Mathematics Primer

Machine Intelligence

Neural Information Processing Group

Summer Term 2018

Outline

Matrix Multiplication, Transpose and Inverse

Consider matrices $\boldsymbol{A} \in \mathbb{R}^{n \times m}$, $\boldsymbol{B} \in \mathbb{R}^{m \times p}$ with elements $(\boldsymbol{A})_{ij} = a_{ij}$, $(\boldsymbol{B})_{ij} = b_{ij}$.

- The **product** $\boldsymbol{AB} \in \mathbb{R}^{n \times p}$ has elements $(\boldsymbol{AB})_{ij} = \sum_{r=1}^{m} a_{ir} b_{rj}$.
- The **transpose** $\boldsymbol{A}^\top$ has elements $(\boldsymbol{A}^\top)_{ij} = a_{ji}$.
- The **inverse** $\boldsymbol{A}^{-1}$ of a square matrix satisfies $\boldsymbol{AA}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$.
- The following identities hold:

$$(\boldsymbol{AB})^\top = \boldsymbol{B}^\top \boldsymbol{A}^\top \tag{1}$$

$$(\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1} \boldsymbol{A}^{-1} \tag{2}$$

$$(\boldsymbol{A}^T)^{-1} = (\boldsymbol{A}^{-1})^T \tag{3}$$

Rank and Trace

Linear independence

A set of vectors $\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n\}$ is **linearly independent**, if $\sum_{i=1}^{n} \alpha_i \boldsymbol{a}_i = 0$ holds only if all $\alpha_i = 0$. This means none of the vectors can be expressed as a linear combination of the others.

Rank

The **rank** $\text{rank}(\boldsymbol{A})$ of a matrix $\boldsymbol{A}$ is the maximum number of linearly independent rows (or columns).

Trace

The **trace** of a square matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is defined as $\text{Tr}(\boldsymbol{A}) = \sum_{i=1}^{n} a_{ii}$.

It holds:

$$\text{Tr}(\boldsymbol{AB}) = \text{Tr}(\boldsymbol{BA}) \tag{4}$$

Determinant

The **determinant** $\det(\boldsymbol{A})$ shows certain properties of a square matrix $\boldsymbol{A}$

- $\det(\boldsymbol{A}) = 0$ iff the rows (or columns) are linearly dependent
- $\det(\boldsymbol{A}) \neq 0$ iff $\boldsymbol{A}$ is invertible

Note:

- Determinant of the identiy matrix: $\det(\boldsymbol{I}) = 1$
- Determinant of a transposed matrix: $\det(\boldsymbol{A}) = \det(\boldsymbol{A}^T)$
- Determinant of a product of two matrices:

$$\det(\boldsymbol{AB}) = \det(\boldsymbol{A})\det(\boldsymbol{B})$$

Determinant calculation (general)

Calculation of the determinant of a $n \times n$-Matrix $\boldsymbol{A}$:

$$\det(\boldsymbol{A}) = \sum_j A_{ij} C_{ij}.$$

Row i can be any row, the result is always the same. The **cofactors** C_{ij} are defined as $C_{ij} = (-1)^{i+j} \det([\boldsymbol{A}]_{\varnothing ij})$, where $[\boldsymbol{A}]_{\varnothing ij}$ is the submatrix that remains when the i-th row and j-th column are removed:

$$[\boldsymbol{A}]_{\varnothing ij} = \begin{pmatrix} A_{11} & A_{12} & \cdots & \varnothing & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & \varnothing & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \varnothing & \ddots & \vdots \\ \varnothing & \varnothing & \varnothing & \varnothing & \varnothing & \varnothing \\ \vdots & \vdots & \ddots & \varnothing & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & \varnothing & \cdots & A_{nn} \end{pmatrix}$$

Determinant calculation (special cases)

$$|\boldsymbol{A}| = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$$

$$\begin{aligned} |\boldsymbol{A}| &= \begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = a\begin{vmatrix} e & f \\ h & i \end{vmatrix} - b\begin{vmatrix} d & f \\ g & i \end{vmatrix} + c\begin{vmatrix} d & e \\ g & h \end{vmatrix} \\ &= aei + bfg + cdh - ced - bdi - afh \end{aligned}$$

Determinant and Inverse

The **inverse** $\boldsymbol{A}^{-1}$ of a square matrix $\boldsymbol{A}$ exists iff $\det(\boldsymbol{A}) \neq 0$ (matrix not singular).
Calculation of the inverse matrix:

$$\boldsymbol{A}^{-1} = \frac{\mathsf{adj}[\boldsymbol{A}]}{\det(\boldsymbol{A})}$$

where the **adjoint** adj$[A]$ of A is the matrix whose elements are the cofactors:

$$(\mathsf{adj}[\boldsymbol{A}])_{ij} = C_{ji}$$

The determinant of an inverse matrix is given by

$$\det(\boldsymbol{A}^{-1}) = \frac{1}{\det(\boldsymbol{A})}$$

Eigendecomposition of a Matrix

Problem: Find the Eigenvectors and Eigenvalues of a $N \times N$ matrix $\boldsymbol{A}$.

- Consider the system of linear equations:

$$\begin{aligned} \boldsymbol{A}\boldsymbol{x} &= \lambda \boldsymbol{x} \\ (\boldsymbol{A} - \lambda \boldsymbol{I})\boldsymbol{x} &= \boldsymbol{0} \end{aligned}$$

- Solutions: N Eigenvectors $\boldsymbol{x} = \boldsymbol{v}_i$ and corresponding Eigenvalues $\lambda = \lambda_i$
- $\boldsymbol{B}\boldsymbol{x} = \boldsymbol{0}$ has non-trivial solutions iff $\det(\boldsymbol{B}) = 0$
- Therefore, non-trivial λ are the roots of the **characteristic polynomial**:

$$p(\lambda) \equiv \det(\boldsymbol{A} - \lambda \boldsymbol{I}) = 0$$

Eigenvalues and Eigenvectors

Characteristic Equation:

$$p(\lambda) \equiv \det(\boldsymbol{A} - \lambda \boldsymbol{I}) = 0$$

- Polynomial of order N
- N (not necessarily distinct) solutions
- Number of non-zero Eigenvalues: rank$(\boldsymbol{A})$
- In general: Eigenvalues are complex
- For symmetric matrices $(\boldsymbol{A} = \boldsymbol{A}^\top)$: Eigenvalues are real
- Determinant: $\det(\boldsymbol{A}) = \prod_{i=1}^{M} \lambda_i$
- Trace: $\mathrm{Tr}(\boldsymbol{A}) = \sum_{i=1}^{M} \lambda_i$

Matrix Gradient

The **gradient** of a function $f : \mathbb{R}^n \to \mathbb{R}$ is given by

$$\nabla f \equiv \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n} \right)^\top$$

Examples:

- linear $f : \boldsymbol{x} \mapsto \boldsymbol{a}^\top \boldsymbol{x} \qquad \nabla f(\boldsymbol{x}) = \boldsymbol{a}$
- quadratic $f : \boldsymbol{x} \mapsto \boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x} \qquad \nabla f(\boldsymbol{x}) = (\boldsymbol{A}^\top + \boldsymbol{A})\boldsymbol{x}$

Consider a scalar-valued function f of the elements of an $n \times m$ matrix $\boldsymbol{W}$, $f : \boldsymbol{W} \mapsto \mathbb{R}, \quad f(\boldsymbol{W}) = f(w_{11}, \ldots, w_{nm})$.
The **matrix gradient** of f w.r.t. $\boldsymbol{W}$ is defined as

$$\frac{\partial f}{\partial \boldsymbol{W}} = \begin{pmatrix} \frac{\partial f}{\partial w_{11}} & \cdots & \frac{\partial f}{\partial w_{n1}} \\ \vdots & & \vdots \\ \frac{\partial f}{\partial w_{1m}} & \cdots & \frac{\partial f}{\partial w_{nm}} \end{pmatrix}$$

Outline

1. Linear Algebra
   - Transpose, Inverse, Rank and Trace
   - Determinant
   - Eigenanalysis
   - Matrix Gradient
2. Analysis
   - Metrics
   - Jacobi and Hessian
   - Taylor Series
   - Optimization
3. Probability Theory
   - Combinatorics
   - Random Variables and Vectors
   - Conditional Probabilities and Independence
   - Expectations and Moments

Definitions from Functional Analysis

Functions, Functionals and Operators

Two sets $\mathcal{M}$ and $\mathcal{N}$ are connected by a **functional dependency** , if to each $x \in \mathcal{M}$ there corresponds a unique element $y \in \mathcal{N}$. This functional dependency is called

- a **function** if $\mathcal{M}$ and $\mathcal{N}$ are sets of numbers
- a **functional** if $\mathcal{M}$ is a set of functions and $\mathcal{N}$ a set of numbers
- an **operator** if both sets are sets of functions

Example: Linear integral operator T with kernel $K(t,x)$:

$$Tf(x) = \int_q^b K(t,x)f(t)dt$$

Infimum and Supremum

Infimum, Supremum

Let D be a subset of $\mathbb{R}$. A number K is called **supremum** (**infimum**) of D, if K is the smallest upper bound (largest lower bound) of D:

$$x \leq K \ (x \geq K), \forall \, x \in D$$

We write: $\sup D = K \ (\inf D = K)$.

Examples:

- For the closed interval $D = [a, b], a \leq b : \sup D = b, \inf D = a$.
- For $D = \{\frac{n}{n+1}, n \in \mathbb{N}\} : \sup D = 1$.

Metric Space

Metric

A metric (or distance function) on a set X is a non-negative mapping

$$d : X \times X \to \mathbb{R}^+$$

$$(x, y) \mapsto d(x, y)$$

with the following characteristics

1. Positive definiteness: $d(x, y) = 0$ iff $x = y$, $d(x, y) > 0$ otherwise
2. Symmetry: $d(x, y) = d(y, x)$, $\forall\, x, y \in X$
3. Triangle inequality: $d(x, z) \leq d(x, y) + d(y, z)$, $\forall\, x, y, z \in X$

- The pair (X, d) forms a **metric space**
- $d(x, y)$ is called the distance between x and y.

Jacobi and Hessian

- The matrix of the partial derivatives of a vector-valued function $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ is known as **Jacobi matrix** and given by

$$\boldsymbol{J}\boldsymbol{f} \equiv \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}$$

- The square matrix of second-order partial derivatives of a scalar-valued function $f : \mathbb{R}^n \to \mathbb{R}$ is called **Hessian matrix** and given by

$$\boldsymbol{H} f \equiv \frac{\partial^2 f}{\partial \boldsymbol{x}^2} = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}$$

Taylor Series

Taylor Series in $\mathbb{R}$

Let $f : I \to \mathbb{R}$ be an infinitely often differentiable function, and $x_0 \in I$. Then the Taylor series around x_0 is defined as

$$\begin{aligned} f(x) &= \sum_{n=0}^{\infty} \frac{1}{n!} \left. \frac{d^n f(x)}{dx^n} \right|_{x_0} (x - x_0)^n \\ &= f(x_0) + f'(x_0) \cdot (x - x_0) + \frac{1}{2} f''(x_0) \cdot (x - x_o)^2 + \dots \end{aligned}$$

Taylor Series in $\mathbb{R}^n$

Let f be an infinitely smooth scalar-valued function with domain in $\mathbb{R}^n$:

$$f(\boldsymbol{x}) = f(\boldsymbol{x}_0) + \boldsymbol{\nabla} f_{(\boldsymbol{x}_0)}^{\top} (\boldsymbol{x} - \boldsymbol{x}_0) + \frac{1}{2} (\boldsymbol{x} - \boldsymbol{x}_0)^{\top} \boldsymbol{H} f_{(\boldsymbol{x}_0)} (\boldsymbol{x} - \boldsymbol{x}_0) + \dots$$

Local Extrema

Let f be a scalar-valued function $\mathbb{R}^n \to \mathbb{R}$.

Critical Points

A point $\boldsymbol{x}_0$, where $\boldsymbol{\nabla} f(\boldsymbol{x}_0) = 0$ is called a critical point of f.

Local Extrema

A critical point $\boldsymbol{x_0}$ of f is

- a minimum of f, if all Eigenvalues of $(\boldsymbol{H} f)(\boldsymbol{x}_0)$ are positive (the Hessian is **positive definite**)
- a maximum of f, if all Eigenvalues of $(\boldsymbol{H} f)(\boldsymbol{x}_0)$ are negative (the Hessian is **negative definite**)
- no extremum of f, in all other cases (the Hessian is **indefinite**)

Convexity

Convex Functions

Let $U \subset \mathbb{R}^N$ be open and convex. A function $f : U \to \mathbb{R}$ is called (strictly) convex, if for all $\boldsymbol{x}_1, \boldsymbol{x}_2 \in U$ with $\boldsymbol{x}_1 \neq \boldsymbol{x}_2$ and all $0 < \lambda < 1$

$$f(\lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2)(<) \leq \lambda f(\boldsymbol{x}_1) + (1-\lambda) f(\boldsymbol{x}_2)$$

Concave Functions

f is called concave, if $-f$ is convex.

The Lagrange Method (Equality Constraints)

Problem: Maximization of a function $f(\boldsymbol{w})$: $\mathbb{R}^n \to \mathbb{R}$ under some **equality** constraints.

$$\max f(\boldsymbol{w}), \qquad \text{s.t.} \quad g_i(\boldsymbol{w}) = 0, \quad \forall i \in \{1, \ldots, k\}$$

Solution: Form the **Lagrangian**

$$\mathcal{L}(\boldsymbol{w}, \lambda_1, \ldots, \lambda_k) = f(\boldsymbol{w}) + \sum_{i=1}^{k} \lambda_i g_i(\boldsymbol{w}),$$

where $\lambda_1, \ldots, \lambda_k$ are called Lagrange multipliers. Find the stationary points (saddle points) of the Lagrangian w.r.t. both $\boldsymbol{w}$ and all the λ_i:

$$\frac{\partial \mathcal{L}(\boldsymbol{w}, \lambda_1, \ldots, \lambda_k)}{\partial \boldsymbol{w}} = \frac{\partial f(\boldsymbol{w})}{\partial \boldsymbol{w}} + \sum_{i=1}^{k} \lambda_i \frac{\partial g_i(\boldsymbol{w})}{\partial \boldsymbol{w}} = \mathbf{0}$$

and

$$\frac{\partial \mathcal{L}(\boldsymbol{w}, \lambda_1, \ldots, \lambda_k)}{\partial \lambda_i} = g_i(\boldsymbol{w}) = 0, \forall i.$$

The Lagrange Method (Inequality Constraints)

Now: Maximization of a function $f(\boldsymbol{w})$ under some **inequality** constraints.

$$\max f(\boldsymbol{w}), \qquad \text{s.t.} \quad h_i(\boldsymbol{w}) \leq 0, \quad \forall i \in \{1, \ldots, k\}$$

Solution: Find the stationary points of the Lagrangian

$$\mathcal{L}(\boldsymbol{w}, \lambda_1, \ldots, \lambda_k) = f(\boldsymbol{w}) + \sum_{i=1}^{k} \lambda_i \, h_i(\boldsymbol{w}),$$

w.r.t. $\boldsymbol{w}$ under the constraints

$$h_i(\boldsymbol{w}) \leq 0, \forall i$$

$$\lambda_i \geq 0, \forall i$$

$$\lambda_i \cdot h_i(\boldsymbol{w}) = 0, \forall i,$$

which are known as the **Karush-Kuhn-Tucker (KKT) conditions**.

Outline

Combinatorics

Consider a set consisting of n elements. The **power set** is the set of all subsets, its cardinality is 2^n.

- **Permutation:** arrangement of n elements in a certain order
  - # **without** repetitions: $P_n = n!$
  - # **with** repetitions ($k \leq n$ repeated elements): $P_n^{(k)} = \frac{n!}{k!}$
- **Combination:** choice of k out of n elements regardless of order
  - # **without** repetitions: $C_n^{(k)} = \binom{n}{k} = \frac{n!}{k!(n-k)!}$
  - # **with** repetitions: $C_n^{(k)} = \binom{n+k-1}{k}$
- **Variation:** choice of k out of n elements taking their order into account
  - # **without** repetitions $V_n^{(k)} = k!\binom{n}{k}$
  - # **with** repetitions: $V_n^{(k)} = n^k$

Random Variable

Consider a set Ω of elementary events w, e.g. all possible outcomes of an experiment. The mapping

$$\Omega \to R \subset \mathbb{R}$$

$$w \to X(w) \equiv X$$

is called a **random variable**.

- If R consists of a finite or countable infinite number of elements, then X is called a **discrete** random variable.
- If $R = \mathbb{R}$ or R consists of intervals from $\mathbb{R}$, then X is called a **continuous** random variable.

Example: Roll dice
w_1: 1 comes up $\to X(w_1) = 1, \ldots, w_6$: 6 comes up $\to X(w_6) = 6$

Distribution of a Random Variable

The **cumulative distribution function (cdf)** or simply **distribution function** of a random variable X at point z is defined as the probability that $X \leq z$:

$$F_X(z) = P\left(X \leq z\right)$$

- Allowing z to vary in $(-\infty, \infty)$ defines the cdf for all values of X.
- $0 \leq F_X \leq 1$, a nondecreasing and continuous function for continuous X.

Example: Roll ideal dice, where $P(X = i) = \frac{1}{6}\ \forall i$

$$F_X(z) = \begin{cases} 0 & \text{for } z < 1 \\ 1/6 & \text{for } 1 \leq z < 2 \\ 2/6 & \text{for } 2 \leq z < 3 \\ & \dots \\ 1 & \text{for } z \geq 6 \end{cases}$$

Probability Density of a Continuous Variable

The **probability density function (pdf)** p_X of a continuous X is obtained as the derivative of its cdf:

$$p_X(z) = \left. \frac{dF_X(x)}{dx} \right|_{x=z}$$

In practice, the cdf is computed from the known pdf using the inverse relationship

$$F_X(z) = \int_{-\infty}^{z} p_X(t)dt$$

Example: Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$

- **cdf** $\quad F(z) \equiv P(X \leq z) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{d}x$
- **pdf** $\quad p(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2\sigma^2}}$

Distribution of a Random Vector

The **distribution function** of a **random vector** $\boldsymbol{X}$:

$$\Omega \rightarrow R^N \subset \mathbb{R}^N$$

$$w \rightarrow \boldsymbol{X}(w) \equiv \boldsymbol{X}$$

at a point $\boldsymbol{z}$ is given by

$$F_{\boldsymbol{X}}(\boldsymbol{z}) = P\left(\boldsymbol{X} \leq \boldsymbol{z}\right)$$

Distribution of a Random Vector

Example

Toss a German 2 Euro and a German 20 Cent coin.

- $w_1 = \{$2 Euro: eagle, 20 Cent: gate$\} \rightarrow \boldsymbol{X}(w_1) = (1, 1)^\top$
- $w_2 = \{$2 Euro: eagle, 20 Cent: number$\} \rightarrow \boldsymbol{X}(w_2) = (1, 2)^\top$
- $w_3 = \{$2 Euro: number, 20 Cent: gate$\} \rightarrow \boldsymbol{X}(w_3) = (2, 1)^\top$
- $w_4 = \{$2 Euro: number, 20 Cent: number$\} \rightarrow \boldsymbol{X}(w_4) = (2, 2)^\top$

$$F_{\boldsymbol{X}}(\boldsymbol{z}) = \begin{cases} 0 & \text{for} \quad (z_1 < 1) \quad \vee \quad (z_2 < 1) \\ 1/4 & \text{for} \quad (1 \leq z_1 < 2) \quad \wedge \quad (1 \leq z_2 < 2) \\ 1/2 & \text{for} \quad (1 \leq z_1 < 2) \quad \wedge \quad (2 \leq z_2) \\ 1/2 & \text{for} \quad (2 \leq z_1) \quad \wedge \quad (1 \leq z_2 < 2) \\ 1 & \text{for} \quad (2 \leq z_1) \quad \wedge \quad (2 \leq z_2) \end{cases}$$

Conditional Probabilities

Conditional Probabilities

Consider two discrete random variables X and Y. The conditional probability of Y given X:

$$P(Y = y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)}, \quad P(X = x) \neq 0$$

Conditional Probability Densities

Consider two continuous random vectors $\boldsymbol{X}$, $\boldsymbol{Y}$ and their joint probability density. The conditional probability density of $\boldsymbol{Y}$ given $\boldsymbol{X}$: Probability for finding $\boldsymbol{Y} \in [\boldsymbol{y}, \boldsymbol{y} + \boldsymbol{dy}]$ if we already know that $\boldsymbol{X} \in [\boldsymbol{x}, \boldsymbol{x} + \boldsymbol{dx}]$.

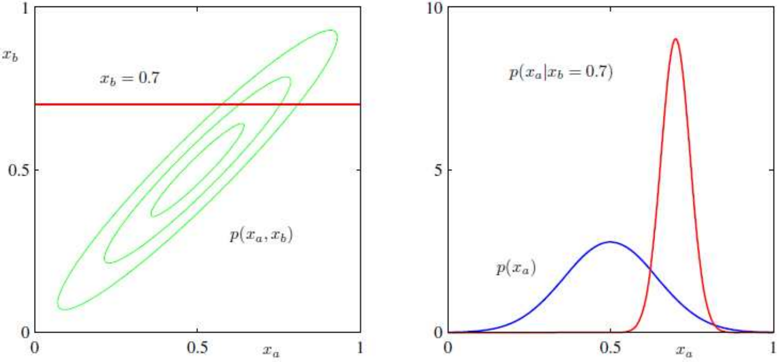
$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{x})} \qquad \text{almost everywhere in } \boldsymbol{X}$$

Independence

Statistical Independence of Continuous Random Vectors

The random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ are statistically independent iff

$$p(\boldsymbol{y}|\boldsymbol{x}) = p(\boldsymbol{y}) \quad \text{or equivalently} \quad p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{x})p(\boldsymbol{y})$$



Marginals

Law of Total Probability (Discrete Random Variables)

Marginalisation over Y:

$$P(X = x) = \sum_k P(X = x, Y = y_k)$$

Marginal Densities (Continuous Random Vectors)

Given the joint density $p_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x}, \boldsymbol{y})$ of two random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$, the marginal density $p_{\boldsymbol{X}}(\boldsymbol{x})$ is obtained by integrating over the other random vector:

$$p_{\boldsymbol{X}}(\boldsymbol{x}) = \int_{-\infty}^{\infty} p_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x}, \tilde{\boldsymbol{y}}) d\tilde{\boldsymbol{y}}$$

Bayes' Theorem

Bayes' Theorem (Discrete Random Variables)

$$
\begin{aligned}
P(Y = y|X = x) &= \frac{P(X = x|Y = y)P(Y = y)}{P(X = x)} \\
&= \frac{P(X = x|Y = y)P(Y = y)}{\sum_k P(X = x|Y = y_k)P(Y = y_k)}
\end{aligned}
$$

Bayes' Theorem (Continuous Random Vectors)

$$
p(\boldsymbol{y}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{y})}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{y})}{\int p(\boldsymbol{x}|\tilde{\boldsymbol{y}})p(\tilde{\boldsymbol{y}})d\tilde{\boldsymbol{y}}}
$$

Decomposition

Factorization of a joint pdf (or cdf), as given by the Chain Rule:

$$p(x_1, \ldots, x_d) = p(x_1)p(x_2|x_1)\ldots p(x_d|x_1, \ldots, x_{d-1})$$

- Special case: Statistical Independence

$$p(x_1, \ldots, x_d) = p(x_1)p(x_2)\ldots p(x_d) = \prod_{k=1}^{d} p(x_k)$$

- Special case: 1st order Markov chain

$$p(x_1, \ldots, x_d) = p(x_d|x_{d-1})p(x_{d-1}|x_{d-2})\ldots p(x_2|x_1)p(x_1)$$

Expectations

- In Practice: Probability density usually unknown
- However: Expectations of functions can be directly estimated from the data

The expectation of a scalar-, vector- or matrix-valued function $\boldsymbol{g}(\boldsymbol{X})$ of a random vector $\boldsymbol{X}$, as defined below, can be estimated from a dataset of k **i.i.d.** samples $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(k)}$:

$$\langle \boldsymbol{g}(\boldsymbol{X}) \rangle \equiv \int_{-\infty}^{\infty} \boldsymbol{g}(\boldsymbol{x})\, p_{\boldsymbol{X}}(\boldsymbol{x})\, d\boldsymbol{x} \approx \frac{1}{k} \sum_{j=1}^{k} \boldsymbol{g}(\boldsymbol{x}^{(j)})$$

- Linearity: $\langle a\boldsymbol{X} + b\boldsymbol{X} + c \rangle = a\langle \boldsymbol{X} \rangle + b\langle \boldsymbol{Y} \rangle + c$
- $p_{\boldsymbol{X}}$ known $\Rightarrow$ Expectations of arbitrary function available
- Expectations for all functions f known $\Rightarrow$ $p_{\boldsymbol{x}}$ can be determined $\Rightarrow$ Statistics of $\boldsymbol{X}$ completely known

Moments

Moments of a random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ are typical expectations used to characterize it. They are obtained when $\boldsymbol{g}(\boldsymbol{X})$ consists of products of components of $\boldsymbol{X}$.

Examples:

- 1st order: $\langle X_i \rangle = \int p(x_i)\, x_i\, dx_i$ ... mean value μ_i, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$
- 2nd order: $\langle X_i X_j \rangle$ ... correlation between X_i, X_j
- 3rd order: $\langle X_i X_j X_k \rangle$ ... e.g. skewness

Correlation Matrix

The correlation matrix of a random vector $\boldsymbol{X}$ contains all second order moments $\langle X_i X_j \rangle$:

$$\boldsymbol{R_X} = \langle \boldsymbol{X} \boldsymbol{X}^\top \rangle$$

- Symmetry: $\boldsymbol{R_X} = \boldsymbol{R_X}^\top$
- Positive semidefinite: $\boldsymbol{a}^\top \boldsymbol{R_X} \boldsymbol{a} \geq 0, \ \forall \boldsymbol{a}$
  $\Rightarrow$ all eigenvalues real and nonnegative
  $\Rightarrow$ all eigenvectors are mutually orthogonal

Covariance Matrix

The covariance matrix of a random vector $\boldsymbol{X}$ is given by

$$\boldsymbol{C_X} \equiv \langle(\boldsymbol{X}-\boldsymbol{\mu_X})(\boldsymbol{X}-\boldsymbol{\mu_X})^\top\rangle = \langle\boldsymbol{X}\boldsymbol{X}^\top\rangle - \boldsymbol{\mu_X}\boldsymbol{\mu_X}^\top = \boldsymbol{R_X} - \boldsymbol{\mu_X}\boldsymbol{\mu_X}^\top$$

and the components C_{ij} are calculated as

$$C_{ij} = \langle X_i X_j\rangle - \mu_i\mu_j = \iint p(x_i, x_j)\, x_i\, x_j\, dx_i\, dx_j - \mu_i\mu_j.$$

- $C_{ii} = \sigma_i^2$ ... variance of X_i
- For zero mean, correlation and covariance matrix are identical

Uncorrelatedness and Independence

Two random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ are **uncorrelated** iff their cross-covariance matrix $\boldsymbol{C}_{\boldsymbol{XY}} = \langle \boldsymbol{X}\boldsymbol{Y}^\top \rangle - \boldsymbol{\mu}_{\boldsymbol{X}}\boldsymbol{\mu}_{\boldsymbol{Y}} = \mathbf{0}$.

- **Uncorrelatedness** implies that

$$\boldsymbol{R}_{\boldsymbol{XY}} = \langle \boldsymbol{X}\boldsymbol{Y}^\top \rangle = \langle \boldsymbol{X} \rangle \langle \boldsymbol{Y}^\top \rangle = \boldsymbol{\mu}_{\boldsymbol{X}}\boldsymbol{\mu}_{\boldsymbol{Y}}{}^\top,$$

  while **independence** implies that

$$\langle \boldsymbol{g}(\boldsymbol{X})\boldsymbol{h}(\boldsymbol{Y}) \rangle = \langle \boldsymbol{g}(\boldsymbol{X}) \rangle \langle \boldsymbol{h}(\boldsymbol{Y}) \rangle \quad \text{for any } \boldsymbol{g}, \boldsymbol{h}$$

  $\Rightarrow$ Independence much stronger property than uncorrelatedness
- Special property of **Gaussian distributions**:
  uncorrelatedness = independence