Homework 1

Ayman YACHAOUI
Kernel methods in machine learning
Data & Knowledge Liberté
École Normale Supérieure de Cachan
ayman.yachaoui@telecom-paristech.fr

1 Kernel examples

Are the following kernels positive definite?

1.1

The kernel K_1 and K_2 are defined as follow:

$$\forall (x,y) \in \mathbb{R}^2 \quad K_1(x,y) = 10^{xy}, \quad K_2(x,y) = 10^{x+y}$$

We know that the kernel K corresponding to the canonical scalar product $\forall (x,y) \in \mathbb{R}^2 \mapsto K(x,y) = \langle x,y \rangle_{\mathbb{R}} = xy$ is a positive definite kernel. Thus by multiplying by the constant $c = \ln 10$ it stays positive definite. Finally taking the exponential we obtain that the kernel $K_1 = \exp(cK)$ is also positive definite.

Now considering K_2, we note $\phi : \mathbb{R} \to \mathbb{R}$ the function such that $\forall x \in \mathbb{R}, \phi(x) = 10^x$.
Then $\forall (x,y) \in \mathbb{R}^2, K_2(x,y) = 10^x 10^y = \langle \phi(x), \phi(y) \rangle_{\mathbb{R}}$ and using Aronzsajn's theorem, K_2 is a positive definite kernel.

1.2

The kernel K_3 is defined as follow:

$$\forall (x,y) \in [0,1[^2 \quad K_3(x,y) = -\log(1-xy)$$

We can also write:

$$\forall (x,y) \in [0,1[^2 \quad K_3(x,y) = \lim_{n \to \infty} \sum_{i=1}^{n} \frac{K(x,y)^i}{i}$$

Finally using the combination rules we have that each element of the sum is a positive definite kernel, so the finite sum is also a positive definite kernel and finally using the pointwise convergence and the limit we get that K_3 is a positive definite kernel too.

1.3

The kernel K_4 is defined as follow for every set $\mathcal{X}$ and $f, g : \mathcal{X} \to \mathbb{R}_+$ two non-negative functions:

$$\forall (x,y) \in \mathcal{X}^2 \quad K_4(x,y) = \min(f(x)g(y), f(y)g(x))$$

Then $\forall a \in \mathbb{R}^n, \forall (x_i)_{1 \leq i \leq n} \in \mathcal{X}^n$, we have:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j K_4(x_i, x_j) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \min(f(x_i)g(x_j), f(x_j)g(x_i))$$

We can suppose $\forall i, f(x_i) > 0$ without loss of generality because if $f(x_i) = 0$ then all the corresponding term are equal to zero and the sum can be reduced. Thus we have:

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j K_4(x_i, x_j) &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j f(x_i) f(x_j) \min\left(\frac{g(x_j)}{f(x_j)}, \frac{g(x_i)}{f(x_i)}\right) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j f(x_i) f(x_j) \int_{t=0}^{+\infty} 1_{t\leq \frac{g(x_j)}{f(x_j)}} 1_{t\leq \frac{g(x_i)}{f(x_i)}} dt \\
&= \int_{t=0}^{+\infty} \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j f(x_i) f(x_j) 1_{t\leq \frac{g(x_j)}{f(x_j)}} 1_{t\leq \frac{g(x_i)}{f(x_i)}} dt \\
&= \int_{t=0}^{+\infty} \left(\sum_{i=1}^{n} a_i f(x_i) 1_{t\leq \frac{g(x_i)}{f(x_i)}}\right)\left(\sum_{j=1}^{n} a_j f(x_j) 1_{t\leq \frac{g(x_j)}{f(x_j)}}\right) dt \\
&= \int_{t=0}^{+\infty} \left(\sum_{i=1}^{n} a_i f(x_i) 1_{t\leq \frac{g(x_i)}{f(x_i)}}\right)^2 dt \\
&\geq 0
\end{aligned}$$

2 Combining kernels

Find the Reproducing Kernel Hilbert Space of the following kernels.

2.1

The kernel K_1 and K_2 are defined as follow:

$$\forall (x, y) \in \mathbb{R}^2 \quad K_1(x, y) = (xy + 1)^2, \quad K_2(x, y) = (xy - 1)^2$$

- First we will look into K_1:
  - First step: Look for an inner-product

$$\begin{aligned}
\forall (x, y) \in \mathbb{R}^2, K_1(x, y) &= x^2 y^2 + 2xy + 1 \\
&= \langle \phi(x), \phi(y) \rangle_{\mathbb{R}^3}
\end{aligned}$$

    By defining ϕ such that:

$$\begin{aligned}
\phi : \quad & \mathbb{R} \to \mathbb{R}^3 \\
& x \mapsto \begin{pmatrix} x^2 \\ \sqrt{2}x \\ 1 \end{pmatrix}
\end{aligned}$$

  - Second step: propose a candidate RKHS
    We know that $\mathcal{H}_1$ contains all the functions

$$\begin{aligned}
f(x) &= \sum_i a_i K_1(x_i, x) \\
&= \sum_i a_i \langle \phi(x_i), \phi(x) \rangle_{\mathbb{R}^3} \\
&= \langle \sum_i a_i \phi(x_i), \phi(x) \rangle_{\mathbb{R}^3}
\end{aligned}$$

    Besides by choosing $x_1 = 0$, $x_2 = \sqrt{2}$ and $x_3 = -\sqrt{2}$ and denoting $e_1 = \phi(x_1) = (0, 0, 1)^\intercal$, $e_2 = \phi(x_2) = (2, 2, 1)^\intercal$ and $e_3 = \phi(x_3) = (2, -2, 1)^\intercal$ we have that e_1, e_2 and e_3 form a basis of $\mathbb{R}^3$ and thus any vector in $\mathbb{R}^3$ may be decomposed as $\sum_i a_i \phi(x_i)$. Our candidate RKHS $\mathcal{H}_1$ will be:

$$f_u(x) = \langle u, \phi(x) \rangle \quad \forall u \in \mathbb{R}^3$$

Endowed with the inner-product $\langle f_u, f_v \rangle_{\mathcal{H}_1} = \langle u, v \rangle_{\mathbb{R}^3}$

- Third step: check that the candidate is a Hilbert space.
  This is trivial here.
- Fourth step: check that $\mathcal{H}_1$ is the RKHS.
  $\mathcal{H}_1$ contains all the function $K_x : t \mapsto K_1(x,t) = \langle \phi(x), \phi(t) \rangle_{\mathbb{R}^3}$.
  For every $x \in \mathbb{R}$ and $f \in \mathcal{H}_1$, the reproducing property holds:

$$\begin{aligned} f_u(x) &= \langle u, \phi(x) \rangle_{\mathbb{R}^3} \\ &= \langle f_u, f_{\phi(x)} \rangle_{\mathcal{H}_1} \\ &= \langle f_u, K_x \rangle_{\mathcal{H}_1} \end{aligned}$$

- We now look into K_2.
  We observe that when choosing $x_1 = 0$ and $x_2 = 1$ the similarity matrix $\mathcal{K}$ is:

$$\mathcal{K} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$$

  Which two eigenvalues are $\frac{1+\sqrt{5}}{2} > 0$ and $\frac{1-\sqrt{5}}{2} < 0$, thus $\mathcal{K}$ is not positive semidefinite and K_2 isn't positive definite.

- Finally we consider $K = K_1 + K_2$

$$\begin{aligned} \forall (x,y) \in \mathbb{R}^2 \quad K(x,y) &= K_1(x,y) + K_2(x,y) \\ &= (xy+1)^2 + (xy-1)^2 \\ &= 2x^2y^2 + 2 \end{aligned}$$

  By introducing ϕ such that:

$$\begin{aligned} \phi : \quad & \mathbb{R} \to \mathbb{R}^2 \\ & x \mapsto \begin{pmatrix} \sqrt{2}x^2 \\ \sqrt{2} \end{pmatrix} \end{aligned}$$

  We follow the same steps as for K_1 and we obtain that the RKHS $\mathcal{H}$ of K is:

$$f_u(x) = \langle u, \phi(x) \rangle \quad \forall u \in \mathbb{R}^2$$

  Endowed with the inner-product $\langle f_u, f_v \rangle_{\mathcal{H}} = \langle u, v \rangle_{\mathbb{R}^2}$

2.2

Let K_1 and K_2 be two positive definite kernels on a set $\mathcal{X}$ and α and β two positive scalars. Then we consider $K = \alpha K_1 + \beta K_2$ and prove first that it is positive definite.

- Symmetrical:
  $\forall (x,y) \in \mathcal{X}^2$:

$$\begin{aligned} K(x,y) &= \alpha K_1(x,y) + \beta K_2(x,y) \\ &= \alpha K_1(y,x) + \beta K_2(y,x) \\ &= K(y,x) \end{aligned}$$

- Positive:
  $\forall (x_i)_{1 \le i \le N} \in \mathcal{X}^N$ and $\forall (a_i)_{1 \le i \le N} \in \mathbb{R}^N$:

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

We will now investigate the RKHS $\mathcal{H}$ of K. We know that it contains all functions of the form:

$$\begin{aligned}\forall x \in \mathcal{X}, \quad K_x : t \mapsto K(x,t) &= \alpha K_1(x,t) + \beta K_2(x,t) \\ &= \alpha K_{1,x}(t) + \beta K_{2,x}(t) \\ &\in \mathcal{H}_1 + \mathcal{H}_2\end{aligned}$$

Thus our candidate RKHS will be $\mathcal{H} = \mathcal{H}_1 + \mathcal{H}_2$ endowed with the following scalar product:

$$\begin{aligned}\forall (f,g) \in \mathcal{H}^2, \exists (f_1,g_1,f_2,g_2) \in \mathcal{H}_1^2 \times \mathcal{H}_2^2 \\ \langle f,g \rangle_{\mathcal{H}} = \frac{1}{\alpha}\langle f_1,g_1 \rangle_{\mathcal{H}_1} + \frac{1}{\beta}\langle f_2,g_2 \rangle_{\mathcal{H}_2}\end{aligned}$$

We now need to prove the reproducing property $\forall f \in \mathcal{H}, \forall x \in \mathcal{X}$:

$$\begin{aligned}f(x) &= f_1(x) + f_2(x) \\ &= \langle f_1, K_{1,x} \rangle_{\mathcal{H}_1} + \langle f_2, K_{2,x} \rangle_{\mathcal{H}_2} \\ &= \frac{1}{\alpha}\langle f_1, \alpha K_{1,x} \rangle_{\mathcal{H}_1} + \frac{1}{\beta}\langle f_2, \beta K_{2,x} \rangle_{\mathcal{H}_2} \\ &= \langle f_1 + f_2, \alpha K_{1,x} + \beta K_{2,x} \rangle_{\mathcal{H}} \\ &= \langle f, K_x \rangle_{\mathcal{H}}\end{aligned}$$

Which ends the proof as $\mathcal{H}$ is a Hilbert space.

3 Uniqueness of the RKHS

Let $K : \mathcal{X}^2 \to \mathbb{R}$ a positive definite function and let $\mathcal{H}$ and $\mathcal{H}'$ be two possible RKHS associated to K. We now focus on $\mathcal{H}$ but the results are exactly the same for $\mathcal{H}'$.

We consider $\mathcal{H}_0$ the pre-Hilbert space spanned by the functions $\{K_x\}_{x \in \mathcal{X}}$, then from the course we know that $\overline{\mathcal{H}_0}$ is a RKHS with K as a reproducing kernel. Beside we have that $\overline{\mathcal{H}_0} \subset \mathcal{H}$ because $\mathcal{H}_0 \subset \mathcal{H}$ by definition and $\mathcal{H}$ is complete.

Let $f \in \mathcal{H}$ such that $\forall g \in \mathcal{H}_0, \langle f,g \rangle_{\mathcal{H}} = 0$. Thus $\forall x \in \mathcal{X}, \langle f, K_x \rangle_{\mathcal{H}} = f(x) = 0$ and $f = 0$.

As a consequence we obtain that $\mathcal{H}_0$ is dense in $\mathcal{H}$ and finally $\overline{\mathcal{H}_0} = \mathcal{H} = \mathcal{H}'$.

Furthermore, for all $(f,g) \in \mathcal{H}_0$ we can find two decompositions such that:

$$f = \sum_i a_i K_{x_i}, \quad g = \sum_i b_i K_{y_i}$$

Then we have:

$$\begin{aligned}\langle f,g \rangle_{\mathcal{H}_0} &= \sum_{i,j} a_i b_j K(x_i,y_j) \\ &= \sum_i a_i g(x_i) \\ &= \sum_i a_i \langle K_{x_i}, g \rangle_{\mathcal{H}} \\ &= \langle \sum_i a_i K_{x_i}, g \rangle_{\mathcal{H}} \\ &= \langle f,g \rangle_{\mathcal{H}} \\ &= \langle f,g \rangle_{\mathcal{H}'}\end{aligned}$$

Therefore we have proved that $\mathcal{H} = \mathcal{H}'$ and $\langle \cdot,\cdot \rangle_{\mathcal{H}} = \langle \cdot,\cdot \rangle_{\mathcal{H}'}$ which ends the proof of the uniqueness of the RKHS.

Homework 2

Ayman YACHAOUI
Kernel methods in machine learning
Data & Knowledge @TelecomParisTech Liberté
École Normale Supérieure de Cachan
ayman.yachaoui@telecom-paristech.fr

1 Dual coordinate ascent algorithms for SVMs

We recall the primal formulation:

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

And its dual formulation:

$$\begin{array}{ll} \max_{\alpha\in\mathbb{R}^n} & 2\alpha^\intercal y - \alpha^\intercal K \alpha \\ \text{s.t.} & 0 \leq y_i \alpha_i \leq \frac{1}{2\lambda n} \end{array}$$

1.1 One variable update rule

Fixing all coordinates but the j-th we have that the objective function g becomes:

$$\begin{aligned} g(\alpha) &= 2\alpha^\intercal y - \alpha^\intercal K \alpha \\ &= 2\sum_{i=1}^{n} \alpha_i y_i - \sum_{i=1}^{n}\sum_{k=1}^{n} \alpha_i \alpha_k K(x_i, x_k) \\ &= 2\sum_{\substack{i=1 \\ i\neq j}}^{n} \alpha_i y_i - \sum_{\substack{i=1 \\ i\neq j}}^{n}\sum_{\substack{k=1 \\ k\neq j}}^{n} \alpha_i \alpha_k K(x_i, x_k) + 2\alpha_j \left(y_j - \sum_{\substack{i=1 \\ i\neq j}}^{n} \alpha_i K(x_i, x_j) \right) - \alpha_j^2 K(x_j, x_j) \end{aligned}$$

Which is a convex function in α_j thus as the gradient vanishes we have:

$$\nabla_{\alpha_j} g(\alpha) = 0 = 2\left(y_j - \sum_{\substack{i=1 \\ i\neq j}}^{n} \alpha_i K(x_i, x_j) \right) - 2\alpha_j K(x_j, x_j)$$

Which yields the optimal solution α_j^*:

$$\alpha_j^* = \frac{y_j - \sum_{\substack{i=1 \\ i\neq j}}^{n} \alpha_i K(x_i, x_j)}{K(x_j, x_j)}$$

But as α_j^* must satisfy $0 \leq y_j \alpha_j^* \leq \frac{1}{2\lambda n}$, if it's outside this bound we chose the closest boundary that respect this inequality and use it as the optimal solution.

1.2 SVM with intercept dual formulation

The objective function is still increasing with respect to $\|f\|_{\mathcal{H}}^2$ and besides it is convex with respect to b which means that we can optimize for f and then for b thus the representer theorem still applies and the solution satisfies:

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

Where $\hat{\alpha}$ solves:

$$\hat{\alpha} = \underset{\alpha \in \mathbb{R}^n}{\arg\min} \left(\frac{1}{n} \sum_{i=1}^{n} \max \left(0, 1 - y_i \left(\sum_{j=1}^{n} \alpha_j K(x_i, x_j) + b \right) \right) + \lambda \alpha^\intercal K \alpha \right)$$

Which yields the SVM primal formulation:

$$\begin{aligned} \min_{\substack{\alpha \in \mathbb{R}^n \\ \xi \in \mathbb{R}^n}} & \frac{1}{n} \sum_{i=1}^{n} \xi_i + \lambda \alpha^\intercal K \alpha \\ \text{s.t.} \quad & \forall i, \quad \xi_i \geq 0 \\ & \forall i, \quad \xi_i \geq 1 - y_i \left(\sum_{j=1}^{n} \alpha_j K(x_i, x_j) + b \right) \end{aligned}$$

We introduce the Lagrangian multipliers $\mu \in \mathbb{R}^n$ and $\nu \in \mathbb{R}^n$ corresponding to the Lagrangian:

$$\mathcal{L}(\alpha, b, \xi, \mu, \nu) = \frac{1}{n} \sum_{i=1}^{n} \xi_i + \lambda \alpha^\intercal K \alpha - \sum_{i=1}^{n} \nu_i \xi_i + \sum_{i=1}^{n} \mu_i \left(1 - \xi_i - y_i \left(\sum_{j=1}^{n} \alpha_j K(x_i, x_j) + b \right) \right)$$

Or in matrix notations:

$$\mathcal{L}(\alpha, b, \xi, \mu, \nu) = \xi^\intercal \left(\frac{\mathbf{1}}{n} - \nu - \mu \right) + \lambda \alpha^\intercal K \alpha - (\text{diag}(y)\mu)^\intercal K \alpha + \mu^\intercal (by + \mathbf{1})$$

Which is a quadratic function in α minimized when the gradient is null:

$$\begin{aligned} \nabla_\alpha \mathcal{L}(\alpha^*) = 0 &= K \left(2\lambda \alpha^* - \text{diag}(y)\mu \right) \\ \alpha^* &= \frac{\text{diag}(y)\mu}{2\lambda} \end{aligned}$$

Besides it is bounded below only when it is constant in ξ and b thus:

$$\begin{aligned} \mu + \nu &= \frac{\mathbf{1}}{n} \\ \mu^\intercal y &= 0 \end{aligned}$$

The dual problem therefore becomes:

$$\begin{aligned} \max_{\mu \in \mathbb{R}^n} & \mu^\intercal \mathbf{1} - \frac{1}{4\lambda} \mu^\intercal \text{diag}(y) K \text{diag}(y) \mu \\ \text{s.t.} \quad & \mu \succeq 0 \\ & \mu \preceq \tfrac{1}{n} \\ & \mu^\intercal y = 0 \end{aligned}$$

And by noticing that $\text{diag}(y)^2 = \mathbf{I}$ we have $\mu = 2\lambda \text{diag}(y)\alpha$ and the SVM dual formulation is:

$$\begin{aligned} \max_{\alpha \in \mathbb{R}^n} & 2\alpha^\intercal y - \alpha^\intercal K \alpha \\ \text{s.t.} \quad & y_i \alpha_i \geq 0 \\ & y_i \alpha_i \leq \tfrac{1}{2\lambda n} \\ & \textstyle\sum_{i=1}^{n} \alpha_i = 0 \end{aligned}$$

We notice that there is a new constraint $\sum_{i=1}^{n} \alpha_i = 0$ which implies that in order to respect it, at least two coordinates need to be updated at the same time.

1.3 Two variables update rule

To update the i-th and j-th coordinates we use the new constraint which yields:

$$\alpha_j = -\sum_{\substack{k=1 \\ k\neq j}}^{n} \alpha_k$$

And by re-injecting it in the objective function g we obtain:

$$\begin{aligned}
g(\alpha) &= 2\alpha^\intercal y - \alpha^\intercal K \alpha \\
&= 2\sum_{k=1}^{n} \alpha_k y_k - \sum_{l=1}^{n}\sum_{k=1}^{n} \alpha_l \alpha_k K(x_l, x_k) \\
&= 2\sum_{\substack{k=1 \\ k\neq j}}^{n} \alpha_k (y_k - y_j) - \sum_{\substack{l=1 \\ l\neq j}}^{n}\sum_{\substack{k=1 \\ k\neq j}}^{n} \alpha_l \alpha_k K(x_l, x_k) - 2\alpha_j \sum_{\substack{k=1 \\ k\neq j}}^{n} \alpha_k K(x_j, x_k) - \alpha_j^2 K(x_j, x_j) \\
&= 2\sum_{\substack{k=1 \\ k\neq j}}^{n} \alpha_k (y_k - y_j) - \sum_{\substack{l=1 \\ l\neq j}}^{n}\sum_{\substack{k=1 \\ k\neq j}}^{n} \alpha_l \alpha_k K(x_l, x_k) + 2\sum_{\substack{l=1 \\ l\neq j}}^{n}\sum_{\substack{k=1 \\ k\neq j}}^{n} \alpha_l \alpha_k K(x_j, x_k) - \sum_{\substack{l=1 \\ l\neq j}}^{n}\sum_{\substack{k=1 \\ k\neq j}}^{n} \alpha_l \alpha_k K(x_j, x_j) \\
&= 2\sum_{\substack{k=1 \\ k\neq j}}^{n} \alpha_k (y_k - y_j) + \sum_{\substack{l=1 \\ l\neq j}}^{n}\sum_{\substack{k=1 \\ k\neq j}}^{n} \alpha_l \alpha_k \left(2K(x_j, x_k) - K(x_l, x_k) - K(x_j, x_j)\right)
\end{aligned}$$

Therefore by finding the point where the gradient vanishes we have:

$$\begin{aligned}
\nabla_{\alpha_i} g(\alpha) = 0 &= 2(y_i - y_j) + 2\alpha_i(2K(x_j, x_i) - K(x_i, x_i) - K(x_j, x_j)) \\
&\quad + \sum_{\substack{k=1 \\ k\neq i,j}}^{n} \alpha_k (2K(x_j, x_k) - K(x_i, x_k) - K(x_j, x_j)) \\
&\quad + \sum_{\substack{k=1 \\ l\neq i,j}}^{n} \alpha_k (2K(x_j, x_i) - K(x_k, x_i) - K(x_j, x_j)) \\
&= 2(y_i - y_j) + 2\alpha_i(2K(x_j, x_i) - K(x_i, x_i) - K(x_j, x_j)) \\
&\quad + 2\sum_{\substack{k=1 \\ k\neq i,j}}^{n} \alpha_k (K(x_j, x_k) - K(x_i, x_k)) + 2(K(x_j, x_i) - K(x_j, x_j)) \left(\sum_{\substack{k=1 \\ k\neq i,j}}^{n} \alpha_k\right)
\end{aligned}$$

Which yields the optimal value:

$$\alpha_i^* = \frac{y_i - y_j + \sum_{\substack{k=1 \\ k\neq i,j}}^{n} \alpha_k (K(x_j, x_k) - K(x_i, x_k)) + (K(x_j, x_i) - K(x_j, x_j)) \left(\sum_{\substack{k=1 \\ k\neq i,j}}^{n} \alpha_k\right)}{K(x_i, x_i) + K(x_j, x_j) - 2K(x_j, x_i)}$$

As we did earlier, if α_i^* doesn't satisfy $0 \leq y_i \alpha_i^* \leq \frac{1}{2\lambda n}$ then we chose the closest bound that satisfies the inequalities and we then compute α_j^* using the linear constraint:

$$\alpha_j^* = -\alpha_i^* - \sum_{\substack{k=1 \\ k\neq i,j}}^{n} \alpha_k$$

And if afterwards α_j^* doesn't satisfy the inequalities $0 \leq y_j \alpha_j^* \leq \frac{1}{2\lambda n}$ we clip it to the closest boundary and we modify α_i^* by using the same rule:

$$\alpha_i^* = -\alpha_j^* - \sum_{\substack{k=1 \\ k\neq i,j}}^{n} \alpha_k$$

2 Kernel mean embedding

2.1 Mean embedding in Hilbert space

For all $y \in \mathcal{X}$ we have:

$$\begin{aligned}\mu(P)(y) &= \mathbb{E}_{X\sim P}[K(X, y)] \\ &= \mathbb{E}_{X\sim P}[\langle K_X, K_y\rangle_{\mathcal{H}}] \\ &= \langle \mathbb{E}_{X\sim P}[K_X], K_y\rangle_{\mathcal{H}}\end{aligned}$$

Thus the reproducing property holds and:

$$\mu(P) = \mathbb{E}_{X\sim P}[K_X] \in \mathcal{H}$$

2.2 Mean equality

For all $f \in \mathcal{H}$ we have:

$$\begin{aligned}\langle f, \mu(P)\rangle_{\mathcal{H}} &= \langle f, \mathbb{E}_{X\sim P}[K_X]\rangle_{\mathcal{H}} \\ &= \mathbb{E}_{X\sim P}[\langle f, K_X\rangle_{\mathcal{H}}] \\ &= \mathbb{E}_{X\sim P}[f(X)]\end{aligned}$$

Thus $\mu(P) = \mu(Q)$ implies $\mathbb{E}_{X\sim P}[f(X)] = \mathbb{E}_{X\sim Q}[f(X)]$.