Homework
Kernel Methods

Quentin Duchemin

February 2019

Exercise 1

We recall some useful results for the exercise :

Theorem 1. *Let $\mathcal{X}$ be a set.*
If $(P_i)_{i\geq 0}$ is a sequence of p.d. kernels that converges pointwisely to a function P, then P is a p.d. kernel.

Theorem 2. *Let $\mathcal{X}$ be a set.*
If $P_1 : \mathcal{X} \to \mathbb{R}$ and $P_2 : \mathcal{X} \to \mathbb{R}$ are p.d. kernels, then $P_1 + P_2$ is a p.d. kernel. A trivial induction gives us that for any finite family of p.d. kernels $(P_i)_{i\in[\![1,n]\!]}$ ($n \in \mathbb{N}$), $\sum_{i=1}^n P_i$ is a p.d. kernel.

Theorem 3. *Let $\mathcal{X}$ be a set.*
If $P : \mathcal{X} \to \mathbb{R}$ is a p.d. kernel, then P^2 (understood as the Hadamard product) is a p.d. kernel. A trivial induction gives us that P^k is a p.d. kernel for all $k \in \mathbb{N}$.

1. • The kernel

$$\begin{aligned} K : \mathbb{R} \times \mathbb{R} &\to \mathbb{R} \\ (x, y) &\mapsto \cos(x - y) \end{aligned}$$

is clearly symmetric since the function cosinus is an <u>even function</u>.

• Let $N \in \mathbb{N}$, $(\alpha_i)_{i=1}^N \in \mathbb{R}^N$ and $(x_i)_{i=1}^N \in \mathbb{R}^N$.
We recall the usual identity for the cosinus of a difference : $\forall (a, b) \in \mathbb{R}^2,\ \cos(a - b) = \cos(a)\cos(b) + \sin(a)\sin(b)$ which leads to :

$$\begin{aligned} \sum_{i=1}^N \sum_{j=1}^N \alpha_i\alpha_j K(x_i, x_j) &= \sum_{i=1}^N \sum_{j=1}^N \alpha_i\alpha_j \cos(x_i - x_j) \\ &= \sum_{i=1}^N \sum_{j=1}^N \alpha_i\alpha_j \left(\cos(x_i)\cos(x_j) + \sin(x_i)\sin(x_j)\right) \\ &= \sum_{i=1}^N \sum_{j=1}^N \alpha_i\alpha_j \cos(x_i)\cos(x_j) + \sum_{i=1}^N \sum_{j=1}^N \alpha_i\alpha_j \sin(x_i)\sin(x_j) \\ &= \left(\sum_{i=1}^N \alpha_i \cos(x_i)\right)^2 + \left(\sum_{i=1}^N \alpha_i \sin(x_i)\right)^2 \\ &\geq 0 \end{aligned}$$

Hence, the kernel K is positive definite.

2. • Let $\mathcal{X} = \{x \in \mathbb{R}^p : ||x||_2 < 1\}$. The kernel

$$\begin{aligned} K : \mathcal{X} \times \mathcal{X} &\to \mathbb{R} \\ (x, y) &\mapsto \frac{1}{1 - x^T y} \end{aligned}$$

is symmetric since $\forall (x,y) \in \mathcal{X}^2$, $x^T y = y^T x$.

- We denote by $\overline{K}$ the linear kernel on $\mathcal{X}$, i.e.

$$\begin{aligned} \overline{K} : \mathcal{X} \times \mathcal{X} &\to \mathbb{R} \\ (x,y) &\mapsto x^T y \end{aligned}$$

We remark that $\forall (x,y) \in \mathcal{X}^2$, the Cauchy-Schwarz inequality gives us $|x^T y| = | < x|y >_{\mathbb{R}^p} | \leq ||x||_2.||y||_2 < 1$ by definition of the set $\mathcal{X}$. This fact allows us to express the kernel K using the Taylor series expansion of the function $f(x) = \frac{1}{1-x} = \sum_{n=0}^{+\infty} x^n$, $\forall x \in]-1,1[$.

Thus $K(x,y) = \lim\limits_{n \to +\infty} \sum_{k=0}^{n} (\overline{K}(x,y))^k$.

  - We know from the course that the Hadamard product of two p.d. kernels is a p.d. kernel. By induction, we get that for all $k \in \mathbb{N}$, the kernel $(x,y) \mapsto \overline{K}(x,y)^k$ is a p.d. kernel $(*)$ (since the linear kernel is a p.d. kernel). This is the theorem (3).
  - We know form the course that the sum of two p.d. kernels is a p.d. kernel. Thus, by induction, for all $n \in \mathbb{N}$, $\sum_{k=0}^{n} (\overline{K}(x,y))^k$ is a p.d. kernel using $(*)$.
  - Using the theorem 1, $K(x,y) = \lim\limits_{n \to +\infty} \sum_{k=0}^{n} (\overline{K}(x,y))^k$ is a p.d. kernel using the previous item.

Hence, the kernel K is positive definite.

3. - Let $(\Omega, \mathcal{A}, P)$ a probability space. The kernel

$$\begin{aligned} K : \mathcal{A} \times \mathcal{A} &\to \mathbb{R} \\ (A,B) &\mapsto P(A \cap B) - P(A)P(B) \end{aligned}$$

is clearly symmetric.

- We remark that for all $(A,B) \in \mathcal{A}^2$,

$$\begin{aligned} P(A \cap B) - P(A)P(B) &= \mathbb{E}[\mathbb{1}_{A \cap B}] - \mathbb{E}[\mathbb{1}_A]\mathbb{E}[\mathbb{1}_B] \\ &= \mathbb{E}[\mathbb{1}_A \mathbb{1}_B] - \mathbb{E}[\mathbb{1}_A]\mathbb{E}[\mathbb{1}_B] \\ &= Cov[\mathbb{1}_A, \mathbb{1}_B] \quad (*) \end{aligned}$$

Let $N \in \mathbb{N}$, $(\alpha_i)_{i=1}^N \in \mathbb{R}^N$ and $(A_i)_{i=1}^N \in \mathcal{A}^N$.

Using $(*)$ and the bilinearity of the Covariance, we have :

$$\begin{aligned} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j K(A_i, A_j) &= \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j Cov[\mathbb{1}_{A_i}, \mathbb{1}_{A_j}] \\ &= Cov\left[\sum_{i=1}^{N} \alpha_i \mathbb{1}_{A_i} \, , \, \sum_{j=1}^{N} \alpha_j \mathbb{1}_{A_j} \right] \\ &= Var\left[\sum_{i=1}^{N} \alpha_i \mathbb{1}_{A_i} \right] \\ &\geq 0 \end{aligned}$$

Hence, the kernel K is positive definite.

4. • Let $\mathcal{X}$ be a set and $f, g : \mathcal{X} \to \mathbb{R}_+$ two non-negative functions.

   The kernel

$$K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$$
$$(x, y) \mapsto \min\{f(x)g(y), f(y)g(x)\}$$

   is clearly symmetric.

   • We adopt the convention that for all $a \in \mathbb{R}$, $\dfrac{a}{0} = 0$. This convention allows us to have for all $(x, y) \in \mathcal{X}$,

$$K(x,y) = \min\{f(x)g(y), f(y)g(x)\} = \frac{1}{g(x)g(y)} \min\left\{\frac{f(x)}{g(x)}, \frac{f(y)}{g(y)}\right\}.$$

   We have used the fact that f and g are non negative. Moreover, the convention adopted makes this equality holds even when $g(x) = 0$ or $g(y) = 0$.

   Using this reformulation we have :

$$\begin{aligned}
K(x,y) &= \min\{f(x)g(y), f(y)g(x)\} \\
&= \frac{1}{g(x)g(y)} \min\left\{\frac{f(x)}{g(x)}, \frac{f(y)}{g(y)}\right\} \\
&= \frac{1}{g(x)g(y)} \int_0^{+\infty} \mathbb{1}_{\{t \leq \frac{f(x)}{g(x)}\}} \mathbb{1}_{\{t \leq \frac{f(y)}{g(y)}\}} dt \\
&= < t \mapsto \frac{1}{g(x)} \mathbb{1}_{\{t \leq \frac{f(x)}{g(x)}\}} \mid t \mapsto \frac{1}{g(y)} \mathbb{1}_{\{t \leq \frac{f(y)}{g(y)}\}} > \qquad (*)
\end{aligned}$$

   where $< .|. >$ denotes the usual scalar product on $L^2(\mathbb{R}_+)$.
   Let $N \in \mathbb{N}$, $(\alpha_i)_{i=1}^N \in \mathbb{R}^N$ and $(x_i)_{i=1}^N \in \mathcal{X}^N$.
   Using $(*)$ and the bilinearity of the scalar product, we have :

$$\begin{aligned}
\sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j K(x_i, x_j) &= \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j < t \mapsto \frac{1}{g(x_i)} \mathbb{1}_{\{t \leq \frac{f(x_i)}{g(x_i)}\}} \mid t \mapsto \frac{1}{g(x_j)} \mathbb{1}_{\{t \leq \frac{f(x_j)}{g(x_j)}\}} > \\
&= < t \mapsto \sum_{i=1}^N \alpha_i \frac{1}{g(x_i)} \mathbb{1}_{\{t \leq \frac{f(x_i)}{g(x_i)}\}} \,\Bigg|\, \sum_{j=1}^N \alpha_j t \mapsto \frac{1}{g(x_j)} \mathbb{1}_{\{t \leq \frac{f(x_j)}{g(x_j)}\}} > \\
&= \left\| t \mapsto \sum_{i=1}^N \alpha_i \frac{1}{g(x_i)} \mathbb{1}_{\{t \leq \frac{f(x_i)}{g(x_i)}\}} \right\|_{L^2}^2 \\
&\geq 0
\end{aligned}$$

   Hence, the kernel K is positive definite.

5. We consider a non-empty finite set E and we define $\forall A, B \subset E,\ K(A, B) = \frac{A \cap B}{A \cup B}$ with the convention $\frac{0}{0} = 0$. We note $n = |E|$.

   We start by doing to useful remarks for what follows.

   - Remark 1: We know that $\forall x \in [0, 1[,\ \sum_{k=0}^{+\infty} x^k = \frac{1}{1-x}$ as the sum of a geometric sequence.
   - Remark 2: If we consider $A, B \subset E$ with A or/and B different from $\emptyset$, $n = |E| > |A^c \cap B^c|$ (where $A^c = E \backslash A$). With the first remark we are allowed to write in this case :

$$\sum_{k=0}^{+\infty} \left(\frac{|A^c \cap B^c|}{n}\right)^k = \frac{1}{1 - \frac{|A^c \cap B^c|}{n}} \qquad (*)$$

Please note that if A or B is the empty set, then $K(A,B) = 0$. Thus, **without loss of generality, we will suppose from now that the subsets of E considered are non empty**. Thus, we have :

$$\begin{aligned} K(A,B) &= \frac{|A \cap B|}{|A \cup B|} \\ &= \frac{|A \cap B|}{n - |A^c \cap B^c|}, \text{ since } (A \cup B)^c = A^c \cap B^c. \\ &= \frac{|A \cap B|}{n} \times \frac{1}{1 - \frac{|A^c \cap B^c|}{n}} \\ &= \frac{|A \cap B|}{n} \times \sum_{k=0}^{+\infty} \left(\frac{|A^c \cap B^c|}{n} \right)^k \end{aligned}$$

We define the functions :

$$\begin{aligned} K_1 : \mathcal{P}(E) \times \mathcal{P}(E) &\to \mathbb{R} \\ (C,D) &\mapsto \frac{|C \cap D|}{n} \end{aligned}$$

and

$$\begin{aligned} K_2 : \mathcal{P}(E) \times \mathcal{P}(E) &\to \mathbb{R} \\ (C,D) &\mapsto \frac{|C^c \cap D^c|}{n} \end{aligned}$$

K_1 and K_2 are two positive definite kernels. In order to justify this claim, we endow $(E, \mathcal{P}(E))$ with the uniform probability distribution denoted $\mathbb{P}$. Then, for all $(C,D) \in \mathcal{P}(E)^2$,

$$K_1(C,D) = \frac{|C \cap D|}{n} = \mathbb{E}\left[\mathbb{1}_C \mathbb{1}_D\right] =< \mathbb{1}_C \mid \mathbb{1}_D > \quad (*)$$

where $< . \mid . >$ denotes the usual scalar product for L^2 random variables.

Thanks to the Aronszajn's theorem, we deduce from $(*)$ that K_1 is a positive definite kernel.

The same argument also holds for K_2 since $K_2(C,D) =< \mathbb{1}_{C^c} \mid \mathbb{1}_{D^c} >$. Thus, K_2 is also a positive definite kernel.

We can now prove that K is a positive definite kernel. Indeed :

- Using the theorem (3) and since K_2 is a p.d. kernel, we have that for all $k \in \mathbb{N}$, K_2^k is a p.d. kernel.
- Then, using the previous item and the theorem (2), we get the for all $N \in \mathbb{N}$, $\sum_{k=1}^{N} K_2^k$ is a p.d. kernel.
- Using the previous item, the theorem (1) and the equality $(*)$, we know that the kernel $K_3 := \sum_{k=0}^{+\infty} K_2^k : (A,B) \mapsto \sum_{k=0}^{+\infty} \left(\frac{|A^c \cap B^c|}{n} \right)^k = \frac{1}{1 - \frac{|A^c \cap B^c|}{n}}$ is a p.d. kernel.
- Finally, since K_1 and K_3 are p.d. kernels and since $K = K_1 K_3$ (hadamard product), we have using the theorem (3) that K is p.d. kernel.

Hence, K is a positive definite kernel.

Exercise 2

1. K_1 and K_2 are two positive kernels and α, β are two positive scalars. We deduce that αK_1 and βK_2 are two positive kernels (as the multiplication by a positive scalar of a positive kernel). Then, we have that $\alpha K_1 + \beta K_2$ is a positive kernel as the sum of two positive kernels (using theorem (2)).

   We denote $\mathcal{H}_1$ (resp. $\mathcal{H}_2$) the RKHS associated with the p.d. kernel K_1 (resp. K_2). We note $< .|. >_1$ (resp. $< .|. >_2$) the scalar product associated with $\mathcal{H}_1$ (resp. $\mathcal{H}_2$).

   - First we look at the topology of $\mathcal{H}_1 + \mathcal{H}_2$. We denote $E = \mathcal{H}_1 \times \mathcal{H}_2$. This set is a Hilbert space if we equip it with the norm $||.||_E : (f_1, f_2) \mapsto \sqrt{\frac{1}{\alpha}||f_1||_1^2 + \frac{1}{\beta}||f_2||_2^2}$,
     We want to compare the topologies of $\mathcal{H}_1 + \mathcal{H}_2$ and E. A direct link between these spaces is the natural surjection

     $$s : E \to \mathcal{H}_1 + \mathcal{H}_2$$
     $$(f_1, f_2) \mapsto f_1 + f_2$$

     We are going to try to make s injective. In order to do so, let's consider $N = s^{-1}(\{0\})$. We will begin by proving that N is a closed subset of E:

     Let $(f_n, -f_n)$ be a sequence of elements of N converging in E to (f, g). By definition of the norm $||.||_E$, $(f_n)_{n \geq 1}$ converges in $\mathcal{H}_1$ to f and $(-f_n)_{n \geq 1}$ converges in $\mathcal{H}_2$ to g. Since convergence in a RKHS implies ponctual convergence, we will have $f = -g$ an therefore $(f, g) \in N$. $\underline{N \text{ is therefore a closed subset of } E}$.

     Since N is closed, E is equal to the direct sum of N and its orthogonal complement $N^\perp$. The restriction $\tilde{s}$ of s to $N^\perp$ will therefore be a bijection.
     Now that we have a linear bijection, we can equip $\mathcal{H} = \mathcal{H}_1 + \mathcal{H}_2$ with an Hilbertian structure inherited from E. With the norm $||.||_\mathcal{H} : f \mapsto ||\tilde{s}^{-1}(f)||_E$, $\mathcal{H}_1 + \mathcal{H}_2$ is indeed a Hilbert space.
   - It is obvious that for all $x \in \mathcal{X}$, $K_x = K(x, .) = \alpha K_1(x, .) + \beta K_2(x, .)$ belongs to $\mathcal{H} = \mathcal{H}_1 + \mathcal{H}_2$ (since $K_1(x, .) \in \mathcal{H}_1$ and $K_2(x, .) \in \mathcal{H}_2$ by the definition of the reproducing kernel of a RKHS).
   - In fact, to prove that $\mathcal{H} = \mathcal{H}_1 + \mathcal{H}_2$ endowed with the norm we just defined is the RKHS of $\alpha K_1 + \beta K_2$, we still need to prove the reproducing property: let $x \in \mathcal{X}$ and $f \in \mathcal{H} = \mathcal{H}_1 + \mathcal{H}_2$. We can write $f = \tilde{s}(f_1, f_2)$ and $K_x = \tilde{s}(A_x, B_x)$ where (f_1, f_2) and (A_x, B_x) live in $N^\perp$. Thus,

     $$< f, K_x >_{\mathcal{H}_1 + \mathcal{H}_2} = < (f_1, f_2), (A_x, B_x) >_E = < (f_1, f_2), (\alpha K_{1x}, \beta K_{2x}) + (A_x - \alpha K_{1x}, B_x - \beta K_{2x}) >_E$$

     but, since $s(A_x - \alpha K_{1x}, B_x - \beta K_{2x}) = A_x - \alpha K_{1x} + B_x - \beta K_{2x} = K_x - K_x = 0$, we have that the vector $(A_x - \alpha K_{1x}, B_x - \beta K_{2x})$ belongs to N. Therefore, it is orthogonal to every element in $N^\perp$, and in particular to (f_1, f_2). Consequently, $< f, K_x >_\mathcal{H} = < (f_1, f_2), (\alpha K_1(x, .), \beta K_2(x, .)) >_E = f_1(x) + f_2(x) = f(x)$
     and the reproducing property is true.

   $\mathcal{H} = \mathcal{H}_1 + \mathcal{H}_2$ is therefore the RKHS of $\alpha K_1 + \beta K_2$.

2. We consider $\psi : \mathcal{X} \to \mathcal{F}$ where $\mathcal{F}$ is a Hilbert space. The kernel

   $$K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$$
   $$(x, x') \mapsto < \psi(x), \psi(x') >_\mathcal{F}$$

   is positive definite as a direct consequence of the Aronzsajn's theorem.

   We are now going to show that the RKHS associated to positive definite kernel K is the image of the operator T defined by :

   $$\forall f \in \mathcal{F}, \quad Tf : \mathcal{X} \to \mathbb{R}$$
   $$x \mapsto (Tf)(x) := < f, \psi(x) >_\mathcal{F}$$

   First, we recall a result seen during the class which will be the cornerstone of the proof :

Theorem 4. *Any kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ positive definite is a reproducing kernel.*

Useful elements of the proof for what follows :

We define $\mathcal{H}_0$ the vector space spanned by the functions K_x for $x \in \mathcal{X}$. The scalar product on $\mathcal{H}_0$ is given by :

$$< f, g >_{\mathcal{H}_0} = \sum_{i,j} a_i b_j K(x_i, x_j)$$

where we have decomposed f and g as $f = \sum_i a_i K_{x_i}$ and $g = \sum_j b_j K_{x_j}$ (we proved in class that the definition is independent of the decomposition). Then, the RKHS $\mathcal{H}_K$ related to the kernel K is obtained by taking the completion of $\mathcal{H}_0$ to a Hilbert space.

Now, we have all the tools to prove our claim :

$$\mathcal{H}_K = Im(T) = \{Tf \ , \ f \in \mathcal{F}\}.$$

- $\mathbf{\mathcal{H}_0 \subset Im(T)}$.
  Indeed, let $x \in \mathcal{X}$. For all $y \in \mathcal{X}$, $K_x(y) =< \psi(x), \psi(y) >_{\mathcal{F}} = (T\psi(x))(y)$. So $Im(T)$ contains all the functions K_x for $x \in \mathcal{X}$. Since $Im(T)$ is a linear space, then linear span of $\{K_x, \ x \in \mathcal{X}\}$, that is $\mathcal{H}_0$, will be in $Im(T)$.
- $\mathbf{T : Span(\psi(x), \ x \in \mathcal{X}) \to \mathcal{H}_0}$ **is isometric.**
  Since for all $x \in \mathcal{X}$, $T\psi(x) = K_x$, we have $T\left(\sum_x \alpha_x \psi(x)\right) = \sum \alpha_x K_x$. Hence,

$$\begin{aligned}
< T\left(\sum_x \alpha_x \psi(x)\right), T\left(\sum_y \beta_y \psi(y)\right) >_{\mathcal{H}_0} &=< \sum_x \alpha_x K_x, \sum_y \beta_y K_y >_{\mathcal{H}_0} \\
&= \sum_{x,y} \alpha_x \beta_y K(x, y) \text{ using the construction of } < .,. >_{\mathcal{H}_0} \text{ recalled in theorem 4} \\
&= \sum_{x,y} \alpha_x \beta_y < \psi(x), \psi(y) >_{\mathcal{F}} \\
&=< \sum_x \alpha_x \psi(x), \sum_y \beta_y \psi(y) >_{\mathcal{F}} .
\end{aligned}$$

  This proves that $T : Span(\psi(x), \ x \in \mathcal{X}) \to \mathcal{H}_0$ is isometric.

  Clearly, $T\left(Span(\psi(x), \ x \in \mathcal{X})\right) = \mathcal{H}_0$.
- $\mathbf{\mathcal{F} = \ker(T) \bigoplus \ker(T)^\perp}$ **with** $\ker(T)^\perp = \overline{Span(\psi(x), \ x \in \mathcal{X})}$.
  - Let $f \in \ker(T)$.
    So, $Tf = 0$ ie $(Tf)(x) =< f, \psi(x) >_{\mathcal{F}} = 0 \ \forall x \in \mathcal{X}$. Since T is linear, this means that $f \perp Span(\psi(x), \ x \in \mathcal{X})$, i.e.

$$\ker(T) \subset Span(\psi(x), \ x \in \mathcal{X})^\perp.$$

  - Let $f \in Span(\psi(x), \ x \in \mathcal{X})^\perp = \{\psi(x), x \in \mathcal{X}\}^\perp$.
    Then, for all x, $0 =< f, \psi(x) >_{\mathcal{F}} = (Tf)(x) \implies Tf = 0$ i.e. $f \in \ker(T)$. Hence :

$$Span(\psi(x), \ x \in \mathcal{X})^\perp \subset \ker(T).$$

    This proves that :

$$\ker(\mathbf{T}) = \mathbf{Span}(\psi(\mathbf{x}), \ \mathbf{x} \in \mathcal{X})^\perp.$$

– By the previous item,

$$\ker(T)^{\perp} = \left(Span(\psi(x),\ x \in \mathcal{X})^{\perp}\right)^{\perp} = \overline{Span(\psi(x),\ x \in \mathcal{X})}$$

This shows in particular that $\ker(T)^{\perp}$ is closed.
We are able to write

$$\mathcal{F} = \ker(T) \bigoplus \ker(T)^{\perp}.$$

- Since $T : Span(\psi(x),\ x \in \mathcal{X}) \to \mathcal{H}_0$ is isometric and surjective, and since $\mathcal{H}_0$ is dense in $\mathcal{H}_K$ (by construction: see theorem (4)), it follows that $T : \underbrace{\overline{Span(\psi(x),\ x \in \mathcal{X})}}_{=\ker(T)^{\perp}} \to \overline{\mathcal{H}_0} = \mathcal{H}_K$ is surjective ((∗), see below for further justification). Hence, we have :

$$\mathcal{H}_K = T(\ker(T)^{\perp}) = T(\ker(T) \bigoplus \ker(T)^{\perp}) = T(\mathcal{F}) = Im(T).$$

Comments

This result of the question 2 allows us to have another point of view on a RKHS. Indeed, we have shown that for a kernel K defined by a feature map ψ, the RKHS related to K is :

$$\mathcal{H}_K = Im(T) = \{x \mapsto < f, \psi(x) >_{\mathcal{F}} \text{ such that } f \in \mathcal{F}\}.$$

This representation implies that the elements of the RKHS are inner products of elements in the feature space and can accordingly be seen as **hyperplanes**.

Further justification for (∗).

$T : Span(\psi(x),\ x \in \mathcal{X}) \to \mathcal{H}_0$ is isometric, and linear. We can thus apply the theorem to extend linear function uniformly continuous (here, T is uniformly continuous because isometric). So, we can extend T as a linear isometry on $\overline{Span(\psi(x),\ x \in \mathcal{X})}$. We still call this new function T. The miracle is that this function T is in fact surjective in H_K.

Indeed, let $g \in \mathcal{H}_K$. Since $\mathcal{H}_0$ is dense in $\mathcal{H}_K$, there exists a sequence $(g_n)_n$ in $\mathcal{H}_0$ such that $||g_n - g||_{\mathcal{H}_0} \to 0$. Since $T : Span(\psi(x),\ x \in \mathcal{X}) \to \mathcal{H}_0$ is surjective, for all $n \in \mathbb{N}$, there exists $f_n \in \mathcal{F}$ such that $Tf_n = g_n$. Since $(g_n)_n$ is convergent, it is in particular a Cauchy sequence and the fact that T is isometric gives us that for all $n, m \in \mathbb{N}$,

$$||g_m - g_n||_{\mathcal{H}_0} = ||Tf_m - Tf_n||_{\mathcal{H}_0} = ||T(f_m - f_n)||_{\mathcal{H}_0} = ||f_m - f_n||_{\mathcal{F}}.$$

Hence, $(f_n)_n$ is a Cauchy sequence in the Hilbert space $\mathcal{F}$. Hence, it converges to some $f \in \mathcal{F}$. But, since $(f_n)_n \in Span(\psi(x),\ x \in \mathcal{X})^{\mathbb{N}}$, we have that $f \in \overline{Span(\psi(x),\ x \in \mathcal{X})}$. Hence, $g \in \mathcal{H}_K$ admits the preimage f by T which belongs to $\overline{Span(\psi(x),\ x \in \mathcal{X})}$.

Exercise 3

1. We recall a theorem studied in class :

   Theorem 5. *The Hilbert space $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ is a RKHS if and only if for any $x \in \mathcal{X}$, the mapping*

$$\begin{aligned} F_x : \mathcal{H} &\to \mathbb{R} \\ f &\mapsto f(x) \end{aligned}$$

   is continuous.

   In our case, $\mathcal{H} = \{f : [0,1] \to \mathbb{R} \text{ absolutely continuous },\ f' \in L^2([0,1]),\ f(0) = 0\}$ endowed with the bilinear form : $\forall f, g \in \mathcal{H}, \quad < f, g >_{\mathcal{H}} = \int_0^1 f'(u)g'(u)du$.

   - H is a prehilbert space of functions
     - $\mathcal{H}$ is a vector space of functions and $< ., . >_{\mathcal{H}}$ is a bilinear form that satisfies $< f, f >_{\mathcal{H}} \geq 0$.
     - f absolutely continuous on $[0,1]$ implies differentiable almost everywhere and $\forall x \in [0,1], \quad f(x) = f(0) + \int_0^x f'(u)du$. Hence:

$$\begin{aligned} \forall f \in \mathcal{H},\ \forall x \in [0,1], \quad |f(x)| = |f(x) - \underbrace{f(0)}_{=0 \text{ since } f\in\mathcal{H}}| = |\int_0^x f'(u)du| \leq \int_0^x \underbrace{|f'(u)|}_{\geq 0} du \leq \int_0^1 |f'(u)|du \\ = \int_0^1 \sqrt{|f'(u)|^2}du \leq \sqrt{\int_0^1 |f'(u)|^2 du} = < f, f >_{\mathcal{H}}^{1/2} \end{aligned} \tag{1}$$

     where the last inequality is obtained by using the Jensen inequality with the concave function $t \mapsto \sqrt{t}$. Therefore $< f, f >_{\mathcal{H}} = 0 \implies f = 0$, showing that $< ., . >_{\mathcal{H}}$ is an inner product. Thus, $\mathcal{H}$ is a preHilbert space.

   - H is a Hilbert space

     Let $(f_n)_{n\in\mathbb{N}}$ a Cauchy sequence of $\mathcal{H}$. Then, $(f'_n)_{n\in\mathbb{N}}$ is a Cauchy sequence of $L^2([0,1])$ (by definition of the norm on $\mathcal{H}$), and thus convergences to some $g \in L^2([0,1])$ for the norm $||.||_{L^2}$ (by completeness).

     Using the inequality (1), for all $x \in [0,1]$, $(f_n(x))_{n\in\mathbb{N}}$ is a Cauchy sequence of $\mathbb{R}$ which is complete and thus converges to some $f(x)$. Moreover,

$$f(x) = \lim_{n\to+\infty} f_n(x) = \lim_{n\to+\infty} \int_0^x f'_n(u)du = \int_0^x g(u)du$$

     where we have used an interversion between limit and integral which is possible thanks to the L^2 convergence of $(f'_n)_n$ to g. This shows that f is absolutely continuous and $f' = g$ almost everywhere, in particular, $f' \in L^2([0,1])$.

     Finally, $f(0) = \lim_{n\to+\infty} f_n(0) = 0$. Therefore, $f \in \mathcal{H}$ and $\lim_{n\to+\infty} ||f_n - f||_{\mathcal{H}} = ||f'_n - g||_{L^2} = 0$.

     We have proved then $\mathcal{H}$ is a Hilbert space.

   - H is a RKHS

     Let $x \in [0,1]$. For all $f \in \mathcal{H}$,

$$|F_x(f)| = |f(x)| \leq ||f||_{\mathcal{H}} \text{ using (1).}$$

     Since the mapping F_x is linear, the above inequality proves that for all $x \in \mathcal{X}$, F_x is continuous. We deduce that $\mathcal{H}$ is a RKHS with the theorem 5.

- Reproducing kernel of H
  Consider the function

$$K : [0,1] \times [0,1] \to \mathbb{R}$$
$$(x,y) \mapsto \min(x,y) = \begin{cases} y & \text{if } y < x \\ x & \text{if } x \leq y \end{cases}$$

  For all $x \in [0,1]$, the function $K_x : t \mapsto K(x,t)$ belongs to $\mathcal{H}$ because :
  - it is absolutely continuous on $[0,1]$ since :
    * K_x has derivative almost everywhere (except in x)
    * K'_x is Lebsgue integrable
    * $\forall t \in [0,1]$, $K_x(t) = K_x(0) + \int_0^t K'_x(u)du$
  - $K'_x = \mathbb{1}_{[0,x]}$ which belongs to $L^2([0,1])$
  - and we finally have $K_x(0) = 0$.

  Moreover for all $x \in [0,1]$ and for all $f \in \mathcal{H}$, $< f, K_x >= \int_0^1 f'(u)K'_x(u)du = \int_0^1 f'(u)\mathbb{1}_{[0,x]}du = \int_0^x f'(u) = f(x) - \underbrace{f(0)}_{=0} = f(x)$. So the reproducing property holds.

Hence, K is the reproducing kernel of the RKHS $\mathcal{H}$.

2. We consider now $\mathcal{H} = \{f : [0,1] \to \mathbb{R} \text{ absolutely continuous },\ f' \in L^2([0,1]),\ f(0) = f(1) = 0\}$ endowed with the bilinear form : $\forall f, g \in \mathcal{H}, \quad < f, g >_{\mathcal{H}} = \int_0^1 f'(u)g'(u)du$.

- H is a prehilbert space of functions
  $\mathcal{H}$ is a vector space of functions and $< .,. >_{\mathcal{H}}$ is an inner product thanks to the previous question. Thus, $\mathcal{H}$ is a preHilbert space.
- H is a Hilbert space
  Let $(f_n)_{n\in\mathbb{N}}$ a Cauchy sequence of $\mathcal{H}$. Then, $(f'_n)_{n\in\mathbb{N}}$ is a Cauchy sequence of $L^2([0,1])$ (by definition of the norm on $\mathcal{H}$), and thus convergences to some $g \in L^2([0,1])$.

  Using the inequality (1), for all $x \in [0,1]$, $(f_n(x))_{n\in\mathbb{N}}$ is a Cauchy sequence of $\mathbb{R}$ which is complete and thus converges to some $f(x)$. Moreover,

$$f(x) = \lim_{n\to+\infty} f_n(x) = \lim_{n\to+\infty} \int_0^x f'_n(u)du = \int_0^x g(u)du$$

  where we have used an interversion between limit and integral which is possible thanks to the L^2 convergence of $(f'_n)_n$ to g. This shows that that f is absolutely continuous and $f' = g$ almost everywhere, in particular, $f' \in L^2([0,1])$.

  Finally, $f(0) = \lim_{n\to+\infty} f_n(0) = 0$ and $f(1) = \lim_{n\to+\infty} f_n(1) = 0$. Therefore, $f \in \mathcal{H}$ and $\lim_{n\to+\infty} ||f_n - f||_{\mathcal{H}} = ||f'_n - g||_{L^2} = 0$.
- H is a RKHS
  The computations derived in the previous question to show that the mapping F_x is continuous for all $x \in [0,1]$ still hold by definition of $\mathcal{H}$ (which is included in the Hilbert space studied in the previous question). Thus, using the theorem 5, we have that $\mathcal{H}$ is a RKHS.
- Reproducing kernel of H
  Consider the function

$$K : [0,1] \times [0,1] \to \mathbb{R}$$
$$(x,y) \mapsto \begin{cases} (1-x)y & \text{if } y < x \\ -x(y-x) + (1-x)x & \text{if } x \leq y \end{cases}$$

  For all $x \in [0,1]$, the function $K_x : t \mapsto K(x,t)$ belongs to $\mathcal{H}$ because :

- it is absolutely continuous on $[0,1]$ since :
  - K_x has derivative almost everywhere (except in x)
  - K'_x is Lebsgue integrable
  - $\forall t \in [0,1]$, $K_x(t) = K_x(0) + \int_0^t K'_x(u)du$
- $K'_x = (1-x)\mathbb{1}_{[0,x]} - x\mathbb{1}_{[x,1]}$ which belongs to $L^2([0,1])$
- and we finally have $K_x(0) = K_x(1) = 0$.

Moreover for all $x \in [0,1]$ and for all $f \in \mathcal{H}$, $< f, K_x >_{\mathcal{H}} = \int_0^1 f'(u)K'_x(u)du = \int_0^x f'(u)(1-x)du - \int_x^1 f'(u)xdu = (1-x)(f(x) - \underbrace{f(0)}_{=0}) - x(\underbrace{f(1)}_{=0} - f(x)) = f(x)$. So the reproducing property holds.

Hence, K is the reproducing kernel of the RKHS $\mathcal{H}$.

3. We consider now $\mathcal{H} = \{f : [0,1] \to \mathbb{R}$ absolutely continuous , $f' \in L^2([0,1])$, $f(0) = f(1) = 0\}$ endowed with the bilinear form : $\forall f, g \in \mathcal{H}, \quad < f, g >_{\mathcal{H}} = \int_0^1 (f(u)g(u) + f'(u)g'(u))du$.

- H is a prehilbert space of functions
  - $\mathcal{H}$ is a vector space of functions and $< .,. >_{\mathcal{H}}$ is a bilinear form that satisfies $< f, f >_{\mathcal{H}} \geq 0$.
  - f absolutely continuous on $[0,1]$ implies differentiable almost everywhere and $\forall x \in [0,1], \quad f(x) = f(0) + \int_0^x f'(u)du$. Hence:

$$\forall f \in \mathcal{H}, \quad |f(x)| = |f(x) - \underbrace{f(0)}_{=0 \text{ since } f \in \mathcal{H}}| = |\int_0^x f'(u)du| \leq \int_0^x \underbrace{|f'(u)|}_{\geq 0} du \leq \int_0^1 |f'(u)|du$$
$$= \int_0^1 \sqrt{|f'(u)|^2}du \underbrace{\leq}_{\text{since } \sqrt{\cdot} \text{ is an increasing function}} \int_0^1 \sqrt{|f'(u)|^2 + |f(u)|^2}du$$
$$\leq \sqrt{\int_0^1 |f'(u)|^2 + |f(u)|^2 du} = < f, f >_{\mathcal{H}}^{1/2} \quad (2)$$

where the last inequality is obtained by using the Jensen inequality with the concave function $t \mapsto \sqrt{t}$. Therefore $< f, f >_{\mathcal{H}} = 0 \implies f = 0$, showing that $< .,. >_{\mathcal{H}}$ is an inner product. Thus, $\mathcal{H}$ is a preHilbert space.

- H is a Hilbert space

  Let $(f_n)_{n\in\mathbb{N}}$ a Cauchy sequence of $\mathcal{H}$.
  - $(f_n)_{n\in\mathbb{N}}$ and $(f'_n)_{n\in\mathbb{N}}$ are Cauchy sequences in $L^2([0,1])$

    $(f_n)_{n\in\mathbb{N}}$ (resp. $(f'_n)_{n\in\mathbb{N}}$) is a Cauchy sequence of $L^2([0,1])$ (by definition of the norm on $\mathcal{H}$), and thus convergences to some $g_0 \in L^2([0,1])$ (resp. $g_1 \in L^2([0,1])$) .
  - **Theorem** : Convergence in $L^2([0,1]) \implies$ Convergence in $\mathcal{D}'([0,1])$

    Let $\phi \in \mathcal{D}([0,1])$ with compact support K_ϕ and $(h_n)_{n\in\mathbb{N}}$ a sequence of $L^2([0,1])$ converging to $h \in L^2([0,1])$. Since $h, h_n \in L^1_{loc}([0,1])$, we can consider the distributions induced by these functions. Moreover, the Cauchy Scharwz inequality gives us :

$$| < h, h_n, \phi >_{\mathcal{D}',\mathcal{D}} | = \left| \int_{[0,1]} (h - h_n)\phi \right| \leq ||h - h_n|_{L^2}||\phi||_{L^2}.$$

    Thus, $(h_n)_n$ converges to h in the distribution sens.
  - $g'_0 = g_1$ in the distribution sens and then in L^2.

    Using the previous item, we get that $f_n \to g_0$ in $\mathcal{D}'([0,1])$ and $f'_n \to g_1$ in $\mathcal{D}'([0,1])$. From $f_n \to g_0$ in $\mathcal{D}'([0,1])$, we deduce that $f'_n \to g'_0$ in $\mathcal{D}'([0,1])$. Using the uniqueness of the limit in $\mathcal{D}'([0,1])$, we have $g'_0 = g_1$ in the distribution sens. Since $g_1 \in L^2([0,1])$, we can deduce that $g'_0 \in L^2([0,1])$, and that the equality $g'_0 = g_1$ is also true in $L^2([0,1])$.

  We have shown that $f_n \to g_0$ and $f'_n \to g'_0$ in L^2. Thus, $f_n \to g_0$ in $\mathcal{H}$. We only need to show that g_0 belongs to $\mathcal{H}$, which is true since :

– The inequality (2) gives that convergence in $\mathcal{H}$ implies pointwise convergence. Thus, $g_0(0) = \lim_{n\to+\infty} f_n(0) = 0$ and $g_0'(1) = \lim_{n\to+\infty} f_n(1) = 0$.

– We have already shown that $g_0' = g_1 \in L^2([0,1])$.

– Finally, g_0 is absolutely continuous since $g_0(x) = \int_0^x g_0'(u)du$.

- H is a RKHS

Let $x \in [0,1]$. For all $f \in \mathcal{H}$,

$$|F_x(f)| = |f(x)| \leq ||f||_{\mathcal{H}} \text{ using (2).}$$

Thus, using the theorem 5, we have that $\mathcal{H}$ is a RKHS.

- Reproducing kernel of H

Consider the function

$$K : [0,1]\times[0,1] \to \mathbb{R}$$
$$(x,y) \mapsto \begin{cases} \left(t \mapsto e^{-t} + (1-e^{-x})\frac{sh(t)}{sh(x)} - 1\right)'(y) & \text{if } y < x \\ 0 & \text{if } x \leq y \end{cases}$$

i.e.

$$K : [0,1]\times[0,1] \to \mathbb{R}$$
$$(x,y) \mapsto \begin{cases} -e^{-y} + (1-e^{-x})\frac{ch(y)}{sh(x)} & \text{if } y < x \\ 0 & \text{if } x \leq y \end{cases}$$

For all $x \in [0,1]$, the function $K_x : t \mapsto K(x,t)$ belongs to $\mathcal{H}$ because :

– it is absolutely continuous on $[0,1]$ since :
  - ∗ K_x has derivative almost everywhere (except in x)
  - ∗ K_x' is Lebsgue integrable
  - ∗ $\forall t \in [0,1]$, $K_x(t) = K_x(0) + \int_0^t K_x'(u)du$

– $\forall y \in [0,1]$, $K_x'(y) = \left(-\sin(y) + \frac{1-\cos(x)}{\sin(x)}\cos(y)\right)\mathbb{1}_{[0,x]}(y)$ which belongs to $L^2([0,1])$

– and we finally have $K_x(0) = K_x(1) = 0$.

Please note that the function K_x has been built such that $\mathcal{P}(K_x) : y \mapsto \int_0^y K_x(t)dt$ is a solution of the equation $g''(y) - g(y) = 1$ on $[0,x]$ with the conditions $g(0) = 0$ and $g(x) = 0$ (∗). Then for all $x \in [0,1]$ and for all $f \in \mathcal{H}$,

$$
\begin{aligned}
< f, K_x >_{\mathcal{H}} &= \int_0^1 K_x(u)f(u) + f'(u)K_x'(u)du \\
&= \int_0^1 K_x(u)f(u)du + \int_0^1 f'(u)K_x'(u)du, \text{ and using an IPP in the first integrale we get} \\
&= \underbrace{\left[\int_0^u K_x(t)dt f(u)\right]_0^1}_{=0 \text{ since } f(0)=f(1)=0} - \int_0^x f'(u)\int_0^u K_x(t)dtdu + \int_0^x f'(u) \underbrace{K_x'(u)}_{=\mathcal{P}(K_x)''(u)} du \\
&= \int_0^x f'(u) \underbrace{\left(\mathcal{P}(K_x)''(u) - \mathcal{P}(K_x)(u)\right)}_{=1 \text{ using } (*)} du \\
&= f(x) - \underbrace{f(0)}_{=0 \text{ since } f\in\mathcal{H}} \\
&= f(x)
\end{aligned}
$$

So the reproducing property holds.

Hence, K is the reproducing kernel of the RKHS $\mathcal{H}$.

Exercise 4: Duality

1. We are considering the following optimization problem

$$\min_{f\in\mathcal{H}_K} \frac{1}{n}\sum_{i=1}^n l_{y_i}(f(x_i)) \text{ such that } ||f||_{\mathcal{H}_K} \leq B.$$

which is equivalent to

$$\min_{f\in\mathcal{H}_K} \frac{1}{n}\sum_{i=1}^n l_{y_i}(f(x_i)) \text{ such that } ||f||^2_{\mathcal{H}_K} \leq B^2. \tag{3}$$

Dualizing the constraint involved in (3), we get that the problem (3) is equivalent to :

$$\min_{f\in\mathcal{H}_K} \sup_{\lambda\geq 0} \frac{1}{n}\sum_{i=1}^n l_{y_i}(f(x_i)) + \lambda(||f||^2_{\mathcal{H}_K} - B^2). \tag{4}$$

Since the function l_y is convex for all $y \in \{-1, +1\}$, we deduce that the optimization problem (4) is a convex optimization problem and qualification holds (since there is no constraint). Thus, **strong duality holds**. Thus, the problem (4) is equivalent to

$$\sup_{\lambda\geq 0} \min_{f\in\mathcal{H}_K} \frac{1}{n}\sum_{i=1}^n l_{y_i}(f(x_i)) + \lambda(||f||^2_{\mathcal{H}_K} - B^2).$$

The KKT conditions give us that there exists $\lambda^* \geq 0$ such that (4) is equivalent to

$$\min_{f\in\mathcal{H}_K} \frac{1}{n}\sum_{i=1}^n l_{y_i}(f(x_i)) + \lambda^*(||f||^2_{\mathcal{H}_K} - B^2) = \min_{f\in\mathcal{H}_K} \frac{1}{n}\Psi(f(x_1), \dots, f(x_n), ||f||^2_{\mathcal{H}_K}), \tag{5}$$

where $\Psi : \mathbb{R}^{n+1} \to \mathbb{R}$ is a function of $n+1$ variables, strictly increasing with respect to the last variable. Since K is the reproducing kernel of the RKHS $\mathcal{H}_K$, we have thanks to the **representer theorem** that a solution f of the optimization problem (5) can be written of the form :

$$f(x) = \sum_{i=1}^{n} \alpha_i K_{x_i}(x), \quad (\alpha_i)_{i=1}^n \in \mathbb{R}^n.$$

Denoting $\mathbf{K}$ the matrix of size $n \times n$: $(K(x_i, x_j))_{1 \leq i,j \leq n}$, we have that :

- $\forall i \in [\![1, n]\!]$, $f(x_i) = (\mathbf{K}\alpha)_i$ where α denote the vector $(\alpha_i)_{i=1}^n$.
- $||f||^2_{\mathcal{H}_K} = \alpha^T \mathbf{K} \alpha$.

The optimization problem (5) is hence equivalent to

$$\min_{\alpha \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} l_{y_i}((\mathbf{K}\alpha)_i) + \lambda^*(\alpha^T \mathbf{K}\alpha - B^2) = \min_{\alpha \in \mathbb{R}^n} R(\mathbf{K}\alpha) + \lambda^*(\alpha^T \mathbf{K}\alpha - B^2), \tag{6}$$

where $R(z) = \frac{1}{n} \sum_{i=1}^{n} l_{y_i}(z_i), \quad \forall z \in \mathbb{R}^n$.

2. We compute the Fenchel-Legendre transform of R. Let $z \in R^n$,

$$\begin{aligned}
R^*(z) &= \sup_{x \in \mathbb{R}^n} < x, z > - R(x) \\
&= \sup_{x \in \mathbb{R}^n} < x, z > - \frac{1}{n} \sum_{i=1}^{n} l_{y_i}(x_i), \text{ here we remark that the problem is separable} \\
&= \sum_{i=1}^{n} \left(\sup_{x_i \in \mathbb{R}} \left[x_i z_i - \frac{1}{n} l_{y_i}(x_i) \right] \right) \\
&= \sum_{i=1}^{n} \frac{1}{n} l^*_{y_i}(n z_i).
\end{aligned}$$

3. We add the slack variable $u = \mathbf{K}\alpha$ in the optimization problem (6). The problem (3) can thus be written as :

$$\min_{\alpha \in \mathbb{R}^n, u \in \mathbb{R}^n} R(u) + \lambda^*(\alpha^T \mathbf{K}\alpha - B^2) \text{ such that } u = \mathbf{K}\alpha. \tag{7}$$

The dual of the problem (7) is :

$$\sup_{\mu \in \mathbb{R}^n} \quad \min_{\alpha \in \mathbb{R}^n, u \in \mathbb{R}^n} R(u) + \lambda^*(\alpha^T \mathbf{K}\alpha - B^2) + \mu^T(\mathbf{K}\alpha - u)$$

which is equivalent to

$$\sup_{\mu \in \mathbb{R}^n} \quad \left(\min_{\alpha \in \mathbb{R}^n} \left[\lambda^*(\alpha^T \mathbf{K}\alpha - B^2) + \mu^T \mathbf{K}\alpha \right] + \min_{u \in \mathbb{R}^n} \left[R(u) - \mu^T u \right] \right)$$

- Since the minimization problem in α is an unconstrained convex optimization problem, an optimal solution is given by setting the gradient to zero which leads to $2\lambda^* \mathbf{K}\alpha = \mathbf{K}\mu$. Thus, all the optimal solution have the form $\alpha = \frac{\mu}{2\lambda^*} + \epsilon$ with $\epsilon \in Ker(\mathbf{K})$, but all those solutions lead to the same function f since $\mathbf{K}(\frac{\mu}{2\lambda^*} + \epsilon) = \mathbf{K}\frac{\mu}{2\lambda^*}$.
- $\min_{u \in \mathbb{R}^n} \left[R(u) - \mu^T u \right] = - \sup_{u \in \mathbb{R}^n} \left[\mu^T u - R(u) \right] = -R^*(\mu)$.

We deduce that the above optimization problem is equivalent to

$$\sup_{\mu \in \mathbb{R}^n} \frac{1}{4\lambda^*} \mu^T \mathbf{K}\mu + \frac{1}{2\lambda^*} \mu^T \mathbf{K}\mu - R^*(\mu) - \lambda^* B^2 = \sup_{\mu \in \mathbb{R}^n} \frac{3}{4\lambda^*} \mu^T \mathbf{K}\mu - R^*(\mu) - \lambda^* B^2.$$

A solution (α, u) from (7) can be easily computed from an optimal solution μ of this dual problem with : $\alpha = \frac{\mu}{2\lambda^*}$ and $u = \mathbf{K}\alpha = \frac{1}{2\lambda^*}\mathbf{K}\mu$. We could have a large choice for α (adding any element of $Ker(\mathbf{K})$) but all of them will lead to the same solution of the original problem defined by : $f(.) = \sum_{i=1}^n \alpha_i K(x_i, .)$.

4. We are now going to the use the previous work to derive the dual problem of the logistic and the squared hinge losses.

- Logistic loss
  We consider the losses $l_y(u) = \ln(1 + e^{-uy})$ for $y \in \{-1, +1\}$. For a given $y \in \{-1, +1\}$, we compute the Fenchel-Legendre transform of l_y:

$$\forall v \in \mathbb{R},\ l_y^*(v) = \sup_{u \in \mathbb{R}} uv - \ln(1 + e^{-uy})$$

First, we remark that

$$l_y^*(v) = \begin{cases} +\infty & \text{if } (v > 0 \text{ and } y = 1) \text{ or } (v < 0 \text{ and } y = -1) \\ +\infty & \text{if } (v < -1 \text{ and } y = 1) \text{ or } (v > 1 \text{ and } y = -1) \\ 0 & \text{if } v = 0 \text{ or } (v = -1 \text{ and } y = 1) \text{ or } (v = 1 \text{ and } y = -1) \end{cases}$$

The justifications are given at the end of this document.
We consider now that we are in one of the two remaining cases: $(-1 < v < 0$ and $y = 1)$ or $(0 < v < 1$ and $y = -1)$.
The function $u \mapsto uv - \ln(1 + e^{-uy})$ is a concave function. We solve the supremum problem by setting the gradient of this function to 0 :

$$v + \frac{ye^{-uy}}{1 + e^{-uy}} = 0 \Leftrightarrow e^{-uy}(v + y) = -v \Leftrightarrow u = \frac{-1}{y}\ln\left(\frac{-v}{v+y}\right) = -y\ln\left(\frac{-v}{v+y}\right).$$

Hence, in those cases, we have $l_y^*(v) = -yv\ln\left(\frac{-v}{v+y}\right) - \ln(1 - \frac{v}{v+y}) = -yv\ln\left(\frac{-v}{v+y}\right) - \ln(\frac{y}{v+y})$.
Thus, the dual problem takes the following form with the logistic losses :

$$\sup_{\mu \in \mathbb{R}^n} \frac{3}{4\lambda^*}\mu^T\mathbf{K}\mu - \frac{1}{n}\sum_{i=1}^n l_{y_i}^*(n\mu_i) - \lambda^* B^2$$

i.e.

$$\sup_{\mu \in \mathbb{R}^n} \frac{3}{4\lambda^*}\mu^T\mathbf{K}\mu - \frac{1}{n}\sum_{i=1}^n \left(-y_i n\mu_i \ln\left(\frac{-n\mu_i}{n\mu_i + y_i}\right) - \ln\left(\frac{y_i}{n\mu_i + y_i}\right)\right) - \lambda^* B^2$$
$$\text{s.t.} \quad -1 < ny_i\mu_i < 0,\ \forall i \in [\![1, n]\!]$$

- Squared hinge loss
  We consider the losses $l_y(u) = \max(0, 1 - yu)^2$ for $y \in \{-1, +1\}$. For a given $y \in \{-1, +1\}$, we compute the Fenchel-Legendre transform of l_y:

$$\forall v \in \mathbb{R},\ l_y^*(v) = \sup_{u \in \mathbb{R}} uv - \max(0, 1 - yu)^2$$

We have :

$$l_y^*(v) = \begin{cases} +\infty & \text{if } (v > 0 \text{ and } y = 1) \text{ or } (v < 0 \text{ and } y = -1) \\ -1 + \frac{(2y+v)^2}{4} & \text{otherwise} \end{cases}$$

Thus, the dual problem takes the following form with the squared hinge losses :

$$\sup_{\mu \in \mathbb{R}^n} \frac{3}{4\lambda^*}\mu^T \mathbf{K}\mu - \frac{1}{n}\sum_{i=1}^{n} l^*_{y_i}(n\mu_i) - \lambda^* B^2$$

i.e.

$$\begin{aligned} &\sup_{\mu \in \mathbb{R}^n} \frac{3}{4\lambda^*}\mu^T \mathbf{K}\mu - \frac{1}{n}\sum_{i=1}^{n}\left(-1 + \frac{(2y_i + n\mu_i)^2}{4}\right) - \lambda^* B^2 \\ &\text{s.t. } y_i\mu_i \leq 0, \ \forall i \in [\![1, n]\!] \end{aligned}$$

i.e.

$$\begin{aligned} &\sup_{\mu \in \mathbb{R}^n} \frac{3}{4\lambda^*}\mu^T \mathbf{K}\mu - y^T\mu - \frac{n}{4}\mu^T\mu - \lambda^* B^2 \\ &\text{s.t. } y_i\mu_i \leq 0, \ \forall i \in [\![1, n]\!] \end{aligned}$$

Justification of the Fenchel-Legendre transforms for the Exercise 4

Logistic Loss

$$l^*_y(v) = \begin{cases} +\infty & \text{if } (v > 0 \text{ and } y = 1) \text{ or } (v < 0 \text{ and } y = -1) \\ +\infty & \text{if } (v < -1 \text{ and } y = 1) \text{ or } (v > 1 \text{ and } y = -1) \\ 0 & \text{if } v = 0 \text{ or } (v = -1 \text{ and } y = 1) \text{ or } (v = 1 \text{ and } y = -1) \end{cases}$$

We justify those points :

- If $v > 0$ and $y = 1$, $\lim\limits_{u \to +\infty} uv - \ln(1 + e^{-uy}) = \lim\limits_{u \to +\infty} uv - \ln(1 + e^{-u}) = +\infty$.
- If $v < 0$ and $y = -1$, $\lim\limits_{u \to -\infty} uv - \ln(1 + e^{-uy}) = \lim\limits_{u \to -\infty} uv - \ln(1 + e^{u}) = +\infty$
- If $v < -1$ and $y = 1$, $uv - \ln(1 + e^{-uy}) = uv - \ln(1 + e^{-u}) = uv + u - \ln(e^u + 1) \underset{u \to -\infty}{\sim} u(v + 1)$. Since $v < -1$, $\lim\limits_{u \to -\infty} uv - \ln(1 + e^{-uy}) = +\infty$.
- If $v > 1$ and $y = -1$, $uv - \ln(1 + e^{-uy}) = uv - \ln(1 + e^{u}) = uv - u - \ln(e^{-u} + 1) \underset{u \to +\infty}{\sim} u(v - 1)$. Since $v > 1$, $\lim\limits_{u \to +\infty} uv - \ln(1 + e^{-uy}) = +\infty$.
- If $v = -1$ and $y = 1$, $uv - \ln(1 + e^{-uy}) = -u - \ln(1 + e^{-u})$ which is always non positive and which takes the value 0 for $u = 0$.
- If $v = 1$ and $y = -1$, $uv - \ln(1 + e^{-uy}) = u - \ln(1 + e^{u}) = -\ln(1 + e^{-u})$ which is always non positive and which takes the value 0 for $u = 0$.

Squared Hinge Loss

$$l^*_y(v) = \begin{cases} +\infty & \text{if } (v > 0 \text{ and } y = 1) \text{ or } (v < 0 \text{ and } y = -1) \\ -1 + \frac{(2y+v)^2}{4} & \text{otherwise} \end{cases}$$

Indeed :

- If $v > 0$ and $y = 1$, $\lim\limits_{u \to +\infty} uv - \max(0, 1 - yu)^2 = \lim\limits_{u \to +\infty} uv - \max(0, 1 - u)^2 = +\infty$.
- If $v < 0$ and $y = -1$, $\lim\limits_{u \to -\infty} uv - \max(0, 1 - yu)^2 = \lim\limits_{u \to -\infty} uv - \max(0, 1 + u)^2 = +\infty$.

- The function $u \mapsto uv - (1-yu)^2 = -1 - u^2 + u(v+2y)$ (since $y^2 = 1$) reaches its maximum at $u^* = \frac{2y+v}{2}$. Let's prove that u^* is such that $1 - yu^* \geq 0$ in the cases ($v \leq 0$ and $y = 1$) and ($v \geq 0$ and $y = -1$). We will then deduce directly that $l_y^*(v) = u^*v - (1-yu^*)^2$ in those cases.

  - If ($v \leq 0$ and $y = 1$),

  $$1 - yu^* \geq 0 \Leftrightarrow 1 \geq u^* \Leftrightarrow 1 \geq \frac{2+v}{2} \Leftrightarrow v \leq 0$$

  - If ($v \geq 0$ and $y = -1$),

  $$1 - yu^* \geq 0 \Leftrightarrow -1 \leq u^* \Leftrightarrow -1 \leq \frac{-2+v}{2} \Leftrightarrow v \geq 0$$

  Hence, $l_y^*(v) = u^*v - (1-yu^*)^2 = -1 + \frac{(2y+v)^2}{4}$.