Part IV
Appendices

Appendix A
Statistics

A.1 Fundamentals

This section provides the fundamentals of probability and statistics. The concept of probability of an event is introduced as a limit of the relative frequency, i.e. of the number of times an experiment has such an event as outcome. Based on such a definition, the rest of this section introduces the *addition law*, defines the *conditionality* and the *statistical independence*.

A.1.1 Probability and Relative Frequency

Consider the simple experiment of tossing an unbiased coin: two mutually exclusive outcomes are possible, head (H) or tail (T), and the result is *random*, i.e., it cannot be predicted with certainty because too many parameters should be taken into account to model the motion of the coin. On the other hand, if the experiment is repeated a sufficient number of times and a whole series of *independent trials under identical conditions* is obtained, the outcome shows some regularities: the fraction of experiments with outcome H, the so-called *relative frequency* of H, is always around 1/2:

$$\frac{n(H)}{n} \simeq \frac{1}{2} \tag{A.1}$$

where $n(H)$ is the number of times that the outcome is H and n is the total number of experiments. The same considerations apply to the T outcome and this is what the common language means whan it says that the *probability* of H or T is 50 percent.

In more general terms, if an experiment has K mutually exclusive possible outcomes $A_1, A_2, \ldots, A_K$, the probability $p(A_i)$ of observing the outcome A_i can be thought of as the following limit:

F. Camastra and A. Vinciarelli, *Machine Learning for Audio, Image and Video Analysis*, Advanced Information and Knowledge Processing,
DOI 10.1007/978-1-4471-6735-8

$$p(A_i) = \lim_{n\to\infty} \frac{n(A_i)}{n} \tag{A.2}$$

(see above for the meaning of symbols). This result is known as the *strong law of large numbers* and it provides the definition of the probability.[1]

A.1.2 The Sample Space

A random experiment is characterized by a set Ω of *mutually exclusive* elementary events ω that correspond to all its possible outcomes. Ω is called a *sample space* and an event A is said to be *associated* with it when it is always possible to decide whether the occurrence of an elementary event ω leads to the occurrence of A or not. As an example consider the rolling of a die;, the sample space contains six elementary events $\omega_1, \ldots, \omega_6$ corresponding to the number of spots on each of the die faces. The event of having an even number of spots is associated to Ω because it is a characteristic that can be clearly attributed to each of the elementary events, and it can be thought of as a set $A = \{\omega_2, \omega_4, \omega_6\}$. In the following, A will refer not only to an event, but also to the corresponding set of its underpinning elementary events and whenever there is no ambiguity, the distinction will not be made.

Based on the above, an event can be defined as a subset of the sample space and this enables to interpret the event properties and relationships in terms of sets and subsets as shown in Fig. A.1. Two events A_i and A_j are said to be *mutually exclusive* when the occurrence of one prevents the other from occurring. This situation is shown in Fig. A.1a where the sets of elementary events corresponding to A_i and A_j are disjoint. When A_i and A_j contain exactly the same elements of the sample space, then the occurrence of one corresponds to the occurrence of the other and the two events are said to be *equivalent* (Fig. A.1b). The *union* $A_i \cup A_j$ of two events is the event including all elements ω of both A_i and A_j, while their *intersection* $A_i \cap A_j$ contains only elementary events belonging to both A_i and A_j, as shown in Fig. A.1c, d, respectively. Two events A_i and A_j are called *complementary* when $A_i = \Omega - A_j = \bar{A}_j$ and the occurrence of one is equivalent to the nonoccurrence of the other. The difference between complementarity and mutual exclusivity is that $\bar{A}_i$ contains all events of Ω that are mutually exclusive with respect to A_i. On the other hand, complementarity and mutual exclusivity are the same property when there are only two events. The event A_i implies A_j when $A_i \subset A_j$, i.e. when the occurrence of A_i corresponds to the occurrence of A_j, but the vice versa is not true. This situation is depicted in Fig. A.1f.

[1] The *Strong law of large numbers* will not be demonstrated in this appendix. However, the interested reader can find the demonstration and related issues in most of the academic statistics textbooks.

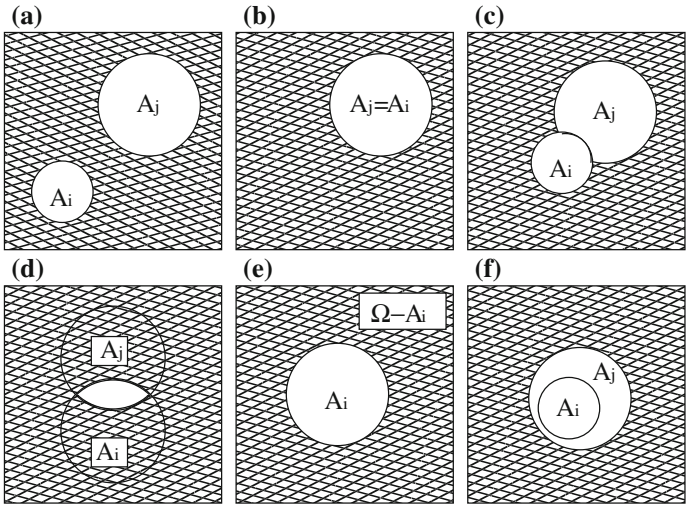

Fig. A.1 Relationships between events. This figure shows the different relationships between events in the sample space. Plot (**a**) shows mutually exclusivity, plot (**b**) shows equivalence, plot (**c**) and (**d**) correspond to union and intersection respectively, plot (**e**) shows the complementarity and plot (**f**) shows the inclusion

A.1.3 The Addition Law

Consider two mutually exclusive events A_i and A_j and the event $A = A_i \cup A_j$. If both A_i and A_j belong to the sample space of an experiment repeated n of times under identical conditions, then the relationship between the respective relative frequencies is as follows:

$$\frac{n(A)}{n} = \frac{n(A_i)}{n} + \frac{n(A_j)}{n}. \tag{A.3}$$

Sect. A.1.1 shows that the relative frequency tends to the probability when $n \to \infty$; thus the above equation corresponds to:

$$P(A) = P(A_i) + P(A_j). \tag{A.4}$$

If the mutually exclusive events are k, then $A = A_1 \cup A_2 \ldots A_k$ and it is possible to write:

$$P(A) = P(A_1 \cup A_2 \ldots A_{k-1}) + P(A_k) \tag{A.5}$$

and the above expression, after applying $k - 2$ times Eq. (A.4), leads to the *addition law for probabilities*:

$$P(A) = P\left(\bigcup_{l=1}^{k} A_l\right) = \sum_{l=1}^{k} P(A_l). \tag{A.6}$$

The above expression is valid only for mutually exclusive events, but an addition law can be obtained also for arbitrary events. This requires to demonstrate some key relationships between probabilitites:

Theorem A.1 *The formulas*

$$0 \leq P(A) \leq 1 \tag{A.7}$$

$$P(A_i - A_j) = P(A_i) - P(A_i \cap A_j) \tag{A.8}$$

$$P(A_j - A_i) = P(A_j) - P(A_i \cap A_j) \tag{A.9}$$

$$P(A_i \cup A_j) = P(A_i) + P(A_j) - P(A_i \cap A_j) \tag{A.10}$$

where $A_i - A_j$ stands for event A_i occurring without event A_j occurring as well, hold for arbitrary events A, A_i and A_j. Moreover, if $A_i \subseteq A_j$, then:

$$P(A_i) \leq P(A_j). \tag{A.11}$$

Equation (A.7) follows from the fact that the probability can be interpreted as a limit of the relative frequency $n(A)/n$. The value of $n(A)$ is the number of times the experiment has A as outcome, thus it cannot be less than 0 and it cannot be more than n. As a consequence:

$$0 \leq \frac{n(A)}{n} \leq 1. \tag{A.12}$$

Such relationships hold also when $n \to \infty$ and this leads to Eq. (A.7).

The events A_i, A_j and $A_i \cup A_j$ can be written as unions of mutually exclusive events as follows:

$$\begin{aligned} A_i &= (A_i - A_j) \cup (A_i \cap A_j) \\ A_j &= (A_j - A_i) \cup (A_i \cap A_j) \\ A_i \cup A_j &= (A_i - A_j) \cup (A_j - A_i) \cup (A_i \cap A_j) \end{aligned}$$

Since all events involved in the above equations are mutually exclusive, the application of the addition law leads to Eqs. (A.8), (A.9) and (A.10), respectively.

When $A_i \subset A_j$, the probability of $A_j - A_i$ is:

$$P(A_j - A_i) = P(A_j) - P(A_i \cap A_j) = P(A_j) - P(A_i) \tag{A.13}$$

because $A_i \cap A_j = A_i$. Since $P(A_j - A_i) \geq 0$,

$$P(A_i) \leq P(A_j). \tag{A.14}$$

which corresponds to Eq. A.12.

After proving the relationships of Theorem A.1, it is possible to avoid the requirement of the mutual exclusivity for the addition law:

Theorem A.2 *Given any n events* $A_1, A_2, \ldots, A_n$, *let*

$$P_1 = \sum_{i=1}^{n} P(A_i) \tag{A.15}$$

$$P_2 = \sum_{1 \leq i \leq j \leq n} P(A_i A_j) \tag{A.16}$$

$$P_3 = \sum_{1 \leq i \leq j \leq k \leq n} P(A_i A_j A_k) \ldots \tag{A.17}$$

where $A_i A_j \ldots A_k$ *is a shorthand for* $A_i \cap A_j \ldots \cap A_k$, *then:*

$$P\left(\bigcup_{l=1}^{n} A_l\right) = P_1 - P_2 + P_3 + \cdots + (-1)^{n+1} P_n. \tag{A.18}$$

When $n = 2$, Eq. (A.18) corresponds to Eq. (A.10); then it is proved. Suppose now that (A.18) holds for $n - 1$; then:

$$P\left(\bigcup_{l=2}^{n} A_l\right) = \sum_{i=2}^{n} P(A_i) - \sum_{2 \leq i \leq j \leq n} P(A_i A_j) + \ldots \tag{A.19}$$

and

$$P\left(\bigcup_{l=2}^{n} A_1 A_l\right) = \sum_{i=2}^{n} P(A_1 A_i) - \sum_{2 \leq i \leq j \leq n} P(A_1 A_i A_j) + \ldots. \tag{A.20}$$

Based on Eq. (A.10), it is possible to write:

$$P\left(\bigcup_{l=1}^{n} A_l\right) = P(A_1) + P\left(\bigcup_{l=2}^{n} A_l\right) - P\left(\bigcup_{l=2}^{n} A_1 A_l\right) \tag{A.21}$$

and by (A.19) and (A.20) this corresponds to:

$$P\left(\bigcup_{l=1}^{n} A_l\right) = P(A_1) + \sum_{i=2}^{n} P(A_i) - \sum_{2\le i\le j\le n} P(A_i A_j) + \cdots +$$

$$+ \sum_{i=2}^{n} P(A_1 A_i) - \sum_{2\le i\le j\le n} P(A_1 A_i A_j) + \cdots = P_1 - P_2 + \ldots + (-1)^{n+1} P_n.$$

The proofs for all n follows by mathematical induction.

A.1.4 Conditional Probability

Given two events A and B, it can be interesting to know how the occurrence of one event influences the occurrence of the other one. This relationship is expressed through the *conditional* probability of A on the hypothesis B, i.e., the probability of observing A when B is know to have occurred:

$$P(A|B) = \frac{P(AB)}{P(B)} \tag{A.22}$$

where $AB = A \cap B$. Since $AB \subseteq B$, then $0 \le P(A|B) \le 1$. When A and B are mutually exclusive, the intersection AB is empty and the conditional probability is null. At the other extreme, if $A \subset B$, then $P(A|B) = 1$ because the event B imply the event A. If $A = \bigcup_k A_k$ and the A_k events are mutually exclusive; then it holds the following *addition law for conditional probabilities*:

$$P(A|B) = \sum_k P(A_k|B). \tag{A.23}$$

It is often convenient to express the probability of an event A as a sum of conditional probabilities with respect to an *exhaustive set* of mutually exclusive events B_k, where exhaustive means that $\bigcup_k B_k = \Omega$:

$$P(A) = \sum_k P(A|B_k)P(B_k). \tag{A.24}$$

Such equation can be demonstrated by observing that $A = \bigcup_k AB_k$ and $P(A)$ can thus be expressed as follows:

$$P(A) = \sum_k P(AB_k) = \sum_k \frac{P(AB_k)}{P(B_k)} P(B_k) \tag{A.25}$$

and, by (A.22), the above expression corresponds to Eq. (A.24).

A.1.5 Statistical Independence

Consider the case of two experiments with different sample spaces Ω_1 and Ω_2. If the experiments are performed always together, it can be interesting to know how the outcome of one experiment is influenced by the outcome of the other one. An example of such a situation is the rolling of two dice; in fact they can be considered as separate experiments leading to separate outcomes. The probability $P(A_1, A_2)$ of having outcome A_1 for the first experiment and A_2 for the second one can be estimated with the relative frequency:

$$P(A_1, A_2) \simeq \frac{n(A_1, A_2)}{n}. \tag{A.26}$$

If the number of trials n is sufficiently high and we take into account only the cases where the outcome of the second experiment is A_2, then we can estimate the probability of observing A_1 as outcome of the first experiment as follows:

$$P(A_1) \simeq \frac{n(A_1, A_2)}{n(A_2)}. \tag{A.27}$$

In fact, as $n \to \infty$, $n(A_2)$ tends to the infinity as well and the left side of the above equation corresponds to the relative frequency of the event A_1. This leads to the following expression for $P(A_1, A_2)$:

$$P(A_1, A_2) \simeq \frac{n(A_1, A_2)}{n} = \frac{n(A_1, A_2)}{n(A_2)} \frac{n(A_2)}{n} \simeq P(A_1)P(A_2) \tag{A.28}$$

when two experiments satisfy the above equation when $n \to \infty$, i.e., when $P(A_1, A_2) = P(A_1)P(A_2)$, they are said *statistically independent*. On the contrary, when $P(A_1, A_2) \neq P(A_1)P(A_2)$, the events are said to be *statistically dependent*.

A.2 Random Variables

This section provides the main notions about random variables and probability distributions. The rest of this section introduces the concepts of *mean value*, *variance*, *probability distribution* and *covariance*.

A.2.1 Fundamentals

A variable ξ is said *random* when its values depend on the events in the sample space of an experiment, i.e. when $\xi = \xi(\omega)$. Random variables are associated to functions

called *probability distributions* that give, for any couple of values x_1 and x_2 (with $x_1 \leq x_2$), the probability $P(x_1 \leq \xi \leq x_2)$ of ξ falling between x_1 and x_2. When ξ assumes values belonging to a finite set or to a countable infinity, the variable is called *discrete* and:

$$P(\xi = x) = p_\xi(x) \tag{A.29}$$

where $p_\xi(x)$ is the probability distribution of ξ. In this case the probability distribution is discrete as well and:

$$P(x_1 \leq \xi \leq x_2) \sum_{x=x_1}^{x_2} p_\xi(x) \tag{A.30}$$

where the sum is carried over all values between x_1 and x_2. If the sum is carried over all possible values of ξ, i.e., over the whole sample space underlying ξ, then the result is 1:

$$\sum_{x=-\infty}^{\infty} p_\xi(x) = 1. \tag{A.31}$$

When a random variable takes values in a continuous range, then it is said *continuous* and its distribution function is continuous as well:

$$P(x_1 \leq \xi \leq x_2) = \int_{x_1}^{x_2} p_\xi(x)dx. \tag{A.32}$$

where $p_\xi(x)$ is called the *probability density function*. If the integration domain covers the whole range of x, i.e., the whole sample space of the experiment underpinning ξ, then the result is 1:

$$\int_{-\infty}^{\infty} p_\xi(x)dx = 1. \tag{A.33}$$

While in the case of discrete variables it is possible to assign a probability to each value that ξ can take, in the case of the random variables it is only possible to have the probability $p_\xi(x)dx$ of ξ falling in a dx wide interval around x, i.e., of $\xi - x$ being smaller than an arbitrary value ϵ.

At each probability distribution function corresponds a *cumulative probability function* $F(x)$ that gives the probability $P(\xi \leq x)$ of ξ being less than x. In the case of discrete variables, $F(x)$ is a staircase function and it corresponds to the following sum:

$$F(x) \sum_{x'=-\infty}^{x} p_\xi(x'). \tag{A.34}$$

In the case of continuous random variables, $F(x)$ is:

$$F(x) = \int_{-\infty}^{x} p_{\xi}(x')dx' \tag{A.35}$$

and it is a continuous function.

Consider now the *random point* $\xi = (\xi_1, \xi_2)$. The probability of ξ corresponding to a point (x_1, x_2) is given by the *joint probability distribution* $p_{\xi_1\xi_2}(x_1, x_2)$:

$$p_{\xi_1\xi_2}(x_1, x_2) = P(\xi_1 = x_1, \xi_2 = x_2). \tag{A.36}$$

The probability $P(x_1' \leq \xi_1 \leq x_1'', x_2' \leq \xi_2 \leq x_2'')$ can be obtained by summing over the corresponding probabilities:

$$P(x_1' \leq \xi_1 \leq x_1'', x_2' \leq \xi_2 \leq x_2'') = \sum_{x_1=x_1'}^{x_1''} \sum_{x_2=x_2'}^{x_2''} p_{\xi_1\xi_2}(x_1, x_2), \tag{A.37}$$

the above is the probability of ξ falling in the region enclosed by the lines $\xi_1 = x_1'$, $\xi_1 = x_1''$, $\xi_2 = x_2'$ and $\xi_2 = x_2''$. When ξ_1 and ξ_2 are continuous variables, the sums are replaced by integrals and the above probability is written as follows:

$$P(x_1' \leq \xi_1 \leq x_1'', x_2' \leq \xi_2 \leq x_2'') = \int_{x_1'}^{x_1''} \int_{x_2'}^{x_2''} p_{\xi_1\xi_2}(x_1, x_2)dx_1dx_2 \tag{A.38}$$

where $p_{\xi_1\xi_2}(x_1, x_2)$ is called *joint probability density*.

The definitions given for two-dimensional random points can be extended to n-dimensional points corresponding to n-tuples of discrete or continuous random variables.

A.2.2 Mathematical Expectation

The *mathematical expectation* or *mean value* $\mathcal{E}[\xi]$ of a discrete random variable ξ corresponds to the following expression:

$$\mathcal{E}[\xi] = \sum_{x=-\infty}^{x=\infty} x p_{\xi}(x) \tag{A.39}$$

where the series is supposed to converge absolutely, i.e., it holds the following:

$$\sum_{x=-\infty}^{x=\infty} |x| p_\xi(x) < \infty. \tag{A.40}$$

A variable $\eta = \phi(x)$, where $\phi(\xi)$ is some function of ξ, is a random variable and $P(\eta = y)$ can be obtained as a sum of the $p_\xi(x)$ over the x values such that $\phi(x) = y$:

$$P(\eta = y) = \sum_{x:\phi(x)=y} p_\xi(x) \tag{A.41}$$

The mathematical expectation $\mathcal{E}[\eta]$ of η can thus be obtained as follows:

$$\mathcal{E}[\eta] = \sum_{y=-\infty}^{y=\infty} yP(\eta = y) = \sum_{y=-\infty}^{y=\infty} y \sum_{x:\phi(x)=y} p_\xi(x) = \sum_{x=-\infty}^{x=\infty} \phi(x) p_\xi(x) \tag{A.42}$$

and the above definition can be extended to a function of an arbitrary number n of random variables $\phi(\xi_1, \xi_2, \ldots \xi_n)$:

$$\mathcal{E}[\phi(\xi_1, \xi_2, \ldots, \xi_n)] = \sum_{x_1=-\infty}^{\infty} \ldots \sum_{x_n=-\infty}^{\infty} \phi(x_1, x_2, \ldots, x_n) p_{\xi_1 \xi_2 \ldots \xi_n}(x_1, \ldots, x_n) \tag{A.43}$$

The mean value of a linear combination of random variables is given by the linear combination of the mean values of the single variables:

$$\mathcal{E}[a\xi_1 + b\xi_2] = a\mathcal{E}[\xi_1] + b\mathcal{E}[\xi_2]. \tag{A.44}$$

In fact, based on Eq. (A.43), we can write:

$$\mathcal{E}[a\xi_1 + b\xi_2] = \sum_{x_1=-\infty}^{\infty} \sum_{x_2=-\infty}^{\infty} (ax_1 + bx_2) p_{\xi_1 \xi_2}(x_1, x_2) =$$

$$= a \sum_{x_1=-\infty}^{\infty} \sum_{x_2=-\infty}^{\infty} x_1 p_{\xi_1 \xi_2}(x_1, x_2) + b \sum_{x_1=-\infty}^{\infty} \sum_{x_2=-\infty}^{\infty} x_2 p_{\xi_1 \xi_2}(x_1, x_2) =$$

$$= a\mathcal{E}[\xi_1] + b\mathcal{E}[\xi_2].$$

When ξ_1 and ξ_2 are independent:

$$\mathcal{E}[\xi_1 \xi_2] = \sum_{x_1=-\infty}^{\infty} \sum_{x_2=-\infty}^{\infty} x_1 x_2 p_{\xi_1}(x_1) p_{\xi_2}(x_2) = \mathcal{E}[\xi_1]\mathcal{E}[\xi_2]. \tag{A.45}$$

When ξ is continuous, then the mathematical expectation is obtained as an integral:

$$\mathcal{E}[\xi] = \int_{-\infty}^{\infty} x p_{\xi}(x) dx. \tag{A.46}$$

For the variable $\eta = \phi(\xi)$, the mathematical expectation is:

$$\mathcal{E}[\eta] = \int_{-\infty}^{\infty} \phi(x) p_{\xi}(x) dx, \tag{A.47}$$

the demonstration follows the same steps as for the corresponding property of discrete variables (see above). The same applies for the mean value of a function $\phi(\xi_1, \ldots, \xi_n)$ of an arbitrary number n of random variables:

$$\mathcal{E}[\phi(\xi_1, \ldots, \xi_n)] = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \phi(\xi_1, \ldots, \xi_n) p_{\xi_1 \ldots \xi_n}(x_1, \ldots, x_n) dx_1 \ldots dx_2. \tag{A.48}$$

The properties demonstrated for the discrete variables can be demonstrated also for the continuous ones by replacing sums with integrals. This is possible because the formal properties of sums and integrals are the same.

A.2.3 Variance and Covariance

The *variance* (or *dispersion*) $D[\xi]$ of a random variable is the mathematical expectation $\mathcal{E}[(\xi - \mu)^2]$ of the quantity $(\xi - \mu)^2$, where $\mu = \mathcal{E}[\xi]$. The variance expression for a discrete variable is

$$D[\xi] = \mathcal{E}[(\xi - \mu)^2] = \sum_{x=-\infty}^{\infty} (x - \mu)^2 p_{\xi}(x) \tag{A.49}$$

while for a continuous variable it is:

$$D[\xi] = \mathcal{E}[(\xi - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 p_{\xi}(x) dx. \tag{A.50}$$

The properties of the variance can be demonstrated without distinguishing between continuous and discrete random variables; in fact they are mostly based on the properties of the mathematical expectation that have the same form for both continuous and discrete variables. It follows from the definition that:

$$D[\xi] = \mathcal{E}[(\xi - \mu)^2] = \mathcal{E}[(\xi^2 - 2\mu\xi + \mu^2)] = \mathcal{E}[\xi^2] - 2\mu\mathcal{E}[\xi] + \mu^2 = \mathcal{E}[\xi^2] - \mu^2, \tag{A.51}$$

then

$$D[c\xi] = \mathcal{E}[c^2\xi^2] - (\mathcal{E}[c\xi])^2 = c^2 D[\xi] \tag{A.52}$$

because $\mathcal{E}[c\xi] = c\mathcal{E}[\xi]$ (see the previous section).

If ξ_1 and ξ_2 are two independent random variables, then:

$$D[\xi_1 + \xi_2] = \mathcal{E}[(\xi_1 + \xi_2 - \mu_1 - \mu_2)^2] =$$

$$= \mathcal{E}[(\xi_1 - \mu_1)^2] + \mathcal{E}[(\xi_2 - \mu_2)^2] + 2\mathcal{E}[(\xi_1 - \mu_1)(\xi_2 - \mu_2)],$$

$$= D[\xi_1] + D[\xi_2] + 2\mathcal{E}[(\xi_1 - \mu_1)]\mathcal{E}[(\xi_2 - \mu_2)],$$

since $\mathcal{E}[(\xi_i - \mu_i)] = 0$, the above corresponds to:

$$D[\xi_1 + \xi_2] = D[\xi_1] + D[\xi_2]. \tag{A.53}$$

Consider a random point $\xi = (\xi_1, \xi_2, \ldots, \xi_n)$, the mathematical expectation of the product $(\xi_i - \mu_i)(\xi_j - \mu_j)$, where μ_i and μ_j are the mean values of ξ_i and ξ_j, respectively, is called *covariance* σ_{ij} of ξ_i and ξ_j:

$$\sigma_{ij} = \mathcal{E}[(\xi_i - \mu_i)(\xi_j - \mu_j)], \tag{A.54}$$

based on the above definition, $\sigma_{ii} = D[\xi_i]$. The $n \times n$ matrix Σ such that $\Sigma_{ij} = \sigma_{ij}$ is called *covariance matrix* of ξ and it has the variances of the ξ_i variables on the main diagonal.

Appendix B
Signal Processing

B.1 Introduction

The goal of this appendix is to provide basic notions about signal processing, the domain involving mathematical techniques capable of extracting from signals information useful for several tasks. The data considered in this book, i.e., audio recordings, images and videos, can be considered as signals and the techniques presented in this appendix are often applied to analyze them. Section B.2 is dedicated to a quick recall of complex numbers because most signal processing techniques include functions defined on the complex domain. Section B.3 is dedicated to the z-transform, a mathematical approach to represent signals through infinite series of powers that make easier to study the effect of systems (see Sect. 2.5). Section B.3.2 introduces the *Fourier transform*, a special case of the z-transform that enables us to analyze the frequency properties of signals. Section B.3.3 presents the Discrete Fourier Transform, a representation for periodic digital signals that can be applied also for finite lenght generic signals and represents sequences through sums of elementary sines and cosines. Section B.4 describes the *discrete cosine transform*, a representation commonly applied in image processing and close to the Discrete Fourier Transform.

The content of this appendix is particularly useful for understanding Chaps. 2, 3 and 12.

B.2 The Complex Numbers

The *complex numbers* are an extension of the real numbers containing all roots of quadratic equations. If j is the solution of the following equation:

$$x^2 = -1 \tag{B.1}$$

F. Camastra and A. Vinciarelli, *Machine Learning for Audio, Image and Video Analysis*, Advanced Information and Knowledge Processing,
DOI 10.1007/978-1-4471-6735-8

then the set **C** of complex numbers is represented in *standard form* as:

$$\{a + bj : a, b \in R\} \tag{B.2}$$

where the symbol : stands for *such that* and **R** is the set of the real numbers. A complex number is typically expressed with a single variable z, the number a is called *real part* $Re(z)$ of z, and b is called the *imaginary part* $Im(z)$ of z. The plan having as coordinates the values of a and b is called *complex* or z plan. Each point of such plan is a complex number and, vice versa, all complex numbers correspond to one point of such plan. The horizontal axis of the z plan is called the *real axis*, while the vertical one is defined *imaginary axis*. The sum and product between complex numbers are defined as follows:

$$(a + bj) + (c + dj) = (a + b) + (c + d)j \tag{B.3}$$

$$(a + bj)(c + dj) = (ac - bd) + (ad + bc)j \tag{B.4}$$

where the fact that $j^2 = -1$ is applied. Two complex numbers z_1 and z_2 that have the same real part a, but imaginary parts b and $-b$, respectively, are said to be *complex conjugates* and this is expressed by writing $z_2 = z_1^*$.

Since the complex numbers can be interpreted as vectors in the z plan, it is possible to define their *modulus*[2] $|z|$ as follows:

$$|z| = \sqrt{a^2 + b^2}. \tag{B.5}$$

The modulus can be calculated as $|z| = \sqrt{zz^*}$ and, as a consequence, $|z| = |z^*|$.

Since the complex numbers can be thought of as vectors in the z plan, it is possible to express them in *polar form* (see Fig. B.1). In fact, if $r = |z|$ and $\tan\theta = b/a$, then $a = r\cos\theta$ and $b = r\sin\theta$, and by the Euler's equation:

$$e^{j\theta} = r\cos\theta + j\sin\theta, \tag{B.6}$$

it is possible to write:

$$z = re^{i\theta}. \tag{B.7}$$

The number r is called the *magnitude* and the angle θ is called *argument* and expressed by $Arg(z)$. The argument of a complex number is not unique because z is not changed by adding integer multiples of 2π to θ. The argument in the interval $]-\pi, \pi]$ (where the] on the left side means that the left extreme is not included) is called *principal value*. The complex conjugate of $z = r(\cos\theta + j\sin\theta)$

[2]The modulus is the distance of the point representing a complex number from the origin of the plane where a and b are the axes.

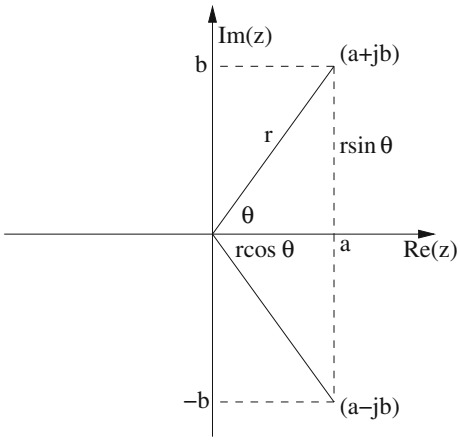

Fig. B.1 Complex plan. The figure shows a complex number and its complex conjugate in the complex plan. The real and imaginary parts can be expressed in terms of r and θ to obtain the polarform

is $r(\cos\theta - j\sin\theta)$, i.e. z^* is obtained by changing θ into $-\theta$. In other words, two complex conjugates have the same magnitude by opposite arguments.

The complex numbers $e^{j\theta}$, with $\theta \in [-\pi, \pi]$, define the so-called *unit circle* in the z plan. The equation $z^N = 1$ has N complex roots with magnitude 1. Since the roots are complex numbers, they can be identified as follows:

$$(e^{j\theta})^N = e^{jN\theta} = 1. \tag{B.8}$$

Since $e^{j\theta} = 1$ when $\theta = 2k\pi$ (where k is an integer), the last equation corresponds to $N\theta = 2k\pi$, then:

$$\theta = \frac{2k\pi}{N} \tag{B.9}$$

and the N roots of 1 are the complex exponentials $e^{j\frac{2k\pi}{N}}$, where $k = 0, 1, \ldots, N-1$.

When $k > N - 1$, the value of the argument is simply increased by multiples of 2π and the roots are the same as those corresponding to the k values between 0 and $N - 1$. The roots of 1 are used to represent periodic signals with the discrete Fourier transform (see Sect. B.3.3).

B.3 The z-Transform

Given a continuous signal $s(t)$, it is possible to obtain, through an A/D conversion including sampling and quantization, a *digital signal* $\{s[0], \ldots, s[N-1]\}$ such that:

$$s[n] = s(nT) = s(n/F) \tag{B.10}$$

where n is an integer, T is called sampling period and F is the sampling frequency. Issues related to sampling (see Sect. 2.3.1) and quantization (see Sect. 2.7) have been discussed in Chap. 2. A digital signal of length N, i.e. including N samples in the sequence $\{s[n]\}$, can be thought of as an infinite length signal such that $s[n] = 0$ for $n < 0$ and $n \geq N$. In the rest of this appendix, digital signals will be referred to as signals and denoted with $s[n]$ whenever there is no ambiguity between sequences and single samples.

The z-*transform* of a digital signal $\{s[n]\}$ is defined by the following pair of equations:

$$S(z) = \sum_{n=-\infty}^{\infty} s[n] z^{-n} \tag{B.11}$$

$$s[n] = \frac{1}{2\pi j} \oint_C S(z) z^{n-1} dz \tag{B.12}$$

where Eq. (B.11) defines the *direct* transform, Eq. (B.12) defines the *inverse* one and C is a closed contour that encircles the z plan origin and lies in the region of existence of $S(z)$ (see below).

The z-transform can be seen as an infinite series of powers of the variable z^{-1} where the $s[n]$ are the coefficients. The series converges to a finite value when the following sufficient condition is met:

$$\sum_{n=-\infty}^{\infty} |s[n]||z^{-n}| < \infty. \tag{B.13}$$

The above equation corresponds to a region of the z plan, called the *region of convergence*, which has the following form:

$$R_1 < |z| < R_2 \tag{B.14}$$

and the values of R_1 and R_2 depend on the characteristics of the sequence $\{s[n]\}$. Consider, for example, a rectangular window $w[n]$ of length N (see Sect. 2.5), the z-transform is:

$$W(z) = \sum_{n=0}^{N-1} z^{-n} = \frac{1 - z^{-N}}{1 - z^{-1}} \tag{B.15}$$

and the region of convergence is $0 < |z| < \infty$. Such a result applies to any finite length sequence.

Consider now the sequence $s[n] = a^n u[n]$, where $u[n]$ is 1 for $n \geq 0$ and 0 otherwise. In this case, the z-transform is:

$$S(z) = \sum_{n=0}^{\infty} a^n z^{-n} = \frac{1}{1 - az^{-1}} \tag{B.16}$$

and the series converges for $|z| > |a|$. This result applies to infinite length sequences which are non-zero only for $n \geq 0$ and it corresponds to a region of convergence of the form $|R_1| < |z| < \infty$.

The case of a sequence different from zero only when $n < 0$ can be studied by considering the case of $s[n] = b^n u[-n-1]$ (where $u[n]$ is the same function as in the previous example):

$$S[n] = \sum_{n=-\infty}^{-1} b^n z^{-n} = \frac{1}{1 - bz^{-1}}. \tag{B.17}$$

Such series converges for $|z| < |b|$ and, in terms of Eq. B.14, this corresponds to the form $0 < |z| < R_2$.

The last example concerns an infinite length sequence which is different from zero for $-\infty < n < \infty$. Such case is a combination of the last two examples and it leads to a region of convergence of the form $R_1 < |z| < R_2$.

B.3.1 z-Transform Properties

The z-transform has several properties that are demonstrated in the following. The first is the linearity:

Theorem B.1 *If $s[n] = as_1[n] + bs_2[n]$, then:*

$$S(z) = aS_1(z) + bS_2(z). \tag{B.18}$$

The demonstration follows directly from the definition of the z-transform:

$$S(z) = \sum_{n=-\infty}^{\infty} (s_1[n] + bs_2[n])z^{-n} = aS_1(z) + bS_2(z) \tag{B.19}$$

Consider the signal $s[n - n_0]$, where n_0 is a constant integer. The effect on the z-transform is described by the following theorem.

Theorem B.2 *The z-transform $S_{n_0}(z)$ of a signal $s_{n_0}[n] = s[n - n_0]$ is related to the z-transform $S(z)$ of $s[n]$ through the following relationship:*

$$S_{n_0}(z) = z^{-n_0} S(z). \tag{B.20}$$

The z-transform of $s_{n_0}[n]$ can be written as:

$$S_{n_0}(z) = \sum_{n=-\infty}^{\infty} s[n - n_0] z^{-n}, \tag{B.21}$$

if $m = n - n_0$, then the last equation becomes:

$$S_{n_0}(z) = \sum_{m=-\infty}^{\infty} s[m] z^{-m-n_0} = z^{-n_0} S(z) \tag{B.22}$$

The elements of a sequence can be weighted with an exponential resulting into a signal $s_a[n] = a^n s[n]$. The effect on the z-transform is as follows:

Theorem B.3 *The z-transform $S_a(z)$ of the signal $s_a[n] = a^n s[n]$ is related to the z-transform $S(z)$ of $s[n]$ through the following relationship:*

$$S_a(z) = S(za^{-1}). \tag{B.23}$$

Following the definition of the z-transform, it is possible to write that:

$$S_a(z) = \sum_{n=-\infty}^{\infty} a^n s[n] z^{-n} = \sum_{n=-\infty}^{\infty} s[n] \left(\frac{z}{a}\right)^{-n} = S(za^{-1}) \tag{B.24}$$

Theorem B.4 *The z-transform $S_n(z)$ of the signal $ns[n]$ is related to the z-transform $S(z)$ of $s[n]$ through the following expression:*

$$S_n(z) = -z \frac{dS(z)}{dz}. \tag{B.25}$$

The expression of $S_n(z)$ is:

$$S_n(z) = \sum_{n=-\infty}^{\infty} ns[n] z^{-n} = z \sum_{n=-\infty}^{\infty} ns[n] z^{-n-1}. \tag{B.26}$$

Since $-nz^{-n-1}$ is the derivative of z^{-n}, the above corresponds to:

$$S_n(z) = -z \sum_{n=-\infty}^{\infty} s[n] \frac{d(z^{-n})}{dz} = -z \frac{dS(z)}{dz} \tag{B.27}$$

Theorem B.5 *The* z*-transform* $S_-(z)$ *of the signal* $s[-n]$ *is related to the* z*-trasnform* $S(z)$ *of the signal* $s[n]$ *through the following expression:*

$$S_-(z) = S(z^{-1}). \tag{B.28}$$

Following the definition of the z-transform:

$$S_-(z) = \sum_{n=-\infty}^{\infty} s[-n] z^{-n}. \tag{B.29}$$

If $m = -n$, the above equation becomes:

$$S_-(z) = \sum_{m=-\infty}^{\infty} s[m] z^m = \sum_{m=-\infty}^{\infty} s[m] \left(\frac{1}{z}\right)^{-m} = S(z^{-1}) \tag{B.30}$$

Theorem B.6 *The* z*-transform* $C(z)$ *of the convolution of two digital signals* $c[n] = s[n] * h[n]$ *corresponds to the product of the* z*-transforms* $S(z)$ *and* $H(z)$ *of* $s[n]$ *and* $h[n]$*, respectively:*

$$C(z) = S(z)H(z). \tag{B.31}$$

The convolution between $s[n]$ and $h[n]$ is $c[n] = \sum_{k=-\infty}^{\infty} s[k]h[n-k]$, thus the z-transform of $c[n]$ is:

$$C(z) = \sum_{n=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} z^{-n} s[k]h[n-k] = \sum_{k=-\infty}^{\infty} s[k] \sum_{n=-\infty}^{\infty} h[n-k] z^{-n}. \tag{B.32}$$

If $n - k = m$, the above expression can be rewritten as:

$$C(z) = \sum_{k=-\infty}^{\infty} s[k] z^{-k} \sum_{m=-\infty}^{\infty} h[m] z^{-m} = S(z)H(z) \tag{B.33}$$

B.3.2 The Fourier Transform

The *Fourier Tranform* (FT) is defined through the following two equations:

$$S(e^{j\omega}) = \sum_{n=-\infty}^{\infty} s[n]e^{-j\omega n} \tag{B.34}$$

$$s[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} S(e^{j\omega})e^{j\omega n}d\omega \tag{B.35}$$

where Eq. (B.34) defines the inverse transform and Eq. (B.35) defines the inverse one. The FT corresponds to the z-transform when $z = e^{j\omega}$, i.e. when z lies on the unit circle of the z plan. Since $|e^{j\omega}| = 1$, the condition for the existence of the FT is (see Eq. (B.13)):

$$\sum_{n=-\infty}^{\infty} |s[n]| < \infty. \tag{B.36}$$

The region of convergence of the above series can be deduced from the examples described in Sect. B.3 by posing $z = e^{j\omega}$. This corresponds to impose as a condition that the unit circle lies in the region of convergence $R_1 < |z| < R_2$. In the case of finite-length sequences, when $R_1 = 0$ and $R_2 = \infty$, the FT exists always. For sequences different from zer only when $n \geq 0$, the region of convergence is $R_1 < |z| < \infty$ and the FT exists when $R_1 < 1$. For infinite-length sequences different from zero when $n < 0$, $R_1 = 0$ and R_2 is a finite constant; thus the FT exists when $R_2 > 1$. For the last example in Sect. B.3, i.e., an infinite length sequence different from zero for both $n < 0$ and $n \geq 0$, both R_1 and R_2 are finite constants and the FT exists when $R_1 < 1 < R_2$.

An important aspect of the FT is that it is a periodic function of ω with period 2π. This can be shown by replacing ω with $\omega + 2\pi$ in Eqs. (B.11) and (B.12), but also by observing that ω determines the position on the unit circle of the z plan (see Fig. B.2). When ω is increased by an integer multiple of 2π, the position on the unit circle is always the same; thus the FT has the same value.

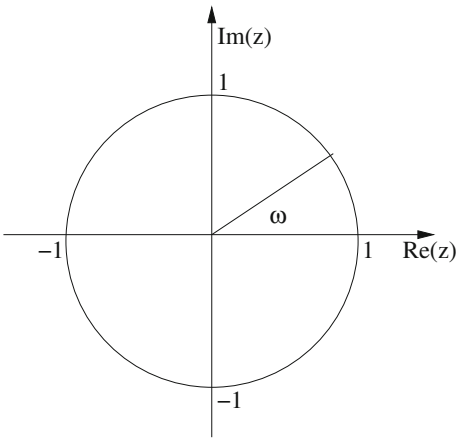

Fig. B.2 Unit circle. The figure shows the unit circle in the z plan. The angle ω identifies a point on the unit circle

The properties demonstrated in Sect. B.3.1 for the z-transform can be extended to the FT by simply replacing z with $e^{j\omega}$. However, the properties hold only when the FTs exist.

B.3.3 The Discrete Fourier Transform

If a digital signal $\hat{s}[n]$ is periodic with period N, i.e., $\hat{s}[n] = \hat{s}[n+N]$ for $-\infty < n < \infty$, then it can be represented by a Fourier series:

$$\hat{S}[k] = \sum_{n=0}^{N-1} \hat{s}[n] e^{-j\frac{2\pi}{N}kn} \tag{B.37}$$

$$\hat{s}[n] = \frac{1}{N} \sum_{k=0}^{N-1} \hat{S}[k] e^{j\frac{2\pi}{N}kn} \tag{B.38}$$

where Eq. (B.37) defines the direct transform and Eq. (B.38) defines the inverse one. The *discrete Fourier transform* (DFT) is an exact representation for any periodic digital signal, but it can be used, with some precautions, to represent finite-length nonperiodic sequences. In fact, consider the z-transform of a digital signal $s[n]$ which is equal to zero for $n < 0$ and $n \geq N$:

$$S(z) = \sum_{n=0}^{N-1} s[n] z^{-n}, \tag{B.39}$$

if $z = e^{j\frac{2\pi}{N}k}$, the above equation becomes:

$$S(e^{j\frac{2\pi}{N}k}) = \sum_{n=0}^{N-1} s[n] e^{-j\frac{2\pi}{N}kn}, \tag{B.40}$$

i.e., it corresponds to the $\hat{S}[k]$ value for a periodic signal $\hat{s}[n]$ obtained by replicating infinite times $s[n]$. In other words, given a finite length signal $s[n]$, it is possible to create an infinite length periodic signal $\hat{s}[n]$ such that $\hat{s}[n + rN] = s[n]$, where r is an integer. The DFT is an exact representation of $\hat{s}[n]$, but it can be used to represent $s[n]$ when only the intervals $0 \leq n \leq N-1$ and $0 \leq k \leq N-1$ are taken into account. Equation (B.40) can be thought of as a sampling of the z-transform on the unit circle of the z plan (see Fig. B.3). For this reason, the properties of the DFT are the same as those of the z-transform with the constraint that $z = \exp(-2j\pi kn/N)$.

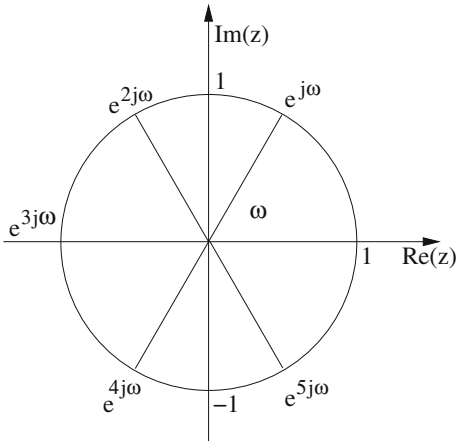

Fig. B.3 DFT interpretation. The DFT can be thought of as a sampling of the z-transform along the unit circle. The figure shows the points corresponding to integer multiples of the angle $\omega = \pi/6$, where the z-transform is sampled in the case of period 6

B.4 The Discrete Cosine Transform

The discrete cosine transform (DCT) is commonly applied in image coding and can be computed via the DFT. Given the N long signal $s[n]$, $0 \leq n < N$, it is possible to obtain a signal $s_e[n]$ of length $2N$ in the following way:

$$s_e[n] = \begin{cases} s[n] & 0 \leq n < N \\ 0 & N \leq n < 2N-1. \end{cases} \tag{B.41}$$

The signal $s_e[n]$ can then be used to create a $2N$ long sequence $y[n]$ defined as:

$$y[n] = s_e[n] + s_e[2N-1-n], \tag{B.42}$$

i.e., a symmetric signal where the first N samples correspond to those of the original $s[n]$ sequence and the remaining N correspond to the same samples, but in a reversed order (see Fig. B.4).

The DFT of $y[n]$ can be written as follows:

$$Y[k] = \sum_{n=0}^{2N-1} y[n] e^{-j\frac{2\pi}{2N}kn}, \tag{B.43}$$

but by definition (see Eq. (B.42)), $y[n] = y[2N-1-n]$, then the DFT of $y[n]$ can be rewritten as:

$$Y[k] = \sum_{n=0}^{N-1} s[n](e^{-j\frac{2\pi}{2N}kn} + e^{-j\frac{2\pi}{2N}k(2N-n-1)}). \tag{B.44}$$

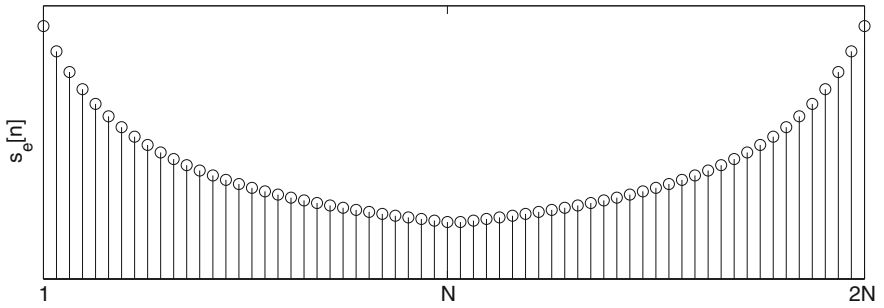

Fig. B.4 Extended signal. The plot shows a signal obtained by adding $s_e[n]$ and $s_e[2N-1-n]$

The N point DCT $C[k]$ of $s[n]$ is then defined as:

$$C(k) = \begin{cases} Y[k]e^{-j\frac{\pi}{2N}kn} & 0 \leq k < N. \\ 0 & otherwise. \end{cases} \tag{B.45}$$

By plugging Eq. (B.43) into the definition of $C(k)$, the result is (for $0 \leq k < N$):

$$C(k) = \sum_{n=0}^{N-1} 2s[n] \cos\left(\frac{(2n+1)k\pi}{2N}\right). \tag{B.46}$$

One of the most important aspects of the DCT is that its coefficients are always real, while in the case of the DFT they are typically complex. The DCT defined in this section is often referred to as *even symmetrical* DCT.

The inverse transform requires as a first step the definition of a $2N$-point DFT $Y[k]$:

$$Y(k) = \begin{cases} C[k]e^{-j\frac{2\pi}{2N}\frac{k}{2}} & 0 \leq k < N \\ 0 & k = N \\ -C[2N-k]e^{-j\frac{2\pi}{2N}\frac{k}{2}} & N+1 \leq k \leq 2N-1. \end{cases} \tag{B.47}$$

This enables us to obtain the inverse DFT as follows:

$$y[n] = \frac{1}{2N} \sum_{k=0}^{2N-1} Y[k]e^{j\frac{2\pi}{2N}kn} \tag{B.48}$$

where $0 \leq n \leq 2N-1$, and the inverse DCT corresponds to the following:

$$s[n] = \begin{cases} y[n] & 0 \leq n \leq N-1 \\ 0 & otherwise. \end{cases} \tag{B.49}$$

By definition (see (Eq. B.47)) $C[k] = C[2N-k]$ $(k = 1, \ldots, N-1)$, then $y[n]$ can be rewritten as a sum over N elements:

$$y[n] = \frac{1}{2N} C[0] + \frac{1}{2N} \sum_{k=1}^{N-1} C[k](e^{-j\frac{\pi k}{2N}(2n+1)} - e^{-j\frac{\pi}{2N}(2N-k)(2n+1)}) = \tag{B.50}$$

$$= \frac{1}{2N} C[0] + \frac{1}{N} \sum_{k=1}^{N-1} C[k] \cos\left(\frac{\pi k(2n+1)}{2N}\right). \tag{B.51}$$

In other words, we can write the inverse DCT as:

$$s[n] \begin{cases} \frac{1}{N} \sum_{k=1}^{N-1} \alpha(k) C[k] \cos\left(\frac{\pi k(2n+1)}{2N}\right) & 0 \leq n \leq N-1 \\ 0 & \textit{otherwise}, \end{cases} \tag{B.52}$$

where $\alpha(k) = 1/2$ for $k = 0$ and $\alpha(k) = 1$ for $k = 1$.

Appendix C
Matrix Algebra

C.1 Introduction

The goal of this appendix is to provide the main notions about matrix algebra and eigenvector calculation. Section C.2 introduces basic definitions and matrix operations, Sect. C.3 shows matrix determinants and their properties and Sect. C.4 presents eigenvalues and eigenvectors.

C.2 Fundamentals

An $m \times n$ matrix is a rectangular array of numbers composed of m rows and n columns:

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix} \tag{C.1}$$

where the a_{ij} entries are called *elements*. The above expression is often written in the following more compact form:

$$A = [a_{ij}]. \tag{C.2}$$

When $m = n$, i.e., the number of columns and rows is the same, the matrix is said to be *square* and the elements a_{ii}, where $i = 1, 2, \dots, n$, form the *main diagonal* of the matrix. The following sum:

F. Camastra and A. Vinciarelli, *Machine Learning for Audio, Image and Video Analysis*, Advanced Information and Knowledge Processing,
DOI 10.1007/978-1-4471-6735-8

$$T(A) = \sum_{k=1}^{n} a_{kk} \tag{C.3}$$

is called the *trace* of the matrix. The trace can be calculated only for square matrices because only these have a main diagonal. Given a matrix $A = [a_{ij}]$, the matrix $A^T = [a_{ji}]$ is called the *transpose* of A and it can be obtained by interchanging rows and columns of A. If $A^T = A$, the matrix is said *symmetric*. The transpose of an $m \times n$ matrix is an $n \times m$ one, thus a matrix cannot be symmetric if $m \neq n$, i.e., if it is not square.

Given two matrices $A = [a_{ij}]$ and $B = [b_{ij}]$, their sum is defined as follows:

$$A + B = [a_{ij} + b_{ij}]. \tag{C.4}$$

In other words, the element ij of the sum corresponds to the sum of the ij elements of A and B. The subtraction $A - B$ corresponds to the matrix where the element ij is $a_{ij} - b_{ij}$:

$$A - B = [a_{ij} - b_{ij}]. \tag{C.5}$$

The multiplication of a matrix A by a scalar c is defined as follows:

$$A = [ca_{ij}], \tag{C.6}$$

the result of such an operation is that each element of A is multiplied by c. The multiplication $A \cdot B$ between two matrices is calculated as follows:

$$A \cdot B = \left[\sum_{k} a_{ik} b_{kj} \right], \tag{C.7}$$

i.e., the element ij corresponds to the dot product of the ith row of A and of the jth column of B. This means that two matrices can be multiplied only when the number of columns of the first one is equal to the number of rows of the second one. A matrix $\mathbf{I}$ such that $A \cdot \mathbf{I} = A$ is called *identity matrix*.

Given two matrices X and Y, if the follwoing holds:

$$XY = YX = \mathbf{I} \tag{C.8}$$

then Y is the *inverse matrix* of X and viceversa. Only square matrices can be inverted because this is the only case where both XY and YX have the same number of rows and columns. In the case of a rectangular matrix X, it is possible to define the so-called *Moore Penrose pseudoinverse* $\hat{X}$:

$$\hat{X} = X^T (X^T X)^{-1}. \tag{C.9}$$

C.3 Determinants

Consider the following 2×2 square matrix:

$$A = \begin{pmatrix} a_{11} & a_{12} \\ c_{21} & a_{22} \end{pmatrix} \tag{C.10}$$

the expression $det(A) = a_{11}a_{22} - a_{12}a_{21}$ is called *determinant* of A. If the matrix is 3×3, then the determinant can be obtained as follows:

$$det(A) = \sum_{i=1}^{3} (-1)^{i+j} a_{ij} det(A_{ij}) = \sum_{j=1}^{3} (-1)^{i+j} a_{ij} det(A_{ij}) \tag{C.11}$$

where i and j can be selected arbitrarily and A_{ij} is a matrix obtained by removing column j and row i from A. In other words, the calculation of the determinant is performed by first selecting one column (or one row) arbitrarily, and then by multiplying its elements ij by $(-1)^{i+j} det(A_{ij})$. In order to simplify the calculations, it is advised to select the row or the column with the highest number of null elements. In fact this minimizes the number of addends of the sums in Eq. (C.11). When A is $n \times n$ with $n > 3$, $det(A)$ can be obtained recursively starting from the above expressions. Only square matrices have a determinant because only in this case the iterative removal of one column and one row brings to a 2×2 matrix.

The above technique is formalized in the *Laplace expansion theorem*:

Theorem C.1 *The determinant of a matrix A can be calculated as follows:*

$$det(A) = \sum M_k A_{n-k} \tag{C.12}$$

where the sum goes over all determinants M_k of order k that can be formed of rows $i_1, \ldots, i_k$ and columns $j_1, \ldots, j_k$, and A_{n-k} is the product of the number $(-1)^{i_1+\cdots+i_k+j_1+\cdots+j_k}$ and the determinant of the matrix remaining from A by deleting the rows $i_1, \ldots, i_k$ and the columns $j_1, \ldots, j_k$ used to form M_k.

The demonstration is omitted and the reader can refer to any academic textbook on matrix algebra.

The determinant of a matrix has several properties that are shown and proved in the following.

Theorem C.2 *If A is an $n \times n$ matrix and c is a scalar, then $det(cA) = c^n det(A)$.*

The proof can be obtained inductively. In the case of the 2×2 matrix of Eq. (C.10), the expression of $det(cA)$ is as follows:

$$det(cA) = ca_{11}ca_{22} - ca_{12}ca_{21} = c^2(a_{11}a_{22} - a_{12}a_{21}) = c^2 det(A). \tag{C.13}$$

If the property holds for an $n-1 \times n-1$ matrix, then the determinant of cA, where A is an $n \times n$ matrix, can be calculated as follows:

$$det(cA) = \sum_{i=1}^{n}(-1)^{i+j} ca_{ij} det(cA_{ij}) = \sum_{i=1}^{n}(-1)^{i+j} ca_{ij} c^{n-1} det(A_{ij}) = c^n det(A), \tag{C.14}$$

and this demonstrates the theorem.

Theorem C.3 *If A and B are $n \times n$ square matrices, then $det(AB) = det(A)det(B)$.*

Theorem C.4 *If A^T is the transpose of A, then $det(A^T) = det(A)$.*

The demonstration can be obtained by induction. When A is a 2×2 matrix like in Eq. (C.10), then A_T is as follows:

$$A = \begin{pmatrix} a_{11} & a_{21} \\ c_{12} & a_{22} \end{pmatrix}. \tag{C.15}$$

By the definition of determinant:

$$det(A) = a_{11}a_{22} - a_{12}a_{21} = det(A^T) \tag{C.16}$$

and the theorem is demonstrated for $n = 2$. If A is now an $n+1 \times n+1$ matrix, its determinant can be obtained as:

$$det(A) = \sum_{i=1}^{n+1}(-1)^{i+j} a_{ij} det(A_{ij}) = \sum_{i=1}^{n+1}(-1)^{i+j} (A^T)_{ji} det(A^T_{ji}) = det(A^T) \tag{C.17}$$

where the last passage is based on the fact that $(A_{ij})^T = A^T_{ji}$.

Theorem C.5 *Consider an $n \times n$ matrix A, $det(A) \neq 0$ if and only if A is nonsingular.*

The first step is to demonstrate that if A is nonsingular, then $det(A) \neq 0$. If A is non singular, A^{-1} exists and $AA^{-1} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. By property C.3:

$$det(AA^{-1}) = det(A)det(A^{-1}) = det(\mathbf{I}) = 1 \tag{C.18}$$

and this is possible only if $det(A) \neq 0$.

The second step is to prove that if $det(A) \neq 0$, then A^{-1} exists. The demonstration can be made by contradiction. If $det(A) = 0$ and A^{-1} exists, then $det(A)det(A^{-1}) = 1$ (see above), but this is not possible because $det(A) = 0$.

C.4 Eigenvalues and Eigenvectors

Consider the square matrix A, the scalar λ is defined *eigenvector* of A if it exists a nonzero vector x (called *eigenvector*) such that:

$$Ax = \lambda x, \tag{C.19}$$

where λ and x are said to form an *eigenpair*. If x is an eigenvector of A, then any other vector cx, where c is a scalar, is an eigenvector of A:

$$A(cx) = cAx = c\lambda x = \lambda(cx). \tag{C.20}$$

The eigenvectors form the basis of a vectorial space called *eigenspace*. When $\lambda = 0$, the eigenspace is called *null space* or *kernel* of A. When A is the identity matrix $\mathbf{I}$, the equation $\mathbf{I}x = x$ is always satisfied, i.e., all n-dimensional vectors are eigenvectors (with $\lambda = 1$) of the $n \times n$ identity matrix.

The eigenpairs can be found by solving the equation $Ax = \lambda x$ that can be rewritten as follows:

$$(A - \lambda \mathbf{I})x = 0. \tag{C.21}$$

where the second member is the null vector. The eigenvectors form the null space of the matrix $A - \lambda \mathbf{I}$ and they can be known once the eigenvalues are available. On the other hand, the above equation can have nonzero solutions only if $A - \lambda \mathbf{I}$ is singular, i.e., if

$$det\,(A - \lambda \mathbf{I}) = 0. \tag{C.22}$$

The above *characteristic equation* involves λ, but not x; however, the eigenvectors can be obtained when the last equation is solved and the eigenvectors are available.

Appendix D
Mathematical Foundations of Kernel Methods

D.1 Introduction

Mercer kernels (or *positive definite kernels*) are the foundations of powerful machine learning algorithms called *kernel methods*. Mercer kernels project implicitly the data in a high-dimensional *feature space* by means of a nonlinear mapping. The kernel theory has been developed during the first four decades of the twentieth century by some of the most brilliant mathematicians of the time. The concept of positive definite kernel has been introduced by [11]. Later on, remarkable contributions have been provided by [2, 3, 13, 16–18, 20]. In machine learning, the use of kernel functions to make computations, has been introduced by [1] in 1964. In 1995 a learning algorithm, *support vector machine* (*SVM*) [6] was introduced. SVM (see Chap. 9) uses Mercer kernels, as a preprocessing, to enpower a linear classifier (*optimal hyperplane algorithm*) so as to make the classifier able to solve nonlinear tasks.

The aim of this appendix is to present an overview of the kernel theory, focusing on the theoretical aspects that are relevant for kernel methods (see Chap. 9), such as the Mercer kernels and the reproducing kernel Hilbert spaces.

The appendix is organized as follows: in Sect. D.2 the definitions of scalar product, norm and metric are recalled; in Sect. D.3 positive definite functions and matrices are presented; Sect. D.4 is devoted to conditionate positive definite kernels and matrices; negative definite functions and matrices are described in Sects. D.5; D.6 presents the connections between negative and definite kernels; Sect. D.7 shows how a metric can be computed by means of a positive definite kernel; Sect. D.8 describes how a positive definite kernel can be represented by means of a Hilbert space. finally some conclusions are drawn in Sect. D.9.

F. Camastra and A. Vinciarelli, *Machine Learning for Audio, Image and Video Analysis*, Advanced Information and Knowledge Processing,
DOI 10.1007/978-1-4471-6735-8

D.2 Scalar Products, Norms and Metrics

The aim of this section is to recall the concepts of inner product, norm and metric [15].

Definition D.1 Let X be a set. A **scalar product** (or **inner product**) is an application $\cdot : X \times X \to \mathbb{R}$ satisfying the following conditions:

(a)	$y \cdot x = x \cdot y$	$\forall x, y \in X$	
(b)	$(x + y) \cdot z = (x \cdot z) + (y \cdot z)$	$\forall x, y, z \in X$	
(c)	$(\alpha x) \cdot y = \alpha(x \cdot y)$	$\forall x, y \in X$	$\forall \alpha \in \mathbb{R}$
(d)	$x \cdot x \geq 0$	$\forall x \in X$	
(e)	$x \cdot x = 0$	$\iff x = 0$	
(f)	$x \cdot (y + z) = (x \cdot y) + (x \cdot z)$	$\forall x, y, z \in X$	

Axioms (a) and (b) imply (f). Using axiom (d) it is possible to associate to the inner product a quadratic form $\|\cdot\|$, called the *norm*, such that:

$$\|x\|^2 = x \cdot x.$$

More generally, the norm can be defined in the following way.

Definition D.2 The seminorm $\|\cdot\| : X \to \mathbb{R}$ is a function that has the following properties:

$$\begin{array}{rcll} \|x\| & \geq & 0 & \forall x \in X \\ \|\alpha x\| & = & |\alpha| \|x\| & \forall \alpha \in \mathbb{R} \quad \forall x \in X \\ \|x + y\| & \leq & \|x\| + \|y\| & \forall x, y \in X \\ x = 0 \implies & & \|x\| = 0 & \forall x \in X \end{array}$$

Besides, if

$$x = 0 \iff \|x\| = 0 \tag{D.1}$$

the function $\|\cdot\| : X \to \mathbb{R}$ is called **norm**.

Norms and inner products are connected by the *Cauchy-Schwarz's inequality*:

$$|x \cdot y| \leq \|x\| \|y\|.$$

Definition D.3 Let X be a set. A function $\rho : X \times X \to \mathbb{R}$ is called a **distance** on X if:

(a)	$\rho(x, y) \geq 0$	$\forall x, y \in X$
(b)	$\rho(x, y) = \rho(y, x)$	$\forall x, y \in X$

(c) $\rho(x, x) = 0 \qquad \forall x \in X$

The (X, ρ) is called a **distance space**.

If ρ satisfies, in addition, the **triangle inequality**

(d) $\rho(x, y) \leq \rho(x, z) + \rho(y, z) \qquad \forall x, y, z \in X$

then ρ is called a **semimetric** on X.

Besides, if

(e) $\rho(x, y) = 0 \qquad \Rightarrow \qquad x = y$

In this case (X, ρ) is called a **metric space**.

It is easy to show that the function $\rho(x, y) = \|x - y\|$ is a *metric*. We conclude the section introducing the concept of L_p spaces.

Definition D.4 Consider countable sequences of real numbers and let $1 \leq p < \infty$. The L_p space is the set of sequences $z = z_1, \ldots, z_n, \ldots$ such that

$$\|z\|_p = \left(\sum_{i=1}^{\infty} |z_i|^p \right)^{\frac{1}{p}} < \infty$$

D.3 Positive Definite Kernels and Matrices

We now introduce the concept of positive definite matrices.

Definition D.5 A $n \times n$ matrix $A = (a_{jk}), \quad a_{jk} \in \mathbb{R}$, is called a **positive definite matrix** iff [3]

$$\sum_{j=1}^{n} \sum_{k=1}^{n} c_j c_k a_{jk} \geq 0 \tag{D.2}$$

for all $n \in \mathbb{N}$, $c_1, \ldots, c_n \subseteq \mathbb{R}$.

The basic properties of positive definite matrices are underlined by the following result.

Theorem 17 *A matrix is positive definite iff is symmetric and has all eigenvalues non-negative.*

A matrix is called *strictly positive definite* if all eigenvalues are positive. The following result (*Sylvester's criterion*), whose proof is omitted, is a useful tool to establish if a matrix is strictly positive definite.

[3] *iff* stands for *if and only if*.

Theorem 18 *Let $A = (a_{jk})$ be a symmetric $n \times n$ matrix. A is strictly positive definite iff*

$$det(a_{jk})_{j,k \leq p} > 0 \qquad p = 1, \ldots, n$$

i.e., all its minors have positive determinants.

Now we introduce the concept of *positive definite kernels*.

Definition D.6 Let X be a nonempty set. A function $\varphi : X \times X \to R$ is called a **positive definite kernel** (or **Mercer kernel**) iff

$$\sum_{j=1}^{n} \sum_{k=1}^{n} c_j c_k \varphi(x_j, x_k) \geq 0$$

for all $n \in \mathbb{N}$, $x_1, \ldots, x_n \subseteq X$ and $c_1, \ldots, c_n \subseteq \mathbb{R}$.

The following result, which we do not prove, underlines the basic properties of positive definite matrices.

Theorem 19 *A kernel φ on $X \times X$*

- *is positive definite iff is symmetric.*
- *is positive definite iff for every finite subset $X_0 \subseteq X$ the restriction of φ to $X_0 \times X_0$ is positive definite.*

Besides, if φ is positive definite, then $\varphi(x, x) \geq 0 \quad \forall x \in X$.

An example of Mercer kernel is the *inner product*, as stated by the following corollary.

Corollary 3 *The inner product is a positive definite (Mercer) kernel.*

Proof Applying the properties of the inner product, we have:

$$\sum_{j=1}^{n} \sum_{k=1}^{n} c_j c_k x_j \cdot x_k = \sum_{j=1}^{n} c_j x_j \cdot \sum_{j=1}^{n} c_j x_j = \| \sum_{j=1}^{n} c_j x_j \|^2 \geq 0$$

For Mercer kernels an inequality analogous to Cauchy Schwarz's one holds, as stated by the following result.

Theorem 20 *For any positive definite kernel φ the following inequality holds*

$$|\varphi(x, y)|^2 \leq \varphi(x, x)\varphi(y, y). \tag{D.3}$$

Proof Without losing generality, we consider the matrix

$$A = \begin{pmatrix} a & b \\ b & d \end{pmatrix}$$

where $a, b, d \in \mathbb{R}$. Then, for $w, z \in \mathbb{R}$ we have:

$$\begin{aligned}(w \quad z)\begin{pmatrix} a & b \\ b & d \end{pmatrix}\begin{pmatrix} w \\ z \end{pmatrix} &= aw^2 + 2bwz + dz^2 \\ &= a\left[w + \frac{b}{a}z\right]^2 + \frac{z^2}{a}\left[ad - b^2\right] \qquad (\forall a \neq 0)\end{aligned}$$

The matrix A is positive definite iff $a \geq 0$, $d \geq 0$ and

$$det\begin{pmatrix} a & b \\ b & d \end{pmatrix} = ad - b^2 \geq 0$$

Therefore for any positive definite kernel φ we have

$$|\varphi(x, y)|^2 \leq \varphi(x, x)\varphi(y, y)$$

Since both sides of the inequality are positive, we get:

$$|\varphi(x, y)| \leq \sqrt{\varphi(x, x)}\sqrt{\varphi(y, y)} \tag{D.4}$$

□

If we define $\|x\|_\varphi \triangleq \sqrt{\varphi(x, x)}$ a *pseudonorm*, the inequality (D.4) becomes

$$|\varphi(x, y)| \leq \|x\|_\varphi \|y\|_\varphi$$

that recalls the Cauchy Schwarz's inequality of the inner product.

The following remark underlines that $\|x\|_\varphi$ is a pseudonorm.

Remark 3 $\|x\|_\varphi$ is not a norm, since $x = 0$ does not imply $\|x\|_\varphi = 0$.

Proof We consider the kernel $\varphi(x, y) = \cos(x - y)$, $x, y \in \mathbb{R}$. φ is a Mercer kernel, since we have:

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \cos(x_i - x_j) &= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \left[\cos(x_i)\cos(x_j) + \sin(x_i)\sin(x_j)\right] \\ &= \left[\sum_{i=1}^{n} c_i \cos(x_i)\right]^2 + \left[\sum_{i=1}^{n} c_i \sin(x_j)\right]^2 \\ &\geq 0\end{aligned}$$

But $\|x\|_\varphi = 1 \quad \forall x$. □

Now we introduce the result that allows to use Mercer kernels **to make inner products**.

Theorem 21 *Let K be a symmetric function such that for all $x, y \in X$, $X \subseteq \mathbb{R}$*

$$K(x, y) \stackrel{\triangle}{=} \Phi(x) \cdot \Phi(y) \tag{D.5}$$

where $\Phi : X \rightarrow F$ and F, which is a Hilbert space,[4] *is called the* **feature space**.

K can be represented in terms of (D.5) iff $K = (K(x_i, x_j))_{i,j=1}^n$ is semi definite positive, i.e., K is a Mercer kernel.

Besides, K defines an **explicit** *mapping if Φ is known, otherwise the mapping is* **implicit**.

Proof We prove the proposition in the case of finite dimension space. Consider a space $X = [x_1, \ldots, x_n]$ and suppose that $K(x, y)$ is a symmetric function on X. Consider the matrix $K = (K(x_i, x_j))_{i,j=1}^n$. Since K is symmetric, an orthogonal matrix $V = [v_1, \ldots, v_n]$ exists such that $K = V \Lambda V^T$, where Λ is a diagonal matrix that has, the eigenvalues λ_i of K, as elements, while v_i are the eigenvectors of K.

Now we consider the following mapping $\Phi : X \rightarrow \mathbb{R}^n$

$$\Phi(x_i) \stackrel{\triangle}{=} (\sqrt{\lambda_t} v_{ti})_{t=1}^n$$

We have:

$$\Phi(x_i) \cdot \Phi(x_j) = \sum_{i=1}^{n} \lambda_t v_{ti} v_{tj} = (V \Lambda V^T)_{ij} = K_{ij} = K(x_i, x_j).$$

The requirement that all the eigenvalues of K are non-negative descends from the definition of Φ since the argument of the square root must be non-negative.

□

For the sake of completeness, we cite Mercer's theorem[5] which is the generalization of the Proposition D.5 for the infinite dimension spaces.

Theorem 22 *Let $X (X \subseteq \mathbb{R}^n)$ be a compact set. If K is a continuous symmetric function such that the operator T_K:*

$$(T_K f)(\cdot) = \int_X K(\cdot, x) f(x) dx \tag{D.6}$$

is positive definite, i.e.,

$$\int_{X \times X} K(x, y) f(x) f(y) dx dy \geq 0 \qquad \forall f \in L_2(X) \tag{D.7}$$

[4] see Sect. D.8.

[5] The theorem was originally proven for $X = [a, b]$. In [8] the theorem was extended to general compact spaces.

then we can expand $K(x, y)$ in a uniformly convergent series in terms of eigenfunctions $\Phi_j \in L_2(X)$ and positive eigenvalues $\lambda_j > 0$,

$$K(x, y) = \sum_{j=1}^{\infty} \lambda_j \Phi_j(x) \Phi_j(y) \tag{D.8}$$

It is necessary to point out the following remark.

Remark 4 The condition (D.7) corresponds to the condition (D.4) of the definition of the Mercer kernels in the finite case.

Now we provide examples of Mercer kernels that define implicit and explicit mapping. The kernel $K(x, y) = cos(x - y)$, $x, y \in \mathbb{R}$ defines an explicit mapping. Indeed, we have

$$K(x, y) = \cos(x - y) = \cos(x)\cos(y) + \sin(x)\sin(y)$$

that is the inner product in a Feature space F defined by the mapping $\Phi : \mathbb{R} \to \mathbb{R}^2$

$$\Phi(x) = \begin{pmatrix} \cos(x) \\ \sin(x) \end{pmatrix}$$

On the contrary, the Gaussian[6] $G = \exp(-\|x - y\|^2)$ is a case of a Mercer kernel with an implicit mapping, since Φ is unknown. The possibility to use Mercer kernels in order to perform inner product makes their study quite important for computer science. In the rest of this section we will present methods to make Mercer kernels.

D.3.1 How to Make a Mercer Kernel

The following theorem shows that Mercer kernels satisfy quite a number of properties.

Theorem 23 *Let φ_1 and φ_2 be Mercer kernels respectively over $X \times X$ and $X \subseteq \mathbb{R}^n$, $a \in \mathbb{R}^+$, $\cdot$ and $\bigotimes$ the inner and the tensor product, respectively.*

Then the following functions are Mercer kernels:

1. $\varphi(x, z) = \varphi_1(x, z) + \varphi_2(x, z)$
2. $\varphi(x, z) = a\varphi_1(x, z)$
3. $\varphi(x, z) = \varphi_1(x, z) \cdot \varphi_2(x, z)$
4. $\varphi(x, z) = \varphi_1(x, z) \bigotimes \varphi_2(x, z)$

[6] For the proof of the positive definiteness of the Gaussian see Corollary 9.

Proof The proofs of the first and the second properties are immediate.

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j [\varphi(x_i, x_j)] = \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \varphi_1(x_i, x_j) + \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \varphi_2(x_i, x_j) \geq 0$$

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j [a\varphi(x_i, x_j)] = a \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \varphi(x_i, x_j) \geq 0.$$

Since the product of positive definite matrices is still positive definite, the third property is immediately proved.

The tensor product of two positive definite matrices is positive definite, since the eigenvalues of the product are all pairs of products of the eigenvalues of the two components. □

The following corollaries provide useful methods in order to make Mercer kernels.

Corollary 4 *Let $\varphi(x, y) : X \times X \to \mathbb{R}$ be positive definite. The following kernels are also positive definite:*

1. $K(x, y) = \sum_{i=0}^{n} a_i [\varphi(x, y)]^n \qquad a_i \in \mathbb{R}^+$
2. $K(x, y) = \exp(\varphi(x, y))$

Proof The first property is an immediate consequence of the Theorem 23. Regarding the second item, the exponential can be represented as:

$$\exp(\varphi(x, y)) = 1 + \sum_{i=1}^{\infty} \frac{[\varphi(x, y)]^i}{i!}$$

and is a limit of linear combinations of Mercer kernels. Since Mercer kernels are closed under the pointwise limit, the item is proved. □

Corollary 5 *Let $f(\cdot) : X \to X$ be a function. Then $\varphi(x, y) = f(x) f(y)$ is positive definite.*

Proof We have:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j f(x_j) f(x_k) = \left(\sum_{i=1}^{n} c_i f(x_i)\right)^2 \geq 0$$

□

The foregoing propositions are very useful to make new Mercer kernels by means of existing Mercer kernels. Nevertheless to prove that a kernel is positive definite is generally not a trivial task. The following propositions, that we do not prove, are useful criteria that allow to state if a kernel is positive definite.

Theorem 24 (Bochner) *If $K(x-y)$ is a continuous positive definite function, then there exists a bounded nondecreasing function $V(u)$ such that $K(x-y)$ is a Fourier Stjelties transform of $V(u)$, that is:*

$$K(x-y) = \int_{-\infty}^{\infty} e^{i(x-y)u} dV(u)$$

If the function $K(x-y)$ satisfies this condition, then it is positive definite.

Theorem 25 (Schoenberg) *Let us call a function $F(u)$* completely monotonic *on $(0,\infty)$, provided that it is in $\mathbb{C}^{\infty}(0,\infty)$ and satisfies the condition:*

$$(-1)^n F^{(n)}(u) \geq 0 \qquad u \in (0,\infty), \qquad n = 0, 1, \ldots$$

Then the function $F(\|x-y\|)$ is positive definite iff $F(\sqrt{\|x-y\|})$ is continuous and completely monotonic.

Theorem 26 (Polya) *Any real, even, continuous function $F(u)$ which is convex on $(0,\infty)$, i.e., satisfies $F(\alpha u_1 + (1-\alpha)u_2 \leq \alpha F(u_1) + (1-\alpha) f(u_2)$ for all u_1, u_2 and $\alpha \in (0,1)$, is positive definite.*

On the basis of these theorems, one can construct different Mercer kernels of the type $K(x-y)$.

D.4 Conditionate Positive Definite Kernels and Matrices

Although the class of Mercer kernels is adequately populated, it can be useful to identify kernel functions that, although non-Mercer kernels, can be used, in similar way, to compute inner products. To this purpose we define the conditionate positive definite matrices and kernels [14].

Definition D.7 A $n \times n$ matrix $A = (a_{ij}) \quad a_{ij} \in \mathbb{R}$ is called a **conditionate positive definite matrix of order** r if it has $n-r$ non-negative eigenvalues.

Definition D.8 We call the kernel φ a **conditionate positive definite kernel of order r** iff is *symmetric* (i.e., $\varphi(x,y) = \varphi(y,x) \quad \forall x, y \in X$) and

$$\sum_{j=1}^{n} \sum_{k=1}^{n} c_j c_k \varphi(x_j, x_k) \geq 0$$

$\forall n \geq 2, x_1, \ldots, x_n \subseteq X$ and $c_1, \ldots, c_n \subseteq \mathbb{R}$, with $\sum_{j=1}^{n} c_j P(x) = 0$ where $P(x)$ is a *polynomial* of order $r-1$.

Examples of conditionate positive kernels are[7]:

$$k(x, y) = -\sqrt{\|x - y\|^2 + \alpha^2} \quad \alpha \in \mathbb{R} \quad \textit{Hardy multiquadric} \quad (r = 1)$$
$$k(x.y) = \|x - y\|^2 \ln \|x - y\| \quad \textit{thin plate spline} \quad (r = 2)$$

As pointed out by [10] conditionally positive definite kernels are admissible for methods that use a kernel to make inner products. This is underlined by the following result.

Theorem 27 *If a conditionate positive definite kernel $k(x, y)$ can be represented as $k(x, y) \stackrel{\triangle}{=} h(\|x - y\|^2)$, then $k(x.y)$ satisfies the Mercer condition (6).*

Proof In [9, 12] it was shown that conditionate positive definite kernels $h(\|x - y\|^2)$ generate semi-norms and $\|.\|_h$ defined by:

$$\|f\|_h^2 = \int h(\|x - y\|^2) f(x) f(y) dx dy \tag{D.9}$$

Since $\|f\|_h^2$ is a seminorm, $\|f\|_h^2 \geq 0$. Since the right side of (D.9) is the Mercer's condition for $h(\|x - y\|^2)$, $h(\|x - y\|^2)$ defines a scalar product in some feature space. Hence $k(x, y)$ can be used to perform an inner product.

□

This result enlarges remarkably the class of kernels, that can be used to perform inner products.

D.5 Negative Definite Kernels and Matrices

We introduce the concept of negative definite matrices.

Definition D.9 A $n \times n$ matrix $A = (a_{ij}) \quad a_{ij} \in \mathbb{R}$ is called a **negative definite matrix** iff

$$\sum_{j=1}^{n} \sum_{k=1}^{n} c_j c_k a_{jk} \leq 0 \tag{D.10}$$

$\forall n > 2, c_1, \ldots, c_n \subseteq \mathbb{R}$.

Since the previous definition involves integers $n > 2$, it is necessary to point out that any 1×1 matrix $A = (a_{11})$ with $a_{11} \in \mathbb{R}$ is called negative definite. The basic properties of negative definite matrices are underlined by the following result.

[7] The conditionate positive definiteness of Hardy multiquadrics is shown in Corollary 7.

Theorem 28 *A matrix is negative definite iff is symmetric and has all eigenvalues* ≤ 0.

A matrix is called *strictly negative definite* if all eigenvalues are negative.

Now we introduce the concept of the *negative definite kernels*.

Definition D.10 We call the kernel φ a **negative definite kernel** iff is symmetric (i.e., $\varphi(x, y) = \varphi(y, x) \ \ \forall x, y \in X$) and

$$\sum_{j=1}^{n} \sum_{k=1}^{n} c_j c_k \varphi(x_j, x_k) \leq 0$$

$\forall n \geq 2, x_1, \ldots, x_n \subseteq X$ and $c_1, \ldots, c_n \subseteq \mathbb{R}$ with $\sum_{j=1}^{n} c_j = 0$.

In analogy with the positive definite kernel, the following result holds:

Theorem 29 *A kernel φ is negative definite iff for every finite subset $X_0 \subseteq X$ the restriction of φ to $X_0 \times X_0$ is negative definite.*

An example of a negative definite kernel is the square of the Euclidean distance.

Corollary 6 *The kernel $\varphi(x, y) = \|x - y\|^2$ is negative definite.*

Proof We have:

$$\begin{aligned}
\sum_{i,j=1}^{n} c_i c_j \varphi(x_j, x_k) &= \sum_{i,j=1}^{n} c_i c_j \|x_i - x_j\|^2 \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j [\|x_i\|^2 - 2(x_i \cdot x_j) + \|x_j\|^2] \\
&= \sum_{j=1}^{n} c_j \sum_{i=1}^{n} c_i \|x_i\|^2 + \sum_{i=1}^{n} c_i \sum_{j=1}^{n} c_j \|x_j\|^2 - 2 \sum_{i,j=1}^{n} c_i c_j (x_i \cdot x_k) \\
&= -2 \sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j (x_i \cdot x_k) \qquad \left(since \sum_{j=1}^{n} c_j = 0 \right) \\
&\leq 0
\end{aligned}$$

since the inner product is positive definite. □

Important properties of negative definite kernels are stated by the following lemma, whose proof [4] is omitted for the sake of brevity.

Lemma D.11 *If $\psi : X \times X \to \mathbb{R}$ is negative definite and satisfies $\psi(x, x) \geq 0$ $\forall x \in X$ then the following kernels are negative definite*

- ψ^α *for* $0 < \alpha < 1$.
- $\log(1 + \psi)$

Consequence of the lemma is that *Hardy multiquadrics* is a conditionate positive definite kernel.

Corollary 7 *The Hardy multiquadrics* $-\sqrt{\alpha^2 + \|x - y\|^2}$ *is a conditionate positive definite kernel of order 1, for* $\alpha \in \mathbb{R}$.

Proof The kernel $\psi(x, y) = \alpha^2 + \|x - y\|^2$ is negative definite,

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j [\alpha^2 + \|x_i - x_j\|^2] &= \alpha^2 \left(\sum_{i=1}^{n} c_i\right)^2 + \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \|x_i - x_j\|^2 \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \|x_i - x_j\|^2 \qquad \left(since \sum_{i=1}^{n} c_i = 0\right) \\ &\leq 0\end{aligned}$$

for Corollary 6.

Therefore, for the previous lemma, $\varphi(x, y) = \psi(x, y)^{\frac{1}{2}}$ is still negative definite. Hence, the opposite of φ, i.e., the Hardy multiquadrics, is a conditionate positive definite kernel of order 1. □

One consequence of Lemma D.11 is the following fundamental result that characterizes negative definite kernels.

Corollary 8 *The Euclidean distance is negative definite. More generally, the kernel* $\psi(x, y) = \|x - y\|^\alpha$ *is negative definite for* $0 < \alpha \leq 2$.

Proof The result is immediate consequence of Corollary 6 and Lemma D.11. □

D.6 Relations Between Positive and Negative Definite Kernels

Positive and negative definite kernels are strictly connected. If K is positive definite then $-K$ is negative definite. On the contrary, if K is negative definite, then $-K$ is a *conditionate positive definite kernel of order 1*. Besides, positive and negative definite functions are related by the following lemma.

Lemma D.12 *Let* X *be a nonempty set,* $x_0 \in X$, *and let* $\psi : X \times X \to \mathbb{R}$ *be a symmetric kernel. Put* $\varphi(x, y) := \psi(x, x_0) + \psi(y, x_0) - \psi(x, y) - \psi(x_0, y_0)$. *Then* φ *is positive definite iff* ψ *is negative definite.*

If $\psi(x_0, x_0) \geq 0$ *and* $\varphi_0(x, y) := \psi(x, x_0) + \psi(y, x_0) - \psi(x, y)$, *then* φ_0 *is positive definite iff* ψ *is negative definite.*

Proof For $c_1, \ldots, c_n \in \mathbb{R}$, $\sum_{j=1}^{n} c_j = 0$ and $x_1, \ldots, x_n \in X$ we have

$$\begin{aligned}\sum_{j=1}^{n}\sum_{i=1}^{n} c_i c_j \varphi(x_i, x_j) &= \sum_{j=1}^{n}\sum_{i=1}^{n} c_i c_j \varphi_0(x_i, x_j) \\ &= -\sum_{j=1}^{n}\sum_{i=1}^{n} c_i c_j \psi(x_i, x_j).\end{aligned}$$

Therefore positive definiteness of φ implies the negative definiteness of ψ.

On the other hand, suppose that ψ is negative definite. Let $c_1, \ldots, c_n \in \mathbb{R}$ and $x_1, \ldots, x_n \in X$. We put $c_0 = -\sum_{j=1}^{n} c_j = 0$. Then

$$\begin{aligned}0 &\geq \sum_{j=0}^{n}\sum_{i=0}^{n} c_i c_j \psi(x_i, x_j) \\ 0 &\geq \sum_{j=1}^{n}\sum_{i=1}^{n} c_i c_j \psi(x_i, x_j) + \sum_{j=1}^{n} c_j c_0 \psi(x_j, x_0) + \sum_{i=1}^{n} c_i c_0 \psi(x_i, x_0) + \|c_0\|^2 \psi(x_0, x_0) \\ 0 &\geq \sum_{j=1}^{n}\sum_{i=1}^{n} c_i c_j [\psi(x_i, x_j) - \psi(x_j, x_0) - \psi(x_i, x_0) + \psi(x_0, x_0)] \\ 0 &\geq -\sum_{j=1}^{n}\sum_{i=1}^{n} c_i c_j \varphi(x_i, x_j)\end{aligned}$$

Hence φ is positive definite. Finally, if $\psi(x_0, x_0) \geq 0$ then

$$\sum_{j=1}^{n}\sum_{i=1}^{n} c_i c_j \varphi_0(x_i, x_j) = \sum_{j=1}^{n}\sum_{i=1}^{n} c_i c_j [\varphi(x_i, x_j)] + \psi(x_0, x_0)\left(\sum_{j=1}^{n} c_j\right)^2 \geq 0$$

□

The following theorem is very important since it allows us to prove that the Gaussian kernel is positive definite.

Theorem 30 (Schoenberg) *Let X be a nonempty set and let $\psi : X \times X \to \mathbb{R}$ be a kernel. Then ψ is negative definite iff* $\exp(-t\psi)$ *is positive definite* $\forall t > 0$.

Proof If $exp(-t\psi)$ is positive definite then $1 - exp(-t\psi)$ is negative definite

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j [1 - exp(-t\psi)] &= \left(\sum_{i=1}^{n} c_i\right)^2 - \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \ \ exp(-t\psi) \\ &= -\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \ \ exp(-t\psi) \ \ \left(since \sum_{i=1}^{n} c_i = 0\right) \\ &\leq 0\end{aligned}$$

since $\exp(-t\psi)$ is definite positive.

The negative definite is also the limit

$$lim_{t\to 0^+} \frac{1}{t}(1 - exp(-t\psi)) = \psi$$

On the other hand, suppose that ψ is negative definite. We show that for $t = 1$, the kernel $\exp(-t\psi)$ is positive definite. We choose $x_0 \in X$ and, for Lemma D.12, we have:

$$-\psi(x, y) = \varphi(x, y) - \psi(x, x_0) - \psi(y, x_0) + \psi(x_0, x)$$

where φ is positive definite. Since

$$\exp(-\psi(x, y)) = \exp(\varphi(x, y)) \exp(-\psi(x, x_0)) \exp(-\psi(y, x_0)) \exp(\psi(x_0, x))$$

we conclude that $\exp(-\psi)$ is positive definite. The generic case $\forall t > 0$, can be derived for induction. □

An immediate consequence of the previous theorem is the following result.

Corollary 9 *The Gaussian* $\exp(-\frac{\|x-y\|^2}{\sigma^2})$ *is positive definite, for* $x, y \in \mathbb{R}^n$ *and* $\sigma \in \mathbb{R}$.

More generally, $\psi(x, y) = \exp(-a\|x - y\|^\alpha)$*, with* $a > 0$ *and* $0 < \alpha \geq 2$*, is positive definite.*

Proof The kernel $\|x - y\|^\alpha$ with $0 < \alpha \geq 2$ is negative definite, as shown in Corollary 8. Therefore for Theorem 30 the Gaussian is positive definite. □

We conclude this section reporting, without proving them, the following results.

Lemma D.13 *A kernel* $\psi : X \times X \to \mathbb{R}$ *is negative definite iff* $(t + \psi)^{-1}$ *is positive definite* $\forall t > 0$.

Theorem 31 *A kernel* $\psi : X \times X \to \mathbb{R}$ *is negative definite iff its Laplace transform* $\mathbb{L}(t\psi)$

$$\mathbb{L}(t\psi) = \int_0^{+\infty} \exp(-ts\psi) d\mu(s)$$

is positive definite $\forall t > 0$.

Consequence of the Lemma D.13, is the following result.

Corollary 10 *Inverse Hardy multiquadrics* $\psi(x, y) = (\alpha^2 + \|x - y\|^2)^{-\frac{1}{2}}, \alpha \in \mathbb{R}$ *is positive definite.*

Proof Since $(\alpha^2 + \|x - y\|^2)^{-\frac{1}{2}}$ is definite negative (see Corollary 7), Inverse Hardy multiquadrics is definite positive for Lemma D.13. □

D.7 Metric Computation by Mercer Kernels

In this section we show how to compute a metric by means of a Mercer kernel. Thanks to a fundamental result [16, 17], it is possible to associate a metric to a kernel. In order to show that we consider, associated to a Mercer kernel K, the kernel $d(x, y)$:

$$d(x, y) \stackrel{\triangle}{=} K(x, x) - 2K(x, y) + K(y, y).$$

The kernel $d(x, y)$ is negative definite.

Corollary 11 *If* $K(x, y)$ *is positive definite then d(x,y) is negative definite. Besides,* $\sqrt{d(x, y)}$ *is negative definite.*

Proof We have:

$$\begin{aligned}
\sum_{i,j=1}^{n} c_j c_i d(x_j, x_i) &= \sum_{i,j=1}^{n} c_j c_i [K(x_i, x_i) - 2K(x_i, x_j) + K(x_j, x_j)] \\
&= \sum_{j=1}^{n} c_j \sum_{i=1}^{n} c_i K(x_i, x_i) - 2 \sum_{i,j=1}^{n} c_j c_i K(x_i, x_j) + \sum_{i=1}^{n} c_i \sum_{j=1}^{n} c_j K(x_j, x_j)] \\
&= -2 \sum_{i=1}^{n} \sum_{j=1}^{n} c_j c_i K(x_i, x_j) \qquad \left(since \sum_{j=1}^{n} c_j = 0 \right) \\
&\leq 0
\end{aligned}$$

since K is definite positive.

Now we show that $d(x, y) \geq 0$

$$\begin{aligned}
d(x, y) &= K(x, x) - 2K(x, x)K(y, y) + K(y, y) \\
&\geq K(x, x) - 2\sqrt{K(x, x)K(y, y)} + K(y, y) \\
&\geq \left[\sqrt{K(x, x)} - \sqrt{K(y, y)}\right]^2 \\
&\geq 0.
\end{aligned}$$

Hence, for Lemma D.11, $\sqrt{d(x, y)}$ is negative definite. □

Now we introduce the result [16, 17].

Theorem 32 (Schoenberg) *Let X be a nonempty set and $\psi : X \times X \to \mathbb{R}$ be negative definite. Then there is a space $H \subseteq \mathbb{R}^X$ and a mapping $x \mapsto \varphi_x$ from X to H such that*

$$\psi(x, y) = \|\varphi_x - \varphi_y\|^2 + f(x) + f(y)$$

where $f : X \to \mathbb{R}$. The function f is non-negative whenever ψ is.
If $\psi(x, x) = 0 \ \ \forall x \in X$ then $f = 0$ and $\sqrt{\psi}$ is a metric on X.

Proof We fix some $x_0 \in X$ and define

$$\varphi(x, y) = \frac{1}{2}\left[\psi(x, x_0) + \psi(y, x_0) - \psi(x, y) - \psi(x_0, y_0)\right]$$

which is positive definite for Lemma D.12. Let H be the associated space for φ and put $\varphi_x(y) = \varphi(x, y)$. Then

$$\begin{aligned}\|\varphi_x - \varphi_y\|^2 &= \varphi(x, x) + \varphi(y, y) - 2\varphi(x, y) \\ &= \psi(x, y) - \frac{1}{2}\left[\psi(x, x) + \psi(y, y)\right]\end{aligned}$$

By setting $f(x) := \frac{1}{2}\psi(x, x)$ we have:

$$\psi(x, y) = \|\varphi_x - \varphi_y\|^2 + f(x) + f(y).$$

The other statements can be derived immediately. □

As pointed out by [7], the negative definiteness of the metric is a property of L_2 spaces. Schoenberg's theorem can be reformulated in the following way:

Theorem 33 *Let X be a L_2 space. Then the kernel $\psi : X \times X \to \mathbb{R}$ is negative definite iff $\sqrt{\psi}$ is a metric.*

An immediate consequence of Schoenberg's theorem is the following result.

Corollary 12 *Let $K(x, y)$ be a positive definite kernel. Then the kernel*

$$\rho_K(x, y) \stackrel{\triangle}{=} \sqrt{K(x, x) - 2K(x, y) + K(y, y)}$$

is a metric.

Proof The kernel $d(x, y) = K(x, x) - 2K(x, y) + K(y, y)$ is negative definite. Since $d(x, x) = 0 \ \ \forall x \in X$, applying Theorem 1 we get that $\rho_K(x, y) \stackrel{\triangle}{=} \sqrt{d(x, y)}$ is a distance. □

Hence, it is always possible to compute a metric by means of a Mercer kernel, even if an implicit mapping is associated with the Mercer kernel. When an implicit mapping is associated to the kernel, it cannot compute the positions $\Phi(x)$ e $\Phi(y)$ in

the feature space of two points x and y; nevertheless it can compute their distance $\rho_K(x, y)$ in the feature space. Finally, we conclude this section, providing metric examples that can be derived by Mercer kernels.

Corollary 13 *The following kernels* $\rho : X \times X \to \mathbb{R}^+$

- $\rho(x, y) = \sqrt{2 - 2\exp(-\|x - y\|^\alpha)}$ *with* $0 < \alpha < 2$
- $\rho(x, y) = \sqrt{(\|x\|^2 + 1)^n + (\|y\|^2 + 1)^n - 2(x \cdot y + 1)^n}$ *with* $n \in \mathbb{N}$

are metrics.

Proof Since $(x \cdot y + 1)^n$ and $\exp(-\|x - y\|^\alpha)$ with $0 < \alpha < 2$ are Mercer kernels, the statement, by means of the Corollary 12, is immediate. □

D.8 Hilbert Space Representation of Positive Definite Kernels

First, we recall some basic definitions in order to introduce the concept of *Hilbert space*.

Definition D.14 A set is a **linear space** (or **vector space**) if the addition and the multiplication by a scalar are defined on X such that, $\forall x, y \in X$ and $\alpha \in \mathbb{R}$

$$\begin{aligned} x + y &\in X \\ \alpha x &\in X \\ 1x &= x \\ 0x &= 0 \\ \alpha(x + y) &= \alpha x + \alpha y \end{aligned}$$

Definition D.15 A sequence x_n in a *normed linear space*[8] is said to be a **Cauchy sequence** if $\|x_n - x_m\| \to 0$ for $n, m \to \infty$.

A space is said to be **complete** when every Cauchy sequence converges to an element of the space.

A complete normed linear space is called a **Banach space**.

A Banach space where an inner product can be defined is called a **Hilbert space**.

Now we pass to represent positive definite kernels in terms of a *reproducing kernel Hilbert space* (*RKHS*).

Let X be a nonempty set and $\varphi : X \times X \to \mathbb{R}$ be positive definite. Let H_0 be the space the subspace of $\mathbb{R}^X$ generated by the functions $\{\varphi_x | x \in X\}$ where $\varphi_x(y) = \varphi(x, y)$.

[8] A *normed linear space* is a linear space where a norm function $\|\cdot\| : X \to \mathbb{R}$ is defined that maps each element $x \in X$ into $\|x\|$.

If $f = \sum_j c_j \varphi_{x_j}$ and $g = \sum_i d_i \varphi_{y_i}$, with $f, g \in H_0$, then

$$\sum_i d_i f(y_i) = \sum_{i,j} c_j d_i \varphi(x_j, y_i) = \sum_j c_j g(x_j) \tag{D.11}$$

The foregoing formula does not depend on the chosen representations of f and g and is denoted $\langle f, g \rangle$. Then the inner product $\langle f, g \rangle = \sum_{i,j} c_i c_j \varphi(x_i, x_j) \geq 0$ since φ is definite positive. Besides, the form $\langle \cdot, \cdot \rangle$ is linear in both arguments.

A consequence of (D.11) is the *reproducing property*

$$\begin{aligned} \langle f, \varphi_x \rangle &= \sum_j c_j \varphi(x_j, x) = f(x) \qquad \forall f \in H_0 \quad \forall x \in X \\ \langle \varphi_x, \varphi_y \rangle &= \varphi(x, y) \qquad \forall x, y \in X \end{aligned}$$

Moreover, using Cauchy Schwarz's inequality, we have:

$$\begin{aligned} \|\langle f, \varphi_x \rangle\|^2 &\leq \langle \varphi_x, \varphi_x \rangle \langle f, f \rangle \\ |f(x)|^2 &\leq \langle f, f \rangle \varphi(x, x) \end{aligned} \tag{D.12}$$

Therefore $\langle f, f \rangle = 0 \Longleftrightarrow f(x) = 0 \quad \forall x \in X$.

Hence, the form $\langle \cdot, \cdot \rangle$ is an *inner product* and H_0 is a *Pre-Hilbertian space*.[9] $\mathbb{H}$, the completion of H_0, is a *Hilbert space*, in which H_0 is a dense subspace. The Hilbert function space $\mathbb{H}$ is usually called the *reproducing kernel Hilbert space* (*RKHS*) associated to the Mercer kernel φ. Hence, the following result has been proved.

Theorem 34 *Let $\varphi : X \times x \to \mathbb{R}$ be a Mercer kernel.*

Then there is a Hilbert space $\mathbb{H} \subseteq \mathbb{R}^X$ and a mapping $x \mapsto \varphi_x$ from X to $\mathbb{H}$ such that

$$\langle \varphi_x, \varphi_y \rangle = \varphi(x, y) \qquad \forall x, y \in X$$

i.e., φ for $\mathbb{H}$ is the **reproducing kernel**.

D.9 Conclusions

In this appendix, the mathematical foundations of the Kernel methods have been reviewed focusing on the theoretical aspects which are relevant for Kernel methods. First we have reviewed Mercer kernels. Then we have described negative kernels underlining the connections between Mercer and negative kernels. We have also described how a positive definite kernel can be represented by means of a Hilbert space. We conclude the appendix providing some bibliographical remarks. Mercer

[9] A Pre-Hilbertian space is a normed, noncomplete space where an inner product is defined.

kernel and RKHS are fully discussed in [3] which also represents a milestone in the kernel theory. A good introduction to the Mercer kernels, more accessible to less experienced readers, can be found in [4]. Finally, the reader can find some mathematical topics of the kernel theory discussed in some handbooks on Kernel methods, such as [19, 21].

References

1. M. Aizerman, E. Braverman, and L. Rozonoer. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821–837, 1964.
2. N. Aronszajn. La theorie generale de noyaux reproduisants et ses applications. *Proc. Cambridge Philos. Soc.*, 39:133–153, 1944.
3. N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68:337–404, 1950.
4. C. Berg, J.P.R. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups*. Springer-Verlag, 1984.
5. C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20(3):273–297, 1995.
6. N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.
7. M. Deza and M. Laurent. Measure aspects of cut polyhedra: l_1-embeddability and probability. Technical report, Departement de Mathematiques et d' Informatique, Ecole Normale Superieure, 1993.
8. N. Dumford and T. J. Schwarz. *Linear Operators Part II: Spectral Theory, Self Adjoint Operators in Hilbert Spaces*. John Wiley, 1963.
9. N. Dyn. Interpolation and approximation by radial and related functions. In *Approximation Theory*, pages 211–234. Academic Press, 1991.
10. F. Girosi. Priors, stabilizers and basis functions: From regularization to radial, tensor and additive splines. Technical report, MIT, 1993.
11. D. Hilbert. Grundzüge einer allgemeinen theorie der linearen integralgleichungen. *Nachr. Göttinger Akad. Wiss. Math. Phys. Klasse*, 1:49–91, 1904.
12. W. R. Madych and S. A. Nelson. Multivariate interpolation and conditionally positive definite functions. *Mathematics of Computation*, 54:211–230, 1990.
13. J. Mercer. Functions of positive and negative type and their connection with the theory of integral equations. *Philos. Trans. Royal Soc.*, A209:415–446, 1909.
14. C. A. Micchelli. Interpolation of scattered data: distance matrices and conditionally positive definite. *Constructive Approximation*, 2:11–22, 1986.
15. W. Rudin. *Real and Complex Analysis*. Mc Graw-Hill, 1966.
16. I. J. Schoenberg. Metric spaces and completely monotone functions. *Ann. of Math.*, 39:811–841, 1938.
17. I. J. Schoenberg. Metric spaces and positive definite functions. *Trans. Amer. Math. Soc.*, 44:522–536, 1938.
18. I. J. Schoenberg. Positive definite functions on spheres. *Duke. Math. J.*, 9:96–108, 1942.
19. B. Schölkopf and A.J. Smola. *Learning with Kernels*. MIT Press, 2002.
20. I. Schur. Bemerkungen zur theorie der beschränkten bilininearformen mit unendlich vielen veränderlichen. *J. Reine Angew. Math.*, 140:1–29, 1911.
21. J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

Index

F. Camastra and A. Vinciarelli, *Machine Learning for Audio, Image and Video Analysis*, Advanced Information and Knowledge Processing,
DOI 10.1007/978-1-4471-6735-8

O

P