# CHAPTER 9
# A REPRODUCING KERNEL HILBERT SPACE FRAMEWORK FOR INFORMATION-THEORETIC LEARNING

## 9.1. Introduction

During the last decade, research on Mercer kernel-based learning algorithms has flourished [1], [2], [3]. These algorithms include for example the support vector machine (SVM) [4] [5], kernel principal component analysis (KPCA) [6] and kernel Fisher discriminant analysis (KFDA) [7]. The common property of these methods is that they operate linearly, as they are explicitly expressed in terms of inner-products in a transformed data space that is a reproducing kernel Hilbert space (RKHS). Most often they correspond to nonlinear operators in the data space, and they are still relatively easy to compute using the so-called "kernel-trick". The kernel trick is no trick at all, it refers to a property of the RKHS that enables the computation of inner-products in a potentially infinite-dimensional feature space, by a simple kernel evaluation in the input space. As we may expect, this is a computational saving step that is one of the big appeals of RKHS. At first glance one may even think that it defeats the "no free lunch theorem" (get something out of nothing), but the fact of the matter is that the price of RKHS is the need for regularization and in the memory requirements as they are memory intensive methods. Kernel-based methods (sometimes also called Mercer kernel methods) have been applied successfully in several applications, e.g. pattern and object recognition [8], time series prediction [9] and DNA and protein analysis [10], to name just a few.

Kernel-based methods rely on the assumption that the projection to the high dimensional feature space simplifies data handling as suggested by Cover's theorem, who showed that the probability of shattering data (i.e. separating it exactly by an

In "*Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*", Jose C. Principe, Springer, 2010. ISBN: 978-1-4419-1569-6

hyperplane) approaches one with a linear increase in space dimension [11]. In the case of the SVM, the assumption is that the data classes become linearly separable, and therefore a separating hyperplane is sufficient for perfect classification. In practice, one can not know for sure if this assumption holds. In fact, one has to hope that the user chooses a kernel (and its free parameter) that shatter the data, and since this is improbable, the need to include the slack variable arises. The innovation of SVMs is exactly on how to train the classifiers with the principle of structural risk minimization [5].

ITL emerged independently of the research on Mercer kernel-based learning algorithms. In information theoretic learning, the starting point is a data set that globally conveys information about a real-world event. The goal is to capture the information in the parameters of a learning machine, using some information theoretic performance criterion. As we have seen, information theoretic criteria are expressed as integrals over functions of probability densities. As a simplifying factor, ITL estimates the $\alpha$-norm of the PDF directly from data, without an explicit PDF estimation. Moreover, information theoretic methods have the advantage over Mercer kernel-based methods that they are easier to interpret.

In this chapter, we will define bottom-up an RKHS for information theoretic learning, named ITL RKHS, defined on the Hilbert space of square integrable PDFs. Then we will provide a geometric perspective of all the ITL quantities presented in the previous chapters. Moreover, we show equivalences between Renyi's quadratic estimators of the statistical quantities and the Mercer kernel methods, which until now have been treated separately. Specifically, we show that Parzen window based estimators for Renyi's quadratic information measures have a dual interpretation as

Mercer kernel-based measures, where they are expressed as functions of mean values in the Mercer kernel feature space. The Mercer kernel plays the same role as the Parzen window of density estimation and they are shown to be equivalent. This means that if the Parzen window size can be reliably determined, then the corresponding Mercer kernel size is simultaneously determined by the same procedure.

Furthermore, we develop a classification rule based on the Euclidean distance between PDFs, and show that this corresponds to a linear classifier in the feature space. By regarding this classifier as a special case of the support vector machine, we provide an information theoretic interpretation of the SVM optimization criterion. This chapter is organized as follows. We start with the definition of the ITL RKHS and show the relation between the RKHS used in ITL and kernel methods. Then an ITL perspective of kernel learning and distances is explained, and a new information theoretic classification rule is derived. Thereafter, we analyze the connection between this classifier and the SVM and other kernel methods.

Finally, a connection between the ITL RKHS and kernel approaches interested in quantifying the statistics of the projected data is established. The ITL RKHS structure offers an elegant and insightful geometric perspective towards information-theoretic learning and to the evaluation of statistics in kernel space.

## 9.2. A RKHS Framework for ITL

This section proposes a reproducing kernel Hilbert space (RKHS) framework for information-theoretic learning (ITL), not based on estimators but directly involving the PDFs. The issue of estimation from data samples is treated in Section 9.4.

The ITL RKHS is uniquely determined by the symmetric non-negative

definite kernel function defined as *the cross information potential* (*CIP*) in ITL. The cross information potential between two PDFs $p(x)$ and $q(x)$, defined in Chapter 2 as $V(p,q) = \int p(x)q(x)dx$, characterizes similarity between two stochastic functions. The information potential used so extensively in Chapters 3, 4, 5 and 6 as an entropic cost function (since it is the argument of the log of Renyi's quadratic entropy) is a special case obtained when $p(x)=q(x)$, i.e. a measure of self-similarity. CIP also appears both in the Euclidean and Cauchy-Schwarz divergence measures as well as in the $QMI_{CS}$ and $QMI_{ED}$ used in Chapters 6, 7 and 8 as a measure of dissimilarity in supervised and unsupervised learning.

We prove the existence of a one-to-one congruence mapping between the ITL RKHS and the inner product space spanned by square integrable probability density functions. All the descriptors and cost functions in the original information-theoretic learning formulation can be re-written as algebraic computations on *deterministic functionals* in the ITL RKHS. We first focus on one-dimensional PDFs, and then consider the extension to multi-dimensions in Section 9.2.4. We form a $L_2$ space spanned by all one-dimensional PDFs, and define an inner product in $L_2$. Since the inner product is symmetric non-negative definite, it uniquely determines the reproducing kernel Hilbert space for ITL which will be denoted as $H_V$. We then prove that the inner product itself is indeed a reproducing kernel in $H_V$.

The $L_2$ space of PDFs

Let $E$ be the set that consists of all square integrable one-dimensional probability density functions over the real numbers, i.e., $f_i(x) \in E$, $\forall i \in I$, where

$\int f_i(x)^2 dx < \infty$ and $I$ is an index set. We then form a linear manifold $\left\{ \sum_{i \in K} \alpha_i f_i(x) \right\}$ for any $K \subset I$ and $\alpha_i \in R$. Close the set topologically according to the convergence in the mean using the norm

$$\left\| f_i(x) - f_j(x) \right\| = \sqrt{\int (f_i(x) - f_j(x))^2 dx} \qquad \forall i, j \in I \tag{9.1}$$

and denote the set of all linear combinations of PDFs and its limit points by $L_2(E)$. $L_2(E)$ is an $L_2$ space on PDFs. Moreover, by the theory of quadratically integrable functions, we know that the linear space $L_2(E)$ forms a Hilbert space if an inner product is imposed accordingly. Given any two PDFs $f_i(x)$ and $f_j(x)$ in $E$, we can define an inner product as

$$\langle f_i(x), f_j(x) \rangle_{L_2} = \int f_i(x) f_j(x) dx \qquad \forall i, j \in I \tag{9.2}$$

Notice that this inner product is exactly the cross information potential defined in Section 2.7.3. This definition of inner product has Eq. (9.1) as the corresponding norm. Hence, $L_2(E)$ equipped with the inner product Eq. (9.2) is a Hilbert space. However, it is not a reproducing kernel Hilbert space because the inner product is not reproducing in $L_2(E)$, i.e., the evaluation of any element in $L_2(E)$ cannot be reproduced via the inner product between two functionals in $L_2(E)$. Next we show that the inner product of Eq. (9.2) is symmetric non-negative definite, and by the Moore-Aronszajn theorem it uniquely determines the RKHS $H_V$.

### RKHS $H_V$ Based on $L_2(E)$

First, we define a bivariate function on the set $E$ as

$$V(f_i, f_j) = \int f_i(x) f_j(x) dx \qquad \forall i, j \in I \tag{9.3}$$

Even though $\boldsymbol{V}$ is defined on $E$ its computation makes use of $L_2(E)$. However, by construction of $L_2(E)$ as the span of $E$, any inner product defined on $L_2(E)$ can be written as an inner product of elements of $E$. In reproducing kernel Hilbert space theory, the kernel function is a measure of similarity between functionals. Notice that Eq. (9.3) corresponds to the definition of the inner product in Eq. (9.2) and the cross information potential between two PDFs, hence it is natural and meaningful to define the kernel function as $V(f_i, f_j)$. Next, we show that Eq. (9.3) is symmetric non-negative definite in $E$.

**Property 9.1 (Non-Negative Definiteness)**: The function in Eq. (9.3) is symmetric non-negative definite in $E \times E \to R$.

**Proof:** The symmetry is obvious. Given any positive integer $N$, any set of $\{f_1(x), f_2(x), \ldots f_N(x)\} \in E$ and any not all zero real numbers $\{\alpha_1, \alpha_2, \ldots, \alpha_N\}$, by definition we have

$$\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j V(f_i, f_j) = \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j \int f_i(x) f_j(x)dx$$

$$= \int \left(\sum_{i=1}^{N}\alpha_i f_i(x)\right)\left(\sum_{j=1}^{N}\alpha_j f_j(x)\right)dx = \int \left(\sum_{i=1}^{N}\alpha_i f_i(x)\right)^2 dx \geq 0. \qquad (9.4)$$

Hence, $V(f_i, f_j)$ is symmetric non-negative definite, and it is also a kernel function. According to the Moore-Aronszajn theorem [12], there is a unique reproducing kernel Hilbert space, denoted by $H_V$, associated with the symmetric non-negative definite function in Eq. (9.3). We construct the RKHS $H_V$ bottom-up. Since this bivariate function is symmetric and non-negative definite, it also has an eigen-decomposition by Mercer's theorem [13] as

$$V(f_i, f_j) = \sum_{k=1}^{\infty} \lambda_k \psi_k(f_i) \psi_k(f_j) \qquad (9.5)$$

where $\{\psi_k(f_i), k = 1, 2, ..\}$ and $\{\lambda_k, k = 1, 2, ..\}$ are sequences of eigenfunctions and corresponding eigenvalues of the kernel function $V(f_i, f_j)$ respectively. The series above converges absolutely and uniformly on $E \times E$. Then we define a space $H_V$ consisting of all functionals $G(\cdot)$ whose evaluation for any given PDF $f_i(x) \in E$ is defined as

$$G(f_i) = \sum_{k=1}^{\infty} \lambda_k a_k \psi_k(f_i), \qquad (9.6)$$

where the sequence $\{a_k,\ k = 1, 2, \ldots\}$ satisfies the following condition

$$\sum_{k=1}^{\infty} \lambda_k a_k^2 < \infty \qquad (9.7)$$

Furthermore we define an inner product of two functionals in $H_V$ as

$$\langle G, F \rangle_{H_V} = \sum_{k=1}^{\infty} \lambda_k a_k b_k \qquad (9.8)$$

where $G$ and $F$ are of form Eq. (9.6), and $a_k$ and $b_k$ satisfy property Eq. (9.7).

It can be verified that the space $H_V$ equipped with the kernel function Eq. (9.3) is indeed a *reproducing kernel Hilbert space* and the kernel function $V(f_i, \cdot)$ is a *reproducing kernel* because of the following two properties:

1. $V(f_i, f_j)$ as a function of $f_i(x)$ belongs to $H_V$ for any given $f_j(x) \in E$ because we can rewrite $V(f_i, f_j)$ as

$$V(f_i, \cdot)(f_j) = \sum_{k=1}^{\infty} \lambda_k b_k \psi_k(f_j), \qquad b_k = \psi_k(f_i)$$

That is, the constants $\{b_k, k = 1, 2, ..\}$ become the eigenfunctions

$\{\psi_k(f_i), k=1,2,..\}$ in the definition of $G$. Therefore,

$$V(f_i,\cdot)\in H_V, \qquad \forall f_i(x)\in E \tag{9.9}$$

2. Given any $G\in H_V$, the inner product between the reproducing kernel and $G$ yields the function itself by the definition Eq. (9.8)

$$\langle G, V(f_i,\cdot)\rangle_{H_V} = \sum_{k=0}^{\infty}\lambda_k a_k b_k = \sum_{k=0}^{\infty}\lambda_k a_k \psi_k(f_i) = G(f_i)$$

This is so called the *reproducing property*.

Therefore, $H_V$ is a reproducing kernel Hilbert space with the kernel function and inner product defined above. By the reproducing property, we can re-write the kernel function in Eq. (9.5) as

$$V(f_i, f_j) = \langle V(f_i,\cdot), V(f_j,\cdot)\rangle_{H_v} \qquad V(f_i,\cdot): f_i \mapsto \sqrt{\lambda_k}\psi_k(f_i), \quad k=1,2,... \tag{9.10}$$

The reproducing kernel nonlinearly maps the original PDF $f_i(x)$ into the RKHS $H_V$. We emphasize here that the reproducing kernel $V(f_i, f_j)$ is deterministic (due to the expected value in the inner product) and data-dependent in the sense that the mean of the norm of nonlinearly transformed vector in the RKHS $H_V$ is dependent on the PDF of the original random variable because

$$\left\|V(f_i,\cdot)\right\|^2 = \langle V(f_i,\cdot), V(f_i,\cdot)\rangle_{H_v} = \int f_i(x)^2\,dx \tag{9.11}$$

Congruence Map Between $H_V$ and $L_2(E)$

We have presented two Hilbert spaces, the Hilbert space $L_2(E)$ of PDFs and the reproducing kernel Hilbert space $H_V$. Even though their elements are very different, there actually exists a one-to-one congruence mapping $\Psi$ (isometric isomorphism) from RKHS $H_V$ onto $L_2(E)$ such that

$$\Psi(V(f_i,\cdot)) = f_i \qquad (9.13)$$

Notice that the mapping $\Psi$ preserves isometry between $H_V$ and $L_2(E)$ since by definitions of inner product Eq. (9.2) in $L_2(E)$ and Eq. (9.10) in $H_V$

$$\langle V(f_i,\cdot), V(f_j,\cdot)\rangle_{H_v} = \langle f_i(x)), f_j(x)\rangle_{L_2} = \langle \Psi(V(f_i,\cdot)), \Psi(V(f_j,\cdot))\rangle_{L_2}$$

that is, the mapping $\Psi$ maintains the inner products in both $H_V$ and $L_2(E)$. In order to obtain an explicit representation of $\Psi$, we define an orthogonal function sequence $\{\xi_m(x), m = 1,2,..\}$ over the real numbers satisfying

$$\int \xi_k(x)\xi_m(x)dx = \begin{cases} 0, & k \neq m \\ \lambda_k, & k = m \end{cases} \quad \text{and} \quad \sum_{k=1}^{\infty} \psi_k(f_i) \int \xi_k(x)dx = 1, \qquad (9.14)$$

where $\{\lambda_k\}$ and $\{\psi_k(f_i)\}$ are the eigenvalues and eigenfunctions evaluated at $f_i$, associated with the kernel function $V(f_i, f_j)$ by Mercer's theorem Eq. (9.5). We achieve an orthogonal decomposition of the probability density function as

$$f(x) = \sum_{k=1}^{\infty} \psi_k(f)\xi_k(x), \quad \forall f \in E. \qquad (9.15)$$

The integration to unit of $f$ is guaranteed by Eq. (9.14) (right). Note that the congruence map $\Psi$ can be characterized as the unique mapping from $H_V$ into $L_2(E)$ satisfying the condition that for every functional $G$ in $H_V$ and every $j$ in $I$

$$\int \Psi(G) f_j(x)dx = \langle G, V(f_j,\cdot)\rangle_{H_v} = G(f_j) \qquad (9.16)$$

It is obvious that $\Psi$ in Eq. (9.13) fulfills the condition Eq. (9.16). Then the congruence map can be represented explicitly as

$$\Psi(G) = \sum_{k=1}^{\infty} a_k \xi_k(x), \quad \forall G \in H_V, \qquad (9.17)$$

where $a_k$ satisfies condition Eq. (9.7). To prove the representation Eq. (9.17) is a valid and unique map, substituting Eq. (9.15) and Eq. (9.17) into Eq. (9.16), we obtain

$$\int \sum_{k=1}^{\infty} a_k \xi_k(x) \sum_{m=1}^{\infty} \psi_m(f_j) \xi_m(x) dx = \sum_{k=1}^{\infty} \sum_{m=1}^{\infty} a_k \psi_m(f_j) \int \xi_k(x) \xi_m(x) dx$$

$$= \sum_{k=1}^{\infty} \lambda_k a_k \psi_k(f_j) = G(f_j). \qquad (9.18)$$

In summary, we provide an explicit representation for the congruence map $\Psi$ from RKHS $H_V$ into $L_2(E)$. These two spaces are equivalent in this geometrical sense. However it should be emphasized that the constituting elements are very different in nature. When using samples (realizations), the RKHS isometry framework offers a natural link between stochastic and deterministic functional analysis. Hence, it is more appealing to use RKHS $H_V$ for information-theoretic learning as we will show in the next section.

## Extension to Multi-Dimensional PDFs

Extension of $H_V$ to multi-dimensional PDFs is straightforward since the definitions and derivations in the previous section can be easily adapted into multi-dimensional probability density functions. Now let $E_m$ be the set that consists of all square integrable $m$-dimensional probability density functions, i.e., $f_{i,m}(x_1,\ldots,x_m) \in E_m$, $\forall i \in I$ and $m \in N$, where $\int f_{i,m}(x_1,\ldots,x_m)^2 dx_1,\ldots,dx_m < \infty$ and $I$ is the index set. We need to change the definition of kernel function Eq. (9.3) to

$$V(f_{i,m}, f_{j,m}) = \int f_{i,m}(x_1,\ldots,x_m) f_{j,m}(x_1,\ldots,x_m) dx_1 \ldots dx_m \qquad \forall i, j \in I \quad (9.19)$$

Then every definition and derivation might as well be modified accordingly in the previous section. Let $\mathrm{H}_{V(m)}$ denote the reproducing kernel Hilbert space determined by the kernel function for $m$-dimensional PDFs. The proposed RKHS framework is consistent with dimensionality of PDFs.

The CIP based on the multi-dimensional PDFs characterizes the information

among different random variables whose domains might not necessary be the same in the whole space. In particular, the two-dimensional PDF CIP can be used to quantify the divergence or the cross-covariance between two random variables, because the joint PDF can be factorized into a product of two marginal PDFs as a special independent case. This is exactly what the definitions of $QMI_{ED}$ and $QMI_{CS}$ in Chapter 2 are based on. We will use the two-dimensional PDF CIP to re-formulate these two quantities in the following section.

## 9.3. ITL Cost Functions in the RKHS Framework

In this section, we re-examine the ITL cost functions introduced in Chapter 2 in the proposed RKHS framework. First, as the kernel function $V(f_i, f_j)$ in $H_V$ is defined as the cross information potential between two PDFs, immediately we have

$$\int p(x)q(x)dx = \langle V(p,\cdot), V(q,\cdot)\rangle_{H_v} \tag{9.20}$$

That is, the cross information potential is the inner product between two transformed functionals in the RKHS $H_V$. The inner product quantifies similarity between two functionals which is consistent with the definition of cross information potential. The information potential can thus be specified as the inner product of the functional with respect to itself

$$\int p(x)^2 dx = \langle V(p,\cdot), V(p,\cdot)\rangle_{H_v} = \left\|V(p,\cdot)\right\|_{H_v}^2 \tag{9.21}$$

The information potential appears as the norm square of nonlinearly transformed functional in the RKHS $H_V$. Therefore, minimizing error entropy in ITL turns out to be maximization of norm square in the RKHS $H_V$ (due to the minus sign in the Renyi's quadratic entropy definition).

More interestingly, the result in Eq. (9.21) presents a new interpretation of

Renyi's quadratic entropy. Since Renyi's quadratic entropy is the negative of the logarithm of the information potential, we obtain

$$H_2(X) = -\log\left\|V(p,\cdot)\right\|_{H_V}^2 \qquad (9.22)$$

This means that there is an information theoretic interpretation for the log of the mean square of the transformed functional in $H_V$.

Based on the reformulations of cross information potential Eq. (9.18) and information potential Eq. (9.19) in RKHS $H_V$, we are ready to re-write the one-dimensional Euclidean and Cauchy-Schwarz distance measures in terms of operations on functionals in $H_V$. First,

$$D_{ED}(p,q) = \left\|V(p,\cdot) - V(q,\cdot)\right\|_{H_V}^2 \qquad (9.23)$$

that is, the Euclidean distance measure is in fact the norm square of the difference between two corresponding functionals in $H_V$. The Cauchy-Schwarz divergence measure can be presented as

$$D_{CS}(p,q) = -\log\left(\frac{\langle V(p,\cdot), V(q,\cdot)\rangle_{H_V}}{\left\|V(p,\cdot)\right\|_{H_V}\left\|V(q,\cdot)\right\|_{H_V}}\right) = -\log(\cos\theta) \qquad (9.24)$$

where $\theta$ is the angle (in $H_V$) between two functional vectors $V(p,\cdot)$ and $V(q,\cdot)$. Therefore, the argument of the log of the Cauchy-Schwarz divergence measure truly depicts the separation of two functional vectors in the RKHS $H_V$. When two vectors lie in the same direction the angle $\theta = 0^\circ$ and $D_{CS}(p,q) = 0$. If two vectors are perpendicular to each other ($\theta = 90^\circ$), $D_{CS}(p,q) = \infty$. The RKHS $H_V$ supplies rich geometric insights into the original definitions of the two divergence measures. Now we see that the geometric interpretation presented in Chapter 2, Section 2.7.4 is in fact accurate in $H_V$.

To extend the same formulation to the Euclidean and Cauchy-Schwarz quadratic mutual information defined in Section 2.7.3, consider the product of marginal PDFs $f_1(x_1)f_2(x_2)$ as a special subset $A_2$ of the 2-dimensional square integrable PDFs set $E_2$ where the joint PDF can be factorized into product of marignals, i.e., $A_2 \subseteq E_2$ . Then both measures characterize different geometric information between the joint PDF and the factorized marginal PDFs. The Euclidean quadratic mutual information ($QMI_{ED}$) can be expressed as

$$I_{ED}(X_1, X_2) = \left\| V(f_{1,2},\cdot) - V(f_1 f_2,\cdot) \right\|_{H_V}^2 \qquad (9.25)$$

where $V(f_{1,2},\cdot)$ is the functional in $H_{V(2)}$ corresponding to the joint PDF $f_{1,2}(x_1, x_2)$, and $V(f_1 f_2,\cdot)$ is for the product of the marginal PDFs $f_1(x_1)f_2(x_2)$. Similarly, the Cauchy-Schwarz quadratic mutual information ($QMI_{CS}$) can be re-written as

$$I_{CS}(X_1, X_2) = -\log \frac{\langle V(f_{1,2},\cdot), V(f_1 f_2,\cdot)\rangle_{H_V}}{\left\| V(f_{1,2},\cdot)\right\|_{H_V} \left\| V(f_1 f_2,\cdot)\right\|_{H_V}} = -\log(\cos\gamma) \qquad (9.26)$$

The angle $\gamma$, measured between $f_{1,2}$ and $f_1 f_2$ is the separation between two functional vectors in $H_{V(2)}$ . When two random variables are independent ( $f_{1,2}(x_1, x_2) = f_1(x_1)f_2(x_2)$, $\gamma = 0^\circ$ and the divergence measure $I_{CS}(f_1, f_2) = 0$ since two sets are equal. If $\gamma = 90^\circ$, two vectors in $H_{V(2)}$ are orthogonal and the joint PDF is singular to the product of marginals. In this case, the divergence measure is infinity.

The proposed RKHS framework provides an elegant and insightful geometric perspective towards information-theoretic learning. All the ITL descriptors can now be re-expressed in terms of algebraic operations on functionals in RKHS $H_V$ .

We can also provide a more mathematical understanding for the ITL operators

and their properties. Let us start with the Cross Information Potential. From a statistical point of view, this quantity is a composite moment (expected value over $p(x)$ of another function $q(x)$) of the r.v. $x$. In Chapter 7 we showed that it is a pseudo-distance that was useful because it was easier to estimate than KL, Renyi's, MI and other distance measures in probability spaces. Now we clearly see that it is the natural metric in $H_v$, because it defines the inner product in ITL RKHS. In the same RKHS we can define obviously other distances, such as $D_{ED}$, $D_{CS}$ and $QMI_{ED}$ and $QMI_{CS}$ that are all dependent on the CIP. But now this picture becomes quite clear.

Another example is the evaluation of the statistical properties of IP. For instance, in Chapter 2 we say that the IP estimator contains higher order statistical information of the input data. This has been recognized in ITL by applying the Taylor expansion to the Gaussian kernel used in the estimate of the information potential definition,

$$\hat{V}(X) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{1}{\sqrt{2\pi\sigma^2}}\sum_{k=0}^{\infty}\frac{1}{k!}\left[-\frac{(x_i - x_j)^2}{2\sigma^2}\right]^k \qquad (9.27)$$

But notice that this result depends on the kernel utilized (the Gaussian kernel only provides sums of even order moments, a polynomial kernel will create a finite sum of moments, etc). From Eq. (9.21) it is clear that the norm maximization in ITL RKHS will include the PDF information, therefore now we have a clean statement that derives from the use of the first order moment of the PDF in $H_V$.

## 9.4. ITL Estimators in RKHS

In this section, we will reinterpret the *estimators* of the information theoretic quantities of information potential and Euclidean and Cauchy Schwarz distance in kernel spaces.

Let $H$ be a Hilbert space of real-valued functions defined on the real numbers $R$, equipped with an inner product $\langle\cdot,\cdot\rangle$ and a real-valued non-negative definite function $\kappa(x,y)$ on $R$x$R$. According to the Moore Aronszajn theorem, $\kappa(x,y)$ is a reproducing kernel, which means that each point in the input space $R$ is mapped to a function in the RKHS defined by the selected kernel $\kappa(\cdot,x)\in H_k$ (since the kernel is so important in the definition we use it as a subscript for $H$). We can define the nonlinear mapping between $R$ and $H_{\mathrm{k}}$ as $\Phi(x)=\kappa(\cdot,x)$, and obtain the reproducing property

$$\left\langle \Phi(x),\Phi(y)\right\rangle_H = \left\langle \kappa(.,x),\kappa(.,y)\right\rangle = \kappa(x,y) \qquad (9.28)$$

Therefore, $\Phi(x)=\kappa(.,x)$ defines the Hilbert space associated with the kernel. For our purposes we will be using here the Gaussian kernel $\kappa(x,y)=G_\sigma(x-y)$, which is a non negative definite function, but many others can also be used. A Gaussian kernel corresponds to an infinite-dimensional Mercer kernel feature space, since the Gaussian has an infinite number of eigenfunctions.

This is very different from the reproducing kernel $V(f_i,f_j)$ which has a norm dependent upon the PDF of the data as shown in Eq. (9.11). The norm of nonlinearly projected vector in the RKHS $H_\kappa$ does not rely on the statistical information of the original data since

$$\left\|\Phi(x)\right\|^2 = \langle\Phi(x),\Phi(x)\rangle_{H_k} = \kappa(0) \qquad (9.29)$$

if we use translation-invariant kernel functions. Moreover, if $x$ is a random variable, $\Phi(x)$ is a function operating on random variable in the RKHS $H_\kappa$. The value of $\kappa(0)$ is a constant regardless of the original data. Consequently, the reproducing kernel Hilbert spaces $H_V$ and $H_\kappa$ determined by $V(f_i,f_j)$ and $\kappa(x,y)$ respectively

are very different in nature, however, there are very interesting links among them as we will see below.

Estimator of the Information Potential

Recall the definition of the information potential *estimator* in Chapter 2, which is presented below for convenience.

$$\hat{V}(X) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G_{\sigma\sqrt{2}}(x_i - x_j)$$

Recall from Chapter 2 that this expression is obtained by integrating a product of two Gaussian functions centered at each sample over the domain which can be written as an inner product as follows

$$G_{\sigma\sqrt{2}}(x_i - x_j) = \kappa(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle_{H_k} \qquad (9.30)$$

Hence, the Parzen window-based estimator for the information potential can be expressed in terms of an inner product in the Mercer kernel space. We can further operate to obtain

$$\begin{aligned} \hat{V}(X) &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G_{\sigma\sqrt{2}}(x_i - x_j) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} \langle \Phi(x_i), \Phi(x_j) \rangle_{H_k} = \\ &= \langle \frac{1}{N}\sum_{j=1}^{N} \Phi(x_i), \frac{1}{N}\sum_{j=1}^{N} \Phi(x_i) \rangle_{H_k} = \mathbf{m}^T\mathbf{m} = \|\mathbf{m}\|^2 \end{aligned} \qquad (9.31)$$

where $N$ samples are assumed available from the data. That is, it turns out that the information potential may be expressed as the squared norm of the mean vector of the data in a Mercer kernel feature space. This connection was previously pointed out in [14] in a study relating orthogonal series density estimation to kernel principal component analysis. Therefore the magic that may still have existed in why we could estimate an entropy without explicit PDF estimation becomes clear now. The argument of the log is a central moment of the projected data, therefore it is likely to

find good estimators that do not require PDF estimation. As we mentioned in Chapter 2, estimating the mean and variance from data does not require PDf estimation, and the same applies to Renyi's quadratic entropy.

Estimators of Quadratic Distances

In Chapter 2 we defined two distances measures using the information potential $D_{ED}$ and $D_{CS}$ and their corresponding estimators based on kernels. Let us assume that the goal is to estimate the distance between two PDFs $p(x)$ and $q(x)$ from which we have respectively $N_1$ and $N_2$ samples. Again for convenience we copy below the estimator for $D_{ED}$.

$$\hat{D}_{ED}(p,q) = \hat{V}_{ED} = \frac{1}{N_1^2}\sum_{i=1}^{N_1}\sum_{i'=1}^{N_1} G_{\sigma\sqrt{2}}(x_i - x_{i'}) - \frac{2}{N_1 N_2}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2} G_{\sigma\sqrt{2}}(x_i - x_j) + \frac{1}{N_2^2}\sum_{i=1}^{N_2}\sum_{j=1}^{N_2} G_{\sigma\sqrt{2}}(x_j - x_{j'}). \quad (9.32)$$

Note that we have for simplicity assumed that the same kernel size σ is appropriate for both estimators. This may not be the case in practice, but it can be incorporated in the subsequent analysis. In analogy to Eq. (9.32), $D_{ED}$ may also be expressed in terms of mean vectors in the Mercer kernel feature space $H_\kappa$ to obtain

$$\hat{D}_{ED}(p,q) = \|\mathbf{m}_1\|^2 - 2\mathbf{m}_1^T\mathbf{m}_2 + \|\mathbf{m}_2\|^2 = \|\mathbf{m}_1 - \mathbf{m}_2\|^2 \qquad (9.33)$$

where $\mathbf{m}_1$ is the kernel feature space mean vector of the data points drawn from $p(x)$, and $\mathbf{m}_2$ is the kernel feature space mean vector of the data points drawn from $q(x)$. Hence, $D_{ED}$ can also be seen to have a geometric interpretation in $H_\kappa$. It measures the square of the norm of the difference vector between the two means $\mathbf{m}_1$ and $\mathbf{m}_2$. In a similar fashion we can obtain the estimator for the Cauchy Schwarz divergence as

$$\hat{D}_{CS}(p,q) = -\log\left(\frac{\mathbf{m}_1^T\mathbf{m}_2}{\|\mathbf{m}_1\|\|\mathbf{m}_2\|}\right) = -\log\left(\cos\angle(\mathbf{m}_1,\mathbf{m}_2)\right) \qquad (9.34)$$

Remember that the information cut explained in Chapter 6 was defined as the argument of the log of $D_{CS}$, therefore it has a dual interpretation as a measure of the

cosine of the angle between cluster mean vectors in the Mercer kernel feature space $H_k$. This metric is very natural in kernel machines because the nonlinear transformation induced by a symmetric reproducing kernel maps the input samples over a sphere in the feature space, since for any $x$, $\|\Phi(x)\|^2=\kappa(0)= 1/(\sqrt{2\pi\sigma})$. Therefore, the distance between $\Phi(x_i)$ and $\Phi(x_j)$ on that sphere (i.e., the geodesic distance) is proportional to the angle between the vectors from the origin to those points

$$d(\Phi(x_i),\Phi(x_j)) \propto \cos^{-1}\left(\frac{\langle\Phi(x_i),\Phi(x_j)\rangle}{\|\Phi(x_i)\|\|\Phi(x_j)\|}\right)=\cos^{-1}(\sqrt{2\pi}\sigma\kappa(x_i-x_j))$$

In other words, the kernel function is in fact computing the cosine of the angle between two points over the sphere (i.e., a distance). In addition, we notice from the previous discussion that the transformed data $\Phi(x_i)$ must lie on some embedded manifold over the positive hyperoctant on the sphere (since the kernel takes only positive values, the angle belongs to the interval $[0, \pi/2]$).

The relationship between the two RKHS $H_v$ and $H_k$ can then be readily perceived thru the ITL descriptors of IP, $D_{ED}$ and $D_{CS}$ and their estimators. In fact, by comparing Eq. (9.22) with (9.31) (and Eq. (9.23) with (9.33) as well as Eq. (9.24) with (9.34)) we see that these statistical quantities in $H_V$ can be *estimated* by the mean operator of the projected functionals in $H_\kappa$, which effectively were derived with the Parzen's non-parametric asymptotically unbiased and consistent PDF estimator employed in the ITL cost functions. Provided one chooses a non-negative definite kernel function as the Parzen window, the statistical quantities in RKHS $H_V$ are related to the estimators in $H_\kappa$ as illustrated in Fig. 9-1.

In more general terms, the reproducing kernel Hilbert spaces $H_\kappa$ and $H_V$ are formally related via the expectation operator, which means that the feature map $V(f_i,\cdot)$ is a transformation of the PDF space into a *deterministic* RKHS $H_V$, while the feature map $\Phi(x)$ nonlinearly projects the sample space into a stochastic RKHS $H_\kappa$. The feature map $V(f_i,\cdot)$ is a descriptor of the stochasticity of the sample space, and immediate algebraic operation can be applied to compute statistics in $H_V$. This means that IP, CIP and all the distances are deterministic scalars that quantify PDF properties. Hence, the proposed ITL RKHS framework provides a function analysis view of statistics, and it seems the natural RKHS to perform statistical inference. Of course, the issue is that if one does not have an analytical description of the data PDF one can not progress further. This is where $H_\kappa$ becomes useful since one can build there the estimators for the above mentioned statistical quantities. But this clearly shows that statistical estimators in $H_\kappa$ operate with the full PDF information.

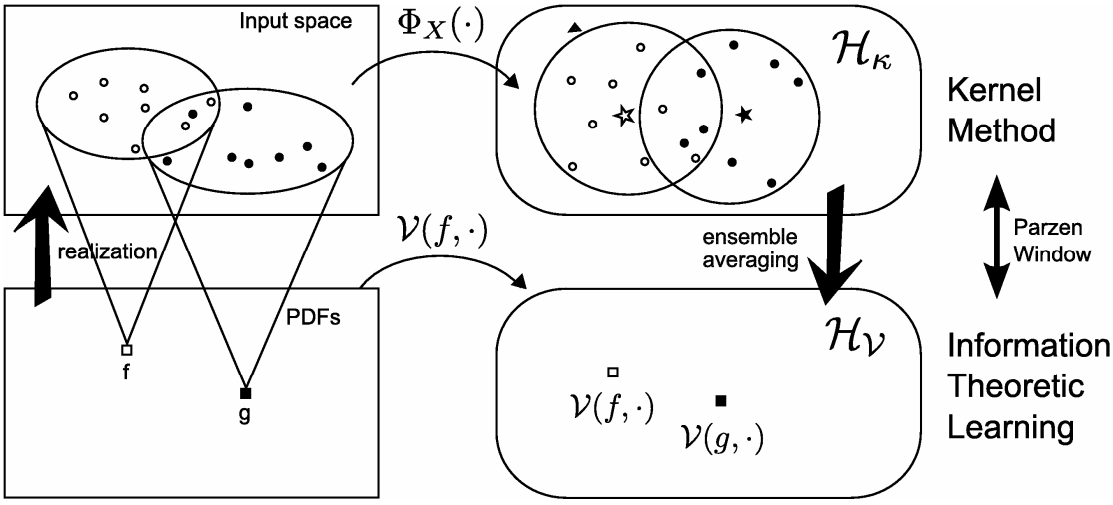


Figure 9-1. The relationship between $H_V$ and $H_\kappa$.

Parzen was the first to introduce the RKHS methodology in statistical signal processing and time series analysis in the late 1950s. The essential idea is that there

exists a congruence map between the RKHS of random variables spanned by the random process and its covariance function $R(t, s) = E[X(t)\ X(s)]$ which determines a unique RKHS, denoted as $H_R$. Note that the kernel includes the second order statistics of the data through the expected value (a data dependent kernel as $V(f_i, f_j)$) and Parzen clearly illustrated that the RKHS offers an elegant functional analysis framework for minimum variance unbiased estimation of regression coefficients, least-squares estimation of random variables, detection of signals in Gaussian noise, and others [15], [16], [17]. In the early 1970s, Kailath and coworkers presented a series of detailed papers on the RKHS approach to detection and estimation problems to demonstrate its superiority in computing likelihood ratios, testing for nonsingularity, bounding signal detectability, and determining detection stability [18], [19], [20], [21], [22]. RKHS concepts have also been extensively applied to a wide variety of problems in optimal approximation including interpolation and smoothing by spline functions in one or more dimensions (curve and surface fitting) [23]. De Figueiredo took a different approach to apply RKHS in nonlinear system and signal analysis [24]. He built the RKHS bottom-up using arbitrarily weighted Fock spaces that played an important role in quantum mechanics [25]. The spaces are composed of polynomials or power series in either scalar or multi-dimensional variables. The generalized Fock spaces have been also applied to nonlinear system approximation, semiconductor device characteristics modeling and neural networks [25].

The conventional mean square error has also been re-written as norm square of projected vectors in the RKHS $H_R$ induced by the covariance function [15]. But $H_R$ only takes the second-order statistics into account, while the RKHS $H_V$ is defined over PDFs and the RKHS $H_\kappa$, depending upon the kernel utilized, will also implicitly

embed all the statistical information of the data. Moreover, notice that $H_V$ although linearly related to the PDF space, it is *nonlinearly related to the data space*, unlike $H_R$ which is linearly related to the data space. This implies that, with appropriate kernels, $H_\kappa$ can be used to estimate second and higher order statistical of the data. In general, mean and covariance operators are necessary to perform statistics in $H_\kappa$ as we will discuss next, but our work shows that the inclusion of the expected value operator in the kernel itself simplifies the analysis when the goal is statistical inference.

## 9.5. Connection Between ITL and Kernel Methods via RKHS $H_V$

In this section, we connect ITL and kernel methods via the proposed RKHS framework. As we have mentioned in the previous section, because the RKHS $H_\kappa$ is induced by the data-independent kernel function, the nonlinearly projected data in $H_\kappa$ is still stochastic and statistical inference is required in order to compute quantities of interest. For instance, in order to compute the statistics over the functionals, the expectation and covariance operators are required. The expected value of functionals in the RKHS $H_\kappa$ is defined as $E[\Phi(x)]$. The cross-covariance is defined as a unique operator $\Sigma_{\mathbf{XY}}$ such that for any functionals $f$ and $g$ in $H_\kappa$

$$\langle g, \sum_{XY} f \rangle_{H_k} = E[g(y) f(x)] - E[g(y)]E[f(x)] = Cov[f(x), g(y)] \quad (9.35)$$

The mean and cross-covariance operators as statistics of functionals in $H_\kappa$ become intermediate steps to compute other quantities such as the maximum mean discrepancy (MMD) [26], kernel independent component analysis (Kernel ICA) [27] and others. But the interesting question is to find out the relationships with both the

ITL RKHS $H_V$ and the ITL estimators of Eq. (9.31), (9.33) and (9.34). We show here that MMD is equivalent to the Euclidean divergence measure, and that Kernel ICA is equivalent to the Cauchy-Schwarz quadratic mutual information. The statistical computations in the RKHS $H_\kappa$ have corresponding algebraic expressions in the RKHS $H_V$.

An ITL Perspective of Maximum Mean Discrepancy

The maximum mean discrepancy (MMD) [26] is a statistical test based on kernel methods to determine whether two samples are from different distributions. Since first order moment of the PDF describes Renyi's quadratic entropy, theoretically, if the expected value of a PDF $p(x)$ for an arbitrary measurable function is the same for both random variables, the two distributions are identical. Since it is not practical to work with such a rich function class, MMD restricts the function class to a unit ball in a reproducing kernel Hilbert space $H_\kappa$ that is associated with the kernel $\kappa(\cdot,\cdot)$. This leads to the following quantity,

$$M(X,Y) = \sup_{\|p\|_{H_k} \le 1} (E[p(x)] - E[p(y)]) \qquad (9.36)$$

where X and Y are the underling random variables of the two distributions, $p$ is a family of measurable functionals in the unit ball of the RKHS $H_\kappa$. The kernel trick can be employed here to compute MMD, that is,

$$p(x) = \langle \Phi(x), p \rangle_{H_k} = \langle \kappa(x,\cdot), p \rangle_{H_k} \qquad (9.37)$$

Substituting Eq. (9.37) into the definition of MMD Eq. (9.36), we obtain

$$M(X,Y) = \left\| m_X - m_Y \right\|_{H_\kappa} \qquad (9.38)$$

where $m_{\mathbf{X}} = E[\Phi(x)]$ and $m_{\mathbf{Y}} = E[\Phi(y)]$ are the statistical expectations of the

functionals $\Phi(x)$ and $\Phi(y)$ in the RKHS $H_\kappa$. Applying $m_{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N}\Phi(x_i)$, an empirical estimate of MMD can be obtained as

$$\left|\hat{M}(X,Y)\right|^2 = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\kappa(x_i,x_j) - \frac{2}{NL}\sum_{i=1}^{N}\sum_{j=1}^{L}\kappa(x_i,y_j) + \frac{1}{L^2}\sum_{i=1}^{L}\sum_{j=1}^{l}\kappa(y_i,y_j) \quad (9.39)$$

where $\{x_i\}_{i=1}^{N}$ and $\{y_j\}_{j=1}^{L}$ are two sets of data samples. The estimate of MMD provides a statistical test to determine whether two sets of data samples are from the same distribution. Comparing with Eq. (9.35) it is easy to show that MMD is equivalent to the Euclidean divergence measure. i.e. $[M(\mathbf{X},\mathbf{Y})]^2 = D_{ED}(f,g)$. Moreover, since the Euclidean divergence measure can be re-written as the norm square difference between two functionals in the RKHS $H_V$, we obtain

$$\left\|m_X - m_Y\right\|_{H_\kappa}^2 = \left\|V(f,\cdot) - V(g,\cdot)\right\|_{H_\kappa}^2 \quad (9.40)$$

The left hand side is the norm square of difference between two functional expectations in the RKHS $H_\kappa$. Since the functional $\Phi(x)$ is still stochastic in $H_\kappa$, the expectation operation is necessary to carry out the computation. On the other hand, the right hand size is the norm square of difference between two functionals in the RKHS $H_V$. Because the functional $V(f,\cdot)$ is deterministic, the computation is algebraic. The feature map $V(f,\cdot)$ for the RKHS $H_V$ is equivalent to the expectation of the feature map $\Phi(x)$ for the RKHS $H_\kappa$ (see Figure 9-1). Therefore, the proposed RKHS framework provides a natural link between stochastic and deterministic functional analysis. The MMD in kernel methods is essentially equivalent to the Euclidean divergence measure in Information-Theoretic Learning.

## An ITL Perspective of Kernel ICA

Kernel ICA is a novel independent component analysis method based on a kernel measure of independence [27]. It assumes an RKHS $H_\kappa$ determined by the kernel $\kappa(x, y)$ and feature map $\Phi(x)$. The feature map $\Phi(x)$ can be derived from the eigen-decomposition of the kernel function $\kappa(x, y)$ according to Mercer's theorem, and forms an orthogonal basis for the RKHS $H_\kappa$. Then the $F$-correlation function is defined as the maximal correlation between the two random variables $f_1(x_1)$ and $f_2(x_2)$, where $f_1$ and $f_2$ range over $H_\kappa$:

$$\rho = \max_{f_1, f_2} corr(f_1(x_1), f_2(x_2)) = \max_{f_1, f_2} \frac{corr(f_1(x_1), f_2(x_2))}{\sqrt{\text{var}(f_1(x_1))\,\text{var}(f_2(x_2))}} \qquad (9.41)$$

Obviously, if the random variables $x_1$ and $x_2$ are independent, then the $F$-correlation is zero. And the converse is also true provided that the RKHS $H_\kappa$ is large enough. This means that $\rho = 0$ implies $x_1$ and $x_2$ are independent. In order to obtain a computationally tractable implementation of $F$-correlation, the reproducing property of RKHS is used to estimate the $F$-correlation. The nonlinear functionals $f_1$ and $f_2$ can be represented by the linear combination of the basis $\{\Phi(x^i)\}_{i=1}^N$ in which $\{x^i\}_{i=1}^N$ is an empirical observation of the random variable $x$ with $N$ samples. That is,

$$f_1 = \sum_{k=1}^{N} \alpha_1^k \Phi(x_1^k), \quad f_2 = \sum_{k=1}^{N} \alpha_2^k \Phi(x_2^k) \qquad (9.42)$$

Substituting Eq. (9.42) and Eq. (9.37) into Eq. (9.41) and using the empirical data to approximate the population value, the $F$-correlation can be estimated as

$$\hat{\rho} = \max_{\alpha_1,\alpha_2} \frac{\alpha_1^T K_1 K_2 \alpha_2}{\sqrt{\left(\alpha_1^T K_1^2 \alpha_1\right)\left(\alpha_2^T K_2^2 \alpha_2\right)}} \qquad (9.43)$$

where $K_1$ and $K_2$ are the Gram matrices associated with the data sets $\{x_1^i\}_{i=1}^N$ and $\{x_2^i\}_{i=1}^N$ defined as $[K_i]_{a,b} = \kappa(x_i^a, x_i^b)$.

Because the cost function in Eq. (9.43) is not a numerically stable estimator in general, a regularization is needed by penalizing the RKHS norms of $f_1$ and $f_2$ in the denominator of Eq. (9.43). The regularized estimator has the same independence characterization property of the $F$-correlation, since it is the numerator, $\alpha_1^T K_1 K_2 \alpha_2$, that characterizes the independence property of two random variables.

We prove here the equivalence between the cost function used in Kernel ICA Eq. (9.43) and the Cauchy-Schwarz quadratic mutual information Eq. (9.34). To prove the equivalence, we use the weighted Parzen window which is defined as

$$\hat{f}(x) = \frac{1}{A}\sum_{i=1}^{N} \alpha_i \kappa(x, x_i) \qquad (9.44)$$

where $A$ is a normalization term such that the integral of $\hat{f}(x)$ equals to 1.

When the Cauchy-Schwarz quadratic mutual information is used as a contrast function in ICA, it should be minimized so that the mutual information between random variables is also minimized. As the logarithm is a monotonic function, minimizing the Cauchy-Schwarz quadratic mutual information is equivalent to maximizing its argument. Therefore, by approximating the population expectation with sample mean for the argument in Eq. (9.34) and estimating the joint and marginal PDFs with weighed Parzen window Eq. (9.44), we obtain

$$\hat{J} = \max_{\alpha_1,\alpha_2} \frac{\alpha_1^T K_1 K_2 \alpha_2}{\sqrt{L\left(1^T K_1^2 \alpha_1\right)\left(1^T K_2^2 \alpha_2\right)}} \qquad (9.45)$$

where $1=[1,1,..,1]^\top$, $[K_i]_{a,b}=\kappa(x_i^a,x_i^b)$, and $L=\sum_{i=1}^{N}\ \sum_{j=1}^{N}\alpha_1^i\kappa(x_1^i,x_1^j)\kappa(x_2^i,x_2^j)\alpha_2^i$.

Comparing the two expressions Eq. (9.43) and (9.45), we notice that they have the same numerators but different normalizations. As we already pointed out, it is the numerators in the Kernel ICA and the Cauchy-Schwarz quadratic mutual information that characterize the dependence measure of two random variables. The denominators only provide normalization. Hence we conclude that the Cauchy-Schwarz quadratic mutual information, estimated via weighed Parzen window, is equivalent to Kernel ICA. Moreover, the coordinates of the nonlinear functionals $f_1$ and $f_2$ in the RKHS $H_\kappa$ Eq. (9.42) have corresponding terms in the weighted Parzen window Eq. (9.44).

In summary, the feature map $\Phi(x)$ works with individual data samples and transforms each data into the RKHS $H_\kappa$ induced by the kernel $\kappa(\cdot,\cdot)$. For applications involving statistical inference on the transformed data, extra operators such as the mean and covariance are required. On the other hand, the feature map $V(f,\cdot)$ deals with PDF directly and transforms each PDF into the RKHS $H_V$ determined by the kernel $V(\cdot,\cdot)$. If the applications are based on the statistics of the transformed functionals, only algebraic computation is needed without defining any extra operators as required in RKHS $H_\kappa$. Therefore the proposed RKHS framework provides a direct and elegant treatment of statistical inference using RKHS technique. Certainly, the RKHS $H_\kappa$ is more flexible in other applications beyond statistical inference since it is based on the available data samples. The RKHS $H_V$ is built directly upon PDFs, and requires Parzen windows to carry out the evaluation of the overall cost functions as we saw in ITL.

## 9.6. An ITL Perspective of MAP and SVM Classifiers

From the previous sections, we obtained a very clear view of the statistical power of mappings onto $H_\kappa$. The square of the projected data vector mean is in fact an estimator of the 2-norm of the PDF of the data as we saw in Chapter 2 when we interpreted quadratic Renyi's entropy. Therefore, a classification rule similar to LDA can be easily implemented with the projected class means without making any Gaussian assumption and benefiting from the high dimensionality of the feature space that provides good performance even for linear classifiers.

9.6.1. Euclidean Distances in $H_\kappa$ and MAP Classifiers

The classification rule is based on $D_{ED}$, which we will analyze theoretically both in the input space and in the Mercer kernel space $H_\kappa$. An interesting property of this new classifier is that it contains the MAP classifier as a special case. We have available the training data points $\{x_i\}$; $i$ =1,…,$N_1$, drawn from $p(x)$, and a corresponding sample from $q(x)$, that is, $\{x_j\}$; $j$ = 1, …, $N_2$. The label information is used to create these two classes. Based on this training data set we wish to construct a classifier, which assigns a test data point $x_0$ to one of the classes $c_1$ or $c_2$. Now, we define

$$\hat{p}'(x) = \frac{1}{N_1+1}\sum_{i=0}^{N_1}\kappa(x,x_i), \qquad \hat{q}'(x) = \frac{1}{N_2+1}\sum_{j=0}^{N_2}\kappa(x,x_j) \qquad (9.46)$$

Hence, $\hat{p}'(x)$ is the Parzen estimator for $p(x)$, assuming $x_0$ is included in the $c_1$ data class. Likewise, $\hat{q}'(x)$ is the Parzen estimator for $q(x)$, assuming $x_0$ is included in the $c_2$ data set. The proposed $D_{ED}$ based strategy is to classify $x_0$ according to the following rule:

$$x_0 \in c_1 : \int \left(\hat{p}'(x) - \hat{q}(x)\right)^2 dx \geq \int \left(\hat{p}(x) - \hat{q}'(x)\right)^2 dx \qquad (9.47)$$

otherwise, assign $x_0$ to $c_2$. In words, the rule assign $x_0$ to the class which, when having

$x_0$ appended to it, makes the estimated distance between the classes the greatest. We will now analyze this simple classification rule in terms of the Mercer kernel feature space $H_\kappa$. Let $m_i', i = 1,2$ be the Mercer kernel feature space mean vector of class $c_i$, assuming $\Phi(x_0)$ is assigned to that class. It is easily shown that

$$\begin{aligned} m_1' &= \frac{N_1}{N_1+1} m_1 + \frac{1}{N_1+1}\Phi(x_0) \\ m_2' &= \frac{N_2}{N_2+1} m_2 + \frac{1}{N_2+1}\Phi(x_0) \end{aligned} \qquad (9.48)$$

In the kernel feature space, the equivalent classification rule of Eq. (9.47) may be expressed as

$$x_0 \in c_1 : \|m_1' - m_2\|^2 \geq \|m_1 - m_2'\|^2 \qquad (9.49)$$

Assume that $P(c_1) = P(c_2)$, that is the prior probabilities for the classes are equal. Let $P(c_1) = N_1/N$ and $P(c_2) = N_2/N$ , which means that we assume that $N_1 = N_2$. In that case, we have

$$\begin{aligned} m_1' &= \beta_1 m_1 + \beta_2 \Phi(x_0) \\ m_2' &= \beta_1 m_2 + \beta_2 \Phi(x_0) \end{aligned} \qquad (9.50)$$

where $\beta_1 = N_1/(N_1+1) = N_2/(N_2+1)$, and $\beta_2 = 1/(N_1+1) = 1/(N_2+1)$. For ease of notation, let $\Phi(x_0) = y$. The left-hand side of Eq.(9.49), becomes

$$\begin{aligned} \|m_1' - m_2\|^2 &= m_1'^T m_1' - 2m_1'^T m_2 + m_2^T m_2 = \\ &= \beta_1^2 \|m_1\|^2 + 2\beta_1\beta_2 m_1^T y + \beta_2^2 \|y\|^2 - 2\beta_1 m_1^T m_2 - 2\beta_2 m_2^T y + \|m_2\|^2 \end{aligned}$$

Similarly, the right-hand side of Eq.(9.46) becomes

$$\begin{aligned} \|m_1 - m_2'\|^2 &= m_1^T m_1 - 2m_1^T m_2' + m_2'^T m_2 = \\ &= \|m_1\|^2 + 2\beta_1\beta_2 m_2^T y + \beta_2^2 \|y\|^2 - 2\beta_1 m_2^T m_1 - 2\beta_2 m_1^T y + \beta_1^2 \|m_2\|^2 \end{aligned}$$

Using these results, the classification rule becomes

$$
\begin{aligned}
x_0 \in c_1: \quad & \|m_1' - m_2\|^2 \geq \|m_1 - m_2'\|^2 \\
& \Leftrightarrow m_1^T y - m_2^T y - \frac{\beta_1^2 - 1}{2\beta_2(\beta_1 + 1)}(\|m_2\|^2 - \|m_1\|^2) \geq 0 \qquad (9.51) \\
& \Leftrightarrow m_1^T y - m_2^T y + b \geq 0
\end{aligned}
$$

where $b = 1/2(\|m_2\|^2 - \|m_1\|^2)$ , and the constant $\frac{\beta_1^2 - 1}{\beta_2(\beta_1 + 1)} = -1$ . The above classification rule has a simple geometrical interpretation. The point *y* is assigned to the class whose mean it is closest, and the class boundary in kernel feature space is a hyperplane given by a vector *w*. Let $w = m_1 - m_2$, and let the midpoint between $m_1$ and $m_2$ be given by $v = 1/2\ (m_1 + m_2)$. Now the class of *y* is determined by examining whether the vector $(y - v)$ encloses an angle smaller than $\pi/2$ with the vector *w* or not. If it does, *y* is closest to $m_1$, and *y* is assigned to $c_1$.

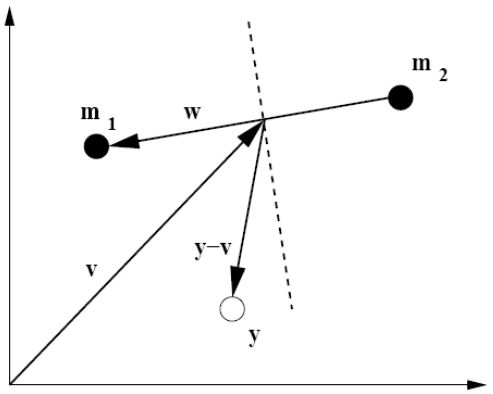


Figure 9-2: ISE-based geometric classification rule: Assign the point y to the class whose mean it is closest to. This can be done by looking at the inner-product between (y − v) and w. It changes sign as the enclosed angle passes through $\pi/2$ . The corresponding decision boundary is given by a hyperplane orthogonal to w (dashed line). From [39].

Hence,

$$
\begin{aligned}
x_0 \in c_1: \quad & w^T(y - v) \geq 0 \\
& w^T y + b \geq 0 \qquad (9.52) \\
& \Leftrightarrow m_1^T y - m_2^T y + b \geq 0
\end{aligned}
$$

Figure 9-2 geometrically illustrates this simple classification rule, which we have derived using the $D_{ED}$ criterion as a starting point. As explained above, in the Mercer kernel space, the value of the inner-product between the class mean values and the new data point determines which class it is assigned to. The threshold value *b*, depends on the squared Euclidean norms of the mean values, which are equivalent to the class information potentials, and hence the class entropies. We now complete the circle, and analyze the Mercer kernel feature space classification rule in terms of Parzen estimators in the input space. Note that

$$m_1^T y = m_1^T \Phi(x_0) = \frac{1}{N_1}\sum_{i=1}^{N_1} \Phi^T(x_i)\Phi(x_0) = \frac{1}{N_1}\sum_{i=1}^{N_1} \kappa(x_0, x_i) = \hat{p}(x_0) \qquad (9.53)$$

Likewise

$$m_2^T y = m_2^T \Phi(x_0) = \frac{1}{N_2}\sum_{j=1}^{N_2} \Phi^T(x_j)\Phi(x_0) = \frac{1}{N_2}\sum_{j=1}^{N_2} \kappa(x_0, x_j) = \hat{q}(x_0) \qquad (9.54)$$

The classification rule hence becomes

$$x_0 \in c_1 : \quad \hat{p}(x_0) - \hat{q}(x_0) + b \geq 0 \qquad (9.55)$$

We remark that this classification rule depends both on the estimated densities at $x_0$, and on the information potentials of the two classes. We have already shown that these information potentials are equivalent to Renyi's quadratic entropies for the classes. In the case that the classes have the same value for the information potential (entropy), which means that the kernel feature space mean values have equal length from the origin, we have $b = 0$, and the current classification rule reduces to the well-known MAP classification rule (for equal priors), where the class probability densities are estimated using Parzen windowing. The same direct connection can not be obtained based on the Cauchy-Schwarz divergence.

The Support Vector Machine

The support vector machine is the most prominent Mercer kernel-based learning algorithm. It is a hyperplane classifier which is based on two crucial properties: 1) the kernel property (kernel trick), which allows for a relatively efficient computation of the SVM algorithm even in infinite dimensional spaces. 2) the maximization of the hyperplane margin, which is a regularizing condition on the hyperplane solution. Basically, it limits the admissible separating hyperplanes to the one maximizing the margin. This regularization has a positive effect on the generalization capability of the classifier [5].

In the following, we give a brief review of the SVM theory. We formulate the problem directly in the Mercer kernel feature space. This Mercer kernel feature space is induced by some kernel function, which hopefully makes the feature space data linearly separable such that it can be separated by a hyperplane. Whether or not the data in fact is linearly separable, heavily depends on the user choosing a proper kernel.

Let $c_1$ and $c_2$ denote two data classes. We are given a training set consisting of $\{x_i\}$; i = 1,….,$N_1$, from $c_1$, and $\{x_j\}$; j = 1,….,$N_2$, from $c_2$. The task is to train a SVM classifier, such that it creates a maximum margin linear classifier in the kernel feature space. After training, the classification rule in feature space is

$$x_0 \in c_1 : \qquad w^{*T}\,\Phi(x_0) + b^* \geq 0 \qquad (9.56)$$

otherwise, $x_0 \in c_2$, where, $x_0$ is a new, previously unseen data point. Presumably, it has either been generated by the process generating the $c_1$ data, or the process generating the $c_2$ data. Regularization by maximizing the margin in feature space corresponds to minimizing the squared norm of the (canonical) separating hyperplane weight vector, that is $\|w^*\|^2$, given the constraints

$$\begin{aligned} w^{*T}\,\Phi(x_i)+b^* &\geq +1, \quad \forall x_i \in c_1 \\ w^{*T}\,\Phi(x_j)+b^* &\leq -1, \quad \forall x_j \in c_2 \end{aligned} \tag{9.57}$$

This is a constrained optimization problem, which is dealt with by introducing Lagrange multipliers $\alpha_i \geq 0, \alpha_j \geq 0$, corresponding to the two classes, and a primal Lagrangian

$$L_P = \frac{1}{2}\|w^*\|^2 - \sum_{i=1}^{N_1}\alpha_i[w^{*T}\,\Phi(x_i)+b^*-1] + \sum_{j=1}^{N_2}\alpha_j[w^{*T}\,\Phi(x_j)+b^*+1] \tag{9.58}$$

The Lagrangian $L_P$ has to be minimized with respect to the primal variables $w^*$ and $b^*$, and maximized with respect to the dual variables $\alpha_i$ and $\alpha_j$. Hence, a saddle point must be found. At the saddle point, the derivatives of $L_P$ with respect to the primal variables must vanish,

$$\frac{\partial L_p}{\partial b^*} = 0, \quad \frac{\partial L_p}{\partial w^*} = 0 \tag{9.59}$$

which leads to

$$\sum_{i=1}^{N_1}\alpha_i = \sum_{j=1}^{N_2}\alpha_j = \Omega, \quad w^* = m_1^* - m_2^* \tag{9.60}$$

where

$$m_1^* = \sum_{i=1}^{N_1}\alpha_i\Phi(x_i), \quad m_2^* = \sum_{j=1}^{N_2}\alpha_j\Phi(x_j) \tag{9.61}$$

By substituting these constraints into Eq. (8.58), the dual Lagrangian

$$L_D = 2\Omega - \frac{1}{2}\left[\sum_{i=1}^{N_1}\sum_{i'=1}^{N_1}\alpha_i\alpha_{i'}\kappa(x_i,x_{i'}) - 2\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\alpha_i\alpha_j\kappa(x_i,x_j) + \sum_{j=1}^{N_2}\sum_{j'=1}^{N_2}\alpha_j\alpha_{j'}\kappa(x_j,x_{j'})\right] \tag{9.62}$$

is obtained, where $\kappa(\cdot,\cdot)$ denotes an inner product between any two training data points in the Mercer kernel feature space $H_\kappa$. $L_D$ must be maximized with respect to the Lagrange multipliers. It can be seen that the solution vector $w^*$ has an expansion in

terms of the training patterns weighted by the Lagrange multipliers. The Karush-Kuhn-Tucker (KKT) conditions

$$\begin{aligned} \alpha_i\left[w^{*T}\,\Phi(x_i)+b^*-1\right]&=0, \quad \forall i=1,\ldots,N_1 \\ \alpha_j\left[w^{*T}\,\Phi(x_j)+b^*+1\right]&=0, \quad \forall j=1,\ldots,N_2 \end{aligned} \tag{9.63}$$

specify the non-zero Lagrange multipliers to be those training patterns which are situated on the margin in feature space. Hence, $w^*$ is a weighted combination of the patterns on the margin. Let us determine the expression for $b^*$ in the SVM theory. For those $b^*$ corresponding to support vectors belonging to $c_1$, we have $b_1^* = 1 - w^{*T}\,\Phi(x_i)$, where $\Phi(x_i)$ is a support vector. By adding all $b_1^*$ values corresponding to $c_1$, we have (remember that only those i's corresponding to support vectors deviate from zero)

$$\begin{aligned} \sum_{i=1}^{N_1}\alpha_i b_1^* &= \sum_{i=1}^{N_1}\alpha_i - w^{*T}\sum_{i=1}^{N_1}\alpha_i\Phi(x_i) \\ \Omega b_1^* &= \Omega - w^{*T}\,m_1^* \\ b_1^* &= 1-\frac{1}{\Omega}\left\|m_1^*\right\|^2+\frac{1}{\Omega}m_1^{*T}\,m_2^* \end{aligned} \tag{9.64}$$

Similarly, for those $b^*$ corresponding to support vectors belonging to $c_2$, we have $b_2^* = -1 - w^{*T}\,\Phi(x_j)$ and we obtain by adding them up

$$\begin{aligned} \sum_{j=1}^{N_2}\alpha_j b_2^* &= -\sum_{j=1}^{N_2}\alpha_j - w^{*T}\sum_{j=1}^{N_2}\alpha_j\Phi(x_j) \\ \Omega b_2^* &= -\Omega - w^{*T}\,m_2^* \\ b_2^* &= -1+\frac{1}{\Omega}\left\|m_1^*\right\|^2-\frac{1}{\Omega}m_1^{*T}\,m_2^* \end{aligned} \tag{9.65}$$

Since $b_1^* = b_2^*$, $b^* = 1/2\left(b_1^* + b_2^*\right)$ which can be written as $b^* = \frac{1}{2\Omega}\left(\left\|m_2^*\right\|^2 - \left\|m_1^*\right\|^2\right)$

ITL Interpretation of the SVM

The classifier developed in Section 9.6.1 is entirely determined by the mean vectors

$m_1$ and $m_2$ of the training data, since both $w$ and $b$ are determined by these vectors. For the classifier to perform well on test data, we are totally dependent on these mean vectors to truly represent the structure of the data. For example, the presence of outliers in the training set may affect the computation of $w$ and $b$ in such a way that the performance of the classifier is degraded. This may be remedied by allowing the contribution of each training data point to the mean vectors to be weighted differently. Let us therefore introduce the weighting components $\alpha_i > 0$ associated with $c_1$, and $\alpha_j > 0$ associated with $c_2$. The weighted mean vectors then become

$$m_1 = \frac{1}{\Omega_1}\sum_{i=1}^{N_1}\alpha_i\Phi(x_i), \qquad m_2 = \frac{1}{\Omega_2}\sum_{j=1}^{N_2}\alpha_j\Phi(x_j) \tag{9.66}$$

By introducing such weighted mean vectors, we also need to introduce some criterion to determine proper weights. Such a criterion should be optimal with respect to classifier performance. The performance of a classifier is measured by its success rate on test data. Hence, the classifier should generalize well. In statistical learning theory, it has been shown that minimization of the squared norm of the hyperplane weight vector, while satisfying the classification constraints on the training data, improves generalization performance.

Based on the arguments above, we may relate the vector $w = m_1 - m_2$ to the SVM weight vector $w^* = m_1^* - m_2^*$. Recall that the SVM is exactly based on regularization by minimization of $\|w^*\|^2$. The minimization is accompanied by the classification constraints of Eq. (9.57), which ensures that the training data is classified correctly. Taking a closer look at the information potentials associated with the weighted mean vectors (Eq. (9.53)), we can write

$$\left\|m_1\right\|^2 = \frac{1}{\Omega_1^2}\sum_{i=1}^{N_1}\sum_{i'=1}^{N_1}\alpha_i\alpha_{i'}\kappa(x_i, x_{i'}) = \int \hat{p}^2(x)dx \tag{9.67}$$

Thus, the weighted mean vector $m_1$ is associated with

$$\hat{p}(x) = \frac{1}{\Omega_1} \sum_{i'=1}^{N_1} \alpha_i \kappa(x, x_i) \qquad (9.68)$$

a weighted Parzen window estimator in the input space. We likewise have the same for the second class. However, in this case, the kernels which constitute these Parzen window estimators are no longer equally important. Recall that to derive the original classification rule based on the $D_{ED}$ of Eq. (9.55) we assumed that $N_1 = N_2$. Using the weighted Parzen window estimators instead, it is easily found that the corresponding assumption becomes $\Omega_1=\Omega_2=\Omega$. Therefore,

$$m_1 = \frac{1}{\Omega} m_1 *, \quad m_2 = \frac{1}{\Omega} m_2 *, \quad w = \frac{1}{\Omega} w * \qquad (9.69)$$

Now, using the weighted Parzen window estimators we may express the SVM optimization problem in an information theoretic framework as follows

$$\min_{\alpha_i, \alpha_j} \|w*\|^2 = \min_{\alpha_i, \alpha_j} \Omega^2 \|w\|^2 = \min_{\alpha_i, \alpha_j} \Omega^2 \|m_1 - m_2\|^2 \qquad (9.70)$$

Since $||m_1 - m_2||^2$ is the Mercer kernel feature space equivalent to the $D_{ED}$, we have

$$\min_{\alpha_i, \alpha_j} \Omega^2 \|m_1 - m_2\|^2 = \min_{\alpha_i, \alpha_j} \Omega^2 \int (\hat{p}(x) - \hat{q}(x))^2 dx \qquad (9.71)$$

The optimization is subject to classification constraints, expressed as

$$\begin{aligned} w*^T \Phi(x_i) + b* \geq +1, \Leftrightarrow \quad & \Omega w^T \Phi(x_i) + \Omega b \geq 1 \\ \Leftrightarrow \quad & \hat{p}(x_i) - \hat{q}(x_i) + b \geq \frac{1}{\Omega} \qquad i = 1,.., N_1 \end{aligned} \qquad (9.72)$$

and

$$\begin{aligned} w*^T \Phi(x_j) + b* \leq -1, \Leftrightarrow \quad & \Omega w^T \Phi(x_j) + \Omega b \leq -1 \\ \Leftrightarrow \quad & \hat{p}(x_j) - \hat{q}(x_j) + b \leq -\frac{1}{\Omega} \qquad j = 1,.., N_2 \end{aligned} \qquad (9.73)$$

Therefore, the SVM classification rule, using the weighted Parzen window estimators,

becomes

$$x_0 \in c_1: \quad w^{*T}\Phi(x_j)+b^* \geq 0, \Leftrightarrow \quad \Omega w^T \Phi(x_0)+\Omega b \geq 0 \\ \Leftrightarrow \quad \hat{p}(x_0)-\hat{q}(x_0)+b \geq 0 \tag{9.74}$$

The weighted Parzen window estimators $\hat{p}(x), \hat{q}(x)$, as defined above, are bona fide density estimators. That is, they are always non-negative and integrate to one. However, since the weights are determined by minimizing $D_{ED}$, which puts emphasis on the points close to the class boundary trying to maximize the overlap between the class PDFs, we do not regard them as proper estimators for the PDFs that generated the data. From SVM theory, we know that in the Mercer kernel feature space, the only non-zero weighting components are those which correspond to data patterns on the margin.

In the input space, it seems that the corresponding non-zero weighting components will be associated with data patterns near the class boundary. We therefore interpret the minimization of the $D_{ED}$ as a sparseness criterion, which tunes the classifier to those patterns which are near the boundary. The other data patterns should be much easier to classify correctly, and are not given any weight in the design of the classifier.

The performance of the classifier is secured by the classification constraints. Note that weighted Parzen window estimators have been previously proposed for improved Parzen window-based Bayes classification [28], [29]. In summary, we have found that one may view the SVM theory in feature space in terms of weighted Parzen density estimation in the input space, where regularization is obtained by minimizing the integrated squared error criterion. Hence, in an information theoretic framework, the support vector machine is formulated by introducing the weights $\alpha_i>0$ and $\alpha_j>0$, and estimating the class densities according to

$$\hat{p}(x) = \frac{1}{\Omega}\sum_{i=1}^{N_1}\alpha_i\kappa(x, x_i), \qquad \hat{q}(x) = \frac{1}{\Omega}\sum_{j=1}^{N_2}\alpha_j\kappa(x, x_j) \tag{9.75}$$

The weights, and hence $\hat{p}(x), \hat{q}(x)$, are learned by enforcing a regularization criterion

$$\min_{\alpha_i, \alpha_j} \Omega^2 \int (\hat{p}(x) - \hat{q}(x))^2 dx \tag{9.76}$$

subject to the classification constraints,

$$\begin{aligned} \hat{p}(x_i) - \hat{q}(x_i) + b &\geq +\frac{1}{\Omega} \quad \forall x_i \in c_1 \\ \hat{p}(x_j) - \hat{q}(x_j) + b &\leq -\frac{1}{\Omega} \quad \forall x_j \in c_2 \end{aligned} \tag{9.77}$$

## 9.7. Case Study: Spike Train Computation in a RKHS

In the previous sections, we defined the ITL RKHS and showed several connections of this perspective with kernel methods. Although the original data space structure was sufficient to perform computations required for machine learning, the RKHS augmented this perspective and provided an elegant mathematical approach to do information theoretic learning. Certain types of data (e.g. point processes), however, do not naturally have the necessary structure to implement machine learning algorithms. For these cases, the structure can be obtained by the construction of an RKHS, and learning problems can easily be formulated and solved in the RKHS. As an application example we illustrate this methodology for spike trains. A spike train $s \in S(T)$ is a simplified representation of a neuron's activity, specified by a sequence of ordered spike times $s = \{t_m \in T : m = 1,\ldots, N\}$ corresponding to the time instants in the interval $T = [0, T]$ at which a neuron emits a spike (i.e., it "fires") [30].

## Kernel function for spike trains

The first step in the construction of the RKHS is to define a symmetric non-negative definite kernel function. There are two basic approaches to do this for spike trains. The first approach follows the ideas from kernel methods, and utilizes the kernel $\kappa$ to build an RKHS on spike times [31][32]. By writing the spike train as a sum of impulses centered at the spike times, and utilizing the linearity of the inner product in the RKHS, it is then possible to extend this RKHS to spike trains. An alternative approach is to build the RKHS by defining the kernel function on statistical descriptors of spike trains. We will follow the later approach since it is more insightful, and closely parallels the construction of the ITL RKHS.

A spike train is a realization of an underlying stochastic point process [33]. In general, to completely characterize a point process, the conditional intensity function must be used. However, for simplicity, here we focus on the *special case of Poisson processes*, which are memoryless and therefore the intensity function (or rate function) completely describes the point process [33]. The general case is considered in Paiva et al. [34]. In a sense, it can be said that intensity functions play for Poisson point processes the same role as PDFs for random variables, since both are complete statistical functional descriptors [35]. Thus, it makes sense to build the RKHS for spike trains following an approach similar to the construction of the ITL RKHS.

Consider two spike trains, $s_i, s_j \in S(T)$, with $i, j \in N$. Denote the intensity of the underlying Poisson processes by $\lambda_{s_i}(t)$ and $\lambda_{s_j}(t)$, respectively, where $t \in [0,T]$ denotes the time coordinate. As with the CIP, we focus first on the case of deterministic statistical descriptors and consider the estimation problem in the next section. For any practical spike train with finite duration $T$, we have that

$$\int_T \lambda_{s_i}^2(t)dt < \infty. \tag{9.78}$$

As a consequence, the intensity functions of spike trains are valid elements of $L_2([0,T])$. Therefore, we can define in this space a kernel function of intensity functions given by the usual inner product in $L_2$,

$$I(s_i, s_j) = \left\langle \lambda_{s_i}, \lambda_{s_j} \right\rangle_{L_2(T)} = \int_T \lambda_{s_i}(t)\lambda_{s_j}(t)dt. \tag{9.79}$$

We shall refer to $I(\cdot,\cdot)$ as the memoryless cross-intensity (mCI) kernel. The proof that the mCI is indeed a symmetric non-negative definite kernel, follows the same steps as the proof of Property 9.1, and is omitted here. Hence, the mCI induces an RKHS, denoted $H_I$. Comparing the definition of the mCI kernel with the CIP kernel, it is clear that both kernels incorporate the statistics descriptors directly into the kernel function. Thus, both are complete statistical operators. As with the CIP kernel, the definition of the mCI naturally induces a norm in the space of the intensity functions,

$$\left\| \lambda_{s_i}(\cdot) \right\|_{L_2(T)} = \sqrt{\left\langle \lambda_{s_i}, \lambda_{s_i} \right\rangle_{L_2(T)}} = \sqrt{\int_T \lambda_{s_i}^2(t)dt} \tag{9.80}$$

which is very useful for the formulation of optimization problems.

Estimation of the memoryless cross-intensity kernel

Spike trains are realizations of underlying point processes, but, as defined, the mCI kernel is a deterministic operator on the point processes rather than on the observed spike trains. Thus, in practice, the kernel function is evaluated with the intensity functions estimated from spike trains. A well known methodology for estimation of the intensity function is kernel smoothing [33]. Given a spike train $s_i$ with spike times $\{t_m^i \in T : m = 1,\ldots, N_i\}$ the estimated intensity function is

$$\hat{\lambda}_{s_i}(t) = \sum_{m=1}^{N_i} h(t - t_m^i), \tag{9.81}$$

where $h$ is the smoothing function. This function must be non-negative and integrate to one over the real line (just like a PDF). Commonly used smoothing functions are the Gaussian, Laplacian and $\alpha$ -functions, among others.

Consider spike trains $s_i, s_j \in S(T)$ with estimated intensity functions $\hat{\lambda}_{s_i}(t)$ and $\hat{\lambda}_{s_j}(t)$ according to Eq. (9.81). Substituting the estimated intensity functions in the definition of the mCI kernel (Eq. (9.79)) yields

$$\hat{I}(s_i, s_j) = \sum_{m=1}^{N_i} \sum_{n=1}^{N_j} \kappa(t_m^i - t_n^j). \tag{9.82}$$

where $\kappa$ is the 'kernel' obtained by the autocorrelation of the smoothing function $h$.

It must be remarked the difference the space of intensity functions from which the mCI kernel function was defined, and the RKHS induced by this kernel. As with the CIP and the ITL RKHS, it can be shown there is a congruence mapping between the two spaces. Therefore, the same result can be obtained from either space.

It is interesting to verify the parallel of concepts and derived operators between the RKHS just defined for spike trains and the ITL RKHS. Yet, the most important result is that the construction of this RKHS provides the structure needed for computation with this data, and which otherwise would not be possible.

## Principal component analysis

To exemplify these developments in an application, we now derive the algorithm to perform principal component analysis (PCA) of spike trains. The derivation of PCA in the RKHS is general, and applicable with other kernel functions. Interestingly, this is the traditional approach in the functional analysis literature [36].

A well known example of discrete PCA done in an RKHS is kernel PCA [37][38].

*Derivation of the algorithm*: Consider a set of spike trains, $\{s_i \in S(T), i = 1,\ldots, N\}$, for which we wish to determine the principal components. Computing the principal components of the spike trains directly is not feasible because we would not know how to define a principal component (PC), however, this is a trivial task in an RKHS. Let $\{\Lambda_{s_i} \in H_I, i = 1,\ldots, N\}$ be the set of elements in the RKHS corresponding to the given spike trains. Denote the mean of the transformed spike trains as

$$\overline{\Lambda} = \frac{1}{N}\sum_{i=1}^{N} \Lambda_{s_i}, \tag{9.83}$$

and the centered transformed spike trains (i.e., with the mean removed) can be obtained as

$$\tilde{\Lambda}_{s_i} = \Lambda_{s_i} - \overline{\Lambda}. \tag{9.84}$$

PCA finds an orthonormal transformation providing a compact description of the data. Determining the principal components of spike trains in the RKHS can be formulated as the problem of finding the set of orthonormal vectors in the RKHS such that the projections of the centered transformed spike trains $\{\tilde{\Lambda}_{s_i}\}$ have *maximum variance*. This means that the principal components can be obtained by solving an optimization problem in the RKHS. A function $\xi \in H_I$ (i.e., $\xi : S(T) \to R$) is a principal component if it maximizes the cost function

$$J(\xi) = \sum_{i=1}^{N} \left[\mathrm{Proj}_{\xi}(\tilde{\Lambda}_{s_i})\right]^2 - \rho\left(\|\xi\|^2 - 1\right) \tag{9.85}$$

where $\mathrm{Proj}_{\xi}(\tilde{\Lambda}_{s_i})$ denotes the projection of the $i$th centered transformed spike train onto $\xi$, and $\rho$ is the Lagrange multiplier to the constraint $\left(\|\xi\|^2 - 1\right)$ imposing that the principal components have unit norm. To evaluate this cost function one needs to be

able to compute the projection and the norm of the principal components. However, the inner product needed for the projection and the norm are naturally defined. Thus, the above cost function can be expressed as

$$J(\xi) = \sum_{i=1}^{N} \left\langle \tilde{\Lambda}_{s_i}, \xi \right\rangle_{H_I}^2 - \rho\left( \left\langle \xi, \xi \right\rangle_{H_I} - 1 \right), \tag{9.86}$$

By the representer theorem [38], $\xi$ is restricted to the subspace spanned by the centered transformed spike trains $\{\tilde{\Lambda}_{s_i}\}$. Consequently, there exist coefficients $b_1,\ldots,b_N \in \mathbb{R}$ such that

$$\xi = \sum_{i=1}^{N} b_j \tilde{\Lambda}_{s_i} = b^T \tilde{\Lambda} \tag{9.87}$$

where $b^T = [b_1,\ldots,b_N]$ and $\tilde{\Lambda}(t) = \left[ \tilde{\Lambda}_{s_1}(t),\ldots,\tilde{\Lambda}_{s_N}(t) \right]^T$. Substituting in Eq. (9.86) yields

$$\begin{aligned} J(\xi) &= \sum_{i=1}^{N} \left( \sum_{j=1}^{N} b_j \left\langle \tilde{\Lambda}_{s_i}, \tilde{\Lambda}_{s_j} \right\rangle \right) \left( \sum_{k=1}^{N} b_k \left\langle \tilde{\Lambda}_{s_i}, \tilde{\Lambda}_{s_k} \right\rangle \right) \\ &\quad + \rho \left( 1 - \sum_{j=1}^{N} \sum_{k=1}^{N} b_j b_k \left\langle \tilde{\Lambda}_{s_i}, \tilde{\Lambda}_{s_k} \right\rangle \right) \\ &= b^T \tilde{I}^2 b + \rho\left(1 - b^T \tilde{I} b\right). \end{aligned} \tag{9.88}$$

where $\tilde{I}$ is the Gram matrix of the centered spike trains; that is, the $N \times N$ matrix with elements

$$\begin{aligned} \tilde{I}_{ij} &= \left\langle \tilde{\Lambda}_{s_i}, \tilde{\Lambda}_{s_j} \right\rangle \\ &= \left\langle \Lambda_{s_i} - \overline{\Lambda}, \Lambda_{s_j} - \overline{\Lambda} \right\rangle \\ &= \left\langle \Lambda_{s_i}, \Lambda_{s_j} \right\rangle - \frac{1}{N} \sum_{l=1}^{N} \left\langle \Lambda_{s_i}, \Lambda_{s_l} \right\rangle - \frac{1}{N} \sum_{l=1}^{N} \left\langle \Lambda_{s_l}, \Lambda_{s_j} \right\rangle + \frac{1}{N^2} \sum_{l=1}^{N} \sum_{n=1}^{N} \left\langle \Lambda_{s_l}, \Lambda_{s_n} \right\rangle. \end{aligned} \tag{9.89}$$

In matrix notation,

$$\tilde{I} = I - \frac{1}{N}(1_N I + I 1_N) + \frac{1}{N^2} 1_N I 1_N, \tag{9.90}$$

where $I$ is the Gram matrix of the inner product of spike trains $I_{ij} = \left\langle \Lambda_{s_i}, \Lambda_{s_j} \right\rangle$, and $1_N$ is the $N \times N$ matrix with all ones. This means that $\tilde{I}$ can be computed directly in terms of $I$ without the need to explicitly remove the mean of the transformed spike trains.

From Eq. (9.88), finding the principal components simplifies to the problem of estimating the coefficients $\{b_i\}$ that maximize $J(\xi)$. Since $J(\xi)$ is a quadratic function its extrema can be found by equating the gradient to zero. Taking the derivative with regards to $b$ (which characterizes $\xi$) and setting it to zero results in

$$\frac{\partial J(\xi)}{\partial b} = 2\tilde{I}^2 b - 2\rho \tilde{I} b = 0, \tag{9.91}$$

and thus corresponds to the eigendecomposition problem

$$\tilde{I} b = \rho b. \tag{9.92}$$

This means that any eigenvector of the centered Gram matrix is a solution of Eq. (9.91). Thus, the eigenvectors determine the coefficients of Eq. (9.87) and characterize the principal components. It is easy to verify that, as expected, the variance of the projections onto each principal component equals the corresponding eigenvalue squared. So, the ordering of $\rho$ specifies the relevance of the principal components.

To compute the projection of a given input spike train $s$ onto the $k$ th principal component (corresponding to the eigenvector with the $k$ th largest eigenvalue) we need only to compute in the RKHS the inner product of $\Lambda_s$ with $\xi_k$. That is,

$$\begin{aligned}\mathrm{Proj}_{\xi_k}(\Lambda_s) &= \left\langle \Lambda_s, \xi_k \right\rangle_{H_I} \\ &= \frac{1}{N}\sum_{i=1}^{N} b_{ki}\left\langle \Lambda_s, \tilde{\Lambda}_{s_i} \right\rangle \\ &= \frac{1}{N}\sum_{i=1}^{N} b_{ki}\left( I(s,s_i) - \frac{1}{N}\sum_{j=1}^{N} I(s,s_j) \right).\end{aligned} \tag{9.93}$$

An alternative approach to derive PCA for spike trains would be to utilize the inner product in the space of intensity functions directly. Basically, the derivation would follow the same steps but now in terms of intensity functions, rather than elements in the RKHS. Nevertheless, due to the congruence between this space and the RKHS induced by the mCI kernel, the result is the same. The key difference is that in Eq. (9.87), the principal components are written as weighted combinations of intensity functions, with weights given by the eigenvectors of the centered Gram matrix. That is, this approach allows the principal components to be obtained as intensity functions. Since intensity functions characterize spike trains, this perspective can be very telling of the underlying data structure.

## Results

To illustrate the algorithm just derived we performed a simple experiment. We generated two template spike trains comprising of 10 spikes uniformly random distributed over an interval of 0.25s. In a specific application these template spike trains could correspond, for example, to the average response of a culture of neurons to two distinct but fixed input stimuli. For the computation of the coefficients of the

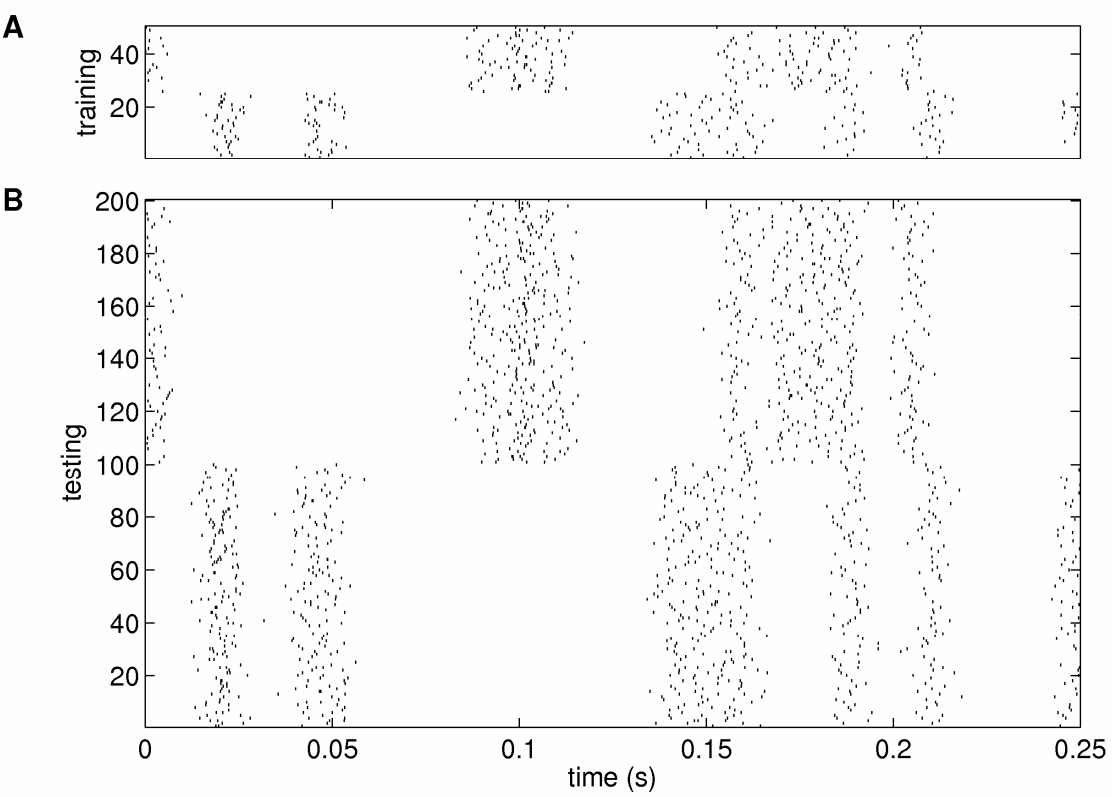


Figure 9.3. (A) Spike trains used for evaluation of the eigendecomposition coefficients of PCA algorithm, and (B) for testing of the result. In either case, the first half of spike trains corresponds to the first template and the remaining to the second template.

eigendecomposition ("training set"), we generated a total of 50 spike trains, half for each template, by randomly copying each spike from the template with probability 0.8 and adding zero mean Gaussian distributed jitter with standard deviation 3ms. For testing of the obtained coefficients, 200 spike trains were generated following the same procedure. The simulated spike trains are shown in Figure 9.3.

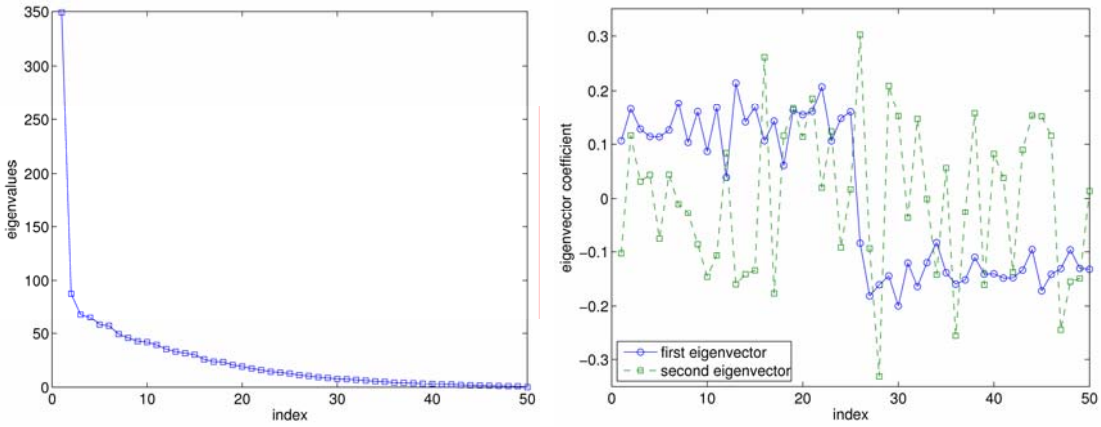


Figure 9.4. Eigenvalues $\{\rho_l, l = 1,..,100\}$ in decreasing order (a) and first two eigenvectors (b) of the eigendecomposition of the centered Gram matrix $\tilde{I}$.

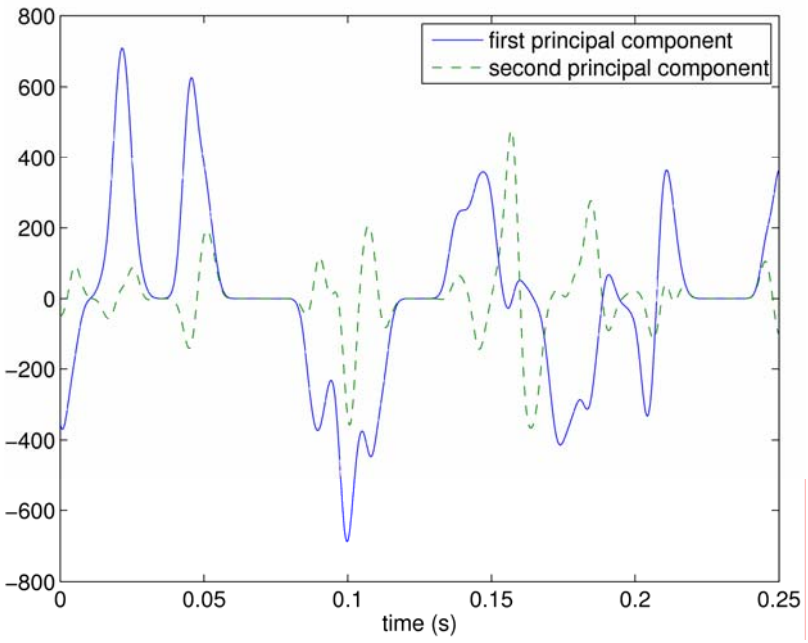


Figure 9.5. First two principal component functions (i.e., eigenfunctions) in the space of intensity functions.

With the PCA algorithm derived previously, we computed the eigendecomposition of the matrix $\tilde{I}$. The evaluation of the mCI kernel was estimated from the spike trains according to Eq. (9.82), and computed with a Gaussian kernel with size 2ms. The eigenvalues $\{\rho_l, l = 1,..,100\}$ and first two eigenvectors are shown in Figure 9.4. The first eigenvalue alone accounts for more than 26% of the variance

of the dataset in the RKHS space. Although this value is not impressive, its importance is clear since it is nearly 4 times higher than the second eigenvalue (6.6%). Furthermore, notice that the first eigenvector clearly shows the separation between spike trains generated from different templates (Figure 9.4(b)). This again can be seen in the first principal component function, shown in Figure 9.5, which reveals the location of the spike times used to generate the templates while discriminating between them with opposite signs. Around periods of time where the spike from both templates overlap the first principal component is zero. As can be seen from the second principal component function, the role of the second eigenvector is to account for the dispersion in the data capable of differentiate spike trains generate from different templates.

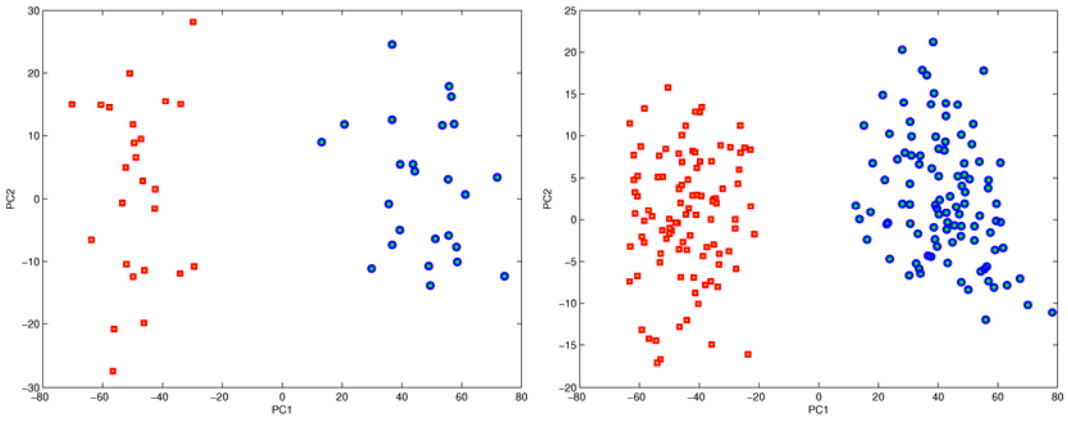


Figure 9.6. Projection of spike trains onto the first two principal components. (a) Shows the projection of the spike trains in the training set and (b) in the testing set. The different point marks differentiate between spike trains corresponding to each one of the classes.

For evaluation and testing, both datasets where projected onto the first two principal components. Figure 9.6 shows the projected spike trains. As noted from the difference between the first and second eigenvalues, the first principal component is the main responsible for the dispersion between classes of the projected spike trains. This happens because the direction of maximum variance is the one that passes

through both clusters of points in the RKHS due to the small dispersion within class. The second principal component seems to be responsible for dispersion due to the jitter noise introduced in the spike trains, and suggests that other principal components play a similar role.

## 9.8. Conclusion

This chapter presented formally the relationship between the ITL descriptors in the preceding chapters and a RKHS that we called $H_v$. The elements of $H_v$ are PDFs, and the kernel is the cross information potential. The inner product between two PDFs was crucial to measure distances for clustering and it also appeared in the divergence measures and quadratic mutual information. From the RKHS perspective we can easily recognize its central role because it defines the natural similarity metric in the space of PDFs.

In the previous chapters we estimated all these quantities directly from samples using the Information Potential, i.e. the double sum of pairwise interactions between the data samples. We saw in this chapter that these estiamtors correspond rather directly to kernel methods. Indeed, when one places a kernel on a data sample we are defining a function that exists in a RKHS defined by the kernel. Therefore, we can reinterpret the information theoretical estimators as kernel operations: the information potential estimator is nothing but the mean square norm of the projected samples, the Cauchy Schwarz divergence estimator is nothing but the log of the cosine of the angles between the projected samples.

Finally, we can establish a relationship between $H_v$ and the $H_\kappa$ defined by the kernel used in the ITL estimators: the mean value (over the data samples) in the

RKHS defined by the kernel exists in the $H_v$. We were able to show that certain statistical operators defined in kernel spaces indeed correspond to ITL quantities (e.g. MDD becomes the Euclidean distance in $H_v$)

It is therefore possible to also interpret well known solutions in kernel methods with ITL descriptors. We showed that the SVM can be interpreted as the maximization of the Euclidean distance between classes estimated with a weighted Parzen window, where the weights are basically the Lagrange multipliers that define the support vectors.

We hoped to convince the reader that the CIP $V(p,q) = \int p(x)q(x)dx$ is the key concept behind all these relations, since it defines $H_v$ and when $p(x)=q(x)$ it defines the quadratic norm of the PDF that yields the estimators for Renyi's entropy. Therefore, synergisms were established between a statistical view of information and its functional view. The information theoretical descriptors presented by Alfred Renyi just take the log of the projected data in $H_v$, which just changes the weighting of the inner product as we established in Chapter 2.

As an application of the RKHS methodology we apply it to a problem where the original space structure of the data does not support operations required for machine learning. One such example is the space of point processes. However, we can define a positive definite function in the point process space that builds a RKHS, where optimization algorithms can be carried out easily. We demonstrate this with PCA, but could likewise have used dissimilarity between spike trains, which as we have seen is associated with divergence.