CHAPTER 1

Preliminary Mathematical Concepts
Chapter 1: Preliminary Mathematical Concepts

Mathematics provides the powerful tools for establishing any theory and idea. The mathematical modeling helps to define the neural networks and their functionalities. The basic operations of the neurons and its topological structures can easily specify with the help of mathematical modeling. The input/output patterns for the neural network can represent with vector and the various learning methods of neural networks are define to employ the mathematical tools like differential equation, gradient methods, matrix operations and many more. In this chapter we are discussing [1-7] the basic preliminaries of mathematical tools, those are used very frequently to understand the neural network architecture and its learning techniques accomplish the task of pattern recognition.

1.1 Sets

A set is a collection of objects or elements, sharing the common characteristics or properties. A set is denoted by listing the elements between braces. The set of positive integers is \( \{1,2,3,\ldots\} \). We also denote a sets with the notation \( \{x \mid \text{conditions on } x\} \) for sets that are more easily described than enumerated. This is read as “the set of elements \( x \) such that \( x \) satisfies…” \( x \in S \) is the notation for “\( x \) is an element of the set \( S \)”. To express the opposite we have \( x \notin S \) for “\( x \) is not an element of the set \( S \)”.

The following is the summary to define the sets adequately:

- The **empty set** is defined as a set with nothing. It may be denoted as \( \{ \} \) or \( \emptyset \).

- Two sets \( A \) and \( B \) are **equal** if they contain exactly the same elements.

- Two sets \( A \) and \( B \) are said be **disjoint**, if they have no elements in common.
• A is **subset** of B if every element in A also in B, it is denoted as \( A \subseteq B \).

• The cardinal number or cardinality of a finite set is non-negative integer representing the number of elements in the set. This is denoted as \( |A| \).

• The **complement** of a set A is defined as a set of all elements (of a special set S, the space or universe) that are not in A and it is denoted as \( \overline{A} \).

• **Intersection** of set A and B is defined as:

\[
A \cap B = \left\{ x \in A \text{ and } x \in B \right\}.
\]

• **Union** of two sets A and B is defined as

\[
A \cup B = \left\{ x \in A \text{ or } x \in B \right\}
\]

• **Difference** of two sets A and B is defined as

\[
A - B = \left\{ x \in A \text{ and } x \in B \right\}
\]

• A partition P of a set A is a collection of mutually exclusive subsets of A, that satisfy:

\[ A_i \cap A_j \text{ unless } i = j \text{ and } \bigcup_i A_i = A \]

1.2 Relations

Relations are based on the notion of set mapping and provide a mathematical formalism for the representation of structure.
If A and B are sets, a relation from A to B is a subset of $A \times B$. Where $A \times B$ denotes the Cartesian product of the set A and B. Given a set

$$A = \{a, b, c, \ldots\} \quad \text{and} \quad B = \{x, y, z, \ldots\},$$

a relation from A to B, namely R, satisfies $R \subseteq A \times B$.

This is defined as a binary relation, since it involves only two sets and provides a way of connecting or relating members of the sets. The relations can also be defined as sets of ordered pair say $(x, y)$, here set of all possible values of $x$ is the **domain** and the set of all possible values of $y$ is called the **range**. The relations have a direction or ordering, this directionality and ordering of a relation defines the following important properties:

- **Reflexive**: $R$ is reflexive if, $\forall a \in A, (a, a) \in R$.

- **Symmetric**: $R$ is symmetric if, $\forall (a, b) \in R, (b, a) \in R$.

- **Transitive**: $R$ is transitive if, $\forall (a, b) \in R \quad \text{and} \quad (b, c) \in R \quad \text{then} \quad (a, c) \in R$

- **Equivalence Relations**: A relation that satisfies all three above mentioned properties, is termed as an equivalence relation.

### 1.3 Functions

A function $f$ from A to B is a relation such that for every $a \in A, \exists$ one and only one $b \in B, \text{such that} \quad (a, b) \in f$. Usually it has been shown as:

$$f : A \Rightarrow B,$$
where A is the domain of function \( f \) and B is the range of the function. If \((a, b) \in f\), we say that \( b \) is the function value of \( a \), and denoted as:

\[ b = f(a) \]

If \( b = f(a) \), then we can write \( a = f^{-1}(b) \) where \( f^{-1} \) is the inverse of \( f \). If \( b = f(a) \) is a one-to-one function, then \( f^{-1}(b) \) is also one-to-one function. In this case, \( a = f^{-1}(f(a)) = f(f^{-1}(a)) \) for each value of \( a \) where both \( f(a) \) and \( f^{-1}(a) \) are defined. If \( b = f(a) \) is a many-to-one function, then \( a = f^{-1}(b) \) is a one-to-many function. \( f^{-1}(b) \) is a multi-valued function.

### 1.4 Scalars and Vectors

A vector is a quantity having both magnitude and a direction. Examples of vector quantities are velocity, force and position. One can represent a vector in \( n \)-dimensional space with an arrow whose initial point is at the origin (figure 1.1). The magnitude is the length of the vector.

![Graphical representation of a vector in three dimensions.](image)

**Figure 1.1:** Graphical representation of a vector in three dimensions.
A scalar has only a magnitude. Examples of scalar quantities are mass, time and speed.

1.4.1 Vector Algebra

Two vectors are equal if they have the same magnitude and direction. The negative of a vector, denoted \(-a\), is a vector of the same magnitude as \(a\) but in the opposite direction.

We add two vectors \(a\) and \(b\) by placing the tail of \(b\) at the head of \(a\) and defining \(a+b\) to be the vector with tail at the origin and head at the head of \(b\) as shown in figure 1.2.

![Vector arithmetic](image)

**Figure 1.2:** Vector arithmetic

The difference, \(a-b\), is defined as the sum of \(a\) and the negative of \(b\), \(a+(-b)\). The result of multiplying \(a\) by scalar \(\alpha\) is a vector of magnitude \(|\alpha||a|\) with the same/opposite direction if \(\alpha\) is positive/negative.

**Zero and Unit Vectors:** The additive identity element for vectors is the zero vector or null vector. This is vector of magnitude zero which is denoted as 0. A unit vector is a vector of magnitude one. If \(a\) is nonzero then \(a/|a|\) is a unit vector in the direction of \(a\).

Unit vectors are often denoted with a caret over-line, \(\hat{n}\).
**Rectangular Unit Vectors:** In \( n \) dimensional Cartesian space, the unit vectors in the directions of the coordinates axes are \( e_1, e_2, \ldots, e_n \). These are called the rectangular unit vectors.

**Linear Dependence Vectors:** A set of \( M \)-dimensional vectors \( \{x_1, x_2, x_3, \ldots, x_M\} \) is said to be linearly dependent if there exist numbers \( \{c_1, c_2, c_3, \ldots, c_M\} \) not all \( 0 \) such that

\[
c_1x_1 + c_2x_2 + c_3x_3 + \ldots + c_Mx_M = 0
\]

**Inner Product:** The inner product of two vectors \( x, y \in \mathbb{R}^M \), \( x = [x_1, x_2, \ldots, x_M]^T \) and \( y = [y_1, y_2, \ldots, y_M]^T \) is defined as

\[
\langle x, y \rangle = \sum_{i=1}^{M} x_i y_i = x^T y
\]  

(1.1)

When the product of two vectors \( \langle x, y \rangle \) is \( 0 \), then vectors are said to be orthogonal.

**1.4.2 Kronecker Delta and Einstein Summation Convention**

The Kronecker Delta tensor is defined as,

\[
\delta_{ij} = \begin{cases} 
1 & \text{if } i = j \\
0 & \text{if } i \neq j 
\end{cases}
\]  

(1.2)

This notation is useful in out work with vectors.

Consider writing a vector in terms of its rectangular components. Instead of using ellipses: \( a = a_1e_1 + \ldots + a_ne_n \), we could write the expression as a sum: \( a = \sum_{i=1}^{n} a_ie_i \), where it is understood that whenever an index is repeated in a term we sum over that index from
1 to \(n\). This is the Einstein summation convention. A repeated index is called a summation index or a dummy index. Other indices can take any value from 1 to \(n\) and are called free indices.

1.5 Matrix

A matrix is a rectangular array of values arranged into rows and columns. Here is a matrix \(A\) of size \(m \times n\):

\[
\begin{pmatrix}
 a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\
 a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\
 \vdots & \vdots & \vdots & \ddots & \vdots \\
 a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn}
\end{pmatrix}
\]

The summation or subtraction of two matrices \(A\) and \(B\) is defined by adding or subtracting corresponding elements as:

\[
A \pm B = \begin{pmatrix}
 a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\
 a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\
 \vdots & \vdots & \vdots & \ddots & \vdots \\
 a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn}
\end{pmatrix} \pm \begin{pmatrix}
 b_{11} & b_{12} & b_{13} & \cdots & b_{1n} \\
 b_{21} & b_{22} & b_{23} & \cdots & b_{2n} \\
 \vdots & \vdots & \vdots & \ddots & \vdots \\
 b_{m1} & b_{m2} & b_{m3} & \cdots & b_{mn}
\end{pmatrix} = \begin{pmatrix}
 a_{11} \pm b_{11} & a_{12} \pm b_{12} & a_{13} \pm b_{13} & \cdots & a_{1n} \pm b_{1n} \\
 a_{21} \pm b_{21} & a_{22} \pm b_{22} & a_{23} \pm b_{23} & \cdots & a_{2n} \pm b_{2n} \\
 \vdots & \vdots & \vdots & \ddots & \vdots \\
 a_{m1} \pm b_{m1} & a_{m2} \pm b_{m2} & a_{m3} \pm b_{m3} & \cdots & a_{mn} \pm b_{mn}
\end{pmatrix}
\]

The summation or subtraction is undefined if the size of the matrices is different.

The multiplication of two matrices is possible only when the number of columns of the first matrix is equal to the number of rows of the second matrix i.e. the product for two matrices \(A\) and \(B\) could be possible only when the order of \(A\) is \(i \times j\) and the order of \(B\) must \(j \times k\), the product matrix will be of the order of the \(I \times k\).
\[(AB)_{ik} = \sum_j A_{ij} B_{jk}\]  \hspace{1cm} (1.3)

The **identity matrix** \(I\) has elements \(I_{ij}\) such as: \(I_{ij} = 1\) when \(i = j\) and \(I_{ij} = 0\) otherwise.

The **transpose** of a matrix \(A = [a_{ij}]_{m \times n}\) is defined as

\[A^T = [a_{ji}]_{n \times m}\]  \hspace{1cm} (1.4)

The **inverse** of a square matrix \(A = [a_{ij}]_{n \times n}\) is defined as

\[A^{-1} = \frac{Adj(A)}{\text{det}(A)}\]  \hspace{1cm} (1.5)

The **rank** of a matrix \(A \in \mathbb{R}^{N \times M}\) is defined as the number of linearly independent rows and columns of \(A\). If \(A\) is a full rank, then its rank is \(N\) or \(M\), which ever is lower.

The **trace** of a matrix is defined as a sum of the diagonal elements of the matrix

\[\text{trace}(A) = \sum_{i=1}^{n} a_{ii}\]  \hspace{1cm} (1.6)

### 1.5.1 Pseudo inverse

The pseudo inverse \(A^+\), also called the Moore–Penrose generalized inverse, of a matrix \(A \in \mathbb{R}^{m \times n}\) is unique, which satisfies

\[AA^+A = A\]

\[A^+AA^+ = A^+\]

\[\left(AA^+\right)^* = AA^+\]

\[\left(A^+A\right)^* = A^+A\]
\[ A^\dagger = (A^\top A)^{-1} A^\top \]  
(1.7)

If \( A^\top A \) is nonsingular, and

\[ A^\dagger = A^\top (A A^\top)^{-1} \]  
(1.8)

if \( A A^\top \) is nonsingular. The pseudo inverse is directly associated with linear LS problems.

When \( A \) is a square nonsingular matrix, the pseudo inverse \( A^\dagger \) is its inverse \( A^{-1} \). For a scalar \( \alpha \), if \( \alpha \neq 0 \), \( \alpha^\dagger = \alpha^{-1} \); if \( \alpha = 0, \alpha^\dagger = 0 \)

### 1.5.2 Vector norms and Matrix Norms

A norm acts as a measure of distance. A **vector norm** on \( \mathbb{R}^n \) is a mapping \( f : \mathbb{R}^n \to \mathbb{R} \) that satisfies such properties: For any \( x, y \in \mathbb{R}^n, a \in \mathbb{R} \),

- \( f(x) \geq 0 \), and \( f(x) = 0 \) iff \( x = 0 \)

- \( f(x + y) \leq f(x) + f(y) \).

- \( f(ax) = |a|f(x) \).

The mapping is denoted as \( f(x) = \|x\| \).

The \( p \)-norm or \( L_p \)-norm is a popular class of vector norms:

\[ \|x\|_p = \left( \sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}} \]  
(1.9)

with \( p \geq 1 \). Usually, the \( L_1, L_2 \) and \( L_\infty \) norms are more useful.
Chapter 1: Preliminary Mathematical Concepts

\[ \|x\|_1 = \sum_{i=1}^{n} |x_i| \]  \hspace{1cm} (1.10)

\[ \|x\|_2 = \left( \sum_{i=1}^{n} |x_i|^2 \right)^{\frac{1}{2}} = (x^T x)^{\frac{1}{2}} \]  \hspace{1cm} (1.11)

\[ \|x\|_\infty = \max_{1 \leq i \leq n} |x_i| \]  \hspace{1cm} (1.12)

The \( L_2 \)-norm is the popular Euclidean norm.

A matrix norm is a generalization of the vector norm by extending \( \mathbb{R}^n \) to \( \mathbb{R}^{m \times n} \).

For a matrix \( A = [a_{ij}]_{m \times n} \), the most frequently used matrix norms are the Frobenius norm

\[ \|A\|_F = \left( \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2 \right)^{\frac{1}{2}} \]  \hspace{1cm} (1.13)

And the matrix \( p \)-norm

\[ \|A\|_p = \sup_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p} = \max_{\|x\|_p = 1} \|Ax\|_p \]  \hspace{1cm} (1.14)

where sup is the supreme operation.

The matrix 2-norm and the Frobenius norm are invariant with respect to orthogonal transforms, that is, for all orthogonal \( Q_1 \) and \( Q_2 \) of appropriate dimensions

\[ \|Q_1AQ_2\|_F = \|A\|_F \]  \hspace{1cm} (1.15)

\[ \|Q_1AQ_2\|_2 = \|A\|_2 \]  \hspace{1cm} (1.16)
1.6 Decomposition

1.6.1 Eigen-value Decomposition

Given a square matrix $A \in \mathbb{R}^{n \times n}$, if there exists a scalar $\lambda$ and a nonzero vector $v$ such that

$$Av = \lambda v$$

then $\lambda$ and $v$ are, respectively, called an eigen-value of $A$ and its corresponding eigenvector. All the eigen-values $\lambda_i, i = 1, \ldots, n$, can be obtained by solving the characteristic equation:

$$\det(A - \lambda I) = 0$$

where $I$ is an $n \times n$ identity matrix. The set of all the eigen-values is called the spectrum of $A$.

If $A$ is nonsingular, $\lambda_i \neq 0$. If $A$ is symmetric, then all the $\lambda_i$'s are real.

The maximum and minimum eigen-values satisfy the Rayleigh quotient

$$\lambda_{\text{max}}(A) = \max_{v \neq 0} \frac{v^T Av}{v^T v}$$

The trace of a matrix is equal to the sum of all its eigen-values and the determinant of a matrix is equal to the product of its eigen-values:

$$\text{tr}(A) = \sum_{i=1}^{n} \lambda_i$$

$$|A| = \prod_{i=1}^{n} \lambda_i$$

1.6.2 Singular Value Decomposition
For a matrix \( A \in \mathbb{R}^{m \times n} \), there exist real unitary matrices \( U = [u_1, u_2, u_3, \ldots, u_m] \in \mathbb{R}^{m \times n} \) and \( V = [v_1, v_2, v_3, \ldots, v_m] \in \mathbb{R}^{n \times n} \) such that
\[
U^T A V = \Sigma
\]
(1.22)
where \( \Sigma \in \mathbb{R}^{m \times n} \) is a real pseudodiagonal \( m \times n \) matrix with \( \sigma_i, i = 1, 2, 3, 4, \ldots, p \),
\[
p = \min(m, n), \sigma_1 \geq \sigma_2 \geq \sigma_3 \geq \ldots \sigma_p \geq 0,
\]
on the diagonal and zeros off the diagonal. \( \sigma_i \)'s are called the singular values of \( A \), \( u_i \) and \( v_i \) are, respectively called the left singular vector and right singular vector for \( \sigma_i \). They satisfy the relations:
\[
A v_i = \sigma_i u_i \quad \text{and} \quad A^T u_i = \sigma_i v_i.
\]
Accordingly \( A \) can be written as
\[
A = U \Sigma V^T = \sum_{i=1}^{r} \lambda_i u_i v_i^T
\]
(1.23)
where \( r \) is the cardinality of the smallest nonzero singular value. In the special case when
\( A \) is a symmetric non-negative definite matrix, \( \Sigma = \text{diag} \left( \frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_p} \right) \),
where \( \lambda_1 \geq \lambda_2 \geq \ldots \lambda_p \geq 0 \) are the real eigenvalues of \( A \), \( \nu_i \) being the corresponding eigenvectors.

The SVD is useful in many situations. The rank of \( A \) can be determined by the number of nonzero singular values. The power of \( A \) can be easily calculated by
\[
A^k = U \Sigma^k V^T, \text{ where } k \text{ is a positive integer.}
\]
The SVD is extensively applied in linear inverse problems. The pseudoinverse of \( A \) can then be described by equation (1.7).
The Frobenius norm can thus be calculated as

$$\| A \|_F = \left( \sum_{i=1}^{p} \sigma_i^2 \right)^{\frac{1}{2}}$$  \hspace{1cm} (1.24)

and the matrix 2-norm is calculated by

$$\| A \|_2 = \sigma_1$$  \hspace{1cm} (1.25)

1.6.3 QR Decomposition

Although the quantity \( (A^T A)^{-1} \) exists, significant numerical difficulties may occur in computing this inverse in instance where \( A^T A \) is nearly singular. For the full-rank or over determined linear LS case, \( m \geq n \), can also be solved by using QR decomposition procedure.

\( A \) is first factorized as

$$A = QR$$  \hspace{1cm} (1.26)

where \( Q \) is an \( m \times m \) orthogonal matrix, that is, \( Q^T Q = I \), and \( R = \begin{bmatrix} \tilde{R} \\ 0 \end{bmatrix} \) is an \( m \times n \) upper triangular matrix with \( \tilde{R} \in R^{m \times n} \).

Inserting (1.26) into the set of linear equation (SLE) \( Ax = b \) and premultiplying by \( Q^T \), we have

$$Rx = Q^T b$$  \hspace{1cm} (1.27)

Denoting \( Q^T b = \begin{bmatrix} \tilde{b} \\ b \end{bmatrix} \), where \( \tilde{b} \in R^n \) and \( b \in R^{m-n} \), we have
\[ \bar{R}x = \bar{b} \]  

(1.28)

Since \( \bar{R} \) is a triangular matrix, \( x \) can be easily solved using backward substitution. This is the procedure used in the GSO procedure.

When rank \( (A) < n \), the rank-deficient LS problem has an infinite number of solutions, the QR decomposition does not necessarily produce an orthonormal basis for \( \text{range}(A) = \{ y \in \mathbb{R}^n : y = Ax \text{ for some } x \in \mathbb{R}^n \} \). The QR decomposition can be applied to produce an orthonormal basis for \( \text{range} (A) \).

The QR decomposition is a basic method for computing the SVD. The QR decomposition itself can be computed by means of the Givens rotation, the Householder transform, or the GSO.

1.6.4 Condition Numbers

The condition number of a matrix \( A \in \mathbb{R}^{m \times n} \) is defined by

\[ \text{cond}_p (A) = \left\| A \right\|_p \left\| A^\dagger \right\|_p \]  

(1.29)

where \( p \) can be selected as 1, 2, \( \infty \), Frobenius, or any other norm. The relation, \( \text{cond}(A) \geq 1 \), always holds. Matrices with small condition numbers are well conditioned, while matrices with large condition number are poorly conditioned or ill-conditioned. The condition number is especially useful in numerical computation, where ill-conditioned matrices are sensitive to rounding errors. For the \( L_2 \)-norm,

\[ \text{cond}_2 (A) = \frac{\sigma_1}{\sigma_2} \]  

(1.30)

where \( p = \min(m,n) \).
1.7 Optimization Methods

1.7.1 Vector Gradient

Let $g$ be a differentiable scalar function of $m$ variables,

$$g = g(w_1, ..., w_n) = g(w)$$  \hspace{1cm} (1.31)

where $w = (w_1, ..., w_m)^T$. Then the vector gradient of $g$ w.r.t. $w$ is the $m$-dimensional vector of partial derivatives of $g$,

$$\nabla g = \nabla_w g = \left( \frac{\partial g}{\partial w_1}, \ldots, \frac{\partial g}{\partial w_m} \right)$$  \hspace{1cm} (1.32)

Similarly, we can define second-order gradient or Hessian matrix,

$$\frac{\partial^2 g}{\partial w^2} = \begin{pmatrix}
\frac{\partial^2 g}{\partial w_1^2} & \ldots & \frac{\partial^2 g}{\partial w_1 \partial w_m} \\
\vdots & \ddots & \vdots \\
\frac{\partial^2 g}{\partial w_m w_1} & \ldots & \frac{\partial^2 g}{\partial w_m^2}
\end{pmatrix}$$  \hspace{1cm} (1.33)

Generalization to the vector valued functions $g(w) = (g_1(w), ..., g_n(w))^T$ leads to a definition of the Jacobian matrix of $g$ w.r.t. $w$,

$$\frac{\partial g}{\partial w} = \begin{pmatrix}
\frac{\partial g_1}{\partial w_1} & \ldots & \frac{\partial g_1}{\partial w_m} \\
\vdots & \ddots & \vdots \\
\frac{\partial g_n}{\partial w_1} & \ldots & \frac{\partial g_n}{\partial w_m}
\end{pmatrix}$$  \hspace{1cm} (1.34)
In this vector convention, the columns of the Jacobian matrix are gradients of the corresponding components functions \(g\), w.r.t. the vector \(w\).

The differentiation rules are analogous to the case of ordinary functions:

\[
\frac{\partial f(w)g(w)}{\partial w} = \frac{\partial f(w)}{\partial w} g(w) + f(w) \frac{\partial g(w)}{\partial w}
\]  

(1.35)

\[
\frac{\partial f(w)}{g(w)} = \left[ \frac{\partial f(w)}{\partial w} g(w) - f(w) \frac{\partial g(w)}{\partial w} \right] / g^2(w)
\]  

(1.36)

\[
\frac{\partial f(g(w))}{\partial w} = f'(g(w)) \frac{\partial g(w)}{\partial w}
\]  

(1.37)

### 1.7.2 Matrix Gradient

Consider a scalar valued function \(g\) of the \(n \times m\) matrix \(W = (w_y)\) (e.g. determinant of a matrix). The matrix gradient w.r.t \(W\) is a matrix of the same dimension as \(W\) consisting of partial derivatives of \(g\) w.r.t. components of \(W\):

\[
\frac{\partial g}{\partial W} = \begin{pmatrix}
\frac{\partial g}{\partial w_{11}} & \cdots & \frac{\partial g}{\partial w_{1n}} \\
\vdots & \ddots & \vdots \\
\frac{\partial g}{\partial w_{m1}} & \cdots & \frac{\partial g}{\partial w_{mn}}
\end{pmatrix}
\]  

(1.38)

### 1.7.3 Taylor Expansion of Multivariate Functions

The well known formula for Taylor series expansion of a scalar function \(g(w)\) reads,

\[
g(w') = g(w) + \frac{dg}{dw}(w'-w) + \frac{1}{2!} \frac{d^2g}{dw^2}(w'-w)^2 + \ldots = g(w) + \sum g'(w)(w'-w)^i
\]  

(1.39)
This can be generalized to the function of $m$ variables

$$g(w') = g(w) + \frac{\partial g}{\partial w}(w' - w) + \frac{1}{2}(w' - w)^T \frac{\partial^2 g}{\partial w^2} (w' - w) + ... \quad (1.40)$$

with derivatives evaluated at $w$. Note that the second term is an inner product of a gradient of $g$ with the vector $w' - w$ and the third term is a quadratic form with the Hessian matrix $\frac{\partial^2 g}{\partial w^2}$.

Similarly for a scalar function of a matrix variable:

$$g(W') = g(W) + \text{trace} \left( \frac{\partial g}{\partial W} (W' - W) \right) + ... \quad (1.41)$$

Reminder: $\text{trace}(A) = \sum_{i=1}^{n} a_{ii}$ \hspace{1cm} (1.42)

i.e. trace is defined as a sum of the diagonal elements of the matrix. The above formula shows 2 terms of Taylor expansion. It uses the extension of definition of an inner product to matrices,

$$AB = \sum_{i=1}^{m} \sum_{j=1}^{m} a_{ij} b_{ij} \quad (1.43)$$

but

$$\text{trace}(A^T B) = \sum_{i=1}^{m} (A^T B)_{ii} = \sum_{i=1}^{m} \sum_{j=1}^{m} a_{ij} b_{ji} \quad (1.44)$$

1.7.4 Unconstrained Optimization
Gradient descent is a method of minimization of a given cost or objective function $J(w)$:

- Start at some initial point $w(0)$;
- Calculate gradient of $J(w)$ at $w(0)$;
- Move in the direction of the negative gradient or steepest descent by some distance;
- Repeat above until consecutive points are sufficiently close.

In mathematical notation the above procedure reads

$$
\Delta w(t) = -\alpha(t) \frac{\partial J(w)}{\partial w} \tag{1.45}
$$

or

$$
\Delta w \alpha = -\frac{\partial J(w)}{\partial w} \tag{1.46}
$$

- Gradient descent moves always downwards in a hilly landscape
- Local minima can trap the movement
- Initialization is important to avoid local minima
- Choice of the learning rate is crucial for speed of convergence;

Stochastic gradient descent

- Data dependent cost functions;
• Statistical model for data;

• Typical form of cost function $J(w) = E[g(w, x)]$, where $E$ denotes expectation
  (i.e. $E[X] = \int x f(x) dx$);

• $x$ is the random vector modeling observation vector;

• (Unknown) pdf of $x$ is $f(x)$ w.r.t. which the expectation is taken;

• Usually only a sample $x(1), x(2), \ldots$ is given

The steepest descent-learning rule becomes,

$$
    w(t) = w(t-1) - \alpha(t) \frac{\partial}{\partial w} E[g(w, x(t))]|_{w=w(t-1)}
$$

(1.47)

or

$$
    w(t) = w(t-1) - \alpha(t) \frac{\partial}{\partial w} \int g(w, \xi) f(\xi) d\xi
$$

(1.48)

For twice differentiable $g(w, x)$ w.r.t. $w$ this equals,

$$
    w(t) = w(t-1) - \alpha(t) \left[ \frac{\partial}{\partial w} g(w, \xi) \right] f(\xi) d\xi
$$

(1.49)

• Approximate the expectation by the sample mean

• Batch-learning: using all available data in each step

On line version is possible - drop the expectation,

$$
    w(t) = w(t-1) - \alpha(t) \frac{\partial}{\partial w} g(w, \xi)|_{w=w(t-1)}
$$

(1.50)
• Calculates direction of next move using only 1 data point (incoming)
• The direction of movement fluctuates highly between the steps but
• The average direction is approximately the steepest descent of the batch version
• Lower computational cost
• Slower - needs more iterations to converge

1.8 Dynamical Systems

For a dynamical system described by a set of ordinary differential equations, the stability of the system can be examined by Lyapunov’s second theorem or the Lipschitz condition.

Lyapunov’s second theorem: For a dynamic system described by a set of differential equations

\[
\frac{dx}{dt} = f(x)
\]

(1.51)

where \( x = (x_1(t), x_2(t), \ldots, x_n(t))^T \) and \( f = (f_1, f_2, \ldots, f_n)^T \). There exists a positive definite function \( E = E(x) \), called a Lyapunov function or energy function, so that

\[
\frac{dE}{dt} = \sum_{j=1}^{n} \frac{\partial E}{\partial x_j} \frac{dx_j}{dt} \leq 0
\]

(1.52)

with \( \frac{dE}{dt} = 0 \) only for \( \frac{dx}{dt} = 0 \), then the system is stable, and the trajectories \( x \) will asymptotically converge to stationary points as \( t \to \infty \).

The stationary points are also known as equilibrium points and attractors. The crucial step in applying the Lyapunov’s second theorem is to find a suitable energy function.
**Chapter 1: Preliminary Mathematical Concepts**

**Lipschitz Condition:** For a dynamic system described by equation(1.51), a sufficient condition that guarantees the existence and uniqueness of the solution is given by the Lipschitz condition

\[ \| f(x_1) - f(x_2) \| \leq \gamma \| x_1 - x_2 \| \]  

(1.53)

where \( \gamma \) is any positive constant, called the Lipschitz constant, and \( x_1, x_2 \) are any two variables in the domain of the function vector \( f \). \( f(x) \) is said to be Lipschitz continuous.

If \( x_1 \) and \( x_2 \) are in some neighborhood of \( x \), then they are said to satisfy the Lipschitz condition locally and will reach a unique solution in the neighborhood of \( x \). The unique solution is a trajectory that will converge to an attractor asymptotically and reach it only at \( t \to \infty \).

### 1.9 Probability

**Definition:**

\( \Omega \): Sample space; contains all possible outcomes of an experiment. \( \Omega \) can be discrete or continuous.

\( \omega \): A single outcome; \( \omega \in \Omega \)

**A:** Specific event of interest or set of outcomes \( A \subseteq \Omega \). An event \( A \) is said to occur if the observed outcome \( \omega \) is an element of \( A \), that is, \( \omega \in A \). Associated that with these definitions is a probability space.

For example, \( A_1 \subseteq A_2 \Rightarrow P(A_1) \leq P(A_2) \)
where $P(A)$ is a probability function that assigns a real, scaler-valued function to each set $A$, with the constraints:

1. $P(A) \geq 0 \quad \forall \quad A \in \Omega$

2. $P(\Omega) = 1$

3. If sets or experiments outcomes $A_1, A_2, \ldots, A_n$ are mutually exclusive, that is, $A_i \cap A_j = \emptyset$, $\forall i, j$, then $P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)$.

Given the set of all outcomes $A_i, i = 1, 2, 3, \ldots, n$, which constitute a partition of $\Omega$, $\sum_{i} P(A_i) = 1$

4. $P(\emptyset) = 0$

### 1.10 Random Variables

Assume we are conducting an experiment that involves asking 10 people whether they like ice cream or not. For each of the surveyed people we record '1' if they say they like it, '0' if they say they don't. The sample space for this experiment is the space of all binary strings of length 10, i.e., it has $2^{10}$ elements. It follows that there are $2^{2^{10}}$ possible events in this sample space. If we are only interested in how many said that they liked ice cream, then we can reduce the sample space to the numbers 0 through 10 which is easier to deal with than the original space. Note that we really defined a function from the sample space into the numbers 0, 1, \ldots, 10 by counting the number of '1's in the binary string encoding of the answers. In general, we define a random variable to be a function
from a sample space into the real numbers. In most experiments, random variables are used implicitly as in "sum of numbers" in a toss of two dice, "number of heads" in 25 coin tosses, etc. Note that in defining a random variable we have defined a new sample space as the range of the variable, \{0, 1, \ldots, 10\} in our ice cream experiment. Given a sample space with a probability function \( P \) and we define a random variable \( X \), we can define a probability function \( P_X \) for the range of \( X \) using \( P \) the following way. Let \( X \) be the range of \( X \), and let \( x \in X \). We observe \( X = x \) if and only if the outcome of the experiment is an \( s \in S \) such that \( X(s) = x \). Hence,

\[
P_X(X = x) = P\left(\{s \in S \mid X(s) = x\}\right) \tag{1.54}
\]

If we concern with event \( A \subseteq X \), we define,

\[
P_X(A) = P\left(\{s \in S \mid X(s) \in A\}\right) \tag{1.55}
\]

We often write \( P(X \in A) \) for \( P_X(A) \) and \( P(X = x) \) for \( P_X(X = x) \) if no confusion can arise.

We can associate several functions with a random variable. The first is the cumulative distribution function or CDF, defined by,

\[
F_X(x) = P_X(X \leq x), \text{ for all } x \tag{1.56}
\]

A random variable is discrete if \( F_X(x) \) is a step function of \( x \), and is continuous if \( F_X(x) \) is continuous. The probability mass function or PMF of a discrete random variable is defined as,
\[ f_x(x) = P_x(X = x), \text{ for all } x \] 

(1.57)

Similarly, the probability density function of a continuous random variable is defined as the function \( f_x \) that satisfies,

\[ F_x(x) = \int_{-\infty}^{x} f_x(t) \, dt, \text{ for all } x \] 

(1.58)

### 1.11 Probability Distribution and Densities

**The Probability Distribution**

The probability distribution \( f(x) \) of a discrete random variable is defined as,

\[ P(X = x) = f(x) \] 

(1.59)

**The Cumulative Distribution Function (CDF)**

The CDF of a discrete random variable is,

\[ F_x(x_0) = \sum_{u \leq x_0} f(u) \] 

(1.60)

The CDF \( F_x(x) \) of a random vector \( x = (x_1, x_2, \ldots, x_n)^T \) at point \( x = x_0 \),

\[ F_x(x_0) = P(x \leq x_0) \] 

(1.61)

**The Multivariate Probability Density Function (PDF)**

The PDF \( p_x(x) \) of a continuous random vector \( x \) is defined as,

\[ p_x(x_0) = \frac{\partial}{\partial x_1} \frac{\partial}{\partial x_2} \ldots \frac{\partial}{\partial x_n} F_x(x) \bigg|_{x=x_0} \] 

(1.62)
Hence,

\[ F_x(x_0) = \int_{-\infty}^{x_0} p_x(x) \, dx \]  

(1.63)

For discrete random variable the corresponding formula is,

\[ F_x(x_0) = \sum_{u \leq x_0} f(u) \]  

(1.64)

where the sum is taken over all \( u \)'s, for which \( u \leq x_0 \).

**The Joint Distribution Function**

The joint distribution function of vectors \( x \) and \( y \) is given by,

\[ F_{x,y}(x_0, y_0) = P(x \leq x_0, y \leq y_0) \]  

(1.65)

where \( x_0, y_0 \) are vectors of dimensions of \( x \) and \( y \), respectively. Thus, the joint distribution function calculates the probability of the event \( x \leq x_0 \) and \( y \leq y_0 \).

The *joint density function* is defined analogously to previous definitions by differentiation of the joint probability distribution w.r.t. all components. It follows that the probability of an event \( (x \leq x_0, y \leq y_0) \) is,

\[ P(x \leq x_0, y \leq y_0) = \int_{-\infty}^{x_0} \int_{-\infty}^{y_0} p_{x,y}(\xi, \eta) \, d\eta \, d\xi \]  

(1.66)

**1.12 Gaussian Distribution**

The Gaussian distribution, known as the *normal distribution*, is the most common assumption for error distribution. The PDF of the normal distribution is defined as
\[ p(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \] (1.67)

for \( x \in \mathbb{R} \), where \( \mu \) is the mean and \( \sigma > 0 \) is the standard deviation. For the Gaussian distribution, 99.73% of the data are within the range of \([\mu - 3\sigma, \mu + 3\sigma]\). The Gaussian distribution has its first-order moment as \( \mu \), second order moment as \( \sigma^2 \), and higher-order moments as zero. If \( \mu = 0 \) and \( \sigma = 1 \), the distribution is called the standard normal distribution. The PDF is also known as the likelihood function. An ML estimator is a set of values \((\mu, \sigma)\) that maximizes the likelihood function for a fixed value of \( x \).

The cumulative distribution function (CDF) is defined as the probability that a random variable is less than or equal to a value \( x \), that is

\[ F(x) = \int_{-\infty}^{x} p(t) dt \] (1.68)

The standard normal CDF, conventionally denoted \( \Phi \), is given by setting \( \mu = 0 \) and \( \sigma = 1 \).

The standard normal CDF is usually expressed by

\[ \Phi(x) = \frac{1}{2} \left[ 1 + \text{erf} \left( \frac{x}{\sqrt{2}} \right) \right] \] (1.69)

where the error function \( \text{erf}(x) \) is a non elementary function, which is defined by

\[ \text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-t^2} dt \] (1.70)

When vector \( x \in \mathbb{R}^n \), the PDF of the normal distribution is then defined by

\[ p(x) = \frac{1}{(2\pi)^{n/2} |\Sigma|} e^{-\frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu)} \] (1.71)

where \( \mu \) and \( \Sigma \) are the mean vector and the covariance matrix, respectively.
1.13 **Cauchy Distribution**

The Cauchy distribution, also known as the Cauchy–Lorentzian distribution, is another popular data-distribution model. The density of the Cauchy distribution is defined as

\[
p(x) = \frac{1}{\pi \sigma \left[ 1 + \left( \frac{x - \mu}{\sigma} \right)^2 \right]} \tag{1.72}
\]

for \( x \in \mathbb{R} \), where \( \mu \) specifies the location of the peak and \( \sigma \) is the scale parameter that specifies the half-width at the half-maximum. When \( \mu = 0 \) and \( \sigma = 1 \), the distribution is called the **standard Cauchy distribution**.

Accordingly, the CDF of the Cauchy distribution is calculated by

\[
F(x) = \frac{1}{\pi} \arctan \left( \frac{x - \mu}{\sigma} \right) + \frac{1}{2} \tag{1.73}
\]

None of the moments is defined for the Cauchy distribution. The median of the distribution is equal to \( \mu \). The Cauchy distribution has a longer tail than the Gaussian distribution, and this makes it more valuable in stochastic search algorithms by searching larger subspaces in the data space.

1.14 **Markov Processes, Markov Chains, and Markov-chain Analysis**

Given a stochastic process \( \{X(t) : t \in \mathbb{T}\} \), where \( t \) is time, \( X(t) \) is a state in the state space \( \mathcal{S} \).

A Markov process is defined as a stochastic process that satisfies the relation characterized by the conditional distribution

\[
P[X(t_0 + t_1) \leq x | X(t_0) = x_0, X(\tau) = x, -\infty < \tau < t_0]\]

\[
= P[X(t_0 + t_1) \leq x | X(t_0) = x_0] \tag{1.74}
\]
for any value of $t_0$ and for $t_1 > 0$. The future distribution of the process is determined by the present value of $X(t_0)$ only.

When $T$ and $S$ are discrete, a Markov process is called a *Markov chain*. Conventionally, time is indexed using integers, and a Markov chain is a set of random variables that satisfy

$$P[X_n = x_n | X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2} \ldots]$$

$$= P[X_n = x_n | X_{n-1} = x_{n-1}]$$

(1.75)

This definition can be extended for multistep Markov chains, where a chain state has conditional dependency on only a finite number of its previous states.

For a Markov chain, $P[X_n = j | X_{n-1} = i]$ is the transition probability of state $i$ to $j$ at time $n-1$. If

$$P[X_n = j | X_{n-1} = i] = P[X_{n+m} = j | X_{n+m-1} = i]$$

(1.76)

for $m \geq 0$ and $i, j \in S$, the chain is said to be *time homogeneous*. In this case, one can denote

$$P_{i,j} = P[X_n = j | X_{n-1} = i]$$

(1.77)

and the transition probabilities can be represented by a matrix, called the *transition matrix*, $P = [P_{ij}]$, where $i, j = 0, 1, \cdots$. For finite $S$, $P$ has a finite dimension.

In the Markov-chain analysis, the transition probability after $k$ step transitions is $P^k$. The *stationary distribution* or *steady-state distribution* is a vector that satisfies

$$P^T \pi^* = \pi^*$$

(1.78)
That is, \( \pi^* \) is the left eigenvector of \( P \) corresponding to the eigenvalue 1. If \( P \) is irreducible and aperiodic, that is, every state is accessible from every other state and in the process none of the states repeats itself periodically, then \( P^k \) converges element wise to a matrix each row of which is the unique stationary distribution \( \pi^* \), with

\[
\lim_{k \to \infty} P^k \pi = \pi^*
\]  

(1.79)

Many modeling applications are Markovian, and the Markov-chain analysis is widely used for convergence analysis for algorithms.

1.15 Bayes Decision Rule for Minimum Error

Consider \( C \) classes, \( \omega_1, \omega_2, \omega_3, \ldots, \omega_C \), with a priori probabilities (the probabilities of each class occurring) \( p(\omega_1), p(\omega_2), \ldots, p(\omega_C) \), assumed known. If we wish to minimize the probability of making an error and we have no information regarding an object other than the class probability distribution then we would assign an object to class \( \omega_j \) if

\[
p(\omega_j) > p(\omega_k) \quad k = 1, \ldots, C; \ k \neq j
\]

(1.80)

This classifies all objects as belonging to one class. For classes with equal probabilities, patterns are assigned arbitrarily between those classes.

However, we do have an observation vector or measurement vector \( x \) and we wish to assign \( x \) to one of the \( C \) classes. A decision rule based on probabilities is to assign \( x \) to class \( \omega_j \) if the probability of class \( \omega_j \) given the observation \( x, p(\omega_j|x) \), is greatest over all classes \( \omega_1, \omega_2, \omega_3, \ldots, \omega_C \). That is, assign \( x \) to class \( \omega_j \) if

\[
p(\omega_j|x) > p(\omega_k|x) \quad k = 1, \ldots, C; \ k \neq j
\]

(1.81)
This decision rule partitions the measurement space into \( C \) regions \( \Omega_1, \ldots, \Omega_c \) such that if \( x \in \Omega_j \) then \( x \) belongs to class \( \omega_j \).

The \textit{a posteriori} probabilities \( p(\omega_j|x) \) may be expressed in terms of the \textit{a priori} probabilities and the class-conditional density functions \( p(x|\omega_j) \) using Bayes' theorem as

\[
p(\omega_j|x) = \frac{p(x|\omega_j)p(\omega_j)}{p(x)}
\]

and so the decision rule (1.81) may be written: assign \( x \) to \( \omega_j \) if

\[
p(x|\omega_j)p(\omega_j) > p(x|\omega_k)p(\omega_k) \quad k = 1, \ldots, C; k \neq j
\]

This is known as Bayes' rule for \textit{minimum error}.

For two classes, the decision rule (1.82) may be written

\[
l_r = \frac{p(x|\omega_1)}{p(x|\omega_2)} > \frac{p(\omega_2)}{p(\omega_1)} \quad \text{implies} \quad x \in \text{class } \omega_1
\]

The fact that the decision rule (1.82) minimizes the error may be seen as follows. The probability of making an error, \( p(\text{error}) \), may be expressed as

\[
p(\text{error}) = \sum_{i=1}^{C} p(\text{error}|\omega_i)p(\omega_i)
\]

where \( p(\text{error}|\omega_i) \) is the probability of misclassifying patterns from class \( \omega_i \). This is given by

\[
p(\text{error}|\omega_i) = \int_{C[\omega_i]} p(x|\omega_i) dx
\]

the integral of the class-conditional density function over \( C[\omega_i] \), the region of measurement space outside \( \Omega_i \) (\( C \) is the complement operator), i.e. \( \sum_{j=1, j \neq i}^{C} \Omega_j \)
Therefore, we may write the probability of misclassifying a pattern as

\[
p(\text{error}) = \sum_{i=1}^{C} \int_{\Omega_i} p(x|\omega_i) p(\omega_i) dx
\]

\[
= \sum_{i=1}^{C} p(w_i) \left( 1 - \int_{\Omega_i} p(x|\omega_i) dx \right)
\]

\[
= 1 - \sum_{i=1}^{C} p(\omega_i) \int_{\Omega_i} p(x|\omega_i) dx
\]  
Equation (1.87)

from which we see that minimizing the probability of making an error is equivalent to maximizing

\[
\sum_{i=1}^{C} p(\omega_i) \int_{\Omega_i} p(x|\omega_i) dx
\]  
Equation (1.88)

the probability of correct classification. Therefore, we wish to choose the regions \( \Omega_i \), so that the integral given in (1.86) is a maximum. This is achieved by selecting \( \Omega_i \) to be the region for which \( p(\omega_i) p(x|\omega_i) \) is the largest over all classes and the probability of correct classification, \( c \), is

\[
C = \int \max_i p(\omega_i) p(x|\omega_i) dx
\]  
Equation (1.89)

where the integral is over the whole of the measurement space, and the Baye's error is

\[
e_B = 1 - \int \max_i p(\omega_i) p(x|\omega_i) dx
\]  
Equation (1.90)
References:


