1. Let \( k_1 \) and \( k_2 \) be positive definite kernels on \( \mathbb{R}^p \). Verify that the following are also valid kernels.

\[ \text{[Hint: it suffices to identify the corresponding feature.]} \]

(a) \( x^\top x' \),

(b) \( ck_1(x, x') \), for \( c \geq 0 \),

(c) \( f(x)k_1(x, x')f(x') \) for any function \( f : \mathbb{R}^p \rightarrow \mathbb{R} \),

(d) \( k_1(x, x') + k_2(x, x') \),

(e) \( k_1(x, x')k_2(x, x') \),

(f) \( \exp (k_1(x, x')) \),

(g) \( \exp \left( -\frac{1}{2\gamma^2} \| x - x' \|_2^2 \right) \).

**Answer:**

(a) \( \varphi(x) = x \).

(b) \( \varphi(x) = \sqrt{c}\varphi_1(x) \), where \( \varphi_1 \) is the feature of \( k_1 \).

(c) \( \varphi(x) = f(x)\varphi_1(x) \)

(d) Positive definite as

\[
\sum_{i,j} \alpha_i\alpha_j (k_1(x_i, x_j) + k_2(x_i, x_j)) = \sum_{i,j} \alpha_i\alpha_j k_1(x_i, x_j) + \sum_{i,j} \alpha_i\alpha_j k_2(x_i, x_j) \geq 0.
\]

The feature is obtained by “stacking” vectors \( \varphi_1 \) and \( \varphi_2 \) together.

(e) By writing \( \varphi_1, \varphi_2 \) for the features of \( k_1 \) and \( k_2 \), we have

\[
k_1(x, x')k_2(x, x') = \varphi_1(x)^\top \varphi_1(x') \varphi_2(x')^\top \varphi_2(x) = \text{Tr} \left( \varphi_1(x')\varphi_2(x')^\top \varphi_2(x)\varphi_1(x)^\top \right) = \text{Tr} \left( \Phi(x')\Phi(x)^\top \right) = \langle \Phi(x'), \Phi(x) \rangle,
\]

where the feature is the outer product matrix \( \Phi(x) = \varphi_1(x)\varphi_2(x)^\top \).

(f) From (b), (d) and (e), since addition and multiplication preserves positive definiteness and since all the coefficients in the Taylor series expansion of the exponential function are non-negative, \( \kappa_m(x, x') = \sum_{i=1}^m \frac{k_i(x, x')}{r_i^2} \) is a valid kernel \( \forall m \in \mathbb{N} \). Fix \( \alpha \) and \( \{x_i\} \). Then \( \alpha_m = \sum_{i,j} \alpha_i\alpha_j \kappa_m(x_i, x_j) \geq 0 \forall m \). But \( \alpha_m \rightarrow \sum_{i,j} \alpha_i\alpha_j \exp \left( k_1(x_i, x_j) \right) \) as \( m \rightarrow \infty \), so \( \sum_{i,j} \alpha_i\alpha_j \exp \left( k_1(x_i, x_j) \right) \geq 0 \) as well.

(g) By (a), (b), (f), \( \exp \left( -\frac{1}{2\gamma^2} x^\top x' \right) \) is a valid kernel, but then by (c) so is \( \exp \left( -\frac{1}{2\gamma^2} \| x - x' \|_2^2 \right) \)

\[
\exp \left( -\frac{1}{2\gamma^2} \| x \|_2^2 \right) \exp \left( \frac{1}{2\gamma^2} x^\top x' \right) \exp \left( -\frac{1}{2\gamma^2} \| x' \|_2^2 \right)
\]
2. Assume that kernel $k$ is not strictly positive definite, but that there exist $\{a_i\}_{i=1}^n$ and $\{x_i\}_{i=1}^n$, such that

$$\sum_{i=1}^n \sum_{j=1}^n a_i a_j k(x_i, x_j) = 0.$$ 

Show that then

$$f(x) = \sum_{i=1}^n a_i k(x_i, x) = 0 \quad \forall x \in \mathcal{X}.$$ 

Hence conclude that the RKHS functions of the form $f(x) = \sum_{i=1}^n a_i k(x_i, x)$ have zero norm if and only if they are identically equal to zero. [Hint: assume contrary for some $x = x_{n+1}$ and consider $\sum_{i=1}^{n+1} \sum_{j=1}^{n+1} a_i a_j k(x_i, x_j)$]

**Answer:** Assume $f(x_{n+1}) \neq 0$. Then $\forall a_{n+1}$

$$0 \leq \sum_{i=1}^{n+1} \sum_{j=1}^{n+1} a_i a_j k(x_i, x_j) = \sum_{i=1}^n \sum_{j=1}^n a_i a_j k(x_i, x_j) + 2 a_{n+1} \sum_{i=1}^n a_i k(x_i, x_{n+1}) + a_{n+1}^2 k(x_{n+1}, x_{n+1})$$

$$= 2 a_{n+1} f(x_{n+1}) + a_{n+1}^2 k(x_{n+1}, x_{n+1}).$$

To minimise the expression in the last line, take $a_{n+1} = -f(x_{n+1})/k(x_{n+1}, x_{n+1})$. But this gives

$$0 \leq -f^2(x_{n+1})/k(x_{n+1}, x_{n+1}).$$

Since $k(x_{n+1}, x_{n+1}) > 0$, it must be that $f(x_{n+1}) = 0$. The conclusion about functions of the form $f(x) = \sum_{i=1}^n a_i k(x_i, x)$ is immediate since $\|f\|_{\mathcal{H}_k}^2 = \sum_{i=1}^n \sum_{j=1}^n a_i a_j k(x_i, x_j)$.

Another way to show this is by simply applying the Cauchy-Schwarz inequality in $\mathcal{H}_k$:

$$|f(x)| = |\langle f, k(x, \cdot) \rangle_{\mathcal{H}_k}| \leq \|f\|_{\mathcal{H}_k} \sqrt{k(x, x)}.$$ 

Thus $\|f\|_{\mathcal{H}_k} = 0$ implies $f(x) = 0$, $\forall x$.

3. **(One-Class SVM)** A Gaussian RBF kernel on $\mathcal{X} = \mathbb{R}^p$ is given by

$$k(x, x') = \exp \left( -\frac{1}{2\sigma^2} \|x - x'\|^2 \right).  \quad (1)$$ 

(i) What is $k(x, x)$ for this kernel? What can you conclude about the norm of the features $\varphi(x)$ of $x$? What values can the angles between $\varphi(x)$ and $\varphi(x')$ take? Sketch the set $\{\varphi(x) : x \in \mathcal{X}\}$ as if the features lived in a 2D space.

**Answer:** $k(x, x) = \|\varphi(x)\|^2 = 1$, and $k(x, x') = \langle \varphi(x), \varphi(x') \rangle > 0$, so the angle between any two feature vectors is not larger than $\pi/2$. 

![Diagram of hypersphere of radius 1](image)
(ii) Let \( \{x_i\}_{i=1}^n \) be a set of points in \( \mathcal{X} = \mathbb{R}^p \) (no labels are given). The one-class Support Vector Machine (SVM) is a method for outlier detection which in its primal form is defined as

\[
\min_{w, \xi, \rho} \frac{1}{2} \|w\|^2 + \frac{1}{\nu n} \sum_{i=1}^n \xi_i - \rho, \quad \text{subject to } \langle w, \varphi(x_i) \rangle \geq \rho - \xi_i, \ \xi_i \geq 0,
\]

where \( \nu \) is a given SVM parameter, features \( \varphi(x) \) correspond to the RBF kernel in (1), and \( \xi_i \)'s are the non-negative slack variables. The fitted hyperplane \( \langle w, \varphi(x) \rangle - \rho \) in the feature space separates the majority of points from the origin (while pushing away from the origin as much as possible) and is used to determine "atypical" \( x \)-instances.

Using the 2D intuition from (i), sketch the corresponding hyperplane in the feature space and annotate with \( \rho \), \( w \) and a non-zero slack \( \xi_j \) for an "outlier" \( x_j \). Would it make sense to use the one-class SVM with a linear kernel?

**Answer:** The hyperplane that separates majority of points from the origin is useful for outlier detection precisely because all feature vectors lie on the unit hypersphere. One-class SVM therefore relies on the properties of the RBF kernel and would not make sense with a linear kernel. With linear kernel, gross outliers in the same half-space as majority of data would still be allowed.

(iii) Write the dual form of the one-class SVM, using Lagrangian duality.

[Hint: setting to zero the derivative of the Lagrangian with respect to \( w \) should give \( w = \sum_{i=1}^n \alpha_i \varphi(x_i) \), where \( \alpha_i \geq 0 \) are the Lagrange multipliers of the constraints \( \langle w, \varphi(x_i) \rangle \geq \rho - \xi_i \)]

**Answer:** Lagrangian is given by

\[
L(w, \xi, \rho, \alpha, \beta) = \frac{1}{2} \|w\|^2 + \frac{1}{\nu n} \sum_{i=1}^n \xi_i - \rho - \sum_{i=1}^n \alpha_i (\langle w, \varphi(x_i) \rangle - \rho + \xi_i) - \sum_{i=1}^n \beta_i \xi_i,
\]

for Lagrange multipliers \( \alpha_i \geq 0, \beta_i \geq 0 \). Differentiating w.r.t. \( w, \xi, \rho \) and setting to zero gives

\[
w = \sum_{i=1}^n \alpha_i \varphi(x_i), \quad \alpha_i + \beta_i = \frac{1}{\nu n}, \quad \sum_{i=1}^n \alpha_i = 1.
\]
Substituting back into Lagrangian gives the dual:

$$\max_{\alpha} -\frac{1}{2} \alpha^T K \alpha, \quad \text{subject to} \quad \sum_{i=1}^{n} \alpha_i = 1, \quad \alpha_i \leq \frac{1}{\nu n}.$$  

4. Derive the Gram matrix $\tilde{K}$ of centred features $\tilde{\varphi}(x_i) = \varphi(x_i) - \frac{1}{n} \sum_{r=1}^{n} \varphi(x_r)$ as a function of kernel values $K_{i,j} = k(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$. Show that it takes the form $HKH$, where $H$ is a matrix you should specify. Verify that $H$ is symmetric and idempotent, i.e., $H^2 = H$.

**Answer:** To get the centred features we need

$$\tilde{K}_{i,j} = \langle \varphi(x_i) - \frac{1}{n} \sum_{r=1}^{n} \varphi(x_r), \varphi(x_j) - \frac{1}{n} \sum_{r=1}^{n} \varphi(x_r) \rangle$$

$$= \langle \varphi(x_i), \varphi(x_j) \rangle + \frac{1}{n^2} \sum_{r=1}^{n} \sum_{s=1}^{n} \langle \varphi(x_r), \varphi(x_s) \rangle - \frac{1}{n} \sum_{r=1}^{n} \langle \varphi(x_r), \varphi(x_r) \rangle$$

$$= K_{i,j} - \frac{1}{n^2} \sum_{r=1}^{n} \sum_{s=1}^{n} K_{r,s} - \frac{1}{n} \sum_{r=1}^{n} K_{i,r} - \frac{1}{n} \sum_{r=1}^{n} K_{r,j},$$

which depends only on $K$. In matrix form, $\tilde{K} = (I - \frac{1}{n} 1 1^T)K(I - \frac{1}{n} 1 1^T)$, where the centering matrix $H = I - \frac{1}{n} 1 1^T$ is clearly symmetric. To check idempotence:

$$(I - \frac{1}{n} 1 1^T)(I - \frac{1}{n} 1 1^T) = (I - \frac{1}{n} 1 1^T - \frac{1}{n} 1 1^T + \frac{1}{n^2} 1^T 1)$$

$$= I - \frac{1}{n} 1 1^T.$$

5. Show that

$$\text{MMD}_k(P, Q) = \sup_{f \in \mathcal{H}_k: \|f\|_{\mathcal{H}_k} \leq 1} |E_{X \sim P} f(X) - E_{Y \sim Q} f(Y)|.$$  

**Answer:**

$$\sup_{f \in \mathcal{H}_k: \|f\|_{\mathcal{H}_k} \leq 1} |E_{X \sim P} f(X) - E_{Y \sim Q} f(Y)| = \sup_{f \in \mathcal{H}_k: \|f\|_{\mathcal{H}_k} \leq 1} \left| \langle f, \mu_k(P) - \mu_k(Q) \rangle_{\mathcal{H}_k} \right|$$

$$\leq 1 \cdot \|\mu_k(P) - \mu_k(Q)\|_{\mathcal{H}_k},$$

by Cauchy-Schwarz. Moreover, the equality holds if $f$ is colinear with $\mu_k(P) - \mu_k(Q)$, i.e. the supremum is attained at

$$f = \frac{\mu_k(P) - \mu_k(Q)}{\|\mu_k(P) - \mu_k(Q)\|_{\mathcal{H}_k}}$$

which is called the witness function.

6. Let $L$ be an unnormalized Laplacian matrix of a graph with $C$ connected components. Verify that
(a) Column vector \( \mathbf{1} \) is the eigenvector of \( \mathbf{L} \) with eigenvalue 0.

(b) \( \mathbf{L} \) is positive semi-definite.

(c) \( v \) is an eigenvector of \( \mathbf{L} \) corresponding to 0-eigenvalue if and only if \( v \in \operatorname{span}\{e_1, \ldots, e_C\} \), where

\[
e_{ci} = \begin{cases} 1, & \text{vertex } i \text{ belongs to the connected component } c, \\ 0, & \text{otherwise.} \end{cases}
\]

**Answer:**

(a) Clear since all rows of \( \mathbf{L} \) sum to zero, so \( \mathbf{L}\mathbf{1} = \mathbf{0} \).

(b) Follows from

\[
\alpha^\top \mathbf{L} \alpha = \alpha^\top \mathbf{D} \alpha - \alpha^\top \mathbf{W} \alpha
= \sum_i \alpha_i^2 \deg(i) - \sum_{i,j} \alpha_i \alpha_j w_{ij}
= \sum_i \alpha_i^2 \sum_j w_{ij} - \sum_{i,j} \alpha_i \alpha_j w_{ij}
= \frac{1}{2} \sum_{i,j} (\alpha_i - \alpha_j)^2 w_{ij} \geq 0.
\]

(c) If \( v \) is a 0-eigenvector of \( \mathbf{L} \), then \( v^\top \mathbf{L} v = \frac{1}{2} \sum_{i,j} (v_i - v_j)^2 w_{ij} = 0 \), so for each pair \( \{i, j\} \), \( v_i = v_j \) or \( w_{ij} = 0 \). For any two connected vertices \( i \) and \( j \), i.e. \( w_{ij} > 0 \), it must be \( v_i = v_j \). But then \( v_i = v_j \) also if there is a path between \( i \) and \( j \). Thus, \( v \) is constant on each connected component, i.e. it lies in \( \operatorname{span}\{e_1, \ldots, e_C\} \). Conversely, denote by \( \mathbf{L}_{cc} \) the submatrix of the Laplacian corresponding to the connected component \( c \) (note that \( \mathbf{L} \) is block-diagonal up to permutation of indices). Then \( \mathbf{L}_{cc} \) is the Laplacian of that connected component and \( \mathbf{L}_{cc} \) has entries \( \mathbf{L}_{cc}\mathbf{1} = 0 \) for vertices in \( c \) and zeros otherwise.

7. Verify that for a given partition \( C_1, C_2, \ldots, C_K \) and column vectors \( h_k \in \mathbb{R}^n \) defined as \( h_{k,i} = \frac{1}{\sqrt{|C_k|}} 1_{\{i \in C_k\}} \), we have

\[
\text{ratio-cut}(C_1, \ldots, C_K) = \sum_{k=1}^K h_k^\top \mathbf{L} h_k.
\]

**Answer:** It is immediate from definition that vectors \( h_k \) are orthonormal. Using the previous problem,

\[
h_k^\top \mathbf{L} h_k = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n (h_{k,i} - h_{k,j})^2 w_{ij}
\]

Note that \( (h_{k,i} - h_{k,j})^2 \) is nonzero if and only if exactly one of \( i, j \) belongs to cluster \( C_k \) (and in this case \( (h_{k,i} - h_{k,j})^2 = 1/|C_k| \)). Indeed, if \( i \notin C_k, j \notin C_k \) then \( h_{k,i} = h_{k,j} = 0 \) and if \( i \in C_k, j \in C_k \) then \( h_{k,i} = h_{k,j} = 1/\sqrt{|C_k|} \). Thus

\[
h_k^\top \mathbf{L} h_k = \frac{1}{|C_k|} \sum_{i \in C_k, j \notin C_k} w_{ij} = \frac{\text{cut}(C_k, \bar{C}_k)}{|C_k|}.
\]