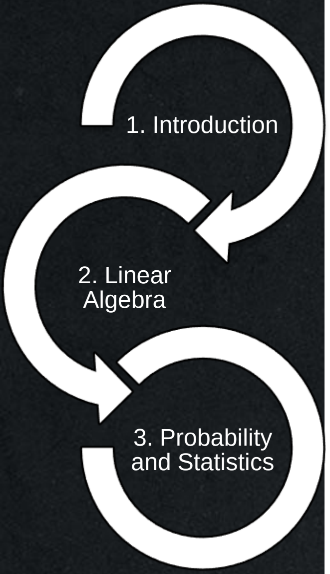
else if $\hat{R} > 1 \rightarrow$ ✗
(non-stationary)



from [2]

03. Probability / Statistics

Information Theory



Important Measures from Information Theory often used within AI algorithms

(Self-) Information of an event: (base is exponential e => unit: nats)

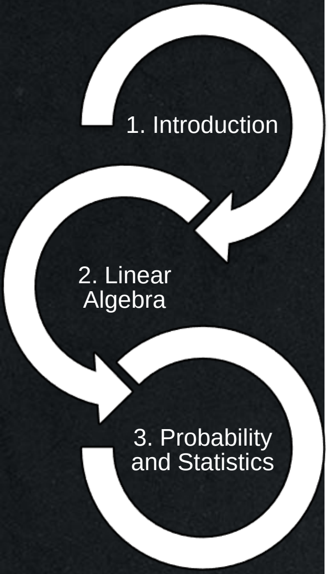
$$I(x) = -\log P(x).$$

Entropy:

$$H(\mathrm{x}) = \mathbb{E}_{\mathrm{x}\sim P}[I(x)] = -\mathbb{E}_{\mathrm{x}\sim P}[\log P(x)]$$

Kullback-Leibler (KL) Divergence:

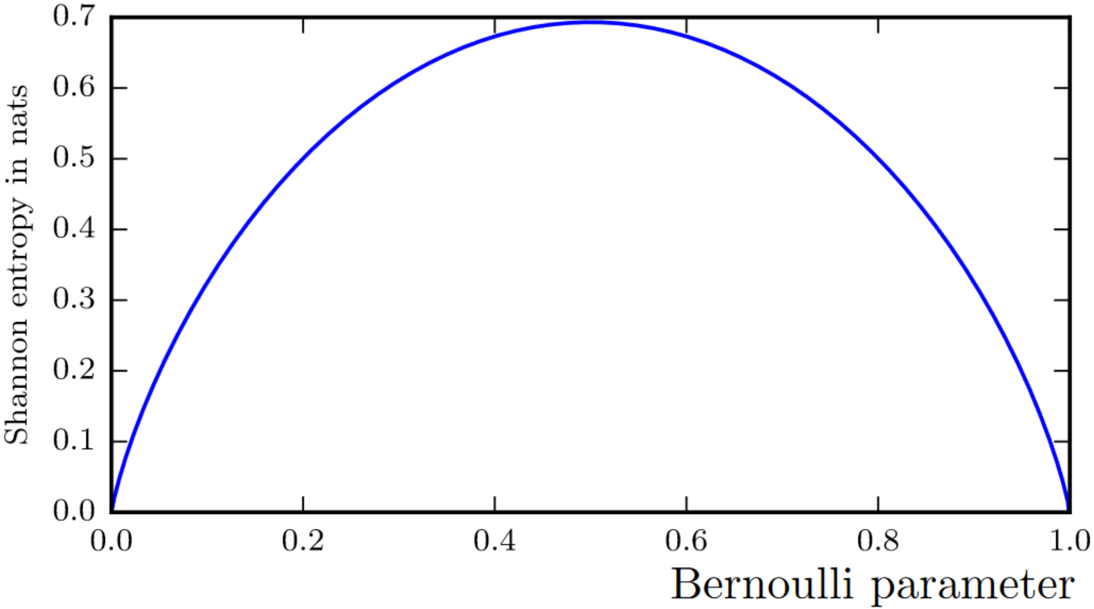
$$D_{\mathrm{KL}}(P\|Q) = \mathbb{E}_{\mathrm{x}\sim P}\left[\log \frac{P(x)}{Q(x)}\right] = \mathbb{E}_{\mathrm{x}\sim P}\left[\log P(x) - \log Q(x)\right]$$



03. Probability / Statistics
Information Theory

Important Measures from Information Theory often used within AI algorithms

Entropy of a Bernoulli Variable:

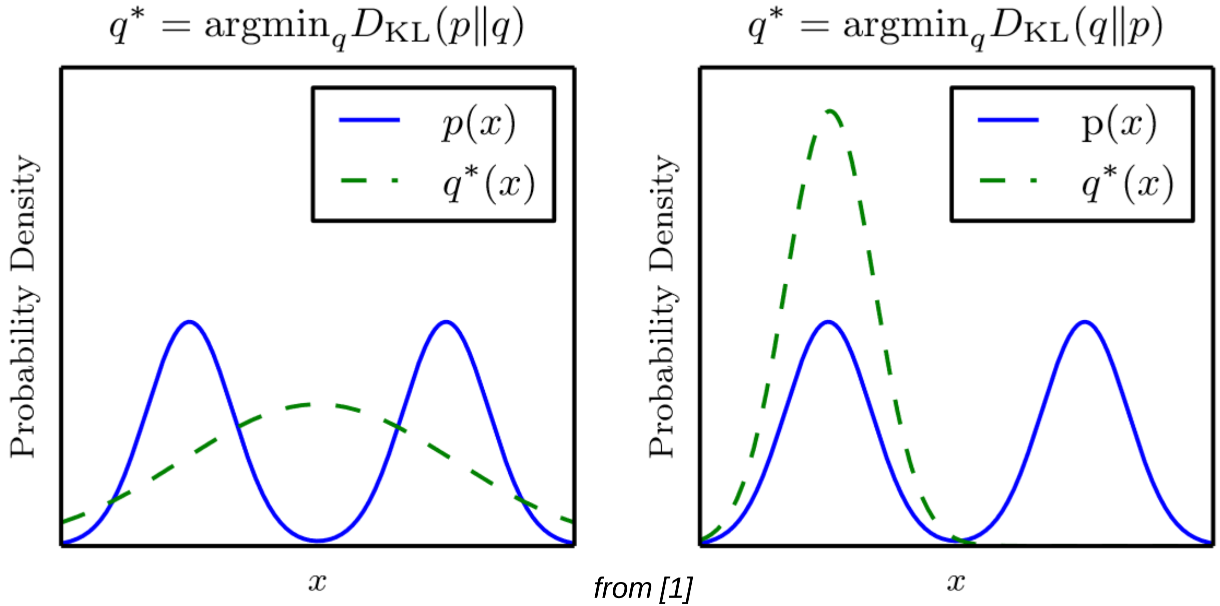


from [1]

03. Probability / Statistics
Information Theory

Important Measures from Information Theory often used within AI algorithms

Note: KL Divergence is Asymmetric

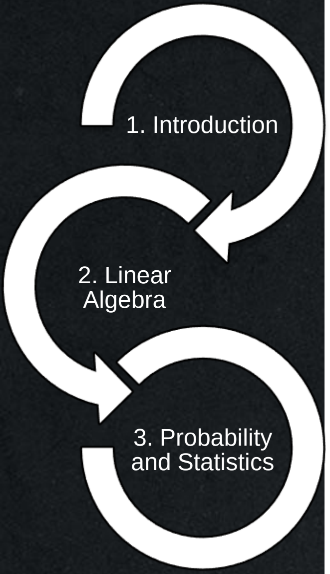


from [1]

03. Probability / Statistics

Summary

from [2]



Probabilities:

- Probabilities **describe our knowledge**
- The less we know, the more we should employ probability theory

Bayesian interpretation: $\Pr(E_i)$ quantifies the likelihood of an event with respect to others in $\mathcal{S}$

Rules/operations events: $\cap, \cup, \subset, \subseteq, \in$

Fundamental Axioms:

1. $0 \leq \Pr(E_i) \leq 1$
2. $\Pr(\mathcal{S}) = 1$
3. Si E_1 et E_2 are mutually exclusives $\Pr(E_1 \cup E_2) = \Pr(E_1) + \Pr(E_2)$

Inclusion-exclusion rule: $\Pr\left(\bigcup_{i=1}^{n} E_i\right) = \ldots$

Bayes Theorem: $\Pr(E_i|A) = \dfrac{\Pr(A|E_i)\Pr(E_i)}{\Pr(A)}$

Probability distributions: PDF, CDF, PMF, CMF

Multivariate Normal: $\mu_1, \sigma_1, \mu_2, \sigma_2, \rho_{12}$

Multivariate probability density function:

- $0 \leq f_{\mathbf{X}}(\mathbf{x})$
- $\int \cdots \int f_{\mathbf{X}}(\mathbf{x})d\mathbf{x} = 1$

Conditional probabilities:

- si $p_{X_1|X_2}(x_1|x_2) = p_{X_1}(x_1)$, $X_1 \perp\!\!\!\perp X_2$
- si $X_1 \perp\!\!\!\perp X_2$ $p_{X_1 X_2}(x_1, x_2) = p_{X_1}(x_1)p_{X_2}(x_2)$

General case: $X_1 \not\perp\!\!\!\perp X_2 \rightarrow$ **Chain rule**

Expectation & Variance:

$$\mathbb{E}[X] = \int x \cdot f_X(x)dx \quad (\text{Continuous R.V.})$$

$$\mathbb{E}[(X - \mu_X)^2] = \sigma_X^2 = \text{var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

Matrix notation: $\mathbf{\Sigma_X} = \mathbf{D_X R_X D_X}$

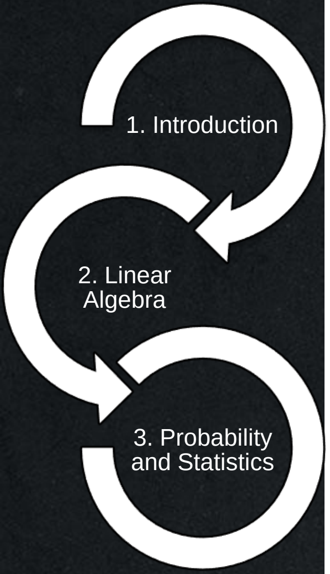
Function of random variables: $Y = g(\mathbf{X})$

$$f_Y(y)dy = f_X(x)dx$$

$$\begin{aligned} \mathbf{M}_Y &= \mathbf{g}(\mathbf{M}_X) = \mathbf{A}\mathbf{M}_X + \mathbf{B} \\ \mathbf{\Sigma}_Y &= \mathbf{A}\mathbf{\Sigma_X}\mathbf{A}^{\mathsf{T}} = \mathbf{J}_{\mathbf{y},\mathbf{x}}\mathbf{\Sigma}_X\mathbf{J}_{\mathbf{y},\mathbf{x}}{}^{\mathsf{T}} \end{aligned}$$

Linearization – First order approximation

$$\begin{aligned} Y = g(\mathbf{X}) &\cong g(\mathbf{M_X}) + \nabla g(\mathbf{M_X})(\mathbf{X} - \mathbf{M_X}) \\ \mu_Y &\cong g(\mathbf{M_X}) \\ \sigma_Y^2 &\cong \nabla g(\mathbf{M_X})\mathbf{\Sigma_X}\nabla g(\mathbf{M_X})^{\mathsf{T}} \end{aligned}$$



03. Probability / Statistics

Summary

from [2]

Univariate Normal:
$X \sim \mathcal{N}(x; \mu, \sigma^2), x \in (-\infty, +\infty)$

if $X \sim \mathcal{N}(x; \mu_X, \sigma_X^2),\ \ Y \sim \mathcal{N}(y; \mu_Y, \sigma_Y^2)$

$$\begin{aligned} Z &= X + Y \\ &\sim \mathcal{N}(z; \mu_Z, \sigma_Z^2) \end{aligned}$$

Multivariate Normal:
$\mathbf{X} \sim \mathcal{N}(x; \mathbf{M_X}, \mathbf{\Sigma_X})$

Normal conditional:
$f_{\mathbf{X}_1|\mathbf{X}_2}(\mathbf{x}_1|\mathbf{X}_2 = \mathbf{x}_2) = \mathcal{N}(\mathbf{x}_1; \mathbf{M}_{1|2}, \mathbf{\Sigma}_{1|2})$

Univariate Lognormal:
$X \sim \ln\mathcal{N}(x; \lambda, \zeta), x \in (0, +\infty)$

if
$X \sim \ln\mathcal{N}(x; \lambda_X, \zeta_X^2),\ \ Y \sim \ln\mathcal{N}(y; \lambda_Y, \zeta_Y^2)$

$$\begin{aligned} Z &= X \cdot Y \\ &\sim \ln\mathcal{N}(z; \lambda_Z, \zeta_Z^2) \end{aligned}$$

Beta:
$X \sim \mathrm{Beta}(x; \alpha, \beta), x \in (0, 1)$

03. Probability / Statistics

Summary

from [2]

Bayes's rule: $$\underbrace{f(\mathbf{x}|\mathcal{D})}_{\text{posterior}} = \frac{\overbrace{f(\mathcal{D}|\mathbf{x})}^{\text{likelihod}} \cdot \overbrace{f(\mathbf{x})}^{\text{prior}}}{\underbrace{f(\mathcal{D})}_{\text{normalization cte.}}}$$

Prior – $f(x)$, $f(\theta)$: based on: Engineering heuristics, Previous posterior PDF, Non-informative prior

Likelihood – $f(\mathcal{D}|x)$ or $f(\mathcal{D}|\theta)$: Conditional probability of a set of observations $\mathcal{D}$ given the values that $\mathbf{x}$ or $\boldsymbol{\theta}$ can take

Evidence – $f(\mathcal{D})$:

$$f(\mathcal{D}) = \underbrace{\int f(\mathbf{y}|\mathbf{x}) \cdot f(\mathbf{x}) d\mathbf{x}}_{\text{continuous case}} = 1$$

MC sampling from the prior

$$\begin{aligned} f(\mathcal{D}) &\approx \frac{1}{S} \sum_{s=1}^{S} \left[\prod_{j=1}^{D} f(y_j|\mathbf{x}_s) \right] \\ \mathbb{E}[\mathbf{x}|\mathcal{D}] &\approx \frac{1}{S} \sum_{s=1}^{S} \left[\mathbf{x}_s \cdot \frac{\prod_{j=1}^{D} f(y_j|\mathbf{x}_s)}{f(\mathcal{D})} \right] \\ \text{var}[\mathbf{x}|\mathcal{D}] &\approx \frac{1}{S} \sum_{s=1}^{S} \left[(\mathbf{x}_s - \mathbb{E}[\mathbf{x}|\mathcal{D}])^2 \cdot \frac{\prod_{j=1}^{D} f(y_j|\mathbf{x}_s)}{f(\mathcal{D})} \right] \end{aligned}$$

Limited to simple cases → MCMC Module

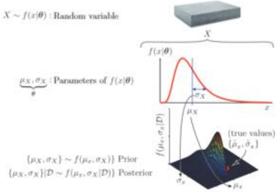
Posterior – $f(\mathbf{x}|\mathcal{D})$: When the number of independent observations $D \to \infty$

$$f(\mathbf{x}|\mathcal{D}) = \frac{f(\mathcal{D}|\mathbf{x})f(\mathbf{x})}{f(\mathcal{D})} \to \underbrace{\delta(\overbrace{\check{\mathbf{x}}}^{\text{true value}})}_{\text{Dirac delta PDF}}$$

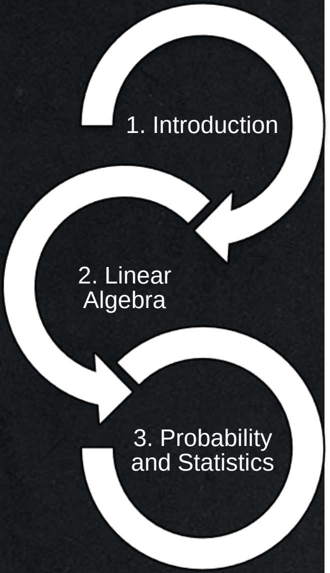
$$f(\boldsymbol{\theta}|\mathcal{D}) = \frac{f(\mathcal{D}|\boldsymbol{\theta})f(\boldsymbol{\theta})}{f(\mathcal{D})} \to \underbrace{\delta(\overbrace{\check{\boldsymbol{\theta}}}^{\text{true value}})}_{\text{Dirac delta PDF}}$$

Posterior Predictive – $f(\mathbf{x}|\mathcal{D})$

$$f(x|\mathcal{D}) = \int f(x; \boldsymbol{\theta}) f(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta}$$



Conjugate priors: For specific combinaisons of prior distribution and likelihood function, the posterior PDF follows the same type of distribution than the prior PDF



References

[1] Goodfellow, I.; Bengio, Y.; Courville, A. (2016) *Deep Learning*, MIT press

[2] Goulet, J. (2020) *Probabilistic Machine Learning for Civil Engineers,* MIT press

[3] Bishop, C. (2006) *Pattern Recognition and Machine Learning*, Springer

[4] Murphy, K. (2012) *Machine Learning: A Probabilistic Perspective*, MIT press

researchgate