MVA "Kernel methods in machine learning" Homework

Mohamed Kerroumi

Exercise 1. Kernels
Show that the following kernels are positive definite:

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

$$\forall x, y \in \mathcal{X} \quad K_4(x, y) = \min(f(x)g(y), f(y)g(x))$$

2. Given a non-empty finite set E, on $\mathcal{X} = \mathcal{P}(E) = \{A : A \subset E\}$:

$$\forall A, B \subset E\,, \quad K\,(A, B) = \frac{|\,A \cap B\,|}{|\,A \cup B\,|}\,,$$

where $|\,F\,|$ denotes the cardinality of F, and with the convention $\frac{0}{0} = 0$.

Solution 1.

1. First, we show that for $\mathcal{X} = \mathbb{R}_+$, $K(x, y) = min(x, y)$ is a positive definite kernel.
Let $(a_1, a_2, ..., a_n) \in \mathbb{R}^n$ and $(x_1, x_2, ..., x_n) \in \mathbb{R}_+^n$, we have

$$\sum_{i=1,j=1}^{n} a_i a_j min(x_i, x_j) = \sum_{i=1,j=1}^{n} a_i a_j \int_{\mathbb{R}} \mathbb{1}_{\leq x_i}(u)\mathbb{1}_{\leq x_j}(u)du = \int_{\mathbb{R}} \Big(\sum_{i=1}^{n} a_i \mathbb{1}_{\leq x_i}(u)\Big)^2 du \geq 0$$

So $K(x, y) = min(x, y)$ is positive definite kernel on $\mathcal{X} = \mathbb{R}_+$
Let $x, y \in \mathbb{R}_+$

$$\begin{aligned}
K(x, y) =& \min(f(x_i)g(x_i), f(x_j)g(x_i)) = \mathbb{1}_{g(x_i)>0} \quad \mathbb{1}_{g(x_j)>0}\Big(f(x_i)g(x_j), f(x_j)g(x_i)\Big) \\
&= \mathbb{1}_{g(x_i)>0} \quad \mathbb{1}_{g(x_j)>0} \quad min\Big(\frac{f(x_i)}{g(x_i)}g(x_i)g(x_j), \frac{f(x_j)}{g(x_j)}g(x_j)g(x_i)\Big) \\
&= \mathbb{1}_{g(x_i)>0} \quad \mathbb{1}_{g(x_j)>0} \quad g(x_i)g(x_j) \quad min\Big(\frac{f(x_i)}{g(x_i)}, \frac{f(x_j)}{g(x_j)}\Big) \\
&= g(x_i)g(x_j) \quad min\Big(\frac{f(x_i)}{g(x_i)}\mathbb{1}_{g(x_i)>0} \quad , \frac{f(x_j)}{g(x_j)}\mathbb{1}_{g(x_j)>0}\Big)
\end{aligned}$$

Let $K(x_i, x_j) = g(x_i)g(x_j)$ and $K'(x_i, x_j) = min\Big(\frac{f(x_i)}{g(x_i)}\mathbb{1}_{g(x_i)>0} \quad , \frac{f(x_j)}{g(x_j)}\mathbb{1}_{g(x_j)>0}\Big)$.
K is a positive definite kernel .
$\forall x \in \mathcal{X} \quad \frac{f(x)}{g(x)}\mathbb{1}_{g(x)>0} \geq 0$, so K' is a positive definite Kernel. Hence K_4 is a p.d kernel as product of p.d kernels.

2. $\mathcal{X} = P(E) = \{A : A \subset E\}$, and $\forall A, B, \quad K(A, B) = \frac{|A \cap B|}{|A \cup B|}$
Let F be the space of measurable functions from $(E, \mathcal{P}(E))$ to $\Big([0, 1], \mathcal{B}([0, 1])\Big)$, and μ the counting measure on E, let f, g be two function of $\mathcal{F}$. we have

$$\langle f, g\rangle = \int f \times g \quad d\mu$$

$$|A \cap B| = \mu(A \cap B) = \int \mathbb{1}_A \mathbb{1}_B \quad d\mu = \langle \mathbb{1}_A, \mathbb{1}_B\rangle$$

So $K_1(A, B) = |A \cap B|$ is a p.d kernel.
We suppose $A \neq \emptyset$ or $B \neq \emptyset$.

$$\frac{1}{|A \cup B|} = \frac{1}{|E| - |A^{\complement} \cap B^{\complement}|} = \frac{1}{|E|}\frac{1}{1 - \frac{|A^{\complement} \cap B^{\complement}|}{|E|}} = \frac{1}{|E|}\sum_0^{+\infty} \frac{|A^{\complement} \cap B^{\complement}|^k}{|E|^k} \quad because \quad \frac{|A^{\complement} \cap B^{\complement}|}{|E|} < 1$$

$K_2(A, B) = \frac{|A^{\complement} \cap B^{\complement}|^k}{|E|^k} = \frac{K_1(A^{\complement}, B^{\complement})}{|E|}$, so K_2 is p.d kernel.
Hence, $\frac{|A^{\complement} \cap B^{\complement}|^k}{|E|^k}$ is p.d kernel as product of p.d kernels.
so $\sum_0^{+\infty} \frac{|A^{\complement} \cap B^{\complement}|^k}{|E|^k}$ is a p.d kernel as a limit of p.d kernels.
Hence $K(A, B) = \frac{|A \cap B|}{|A \cup B|}$ is a p.d kernel.
Now if $|A_i| = 0$ and $|A_j| = 0$, hence their terms in the expression $\sum_{k,l} a_k a_l K(A_k, A_l)$ is $a_i^2 + a_j^2 + 2a_i a_j \geq 0$, so they don't affect the positivity of the expression, hence $K(A, B)$ is a p.d kernel for all A,B in $\mathcal{P}(E)$.

Exercise 2. Kernels encoding equivalence classes.
Consider a similarity measure $K : \mathcal{X} \times \mathcal{X} \to \{0, 1\}$ with $K(x, x) = 1$ for all x in $\mathcal{X}$. Prove that K is p.d. if and only if, for all x, x', x'' in $\mathcal{X}$,

- $K(x, x') = 1 \Leftrightarrow K(x', x) = 1$, and
- $K(x, x') = K(x', x'') = 1 \Rightarrow K(x, x'') = 1$.

Solution 2.

- Let K be a p.d kernel, and $x, x', x'' \in \mathcal{X}$
$K(x, x') = 1 \iff K(x', x) = 1$ because K is symmetric.
We suppose $K(x, x') = K(x', x) = 1$ and $K(x, x'') \neq 0$ so $K(x, x'') = 0$

let $(a_1, a_2, a_3) \in \mathbb{R}^3$, K is a p.d so:

$$a_1^2K(x,x)+a_2^2K(x',x')+a_3^2K(x'',x'')+2a_1a_2K(x,x')+2a_1a_3K(x,x'')+2a_2a_3K(x',x'') \geq 0$$
$$a_1^2 + a_2^2 + a_3^2 2a_1a_2 + 2a_2a_3 \geq 0$$
$$(a_1 + a_2)^2 + (a_2 + a_3)^2 \geq a_2^2$$

If we choose $a_1 = a_3 = -a_2$ and $a_2 \neq 0$, we will have:

$$0 \geq a_2 > 0 \quad \text{Which is impossible}$$

Hence $K(x, x'') = 1$

- Suppose $K(x, x') = 1 \Leftrightarrow K(x', x) = 1$, and $K(x, x') = K(x', x'') = 1 \Rightarrow K(x, x'') = 1$.
  We will show that K is a p.d.
  $\forall$x,y $\in \mathcal{X}$, we define $x \sim y \quad$ if $\quad K(x, y) = 1$
  $\sim$ is an equivalency relation:
  Reflexion property: $X \sim x$ because $K(x, x) = 1$
  Symmetric property: $X \sim y \iff K(x, y = 1) \iff K(y, x) = 1 \iff y \sim x$
  Transitive property:

$$x' \quad x''' \iff K(x, x') = K(x', x'') = 1 \implies K(x, x'') = 1 \implies x \sim x''$$

Let $(a_1, a_2, ..., a_n) \in \mathbb{R}^n$ and $(x_1, x_2, ..., x_n) \in \mathcal{X}^n$
We can find in $(x_1, x_2, ...x_n)$ p equivalence classes.

$$\sum_{i=1,j=1}^{n} a_i a_j K(x_i, x_j) = \sum_{k=1}^{p} \sum_{x_i \in C_k, x_j \in C_k} a_i a_j$$
$$\sum_{x_i \in C_k, x_j \in C_k} a_i a_j = \Big(\sum_{x_i \in C_k} a_i \Big)^2$$

So $\sum_{i=1,j=1}^{n} a_i a_j K(x_i, x_j) \geq 0$
Hence K is a p.d kernel.

Exercise 3. COCO

Given two sets of real numbers $X = (x_1, \ldots, x_n) \in \mathbb{R}^n$ and $Y = (y_1, \ldots, y_n) \in \mathbb{R}^n$, the covariance between X and Y is defined as

$$cov_n(X, Y) = \mathbf{E}_n(XY) - \mathbf{E}_n(X)\mathbf{E}_n(Y)\,,$$

where $\mathbf{E}_n(U) = (\sum_{i=1}^{n} u_i)/n$. The covariance is useful to detect linear relationships between X and Y. In order to extend this measure to potential nonlinear relationships between X and Y, we consider the following criterion:

$$C_n^K(X, Y) = \max_{f,g \in \mathcal{B}_K} cov_n(f(X), g(Y))\,,$$

where K is a positive definite kernel on $\mathbb{R}$, $\mathcal{B}_K$ is the unit ball of the RKHS of K, and $f(U) = (f(u_1), \ldots, f(u_n))$ for a vector $U = (u_1, \ldots, u_n)$.

1. Express simply $C_n^K(X,Y)$ for the linear kernel $K(a,b) = ab$.
2. For a general kernel K, express $C_n^K(X,Y)$ in terms of the Gram matrices of X and Y.

Solution 3.

1. If we consider the linear kernel $K(a,b) = ab$
$$\begin{aligned} C_n^K(X,Y) &= \max_{f,g\in\mathcal{B}_K} cov_n(f(X),g(Y)) \\ &= \max_{a,b,a^2\leq 1,b^2\leq 1} cov_n\Big((ax_1,ax_2,...,ax_n),(by_1,by_2,...,by_n)\Big) \\ &= \max_{a,b,|a|\leq 1,|b|\leq 1} \frac{ab}{n}\sum_{i=1}^n x_iy_i - \frac{ab}{n^2}(\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i) \\ &= \max_{a,b,|a|\leq 1,|b|\leq 1} ab \times cov_n(X,Y) \\ &= |cov_n(X,Y)| \end{aligned}$$

2. $C_n^K(X,Y) = \max_{f,g\in\mathcal{B}_K} cov_n(f(X),g(Y))$
Let f^* and g^* be the solutions of the maximization problem
f^* is the solution to the problem $\max_{f,g\in\mathcal{B}_K} cov_n(f(X),g^*(Y))$
g^* is the solution to the problem $\max_{f,g\in\mathcal{B}_K} cov_n(f^*(X),g(Y))$
By the representer theorem:

$$\exists F = (f_1,f_2,...,f_n)\in\mathbb{R}^n \quad \text{such as} \quad f(x) = \sum_{i=1}^n f_iK(x_i,x)$$

$$\exists G = (g_1,g_2,...,g_n)\in\mathbb{R}^n \quad \text{such as} \quad g(y) = \sum_{i=1}^n g_iK(y_i,y)$$

$\|f\|_H^2 = \sum_{i=1,j=1}^n f_if_jK(x_i,x_j) = F^\top K_XF$ and $\|g\|_H^2 = G^\top K_YG$
$f(x_i) = [K_XF]_i$ and $g(y_i) = [K_YG]_i$ So
$$\begin{aligned} C_n^K(X,Y) &= \max_{f,g\in\mathcal{B}_K} cov_n(f(X),g(Y)) \\ &= \max_{f,g\in\mathcal{B}_K} \frac{1}{n}\sum_{i=1}^n f(x_i)g(x_i) - \frac{1}{n^2}\sum_{i=1}^n f(x_i)\sum_{i=1}^n g(x_i) \\ &= \max_{F^\top K_XF\leq 1,G^\top K_YG\leq 1} \frac{1}{n}\Big((K_XF)^\top(K_YG) - \frac{1}{n}(K_XF)^\top\mathbb{1}_n\mathbb{1}_n^\top K_YG\Big) \\ &= \max_{F^\top K_XF\leq 1,G^\top K_YG\leq 1} \frac{1}{n}F^\top K_XK_YG - \frac{1}{n^2}F^\top K_X\mathbb{1}_n\mathbb{1}_n^\top K_YG \qquad K_x, \\ &= \max_{F^\top K_XF\leq 1,G^\top K_YG\leq 1} \frac{1}{n}F^\top K_X\Big(I - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^\top\Big)K_YG \end{aligned}$$

Let $H = I - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^\top$

$$= \max_{F^\top K_XF\leq 1,G^\top K_YG\leq 1} \frac{1}{n}F^\top K_XHK_YG$$

K_y, are non negative matrices, so admit a Square root non negative, so,

$$C_n^K(X,Y) = \max_{\|K_x^{\frac{1}{2}}F\|\leq 1, \|K_y^{\frac{1}{2}}G\|\leq 1} \frac{1}{n} F^\top K_x^{\frac{1}{2}} K_x^{\frac{1}{2}} H K_y^{\frac{1}{2}} K_y^{\frac{1}{2}} G$$

Let $\bar{F} = K_x^{\frac{1}{2}} F$ and $\bar{G} = K_y^{\frac{1}{2}} G$

$$C_n^K(X,Y) = \max_{\|\bar{F}\|\leq 1, \|\bar{G}\|\leq 1} \frac{1}{n} \bar{F}^\top K_x^{\frac{1}{2}} H K_y^{\frac{1}{2}} \bar{G}$$

$$C_n^K(X,Y) = \frac{1}{n} \| \ K_x^{\frac{1}{2}} H K_y^{\frac{1}{2}} \|_2$$

Exercise 4. Dual coordinate ascent algorithms for SVMs

1. We recall the primal formulation of SVMs seen in the class (slide 142).

$$\min_{f\in\mathcal{H}} \frac{1}{n} \sum_{i=1}^n \max(0, 1 - y_i f(\mathbf{x}_i)) + \lambda \|f\|_{\mathcal{H}}^2,$$

and its dual formulation (slide 152)

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^n} 2\boldsymbol{\alpha}^\top \mathbf{y} - \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha} \quad \text{such that} \quad 0 \leq y_i \alpha_i \leq \frac{1}{2\lambda n}, \quad \text{for all } i.$$

The coordinate ascent method consists of iteratively optimizing with respect to one variable, while fixing the other ones. Assuming that you want to maximize the dual by following this approach. Find (and justify) the update rule for α_j.

2. Consider now the primal formulation of SVMs with intercept

$$\min_{f\in\mathcal{H}, b\in\mathbb{R}} \frac{1}{n} \sum_{i=1}^n \max(0, 1 - y_i (f(\mathbf{x}_i) + b)) + \lambda \|f\|_{\mathcal{H}}^2,$$

Can we still apply the representer theorem? Why? Derive the corresponding dual formulation by using Lagrangian duality. Can we apply the coordinate ascent method to this dual? If yes, what are the update rules?

3. Consider a coordinate ascent method to this dual that consists of updating two variables (α_i, α_j) at a time (while fixing the $n-2$ other variables). What are the update rules for these two variables?

Solution 4.

1. Update of α_j using coordinate ascent method:
   Let $e_j = (0, 0, ..., 0, 1, 0, ..., 0)$ the n_tuple equal to 0, except the j'th component. The update rule is

$$\alpha_j \to \alpha_j + \delta^* \quad \text{Such that} \quad \delta^* = argmax_{\delta, 0 \leq (\alpha_j + \delta) y_j \leq \frac{1}{2\lambda n}} f(\alpha + \delta e_j)$$

   Where

$$\begin{aligned} f(\alpha + \delta e_j) &= 2(\alpha + \delta e_j)^\top y - (\alpha + \delta e_j)^\top K (\alpha + \delta e_j) \\ &= 2\alpha^\top y + 2\delta e_j^\top y - \alpha^\top K \alpha - 2\delta e_j^\top K \alpha - \delta^2 K_{jj} \end{aligned}$$

   f is a quadratic function, so to compute the $argmax$ we set the gradient w.r.t to δ equal to 0.

$$\nabla_\delta f(\alpha + \delta e_j) = 2y_j - 2[K\alpha]_j - 2\delta K_{jj} = 0$$

   - If $K_{jj} \neq 0$:

$$\delta^* = \frac{y_j - [K\alpha]_j}{K_{jj}}$$

     δ^* must verify the condition $\quad 0 \leq (\alpha_j + \delta^*) y_j \leq \frac{1}{2\lambda n} \quad$ .Hence:

$$\delta^* = \begin{cases} \max\left(\min\left(\frac{1}{2\lambda n} - \alpha_j, \frac{y_j - [K\alpha]_j}{K_{jj}}\right), -\alpha_j\right) & \text{if} \quad y_j = 1 \\ \max\left(\min\left(-\alpha_j, \frac{y_j - [K\alpha]_j}{K_{jj}}\right), -\alpha_j - \frac{1}{2\lambda n}\right) & \text{if} \quad y_j = -1 \end{cases}$$

   - If $K_{jj} = 0$:

$$\nabla_\delta f(\alpha + \delta e_j) = 2y_j - 2[K\alpha]_j$$

     So $f(\alpha + \delta e_j)$ is an affine function, w.r.to δ hence :
     If

$$y_i \geq [K\alpha]_j \quad \Rightarrow \quad \delta^* + \alpha_j = \frac{1}{2\lambda n} \quad \Rightarrow \quad \delta^* = \frac{1}{2\lambda n} - \alpha_j$$

     If

$$y_i \leq [K\alpha]_j \quad \Rightarrow \quad \delta^* + \alpha_j = 0 \quad \Rightarrow \quad \delta^* = -\alpha_j$$

   and

$$\alpha_j \leftarrow \alpha_j + \delta^*$$

2. The primal problem :

$$\min_{f \in \mathcal{H}, b \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \max(0, 1 - y_i(f(\mathbf{x}_i) + b)) + \lambda \|f\|_{\mathcal{H}}^2,$$

   We can still apply the representer theorem, because in the demonstration of the theorem in the course, we saw that we can decompose $f = f_H + f_\perp$ where $f_H \in \mathcal{H}$ and $f_\perp$ is perpendicular to H, we can do the same thing for this problem, we can write f as $\quad f = f_S + f_\perp \quad$ where $f_S = f_0 + b$ where $f_0 \in \mathcal{H}$ and $f_\perp$ perpendicular to $\quad \mathcal{H} + \mathbb{R} \quad$ so instead

of minimizing over $\mathcal{H}$ we minimize over $\mathcal{H} + \mathbb{R}$ with function of the form $f(x) + b$, then we can apply the representer theorem.

$$\underset{f\in\mathcal{H},b\in\mathbb{R}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}\max(0, 1 - y_i(f(\mathbf{x}_i) + b)) + \lambda\|f\|_{\mathcal{H}}^2 = \Big(\sum_{i=1}^{n}\alpha_i^* K_i, b^*\Big)$$

So the problem become :

$$\min_{\alpha\in\mathbb{R}^n,b\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n}\varphi_{Hinge}\Big(y_i([K\alpha]_i + [b\mathbb{1}_n]_i\Big) + \lambda\alpha^\top K\alpha$$

which is equivalent to :

$$\min_{\alpha\in\mathbb{R}^n,b\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n}\varphi_{Hinge}\Big(y_i([K\alpha + b\mathbb{1}_n]_i\Big) + \lambda\alpha^\top K\alpha$$

In the course, we saw that the Lagrangian of this problem is :

$$L(\alpha, b, \xi, \mu, v) = \frac{1}{n}\xi^\top\mathbb{1}_n + \lambda\alpha^\top K\alpha - (diag(y)\mu)^\top(K\alpha + \beta\mathbb{1}_n) - (\mu + v)^\top\xi + \mu^\top\mathbb{1}_n$$

- Minimization w.r.to b implies:

$$\sum_{i=1}^{n} y_i\mu_i = 0$$

- Minimization w.r.to α implies:

$$\nabla_\alpha L(\alpha, b, \xi, \mu, v) = 2\lambda K\alpha - K diag(y)\mu = 0$$
$$\alpha^* = \frac{diag(y)\mu}{2\lambda}$$

  So $\quad \sum_{i=1}^{n}\alpha_i = 0$

- Minimization w.r.to ξ implies:

$$\nabla L_\xi = \frac{1}{n} - \mu - v = 0$$

  so $\quad \mu + v = \frac{1}{n}$

Hence, as we saw in course, the dual formulation of SVM with intercept is :

$$\boxed{\begin{array}{lll} \max\limits_{\alpha\in\mathbb{R}^n} & 2\alpha^\top y - \alpha^\top K\alpha & \\ & 0 \le y_i\alpha_i \le \dfrac{2}{2\lambda n} & \forall i \\ & \sum\limits_{i=1}^{n}\alpha_i = 0 & \end{array}}$$

We can't apply the coordinate ascent method for this dual because if we modify only one variable α_j and keep all the $n-1$ other variables fixed, we will break the constraint $\sum_{i=1}^{n}\alpha_i = 0$.

Exercise 5. Duality

Let $(x_1, y_1), \ldots, (x_n, y_n)$ a training set of examples where $x_i \in \mathcal{X}$, a space endowed with a positive definite kernel K, and $y_i \in \{-1, 1\}$, for $i = 1, \ldots, n$. $\mathcal{H}_K$ denotes the RKHS of the kernel K. We want to learn a function $f : \mathcal{X} \mapsto \mathbb{R}$ by solving the following optimization problem:

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

where ℓ_y is a convex loss functions (for $y \in \{-1, 1\}$) and $B > 0$ is a parameter.

1. Show that there exists $\lambda \geq 0$ such that the solution to problem (1) can be found be solving the following problem:
$$\min_{\alpha \in \mathbb{R}^n} R(K\alpha) + \lambda \alpha^\top K \alpha \,, \tag{2}$$
where K is the $n \times n$ Gram matrix and $R : \mathbb{R}^n \mapsto \mathbb{R}$ should be explicited.

2. Compute the Fenchel-Legendre transform[1] R^* of R in terms of the Fenchel-Legendre transform ℓ_y^* of ℓ_y.

3. Adding the slack variable $u = K\alpha$, the problem (2) can be rewritten as a constrained optimization problem:
$$\min_{\alpha \in \mathbb{R}^n, u \in \mathbb{R}^n} R(u) + \lambda \alpha^\top K \alpha \quad \text{such that} \quad u = K\alpha \,. \tag{3}$$
Express the dual problem of (4) in terms of R^*, and explain how a solution to (3) can be found from a solution to the dual problem.

4. Explicit the dual problem for the logistic and squared hinge losses:
$$\ell_y(u) = \log(1 + e^{-yu}) \,.$$
$$\ell_y(u) = \max(0, 1 - yu)^2 \,.$$

Solution 5.

1. Let $(x1, x2, ..., x_n) \in \mathcal{X}^n$ and $(y_1, y_2, ..., y_n) \in \{0, 1\}^n$, we want to solve the following optimization problem:
$$\min_{f \in \mathcal{H}_k} \frac{1}{n} \sum_{i=1}^{n} l_{y_i}(f(x_i)) \quad \text{such as} \quad \|f\|_{\mathcal{H}_k} \leq B$$

[1] For any function $f : \mathbb{R}^N \mapsto \mathbb{R}$, the *Fenchel-Legendre transform* (or *convex conjugate*) of f is the function $f^* : \mathbb{R}^N \mapsto \mathbb{R}$ defined by
$$f^*(u) = \sup_{x \in \mathbb{R}^N} x^\top u - f(x) \,.$$

The Lagrangian of this problem is :

$$L(f,\lambda) = \frac{1}{n}\sum_{i=1}^{n} l_{y_i}(f(x_i)) \quad + \quad \lambda(\|f\|^2_{\mathcal{H}_k} - B^2)$$

l_{y_i} is convex and $\|f\|^2_{\mathcal{H}_k}$ is convex w.r.t f.
So $\frac{1}{n}\sum_{i=1}^{n} l_{y_i}(f(x_i))$ and $\|f\|^2_{\mathcal{H}_k} - B^2$ are convex.
The problem is strictly feasible so by the Slater's constraint qualification, Strong duality holds. So there exists a $\lambda \geq 0$ such as the solution of the problem (1) is the solution of:

$$L(f,\lambda) = \frac{1}{n}\sum_{i=1}^{n} l_{y_i}(f(x_i)) \quad + \quad \lambda(\|f\|^2_{\mathcal{H}_k} - B^2)$$

By the representer theorem, the solution can be explained as :

$$f(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x)$$

By removing the constant, the problem becomes :

$$\min_{\alpha\in\mathbb{R}^n} \quad \frac{1}{n}\sum_{i=1}^{n} l_{y_i}\Big(\sum_{j=1}^{n} \alpha_j K(x_i,x_j)\Big) \quad + \quad \lambda\Big(\sum_{i=1,j=1}^{n} \alpha_i\alpha_j K(x_i,x_j)\Big)$$

$$\min_{\alpha\in\mathbb{R}^n} \quad \frac{1}{n}\sum_{i=1}^{n} l_{y_i}\Big([K\alpha]_i\Big) \quad + \quad \lambda\alpha^\top K\alpha$$

So $R(u) = \frac{1}{n}\sum_{i=1}^{n} l_{y_i}(u_i)$

2. R^* in terms of l^*_y
   Let $u \in \mathbb{R}^n$ and $x \in \mathbb{R}^n$

$$\begin{aligned}
R^*(x) &= \sup_{x\in\mathbb{R}^n} \quad x^\top u - R(u) \\
&= \sup_{x\in\mathbb{R}^n} \quad \sum_{i=1}^{n} x_i u_i - \frac{1}{n}\sum_{i=1}^{n} l_{y_i}(u_i) \\
&= \sup_{x\in\mathbb{R}^n} \quad \sum_{i=1}^{n} \frac{1}{n}\Big(n x_i y_i - l_{y_i}(u_i)\Big) \\
&= \sum_{i=1}^{n} \quad \frac{1}{n}\sup_{x_i\in\mathbb{R}} \Big(n x_i y_i - l_{y_i}(u_i)\Big) \\
&= \frac{1}{n}\sum_{i=1}^{n} \quad l^*_{y_i}(n x_i)
\end{aligned}$$

3. After adding slack variables, the problems becomes,

$$\min_{\alpha\in\mathbb{R}^n, u\in\mathbb{R}^n} R(u) + \lambda\alpha^\top K\alpha \quad \text{such that} \quad u = K\alpha\,. \tag{4}$$

Let $\mu \in \mathbb{R}^n$, the lagrangian of equation(4) is :

$$\begin{aligned} L(\alpha, u, \mu) &= R(u) + \lambda\alpha\top K\alpha + \mu^\top(K\alpha - u) \\ &= R(u) - \mu^\top u + \lambda\alpha^\top K\alpha + \mu^\top K\alpha \end{aligned}$$

$$\begin{aligned} \min_u L(\alpha, u, \mu) &= \min_u R(u) - \mu^\top u + \lambda\alpha^\top K\alpha + \mu^\top K\alpha \\ &= -R^*(\mu) + \lambda\alpha^\top K\alpha + \mu^\top K\alpha \end{aligned}$$

$$\min_{u,\alpha} L(\alpha, u, \mu) = -R^*(\mu) + \min_\alpha(\lambda\alpha^\top K\alpha + \mu^\top K\alpha)$$

The function $\alpha \to \lambda\alpha^\top K\alpha + \mu^\top K\alpha$ is a quadratic function and $K \geq 0$
So $\nabla_\alpha(\lambda\alpha^\top K\alpha + \mu^\top K\alpha) = 2\lambda K\alpha + K\mu = 0$
Hence

$$K\alpha^* = \frac{-1}{2\lambda}K\mu$$

and we have :

$$\lambda\alpha^{*\top}K\alpha^* = \lambda\alpha^{*\top}\Big(\frac{-1}{2\lambda}K\mu\Big) = \frac{-1}{2}\alpha^{*\top}K\mu = \frac{-1}{2}\mu^\top K\alpha^* = \frac{1}{4\lambda}\mu^\top K\mu$$
$$\mu^\top K\alpha = \frac{-1}{2\lambda}\mu^\top K\mu$$

So $\min_{\alpha,u} L(\alpha, u, \mu) = -R^*(\mu) - \frac{1}{4\lambda}\mu^\top K\mu$
Hence the dual problem is :

$$\boxed{\min_{\mu\in\mathbb{R}^n} R^*(\mu) + \frac{1}{4\lambda}\mu^\top K\mu}$$

Once we find μ^* a solution of the dual, we get α the solution of the primal by the formula :

$$\alpha = \frac{1}{2\lambda}(y - \mu^*) \quad \text{with} \quad y \in Ker(K)$$

4. Suppose $l_y(u) = \log(1 + e^{-yu})$

$l_y^*(x) = \sup_{u\in\mathbb{R}}\Big(xu - \log(1+e^{-yu})\Big) = \sup_{u\in\mathbb{R}}\Big((xy)(yu) - \log(1+e^{-yu})\Big) = \sup_{u\in\mathbb{R}}\Big((yx+1)(yu) - \log(1+e^{yu})\Big)$

because $y^2 = 1$

- If $xy = 0$ we have $l_y^*(x) = \sup_{u\in\mathbb{R}}\Big(-\log(1 + e^{-yu})\Big) = 0$
- If $xy > 0$ we chose u such as $yu \to +\infty$ so $\sup_{v\in\mathbb{R}}\Big((xy)(v) - \log(1 + e^{-v})\Big) = +\infty$
  So in this case $l_y^*(x) = +\infty$

- If $xy < -1$ we chose u such as $yu \to -\infty$ so $(yx+1)(yu) \to \infty$ and $\log(1+e^{yu}) \to 0$ so $l_y^*(u) = +\infty$
- If $xy = -1$ we have $l_y^*(u) = \sup_{u\in\mathbb{R}} \Big(-\log(1+e^{yu})\Big) = 0$
- If $-1 < xy < 0$:

$$\nabla_u\Big(xu - \log(1-e^{-yu})\Big) = x + \frac{ye^{-yu}}{1+e^{-yu}} = 0 \iff \quad xy + \frac{1}{1+e^{yu}} = 0$$

Hence $u^* = y\log(-1-\frac{1}{xy})$

$$\begin{aligned} l_y^*(x) &= xy\log(-1-\frac{1}{xy}) - \log\left(1+exp(-log(-1-\frac{-1}{yx}))\right) \\ &= yv\log(\frac{1+yx}{-yx}) - log(\frac{1}{1+yx}) \\ &= (xy+1)\log(xy+1) - xy\log(-xy) \end{aligned}$$

So :

$$l_y^*(x) = \begin{cases} (1+xy)\log(xy+1) - xy\log(-xy) & \text{if} \quad -1 \le xy \le 0 \\ +\infty \quad \text{otherwise} \end{cases}$$

So the dual problem becomes :

$$\boxed{\begin{array}{l} \min\limits_{\mu\in\mathbb{R}^n} \dfrac{1}{4\lambda}\mu^\top K\mu + \dfrac{1}{n}\sum\limits_{i=1}^{n}(n\mu_i y_i+1)\log(n\mu_i y_i+1) - n\mu_i y_i\log(-n\mu_i y_i) \\ sb.to \quad \dfrac{-1}{n} \le y_i\mu_i \le 0 \end{array}}$$

We suppose $l_y(u) = \max(0, 1-yu)^2$

$$l_y^*(x) = \sup_{u\in\mathbb{R}}\Big(xu - \max(0,1-yu)^2\Big) = \sup_{u\in\mathbb{R}}\Big((xy)(yu) - \max(0,1-yu)^2\Big) = \sup_{v\in\mathbb{R}}\Big(xyv - \max(0,1-v)^2\Big)$$

If $xy > 0$ we have $l_y^*(x) = +\infty$
We suppose $xy \le 0$:

- if $v \ge 1 \quad l_y^*(x) = \sup_{v\ge 1} xyv = xy$
- if $v \le 1 \quad l_y^*(x) = \sup_{v\le 1} xyv - (1-v)^2 = \sup_{v\le 1} xyv - (v-1)^2$

$$\nabla_v(xyv - (v-1)^2) = xy - 2v + 2 = 0$$

So $v^* = 1 + \frac{xy}{2} \quad$ hence $\quad l_y^*(x) = xy + \frac{(xy)^2}{2} - \frac{(xy)^2}{4} = xy + \frac{(xy)^2}{4}$ so we deduce that:

$$l_y^*(x) = \begin{cases} xy + \frac{(xy)^2}{4} & \text{if} \quad xy \le 0 \\ +\infty \quad \text{otherwise} \end{cases}$$

So the dual problem is :

$$\min_{\mu\in\mathbb{R}^n} \frac{1}{4\lambda}\mu^\top K\mu + \frac{1}{n}\sum_{i=1}^{n} n\mu_i y_i + \frac{n^2\mu_i^2 y_i^2}{4}$$

$$sb.to \quad y_i\mu_i \leq 0$$

$$\boxed{\begin{array}{l} \min\limits_{\mu\in\mathbb{R}^n} \quad \dfrac{1}{4\lambda}\mu^\top K\mu + \displaystyle\sum_{i=1}^{n} \mu_i y_i + \frac{n\mu_i^2}{4} \\ \quad sb.to \quad y_i\mu_i \leq 0 \end{array}}$$