Homework for the Course
“Machine Learning with Kernel Methods”

Lilian Besson
ENS de Cachan
Master MVA

Basile Clement
ENS Ulm
Master MVA

Nissim Zerbib
ENS Ulm
Master MVA

For all: `first.last@ens-cachan.fr`

If needed, have a look to the reference page for the course[1]*, and to the homework assignment.*

<u>*Note*</u>: In all this document, the $\mathbb{N}$ and $\mathbb{R}_+$ sets have their usual **English** meaning, that is, the set of positive integers and positive reals – 0 being in none of those sets.

1 Problem 1: Combination rules for kernels

Let K_1, K_2 be two positive definite (p.d.) kernels on a set $\mathcal{X}$.

1.1 Question 1.1

Let $\alpha, \beta \geq 0$ be two real numbers, and define $K \overset{\text{def}}{=} \alpha K_1 + \beta K_2$.

The symmetry of K is immediate from the symmetry of K_1 and K_2; moreover for $n \in \mathbb{N}$, real weights $\{a_i\}_{1 \leq i \leq n} \in \mathbb{R}$, and a dataset $\{x_i\}_{1 \leq i \leq n} \in \mathcal{X}$, we have:

$$\begin{aligned}
\sum_{i,j=1}^{n} a_i a_j K(x_i, x_j) &= \sum_{i,j=1}^{n} a_i a_j \left(\alpha K_1 + \beta K_2\right)(x_i, x_j) \\
&= \left(\sum_{i,j=1}^{n} \alpha a_i a_j K_1\left(x_i, x_j\right)\right) + \left(\sum_{i,j=1}^{n} \beta a_i a_j K_2\left(x_i, x_j\right)\right) \\
&= \alpha \underbrace{\left(\sum_{i,j=1}^{n} a_i a_j K_1\left(x_i, x_j\right)\right)}_{\geq 0} + \beta \underbrace{\left(\sum_{i,j=1}^{n} a_i a_j K_1\left(x_i, x_j\right)\right)}_{\geq 0} \geq 0.
\end{aligned}$$

Thus, K is a positive definite kernel.

1.2 Question 1.2

Define $K : (x, y) \mapsto K_1(x, y) K_2(x, y)$, whose symmetry is immediate from the symmetry of K_1 and K_2.

[1] `http://lear.inrialpes.fr/people/mairal/teaching/2015-2016/MVA/`

Let $n \in \mathbb{N}$, and a dataset $x_1, \dots, x_n$ in $\mathcal{X}$. Denote by $[K_1]$ (resp. $[K_2]$) the positive semidefinite similarity matrice of K_1 (resp. K_2) w.r.t. the dataset. Since $[K_2]$ is positive semidefinite, it has a positive semidefinite square root S: $[K_2] = S^2$, that is, for each i, j, $[K_2]_{ij} = \sum\limits_{m=1}^{n} S_{im} S_{mj}$.

Therefore, for any weights $a_1, \dots, a_n \in \mathbb{R}^n$, we have

$$\begin{aligned}
\sum_{i,j=1}^{n} a_i a_j [K_1]_{ij} [K_2]_{ij} &= \sum_{i,j=1}^{n} a_i a_j [K_1]_{ij} \left(\sum_{m=1}^{n} S_{im} S_{mj} \right) \\
&= \sum_{m=1}^{n} \left(\sum_{i,j=1}^{n} (a_i S_{im}) (a_j S_{mj}) [K_1]_{ij} \right)
\end{aligned}$$

And because $[K_1]$ is positive semidefinite, for each m, the inner sum is nonnegative (using the weights $a_i{}' \stackrel{\text{def}}{=} a_i S_{im}$):

$$= \sum_{m=1}^{n} \underbrace{\left(\sum_{i,j=1}^{n} (a_i S_{im}) (a_j S_{mj}) [K_1]_{ij} \right)}_{\geq 0} \geq 0$$

Since the similarity matrix $[K]$ is defined by $[K]_{ij} = [K_1]_{ij} [K_2]_{ij}$, we have just shown that it was positive semidefinite, for all n and dataset $x_1, \dots, x_n$. Therefore the product kernel K is indeed a p.d. kernel.

1.3 Question 1.3

Let $(K_n)_{n>0}$ be a sequence of p.d. kernels. Assume that, for all $x, y \in \mathcal{X}$, $(K_n(x, y))_{n>0}$ converges to a value $K(x, y) \in \mathbb{R}$. First of all, by uniqueness of the limit, we can indeed define the pointwise limit K as a *function*. Let us show that it is also a p.d. kernel.

Its symmetry is immediate from the symmetry of all of the K_n, so let's take $m \in \mathbb{N}$, $\{a_i\}_{1 \leq i \leq m} \in \mathbb{R}$, and $\{x_i\}_{1 \leq i \leq m} \in \mathcal{X}$, and prove that the similarity matrix is positive semidefinite. We have:

$$\sum_{i,j=1}^{m} a_i a_j K(x_i, x_j) = \sum_{i,j=1}^{m} a_i a_j \left(\lim_{n \to +\infty} K_n(x_i, x_j) \right)$$

By linearity of the limit for each i, j, and because a_i, a_j don't depend on n:

$$= \sum_{i,j=1}^{m} \lim_{n \to +\infty} (a_i a_j K_n(x_i, x_j))$$

But the sum is finite, and m does not depend on n either, so:

$$= \lim_{n \to +\infty} \underbrace{\left(\sum_{i,j=1}^{m} a_i a_j K_n(x_i, x_j) \right)}_{\geq 0} \geq 0.$$

And so, K is also positive definite, as expected.

1.4 Question 1.4

First, we write e^{K_1} as a pointwise convergent series:

$$\mathrm{e}^{K_1}(x, y) = \sum_{t=0}^{+\infty} \frac{K_1(x, y)^t}{t!}.$$

Thanks to 1.2, for each $t \geq 0$, K_1^t is a p.d. kernel (by an immediate recurrence), and thanks to 1.1, for each $T \geq 0$, $\sum_{t=0}^{T} K_1^t / t!$ is also a p.d. kernel (by another immediate recurrence). Finally, thanks to 1.3, we have a pointwise convergent series of p.d. kernels, and so e^{K_1} is also a p.d. kernel.

This conclude the problem 1.

2 Problem 2: Positive Definite Kernels

In this exercise, we will use the course results (proved in section 1) that the sum, product, exponentiation, scaling by a non-negative constant and point-wise limit (when it exists) of p.d. kernels are all p.d. kernels. We also remark that if K is a p.d. kernel on $\mathcal{X}$, then its restriction to a subset of $\mathcal{X}$ is also a p.d. kernel on the restriction (the proof is immediate using Aronszajn's theorem[2]).

We wrote a small numerical script to test the positiveness of the kernels, by trying many random values for n, $a_i \in \mathbb{R}$, and x_i in their respective domain, and it suggested that all the kernels were positive, except for $K_3(x,y) = \log(1+xy)$, $K_5(x,y) = \cos(x+y)$, $K_8(x,y) = \max(x,y)$ and $K_{11}(x,y) = \mathrm{LCM}(x,y)$. Our script found numerical counter-examples for these kernels that we used as inspiration for the actual counter-examples provided in this document. The Python script and a full example of its output are available at `https://bitbucket.org/snippets/lbesson/ay5yE`.

Let's prove or disprove the positive definiteness of each kernel, one by one:

1. $K(x,y) = \frac{1}{1-xy} = \sum\limits_{k=0}^{\infty} (xy)^k$ with $\mathcal{X} = \{-1, 1\}$ **is a p.d. kernel** as limit of a sum of restricted (to $\{-1,1\}^2$) polynomial kernels.
2. $K(x,y) = 2^{xy} = e^{xy \ln 2}$ with $\mathcal{X} = \mathbb{N}$ **is a p.d. kernel** as exponential of a scaled (by a factor $\ln 2 > 0$) and restricted (to $\mathbb{N}^2$) linear kernel.
3. $K(x,y) = \log(1+xy)$ with $\mathcal{X} = \mathbb{R}_+$ **is not a p.d. kernel**.
   For instance, let us consider the two points $x = 1$ and $y = 2$. The similarity matrix is $[K] = \begin{bmatrix} \log(1+1) & \log(1+2) \\ \log(1+2) & \log(1+4) \end{bmatrix} = \begin{bmatrix} \log 2 & \log 3 \\ \log 3 & \log 5 \end{bmatrix}$ and we have $(-2 \quad 1)\,[K] \begin{pmatrix} -2 \\ 1 \end{pmatrix} = 4\log 2 - 4\log 3 + \log 5 = \log \frac{5\times 2^4}{3^4} = \log \frac{80}{81} < 0$.
4. $K(x,y) = \mathrm{e}^{-(x-y)^2} = \mathrm{e}^{2xy}\mathrm{e}^{-x^2}\mathrm{e}^{-y^2}$ with $\mathcal{X} = \mathbb{R}$ **is a p.d. kernel** as product of the p.d. kernels $K_1 : (x,y) \mapsto \mathrm{e}^{2xy}$ and $K_2 : (x,y) \mapsto \mathrm{e}^{-x^2}\mathrm{e}^{-y^2}$. K_1 is a p.d. kernel as exponential of a scaled (by a factor $2 > 0$) linear kernel, and K_2 is a p.d. kernel by applying Aronszajn's theorem to the mapping $\Phi : x \mapsto \mathrm{e}^{-x^2}$ from $\mathcal{X}$ into the Euclidean space $\mathbb{R}$.
5. $K(x,y) = \cos(x+y)$ with $\mathcal{X} = \mathbb{R}$ **is not a p.d. kernel**.
   For instance, let us consider the two points $x = \frac{\pi}{2}$ and $y = 0$. The similarity matrix is $[K] = \begin{bmatrix} \cos(\pi) & \cos(\frac{\pi}{2}) \\ \cos(\frac{\pi}{2}) & \cos(0) \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$ and we have $(1 \quad 0)\,[K] \begin{pmatrix} 1 \\ 0 \end{pmatrix} = -1 < 0$.
6. $K(x,y) = \cos(x-y) = \cos x \cos y + \sin x \sin y$ with $\mathcal{X} = \mathbb{R}$ **is a p.d. kernel**, by applying Aronszajn's theorem to the mapping $\Phi : x \mapsto (\cos x \quad \sin x)$ from $\mathcal{X}$ into the Euclidean space $\mathbb{R}^2$ (but its image is just the unit circle).
7. $K(x,y) = \min(x,y)$ with $\mathcal{X} = \mathbb{R}_+$ **is a p.d. kernel**.
   First, let us define $\Phi : \mathbb{R} \mapsto L^2(\mathbb{R}_+)$ that maps a real number x, to the square-integrable step function $\Phi(x) = \mathbb{1}_{[0,x]}$ (i.e., $t \mapsto 1$ if $t \leq x$, 0 otherwise). Two interesting properties of these step functions are that for any two real non-negative numbers a and b we have $\int_0^\infty \Phi(a) = \int_0^a 1 = a$ and – since $[0,a] \cap [0,b] = [0, \min(a,b)]$ – $\Phi(\min(a,b)) = \Phi(a)\Phi(b)$; thus we have:

$$\begin{aligned} \min(a,b) &= \int_0^{+\infty} \mathbb{1}_{[0,a]} \mathbb{1}_{[0,b]} \\ &= \langle \Phi(a), \Phi(b) \rangle_{L^2(\mathbb{R}_+)}. \end{aligned}$$

   Applying Aronszajn's theorem to Φ then proves that K is a p.d. kernel.

[2] See slide #42 of the course, for more details on Aronszajn's theorem.

8. $K(x,y) = \max(x,y)$ with $\mathcal{X} = \mathbb{R}_+$ **is not a p.d. kernel**.
   For instance, let's consider the two points $x = 2$ and $y = 1$. The similarity matrix is $[K] = \begin{bmatrix} 2 & \max(1,2) \\ \max(1,2) & 1 \end{bmatrix} = \begin{bmatrix} 2 & 2 \\ 2 & 1 \end{bmatrix}$ and we have $(1 \quad -1)\,[K] \begin{pmatrix} 1 \\ -1 \end{pmatrix} = 2 - 2 - (2-1) = -1 < 0$.
9. $K(x,y) = \min(x,y)/\max(x,y)$ with $\mathcal{X} = \mathbb{R}_+$ **is a p.d. kernel**.
   Let's first prove that $K_2(x,y) = 1/\max(x,y)$ is a p.d. kernel. For any $x, y > 0$, we have $1/\max(x,y) = \min(1/x, 1/y)$, so this is similar to the previous question and follows from applying Aronszajn's theorem to the mapping $\Phi : \mathbb{R}_+ \to L^2(\mathbb{R}_+), x \mapsto \mathbb{1}_{[0,1/x]}$.
   Then, $K(x,y) = \min(x,y)/\max(x,y) = \min(x,y) \cdot K_2(x,y)$ is a p.d. kernel as product of two p.d. kernels.
10. $K(x,y) = \mathrm{GCD}(x,y)$ with $\mathcal{X} = \mathbb{N}$ **is a p.d. kernel**.
    Indeed, let's define the mapping $\Phi : \mathbb{N} \mapsto \bigoplus_{i \in \mathbb{N}} L^2(\mathbb{R}_+)$ such that if $x = \prod_{i \in \mathbb{N}} p_i^{\alpha_i}$ is the (unique) prime factorization of $x \in \mathbb{N}$, $\Phi(x)_i = \sqrt{\log p_i}\,\mathbb{1}_{[0,\alpha_i]}$ for all $i \in \mathbb{N}$.
    To ensure $\Phi(x) \in \bigoplus_{i \in \mathbb{N}} L^2(\mathbb{R}_+)$, we check that $\sum_{i=0}^{+\infty} \|\Phi(x)_i\|^2 = \sum_{i=0}^{+\infty} \alpha_i \log p_i = \log x < \infty$.
    We also remark that since the p_i are prime numbers, $p_i > 1$ and thus $\log p_i > 0$, i.e. $\sqrt{\log p_i}$ is well-defined.
    Now, for any two integers x and y with respective prime factorizations $x = \prod_{i \in \mathbb{N}} p_i^{\alpha_i}$ and $y = \prod_{i \in \mathbb{N}} p_i^{\beta_i}$, we have:
    $$\langle \Phi(x), \Phi(y) \rangle = \sum_{i \in \mathbb{N}} \int_0^\infty \sqrt{\log p_i}\,\mathbb{1}_{[0,\alpha_i]} \sqrt{\log p_i}\,\mathbb{1}_{[0,\beta_i]} = \sum_{i \in \mathbb{N}} \min(\alpha_i, \beta_i) \log p_i.$$
    Since $\mathrm{GCD}(x,y) = \prod_{i \in \mathbb{N}} p_i^{\min(\alpha_i,\beta_i)}$, applying Aronszajn's theorem to Φ proves that $\log \mathrm{GCD}$ is a p.d. kernel – from which we conclude immediately that $K = e^{\log \mathrm{GCD}}$ is a p.d. kernel as exponential of a p.d. kernel.
11. $K(x,y) = \mathrm{LCM}(x,y)$ with $\mathcal{X} = \mathbb{N}$ **is not a p.d. kernel**.
    For instance, let's consider the two points $x = 2$ and $y = 1$ (again). The similarity matrix is $[K] = \begin{bmatrix} 2 & \mathrm{LCM}(1,2) \\ \mathrm{LCM}(1,2) & 1 \end{bmatrix} = \begin{bmatrix} 2 & 2 \\ 2 & 1 \end{bmatrix}$ that was shown to not be positive semidefinite in question 8.
12. $K(x,y) = \mathrm{GCD}(x,y)/\mathrm{LCM(x,y)}$ with $\mathcal{X} = \mathbb{N}$ **is a p.d. kernel**.
    Let's start by remarking that $K_2(x,y) = 1/(x \cdot y)$ is a p.d. kernel, thanks to Aronszajn's theorem applied to the mapping $\Phi : \mathbb{N} \to \mathbb{R}, x \mapsto 1/x$.
    Then we show that $1/\mathrm{LCM}(x,y) = \mathrm{GCD}(x,y)/(x \cdot y)$, is a p.d. kernel as product of two p.d. kernels (see question 10 for the p.d. nature of GCD).
    Finally, we use exactly the same trick as above for the $\min / \max$ kernel, here $K(x,y) = \mathrm{GCD}(x,y)/\mathrm{LCM(x,y)} = \mathrm{GCD}(x,y) \cdot 1/\mathrm{LCM(x,y)}$ is a p.d. kernel as product of two p.d. kernels again.

This conclude the problem 2.

---

3 Problem 3: Covariance Operators in RKHS

For this whole exercise, for $n \in \mathbb{N}$ we will denote by $U = \frac{1}{n} \cdot \mathbb{1} \cdot \mathbb{1}^T$ the $n \times n$ matrix with all coefficients equal to $\frac{1}{n}$.

3.1 Question 3.1

Since the regular product on $\mathbb{R}$ is a scalar product, the RKHS for the linear kernel is $\mathbb{R}$ with an identity embedding. As such, we have, for $f, g \in \mathbb{R}$:

$$\begin{aligned}
\mathrm{cov}_n(f(X), g(X)) &= \mathbb{E}_n[fXgY] - \mathbb{E}_n[fX]\mathbb{E}_n[gY] \\
&= \frac{1}{n}\sum_{i=1}^n fX_i gY_i - (\frac{1}{n}\sum_{i=1}^n fX_i)(\frac{1}{n}\sum_{i=1}^n gY_i) \\
&= \frac{fg}{n}\left(\sum_{i=1}^n X_iY_i - \sum_{i=1}^n X_i \sum_{j=1}^n \frac{1}{n} Y_j\right) \\
&= \frac{fg}{n}\left(X^TY - X^TUY\right).
\end{aligned}$$

Since the constraints $f, g \in \mathcal{B}_K$ mean that $|f| \leq 1$ and $|g| \leq 1$ in this simple case, we then deduce that:

$$\begin{aligned}
C_n^K(X,Y) &= \max_{f,g\in\mathcal{B}_K} \frac{fg}{n} X^T(I-U)Y \\
&= \frac{|X^T(I-U)Y|}{n}.
\end{aligned}$$

3.2 Question 3.2

Let's first show that we can restrict ourselves to f and g with representations of form $f = \sum_{i=1}^n F_i K_{X_i}$ and $g = \sum_{i=1}^n G_i K_{Y_i}$. Indeed, suppose that we have a solution (f^*, g^*) for the maximization problem defining C_n^K; then f^* is also solution of the maximization problem $\max_{f\in\mathcal{B}_K} \mathrm{cov}_n(f(X), g^*(Y))$. Since $\mathrm{cov}_n(f(X), g^*(Y)) = \frac{1}{n}\sum_{i=1}^n \langle f, X_i\rangle g^*(Y_i) - \frac{1}{n^2}\sum_{i,j=1}^n \langle f, X_i\rangle g^*(Y_i)$ is linear in f, this optimization problem is a convex optimization problem in f for which strong duality holds (take $f = 0$ to check for Slater's condition).

Since the dual problem satisfies the conditions of the representer theorem, we conclude that f^* admits a representation of the aforementioned form. Using an f^* with this form, we apply the same reasoning to g^* to obtain an optimal pair (f^*, g^*) where both f^* and g^* have the aforementioned forms.

If we design by F the vector $(F_1, \ldots, F_n)$, we have that $f(X_i) = [K_XF]_i$ and $\|f\|^2 = F^TK_XF$ (and similar relations for G, g and Y, mutatis mutandi); thus we can write:

$$\begin{aligned}
\mathrm{cov}_n(X,Y) &= \frac{1}{n}\sum_{i=1}^n f(X_i)g(Y_i) - \frac{1}{n}\sum_{i=1}^n f(X_i)\frac{1}{n}\sum_{j=1}^n g(Y_j) \\
&= \frac{1}{n}\sum_{i=1}^n [K_XF]_i[K_YG]_i - \frac{1}{n}\sum_{i=1}^n [K_XF]_i \sum_{j=1}^n \frac{1}{n}[K_YG]_j \\
&= \frac{1}{n}\left((K_XF)^TK_YG - (K_XF)^TUK_YG\right) \\
&= \frac{1}{n}F^TK_X(I-U)K_YG.
\end{aligned}$$

So we can rewrite $n \times C_n^K$ as the solution of the maximization of $F^TK_X(I-U)K_YG$ subject to $F^TK_XF \leq 1$ and $G^TK_YG \leq 1$. Recalling that K_X and K_Y are positive semi-definite matrices, and as such admit a positive semi-definite square root, we can rewrite this again as the maximization of $(K_X^{1/2}F)^T K_X^{1/2}(I-U)K_Y^{1/2}(K_Y^{1/2}G)$, subject to $\|K_X^{1/2}F\| \leq 1$ and $\|K_Y^{1/2}G\| \leq 1$.

We now claim that this is equivalent to the maximization of $\tilde{F}^T K_X^{1/2}(I-U)K_Y^{1/2}\tilde{G}$, subject to $\|\tilde{F}\| \leq 1$ and $\|\tilde{G}\| \leq 1$. This is trivial when $K_X^{1/2}$ and $K_Y^{1/2}$ are invertible; however the general case requires more care. We have:

- For any F, G such that $\|K_X^{1/2}F\| \leq 1$ and $\|K_Y^{1/2}G\| \leq 1$, we can define $\tilde{F} = K_X^{1/2}F$ and $\tilde{G} = K_Y^{1/2}G$ satisfying $\|\tilde{F}\| \leq 1$ and $\|\tilde{G}\| \leq 1$ and such that $\tilde{F}^T K_X^{1/2}(I-U)K_Y^{1/2}\tilde{G} = F^T K_X(I-U)K_Y G$.
- Conversely, recall that, as real-valued symmetric matrices, $K_X^{1/2}$ and $K_Y^{1/2}$ are diagonalizable in an orthogonal basis. As such, for any $\tilde{F}, \tilde{G}$ such that $\|\tilde{F}\| \leq 1$ and $\|\tilde{G}\| \leq 1$, we can write $\tilde{F} = K_X^{1/2}F + k_F$ and $\tilde{G} = K_Y^{1/2}G + k_G$ for some vectors F, G, k_F and k_G such that $K_X^{1/2}k_F = K_Y^{1/2}k_G = 0$, and $\langle k_F, K_X^{1/2}F\rangle = \langle k_G, K_Y^{1/2}G\rangle = 0$. By orthogonality, we have $\|K_X^{1/2}F\| = \|\tilde{F}\| - \|k_F\| \leq 1 - \|k_F\| \leq 1$ and similarly $\|K_Y^{1/2}G\| \leq 1$; moreover $\tilde{F}^T K_X^{1/2}(I-U)K_Y^{1/2}\tilde{G} = (K_X^{1/2}(K_X^{1/2}F + k_F))^T(I-U)K_Y^{1/2}(K_Y^{1/2}G + k_G) = F^T K_X(I-U)K_Y G$.

The, these two optimization problems are indeed equivalent, and we have

$$\begin{aligned} n \times C_n^K(X,Y) &= \max_{\|\tilde{F}\|\leq 1, \|\tilde{G}\|\leq 1} \tilde{F}^T K_X^{1/2}(I-U)K_Y^{1/2}\tilde{G}, \\ &= \max_{\|\tilde{G}\|\leq 1} \max_{\|\tilde{F}\|\leq 1} \tilde{F}^T K_X^{1/2}(I-U)K_Y^{1/2}\tilde{G}. \end{aligned}$$

Considering a fixed $\tilde{G}$; if we define $M_G = K_X^{1/2}(I-U)K_Y^{1/2}\tilde{G}$, $\tilde{F}$ is a solution of $\arg\max_{\|\tilde{F}\|\leq 1} \tilde{F}^T M_G$. This is simply the maximization of a scalar product on the unit ball, reached on $\tilde{F} = \frac{M_G}{\|M_G\|}$. Plugging this back in, we get

$$n \times C_n^K(X,Y) = \max_{\|\tilde{G}\|\leq 1} \tilde{F}^T M_G = \max_{\|\tilde{G}\|\leq 1} \frac{M_G^T M_G}{\|M_G\|} = \max_{\|\tilde{G}\|\leq 1} \|K_X^{1/2}(I-U)K_Y^{1/2}\tilde{G}\|.$$

We recognize the spectral norm, and conclude that:

$$C_n^K(X,Y) = \frac{1}{n}\|K_X^{1/2}(I-U)K_Y^{1/2}\|_2.$$

This concludes the problem 3.

---

4 Problem 4: Some Basic Learning Bounds

4.1 Question 4.1

Let us take $x \in \mathcal{X}$ and f, g in $\mathcal{H}_K$. We have the following:

$$|R_\phi(f,x) - R_\phi(g,x)| = \left|\phi(f(x)) - \phi(g(x)) + \lambda(\|f\|^2_{\mathcal{H}_K} - \|g\|^2_{\mathcal{H}_K})\right|$$

(The norms are still $\|\cdot\|_{\mathcal{H}_K}$ but for sake of readability we simplify the notations here.)

$$\leq |\phi(f(x)) - \phi(g(x))| + \lambda\Big|(\|f\| + \|g\|)(\|f\| - \|g\|)\Big|$$

And ϕ is L-Lipschitz

$$\leq L|f(x) - g(x)| + \lambda(\|f\| + \|g\|)\Big|\|f\| - \|g\|\Big|$$

For this right term $\Big|\|f\| - \|g\|\Big|$, we use the left triangle inequality

$$\leq L|\langle f, K_x\rangle - \langle g, K_x\rangle| + \lambda(R + R)\|f - g\|$$

By linearity of the scalar product

$$\begin{aligned}
&= L|\langle f - g, K_x\rangle| + 2\lambda R\|f - g\| \\
&\leq L\|f - g\|\|K_x\| + 2\lambda R\|f - g\| \\
&\leq (\kappa L + 2\lambda R)\|f - g\|_{\mathcal{H}_K}.
\end{aligned}$$

Hence we can take $C_1 := \kappa L + 2\lambda R$.

4.2 Question 4.2

ϕ is a convex function on the open interval equal to the real line, so it is continuous. Thus ϕ admits in each point $u \in \mathbb{R}$ a subgradient (subderivative here) $\alpha_u \in \mathbb{R}$ such that for every $v \in \mathbb{R}$:

$$\phi(v) \geq \phi(u) + \alpha_u(v - u).$$

Note that the composition $g \mapsto \phi(g(x)) = \phi(\langle g, K_x\rangle)$ admits a subgradient in every $g \in \mathcal{H}_K$ of the form $\{\alpha_{g(x)}K_x \,:\, \alpha_{g(x)} \text{ subdifferential of } \phi \text{ in } g(x)\}$.

Hence

$$\phi(f(x)) \geq \phi(f_x(x)) + \left\langle \alpha_{f_x(x)}K_x, f - f_x\right\rangle.$$

Since $\|f\|^2 = \|f_x\|^2 + 2\langle f_x, f - f_x\rangle + \|f - f_x\|^2$, we can add both equations and obtain:

$$\phi(f(x)) + \lambda\|f\|^2 \geq \phi(f_x(x)) + \lambda\|f_x\|^2 + \left\langle(\alpha_{f_x(x)}K_x + 2\lambda f_x), f - f_x\right\rangle + \lambda\|f - f_x\|^2. \quad (1)$$

for every element $\alpha_{f_x(x)}$ of the subdifferential of ϕ at $f_x(x)$.

By optimality condition, there exists α^* in the subdifferential of this function such that $\alpha^* K_x + 2\lambda f_x = 0$. Using this α^* in equation (1):

$$\begin{aligned}
R_\phi(f,x) - R_\phi(f_x,x) &\geq \langle 0, f - f_x\rangle + \lambda\|f - f_x\|^2 \\
&\geq C_2\|f - f_x\|^2_{\mathcal{H}_K}.
\end{aligned}$$

where $C_2 := \lambda$.

4.3 Question 4.3

For every $f \in B_R$ and for each realization of the random variable X:

$$\psi(f, X)^2 = |R_\phi(f, X) - R_\phi(f_X, X)|^2$$

By question 4.1

$$\leq C_1^2 \|f - f_X\|_{\mathcal{H}_k}^2$$

By question 4.2

$$\leq C_1^2 \frac{1}{C_2} \psi(f, X)$$
$$\leq \frac{C_1^2}{C_2} \psi(f, X).$$

Hence by taking expectations, we obtain that for every $f \in B_R$:

$$\mathbb{E}\left[\psi(f, X)^2\right] \leq C \cdot \mathbb{E}\left[\psi(f, X)\right].$$

where $C := \frac{C_1^2}{C_2} = \frac{(\kappa L + 2\lambda R)^2}{\lambda}$.

This conclude the problem 4.

References

We mainly used the course slides, the Wikipedia page on Aronszajn theorem, and this tutorial on RKHS (`http://math.unm.edu/~alvaro/rkhs_tutorial.pdf`).