Mathematical Tools: Probability Theory, Algebra, ...

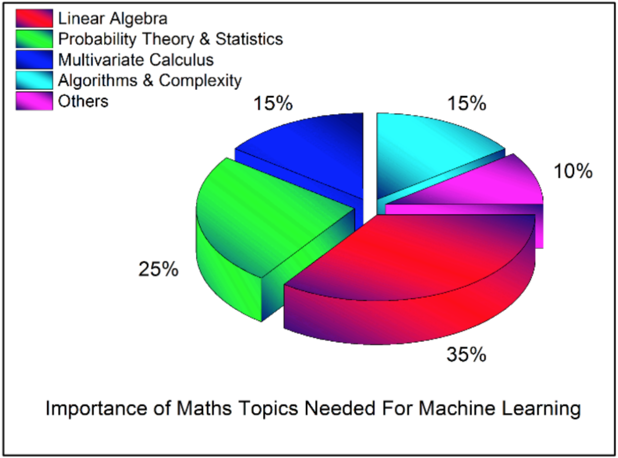
Mário A. T. Figueiredo

Instituto Superior Técnico & Instituto de Telecomunicações

Lisboa, **Portugal**

LxMLS 2022: Lisbon Machine Learning School

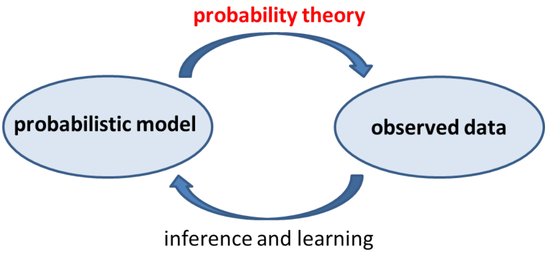
July 24, 2022



(by Wale Akinfaderin, `https://tinyurl.com/y6hc9sh8`)

Part I: Probability Theory

Probability theory



- Probability theory has its roots in games of chance

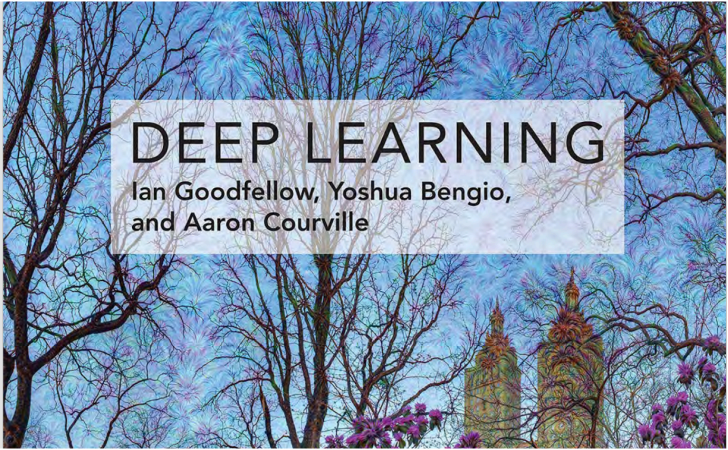
- Great names of science: Bayes, Bernoulli(s), Boltzman, Cardano, Cauchy, Fermat, Huygens, Kolmogorov, Laplace, Pascal, Poisson, ...
- Tool to handle uncertainty, information, knowledge, observations, ...
- ...thus also learning, decision making, inference, science,...

Still important today, in the Deep Learning age?

CONTENTS

What book is this from?

Do we still need this?



What is probability?

Example: $\mathbb{P}(\text{randomly drawn card is } \clubsuit) = 13/52$.

Example: $\mathbb{P}(\text{getting 1 in throwing a fair die}) = 1/6$.

- Classical definition: $\mathbb{P}(A) = \dfrac{N_A}{N}$

  ...with N mutually exclusive equally likely outcomes,
  N_A of which result in the occurrence of A. *Laplace, 1814*

- Frequentist definition: $\mathbb{P}(A) = \lim_{N\to\infty} \dfrac{N_A}{N}$

  ...relative frequency of occurrence of A in infinite number of trials.

- Subjective probability: $\mathbb{P}(A)$ is a degree of belief. *de Finetti, 1930s*

  .

  ...gives meaning to $\mathbb{P}(\text{"it will rain today"})$, or
  $\mathbb{P}(\text{"Patient A has disease } x\text{"})$

MENU /

aeon

The concept of probability is not as simple as you think

Nevin Climenhaga

A summary of some interpretations of probability

	Classical	Frequentist	Subjective	Propensity
Main hypothesis	Principle of indifference	Frequency of occurrence	Degree of belief	Degree of causal connection
Conceptual basis	Hypothetical symmetry	Past data and reference class	Knowledge and intuition	Present state of system
Conceptual approach	Conjectural	Empirical	Subjective	Metaphysical
Single case possible	Yes	No	Yes	Yes
Precise	Yes	No	No	Yes
Problems	Ambiguity in principle of indifference	Circular definition	Reference class problem	Disputed concept

"The mathematics of probability can be developed on an entirely axiomatic basis, independent of any interpretation." (wikipedia)

Key concepts: Sample space and events

- Sample space $\mathcal{X}$ = set of possible outcomes of a random experiment.

  Examples:

  - Tossing two coins: $\mathcal{X} = \{HH, TH, HT, TT\}$
  - Roulette: $\mathcal{X} = \{1, 2, ..., 36\}$
  - Draw a card from a shuffled deck: $\mathcal{X} = \{A\clubsuit, 2\clubsuit, ..., Q\diamondsuit, K\diamondsuit\}$.

- An event A is a subset of $\mathcal{X}$: $A \subseteq \mathcal{X}$ (also written $A \in 2^{\mathcal{X}}$).

  Examples:

  - "exactly one H in 2-coin toss": $A = \{TH, HT\}$.
  - "odd number in the roulette": $B = \{1, 3, ..., 35\}$.
  - "drawn a $\heartsuit$ card": $C = \{A\heartsuit, 2\heartsuit, ..., K\heartsuit\}$

Key concepts: Sample space and events

- Sample space $\mathcal{X}$ = set of possible outcomes of a random experiment.

  (More delicate) examples:
  - Distance travelled by tossed die: $\mathcal{X} = \mathbb{R}_+$
  - Location of the next rain drop on a given square tile: $\mathcal{X} = \mathbb{R}^2$

- Properly handling the continuous case requires deeper concepts:
  - Sigma algebras, Borel sets, measurable functions, ...

...mathematically **heavier** stuff, not covered here

Kolmogorov's Axioms for Probability

- Probability is a function that maps events A into the interval $[0, 1]$.

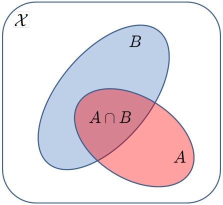
  Kolmogorov's axioms (1933) for probability

  - For any A, $\mathbb{P}(A) \geq 0$
  - $\mathbb{P}(\mathcal{X}) = 1$
  - If $A_1, A_2 ... \subseteq \mathcal{X}$ are disjoint events, then $\mathbb{P}\Big(\bigcup_i A_i\Big) = \sum_i \mathbb{P}(A_i)$

- From these axioms, many results can be derived.

  Examples:

  - $\mathbb{P}(\emptyset) = 0$
  - $C \subset D \;\Rightarrow\; \mathbb{P}(C) \leq \mathbb{P}(D)$
  - $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$
  - $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$ (union bound)

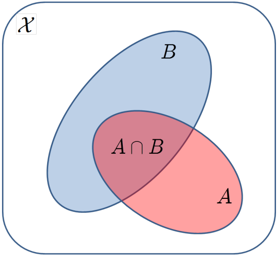


Conditional Probability and Independence

- If $\mathbb{P}(B) > 0$, $\mathbb{P}(A|B) = \dfrac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$ (conditional prob. of A, given B)

- ...satisfies all of Kolmogorov's axioms:
  - For any $A \subseteq \mathcal{X}$, $\mathbb{P}(A|B) \geq 0$
  - $\mathbb{P}(\mathcal{X}|B) = 1$
  - If A_1, A_2, ... $\subseteq \mathcal{X}$ are disjoint,

$$\mathbb{P}\Big(\bigcup_i A_i \Big| B\Big) = \sum_i \mathbb{P}(A_i|B)$$



- Independence: A, B are independent ($A \perp\!\!\!\perp B$):

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B).$$

Conditional Probability and Independence

- If $\mathbb{P}(B) > 0$, $\quad \mathbb{P}(A|B) = \dfrac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$

- Events A, B are independent ($A \perp\!\!\!\perp B$) $\Leftrightarrow$ $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$.

- Relationship with conditional probabilities:

$$A \perp\!\!\!\perp B \;\Leftrightarrow\; \mathbb{P}(A|B) = \mathbb{P}(A)$$

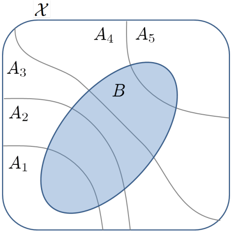
- Example: $\mathcal{X} =$ "52 cards", $A = \{4\heartsuit, 4\clubsuit, 4\diamondsuit, 4\spadesuit\}$, and $B = \{A\heartsuit, 2\heartsuit, ..., K\heartsuit\}$; then, $\mathbb{P}(A) = 1/13$, $\mathbb{P}(B) = 1/4$

$$\begin{aligned} \mathbb{P}(A \cap B) &= \mathbb{P}(\{4\heartsuit\}) = \frac{1}{52} \\ \mathbb{P}(A)\,\mathbb{P}(B) &= \frac{1}{13}\,\frac{1}{4} = \frac{1}{52} \\ \mathbb{P}(A|B) &= \mathbb{P}(\text{“}4\text{”}\,|\,\text{“}\heartsuit\text{”}) = \frac{1}{13} = \mathbb{P}(A) \end{aligned}$$

Bayes Theorem

- Law of total probability: if $A_1, ..., A_n$ are a partition of $\mathcal{X}$

$$\mathbb{P}(B) = \sum_i \mathbb{P}(B|A_i)\mathbb{P}(A_i)$$
$$= \sum_i \mathbb{P}(B \cap A_i)$$



- Bayes' theorem: if $\{A_1, ..., A_n\}$ is a partition of $\mathcal{X}$

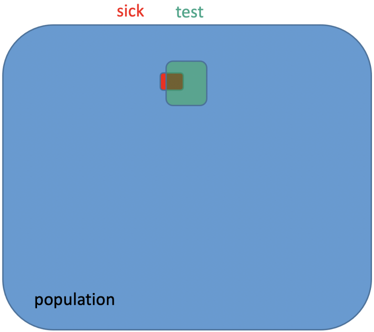
$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B \cap A_i)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A_i)\,\mathbb{P}(A_i)}{\sum_j \mathbb{P}(B|A_j)\mathbb{P}(A_j)}$$

Bayesian inference

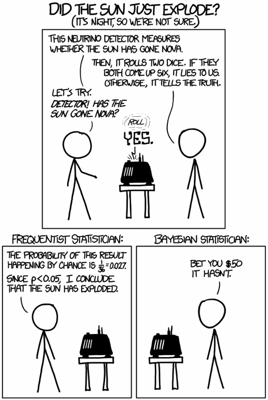
Bayes (1763)

Laplace (1812)

$$\mathbf{P}(\text{sick} \mid \text{test}) = \frac{\mathbf{P}(\text{test}, \text{sick})}{\mathbf{P}(\text{test})} = \frac{\overbrace{\mathbf{P}(\text{test} \mid \text{sick})}^{\text{sensitivity}}\ \overbrace{\mathbf{P}(\text{sick})}^{\text{prevalence}}}{\mathbf{P}(\text{test} \mid \text{sick})\,\mathbf{P}(\text{sick}) + \underbrace{\mathbf{P}(\text{test} \mid \text{not sick})}_{\text{false positive}}\ \underbrace{\mathbf{P}(\text{not sick})}_{(1-\text{prevalence})}}$$

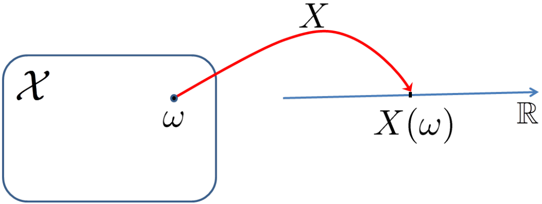


 = prevalence = **P**(sick)



Random Variables

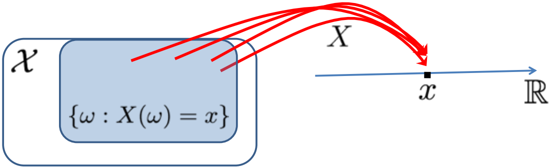
- A (real) random variable (RV) is a function: $X : \mathcal{X} \rightarrow \mathbb{R}$



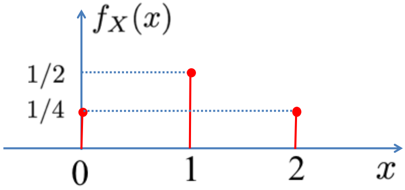
  - Discrete RV: range of X is countable (*e.g.*, $\mathbb{N}$ or $\{0, 1\}$)
  - Continuous RV: range of X is uncountable (*e.g.*, $\mathbb{R}$ or $[0,\ 1]$)
  - Example: number of heads in tossing two coins,
    $\mathcal{X} = \{HH, HT, TH, TT\}$,
    $X(HH) = 2$, $X(HT) = X(TH) = 1$, $X(TT) = 0$.
    Range of $X = \{0, 1, 2\}$.
  - Example: distance traveled by a tossed coin; range of $X = \mathbb{R}_+$.

Discrete Random Variables

- Probability mass function: $f_X(x) = \mathbb{P}(\{\omega \in \mathcal{X} : X(\omega) = x\})$



- Example: number of heads in tossing 2 coins; range$(X) = \{0, 1, 2\}$.



Important Discrete Random Variables

- Uniform: $X \in \{x_1, ..., x_K\}$, pmf $f_X(x_i) = 1/K$.

  Example: a fair roulette $X \in \{1, ..., 36\}$, with $f_X(x) = 1/36$

  Example: a fair die $X \in \{1, ..., 6\}$, with $f_X(x) = 1/6$

- Bernoulli RV: $X \in \{0, 1\}$, pmf $f_X(x) = \left\{ \begin{array}{ccc} p & \Leftarrow & x = 1 \\ 1-p & \Leftarrow & x = 0 \end{array} \right.$

  Compact form: $f_X(x) = p^x (1-p)^{1-x}$.

  Example: a coin toss; heads = 0, tails = 1

  fair, if $p = 1/2$; unfair, if $p \neq 1/2$

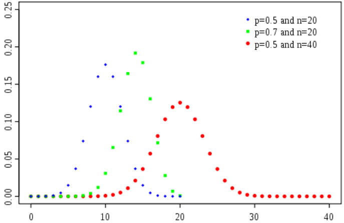
Important Discrete Random Variables

- Binomial RV: $X \in \{0, 1, ..., n\}$ (sum of n Bernoulli RVs)

$$f_X(x) = \text{Binomial}(x; n, p) = \binom{n}{x} p^x \, (1-p)^{(n-x)}$$

Binomial coefficients
("n choose x"):

$$\binom{n}{x} = \frac{n!}{(n-x)!\, x!}$$



Example: number of heads in n coin tosses.

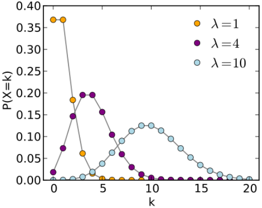
Other Important Discrete Random Variables

- Geometric(p): $X \in \mathbb{N}$, pmf $f_X(x) = p(1-p)^{x-1}$.

  Example: number of coin tosses until first heads.

- Poisson(λ):

$$X \in \mathbb{N} \cup \{0\},$$
$$\text{pmf } f_X(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$



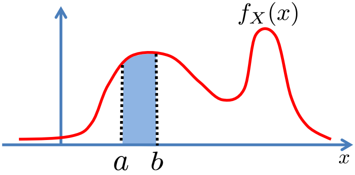
"...probability of the number of independent occurrences in a fixed (time/space) interval, if these occurrences have known average rate"

Examples: number of rain drops per second on a given area, number of calls per hour in a call center, number of tweets per day by DT, ...

Continuous Random Variables

- Probability density function (pdf, continuous RV): $f_X(x)$

$$\int_{-\infty}^{\infty} f_X(x) = 1 \qquad \mathbb{P}(X \in [a,b]) = \int_a^b f_X(x)\, dx$$

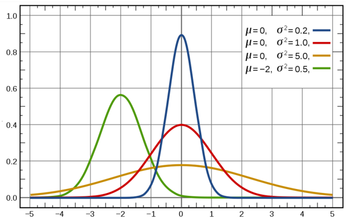


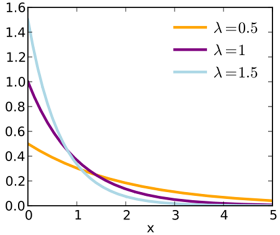
- Notice: $\mathbb{P}(X = c) = 0$

Important Continuous Random Variables

- Uniform: $f_X(x) = \text{Uniform}(x; a, b) = \begin{cases} \frac{1}{b-a} & \Leftarrow \quad x \in [a,\, b] \\ 0 & \Leftarrow \quad x \notin [a,\, b] \end{cases}$

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





- Exponential: $f_X(x) = \text{Exp}(x; \lambda) = \begin{cases} \lambda e^{-\lambda\, x} & \Leftarrow \quad x \geq 0 \\ 0 & \Leftarrow \quad x < 0 \end{cases}$

Expectation of (Real) Random Variables

- Expectation: $\mathbb{E}(X) = \begin{cases} \displaystyle\sum_i x_i \, f_X(x_i) & X \in \{x_1, ...x_K\} \subset \mathbb{R} \\ \displaystyle\int_{-\infty}^{\infty} x \, f_X(x) \, dx & X \text{ continuous} \end{cases}$

- Example: Bernoulli, $f_X(x) = p^x \, (1-p)^{1-x}$, for $x \in \{0, \, 1\}$.

$$\mathbb{E}(X) = 0 \, (1-p) + 1 \, p = p.$$

- Example: Binomial, $f_X(x) = \binom{n}{x} p^x \, (1-p)^{n-x}$, for $x \in \{0, ..., n\}$.

$$\mathbb{E}(X) = n \, p.$$

- Example: Gaussian, $f_X(x) = \mathcal{N}(x; \mu, \sigma^2)$. $\quad \mathbb{E}(X) = \mu$.

- Linearity of expectation:

$$\mathbb{E}(\alpha \, X + \beta \, Y) = \alpha \, \mathbb{E}(X) + \beta \, \mathbb{E}(Y), \qquad \alpha, \beta \, \in \mathbb{R}$$

Expectation of Functions of RVs

- $$\mathbb{E}(g(X)) = \begin{cases} \displaystyle\sum_i g(x_i) f_X(x_i) & X \text{ discrete, } g(x_i) \in \mathbb{R} \\ \displaystyle\int_{-\infty}^{\infty} g(x)\, f_X(x)\, dx & X \text{ continuous} \end{cases}$$

- Example: variance, $\text{var}(X) = \mathbb{E}\Big(\big(X - \mathbb{E}(X)\big)^2\Big) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$

- Example: Bernoulli variance, $\mathbb{E}(X^2) = \mathbb{E}(X) = p$, thus $\text{var}(X) = p(1-p)$.

- Example: Gaussian variance, $\mathbb{E}\big((X-\mu)^2\big) = \sigma^2$.

- Probability as expectation of indicator, $\mathbf{1}_A(x) = \begin{cases} 1 & \Leftarrow & x \in A \\ 0 & \Leftarrow & x \notin A \end{cases}$

$$\mathbb{P}(X \in A) = \int_A f_X(x)\, dx = \int \mathbf{1}_A(x)\, f_X(x)\, dx = \mathbb{E}(\mathbf{1}_A(X))$$

The importance of the Gaussian

The importance of the Gaussian

Take n independent RVs $X_1, ..., X_n$, with $\mathbb{E}[X_i] = \mu_i$ and $\text{var}(X_i) = \sigma_i^2$

- Their sum, $Y_n = \sum_{i=1}^{n} X_i$ satisfies:

$$\mathbb{E}[Y_n] = \sum_{i=1}^{n} \mu_i \equiv \mu \qquad\qquad \text{var}(Y_n) = \sum_i \sigma_i^2 \equiv \sigma^2$$

- Let $Z_n = \dfrac{Y_n - \mu}{\sigma}$, thus $\mathbb{E}[Z_n] = 0$ and $\text{var}(Z_n) = 1$

- Central limit theorem: under mild conditions,

$$\lim_{n\to\infty} Z_n \sim \mathcal{N}(0,1)$$

Two (or More) Random Variables

- Joint pmf of two discrete RVs: $f_{X,Y}(x,y) = \mathbb{P}(X = x \wedge Y = y)$.

  Extends trivially to more than two RVs.

- Joint pdf of two continuous RVs: $f_{X,Y}(x,y)$, such that

$$\mathbb{P}\big((X,Y) \in A\big) = \iint_A f_{X,Y}(x,y)\,dx\,dy, \qquad A \in \sigma(\mathbb{R}^2)$$

  Extends trivially to more than two RVs.

- Marginalization: $f_Y(y) = \begin{cases} \displaystyle\sum_x f_{X,Y}(x,y), & \text{if } X \text{ is discrete} \\ \displaystyle\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx, & \text{if } X \text{ continuous} \end{cases}$

- Independence:

$$X \perp\!\!\!\perp Y \;\Leftrightarrow\; f_{X,Y}(x,y) = f_X(x)\,f_Y(y) \;\underset{\not\Leftarrow}{\Rightarrow}\; \mathbb{E}(X\,Y) = \mathbb{E}(X)\,\mathbb{E}(Y).$$

Conditionals and Bayes' Theorem

- Conditional pmf (discrete RVs):

$$f_{X|Y}(x|y) = \mathbb{P}(X = x|Y = y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

- Conditional pdf (continuous RVs): $f_{X|Y}(x|y) = \dfrac{f_{X,Y}(x, y)}{f_Y(y)}$

  ...the meaning is technically delicate.

- Bayes' theorem: $f_{X|Y}(x|y) = \dfrac{f_{Y|X}(y|x)\, f_X(x)}{f_Y(y)}$ (pdf or pmf).

- Also valid in the mixed case (*e.g.*, X continuous, Y discrete).

Joint, Marginal, and Conditional Probabilities: An Example

- A pair of binary variables $X, Y \in \{0, 1\}$, with joint pmf:

$f_{X,Y}(x,y)$	$Y = 0$	$Y = 1$
$X = 0$	1/5	2/5
$X = 1$	1/10	3/10

- Marginals: $f_X(0) = \frac{1}{5} + \frac{2}{5} = \frac{3}{5}, \qquad f_X(1) = \frac{1}{10} + \frac{3}{10} = \frac{4}{10},$

$$f_Y(0) = \frac{1}{5} + \frac{1}{10} = \frac{3}{10}, \quad f_Y(1) = \frac{2}{5} + \frac{3}{10} = \frac{7}{10}.$$

- Conditional probabilities:

$f_{X\|Y}(x\|y)$	$Y = 0$	$Y = 1$
$X = 0$	2/3	4/7
$X = 1$	1/3	3/7

$f_{Y\|X}(y\|x)$	$Y = 0$	$Y = 1$
$X = 0$	1/3	2/3
$X = 1$	1/4	3/4

An Important Multivariate RV: Multinomial

- Multinomial: $X = (X_1, ..., X_K)$, $X_i \in \{0, ..., n\}$, s.t. $\sum_i X_i = n$,

$$f_X(x_1, ..., x_K) = \begin{cases} \binom{n}{x_1\ x_2\ \cdots\ x_K} p_1^{x_1}\, p_2^{x_2} \cdots p_K^{x_K} & \Leftarrow \quad \sum_i x_i = n \\ 0 & \Leftarrow \quad \sum_i x_i \neq n \end{cases}$$

$$\binom{n}{x_1\ x_2\ \cdots\ x_K} = \frac{n!}{x_1!\, x_2! \cdots x_K!}$$

Parameters: $p_1, ..., p_K \geq 0$, such that $\sum_i p_i = 1$.

- Generalizes the binomial from binary to K-classes.

- Example: tossing n independent fair dice, $p_1 = \cdots = p_6 = 1/6$.
  $x_i =$ number of outcomes with i dots (of course, $\sum_i x_i = n$)

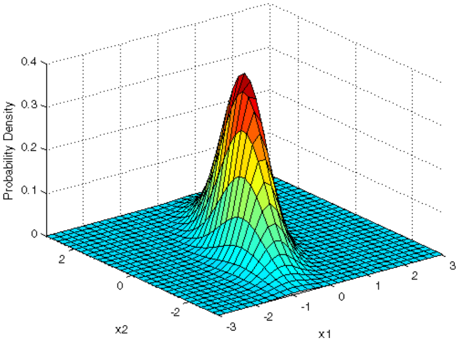
- Example: bag of words (BoW) multinomial model with vocabulary of K words

An Important Multivariate RV: Gaussian

- Multivariate Gaussian: $X \in \mathbb{R}^n$,

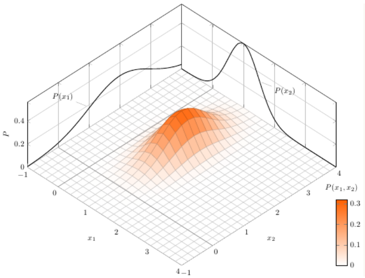
$$f_X(x) = \mathcal{N}(x; \mu, C) = \frac{1}{\sqrt{\det(2\,\pi\,C)}} \exp\left(-\frac{1}{2}(x-\mu)^T C^{-1}(x-\mu)\right)$$

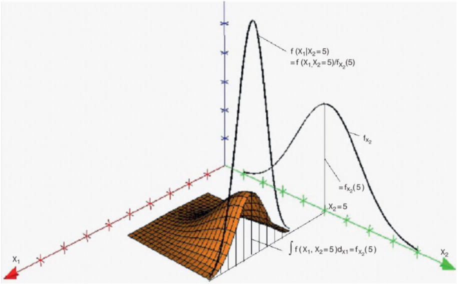
- Parameters: vector $\mu \in \mathbb{R}^n$ and matrix $C \in \mathbb{R}^{n\times n}$.
  Expected value: $\mathbb{E}(X) = \mu$. Meaning of C: later.



Key Properties of Multivariate Gaussian

- Marginals are Gaussian.
- Conditionals are Gaussian.





Transformations

$X \sim f_X$ and $Y = g(X) \;\Rightarrow\; f_Y =?$

- Discrete case:
  $f_Y(y) = \mathbb{P}(g(X) = y) = \mathbb{P}\big(\{x : g(x) = y\}\big) = \mathbb{P}\big(g^{-1}(y)\big)$

- Continuous case (for g strictly monotonic, thus invertible):

$$f_Y(y) = f_X\left(g^{-1}(y)\right)\Big|\frac{d\,g^{-1}(y)}{d\,y}\Big|$$

- Continuous multivariate case (invertible):

$$f_Y(y) = f_X\left(g^{-1}(y)\right)\left|\text{det} J_{g^{-1}}(y)\right|$$

  where $\text{det} J_{g^{-1}}(y)$ is the determinant of the Jacobian of g^{-1} at y.

Covariance, Correlation, and all that...

- Covariance between two RVs:

$$\text{cov}(X,Y) = \mathbb{E}\Big[\big(X-\mathbb{E}(X)\big)\,\big(Y-\mathbb{E}(Y)\big)\Big] \;=\; \mathbb{E}(X\,Y) - \mathbb{E}(X)\,\mathbb{E}(Y)$$

- Relationship with variance: $\text{var}(X) = \text{cov}(X,X)$.

- Correlation: $\text{corr}(X,Y) = \rho(X,Y) = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)}\sqrt{\text{var}(Y)}} \in [-1,\,1]$

- $X \perp\!\!\!\perp Y \;\Leftrightarrow\; f_{X,Y}(x,y) = f_X(x)\,f_Y(y) \;\underset{\not\Leftarrow}{\Rightarrow}\; \text{cov}(X,\,Y) = 0.$

- Covariance matrix of multivariate RV, $X \in \mathbb{R}^n$:

$$\text{cov}(X) = \mathbb{E}\Big[\big(X-\mathbb{E}(X)\big)\big(X-\mathbb{E}(X)\big)^T\Big] = \mathbb{E}(X\,X^T) - \mathbb{E}(X)\mathbb{E}(X)^T$$

- Covariance of Gaussian RV, $f_X(x) = \mathcal{N}(x;\mu,C) \;\Rightarrow\; \text{cov}(X) = C$

More on Expectations and Covariances

Let $A \in \mathbb{R}^{n\times n}$ be a matrix and $a \in \mathbb{R}^n$ a vector.

- If $\mathbb{E}(X) = \mu$ and $Y = AX$, then $\mathbb{E}(Y) = A\mu$;
- If $\mathbb{E}(X) = \mu$ and $Y = X + \gamma$, then $\mathbb{E}(Y) = \mu + \gamma$;
- If $\text{cov}(X) = C$ and $Y = AX$, then $\text{cov}(Y) = ACA^T$;
- If $\text{cov}(X) = C$ and $Y = a^T X \in \mathbb{R}$, then $\text{var}(Y) = a^T C a \geq 0$;
- If $\text{cov}(X) = C$ and $Y = C^{-1/2}X$, then $\text{cov}(Y) = I$;

Combining the 2-nd and the 5-th facts: standardization:

$$\mathbb{E}(X) = \mu,\;\; \text{cov}(X) = C,\;\; Y = C^{-\frac{1}{2}}(X-\mu) \quad \Rightarrow \quad \mathbb{E}(Y) = 0,\;\; \text{cov}(Y) = I$$

Combining the 2-nd and the 3-rd facts: reparametrization trick:

$$\mathbb{E}(X) = 0,\;\; \text{cov}(X) = I,\;\; Y = AX+\mu \quad \Rightarrow \quad \mathbb{E}(Y) = \mu,\;\; \text{cov}(Y) = AA^T$$

Exponential Families

A pdf or pmf $f_X(x|\eta)$, with parameter(s) η, for $X \in \mathcal{X}$, is in an exponential family if

$$f_X(x|\eta) = \frac{1}{Z(\eta)}\, h(x)\, \exp\big(\eta^T \phi(x)\big)$$

where $\eta^T \phi(x) = \sum_j \eta_j \phi_j(x)$ and

$$Z(\eta) = \int_{\mathcal{X}} h(x)\, \exp\big(\eta^T \phi(x)\big)\, dx.$$

- Canonical parameter(s): η
- Sufficient statistics: $\phi(x)$
- Partition function: $Z(\eta)$

Examples: Bernoulli, Poisson, binomial, multinomial, Gaussian, exponential, beta, Dirichlet, Laplacian, log-normal, Wishart, ...

Exponential Families

$$f_X(x|\eta) = \frac{1}{Z(\eta)}\, h(x)\, \exp\big(\eta^T \phi(x)\big)$$

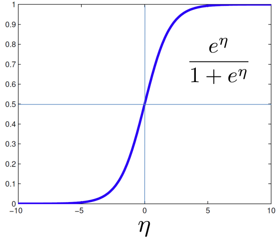
- Example: Bernoulli pmf $f_X(x) = p^x (1-p)^{1-x}$,

$$f_X(x) = \exp\big(x \log p + (1-x)\log(1-p)\big) = (1-p)\exp\big(x \log \tfrac{p}{1-p}\big),$$

thus $\eta = \log \frac{p}{1-p}$, $\phi(x) = x$, $Z(\eta) = 1 + e^\eta$, and $h(x) = 1$.

Notice that $p = \frac{e^\eta}{1+e^\eta}$

(logistic transformation)



More on Exponential Families

- Independent identically distributed (i.i.d.) observations:

$$X_1, ..., X_m \sim f_X(x|\eta) = \frac{1}{Z(\eta)}\, h(x)\, \exp\bigl(\eta^T \phi(x)\bigr)$$

then

$$f_{X_1,...,X_m}(x_1, ..., x_m|\eta) = \frac{1}{Z(\eta)^m} \left(\prod_{j=1}^{m} h(x_i)\right) \exp\left(\eta^T \sum_{j=1}^{m} \phi(x_j)\right)$$

- Expected sufficient statistics:

$$\frac{d \log Z(\eta)}{d\eta} = \frac{\frac{dZ(\eta)}{d\eta}}{Z(\eta)} = \frac{1}{Z(\eta)} \int \phi(x) h(x) \exp\bigl(\eta^T \phi(x)\bigr)\, dx = \mathbb{E}\bigl(\phi(X)\bigr)$$

Important Inequalities

- Markov's ineqality: if $X \geq 0$ is an RV with expectation $\mathbb{E}(X)$, then

$$\mathbb{P}(X > t) \leq \frac{\mathbb{E}(X)}{t}$$

Simple proof:

$$t\,\mathbb{P}(X > t) = \int_t^\infty t\, f_X(x)\, dx \leq \int_t^\infty x\, f_X(x)\, dx = \mathbb{E}(X) - \underbrace{\int_0^t x\, f_X(x)\, dx}_{\geq 0} \leq \mathbb{E}(X)$$

- Chebyshev's inequality: $\mu = \mathbb{E}(Y)$ and $\sigma^2 = \text{var}(Y)$, then

$$\mathbb{P}(|Y - \mu| \geq s) \leq \frac{\sigma^2}{s^2}$$

...simple corollary of Markov's inequality, with $X = |Y - \mu|^2,\ t = s^2$

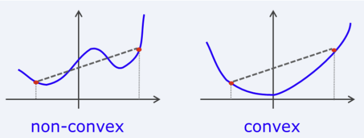
Important Inequalities

- Cauchy-Schwartz's inequality for RVs:

$$\mathbb{E}(|X\,Y|) \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}$$

- Recall that a real function g is convex if, for any x, y, and $\alpha \in [0,1]$

$$g(\alpha x + (1-\alpha)y) \leq \alpha g(x) + (1-\alpha)g(y)$$



Jensen's inequality: if g is a real convex function, then

$$g(\mathbb{E}(X)) \leq \mathbb{E}(g(X))$$

Examples: $\mathbb{E}(X)^2 \leq \mathbb{E}(X^2) \;\Rightarrow\;$ var$(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 \geq 0.$
$\mathbb{E}(\log X) \leq \log \mathbb{E}(X)$, for X a positive RV.

Information, entropy, and all that...

Entropy of a discrete RV $X \in \{1, ..., K\}$: $\boxed{H(X) = -\sum_{x=1}^{K} f_X(x) \log f_X(x)}$

- Positivity: $H(X) \geq 0$;
  $H(X) = 0 \;\Leftrightarrow\; f_X(i) = 1$, for exactly one $i \in \{1, ..., K\}$.

- Upper bound: $H(X) \leq \log K$;
  $H(X) = \log K \;\Leftrightarrow\; f_X(x) = 1/K$, for all $x \in \{1, ..., K\}$

- Measure of uncertainty/randomness of X

- With $\log_2$, units are bits/symbol

- Central role in information/coding theory: lower bound on expected number of bits to code X

- Widely used: physics, biological sciences (computational biology, neurosciences, ecology, ...), economics, finances, social sciences, ...

Entropy and all that...

Continuous RV X, differential entropy: $\boxed{h(X) = -\int f_X(x) \log f_X(x)\, dx}$

- $h(X)$ can be positive or negative (unlike in the discrete case)

  Example: for $f_X(x) = \text{Uniform}(x; a, b)$,

$$h(X) = \log(b - a).$$

- Gaussian upper bound: $f_X(x) = \mathcal{N}(x; \mu, \sigma^2)$, then

$$h(X) = \frac{1}{2} \log(2\pi e \sigma^2).$$

  For any RV Y with $\text{var}(Y) = \sigma^2$, then $h(Y) \leq \frac{1}{2} \log(2\pi e \sigma^2)$.

  ...yet another reason for why the Gaussian is important.

Kullback-Leibler divergence

Kullback-Leibler divergence (KLD) between two pmf:

$$D(f_X \| g_X) = \sum_{x=1}^{K} f_X(x) \log \frac{f_X(x)}{g_X(x)}$$

Positivity: $D(f_X \| g_X) \geq 0$
$D(f_X \| g_X) = 0 \;\Leftrightarrow\; f_X(x) = g_X(x),$ for $x \in \{1, ..., K\}$

KLD between two pdf:

$$D(f_X \| g_X) = \int f_X(x) \log \frac{f_X(x)}{g_X(x)} \, dx$$

Positivity: $D(f_X \| g_X) \geq 0$
$D(f_X \| g_X) = 0 \;\Leftrightarrow\; f_X(x) = g_X(x),$ almost everywhere

Issues: not symmetric; $D(f_X \| g_X) = +\infty$ if $g_X(x) = 0$ and $f_X(x) \neq 0$

Mutual information

Mutual information (MI) between two random variables:

$$I(X;Y) = D\big(f_{X,Y} \| f_X\, f_Y\big)$$

Positivity: $I(X;Y) \geq 0$
$I(X;Y) = 0 \;\Leftrightarrow\; X, Y$ are independent.

MI = measure of dependency between two random variables

MI = number of bits of information that X has about Y

Bound: $I(X;Y) \leq \min\{H(X), H(Y)\}$

Deterministic function: if $Y = \phi(X)$, then $I(X;Y) = H(Y) \leq H(X)$

Recommended Reading (Probability and Statistics)

- A. Maleki and T. Do, "Review of Probability Theory", Stanford University, 2017 (`https://tinyurl.com/pz7p9g5`)
- K. Murphy, "Machine Learning: A Probabilistic Perspective", MIT Press, 2012 (Chapter 2).
- L. Wasserman, "All of Statistics: A Concise Course in Statistical Inference", Springer, 2004.

Part II: Algebra and a Few Other Things

Notation: Matrices and Vectors

- $A \in \mathbb{R}^{m \times n}$ is a matrix with m rows and n columns.

$$A = \begin{bmatrix} A_{1,1} & \cdots & A_{1,n} \\ \vdots & \ddots & \vdots \\ A_{m,1} & \cdots & A_{m,n} \end{bmatrix}.$$

- $x \in \mathbb{R}^n$ is a vector with n components,

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

- A (column) vector is a matrix with n rows and 1 column.

- A matrix with 1 row and n columns is called a row vector.

Matrix Transpose and Products

- Given matrix $A \in \mathbb{R}^{m\times n}$, its transpose A^T is such that $(A^T)_{i,j} = A_{j,i}$.
- A matrix A is symmetric if $A^T = A$.
- Given matrices $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times p}$, their product is

$$C = A\,B \;\in \mathbb{R}^{m\times p} \quad \text{where} \;\; C_{i,j} = \sum_{k=1}^{n} A_{i,k}\, B_{k,j}$$

- Inner product between vectors $x, y \in \mathbb{R}^n$:

$$\langle x, y\rangle = x^T y = y^T x = \sum_{i=1}^{n} x_i\, y_i \in \mathbb{R}.$$

- Outer product: $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$: $x\,y^T \in \mathbb{R}^{n\times m}$, where

$$(x\,y^T)_{i,\,j} = x_i\, y_j$$

Properties of Matrix Products and Transposes

- Given matrices $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times p}$, their product is

$$C = A\,B \;\in \mathbb{R}^{m\times p} \quad \text{where} \;\; C_{i,j} = \sum_{k=1}^{n} A_{i,k}\, B_{k,j}$$

- Matrix product is associative: $(AB)C = A(BC)$.
- In general, matrix product is not commutative: $AB \neq BA$.
- Transpose of product: $(A\,B)^T = B^T A^T$.
- Transpose of sum: $(A+B)^T = A^T + B^T$.

Special Matrices

- The identity matrix $I \in \mathbb{R}^{n\times n}$ is a square matrix such that

$$I_{ij} = \left\{ \begin{array}{ll} 1, & i = j \\ 0, & i \neq j \end{array} \right. \qquad I = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

- Neutral element of matrix product: $A\,I = I\,A = A$.
- Diagonal matrix: $(i \neq j) \Rightarrow A_{i,j} = 0$.
- Upper triangular matrix: $(j < i) \Rightarrow A_{i,j} = 0$.
- Lower triangular matrix: $(j > i) \Rightarrow A_{i,j} = 0$.

Eigenvalues, eigenvectors, determinant, trace

- A vector $x \in \mathbb{R}^n$ is an eigenvector of matrix $A \in \mathbb{R}^{n \times n}$ if

$$A\,x = \lambda\,x,$$

  where $\lambda \in \mathbb{R}$ is the corresponding eigenvalue.

- The eigenvalues of a diagonal matrix are the elements in the diagonal. (quiz: what are the eigenvectors?)

- Matrix trace: $\text{trace}(A) = \sum_i A_{i,i} = \sum_i \lambda_i$

- Matrix determinant: $|A| = \det(A) = \prod_i \lambda_i$

- Properties of determinant: $|AB| = |A||B|, \;\; |A^T| = |A|,$ $|\alpha\, A| = \alpha^n |A|$

- Properties of the trace: $\text{trace}(A+B) = \text{trace}(A) + \text{trace}(B),$
  $\text{trace}(ABC) = \text{trace}(CAB) = \text{trace}(BCA)$ (cyclic permutations)

Matrix Inverse

- Matrix $A \in \mathbb{R}^{n\times n}$ in invertible if there is $B \in \mathbb{R}^{n\times n}$ s.t. $AB = BA = I$.
- ...matrix B, such that $AB = BA = I$, denoted $B = A^{-1}$.
- Matrix $A \in \mathbb{R}^{n\times n}$ is invertible $\Leftrightarrow \det(A) \neq 0$.
- Determinant of inverse: $\det(A^{-1}) = \dfrac{1}{\det(A)}$.
- Solving system $Ax = b$, if A is invertible: $x = A^{-1}b$.
- Properties: $(A^{-1})^{-1} = A, \;\; (A^{-1})^T = (A^T)^{-1}, \;\; (A\,B)^{-1} = B^{-1}A^{-1}$
- There are many algorithms to compute A^{-1}; general case, computational cost $O(n^3)$.

Quadratic Forms and Positive (Semi-)Definite Matrices

- Given matrix $A \in \mathbb{R}^{n\times n}$ and vector $x \in \mathbb{R}^n$,

$$x^T A\, x = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{i,j}\, x_i\, x_j \ \in \ \mathbb{R}$$

  is called a quadratic form.

- A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is positive semi-definite (PSD) if, for any $x \in \mathbb{R}^n$, $x^T A\, x \geq 0$.

- A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is positive definite (PD) if, for any $x \in \mathbb{R}^n$, $(x \neq 0) \Rightarrow x^T A\, x > 0$.

- Matrix $A \in \mathbb{R}^{n\times n}$ is PSD $\Leftrightarrow$ all $\lambda_i(A) \geq 0$.

- Matrix $A \in \mathbb{R}^{n\times n}$ is PD $\Leftrightarrow$ all $\lambda_i(A) > 0$.

A Bit More Formal: Vector Spaces

- A vector space over a field $\mathbb{F}$ (*e.g.*, $\mathbb{R}$) is a set $\mathbb{V}$ and a pair of operations, $+ : \mathbb{V} \times \mathbb{V} \rightarrow \mathbb{V}$ and $\cdot : \mathbb{F} \times \mathbb{V} \rightarrow \mathbb{V}$, that satisfy the following axioms, $\forall x, y, z \in \mathbb{V}$ and $\forall \alpha, \beta \in \mathbb{F}$:

  ✓ $+$ is associative and commutative;

  ✓ $\exists 0 \in \mathbb{V}$, such that $0 + x = x$;

  ✓ $\exists -x \in \mathbb{V}$, such that $-x + x = 0$;

  ✓ $\alpha \cdot (\beta \cdot x) = (\alpha \cdot \beta) \cdot x$;

  ✓ $1 \cdot x = x$, where $1 \in \mathbb{F}$ is such that $1 \cdot \alpha = \alpha$;

  ✓ $\alpha \cdot (x + y) = \alpha \cdot x + \alpha \cdot y$;

  ✓ $(\alpha + \beta) \cdot x = \alpha \cdot x + \beta \cdot x$.

- Elements of $\mathbb{V}$ are called vectors; elements of $\mathbb{F}$ are scalars.
- Standard compact notation: $\alpha\, x \equiv \alpha \cdot x$.

Vector Space: Examples

- "Usual vectors" $(\mathbb{R}^n, +, \cdot)$ over field $\mathbb{R}$
  - ✓ $x = (x_1, ..., x_n)$, $y = (y_1, ..., y_n)$, $x + y = (x_1 + y_1, ..., x_n + y_n)$;
  - ✓ $x = (x_1, ..., x_n)$, $\alpha\, x = (\alpha\, x_1, ..., \alpha\, x_n)$
- Real matrices $(\mathbb{R}^{m\times n}, +, \cdot)$ over field $\mathbb{R}$
  - ✓ usual matrix addition and multiplication by scalar;
- Complex matrices $(\mathbb{C}^{m\times n}, +, \cdot)$ over field $\mathbb{C}$ (complex numbers).
- Binary vectors $(\{0,1\}^n, +, \cdot)$ over $GF(2) = \{0,1\}$ (Galois field),
  - ✓ $+$ is *modulo-2* addition: $0+0=0$, $0+1=1+0=1$, $1+1=0$.
  - ✓ $\cdot$ is standard multiplication: $0\cdot 0 = 0\cdot 1 = 1\cdot 0 = 0$, $1\cdot 1 = 1$.
- Set of all functions $f : \Omega \to \mathbb{R}$, with point-wise addition and multiplication, is a vector space over $\mathbb{R}$.

Norm on Vector Space $\mathbb{V}$

- A norm is a function $\|\cdot\| : \mathbb{V} \to \mathbb{R}_+$ satisfying,

  $\forall x, y \in \mathbb{V}$ and $\forall \alpha \in \mathbb{R}$,

  ✓ homogeneity, $\|\alpha\, x\| = |\alpha|\; \|x\|$;

  ✓ triangle inequality, $\|x + y\| \leq \|x\| + \|y\|$;

  ✓ definiteness, $\|x\| = 0 \Leftrightarrow x = 0$.

- A seminorm may not satisfy definiteness.

- Classical example in $\mathbb{R}^n$: Euclidean norm: $\|x\| = \sqrt{\sum_{i=1}^{n} x_i^2}$.

- Two norms $\|\cdot\|$ and $\|\cdot\|'$ are equivalent if $\exists \alpha, \beta > 0$ such that

$$\forall x \in \mathbb{V}, \quad \alpha\|x\| \leq \|x\|' \leq \beta\|x\|;$$

  if $\mathbb{V}$ is finite-dimensional, all norms in $\mathbb{V}$ are equivalent.

Other Norms

- The ℓ_p norm of a vector $x \in \mathbb{R}^n$, where $p \geq 1$,

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}.$$

Notable cases:

  - ℓ_2 (Euclidean) norm.
  - ℓ_1 norm, $\|x\|_1 = \sum_i |x_i|$.
  - ℓ_∞ norm, $\lim_{p\to\infty} \|x\|_p = \max\{|x_1|, ..., |x_n|\} \equiv \|x\|_\infty$
  - ℓ_0 "norm" (not a norm), $\lim_{p\to 0} \|x\|_p = \#\{i : x_i \neq 0\} \equiv \|x\|_0$

- Some equivalences:

$$\begin{array}{ccccc} \|x\|_2 & \leq & \|x\|_1 & \leq & \sqrt{n}\,\|x\|_2 \\ \|x\|_\infty & \leq & \|x\|_2 & \leq & \sqrt{n}\,\|x\|_\infty \\ \|x\|_\infty & \leq & \|x\|_1 & \leq & n\,\|x\|_\infty \end{array}$$

Inner Product on Vector Space $\mathbb{V}$ Over $\mathbb{R}$

- An inner product is a function $\langle \cdot, \cdot \rangle : \mathbb{V} \times \mathbb{V} \to \mathbb{R}$, satisfying, $\forall x, y \in \mathbb{V}$ and $\forall \alpha \in \mathbb{R}$,
  - ✓ symmetry, $\langle x, y \rangle = \langle y, x \rangle$
  - ✓ (bi)linearity, $\langle \alpha x + \beta z, y \rangle = \alpha \langle x, y \rangle + \beta \langle z, y \rangle$;
  - ✓ definiteness, $\langle x, x \rangle \geq 0$ and $\langle x, x \rangle = 0 \Leftrightarrow x = 0$.
- Standard inner product in $\mathbb{R}^n$: $\langle x, y \rangle = \sum_{i=1}^n x_i \, y_i = x^T y$
- Also an inner product in $\mathbb{R}^n$: $\langle x, y \rangle = x^T M y$, where M is PD
- Norm (is it?) induced by an inner product $\|x\| = \sqrt{\langle x, x \rangle}$
  - ✓ for $\langle x, y \rangle = x^T y$, then $\|x\|^2 = x^T x = \sum_{i=1}^n x_i^2$ (Euclidean norm)
  - ✓ for $\langle x, y \rangle = x^T M y$, then $\|x\|_M^2 = x^T M x$ (Mahalanobis norm)

Key Properties of Inner Products

- If $\|\cdot\|$ is induced by inner product $\langle\cdot,\cdot\rangle$ (that is $\|x\|^2 = \langle x, x\rangle$), then

$$\|x+y\|^2 = \langle x+y, x+y\rangle = \|x\|^2 + \|y\|^2 + 2\langle x, y\rangle$$

- Cauchy—Schwarz inequality: $\boxed{|\langle x, y\rangle| \leq \|x\|\,\|y\|}$

  "Des démonstrations qui font boum!" ("Proofs that make boom!")
  [Jean-Baptiste Hiriart-Urruty]

$$0 \leq \frac{1}{2}\left\|\frac{x}{\|x\|} \pm \frac{y}{\|y\|}\right\|^2 = 1 \pm \frac{\langle x, y\rangle}{\|x\|\,\|y\|} \Leftrightarrow \begin{cases} \langle x, y\rangle \leq \|x\|\,\|y\| \\ -\langle x, y\rangle \leq \|x\|\,\|y\| \end{cases}$$

- Corollary: $\|\cdot\|$ is indeed a norm, as it satisfies the triangle inequality:

$$\|x+y\|^2 \leq \|x\|^2 + \|y\|^2 + 2|\langle x, y\rangle| \leq \|x\|^2 + \|y\|^2 + 2\|x\|\,\|y\| = \big(\|x\| + \|y\|\big)^2$$

- Hilbert space: complete vector space equipped with an inner product.

Basis and Dimension of a Vector Space $\mathbb{V}$

- Basis: collection of vectors $B = \{b_1, b_2, ...\} \subset \mathbb{V}$ satisfying:
  - ✓ linear independence: for any finite linear combination
  $$\alpha_1 b_1 + ... + \alpha_m b_m = 0 \;\Rightarrow\; \alpha_1 = \cdots = \alpha_m = 0$$
  - ✓ spanning ability: any vector $v \in \mathbb{V}$ can be written as
  $$v = \alpha_1 b_1 + ... + \alpha_m b_n;$$
  in other words, $\mathbb{V} = \text{span}(B)$.
- Dimension of $\mathbb{V}$: $\dim(\mathbb{V}) = \#B$
- Orthogonal basis: $i \neq j \Rightarrow \langle b_i, b_j \rangle = 0$
- Orthonormal basis: orthogonal and $\|b_i\| = 1$, $\forall\, b_i \in B$.

Rank, Range, and Null Space

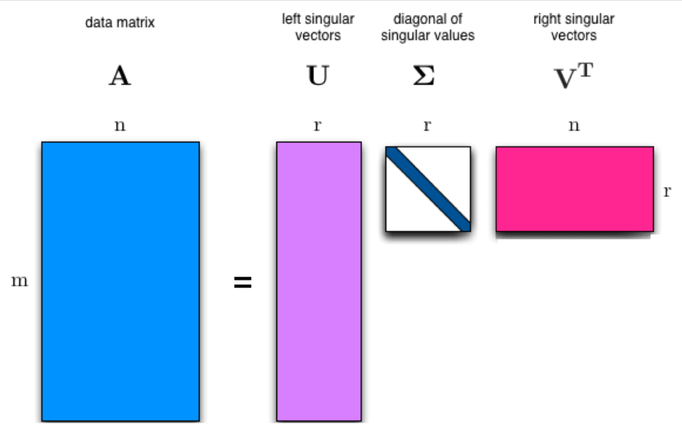
- Consider some real matrix $A \in \mathbb{R}^{m \times n}$
- Range of A: $\mathcal{R}(A) = \big\{ y \in \mathbb{R}^m : \exists\, x \in \mathbb{R}^n \text{ such that } y = Ax \big\} \subseteq \mathbb{R}^m$
- Null space of A: $\mathcal{N}(A) = \big\{ x \in \mathbb{R}^n : Ax = 0 \big\} \subseteq \mathbb{R}^n$.
- Both $\mathcal{R}(A)$ and $\mathcal{N}(A)$ are vector spaces.
- Dimension theorem: $\dim\big(\mathcal{R}(A)\big) + \dim\big(\mathcal{N}(A)\big) = n$
- Rank: $\text{rank}(A) = \dim\big(\mathcal{R}(A)\big) \leq \min\{m, n\}$
- $\text{rank}(A) = n - \dim\big(\mathcal{N}(A)\big)$

Singular Value Decomposition (SVD)

- Any rank-r matrix $A \in \mathbb{R}^{m \times n}$ can be written as $A = U \Lambda V^T$
  - ✓ columns of $U \in \mathbb{R}^{m \times r}$ are an orthonormal basis of $\mathcal{R}(A)$;
  - ✓ columns of $V \in \mathbb{R}^{n \times r}$ are an orthonormal basis of $\mathcal{R}(A^T)$;
  - ✓ $\Lambda = \text{diag}(\sigma_1, ..., \sigma_r)$ is a $r \times r$ diagonal matrix;
  - ✓ $\sigma_1, ..., \sigma_r$ are called singular values.
  - ✓ $\sigma_1, ..., \sigma_r$ are square roots of the eigenvalues of $A^T A$ or $A A^T$.
- Orthonormality of U and V: $U^T U = I$ and $V^T V = I$.
- Transposition: $A^T = (U \Lambda V^T)^T = V \Lambda U^T$.

Singular Value Decomposition (SVD)

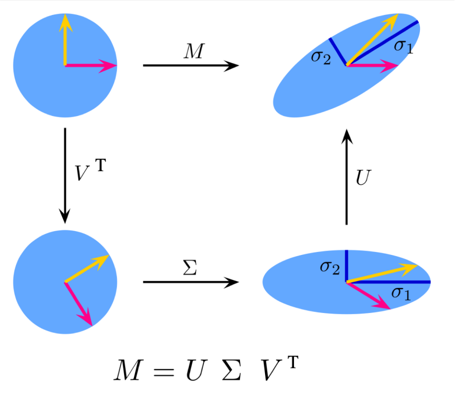
- $A = U\Lambda V^T$, where $U \in \mathbb{R}^{m\times r}$ and $V \in \mathbb{R}^{n\times r}$.



Picture credits: Mukesh Mithrakumar

Singular Value Decomposition (SVD)

- $A = U\Lambda V^T$, where $U \in \mathbb{R}^{m\times r}$ and $V \in \mathbb{R}^{n\times r}$.

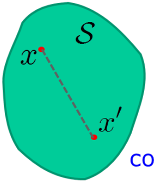


Picture credits: Wikipedia

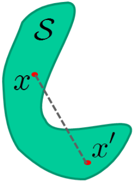
Convex Sets

Convex and strictly convex sets

$\mathcal{S}$ is convex if $\;x, x' \in \mathcal{S} \;\Rightarrow \forall \lambda \in [0,1], \;\; \lambda x + (1-\lambda) x' \in \mathcal{S}$

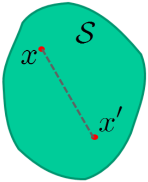


convex

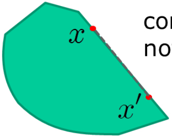


non-convex

$\mathcal{S}$ is strictly convex if $x, x' \in \mathcal{S} \;\Rightarrow \forall \lambda \in (0,1), \;\; \lambda x + (1-\lambda) x' \in \mathrm{int}(\mathcal{S})$



strictly convex

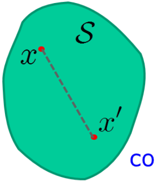


convex, but not strictly

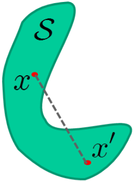
Convex Sets

Convex and strictly convex sets

$\mathcal{S}$ is convex if $x, x' \in \mathcal{S} \;\Rightarrow \forall \lambda \in [0,1], \;\; \lambda x + (1-\lambda) x' \in \mathcal{S}$

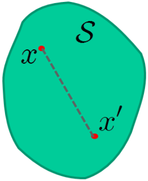


convex

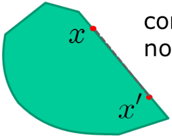


non-convex

$\mathcal{S}$ is strictly convex if $x, x' \in \mathcal{S} \;\Rightarrow \forall \lambda \in (0,1), \;\; \lambda x + (1-\lambda) x' \in \mathrm{int}(\mathcal{S})$



strictly convex



convex, but not strictly

Convex Functions

Convex and strictly convex functions

Extended real valued function: $f : \mathbb{R}^N \rightarrow \bar{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$

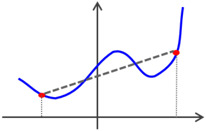
Domain of a function: $\mathrm{dom}(f) = \{x : f(x) \neq +\infty\}$

f is a convex function if

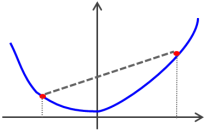
$$\forall \lambda \in [0,1], x, x' \in \mathrm{dom}(f) \;\; f(\lambda x + (1-\lambda)x') \leq \lambda f(x) + (1-\lambda) f(x')$$

f is a strictly convex function if

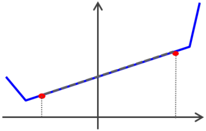
$$\forall \lambda \in (0,1), x, x' \in \mathrm{dom}(f) \;\; f(\lambda x + (1-\lambda)x') < \lambda f(x) + (1-\lambda) f(x')$$



non-convex



convex
strictly convex



convex, not strictly

Recommended Reading

- Z. Kolter and C. Do, "Linear Algebra Review and Reference", Stanford University, 2015 (`https://tinyurl.com/44x2qj4`)

Concluding...

Enjoy LxMLS 2022!