1 Motivation: Kernels in Machine Learning

Kernels have been found to be extremely useful in basic machine learning tasks like classification, regression and others. A full treatment of the role of kernels in machine learning can be found in the book by Schölkopf and Smola [6]. In these notes, in order to motivate the study of kernel functions, we examine in outline the case of classification, and that too only classification using support vector machines. We will formally define kernels only in Section 2.

1.1 Motivating example: Support Vector Machines

Classification is a basic problem in machine learning. In the so-called supervised setting of the 2-class classification problem we are given a training set which consists of points from a domain, each labelled with 1 or -1, denoting which class they belong to. The goal is to find an easily computable function of the domain that maps unlabelled elements of the domain (query points) to a label. In the case where the domain is the $d$-dimensional space $\mathbb{R}^d$ a natural way of approaching this problem is separation by a hyperplane, a method that is attributed to Vapnik in [6]. We state this formally.

**Problem 1.1** (Two-class classification using hyperplanes). Given a training set $S \subseteq \mathbb{R}^d \times \{-1, 1\}$, find a hyperplane $\mathbf{w}^T \mathbf{x} + b = 0$ such that for

\[(x, 1) \in S : \mathbf{w}^T \mathbf{x} + b \geq 0,\]

and for

\[(x, -1) \in S : \mathbf{w}^T \mathbf{x} + b < 0.\]
To understand why the conditions amount to “separation” by a hyperplane we recall that \( w^T x \) is actually the dot product between the vectors \( w \) and \( x \) and can be thought of as the projection of \( x \) on \( w \) scaled by the constant \( c \). Hence \( w^T x + b = 0 \) is satisfied by all points for which the scaled projection has value exactly \(-b\) and we can separate all points of \( \mathbb{R}^d \) into two sets, those on the positive “side” of the hyperplane and those on the negative “side” of it, i.e., those whose scaled projection is greater that \(-b\) and those whose scaled projection is smaller.

We note that there may be cases where there is no solution to Problem 1.1, but for our purposes we assume we are dealing with cases where a solution exists. This problem has been approached in the literature by considering its optimization version which seeks to find a hyperplane that not only separates the two classes but has the maximum distance from every training point. This can be stated formally.

**Problem 1.2** (Optimal two-class classification using hyperplanes). Given a training set \( S \subseteq \mathbb{R}^d \times \{-1, 1\} \), and a constant \( c > 0 \),

\[
\maximize \min_{(x,y) \in S} y(w^T x + b),
\]

subject to

\[\|w\| = c.\]

In the machine learning literature \((w^T x + b)/c\) is known as the geometric margin of the point \( x \) and \( y \cdot (w^T x + b)/c \), which is always positive, is known as its functional margin. The objective is to maximize the smallest functional margin. Since the scaling factor is in our control, we can restate this problem by demanding that the functional margin be fixed.

**Problem 1.3** (Alternate formulation of Problem 1.2). Given a training set \( S \subseteq \mathbb{R}^d \times \{-1, 1\} \),

\[
\maximize \frac{1}{2} \|w\|^2
\]

subject to

\[\forall (x,y) \in S : y(w^T x + b) \geq 1.\]

The \(1/2\) and the square in the objective function are to make the dual form more convenient. Going through the method of Lagrange multipliers we get the dual problem:

**Problem 1.4** (Dual version of Problem 1.2). Given a training set \( S \subseteq \mathbb{R}^d \times \{-1, 1\} \),

\[
\maximize \sum_{i=1}^{\|S\|} \alpha_i - \frac{1}{2} \sum_{i=1}^{\|S\|} \sum_{j=1}^{\|S\|} y_i y_j \alpha_i \alpha_j x_i^T x_j
\]

subject to

\[\alpha_i \geq 0, 1 \leq i \leq \|S\|, \text{ and } \sum_{i=1}^{\|S\|} \alpha_i y_i = 0.\]
A detailed treatment of how the dual problem is derived can be seen in [4]. For our purposes we observe that the objective function in Problem 1.4 depends entirely on pairwise dot products between training points. This leads to the following insight:

If we transform \( S \subset \mathbb{R}^d \) through a transformation \( \kappa : \mathbb{R}^d \to X \), we can find a separating hyperplane in \( X \) by solving Problem 1.4 with \( x_i^T x_j \) replaced by \( \langle \kappa(x_i), \kappa(x_j) \rangle \), where \( \langle \cdot, \cdot \rangle \) is an inner product defined on \( X \).

Such a transformation could be advantageous if the training set is not separable in \( \mathbb{R}^d \) but its transformed version is separable in \( X \) (with hyperplanes defined via the inner product defined on \( X \)). Even if the transformed version is not separable, it may be more amenable to low-error classification using regularization or other techniques.

However, computing the dot product in the transformed case could be computationally more demanding, especially if the \( X \) has greater dimension than \( d \). This leads us to the following conclusion

For a “useful” transformation \( \kappa : \mathbb{R}^d \to X \), if we can find an efficiently computable function \( \varphi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R} \) such that for all \( x, y \in \mathbb{R}^d \), \( \varphi(x, y) = \langle \kappa(x), \kappa(y) \rangle \), then we can solve Problem 1.4 in \( X \) efficiently.

We will see ahead that there is a class of functions called kernels that satisfy the condition that they can be represented as inner products in a transformed space. Whether this transformed space is useful or not depends on the application to which the machine learning method is being applied. The question of usefulness is outside the scope of these notes.

### 1.2 Transforming the data: Some examples

**Example 1.5.** Suppose \( S \subseteq \mathbb{R}^d \), let \( \kappa(x) = (x(i)x(j) : 1 \leq i, j \leq d) \).

This transformation maps a \( d \)-dimensional vector to a \( d^2 \) dimensional vector comprising all the two coordinate products of the original vector. Prima facie it appears that this means that computing a dot product in the transformed space would be more expensive, \( d^2 \) operations as opposed to the \( d \) operations required to compute a dot product in \( \mathbb{R}^d \). However we observe an interesting property of the transformation. Given \( x, z \in S \),

\[
\phi(x) \phi(z)) = \sum_i \sum_j x(i)x(j)z(i)z(j) \\
= \sum_i x(i)z(i) \sum_j x(j)z(j) \\
= (x^T z)^2
\]

This means that instead of performing \( d^2 \) operations to compute an inner product in the transformed space we can simply compute \( d + 1 \) operations because we have identified a (more efficiently computable) function that gives us the value of the inner product in the transformed space. A similar transformation is the next one:

**Example 1.6.** Suppose \( S \subseteq \mathbb{R}^2 \). For some \( c \in \mathbb{R} \), let \( \phi(x) = (x(i)x(j) : 1 \leq i, j \leq d; \sqrt{2c}x(i) : 1 \leq i \leq d; c) \)
We omit the calculation but in this case we find that

\[ \phi(x)\phi(z)) = (x^T z + c)^2 \]

. We can further generalize this.

**Exercise 1.1.** Find the mapping \( \phi \) such that the for \( x, z \in \mathbb{R}^d \), \( \phi(x)^T \phi(z) = (x^T z + c)^d \).

The efficiency gained by identifying these function is of no use unless we can say that such a mapping is useful. Schölkopf and Smola [6] claim that the kinds of transformations we have seen above, they call them “monomial mappings”, are useful in the case of pattern analysis/classification in images where each coordinate of the data vector corresponds to a pixel.

We now turn to the general class of functions that Mercer showed in 1909 can be represented as inner products in a transformed space: positive definite kernels.

## 2 Kernels

The material in this and subsequent sections closely follows the presentation in the book by Berg, Christensen and Ressel [1]. In general we will assume that all scalars are drawn from complex numbers, \( \mathbb{C} \), clarifying explicitly when we are restricting to the reals. We will use \( \overline{c} \) to denote the complex conjugate of \( c \in \mathbb{C} \). For an \( n \times m \) complex matrix \( A \), we will use \( A^* \) to denote its conjugate transpose, i.e., \( A^*_{ij} = \overline{A_{ji}} \).

### 2.1 Definitions

We first begin with some basic definitions from linear algebra.

**Definition 2.1 (Positive definite matrix).** An \( n \times n \) matrix \( A \) is called positive definite (p.d.) if

\[ \sum_{j,k=1}^n c_j \overline{c_k} A_{jk} \geq 0 \]  

(1)

for every \( \{c_1, \ldots, c_n\} \subseteq \mathbb{C} \). If the inequality (1) holds strictly for every \( \{c_1, \ldots, c_n\} \subseteq \mathbb{C} \) then the matrix is called strictly positive definite.

**Remarks.**

1. In the linear algebra literature the term positive semidefinite is often used for matrices that satisfy (1) and positive definite is used for matrices that satisfy (1) strictly.

2. Implicit in Definition 2.1 is the fact that the sum on the LHS of (1) is real. If it were complex then comparison with 0 would be meaningless. We recall that if \( A \) is Hermitian then the LHS is guaranteed to be real.

3. If we consider the \( n \times 1 \) vector \( c \) such that \( c(i) = c_i \) then (1) can be rewritten as \( c^* A c \geq 0 \).

Clearly, we can define negative definite matrices by reversing the direction of the \( \geq \) sign in (1), but we define an interesting class of matrices that contain the negative definite matrices.
Definition 2.2 (Conditionally negative definite matrix). An $n \times n$ matrix $A$ is called conditionally negative definite (c.n.d.) if it is Hermitian and if

$$\sum_{j,k=1}^{n} c_j \bar{c}_k A_{jk} \leq 0$$ (2)

for every $\{c_1, \ldots, c_n\} \subseteq \mathbb{C}$ such that $\sum_{i=1}^{n} c_i = 0$.

Note that here we don’t need to assume that the LHS of (2) is real since we have explicitly specified that $A$ is Hermitian. With these definitions in hand we are ready to define two interesting classes of bivariate functions. In the following we will abbreviate positive definite as p.d. and conditionally negative definite as c.n.d.

Definition 2.3 (Positive definite kernel). Let $X$ be a non empty set. $\varphi : X \times X \rightarrow \mathbb{C}$ is a positive definite kernel if

$$\sum_{j,k=1}^{n} c_j c_k \varphi(x_j, x_k) \geq 0$$ (3)

$\forall n \in \mathbb{N}, \{x_1, x_2, \ldots, x_n\} \subseteq X$, and $\{c_1, c_2, \ldots, c_n\} \subseteq \mathbb{C}$.

Example 2.4. For $d > 0$, $\varphi : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ defined as $\varphi(x, y) = x^T y$ is a p.d. kernel.

Proof. For some $n \in \mathbb{N}$, suppose $\{x_1, \ldots, x_n\} \subseteq \mathbb{R}^d$ and $\{c_1, c_2, \ldots, c_n\} \subseteq \mathbb{C}$. Then

$$\sum_{j,k=1}^{n} c_j \bar{c}_k \varphi(x_j, x_k) = \sum_{j,k=1}^{n} c_j \bar{c}_k x_j^T x_k = \sum_{j,k=1}^{n} (c_k x_k)^* (c_j x_j) = \left(\sum_{j=1}^{n} c_j x_j\right)^* \left(\sum_{j=1}^{n} c_j x_j\right) = \left|\sum_{j=1}^{n} c_j x_j\right|^2 > 0.$$

Example 2.4 shows that all bivariate functions that are dot products of a Euclidean space are p.d. kernels. Later we will see that this is true in general for the inner product of an inner product (pre-Hilbert) space.

Definition 2.5 (Conditionally negative definite kernel). Let $X$ be a non empty set. $\psi : X \times X \rightarrow \mathbb{C}$ is a conditionally negative definite kernel if it is Hermitian, i.e., $\psi(x, y) = \psi(y, x) \forall x, y \in X$, and if

$$\sum_{j,k=1}^{n} c_j \bar{c}_k \psi(x_j, x_k) \leq 0$$ (4)

$\forall n \geq 2, \{x_1, x_2, \ldots, x_n\} \subseteq X$, and $\{c_1, c_2, \ldots, c_n\} \subseteq \mathbb{C}$ such that $\sum c_i = 0$.

Example 2.6. $\psi : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ defined as $\psi(x, y) = (x - y)^2$ is a c.n.d. kernel.
Proof. Since $\psi$ is symmetric so the condition that the kernel should be Hermitian is satisfied. Given $n \geq 2$, let $\{x_1, \ldots, x_n\} \subseteq \mathbb{R}$ and $\{c_1, c_2, \ldots, c_n\} \subseteq \mathbb{C}$ such that $\sum c_i = 0$. Now,

$$\sum_{j,k=1}^{n} c_j \overline{c_k} (x_j - x_k)^2 = \sum_{j=1}^{n} c_j x_j^2 \sum_{k=1}^{n} \overline{c_k} + \sum_{k=1}^{n} \overline{c_k} x_k^2 \sum_{j=1}^{n} c_j - 2 \sum_{j,k=1}^{n} c_j \overline{c_k} x_j x_k$$

By the condition $\sum c_i = 0$ the first two terms on the RHS are 0 and the third term can be rewritten as

$$-2 \left( \sum_{j=1}^{n} c_j x_j \right) \left( \sum_{j=1}^{n} \overline{c_j} x_j \right) = -2 \left| \sum_{j=1}^{n} c_j x_j \right|^2 \leq 0.$$

\[ \square \]

Remarks.

1. If the inequality (3) (resp. (4)) is strict then the kernel is called strictly positive definite (resp. strictly negative definite).

2. In both Definitions 2.3 and 2.5 if we had imposed the restriction that all the $x_i$s are distinct we would not get a weaker definition since

$$\sum_{j,k=1}^{n} c_j \overline{c_k} \varphi(x_j, x_k) = \sum_{j,k=1}^{p} d_j \overline{d_k} \varphi(x_{\alpha j}, x_{\alpha k})$$

where the $x_{\alpha i}$ are distinct and $d_k := \sum_{i:x_i=x_{\alpha k}} c_i$ for $k = 1, \ldots, p$.

3. If $\sigma : X \rightarrow X$ is a bijection then $\varphi$ is a p.d. (respectively c.n.d.) kernel iff $\varphi \circ (\sigma \times \sigma)$ is a p.d. (resp. c.n.d.) kernel. This comes from the fact that all possible finite subsets of $X$ satisfy the condition (3) (resp. (4)), and so if this condition is satisfied by distinct $x_1, \ldots, x_n$ for some $n \in \mathbb{N}$ (resp $n \geq 2$) then it is satisfied by $\sigma(x_1), \ldots, \sigma(x_n)$ which is also a set of distinct elements from $X$.

4. If $X$ is finite with $|X| = n$, then $\varphi$ is a p.d. (respectively c.n.d.) kernel iff the $n \times n$ matrix $A$ with $A_{ij} = \varphi(x_j, x_k)$, $1 \leq j, k \leq n$ is p.d. (resp. c.n.d.) This can be deduced from the fact that all the principal submatrices of a p.d. matrix are p.d. and from a similar argument for the c.n.d. case.

5. For the c.n.d. case we restrict $n$ to be at least 2 since if $n = 1$ then $c_1$ will have to be 0 to satisfy the condition of summing to 0, i.e., the definition will be trivially true for all bivariate functions.
2.2 Some properties of p.d. and c.n.d. kernels

2.2.1 Basic properties

We prove some foundational properties of the classes of kernels defined above. First we show that the definition of positive definiteness for a kernel implies that the kernel is Hermitian.

**Proposition 2.7** (P.d. kernels are Hermitian). If \( \varphi \) is a p.d. kernel defined on \( X \times X \) then all its diagonal elements are positive, i.e., \( \varphi(x, x) \geq 0 \) for all \( x \in X \) and \( \varphi \) is Hermitian, i.e. \( \varphi(x, y) = \overline{\varphi(y, x)} \) for all \( x, y \in X \).

**Proof.** It is easy to see that \( \varphi(x, x) \) is real and \( \geq 0 \) by setting \( n = 1 \) in Definition 2.3.

We now show that \( \varphi \) is Hermitian. For some \( x, y \in X \) such that \( x \neq y \) let \( A = \begin{pmatrix} \varphi(x, x) & \varphi(x, y) \\ \varphi(y, x) & \varphi(y, y) \end{pmatrix} \) and let \( c_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \). Since \( \varphi \) is p.d. we have that

\[
   c_1^* A c_1 = \varphi(x, x) + \varphi(y, y) + \varphi(x, y) + \varphi(y, x) \geq 0.
\]

Since we have already shown that \( \varphi(x, x) \) and \( \varphi(y, y) \) are real, so this proves that \( \text{Im} \varphi(x, y) = -\text{Im} \varphi(y, x) \).

Similarly, if we take \( c_2 = \begin{pmatrix} 1 \\ -i \end{pmatrix} \), since \( \varphi \) is p.d. we have that

\[
   c_2^* A c_2 = \varphi(x, x) + \varphi(y, y) + i(\varphi(y, x) - \varphi(x, y)) \geq 0,
\]

which can only be true if \( \text{Re} \varphi(x, y) = \text{Re} \varphi(y, x) \).

\( \square \)

**Remark.** As we have just seen, a p.d. kernel can be shown to be Hermitian. However the same argument doesn’t hold for c.n.d. kernels since the vectors \( c_1 \) and \( c_2 \) don’t have the property that their coordinates add up to 0. So we have to add the requirement that c.n.d. kernels are Hermitian in Definition 2.5.

Next we show that for real-valued kernels, symmetry and positive (resp. conditionally negative) definiteness w.r.t. real vectors is good enough to show that they are p.d. (resp. c.n.d.).

**Proposition 2.8** (Real-valued kernels). A real-valued kernel \( \varphi : X \times X \to \mathbb{R} \) is p.d. (resp. c.n.d.) iff \( \varphi \) is symmetric, i.e., \( \varphi(x, y) = \varphi(y, x) \) for all \( x, y \in X \), and

\[
   \sum_{j,k=1}^{n} r_j r_k \varphi(x_j, x_k) \geq 0 \quad \text{(resp.} \leq 0) \quad (5)
\]

for all \( n \in \mathbb{N} \) (resp. \( n \geq 2 \)), \( \{x_1, \ldots, x_n\} \subseteq X \), \( \{r_1, \ldots, r_n\} \subseteq \mathbb{R} \) (resp. additionally \( \sum_{i=1}^{n} r_i = 0 \)).

**Proof.** Given \( \{x_1, \ldots, x_n\} \subseteq X \), consider \( \{c_1, \ldots, c_n\} \subseteq \mathbb{C} \). For each \( j \), let \( a_j = \text{Re} \, c_j \) and \( b_j = \text{Im} \, c_j \). Then we have that

\[
   \sum_{j,k=1}^{n} c_j \overline{c_k} \varphi(x_j, x_k) = \sum_{j,k=1}^{n} (a_j a_k + b_j b_k) \varphi(x_j, x_k) + i \sum_{j,k=1}^{n} (b_j a_k - a_j b_k) \varphi(x_j, x_k)
\]
Symmetry ensures that the imaginary part of the RHS is 0. The real part of the RHS is non-negative for the p.d. case and non-positive for the c.n.d. case because it is the sum of two summations that can be written in the form of the LHS of (5).

We present an important application of Proposition 2.8.

**Example 2.9.** The real-valued kernel $\psi: \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, defined as $\psi(x, y) = (x - y)^2$ is conditionally negative definite (c.n.d.)

**Proof.** Proposition 2.8 stipulates that $\psi$ should be symmetric, which is clearly the case. So now, nor some $n$, we consider $\{x_1, ..., x_n\} \subseteq \mathbb{R}$ and $\{c_1, ..., c_n\} \subseteq \mathbb{R}$ s.t. $\sum_i c_i = 0$.

$$
\sum_{j,k} c_j c_k (x_j - x_k)^2 = \sum_{j,k} c_j c_k (x_j^2 - x_k^2 - 2x_j x_k) \\
= \sum_k c_k \sum_j c_j x_j^2 + \sum_j c_j \sum_k c_k x_k^2 - 2 \sum_{j,k} c_j c_k x_j x_k \\
= 0 + 0 - 2 \left( \sum_i c_i x_i \right)^2 \leq 0
$$

The last simplification follows from $\sum_i c_i = 0$ because of which the first two terms on the RHS equal zero.

We note from the proof that it is fairly straightforward to prove the same property if $\psi$ is defined as the square of the Euclidean distance between two points in $\mathbb{R}^d$, i.e., the square of the euclidean distance gives us a c.n.d. kernel. Later we will see this can be extended to general Hilbert spaces.

### 2.2.2 Relating diagonal and off-diagonal elements

We now present two results that relate the diagonal elements of p.d. and c.n.d. kernels with their off-diagonal elements.

**Proposition 2.10.** For a c.n.d. kernel $\psi: X \times X \to \mathbb{C}$, $\psi(x, x) + \psi(y, y) \leq 2 \text{Re} \, \psi(x, y)$ for all $x, y \in X$.

**Proof.** Define $A = \begin{pmatrix} \psi(x, x) & \psi(x, y) \\ \psi(y, x) & \psi(y, y) \end{pmatrix}$ and let $c = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$. Since $\psi$ is c.n.d. we have that $c^* A c_1 = \psi(x, x) + \psi(y, y) - \psi(x, y) - \psi(y, x) \leq 0$.

Since $\psi$ is Hermitian by definition, i.e., $\psi(y, x) = \overline{\psi(x, y)}$, the result follows.

**Proposition 2.11.** For a p.d. kernel $\varphi: X \times X \to \mathbb{C}$, $|\varphi(x, y)|^2 \leq \varphi(x, x) \varphi(y, y)$ for all $x, y \in X$.
Proof. For \( x, y \in X \) define \( A = \begin{pmatrix} a & b \\ b & d \end{pmatrix} \) where \( a = \varphi(x, x) \), \( d = \varphi(y, y) \), and \( b = \varphi(y, x) \). Let us consider a vector \( c = \begin{pmatrix} w \\ z \end{pmatrix} \). We have that
\[
(w, z) \begin{pmatrix} a & b \\ b & d \end{pmatrix} \begin{pmatrix} w \\ z \end{pmatrix} = a |w|^2 + 2 \text{Re} (bz \overline{w}) + d |z|^2.
\]
When \( a \neq 0 \), we can rewrite the RHS as
\[
a \left| w + \frac{b}{a} z \right|^2 + \frac{|z|^2}{a} (ad - |b|^2).
\]
Since this quantity should be \( \geq 0 \) for all choices of \( w \) and \( z \), we can deduce that \( a > 0 \) since if this is not so we can choose \( z = 0 \) to contradict the fact that \( \varphi \) is p.d. By a similar argument when \( d \neq 0 \) we can rewrite the quantity in (6) with the roles of \( a \) and \( d \) reversed to deduce that \( d \geq 0 \).

Examining (6) we note that if we choose \( w = -bz/a \) the first term becomes 0, so if \( \varphi \) is p.d. it is necessary that \( ad - |b|^2 \geq 0 \) which, taken with the fact that \( a, d \geq 0 \), proves the result.

2.2.3 Creating kernels from univariate functions

A univariate complex-valued function can be used to create both a p.d. kernel and a c.n.d. kernel. The following proposition shows how.

**Proposition 2.12.** Given a function \( f : X \to \mathbb{C} \), \( \varphi(x, y) = f(x) \overline{f(y)} \) is a p.d. kernel and \( \psi(x, y) = f(x) + \overline{f(y)} \) is a c.n.d. kernel.

**Proof.** Consider first the p.d. case. For some choice of \( n \in \mathbb{N} \), \( \{x_1, \ldots, x_n\} \subseteq X \) and \( \{c_1, \ldots, c_n\} \subseteq \mathbb{C} \), we have
\[
\sum_{j=1}^{n} \sum_{k=1}^{n} c_j \overline{c}_k f(j) \overline{f(k)} = \sum_{j=1}^{n} c_j f(j) \sum_{k=1}^{n} \overline{c}_k f(k) = \left| \sum_{j=1}^{n} c_j f(j) \right|^2 \geq 0.
\]
For the c.n.d. case we assume \( \sum_{i=1}^{n} c_i = 0 \). So,
\[
\sum_{j=1}^{n} \sum_{k=1}^{n} c_j \overline{c}_k (f(j) + \overline{f(k)}) = \sum_{j=1}^{n} c_j f(j) \sum_{k=1}^{n} \overline{c}_k + \sum_{k=1}^{n} \overline{c}_k f(k) \sum_{j=1}^{n} c_j = 0 \]

**Corollary 2.13.** A constant function \( \varphi(x, x) = c \) is p.d. if and only if \( c \geq 0 \) and c.n.d. if and only if \( c \in \mathbb{R} \).
Proof. Clearly if $c < 0$ then the function cannot be p.d. But if $c \geq 0$ then we can choose $f(x) = \sqrt{c}$ and apply Prop. 2.12 to prove that $\varphi(x, y) = f(x)\overline{f(y)}$ is p.d.

For the c.n.d. case, note that a constant function cannot have a non-zero imaginary part since that would violate the condition that the function is Hermitian. For any $c \in \mathbb{R}$, if we set $f(x) = s$ for any $s \in \mathbb{C}$ s.t. $\text{Re } s = c/2$, then, by Prop. 2.12, $f(x) + \overline{f(y)} = c$ is c.n.d. \qed

2.2.4 Combining kernels

We now discuss the properties of kernels created by combining other kernels. First we define some terms.

Definition 2.14 (Convex cone). Suppose $X$ is a vector space with associated scalar field $\mathbb{F} \in \{\mathbb{C}, \mathbb{R}\}$. A set $S \subseteq X$ is called a convex cone if $\forall x, y \in S$ and $\alpha, \beta \geq 0$, $\alpha x + \beta y \in S$.

Definition 2.15 (Pointwise convergence). If $\mathcal{F}$ is the set of functions from a set $X$ to $\mathbb{C}$ and $\{f_n\}_{n \geq 0}$ is a sequence of functions from $\mathcal{F}$, we say $\{f_n\}_{n \geq 0}$ converges pointwise to $f \in \mathcal{F}$ if for all $x \in X$

$$\lim_{n \to \infty} f_n(x) = f(x).$$

Definition 2.16 (Closure under pointwise convergence). Let $\mathcal{F} \subseteq \mathcal{F}$ be a set of functions from $X$ to $\mathbb{C}$. Then we say $\mathcal{F}$ is closed under the topology of pointwise convergence if for every sequence $\{f_n\}_{n \geq 0}$ in $\mathcal{F}$ converges pointwise to a function $f \in \mathcal{F}$.

We illustrate the notion of closure under pointwise convergence by an example where it does not hold.

Example 2.17. Let $\mathcal{F}$ be the set of functions from $\mathbb{R}$ to $\mathbb{R}$ and let $\mathcal{L}$ be the set of functions from $[0, 1]$ to $[0, 1)$. Consider the sequence $\{f_n\}_{n \geq 0}$ such that

$$f_i(x) = x^2 - \frac{1}{i}.$$ 

Clearly this sequence converges pointwise to $f(x) = x^2$. But since $f(1) =, f \notin \mathcal{L}$ although every function in the sequence is in $\mathcal{L}$. Hence, $\mathcal{L}$ is not closed under the topology of pointwise convergence.

Definition 2.18 (Gram matrix). Suppose $X$ is a non-empty set and $\varphi : X \times X \to \mathbb{C}$ is a kernel defined on $X$. Given $V = \{v_1, v_2, \ldots, v_n\} \subseteq X$ we will refer matrix $G$ with entries $G_{ij} = \varphi(v_i, v_j)$ as the Gram matrix of $V$ w.r.t $\varphi$.

Note that this is a somewhat more general use of the term Gram matrix than is usually encountered in the linear algebra literature.

We now proceed to the properties of combinations of kernels.

Proposition 2.19 (Convex combination). P.d. (c.n.d) kernels form a convex cone that is closed in the topology of pointwise convergence.
Proof. If \( \varphi_1 \) and \( \varphi_2 \) are p.d. kernels defined on \( X \times X \) and we consider \( \{x_1, \ldots, x_n\} \subseteq X \) and \( \{c_1, \ldots, c_n\} \subseteq \mathbb{C} \) for some \( n \in \mathbb{N} \), clearly if \( \alpha, \beta \geq 0 \) then
\[
\sum_{j,k} c_j \overline{c}_k \left( \alpha \varphi_1(x_j, x_k) + \beta \varphi_2(x_j, x_k) \right) = \alpha \sum_{j,k} c_j \overline{c}_k \varphi_1(x_j x_k) + \beta \sum_{j,k} c_j \overline{c}_k \varphi_2(x_j x_k) \geq 0.
\]
A similar argument shows that c.n.d. kernels form a convex cone.

Now consider a sequence of p.d. kernels \( \{\varphi_n\}_{n \geq 0} \). Choose \( \{c_1, \ldots, c_n\} \subseteq \mathbb{C} \) and \( \{x_1, \ldots, x_n\} \subseteq X \) for some \( n \in \mathbb{N} \) and consider
\[
\lim_{n \to \infty} a_n = \sum_{j,k} c_j \overline{c}_k \varphi_n(x_j, x_k).
\]
Since each \( a_n \geq 0 \), this limit, whenever it exists, will also be non-negative. Hence we have shown that the set of p.d. kernels is closed under pointwise convergence. A similar argument can be made for the set of c.n.d. kernels. □

Remark. Closure under pointwise convergence does not hold for strictly p.d. (c.n.d.) kernels. For the p.d. case we can see this above since although each \( a_n \) above may be strictly positive, the limit could be 0.

Theorem 2.20 (Schur’s theorem for products of p.d. kernels). If \( \varphi_1, \varphi_2 : X \times X \to \mathbb{C} \) are p.d. kernels then \( \varphi_1 \varphi_2 : X \times X \to \mathbb{C} \) is also p.d.

Proof. All we need to show is that if \( A = (a_{jk}) \) and \( B = (b_{jk}) \) are p.d. then \( C = (a_{jk}b_{jk}) \) is p.d. By the definition of p.d. kernels and by Proposition 2.7 a Gram matrix derived from a p.d. kernel is Hermitian and positive definite. So if we can prove that \( C \) is p.d. then we can apply this to the Gram matrices of a set \( \{x_1, \ldots, x_n\} \subseteq X \) w.r.t. \( \varphi_1 \) and \( \varphi_2 \) for some \( n \in \mathbb{N} \). This will prove that the Gram matrix of the set w.r.t. \( \varphi_1 \varphi_2 \) is also p.d. and we are done.

The Spectral theorem for hermitian matrices says that \( A \) is diagonalizable. Additionally, it has non negative eigenvalues because it is derived from a p.d. kernel, i.e., there is a unitary matrix \( Q \) and a diagonal matrix \( \Lambda \) with non-negative entries such that
\[
A = Q^* \Lambda Q = Q^* \Lambda^{1/2} \Lambda^{1/2} Q = (\Lambda^{1/2} Q)^* (\Lambda^{1/2} Q)
\]
Therefore, we have \( n \) functions \( f_1, \ldots, f_n : \{1, \ldots, n\} \to \mathbb{C} \) such that
\[
a_{jk} = \sum_{p=1}^{n} f_p(j) \overline{f}_p(k)
\]
Therefore, for any \( \{c_1, \ldots, c_n\} \subseteq \mathbb{C} \),
\[
\sum_{j,k} c_j \overline{c}_k a_{jk} b_{jk} = \sum_{p=1}^{n} \sum_{j,k=1}^{n} c_j f_p(j) \overline{c}_k f_p(k) b_{jk}
\]
Let us analyse the inner summation on the RHS for any value of \( p \). Put \( \tilde{c}_j = c_j f_p(j) \) and \( \tilde{c}_k = c_k f_p(k) \), then the inner summation simplifies to \( \sum_{j,k} \tilde{c}_j \overline{\tilde{c}}_k b_{jk} \) which is non-negative because \( B \) is p.d. Summing up over non-negative terms for all \( p \in \{1, \ldots, n\} \), we have
\[
\sum_{j,k} c_j \overline{c}_k a_{jk} b_{jk} \geq 0
\]
Hence, \( C \) is p.d. □
Let us consider first a simple application of Theorem 2.20.

**Proposition 2.21.** If \( \varphi : X \times X \to \mathbb{C} \) is a p.d. kernel show that \( \overline{\varphi} \), \( \text{Re} \, \varphi \), and \( |\varphi|^2 \) are also p.d.

**Proof.** Since \( \varphi \) is p.d., it follows that for \( n \in \mathbb{N} \), \( \{x_1, \ldots, x_n\} \subseteq X \) and \( \{c_1, \ldots, c_n\} \subseteq \mathbb{C} \),

\[
\sum_{j,k} c_j \overline{c}_k \varphi(x_j, x_k) \geq 0.
\]

Taking the complex conjugate of this, we have

\[
0 \leq \sum_{j,k} c_j \overline{c}_k \varphi(x_j, x_k) = \sum_{j,k} \overline{c}_j c_k \overline{\varphi}(x_j, x_k),
\]

which proves that \( \overline{\varphi} \) is p.d. Now, from Proposition 2.19 it follows that \( \text{Re} \, \varphi = (\varphi + \overline{\varphi})/2 \) is also p.d. From Theorem 2.20 it follows that \( |\varphi|^2 \) is p.d. \( \square \)

**Remark.** In general \( |\varphi| \) is not guaranteed to be p.d.

Theorem 2.20 has some important consequences, two of which we discuss next.

**Corollary 2.22** (Tensor products of p.d. kernels are p.d.). Let \( \varphi_1 : X \times X \to \mathbb{C} \) and \( \varphi_2 : Y \times Y \to \mathbb{C} \) be p.d. kernels. Then their tensor product \( \varphi_1 \otimes \varphi_2 : (X \times Y) \times (X \times Y) \to \mathbb{C} \) defined by \( \varphi_1 \otimes \varphi_2(x_1, y_1, x_2, y_2) = \varphi_1(x_1, x_2) \varphi_2(y_1, y_2) \) is also p.d.

**Proof.** Since \( \varphi_1 \) and \( \varphi_2 \) are defined on different spaces we cannot directly apply Theorem 2.20, so we define \( \tilde{\varphi}_1 \) and \( \tilde{\varphi}_2 \) that transfer them over to the product space, i.e., \( \tilde{\varphi}_1, \tilde{\varphi}_2 : (X \times Y) \times (X \times Y) \to \mathbb{C} \). We define these as follows:

\[
\tilde{\varphi}_1(x_1, y_1, x_2, y_2) = \varphi_1(x_1, x_2)
\]

\[
\tilde{\varphi}_2(x_1, y_1, x_2, y_2) = \varphi_2(y_1, y_2)
\]

We know that \( \tilde{\varphi}_1 \) and \( \tilde{\varphi}_2 \) are p.d. kernels because \( \varphi_1 \) and \( \varphi_2 \) are p.d. By Theorem 2.20, their product is also p.d., and thus the tensor product, \( \varphi_1 \otimes \varphi_2 \) is p.d. \( \square \)

**Corollary 2.23.** Let \( \varphi : X \times X \to \mathbb{C} \) be p.d. such that \( |\varphi(x, y)| < \rho \), \( \forall (x, y) \in X \times X \). Then if \( f(z) = \sum_{n=0}^{\infty} a_n z^n \) is holomorphic in \( \{z \in \mathbb{C} : |z| < \rho\} \) and \( a_n \geq 0 \) for all \( n \geq 0 \), the composed kernel \( f \circ \varphi \) is again p.d. In particular, if \( \varphi \) is p.d. then so is \( \exp(\varphi) \).

**Proof.** Define \( f_n = \sum_{i=0}^{\infty} a_i \varphi^n \). By Theorem 2.20, \( \varphi^n \) is pd for all \( n \geq 0 \). Since \( a_i \geq 0 \) for all \( i \geq 0 \), \( f_n \) is a convex combination of p.d. kernels and is therefore, by Proposition 2.19, itself p.d. \( f(z) \) is holomorphic in \( \{z \in \mathbb{C} : |z| < \rho\} \). Therefore, if \( |\varphi(x, y)| < \rho \) for all \( x, y \in X \), \( f_n \) will converge pointwise to \( \sum_{n=0}^{\infty} a_n \varphi^n \) which is guaranteed to be a p.d. kernel by Proposition 2.19. In particular \( \exp(\varphi) \) is p.d. for any p.d. kernel \( \varphi \) because the radius of convergence of \( \sum_{n=0}^{\infty} z^n/n! \) is \( \rho = \infty \). \( \square \)

### 2.2.5 A Sylvester-like criterion for kernels

For strictly positive definite matrices, the following result relates the positive definite property to the value of the determinants of their principal submatrices.

**Theorem 2.24** (Sylvester’s criterion for strictly p.d. matrices). An \( n \times n \) Hermitian matrix \( A = (a_{jk} : j, k \leq n) \) is strictly p.d. iff \( \det(a_{jk} : j, k \leq p) > 0 \) for \( 1 \leq p \leq n \).
Proof. First, let’s assume that $A$ is strictly p.d. Therefore, by the Spectral Theorem for p.d. Hermitian matrices, $A$ can be diagonalized as $Q^*AQ$ where $Q$ is a unitary matrix and $A$ is a diagonal matrix. Since all the entries of $A$ are non-negative (in fact strictly positive), we can write $A = B \ast B$ where $B = QA^{1/2}$. Therefore, $\det(A) = |\det(B)|^2 \geq 0$. Further $\det(A) \neq 0$, since other $\det(A - 0.1) = 0$ i.e. 0 is an eigen value for $A$, which contradicts the assumption that $A$ is strictly p.d. Hence, $\det(A) > 0$.

Since $A$ being strictly p.d. implies that $(a_{jk} : j, k \leq p)$ is strictly p.d. for every $p$ s.t. $1 \leq p \leq n$, we can use a similar argument to prove that $\det(a_{jk} : j, k \leq p) > 0$ for $1 \leq p \leq n$.

Now, we assume that $\det(a_{jk} : j, k \leq p) > 0$ for $1 \leq p \leq n$ we will show that $A$ is strictly p.d. The proof is by induction. Clearly this is true for $p = 1$ which is the base case for the induction. As our induction hypothesis we assume that $\det(a_{jk} : j, k \leq n - 1) > 0$ implies that $(a_{jk} : j, k \leq n - 1)$ is strictly p.d.

Now, let us consider the matrix $A = (a_{jk} : j, k \leq n)$ and transform it to $A' = (a'_{jk} : j, k \leq n - 1)$ such that

$$a'_{jk} = a_{jk} - \left(\frac{a_{1k}}{a_{11}}\right)a_{j1},$$

i.e., we multiply the first column by $a_{1k}/a_{11}$ and subtract it from the $k$th column, for $2 \leq k \leq n$. This gives us

$$A' = \begin{bmatrix} a_{11} & 0 & \ldots & 0 \\ a_{21} & a_{22} & \ldots & a'_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a'_{n2} & \ldots & a'_{nn} \end{bmatrix}$$

Since $A'$ is derived from $A$ through elementary column transformations, $\det(a_{jk} : j, k \leq p) = \det(a'_{jk} : j, k \leq p)$, $2 \leq p \leq n$. Further, $\det(a'_{jk} : j, k \leq p) = \det(b_{jk} : j, k \leq p)$ where $B = (b_{jk} : j, k \leq n)$ is defined as

$$B = \begin{bmatrix} a_{11} & 0 & \ldots & 0 \\ 0 & a'_{22} & \ldots & a'_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a'_{n2} & \ldots & a'_{nn} \end{bmatrix}$$

Let $C = (c_{jk} : j, k \leq n - 1)$ be defined as follows.

$$C = \begin{bmatrix} a'_{22} & \ldots & a'_{2n} \\ \vdots & \ddots & \vdots \\ a'_{n2} & \ldots & a'_{nn} \end{bmatrix}$$

Clearly $\det(c_{jk} : j, k \leq p) = 1/a_{11} \cdot \det(b_{jk} : j, k \leq p + 1)$, $1 \leq p \leq n - 1$. Since $a_{11} > 0$ and $\det(b_{jk} : j, k \leq p + 1) = \det(a_{jk} : j, k \leq p + 1) > 0$ for all $1 \leq p \leq n - 1$ we know that $\det(c_{jk} : j, k \leq p) > 0$ for $1 \leq p \leq n - 1$, i.e., we have shown that all the principal submatrices of $C$ have strictly positive. We are now in a position to apply the induction hypothesis on $C$ to establish that it is strictly p.d. provided it is Hermitian. Consider $a'_{jk}$ such that $j, k \in \{2, \ldots, n\}$. We have,

$$a'_{jk} = a_{jk} - \left(\frac{a_{1k}}{a_{11}}\right)a_{j1} = a_{jk} - \left(\frac{a_{1k}}{a_{11}}\right)a_{j1} = a_{k1} - \left(\frac{a_{k1}}{a_{11}}\right)a_{1j} = a'_{kj}.$$
Hence $C$ is Hermitian and so, by the induction hypothesis, it is strictly p.d.

With this in hand we will attempt to show that $A$ itself if strictly p.d. Consider an arbitrary non-zero vector $c = (c_1, c_2, \ldots, c_n) \in \mathbb{C}^n$

\[
\sum_{j,k=1}^{n} c_j \overline{c}_k a_{jk} = \sum_{j,k=2}^{n} c_j \overline{c}_k \left( a'_{jk} + \frac{a_{1k}a_{j1}}{a_{11}} \right) + \sum_{j=2}^{n} c_j \overline{c}_1 a_{j1} + \sum_{k=2}^{n} c_1 \overline{c}_k a_{1k} + |c_1|^2 a_{11} 
\]

\[
= \sum_{j,k=2}^{n} c_j \overline{c}_k a'_{jk} + \frac{1}{a_{11}} \left( |\sum_{j=2}^{n} c_j a_{j1}|^2 + 2a_{11} \text{Re} \left( c_1 \sum_{k=2}^{n} \overline{c}_k a_{1k} \right) \right) + (|c_1|a_{11})^2 
\]

Now if there is $i \in \{2, \ldots, n\}$ such that $c_i \neq 0$ then the first term is positive and second term is non-negative. On the other hand if $c_i = 0$ for all $i \in \{2, \ldots, n\}$ then the first term is zero and the second term is positive since $c_1 \neq 0$.

\[\square\]

**Remark:** Sylvester’s criterion does not hold for p.d. matrices that are not strictly p.d. For example, consider

\[
A = \begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix}
\]

Here, $A$ is a Hermitian $2 \times 2$ matrix such that the determinants of both principal submatrices are 0 but $A$ is not p.d.

We now turn to establishing a Sylvester-like criterion for p.d. kernel.

**Theorem 2.25.** If $\varphi : X \times X \to \mathbb{C}$ is a kernel then $\varphi$ is p.d. iff

\[
det(\varphi(x_j, x_k) : j, k \leq n) \geq 0
\]

for all $n \in \mathbb{N}$ and $x_1, \ldots, x_n \in X$

**Proof.** If $\varphi$ is p.d. then, by Proposition 2.7, the $n \times n$ matrix given by $A = (\varphi(x_j, x_k) : j, k \leq n)$ is a p.d. Hermitian matrix and we can apply an argument similar to the one made in the proof of Theorem 2.24 to prove that all its principal submatrices have non-negative determinant.

So, let us turn to the other direction, i.e., let $\varphi : X \times X \to \mathbb{C}$ be a kernel such that for any $n \in \mathbb{N}$ and $x_1, \ldots, x_n \in X, \det(\varphi(x_j, x_k) : j, k \leq n) \geq 0$. Define $\varphi_\varepsilon = \varphi + \varepsilon I_\Delta$ where $\varepsilon > 0$ and $\Delta$ is the diagonal in $X \times X$, i.e., $\varphi_\varepsilon$ adds a small positive constant to $\varphi(x, x)$ for each $x \in X$.

Computing the determinant of the matrix created by modified version of $\varphi$, and assuming that $x_1, \ldots, x_n$ are distinct, we see that

\[
det(\varphi_\varepsilon(x_j, x_k) : j, k \leq n) = \sum_{p=0}^{n} d_p \varepsilon^p,
\]

where $d_n = 1$ and

\[
d_p = \sum_{A \subseteq \{1, \ldots, n\}} \det(\varphi(x_j, x_k) : j, k \in A)_{|A|=(n-p)}
\]
for $0 \leq p \leq (n-1)$. Since each of the terms in the sum defining $d_p$ is non-negative, therefore $d_p$ is non-negative for $0 \leq p \leq (n-1)$. This implies that $\det(\varphi_e(x_j, x_k) : j, k \leq n) \geq \varepsilon^n > 0$. Since the same argument can be used to establish that $\det(\varphi_e(x_j, x_k) : j, k \leq p) > 0$ for $1 \leq p \leq n-1$, we get by Theorem 2.24 that $\varphi_e$ is a p.d. kernel.

Clearly $\varphi$ is the pointwise limit $\varepsilon \downarrow 0$, and since the convex cone of p.d. kernels is closed under the topology of pointwise convergence (Proposition 2.19) we see that $\varphi$ is p.d. $\square$

2.3 Relating p.d. and c.n.d. kernels

We now discuss the relationships between p.d. and c.n.d. kernels. Clearly if $\varphi$ is p.d. then $-\varphi$ is c.n.d. Although the converse need not be true, the following lemma gives a useful relationship in the opposite direction.

**Lemma 2.26.** Let $X$ be a non-empty set, $x_0 \in X$ and $\psi : X \times X \to \mathbb{C}$ be a Hermitian kernel. Let $\varphi : X \times X \to \mathbb{C}$ be such that $\varphi(x, y) = \psi(x, x_0) + \overline{\psi(y, x_0)} - \psi(x, y) - \psi(x_0, y_0)$. Then $\varphi$ is p.d. iff $\psi$ is c.n.d. Further, if $\psi(x_0, x_0) \geq 0$ and $\varphi_0(x, y) = \psi(x, x_0) + \overline{\psi(y, x_0)} - \psi(x, y)$ then $\varphi_0$ is p.d. iff $\psi$ is c.n.d.

**Proof.** (i) Let $\varphi$ be a p.d. kernel. For $n \geq 2$ consider $c_1, \ldots, c_n \in \mathbb{C}$ such that $\sum_{i=1}^{n} c_i = 0$ and $x_1, x_2, \ldots, x_n \in X$. Since $\varphi$ is p.d. we have that

$$
\sum_{j,k=1}^{n} c_j \overline{c_k} (\psi(x_j, x_0) + \overline{\psi(x_k, x_0)} - \psi(x_j, x_k) - \psi(x_0, y_0)) \geq 0.
$$

Since $\sum_{i=1}^{n} c_i = 0$, three of the terms on the LHS become 0 and we are left with

$$
- \sum_{j,k=1}^{n} \psi(x_j, x_k) \geq 0,
$$

which proves that $\psi$ is c.n.d. By a similar argument we can show that $\psi$ is c.n.d. if $\varphi_0$ is p.d.

(ii) Let $\psi$ be a c.n.d. kernel. For some $n \in \mathbb{N}$ consider $c_1, c_2, \ldots, c_n \in \mathbb{C}$ and $x_1, x_2, \ldots, x_n \in X$. Choose $x_0 \in X$ and set $c_0 = -\sum_{i=1}^{n} c_i$. Since $\psi$ is c.n.d. we know that

$$
\sum_{j,k=0}^{n} c_j \overline{c_k} \psi(x_j, x_k) \leq 0.
$$

We split the sum on the LHS, separating out the terms involving $c_0$ and $x_0$. This gives us that

$$
\sum_{j,k=1}^{n} c_j \overline{c_k} \psi(x_j, x_k) + \sum_{j=1}^{n} c_j \overline{c_0} \psi(x_j, x_0) + \sum_{k=1}^{n} c_0 \overline{c_k} \psi(x_0, x_k) + |c_0|^2 \psi(x_0, x_0) \leq 0.
$$

Replacing $c_0$ with $-\sum_{i=1}^{n} c_i$ on the LHS we get

$$
\sum_{j,k=1}^{n} c_j \overline{c_k} (\psi(x_j, x_k) - \psi(x_j, x_0) - \psi(x_0, x_k) + \psi(x_0, y_0)) \leq 0,
$$

(7)
i.e.,
\[- \sum_{j,k=1}^{n} c_j c_k \varphi(x_j, x_k) \leq 0.\]
So, \( \varphi \) is p.d. Note that if \( \psi(x_0, y_0) \geq 0 \) then by (7) we have,
\[
\sum_{j,k=1}^{n} c_j c_k (\psi(x_j, x_k) - \psi(x_j, x_0) - \psi(x_0, x_k)) \leq 0,
\]
i.e., \( \varphi_0 \) is p.d.

The form of Lemma 2.26 doesn’t appear so easy to work with but it has some nice consequences. We present one such here, which is attributed ot Schönberg in [1].

**Theorem 2.27.** If \( \psi : X \times X \to \mathbb{C} \) is a kernel on a non-empty set \( X \) then \( \psi \) is c.n.d. iff \( \exp(-t\psi) \) is a p.d. kernel \( t > 0 \).

**Proof.** (i) First assume that for \( t > 0 \), \( \exp(-t\psi) \) is a p.d. kernel. Therefore, \( (1 - \exp(-t\psi)) \) is a c.n.d. kernel and so is \( \left(1 - \frac{\exp(-t\psi)}{t}\right) \). Since
\[
\psi = \lim_{t \to 0} \frac{1}{t} (1 - \exp(-t\psi)),
\]
by the fact that the convex cone of c.n.d. kernels is closed under pointwise convergence (Proposition 2.19) we infer that \( \psi \) is a c.n.d. kernel.

(ii) Let us assume that \( \psi \) is a c.n.d. kernel. It is sufficient to show that \( \exp(\psi) \) is p.d. as we can replace \( \psi \) by \( t\psi \) because if \( \psi \) is c.n.d. then \( t\psi \) is c.n.d. whenever \( t > 0 \). Let us choose \( x_0 \in X \) and \( \varphi \) as in Lemma 2.26 to get,
\[-\psi(x, y) = \varphi(x, y) - \psi(x, x_0) - \overline{\psi(y, x_0)} + \psi(x_0, x_0).\]
Taking exponents on both sides we have
\[
\exp(-\psi(x, y)) = \exp(\varphi(x, y)) \cdot \exp(\psi(x, x_0)) \cdot \exp(\overline{\psi(y, x_0)}) \cdot \exp(\psi(x_0, x_0))
\]
Examining the RHS we see that it is a product of four terms. The first of these is the exponent of a p.d. kernel and therefore is p.d. by Corollary 2.23. The product \( \exp(\psi(x, x_0)) \cdot \exp(\overline{\psi(y, x_0)}) \) is a p.d. kernel by Proposition 2.12. And \( \exp(\psi(x_0, x_0)) \) is a p.d. kernel trivially as it is a positive constant. By Schur’s theorem (Theorem 2.20), the product of these three p.d. kernels is also a p.d. kernel.

**Remarks.**

1. Applying Theorem 2.27 to Example 2.9 extended to \( d \)-dimensions tells us that the Gaussian Radial Basis Function, \( \varphi(x, y) = e^{-c\|x-y\|^2} \), a very popular kernel used widely in Machine Learning applications, is indeed p.d.

2. It is also possible to show that \( \psi \) is c.n.d. iff \( 1/(t + \psi) \) is p.d. for all \( t > 0 \) but we omit the proof here, refering the reader to [1] for this and other interesting relationships between p.d. and c.n.d. kernels.
3 Hilbert spaces and kernels

In this section we will establish the relationship the role of Hilbert spaces in the study of p.d. kernels. We will discuss Mercer’s theorem and also present the Reproducing Kernel Hilbert Space associated with a p.d. kernel.

3.1 Inner products and their associated norms

To make these notes self-contained we first present some basic definitions and results to develop the definition of a Hilbert space. The primary source of the material in this section is the chapter by Heil [3].

Definition 3.1 (Semi-Inner Product). Let $X$ be a vector space over $\mathbb{C}$. A function $\langle \cdot, \cdot \rangle : X \times X \rightarrow \mathbb{C}$ is called a semi-inner product if:

1. $\langle x, x \rangle \geq 0$ for all $x \in X$,
2. (Hermitian) $\langle x, y \rangle = \overline{\langle y, x \rangle}$ for all $x, y \in X$ and
3. (Linearity in the first variable) $\langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle$, for all $x, y, z \in X$ and $\alpha, \beta \in \mathbb{C}$.

Remarks.

- (Anti-linearity in the second variable). From the Hermitian property and linearity in the first variable it is easy to see that for all $x, y, z \in X$ and $\alpha, \beta \in \mathbb{C}$
  \[ \langle x, \alpha y + \beta z \rangle = \overline{\alpha} \langle x, y \rangle + \overline{\beta} \langle x, z \rangle. \]

- It is also easy to deduce that $\langle x, 0 \rangle = 0 = \langle 0, x \rangle$ for all $x \in X$. Further, $\langle 0, 0 \rangle = 0$.

Semi inner-products have a property of interest to us: they are p.d. kernels.

Proposition 3.2. A semi-inner product $\langle \cdot, \cdot \rangle : X \times X \rightarrow \mathbb{C}$ is a p.d. kernel.

Proof. Let $n \in N$, $\{x_1, \ldots, x_n\} \subseteq X$ and $\{c_1, \ldots, c_n\} \subseteq \mathbb{C}$. Since a semi-inner product is linear in the first variable and anti-linear in the second variable we get,
\[
\sum_{j,k=1}^{n} c_j \overline{c_k} \langle x_j, x_k \rangle = \sum_{j,k=1}^{n} \langle c_j x_j, c_k x_k \rangle = \left\langle \sum_{j=1}^{n} c_j x_j, \sum_{j=1}^{n} c_j x_j \right\rangle \geq 0.
\]

Definition 3.3 (Inner Product). If a semi-inner product space $\langle \cdot, \cdot \rangle$ satisfies
\[
\langle x, x \rangle = 0 \implies x = 0
\]
then it is called an inner inner product.
Definition 3.4 (Inner product space). A vector space $X$ with inner product $\langle \cdot, \cdot \rangle$ defined on it is called an inner product space or a pre-Hilbert space.

We now discuss some examples of inner products.

Example 3.5. Prove that the dot product $\mathbf{x} \cdot \mathbf{y} = x_1\bar{y}_1 + \cdots + x_n\bar{y}_n$ is an inner product on $\mathbb{C}^n$.

Proof. First we see that $\mathbf{x} \cdot \mathbf{x} = x_1\bar{x}_1 + \cdots + x_n\bar{x}_n = \sum |x_k|^2 \geq 0$.

This further implies $\mathbf{x} \cdot \mathbf{x}$ is 0 only if $\mathbf{x} = \mathbf{0}$.

For the Hermitian property note that $\mathbf{y} \cdot \mathbf{x} = \sum_{i=1}^{n} y_i x_i = \sum_{i=1}^{n} \mathbf{x}_i \bar{y}_i = \mathbf{x} \cdot \mathbf{y}$.

Linearity in the first variable is easy to see. $\square$

Example 3.6. If $w_1, \ldots, w_n \geq 0$ are fixed scalars, then the weighted dot product $\langle \mathbf{x}, \mathbf{y} \rangle = x_1\bar{y}_1 w_1 + \cdots + x_n\bar{y}_n w_n$ is a semi-inner product on $\mathbb{C}^n$. It is an inner product if $w_i > 0$ for each $i$.

This example can be established just like Example 3.5, except here we note that even if there is a $w_i = 0$ then $\langle \mathbf{x}, \mathbf{x} \rangle$ is 0 for any $\mathbf{x}$ which is non-zero only in the $i$th coordinate. Hence $\langle \cdot, \cdot \rangle$ is only a semi-inner product unless $w_i > 0$ for all $i$.

Example 3.7. Let $A$ be a $n \times n$ p.d. matrix. $\langle \mathbf{x}, \mathbf{y} \rangle = A \mathbf{x} \cdot \mathbf{y}$ is a semi-inner product on $\mathbb{C}^n$ where $\cdot$ is as defined in Example 3.5. Moreover, $\langle \cdot, \cdot \rangle$ is an inner product if $A$ is strictly p.d.

Proof. Clearly $\langle \mathbf{x}, \mathbf{x} \rangle$ is non-negative if $A$ is p.d. and positive for all $\mathbf{x} \neq \mathbf{0}$ if $A$ is strictly p.d. Linearity in the first variable follows easily since $A$ is a linear transformation. It remains to show the Hermitian property.

$$A \mathbf{y} \cdot \mathbf{x} = \sum_{i=1}^{n} \left( \sum_{j=1}^{n} A_{ij} y_j \right) x_i = \sum_{i,j=1}^{n} A_{ij} y_j x_i.$$  

Since all p.d. matrices are Hermitian (by arguments similar to that in the proof of Proposition 2.7), the last term above can be rewritten as $\sum_{i,j=1}^{n} A_{ij} \bar{y}_j x_i = \sum_{j=1}^{n} \left( \sum_{i=1}^{n} A_{ji} x_i \right) y_j = A \mathbf{x} \cdot \mathbf{y}$. $\square$

The following example shows that all semi-inner products on $\mathbb{C}^n$ can be represented by a p.d. matrix.
Example 3.8. For an arbitrary semi-inner product $\langle \cdot, \cdot \rangle$ on $\mathbb{C}^n$, there is a p.d matrix $A$ such that $\langle x, y \rangle = Ax \cdot y$.

Proof. Let $e^1, \ldots, e^n$, where $e^i$ is the vector that is 1 in the $i$th coordinate and 0 elsewhere, be the standard basis of $\mathbb{C}$. We hypothesise that the $A$ is the Gram matrix of this basis w.r.t the given inner product, i.e., $A_{ij} = \langle e^i, e^j \rangle$, $1 \leq i, j \leq n$. To verify this note, using the properties that any semi-inner product is linear in the first variable and antilinear in the second variable, that

$$\langle x, y \rangle = \left\langle \sum_{i=1}^n x_i e^i, \sum_{i=1}^n y_i e^i \right\rangle = \sum_{i,j=1}^n x_i \overline{y_j} \langle e^i, e^j \rangle.$$ 

It is easy to see that $A$ is p.d. since $x^T Ax = \langle x, x \rangle$ which is non-negative since $\langle \cdot, \cdot \rangle$ is an semi-inner product.

We now discuss some basic properties of inner products. With every semi-inner product $\langle \cdot, \cdot \rangle$ we will associate a univariate function:

$$\| x \| = \langle x, x \rangle^{1/2}.$$ 

Lemma 3.9 (Polar Identity). If $\langle \cdot, \cdot \rangle$ is a semi-inner product on a vector space $X$ then for all $x, y \in X$

$$\| x + y \|^2 = \| x \|^2 + \| y \|^2 + 2 \Re \langle x, y \rangle.$$ 

Proof. $\| x + y \|^2 = \langle x + y, x + y \rangle = \| x \|^2 + \| y \|^2 + \langle x, y \rangle + \langle y, x \rangle$ The result follows by using the Hermitian property, $\langle x, y \rangle = \overline{\langle y, x \rangle}$.

Theorem 3.10 (Cauchy-Bunyakovsky-Schwarz Inequality). If $\langle \cdot, \cdot \rangle$ is a semi-inner product on a vector space $X$ then for all $x, y \in X$

$$| \langle x, y \rangle | \leq \| x \| \| y \|.$$ 

Moreover, equality holds if $x = \lambda y$ for some $\lambda \in \mathbb{C}$.

Proof. If either of $x$ or $y$ are 0 the result follows trivially so we assume $x, y$ are both non-zero. Note that if $x \in X$ and $\alpha \in \mathbb{C}$ then by the linearity and anti-linearity properties of semi-inner products $\| \alpha x \| = |\alpha| \| x \|$. With this in mind, by Lemma 3.9, for any $t \in \mathbb{C}$ we get that

$$0 \leq \| x + ty \|^2 \leq \| x \|^2 + 2 \Re \langle x, ty \rangle + |t|^2 \| y \|^2$$

Setting $t = \frac{\langle x, y \rangle}{\langle y, y \rangle}$ we get

$$\| x \|^2 - 2 \Re \frac{|\langle x, y \rangle|^2}{\| y \|^2} = \| x \|^2 - 2 \frac{|\langle x, y \rangle|^2}{\| y \|^4} + \frac{|\langle x, y \rangle|^2}{\| y \|^2} = \| x \|^2 - \frac{|\langle x, y \rangle|^2}{\| y \|^2} \geq 0,$$

and the result follows. If $x = \lambda y$ for some $\lambda \in \mathbb{C}$, then

$$\langle \lambda y, \lambda y \rangle - \frac{|\langle \lambda y, y \rangle|^2}{\| y \|^2} = |\lambda|^2 \| y \|^2 - |\lambda|^2 \frac{\| y \|^4}{\| y \|^2} = 0$$

$\square$
Remark. Both the Polar identity and the Cauchy-Bunyakovsky-Schwarz Inequality don’t need \(\langle\cdot,\cdot\rangle\) to be an inner product, semi-inner product is enough.

Now, let us see some two examples of inner products in spaces which are not finite dimensional.

Example 3.11 (Square summable sequences). Let \(I\) be a countable index set (eg. set of natural numbers or integers). Let \(w : I \to [0, \infty)\). Let

\[
\ell^2_w = \ell^2_w(I) = \left\{ x = \{x_i\}_{i \in I} : \sum_{i \in I} |x_i|^2 w(i) < \infty \right\}
\]

then

\[
\langle x, y \rangle = \sum_{i \in I} x_i \bar{y}_i w(i)
\]
defines a semi-inner product on \(\ell^2_w\). If \(w(i) > 0\) for all \(i \in I\) then it is an inner product.

Proof. The non-negativity of diagonal elements is easy to prove and, like in Example 3.6 we can argue that \(\langle x, x \rangle = 0\) whenever \(x = 0\) only if every \(w_i > 0\). Linearity in the first variable is also easy to establish. But a complication arises with the Hermitian property. We know that

\[
\overline{\langle x, y \rangle} = \sum_{i \in I} \bar{x}_i y_i w(i) = \sum_{i \in I} x_i \bar{y}_i w(i),
\]

but there is no guarantee that the last sum is finite. To see that it is, let \(I = \mathbb{N}\) and consider the partial sum of the first \(n\) terms of the sum. By Example 3.6 and by the Cauchy-Bunyakovski-Schwarz inequality (Theorem 3.10) we know that

\[
\sum_{i=1}^n x_i y_i w(i) \leq \sum_{i=1}^n |x_i|^2 w_i \sum_{i=1}^n |y_i|^2 w_i.
\]

Taking limits on both sides we get that the RHS is finite since \(x, y \in \ell^2_w\), and hence the LHS converges to a finite quantity and hence is equal to \(\langle y, x \rangle\). \(\square\)

Example 3.12 (Square integrable functions). Let \((X, \Omega, \mu)\) be a measure space. Define

\[
L^2(X, \mu) = \left\{ f : X \to \mathbb{F} : \int_X |f(x)|^2 d\mu(x) < \infty \right\}
\]

where we identify functions that are equal almost everywhere, i.e.,

\[
f = g \iff \mu\{x \in X : f(x) \neq g(x)\} = 0.
\]

Then

\[
\langle f, g \rangle = \int_X f(x) \overline{g(x)} d\mu(x)
\]
defines an inner product on \(L^2(X, \mu)\).
We now introduce the notion of a norm and show that an inner product naturally induces a norm.

**Definition 3.13 (Semi Norm).** Given a vector space $X$, a function $f : X \to \mathbb{R}$ is said to be a semi norm, if for all $x, y \in X$ and $\alpha \in \mathbb{C}$

1. $f(x) \geq 0$,  
2. $f(\alpha x) = |\alpha|f(x)$, and  
3. $f(x + y) \leq f(x) + f(y)$. (Triangle Inequality)

Further, if $f(x) = 0 \implies x = 0$, then $f$ is called a norm.

**Definition 3.14 (Normed Linear Space).** Any vector space equipped with a norm is called as a normed linear space or a normed vector space.

**Proposition 3.15.** $\| \cdot \|$ is a semi-norm if $\langle \cdot, \cdot \rangle$ is a semi-inner product and a norm if $\langle \cdot, \cdot \rangle$ is an inner product.

**Proof.** The non-negativity of $\| \cdot \|$ follows easily. The second property follows from the fact that a semi-inner product is linear in the first variable and anti-linear in the second variable. We now argue for the Triangle inequality.

From the Polar identity (Lemma 3.9) we know that  
$$\|x + y\|^2 = \|x\|^2 + 2\text{Re} \langle x, y \rangle + \|y\|^2.$$  
Since $\text{Re} \alpha \leq |\alpha|$ for any $\alpha \in \mathbb{C}$, we can say that  
$$\|x + y\|^2 \leq \|x\|^2 + 2|\langle x, y \rangle| + \|y\|^2.$$  
Applying the Cauchy-Bunyakovsky-Schwarz Inequality (Theorem 3.10), we get  
$$\|x + y\|^2 \leq \|x\|^2 + 2\|x\|\|y\| + \|y\|^2 = (\|x\| + \|y\|)^2.$$  

Thus we see that an inner product induces a norm. Now we will see that just like the inner product is a p.d. kernel, the norm provides a natural c.n.d. kernel. The following proposition is a generalization of Example 2.9.

**Proposition 3.16.** $\psi(x, y) = \|x - y\|^2$ is a c.n.d. kernel if $\|x\|$ is a semi norm.

**Proof.** Given $n \geq 2$, $\{x_1, \ldots, x_n\} \subseteq X$ and $\{c_1, \ldots, c_n\} \subseteq \mathbb{C}$ such that $\sum_{i=1}^n c_i = 0$ by the Polar Identity (Lemma 3.9) we have that  
$$\sum_{j, k=1}^n c_j \overline{c_k} \|x_j - x_k\|^2 = \sum_{j, k=1}^n c_j \overline{c_k} \|x_j\|^2 - 2\sum_{j, k=1}^n c_j \overline{c_k} \text{Re} \langle x_j, x_k \rangle + \sum_{j, k=1}^n c_j \overline{c_k} \|x_k\|^2,$$  
where the first and the third term are 0 because $\sum_{i=1}^n c_i = 0$. From Proposition 3.2 we know that $\langle \cdot, \cdot \rangle$ is a p.d. kernel and so, from Proposition 2.21, we know that $\text{Re} \langle \cdot, \cdot \rangle$ is a p.d. kernel. Therefore  
$$\sum_{j, k=1}^n c_j \overline{c_k} \|x_j - x_k\|^2 = -2\sum_{j, k=1}^n c_j \overline{c_k} \text{Re} \langle x, y \rangle \leq 0$$  

$\square$
3.2 Hilbert spaces

Hilbert spaces are inner product spaces with certain convergence properties. To make the definition concrete, we first need to define notions of convergence and investigate some properties of inner product spaces under these notions. Note that Definition 3.17 and Proposition 3.18 hold for any linear space which has a norm associated with it, i.e., a more general class of spaces than inner product spaces.

**Definition 3.17** (Convergence in a normed space). Let $X$ be a normed linear space and let $\{f_n\}_{n=1}^\infty$ be a sequence of elements of $X$.

1. We say that $\{f_n\}$ converges to $f \in X$, denoted $f_n \to f$, if

$$\lim_{n \to \infty} \|f - f_n\| = 0,$$

i.e.,

$$\forall \varepsilon > 0, \exists N > 0 \text{ such that } n > N \implies \|f - f_n\| < \varepsilon.$$

2. We say that $\{f_n\}$ is Cauchy if

$$\forall \varepsilon > 0, \exists N > 0 \text{ such that } m, n > N \implies \|f_m - f_n\| < \varepsilon.$$

**Proposition 3.18** (Convergence properties of normed linear spaces). Let $X$ be a normed linear space and $x, y \in X$.

1. Reverse Triangle Inequality: $\|x - y\| \geq \|x\| - \|y\|$.

2. Continuity of the norm: $x_n \to x \implies \|x_n\| \to \|x\|$.

3. Continuity of the inner product: If $X$ is an inner product space then

$$x_n \to x, y_n \to y \implies \langle x_n, y_n \rangle \to \langle x, y \rangle.$$

4. All convergent sequences are bounded, and the limit of a convergent sequence is unique.

5. Cauchy sequences are bounded.

6. Every convergent sequence is Cauchy.

7. There exist inner product spaces for which not every Cauchy sequence is convergent.

**Proof.** 1. By the triangle inequality we have

$$\|x\| + \|y - x\| \geq \|x + (y - x)\| = \|y\|$$

and

$$\|x - y\| + \|y\| \geq \|(x - y) + y\| = \|x\|.$$ 

When we move $\|x\|$ and $\|y\|$ to the RHS in the inequalities above, we have

$$\|y - x\| \geq \|y\| - \|x\|$$

and

$$\|x - y\| \geq \|x\| - \|y\|.$$ 

We know that $\|y - x\| = \|x - y\|$, therefore, we have $\|x - y\| \geq \|\|x\| - \|y\||$. 

22
2. For some $\epsilon > 0$, if $\|x_n - x\| < \epsilon$ then by the reverse triangle inequality above, we have $\|x_n\| - \|x\| \leq \|x_n - x\| < \epsilon$.

3. We can write

$$|\langle x_n, y_n \rangle - \langle x, y \rangle| = |\langle x_n, y_n \rangle - \langle x_n, y \rangle + \langle x_n, y \rangle - \langle x, y \rangle|$$

$$\leq \|x_n\| \cdot \|y_n - y\| + \|x_n - x\| \cdot \|y\|$$

where the last step follows from the Cauchy-Bunyakovskiy-Schwarz inequality (Theorem 3.10). Since the norm is continuous, i.e., $\|x_n\| \to \|x\|$ and $\|y_n\| \to \|y\|$, therefore,

$$|\langle x_n, y_n \rangle - \langle x, y \rangle| \to 0.$$ 

4. (i) Let $f_n$ be a convergent sequence and $f_n \to f$. If we take $\epsilon = 1$, we have an $N > 0$ such that $|f_n - f| < 1$ for all $n > N$. Using the reverse triangle inequality, for $n > N$,

$$|f_n| - |f| < 1 \implies |f_n| < |f| + 1$$

Let $M = \max\{|f| + 1, |f_1|, \ldots, |f_N|\}$, we have $|f_n| \leq M$ for all $n > N$.

(ii) Uniqueness of the limit can be proved by assuming the contrary and showing a contradiction.

5. By the definition of Cauchy sequence for every $\epsilon > 0$ there is an $N > 0$ such that for $n, m > N$, $|f_n - f_m| < \epsilon$. Using the triangle inequality, we have

$$|f_n| = |f_n - f_m + f_m| \leq |f_n - f_m| + |f_m|.$$ 

Taking $\epsilon = 1$, we have that $|f_n - f_m| < 1$ for $n, m > N$. Now taking $m = N + 1$, we get $|f_n| < 1 + |f_{N+1}|$ This bounds all the terms beyond the $N$th term. Taking $M = \max\{|f_1|, \ldots, |f_N|, |f_{N+1}| + 1\}$, we have $|f_n| \leq M$ for all $n > 0$.

6. Let $f_n$ be a convergent sequence, we have $|f_n - f| < \epsilon$, for all $\epsilon > 0$. Using the triangle inequality,

$$|f_n - f_m| = |f_n - f + f - f_m| \leq |f_n - f| + |f - f_m|$$

for all $n, m > N$ for some $N > 0$. Replacing $\epsilon$ by $\epsilon/2$, we have $|f_n - f_m| < \epsilon$. Therefore, $f_n$ is a Cauchy sequence.

7. Consider $L^2([0,1]) \cap C([0,1])$, where $L^2$ is a set of continuous square integrable functions on $[0,1]$ and $C([0,1])$ is a set of continuous functions on $[0,1]$. Define a sequence

$$f_n(x) = \begin{cases} 
1, & x = 0 \\
1 - nx, & 0 \leq x \leq \frac{1}{n} \\
0, & x \geq \frac{1}{n}
\end{cases}$$
Using the inner product defined in Example 3.12, and assuming $n > m$,

$$\|f_n - f_m\| = \int_{x=0}^{1} (f_n(x) - f_m(x))^2 dx$$

$$= \int_{0}^{\frac{n-m}{m}} (n-m)^2 x^2 dx + \int_{\frac{n-m}{m}}^{1} m^2 x^2 dx + 0$$

$$= (n-m)^2 \left( \frac{1}{3n^2} \right) + \frac{m^2}{3} \left( \frac{1}{m^3} - \frac{1}{n^3} \right)$$

$$\leq \frac{1}{3m}$$

Therefore, the sequence is Cauchy. As $n \to \infty$, $f_n$ becomes discontinuous and hence \{f_n\} is not convergent.

Now we are ready to define a Hilbert space.

**Definition 3.19 (Hilbert Space).** An inner product space $H$ is called a Hilbert space if it is complete, i.e., if every Cauchy sequence is convergent, i.e., \{f_n\}_{n=1}^\infty is Cauchy in $H$ implies that there is an $f \in H$ such that $f_n \to f$.

**Remark.** A normed linear space is called a Banach space if every Cauchy sequence converges. Hence every Hilbert space is also a Banach space.

### 3.3 Reproducing Kernel Hilbert Spaces

We now show that for every p.d. kernel $\varphi : X \times X \to \mathbb{C}$ we can construct a Hilbert space, $H$, associated with it in the sense that there is a mapping $f : X \to H$ such that, for every $x, y \in X$, $\varphi(x, y) = \langle f(x), f(y) \rangle$, where $\langle \cdot, \cdot \rangle$ is the inner product defined on $H$. This $H$ will be called the Reproducing Kernel Hilbert Space associated with $\varphi$, for reasons that will become clear shortly.

First, given $x \in X$ and p.d. kernel $\varphi : X \times X \to \mathbb{C}$, define univariate function $\varphi_x : X \to \mathbb{C}$ as follows: for $y \in X$, $\varphi_x(y) = \varphi(x, y)$.

If $\mathbb{C}^X$ is the set of functions from $X$ to $\mathbb{C}$, let $H_0$ be the linear subspace of $\mathbb{C}^X$ generated by $\{\varphi_x : x \in X\}$, i.e., if $f, g \in H_0$ then for some $n \in \mathbb{N}$, $\{x_1, \ldots, x_n\} \subseteq X$, $\{c_1, \ldots, c_n\} \subseteq \mathbb{C}$, $f = \sum_{j=1}^{n} c_j \varphi_{x_j}$, and, similarly, $g = \sum_{k=1}^{m} d_k \varphi_{y_k}$ for some $m \in \mathbb{N}$, $\{y_1, \ldots, y_m\} \subseteq X$, and $\{d_1, d_2, \ldots, d_n\} \subseteq \mathbb{C}$. We define a bivariate function $\langle \cdot, \cdot \rangle : H_0 \times H_0 \to \mathbb{C}$ as follows:

$$\langle f, g \rangle = \sum_{j,k} c_j \overline{d_k} \varphi(x_j, y_k).$$

**The reproducing property.** To appreciate the definition of $\langle \cdot, \cdot \rangle$, let us define the $n \times m$ matrix $M = (\varphi(x_i, y_j) : i \leq n, j \leq m)$ and consider the row vectors $c = (c_1, \ldots, c_n)$ and $d = (d_1, \ldots, d_m)$. Then

$$\langle f, g \rangle = c M d^*.$$
Note that $cM$ is a $1 \times m$ row vector whose $i$th coordinate is

$$\sum_{j=1}^{n} c_j \varphi(x_j, y_i) = f(y_i).$$

Therefore

$$\langle f, g \rangle = \sum_{i=1}^{m} d_i f(y_i).$$

If $g = \varphi_y$, then we have that $\langle f, \varphi_y \rangle = f(y)$. And, in particular, if $f = \varphi_x$ and $g = \varphi_y$ we get the so called “reproducing property”,

$$\langle \varphi_x, \varphi_y \rangle = \varphi_x(y) = \varphi(x, y).$$

Viewed from the other direction, using the Hermitian property of $\varphi$, note that $Md^*\varphi$ is a $n \times 1$ column vector whose $j$th coordinate is

$$\sum_{i=1}^{m} d_i \varphi(y_i, x_j) = \sum_{i=1}^{m} d_i \varphi(y_i, x_j) = g(x_j),$$

which tells us that

$$\langle f, g \rangle = \sum_{i=1}^{n} c_i g(x_i).$$

**Proposition 3.20.** $\langle \cdot, \cdot \rangle$ is an inner product on $H_0$.

**Proof.** $\langle f, f \rangle \geq 0$ since $\langle f, f \rangle = \sum_{j,k} c_j c_k \varphi(x_j, x_k) \geq 0$ as $\varphi$ is a p.d. kernel. Linearity and the fact that the function is Hermitian are easy to see and so $\langle \cdot, \cdot \rangle$ is a semi-inner product.

By the reproducing property we have that

$$|f(x)|^2 = |\langle f, \varphi_x \rangle|^2 \leq \|f\|^2 \|\varphi_x\|^2 = \langle f, f \rangle \varphi(x, x)$$

since $\varphi$ is a p.d. kernel, $\varphi(x, x) \geq 0$. And $|f(x)|^2 > 0$ unless $f$ is identically 0. Therefor $\langle f, f \rangle > 0$ unless $f$ is identically 0 (i.e. the zero for the vector space $H_0$) .

So, we see that $H_0$ has an inner product defined on it, and this inner product has the “reproducing property”. However it is not clear that $H_0$ is a Hilbert space. Nonetheless $H_0$ can be completed to give a Hilbert space $H$ in a standard way, and this $H$ is called the reproducing kernel Hilbert space associated with $\varphi$.

### 3.4 Mercer’s theorem

Finally, we discuss Mercer’s theorem. We have already seen in Proposition 3.2 that every semi-inner product is a p.d. kernel. Mercer’s theorem gives us the converse: that every p.d. kernel is
an inner product in some Hilbert space. Arguably the construction provided in Section 3.3 has already established this fact. Here we present an alternate way of proving this.

Generalizing Definition 2.3 we say that a bivariate function $\varphi : X \times X \to \mathbb{C}$ is said to satisfy Mercer’s condition if for all square integrable functions $f$, we have

$$\int_{X \times X} \varphi(x,y) f(x) \overline{f(y)} dxdy \geq 0.$$  

Clearly a bivariate function that satisfies this definition is also a p.d. kernel as per Definition 2.3. Mercer’s theorem says that any bivariate function satisfying Mercer’s condition can be expressed as an inner product of a separable Hilbert space. Here we prove a much simpler version of the theorem for p.d. kernels defined on finite sets, noting that the basic arguments are similar to that of the full theorem.

**Theorem 3.21** (Mercer’s Theorem for finite sets). Given a finite set $X$, a Hermitian function $\varphi : X \times X \to \mathbb{C}$ is a p.d. kernel iff there exists a Hilbert space $(H, \langle \cdot , \cdot \rangle_H)$ and a mapping $f : X \to H$ such that $\varphi(x,y) = \langle f(x), f(y) \rangle_H$ for all $x, y \in X$.

**Proof.** Proposition 3.2 establishes that an inner product is p.d. So we turn to the other direction. Assume $\varphi$ is a p.d. kernel. Consider $A = \{\varphi(x_i, x_j) : i, j \leq |X|\}$, i.e., the Gram matrix of $X$ w.r.t. $\varphi$. Let $f : \mathbb{C}^{|X|} \to \mathbb{C}^{|X|}$ and define $f(x) = e_x$, where $e_x$ is a vector in $\mathbb{C}^{|X|}$ s.t. $e_x = 1$ at $x$th position and 0 otherwise. Consider the inner product $\langle \cdot , \cdot \rangle_A$ as defined in Example 3.7. Then $\langle f(x), f(y) \rangle_A = \langle e_x, e_y \rangle_A = e_x A e_y = A_{xy} = \varphi(x, y)$.

\[ \square \]

4 Random Fourier Features

In this section we discuss the technique introduced by Rahimi and Recht [5] for efficient kernel computation. Our treatment is based on the lecture notes by Stephen Tu [7].

4.1 Fourier Transforms and Translation invariant kernels

We begin by defining the Fourier transform of a function. Before that we recall that $L^1(X) = \{f : X \to \mathbb{C} : \int_X |f(x)| dx < \infty\}$ is the set of summable complex-valued functions defined on a set $X$. Within this class we identify those functions which take non-negative real values and have the property that $\int_X f(x) dx = 1$. We call these probability distributions.

**Definition 4.1** (Fourier transform/Characteristic function). Given a $d > 0$ and $f \in L^1(\mathbb{R}^d)$ the Fourier transform of $f$, $\hat{f} : \mathbb{R}^d \to \mathbb{C}$, is defined as

$$\hat{f}(\xi) = \int_{\mathbb{R}^d} f(x)e^{-i\xi^T x} dx.$$  

If $f$ is a probability distribution and $Y$ is a random variable with distribution $f$ then $\hat{f}(\xi) = E[e^{i\xi Y}]$ is called the characteristic function of $Y$.  

26
Remark. In the case when $f$ is a probability distribution it is easy to check that $\hat{f}(0) = 1$. Conversely if $f$ takes non-negative values and $\hat{f}(0) = 1$ then $f$ is a probability distribution.

The Fourier transform has an important positive definiteness property.

Proposition 4.2. Suppose, for some $d > 0$, that $\hat{f}$ is the Fourier transform of some $f \in L^1(\mathbb{R}^d)$. Then the bivariate function $\varphi : \mathbb{R}^d \times \mathbb{R}^d$ defined as

$$\varphi(\xi_1, \xi_2) = \hat{f}(\xi_1 - \xi_2),$$

is a p.d. kernel.

Proof. Consider some $n \in \mathbb{N}$, $\{c_1, \ldots, c_n\} \in \mathbb{C}$ and $\{\xi_1, \ldots, \xi_n\} \in \mathbb{R}^d$. Then

$$\sum_{j,k=1}^{n} c_j \bar{c}_k \varphi(\xi_j, \xi_k) = \sum_{j,k=1}^{n} c_j \bar{c}_k \hat{f}(\xi_j - \xi_k) = \sum_{j,k=1}^{n} c_j \bar{c}_k \int_{\mathbb{R}^d} f(x) e^{-i(\xi_j - \xi_k)^T x} \, dx = \int_{\mathbb{R}^d} f(x) \left\{ \sum_{j,k=1}^{n} c_j \bar{c}_k e^{-i(\xi_j - \xi_k)^T x} \right\} \, dx = \int_{\mathbb{R}^d} f(x) \left\{ \sum_{j,k=1}^{n} c_j e^{-i\xi_j^T x} c_k e^{-i\xi_k^T x} \right\} \, dx = \int_{\mathbb{R}^d} f(x) \left| \sum_{j=1}^{n} c_j e^{-i\xi_j^T x} \right|^2 \, dx > 0$$

The bivariate function derived from a univariate function in Proposition 4.2 suggests the following definition.

Definition 4.3 (Translation invariant kernel). A kernel $\varphi : X \times X \to \mathbb{C}$ is called translation invariant if $\varphi(x_1, x_2) = g(x_1 - x_2)$ for some $g : X \to \mathbb{C}$.

Proposition 4.2 provides a natural way of proving that a translation invariant kernel is p.d.: we can simply show that it is the Fourier transform of some function. Let us discuss this in the context of some examples.

Example 4.4. The following are examples of translation-invariant p.d. kernels.

1. Gaussian kernel: $e^{-\frac{\|x_1-x_2\|^2}{2\sigma^2}}$ for $\sigma > 0$

2. Laplacian kernel: $e^{-\frac{\|x_1-x_2\|}{\sigma}}$ for $\sigma > 0$

3. Sinc kernel: $\sin \left( \frac{a(x_1-x_2)}{\pi(x_1-x_2)} \right)$ for $a > 0$

Remarks.
1. The fact that the Gaussian kernel is p.d. follows from Schonberg’s Theorem (Theorem 2.27) and Example 2.9 extended to \(d\)-dimensions. Later in this lecture we will present an alternate proof for the Gaussian kernel that goes via Proposition 4.2.

2. From Proposition 3.16 we know that the 1-norm is c.n.d. so again applying Schonberg’s Theorem we see that the Laplacian kernel is p.d.

3. For the sinc kernel note that if we define the function

\[
\text{rect}(x) = \begin{cases} 
1, & -\frac{1}{2} \leq x \leq \frac{1}{2} \\
0, & \text{otherwise}
\end{cases}
\]

and, for \(a, b > 0\), \(f_{a,b}(x) = b \cdot \text{rect}(x/a)\) then

\[
\hat{f}_{a,b}(\xi) = b \int_{-a/2}^{a/2} e^{-i\xi x} dx \\
= \frac{b}{i\xi} \left( e^{-ia\xi} - e^{ia\xi} \right) \\
= \frac{2b}{\xi} \sin \left( \frac{a\xi}{2} \right).
\]

Clearly then \(\sin(a\xi)/\pi\xi\) is \(\hat{f}_{a,1/2\pi}\) and hence, by Proposition 4.2, the Sinc kernel is p.d.

The converse of Proposition 4.2 is a celebrated theorem due to Bochner which states, roughly, that every translation-invariant p.d. kernel can be written as a Fourier transform of some function. Since stating this theorem in its full generality is not possible given the theory developed so far, we omit the statement.

From Bochner’s theorem we know that the Gaussian kernel should be expressible as a Fourier transform of some function. In this special case it is possible to explicitly determine the function without resorting to Fourier inversions. In fact we can show the remarkable fact that the Gaussian kernel can be expressed as the characteristic function of a random variable that is itself Gaussian.

**Proposition 4.5.** For \(\sigma \in \mathbb{R}\), if \(Y \in \mathbb{R}^d\) is a normally distributed multivariate random variable with mean 0 and covariance matrix \(1/\sigma^2\cdot I\) then

\[
e^{-\frac{\|\xi\|^2}{2\sigma^2}} = E[e^{-i\xi Y}].
\]

**Proof.** First we characterize the Fourier transform of a univariate Gaussian.

**Lemma 4.6.** Let \(f(x) = e^{-\frac{x^2}{2z}}\) for some positive \(z > 0\). Then

\[
\hat{f}(\xi) = (2\pi)^{\frac{1}{2}} z^{-\frac{1}{2}} e^{-\frac{\xi^2}{2z}}
\]

**Proof of Lemma 4.6.** Differentiating under the Fourier transform integral we have

\[
\frac{d}{d\xi} \hat{f}(\xi) = \frac{d}{d\xi} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2z}} e^{-i\xi x} dx = \int_{-\infty}^{\infty} e^{-\frac{x^2}{2z}} (-i\xi) e^{-i\xi x} dx = \frac{i}{z} \int_{-\infty}^{\infty} e^{-\xi x} \frac{d}{dx} e^{-\frac{x^2}{2z}}.
\]

28
Integrating by parts we get
\[
\frac{d}{d\xi} \hat{f}(\xi) = \frac{i}{z} e^{-i\xi x} \left[ e^{-i\xi x} e^{-\frac{z^2\xi^2}{2}} - \int_{-\infty}^{\infty} (-i\xi) e^{-i\xi x} e^{-\frac{z^2\xi^2}{2}} \right] = \frac{\xi}{z} \hat{f}(\xi).
\]

This is an ordinary differential equation and \(\hat{f}(\xi) = Ce^{\frac{-\xi^2}{2\gamma^2}}\) is a solution for any constant \(C\). By satisfying the boundary condition, we have \(C = \hat{f}(0) = \int e^{-\frac{x^2}{2}} dx\). Since \(\hat{f}(0)\) is the normalization constant of a \(N(0, \frac{1}{z})\) distribution, therefore, \(C = (2\pi)^{\frac{1}{2}}\zeta\frac{1}{2}\).

Now consider the \(d\)-dimensional Gaussian function \(f(x) = e^{-\frac{z^2||x||^2}{2}}\) for \(z > 0\). Using Lemma 4.6 and taking a Fourier transform we have
\[
\hat{f}(x) = \int_{\mathbb{R}^d} e^{-\frac{z^2||\xi||^2}{2}} e^{-i\xi x} \, dx = \prod_{i=1}^{d} \int_{-\infty}^{\infty} e^{-\frac{z^2\xi^2}{2}} e^{-i\xi x_i} \, dx_i = \prod_{i=1}^{d} \hat{f}(\xi_i) = (2\pi)^{\frac{d}{2}} z^{-\frac{d}{2}} e^{-\frac{||x||^2}{2z^2}}
\]
Setting \(z = \sigma^2\) and \(\gamma = 1/\sigma\), we have
\[
e^{-\frac{||x||^2}{2\sigma^2}} = \frac{1}{(2\pi\gamma^2)^{-\frac{d}{2}}} \int_{\mathbb{R}^d} e^{-i\xi x} e^{-\frac{||\xi||^2}{2\gamma^2}} \, dx
\]
We now see that RHS is simply the expectation over a random variable distributed as a multivariate Gaussian for \(N(0, \gamma^2 I)\). \(\square\)

4.2 The Random Fourier Feature method of Rahimi and Recht [5]

In [5], Rahimi and Recht suggested the following mapping:

**Definition 4.7** (Random Fourier Features). Given a set of training points \(\{x_1, ..., x_n\} \subseteq \mathbb{R}^d\) such that \(d > 1\). For some \(\sigma \in \mathbb{R}\) and \(D > 0\), we construct a random feature map \(\phi : \mathbb{R}^d \rightarrow [0, 1]^D\) as

\[
\phi(x) = \begin{bmatrix}
\cos(w_1^T + b_1) \\
\vdots \\
\cos(w_D^T + b_D)
\end{bmatrix}, \text{ } w_i \stackrel{iid}{\sim} N(0, \sigma^2 I), b_i \stackrel{iid}{\sim} \text{Unif}([0, 2\pi])
\]

Note that all \(D\) dimensions are independent of each other.

**Fact 4.1.** Suppose \(x, y \in \mathbb{R}^d\). If \(w \in \mathbb{R}^d\) is chosen randomly with distribution \(N(0, 1/\sigma^2 \cdot I)\) and \(b\) is chosen uniformly at random from \([0, 2\pi]\) then

\[
E_{w,b}[\phi(x)w, \phi(y)w, b] = e^{-\frac{||x-y||^2}{2\sigma^2}}.
\]

**Proof.** From Proposition 4.5 we know that \(\varphi(x, y) = e^{-\frac{||x-y||^2}{2\sigma^2}} = E_w[\varphi(x, y)w]\). Note that
\[
e^{-i(y-x)w} = (e^{i\varphi(x, y)} - i \sin(w^T x) - i \sin(w^T y))
\]
\[
= \cos(w^T x) + i \sin(w^T x)) (\cos(w^T y) - i \sin(w^T y))
\]
\[
= \cos(w^T x) \cos(w^T y) + \sin(w^T x) \sin(w^T y)
\]
\[
+ i(\sin(w^T x) \cos(w^T y) - \cos(w^T x) \sin(w^T y))
\]
\[
= \cos(w^T (x-y)) + i(\sin(w^T (x-y)).
\]
Taking expectations on both sides we and recalling that the Gaussian kernel is real-valued we get that $E_w \left[ e^{-i(y-x)w} \right] = E_w \left[ \cos(w^T(x - y)) \right]$. Further note that

$$E_{w,b} [\phi(x) \cdot \phi(y)] = E_{w,b} \left[ \cos(w^T(x - y)) \right] + E_{w,b} \left[ \cos(w^T(x - y) + 2b) \right]$$

Since $\int_0^{2\pi} \cos(a + 2x)dx = 0$ for any $a \in \mathbb{R}$, the second term on the RHS is 0 and so we have that

$$E_{w,b} [\phi(x) \cdot \phi(y)] = E_w \left[ \cos(w^T(x - y)) \right] = E_w \left[ e^{-i(x-y)w} \right] = \varphi(x,y).$$

From Fact 4.1 it is clear that each coordinate of the random mapping $\phi$ gives us a Gaussian kernel computation in expectation. However a single random choice may be far from the mean. Rahimi and Recht choose $D$ coordinates to reduce the error, i.e., they try to empirically estimate the mean. They define an inner product on $[0,1]^D$ as follows: $\langle x, y \rangle = \frac{1}{D} \sum_{i=1}^{D} x(i)y(i)$. Fact 4.1 tells us that $\langle x, y \rangle$ is in fact an empirical estimate of the value $\varphi(x,y)$. This allows us to use simple dot products in $\mathbb{R}^D$ to estimate the kernel computation in $\mathbb{R}^d$ required for a learning task, as discussed in Section 1.

In Section 5 we will discuss concentration results that help us bound the quality of the empirical estimate in terms of the parameter $D$. We will also show that the Gram matrix defined by $\phi$ on the training set via the inner product $\langle \cdot, \cdot \rangle$ is a good approximation of the Gram matrix defined by the Gaussian kernel.

## 5 Concentration inequalities

In this section we introduce some of the theory of concentration for random variables and random matrices. The theory is vast, so we only discuss that part of it that is relevant to proving some properties of Rahimi and Recht’s method [5] presented in Section 4.2

### 5.1 Concentration for random variables

We now discuss the phenomenon of “concentration”, i.e., the fact that sums of independent random variables are often close to the expectation of their sums. Specifically we are interested in proving Hoeffding’s inequality which applies to random variables that take values in a bounded range. The material in this section is based on the presentation in the book by Boucheron, Lugosi and Massart [2].

#### 5.1.1 The Cramer-Chernoff method

Hoeffding’s inequality and several other concentration results depend on the following simple result that we call the “extended” Markov’s inequality.
Lemma 5.1 (Extended Markov’s Inequality). If $X$ is a real-valued random variable and $\phi : \mathbb{R} \to \mathbb{R}_+$ is a non-decreasing non-negative valued function, then

$$P(X \geq t) \leq \frac{E[\phi(X)]}{\phi(t)}.$$

**Proof.** Observing first that since $\phi$ is non-decreasing, we have, for any random variable $X$ and $t \in \mathbb{R}$, that $P(X > t) \leq P(\phi(X) > \phi(t))$.

Now, suppose $Y$ is a non-negative valued random variable with probability density function $f$, then

$$E[Y] = \int_{0}^{\infty} xf(x)dx \geq \int_{t}^{\infty} xf(x)dx \geq t \int_{t}^{\infty} f(x)dx = t \cdot P(Y \geq t).$$

The result follows. \qed

**Remark.** Chebyshev’s inequality follows from Lemma 5.1. Taking $Y = Z - E[Z]$ and $\phi(x) = x^2$ gives us

$$P(|Z - E[Z]| \geq t) \leq \frac{\text{Var}(Z)}{t^2},$$


**Definition 5.2** (Moments of a Random Variable). Given a random variable $X$, $E[X^k], k \geq 0$ is called the $k^{th}$ moment of $X$.

**Remarks.**

1. $\text{Var}(X)$ i.e the variance of $X$ contains 2 moments in it since $\text{Var}(X) = E[X^2] - E[X]^2$.

2. It is not difficult to see that if $E[|X|^k] < \infty$ for $k > 1$ then $E[|X|^{k-1}] < \infty$.

3. However the converse of the previous remark is not true. For example consider a discrete random variable $X, X > 0$, such that $P(X = i) = \frac{c}{i^3}$ where $c = 1/(\sum_{i=1}^{\infty} 1/i^3)$. In this case $E[X] = \sum_{i=1}^{\infty} \frac{c}{i} < \infty$ but $E[X^2] = \sum_{i=1}^{\infty} \frac{c}{i}$ is unbounded.

4. The more higher moments we have, the tighter the bound we can get using Lemma 5.1. In general, for $k \geq 1$,

$$P(|Z - E[Z]| \geq t) \leq \frac{E[|Z - E[Z]|^k]}{t^k}.$$

When all moments are finite it is convenient to wrap up all the moment information in a single function called the moment generating function.

**Definition 5.3** (Moment generating function and Cumulant generating function). Given a real-valued random variable $Z$, we call

$$M_Z(\lambda) = E[e^{\lambda Z}]$$

the moment generating function of $Z$. Further we call

$$\psi_Z(\lambda) = \log M_Z(\lambda) = E[e^{\lambda Z}]$$

the cumulant generating function of $Z$. 

31
**Remark.** The $k$-th moment of $Z$ can be retrieved by evaluating the $k$-th derivative of $M_Z(\lambda)$ at $\lambda = 0$.

The cumulant generating function has an important additivity property.

**Proposition 5.4.** If $Z_1, \ldots, Z_n$ are independent real-valued random variables and $Z = \sum_{i=1}^n Z_i$ then

$$\psi_Z(\lambda) = \sum_{i=1}^n \psi_{Z_i}(\lambda).$$

**Proof.** The result follows by observing that

$$M_Z(\lambda) = \mathbb{E} \left[ e^{\lambda \sum_{i=1}^n Z_i} \right] = \prod_{i=1}^n \mathbb{E} \left[ e^{\lambda Z_i} \right] = \prod_{i=1}^n \mathbb{E} \left[ e^{\lambda Z_i} \right],$$

where the last equality holds only in the case that $Z_1, \ldots, Z_n$ are independent. \hfill $\square$

With these definitions we are ready to state Chernoff’s inequality.

**Lemma 5.5 (Chernoff’s Inequality).** For a real valued random variable $Z$ and any $t > 0$,

$$P(Z > t) \leq e^{-\psi^*(t)},$$

where

$$\psi^*(t) = \sup_{\lambda > 0} (\lambda t - \psi_Z(\lambda)) = \sup_{\lambda > 0} (\lambda t - \log \mathbb{E}[e^{\lambda Z}]).$$

If $t > \mathbb{E}[Z]$ then

$$\psi^*(t) = \sup_{\lambda \in \mathbb{R}} (\lambda t - \log \mathbb{E}[e^{\lambda Z}]).$$

**Proof.** Noting that the function $\phi(x) = e^{\lambda x}$ takes non-negative values and is non-decreasing whenever $\lambda > 0$, Markov’s inequality (Lemma 5.1) gives us that

$$P(Z > t) \leq e^{-\lambda t \mathbb{E}[e^{\lambda Z}]}$$

for every $\lambda > 0$. The RHS is minimized by choosing the value of $\lambda$ that maximizes $\lambda t - \log \mathbb{E}[e^{\lambda Z}]$.

In case $t > \mathbb{E}[Z]$ then, by Jensen’s inequality, we have that

$$\log \mathbb{E}[e^{\lambda Z}] \geq \mathbb{E} \left[ \log e^{\lambda Z} \right] = \lambda \mathbb{E}[Z].$$

This tells us that if $\lambda < 0$ then

$$\lambda t - \log \mathbb{E}[e^{\lambda Z}] \leq \lambda t - \lambda \mathbb{E}[Z] < 0,$$

and so the maximization can be formally extended over the entire real line. \hfill $\square$

**Remark.** From Proposition 5.4 we can conclude that if $Z = \sum_{i=1}^n Z_i$ for an independent collection $Z_1, \ldots, Z_n$ then

$$\psi^*(t) = \sup_{\lambda > 0} \left( \lambda t - \sum_{i=1}^n \log \mathbb{E}[e^{\lambda Z_i}] \right).$$
5.1.2 Hoeffding’s inequality

In this section we will prove Hoeffding’s inequality and show how it is used in [5]. Hoeffding’s inequality is an application of the Cramer-Chernoff method to sums of independent random variables that take values within a bounded range. The logarithm of the deviation of the sum from its expectation varies inversely with the size of the range.

**Theorem 5.6** (Hoeffding’s inequality). Let \( X_1, X_2, ..., X_n \) be independent random variables such that \( X_i \) takes its values in \([a_i, b_i]\), \(1 \leq i \leq n\). Let \( S = \sum_{i=1}^{n} (X_i - E[X_i])\). Then, for every \( t > 0\),

\[
P(S \geq t) \leq \exp \left( -\frac{2t^2}{\sum_{i=1}^{n} (b_i - a_i)^2} \right).
\]

**Proof.** The proof relies on the following characteristic of the cumulant generating function of a bounded random variable.

**Lemma 5.7** (Hoeffding’s lemma). If \( Z \) is a random variable with \( E[Z] = 0 \) that takes values in \([a, b]\) then

\[
\psi_Z(\lambda) \leq \frac{\lambda^2(b-a)^2}{8}
\]

for all \( \lambda \in \mathbb{R} \).

**Proof of Lemma 5.7.** Since \( \psi_Z(\lambda) \) is at least twice differentiable at \( \lambda = 0 \), by Taylor’s theorem we know that there is a \( \theta \in [0, \lambda] \) such that

\[
\psi_Z(\lambda) = \psi_Z(0) + \psi'_Z(0) + \psi''_Z(\theta).
\]

At \( \lambda = 0 \) this reduces to \( E[Z] \) which is 0 by assumption. So, we turn to the second derivative. Differentiating \( \psi''_Z(\lambda) \) again we get

\[
\psi''_Z(\lambda) = \frac{E[Z^2 e^{\lambda Z}]}{E[e^{\lambda Z}]} - \left( \frac{E[Z e^{\lambda Z}]}{E[e^{\lambda Z}]} \right)^2 = e^{-\psi_Z(\lambda)} E[Z^2 e^{\lambda Z}] - (e^{-\psi_Z(\lambda)} E[Z e^{\lambda Z}])^2.
\]

We will show that the RHS is the variance of a random variable that takes values in \([a, b]\). For any given \( \lambda \in \mathbb{R} \) define a random variable \( X_\lambda \) which takes value \( x \) with probability \( e^{-\psi_Z(\lambda)} \cdot e^{\lambda x} \cdot P(Z = x) \). By the definition of \( \psi_Z(\lambda) \) this is a probability distribution. Since \( P(Z = x) \) is 0 outside \([a, b]\), clearly \( X_\lambda \) takes values only in \([a, b]\). The variance of \( X_\lambda \) is

\[
\text{Var}(X_\lambda) = \int_a^b x^2 e^{-\psi_Z(\lambda)} \cdot e^{\lambda x} \cdot P(Z = x) \, dx - \int_a^b x e^{-\psi_Z(\lambda)} \cdot e^{\lambda x} \cdot P(Z = x) \, dx,
\]

which is equal to \( \psi''_Z(\lambda) \).
For any r.v. $Y$ that takes values in $[a, b]$, we know that $|Y - (b + a)/2| \leq (b - a)/2$ with probability 1 since the distance of any $x \in [a, b]$ from the mid-point of the interval is at most half the length of the interval. This immediately tells us that

$$\text{Var}(Y) = \text{Var}\left(\left|Y - \frac{b + a}{2}\right|\right) \leq \frac{(b - a)^2}{4}.$$ 

Consider the $\theta$ defined in (8). Since $\psi''_Z(\theta) = \text{Var}(X_{\theta})$ and $X_{\theta}$ is also a r.v. that takes values in $[a, b]$, it’s variance is also at most $(b - a)^2/4$ and so the result follows.

Note that if $X$ takes values in $[a, b]$ then so does $X - E[X]$. Hence for each $i$, $1 \leq i \leq n$, by Lemma 5.7, $\psi_{X - X_i}(\lambda) \leq \lambda^2(b_i - a_i)^2/8$. So, since $X_1, \ldots, X_n$ are independent Proposition 5.4 tells us that

$$\psi_S(\lambda) \leq \frac{\lambda^2}{8} \sum_{i=1}^n (b_i - a_i)^2,$$ 

and so, by Chernoff’s inequality (Lemma 5.5)

$$P(S > t) \leq \exp\left\{-\sup_{\lambda \in \mathbb{R}} \left(\lambda t - \sum_{i=1}^n \frac{\lambda^2(b_i - a_i)^2}{8}\right)\right\}.$$ 

By elementary calculus we find that the supremum on the right is achieved at

$$\lambda^* = \frac{4t}{\sum_{i=1}^n (b_i - a_i)^2}.$$ 

Putting this value back into the previous expression gives us the result.

**Example 5.8 (A tail bound for Random Fourier Feautures).** Given a set of training points $T = \{x_1, \ldots, x_n\} \subseteq \mathbb{R}^d$ such that $d > 1$. For some $\sigma \in \mathbb{R}$ and $D > 0$, we construct a random feature map $\phi : \mathbb{R}^d \to [0, 1]^D$ as given in Definition 4.7. Define an inner product on $[0, 1]^D$ as follows: $\langle x, y \rangle = \frac{1}{D} \sum_{k=1}^D x(k)y(k)$. Then, for all $x_i, x_j \in T$, and any $\delta, \epsilon > 0$

$$P\left(\left|\langle \phi(x_i), \phi(x_j) \rangle - e^{-\frac{|x_i - x_j|^2}{2\sigma^2}}\right| > \epsilon\right) < \delta,$$

if

$$D \geq \frac{16}{\epsilon^2} \log \frac{n}{\delta}.$$ 

**Proof.** From Fact 4.1 we know that $E[\phi(x_i)(k)\phi(x_j)(k)] = e^{-\frac{|x_i - x_j|^2}{2\sigma^2}}$. Set $Z_k = \phi(x_i)(k)\phi(x_j)(k)$, $1 \leq k \leq D$. These form an independent collection and each $Z_k$ takes values in $[-2, 2]$. So, by Hoeffding’s inequality (Theorem 5.6) we have that

$$P\left(\sum_{k=1}^D Z_k - E[Z_k] > D\epsilon\right) \leq e^{-\frac{D\epsilon^2}{8}}.$$ 

Setting

$$e^{-\frac{D\epsilon^2}{8}} < \frac{\delta}{n^2},$$

34
and solving for $D$ we get

$$D > \frac{16}{\varepsilon^2} \log \frac{n}{\delta}.$$ 

Since the probability of the estimate being far from the mean is at most $\delta/n^2$ for a single pair therefore, by the union bound, the estimate for any of the $n^2$ pairs $(x_i, x_j), 1 \leq i, j \leq n$ far from the mean with probability at most $\delta$. 

References


