Mathematics Primer

Machine Intelligence

Neural Information Processing Group

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
4. References

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
4. References

Matrix Multiplication, Transpose and Inverse

Consider matrices $\underline{\mathbf{A}} \in \mathbb{R}^{N\times M}$, $\underline{\mathbf{B}} \in \mathbb{R}^{M\times p}$ with elements $(\underline{\mathbf{A}})_{ij} = a_{ij}$, $(\underline{\mathbf{B}})_{ij} = b_{ij}$.

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

$$(\underline{\mathbf{A}}\,\underline{\mathbf{B}})^\top = \underline{\mathbf{B}}^\top\underline{\mathbf{A}}^\top$$

$$(\underline{\mathbf{A}}\,\underline{\mathbf{B}})^{-1} = \underline{\mathbf{B}}^{-1}\underline{\mathbf{A}}^{-1}$$

$$(\underline{\mathbf{A}}^\top)^{-1} = (\underline{\mathbf{A}}^{-1})^\top$$

Rank and Trace

Linear independence

A set of vectors $\{\underline{\mathbf{a}}_1, \ldots, \underline{\mathbf{a}}_N\}$ is **linearly independent**, if $\sum_{i=1}^{N} \alpha_i \underline{\mathbf{a}}_i = 0$ 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** $\mathrm{rank}(\underline{\mathbf{A}})$ of a matrix $\underline{\mathbf{A}}$ is the maximum number of linearly independent rows (or columns).

Trace

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

It holds:

$$\mathrm{Tr}(\underline{\mathbf{A}}\,\underline{\mathbf{B}}) = \mathrm{Tr}(\underline{\mathbf{B}}\,\underline{\mathbf{A}})$$

Determinant

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

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

Note:

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

$$\det(\underline{\mathbf{A}}\,\underline{\mathbf{B}}) = \det(\underline{\mathbf{A}})\det(\underline{\mathbf{B}})$$

Determinant calculation (general)

Calculation of the determinant of an $N \times N$-Matrix $\underline{\mathbf{A}}$:

$$\det(\underline{\mathbf{A}}) = \sum_j a_{ij} C_{ij}.$$

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

$$[\underline{\mathbf{A}}]_{\varnothing ij} = \begin{pmatrix} a_{11} & a_{12} & \cdots & \varnothing & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & \varnothing & \cdots & a_{2N} \\ \vdots & \vdots & \ddots & \varnothing & \ddots & \vdots \\ \varnothing & \varnothing & \varnothing & \varnothing & \varnothing & \varnothing \\ \vdots & \vdots & \ddots & \varnothing & \ddots & \vdots \\ a_{N1} & a_{N2} & \cdots & \varnothing & \cdots & a_{NN} \end{pmatrix}$$

Determinant calculation (special cases)

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

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

Determinant and Inverse

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

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

where the **adjoint** $\mathsf{adj}[\underline{\mathbf{A}}]$ of $\underline{\mathbf{A}}$ is the matrix whose elements are the cofactors:

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

The determinant of an inverse matrix is given by

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

Eigendecomposition of a Matrix

Problem: Find the Eigenvectors and Eigenvalues of an $N \times N$ matrix $\underline{\mathbf{A}}$.

- Consider the system of linear equations:

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

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

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

Eigenvalues and Eigenvectors

Characteristic Equation:

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

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

Eigendecomposition of a Matrix in $\mathbb{R}^3$

Example

$$\underline{\mathbf{A}} = \begin{pmatrix} 0 & -1 & 1 \\ -3 & -2 & 3 \\ -2 & -2 & 3 \end{pmatrix}$$

- Eigenvalues: $\det(\underline{\mathbf{A}} - \lambda \underline{\mathbf{I}}) = -\lambda^3 + \lambda^2 + \lambda - 1 = 0$
  $\Rightarrow \lambda_1 = 1, \lambda_2 = 1, \lambda_3 = -1$
- Find each eigenvector $\underline{\mathbf{x}} = \underline{\mathbf{v}}_i$ associated with each eigenvalue λ_i:

$$(\underline{\mathbf{A}} - \lambda_i \underline{\mathbf{I}})\underline{\mathbf{x}} = \begin{pmatrix} 0 - \lambda_i & -1 & 1 \\ -3 & -2 - \lambda_i & 3 \\ -2 & -2 & 3 - \lambda_i \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = 0$$

$\Rightarrow \lambda_1, \lambda_2$ Eigenspace: $\{(x_1, x_2, x_3) | -x_1 - x_2 + x_3 = 0\}$
$\Rightarrow \lambda_3$ Eigenspace: $\{(t, 3t, 2t) | t \in \mathbb{R}\}$

Eigendecomposition of a Matrix in $\mathbb{R}^3$

Example

$$\underline{\mathbf{A}} = \begin{pmatrix} 0 & -1 & 1 \\ -3 & -2 & 3 \\ -2 & -2 & 3 \end{pmatrix}$$

- The rank of $\underline{\mathbf{A}}$ is $rank(\underline{\mathbf{A}}) = 3$
- The number of non-zero Eigenvalues is $3\checkmark$
- The trace of $\underline{\mathbf{A}}$ is $\text{Tr}(\underline{\mathbf{A}}) = 0 - 2 + 3 = 1$
- The sum of the Eigenvalues is $-1 + 1 + 1 = 1\checkmark$
- The determinant of $\underline{\mathbf{A}}$ is $\det(\underline{\mathbf{A}}) = -1$
- The product of the Eigenvalues is $(-1) \cdot 1 \cdot 1 = -1\checkmark$

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

Consider a <u>scalar-valued</u> function f of the elements of an $N \times M$ matrix $\underline{\mathbf{W}}$, $f : \underline{\mathbf{W}} \mapsto \mathbb{R}\,, \quad f(\underline{\mathbf{W}}) = f(w_{11}, \ldots, w_{NM})$.
The **matrix gradient** of f w.r.t. $\underline{\mathbf{W}}$ is defined as

$$\frac{\partial f}{\partial \underline{\mathbf{W}}} = \begin{pmatrix} \frac{\partial f}{\partial w_{11}} & \cdots & \frac{\partial f}{\partial w_{N1}} \\ \vdots & & \vdots \\ \frac{\partial f}{\partial w_{1M}} & \cdots & \frac{\partial f}{\partial w_{NM}} \end{pmatrix}$$

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
4. References

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 $\underline{\mathbf{f}} : \mathbb{R}^N \to \mathbb{R}^M$ is known as **Jacobi matrix** and is given by

$$\underline{\mathbf{J}}_{\underline{\mathbf{f}}} \equiv \frac{\partial \underline{\mathbf{f}}}{\partial \underline{\mathbf{x}}} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_N} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_M}{\partial x_1} & \cdots & \frac{\partial f_M}{\partial x_N} \end{pmatrix}$$

- 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 is given by

$$\underline{\mathbf{H}}_f \equiv \frac{\partial^2 f}{\partial \underline{\mathbf{x}}^2} = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_N} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_N} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_N \partial x_1} & \frac{\partial^2 f}{\partial x_N \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_N^2} \end{pmatrix}$$

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

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

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

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

$$f(\underline{\mathbf{x}}) = f(\underline{\mathbf{x}}_0) + \underline{\nabla} f^{\top}_{(\underline{\mathbf{x}}_0)} (\underline{\mathbf{x}} - \underline{\mathbf{x}}_0) + \frac{1}{2} (\underline{\mathbf{x}} - \underline{\mathbf{x}}_0)^{\top} \underline{\mathbf{H}}_{f(\boldsymbol{x}_0)} (\underline{\mathbf{x}} - \underline{\mathbf{x}}_0) + \ldots$$

Local Extrema

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

Critical Points

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

Local Extrema

A critical point $\underline{\mathbf{x}}_{\mathbf{0}}$ of f is

- a minimum of f, if all Eigenvalues of $(\underline{\mathbf{H}}_f)(\underline{\mathbf{x}}_0)$ are positive (the Hessian is **positive definite**)
- a maximum of f, if all Eigenvalues of $(\underline{\mathbf{H}}_f)(\underline{\mathbf{x}}_0)$ 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 $\underline{\mathbf{x}}_1, \underline{\mathbf{x}}_2 \in U$ with $\underline{\mathbf{x}}_1 \neq \boldsymbol{x}_2$ and all $0 < \lambda < 1$

$$f\left(\lambda \underline{\mathbf{x}}_1 + (1-\lambda)\underline{\mathbf{x}}_2\right)(<) \leq \lambda f(\underline{\mathbf{x}}_1) + (1-\lambda) f(\underline{\mathbf{x}}_2)$$

Concave Functions

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

The Lagrange Method (Equality Constraints)

Problem: Maximization of a function $f(\underline{\mathbf{w}})$: $\mathbb{R}^N \rightarrow \mathbb{R}$ under some **equality** constraints $g_i(\underline{\mathbf{w}}) = 0 \; \forall i \in \{1, \dots, k\}$.

$$f(\underline{\mathbf{w}}) \overset{!}{=} \max, \qquad \text{s.t.} \quad g_i(\underline{\mathbf{w}}) = 0, \quad \forall i \in \{1, \dots, k\}$$

Solution: Form the **Lagrangian**

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

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

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

and

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

The Lagrange Method (Inequality Constraints)

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

$$f(\underline{\mathbf{w}}) \overset{!}{=} \max, \qquad \text{s.t.} \quad h_i(\underline{\mathbf{w}}) \leq 0, \quad \forall i \in \{1, \dots, k\}$$

Solution: Find the stationary points of the Lagrangian

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

w.r.t. $\underline{\mathbf{w}}$ under the constraints

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

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

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

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

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
4. References

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

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

$$w \rightarrow X(w) \equiv X$$

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 $\rightarrow X(w_1) = 1$, $\ldots$, w_6: 6 comes up $\rightarrow X(w_6) = 6$

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

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

Distribution of a Random Vector

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

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

$$\underline{\mathbf{w}} \to \underline{\mathbf{X}}(\underline{\mathbf{w}}) \equiv \underline{\mathbf{X}}$$

at a point $\underline{\mathbf{z}}$ is given by

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

Distribution of a Random Vector
Example

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

- $\underline{\mathbf{w}}^{(1)} = \{$2 Euro: eagle, 20 Cent: gate$\} \rightarrow \underline{\mathbf{X}}(\underline{\mathbf{w}}^{(1)}) = (1,1)^\top$
- $\underline{\mathbf{w}}^{(2)} = \{$2 Euro: eagle, 20 Cent: number$\} \rightarrow \underline{\mathbf{X}}(\underline{\mathbf{w}}^{(2)}) = (1,2)^\top$
- $\underline{\mathbf{w}}^{(3)} = \{$2 Euro: number, 20 Cent: gate$\} \rightarrow \underline{\mathbf{X}}(\underline{\mathbf{w}}^{(3)}) = (2,1)^\top$
- $\underline{\mathbf{w}}^{(4)} = \{$2 Euro: number, 20 Cent: number$\} \rightarrow \underline{\mathbf{X}}(\underline{\mathbf{w}}^{(4)}) = (2,2)^\top$

$$F_{\underline{\mathbf{X}}}(\underline{\mathbf{z}}) = \begin{cases} 0 & \text{for} \quad (z_1 < 1) \vee (z_2 < 1) \\ 1/4 & \text{for} \quad (1 \leq z_1 < 2) \wedge (1 \leq z_2 < 2) \\ 1/2 & \text{for} \quad (1 \leq z_1 < 2) \wedge (2 \leq z_2) \\ 3/4 & \text{for} \quad (2 \leq z_1) \wedge (1 \leq z_2 < 2) \\ 1 & \text{for} \quad (2 \leq z_1) \wedge (2 \leq z_2) \end{cases}$$

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 $\underline{\mathbf{X}}$, $\underline{\mathbf{Y}}$ and their joint probability density. The conditional probability density of $\underline{\mathbf{Y}}$ given $\underline{\mathbf{X}}$: Probability for finding $\underline{\mathbf{Y}} \in [\underline{\mathbf{y}}, \underline{\mathbf{y}} + d\underline{\mathbf{y}}]$ if we already know that $\underline{\mathbf{X}} \in [\underline{\mathbf{x}}, \underline{\mathbf{x}} + d\underline{\mathbf{x}}]$.

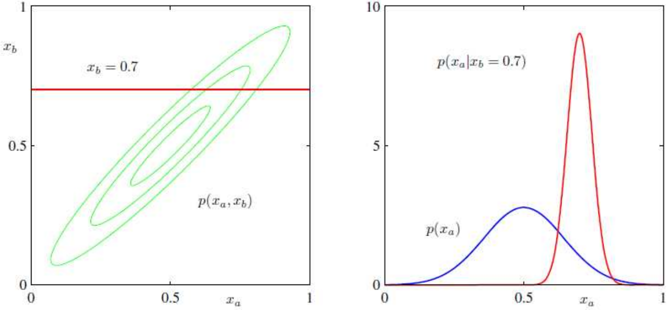
$$p(\underline{\mathbf{y}}|\underline{\mathbf{x}}) = \frac{p(\underline{\mathbf{x}}, \underline{\mathbf{y}})}{p(\underline{\mathbf{x}})} \qquad \text{almost everywhere in } \underline{\mathbf{X}}$$

Independence

Statistical Independence of Continuous Random Vectors

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

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



source: Bishop, 2006, Ch. 2.3.2

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

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

Bayes' Theorem

Bayes' Theorem (Discrete Random Variables)

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

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

Bayes' Theorem (Continuous Random Vectors)

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

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 $\underline{\mathbf{g}}(\underline{\mathbf{X}})$ of a random vector $\underline{\mathbf{X}}$, as defined below, can be estimated from a dataset of k **i.i.d.** samples $\underline{\mathbf{x}}^{(1)}, \underline{\mathbf{x}}^{(2)}, \ldots, \underline{\mathbf{x}}^{(k)}$:

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

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

Moments

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

Examples:

- 1st order: $\langle X_i \rangle = \int p(x_i)\, x_i\, dx_i$ ... mean value μ_i, $\underline{\mu} = (\mu_1, \ldots, \mu_n)$
- 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 $\underline{\mathbf{X}}$ contains all second order moments $\langle X_i X_j \rangle$:

$$\underline{\mathbf{R}}_{\underline{\mathbf{X}}} = \langle \underline{\mathbf{X}}\, \underline{\mathbf{X}}^\top \rangle$$

- Symmetry: $\underline{\mathbf{R}}_{\underline{\mathbf{X}}} = \underline{\mathbf{R}}_{\underline{\mathbf{X}}}^\top$
- Positive semidefinite: $\underline{\mathbf{a}}^\top \underline{\mathbf{R}}_{\underline{\mathbf{X}}} \underline{\mathbf{a}} \geq 0,\ \forall \underline{\mathbf{a}}$
  $\Rightarrow$ all eigenvalues real and nonnegative
  $\Rightarrow$ all eigenvectors are mutually orthogonal

Covariance Matrix

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

$$\underline{\mathbf{C}}_{\underline{\mathbf{X}}} \equiv \langle (\underline{\mathbf{X}} - \underline{\mu}_{\underline{\mathbf{X}}})(\underline{\mathbf{X}} - \underline{\mu}_{\underline{\mathbf{X}}})^{\top} \rangle = \langle \underline{\mathbf{X}}\, \underline{\mathbf{X}}^{\top} \rangle - \underline{\mu}_{\underline{\mathbf{X}}} \underline{\mu}_{\underline{\mathbf{X}}}^{\top} = \underline{\mathbf{R}}_{\underline{\mathbf{X}}} - \underline{\mu}_{\underline{\mathbf{X}}} \underline{\mu}_{\underline{\mathbf{X}}}^{\top}$$

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 (“centered”), the correlation and covariance matrices are identical.

Uncorrelatedness and Independence

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

- **Uncorrelatedness** implies that

  $$\underline{\mathbf{R}}_{\underline{\mathbf{X}}\,\underline{\mathbf{Y}}} = \langle \underline{\mathbf{X}}\,\underline{\mathbf{Y}}^\top \rangle = \langle \underline{\mathbf{X}} \rangle \langle \underline{\mathbf{Y}}^\top \rangle = \underline{\mu}_{\underline{\mathbf{X}}}\underline{\mu}_{\underline{\mathbf{Y}}}^\top,$$

  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

References

Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.