Yoav Zemel
Guillaume Obozinski
Ho Yun
Shivang Sachar

Statistical Machine Learning

Exercise sheet 9

Exercise 9.1 (Linear kernel) Consider the function $K : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ defined by $K(\mathbf{x}, \mathbf{y}) \mapsto \mathbf{x}^\top \mathbf{y}$.

(a) Show that K is a symmetric positive-definite function, and by Aronszajn's theorem, a reproducing kernel.

Solution: Notice that, for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$, $K(\mathbf{x}, \mathbf{y}) = K(\mathbf{y}, \mathbf{x})$ and for $\{\mathbf{x}_j\}_{j=1}^n \subset \mathbb{R}^p$ and $\{\alpha_i\}_{i=1}^n \subset \mathbb{R}$,

$$\sum_{i,j=1}^n \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) = \sum_{i,j=1}^n \alpha_i \alpha_j \mathbf{x}_i^\top \mathbf{x}_j = \left\| \sum_{i=1}^n \alpha_i \mathbf{x}_i \right\|^2 \geq 0$$

(b) Let $\mathcal{H}$ be the RKHS with reproducing kernel K defined above. Show that $f \in \mathcal{H}$ if and only if f is a linear function, that is, there exists $\tilde{\mathbf{f}} \in \mathbb{R}^p$ such that $f(\mathbf{x}) = \mathbf{x}^\top \tilde{\mathbf{f}} = K(\mathbf{x}, \tilde{\mathbf{f}})$.

(Hint: One direction is very easy. For the other, you can first show that all the functions $K(\mathbf{x}, \cdot)$ live a finite dimensional space and therefore have a canonical basis on which can we decompose them, and then use the kernel reproducing property to extend this to all functions in $\mathcal{H}$).

Solution:

- We first show that any linear function is in $\mathcal{H}$. Indeed, for any $\mathbf{w} \in \mathbb{R}^p$, the function $\mathbf{x} \mapsto \mathbf{w}^\top \mathbf{x}$ is exactly the function $\mathbf{x} \mapsto K(\mathbf{w}, \mathbf{x})$ that we denoted $K(\mathbf{w}, \cdot)$, and we clearly have $K(\mathbf{w}, \cdot) \in \mathcal{H}$.
- We now show that any function in $\mathcal{H}$ is a linear function. Intuitively and informally, this should be true because a RKHS is a vector space and taking linear combinations of linear functions only produces linear functions, and taking limits of sequences of linear functions also produces linear functions.

The following proof takes a more abstract approach:

Let $\{\mathbf{e}_j\}_{j=1}^p$ is a basis of $\mathbb{R}^p$. We will first show that any $K(\cdot, \mathbf{x})$ is a linear combination of the $K(\cdot, \mathbf{e}_i)$. Indeed,

$$K(\mathbf{y}, \mathbf{x}) = \mathbf{y}^\top \mathbf{x} = \mathbf{y}^\top \sum_{j=1}^p (\mathbf{x}^\top \mathbf{e}_j) \mathbf{e}_j = \sum_{j=1}^p (\mathbf{x}^\top \mathbf{e}_j)(\mathbf{y}^\top \mathbf{e}_j) = \sum_{j=1}^p (\mathbf{x}^\top \mathbf{e}_j) K(\mathbf{y}, \mathbf{e}_j),$$

where $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y}$ for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$. Thus, $K(\cdot, \mathbf{x}) = \sum_{j=1}^p (\mathbf{x}^\top \mathbf{e}_j) K(\cdot, \mathbf{e}_j)$.

Now, if $f \in \mathcal{H}$, then, if $\langle \cdot, \cdot \rangle$ denotes the dot product in $\mathcal{H}$, we have:

$$f(\mathbf{x}) = \langle f, K(\cdot, \mathbf{x}) \rangle = \langle f, \sum_{j=1}^{p} (\mathbf{x}^\top \mathbf{e}_j) K(\cdot, \mathbf{e}_j) \rangle = \sum_{j=1}^{p} (\mathbf{x}^\top \mathbf{e}_j) \langle f, K(\cdot, \mathbf{e}_j) \rangle = \mathbf{x}^\top [\sum_{j=1}^{p} \langle f, K(\cdot, \mathbf{e}_j) \rangle \, \mathbf{e}_j].$$

Thus, $f = K(\cdot, \tilde{\mathbf{f}})$ for $\tilde{\mathbf{f}} = \sum_{j=1}^{p} \langle f, K(\cdot, \mathbf{e}_j) \rangle \, \mathbf{e}_j$, which in particular proves that f is a linear function.

(c) Using only elementary linear algebra (that is, without using any facts about reproducing kernels), show that $\mathcal{H}$ forms a Hilbert space under the inner product $\langle K(\cdot, \mathbf{x}), K(\cdot, \mathbf{y}) \rangle = K(\mathbf{x}, \mathbf{y})$.

Solution: Notice that there is a bijective correspondence between $f \in \mathcal{H}$ and $\tilde{\mathbf{f}} \in \mathbb{R}^p$ given by $f(\mathbf{x}) = \mathbf{x}^\top \tilde{\mathbf{f}}$. Additionally, for $g \in \mathcal{H}$ given by $g(\mathbf{x}) = \mathbf{x}^\top \tilde{\mathbf{g}}$ and $\alpha, \beta \in \mathbb{R}$ we have that: $(\alpha f + \beta g)(\mathbf{x}) = \mathbf{x}^\top (\alpha \tilde{\mathbf{f}} + \beta \tilde{\mathbf{g}})$ and $\langle f, g \rangle = \langle K(\cdot, \tilde{\mathbf{f}}), K(\cdot, \tilde{\mathbf{g}}) \rangle = \tilde{\mathbf{f}}^\top \tilde{\mathbf{g}}$. It follows that $\mathcal{H}$ under its inner product is isomorphic to the Euclidean space $\mathbb{R}^p$ under the inner product $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbb{R}^p} = \mathbf{x}^\top \mathbf{y}$, and is thus a Hilbert space, just like the latter.

Alternatively, we could have used the fact that every finite-dimensional inner-product space is complete, and therefore, a Hilbert space.

Exercise 9.2 (Squared loss regression in RKHS) Let $\mathcal{H}$ denote the RKHS associated to a Mercer kernel K.

(a) *Preliminary questions*

(i) Let $\mathbf{K}$ be a positive semi definite matrix. Show that $\mathbf{K}$ and $(\mathbf{K} + \lambda \mathbf{I})^{-1}$ commute.

Solution: Both $\mathbf{K}$ and $\mathbf{K} + \lambda \mathbf{I}$ are diagonal in the orthogonal eigenbasis of $\mathbf{K}$ and so is $(\mathbf{K} + \lambda \mathbf{I})^{-1}$, this means that these matrices are co-diagonalizable and therefore they commute.

Alternatively, one can write $\mathbf{K}(\mathbf{K} + \lambda \mathbf{I})^{-1} = (\mathbf{K} + \lambda \mathbf{I} - \lambda \mathbf{I})(\mathbf{K} + \lambda \mathbf{I})^{-1} = \mathbf{I} - \lambda (\mathbf{K} + \lambda \mathbf{I})^{-1}$ and similarly $(\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{K}$ is equal to the same expression.

(ii) Deduce from the previous question that if $\mathbf{h} \in \ker(\mathbf{K})$ then so does $(\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{h}$.

Solution: We need to show that if $\mathbf{h}' = (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{h}$ then $\mathbf{K} \mathbf{h}' = 0$ but indeed, using the result of the previous question, $\mathbf{K} \mathbf{h}' = \mathbf{K}(\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{h} = (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{K} \mathbf{h} = 0$.

(iii) Let $\mathbf{K} = \big(K(\mathbf{x}_i, \mathbf{x}_j)\big)_{1 \le i,j \le n}$ with K the above defined Mercer kernel. Show that if $\mathbf{h} \in \ker(\mathbf{K})$ then the function $\sum_{i=1}^{n} h_i K(x_i, \cdot)$ is constant and equal to 0.

Solution: We compute the norm of this function in $\mathcal{H}$:

$$\| \sum_{i=1}^{n} h_i K(x_i, \cdot) \|_{\mathcal{H}}^2 = \big\langle \sum_{i=1}^{n} h_i K(x_i, \cdot), \sum_{j=1}^{n} h_j K(x_j, \cdot) \big\rangle_{\mathcal{H}} = \sum_{i=1}^{n} \sum_{j=1}^{n} h_i h_j \langle K(x_i, \cdot), K(x_j, \cdot) \rangle_{\mathcal{H}}.$$

But then using $K(x_i, x_j) = \langle K(x_i, \cdot), K(x_j, \cdot) \rangle_{\mathcal{H}}$, we have

$$\| \sum_{i=1}^{n} h_i K(x_i, \cdot) \|_{\mathcal{H}}^2 = \mathbf{h}^\top \mathbf{K} \mathbf{h} = 0,$$

since $\mathbf{h} \in \ker(\mathbf{K})$. This shows that this function is the zero function.

(b) Show that the solution to the regression problem

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

is $\widehat{f}(\mathbf{x}) = \sum_{i=1}^n \widehat{\alpha}_i K(\mathbf{x}, \mathbf{x}_i)$ with $\widehat{\boldsymbol{\alpha}} = (\mathbf{K} + n\lambda\mathbf{I})^{-1}\boldsymbol{y}$, where $\mathbf{K}$ is the Gram matrix associated to K.

Solution: By the representer theorem, the function solution is of the form $\widehat{f}(\mathbf{x}) = \sum_{i=1}^n \widehat{\alpha}_i K(\mathbf{x}, \mathbf{x}_i)$, so we only need to find the form of the vector $\widehat{\boldsymbol{\alpha}}$. By substituting $\widehat{f}(\mathbf{x}) = \sum_{i=1}^n \widehat{\alpha}_i K(\mathbf{x}, \mathbf{x}_i)$ we get,

$$\begin{aligned}
\widehat{\boldsymbol{\alpha}} &= \operatorname*{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^n} \frac{1}{n}\sum_{i=1}^{n}\Bigg\{y_i - \underbrace{\sum_{j=1}^{n}\alpha_j K(\mathbf{x}_i, \mathbf{x}_j)}_{=(\mathbf{K}\boldsymbol{\alpha})_i}\Bigg\}^2 + \lambda\boldsymbol{\alpha}^\top\mathbf{K}\boldsymbol{\alpha} \\
&= \operatorname*{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^n} \frac{1}{n}\|\boldsymbol{y} - \mathbf{K}\boldsymbol{\alpha}\|^2 + \lambda\boldsymbol{\alpha}^\top\mathbf{K}\boldsymbol{\alpha};
\end{aligned}$$

since $\|f\|_{\mathcal{H}}^2 = \boldsymbol{\alpha}^\top\mathbf{K}\boldsymbol{\alpha}$. To solve this optimization problem we differentiate with respect to $\boldsymbol{\alpha}$, giving the normal equation

$$-\mathbf{K}^\top(\boldsymbol{y} - \mathbf{K}\boldsymbol{\alpha}) + n\lambda\mathbf{K}\boldsymbol{\alpha} = \mathbf{0} \quad \Leftrightarrow \quad \mathbf{K}(-\boldsymbol{y} + \mathbf{K}\boldsymbol{\alpha} + n\lambda\boldsymbol{\alpha}) = \mathbf{0},$$

using the symmetry of the Gram matrix $\mathbf{K}$. Thus, $(-\boldsymbol{y} + \mathbf{K}\boldsymbol{\alpha} + n\lambda\boldsymbol{\alpha}) \in \operatorname{Ker} \mathbf{K}$, that is $\boldsymbol{\alpha} = \widehat{\boldsymbol{\alpha}} + \boldsymbol{h}$, where $\widehat{\boldsymbol{\alpha}} = (\mathbf{K} + n\lambda\mathbf{I})^{-1}\boldsymbol{y}$ and $\boldsymbol{h} = (\mathbf{K} + n\lambda\mathbf{I})^{-1}\boldsymbol{h}'$ such that $\boldsymbol{h}' \in \operatorname{Ker} \mathbf{K}$.

But by (a.ii), $\boldsymbol{h} \in \ker(\mathbf{K})$. And so, $\widehat{f}(\mathbf{x}) = \sum_{i=1}^n \widehat{\alpha}_i K(\mathbf{x}, \mathbf{x}_i) + \sum_{i=1}^n h_i K(\mathbf{x}, \mathbf{x}_i) = \sum_{i=1}^n \widehat{\alpha}_i K(\mathbf{x}, \mathbf{x}_i)$ because $\sum_{i=1}^n h_i K(\cdot, \mathbf{x}_i) = 0$, by (a.iii).

(c) Using the above result show that the solution to the ridge regression problem with no intercept,

$$\widehat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta}} \frac{1}{n}\|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_2^2,$$

where $\boldsymbol{y} \in \mathbb{R}^n$ and the design matrix $\mathbf{X}$ is $n \times p$ is given by $\widehat{\boldsymbol{\beta}} = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + n\lambda\mathbf{I})^{-1}\boldsymbol{y}$.

Solution: Let $K : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ given by $K(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top\mathbf{y}$. Then

$$\|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|^2 = \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 = \sum_{i=1}^{n}\left(y_i - K(\boldsymbol{\beta}, \mathbf{x}^i)\right)^2$$

where $\mathbf{x}^i$ is ith column of $\mathbf{X}^\top$ and $\|\boldsymbol{\beta}\|_2^2 = \|K(\cdot, \boldsymbol{\beta})\|_{\mathcal{H}}^2$. Using Exercise 10.1 (b), the problem can be equivalently stated as:

$$\widehat{f} = \operatorname*{argmin}_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}\left(y_i - f(\mathbf{x}^i)\right)^2 + \lambda\|f\|_{\mathcal{H}}^2,$$

with $\widehat{f} = K(\cdot, \widehat{\boldsymbol{\beta}})$. By the result of Kimmeldorf and Wahba, it follows that $\widehat{f}(\mathbf{x}) = \sum_{i=1}^n \widehat{\alpha}_i K(\mathbf{x}, \mathbf{x}^i) = K(\mathbf{x}, \sum_{i=1}^n \widehat{\alpha}_i\mathbf{x}^i)$ with $\widehat{\boldsymbol{\alpha}} = (\mathbf{K} + n\lambda\mathbf{I})^{-1}\boldsymbol{y} = (\mathbf{X}\mathbf{X}^\top + n\lambda\mathbf{I})^{-1}\boldsymbol{y}$. It follows that $\widehat{\boldsymbol{\beta}} = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + n\lambda\mathbf{I})^{-1}\boldsymbol{y}$.

Exercise 9.3 (Ridge regression and kernel trick) Consider again, the ridge regression problem with no intercept,

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \ \frac{1}{n}\|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_2^2,$$

where $\boldsymbol{y} \in \mathbb{R}^n$ and the design matrix $\mathbf{X}$ is $n \times p$.

(a) Using what you know about ridge regression and the identity, $(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)\mathbf{X}^\top = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_n)$, show that $\widehat{\boldsymbol{\beta}} = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + n\lambda\mathbf{I}_n)^{-1}\boldsymbol{y}$ as in the previous problem.

Solution: The conclusion follows by multiplying both sides of the identity by $(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_n)^{-1}\boldsymbol{y}$ from right and by $(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)^{-1}$.

(b) Thus, there are two methods for computing $\widehat{\boldsymbol{\beta}}$: $\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + n\lambda\mathbf{I}_n)^{-1}\boldsymbol{y}$ and $(\mathbf{X}^\top\mathbf{X} + n\lambda\mathbf{I}_n)^{-1}\mathbf{X}^\top\boldsymbol{y}$. What is the computational complexity of applying each method? When should one be favored over the other?

Solution: The two costly operations involved are matrix multiplication and that of solving linear equations. The complexity of solving $\mathbf{A}\mathbf{x} = \mathbf{y}$, where $\mathbf{A}$ is a $n \times n$ matrix and $\mathbf{x}, \mathbf{y}$ matrices, is $n \times 1$ is $O(n^3)$ while that of multiplying a $p \times q$ matrix $\mathbf{B}$ with a $q \times r$ matrix $\mathbf{C}$ is $O(pqr)$.

Using these facts, one can compute that the complexity of evaluating $\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + n\lambda\mathbf{I})^{-1}\boldsymbol{y}$ is $O(n^3 + n^2p)$ while that of evaluating $(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^\top\boldsymbol{y}$ is $O(p^3 + p^2n)$. When $p < n$, the latter is a better method for calculating $\widehat{\boldsymbol{\beta}}$.

Exercise 9.4 (RKHS and the representer theorem) Suppose that K has an eigen-expansion

$$K(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{\infty} \gamma_j \phi_j(\mathbf{x})\phi_j(\mathbf{y}), \tag{1}$$

where $\gamma_j \geq 0$ are eigenvalues that satisfy $\sum_{j=1}^{\infty} |\gamma_j|^2 < \infty$ and $\{\phi_j\}_{j=1}^{\infty}$ forms the orthogonal basis of the function space $\mathcal{H}$. The space $\mathcal{H}$ has the form

$$\mathcal{H} = \left\{ f : \mathbb{R}^p \to \mathbb{R} : f(\mathbf{x}) = \sum_{i=1}^{\infty} c_i\phi_i(\mathbf{x}) \text{ for all } \mathbf{x} \text{ and } \sum_{i=1}^{\infty} c_i^2/\gamma_i < \infty \right\}$$

For $f(\mathbf{x}) = \sum_{i=1}^{\infty} c_i\phi_i(\mathbf{x})$ and $g(\mathbf{x}) = \sum_{i=1}^{\infty} d_i\phi_i(\mathbf{x})$ in $\mathcal{H}$, define

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{i=1}^{\infty} \frac{c_i d_i}{\gamma_i}.$$

NOTE: In the following problems, do not use any results about reproducing kernels.

(a) Show that $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is an inner product.

Solution: It suffices to verify that for $\alpha, \beta \in \mathbb{R}$ and $f(\mathbf{x}) = \sum_{i=1}^{\infty} c_i\phi_i(\mathbf{x})$, $g(\mathbf{x}) = \sum_{i=1}^{\infty} d_i\phi_i(\mathbf{x})$ and $h(\mathbf{x}) = \sum_{i=1}^{\infty} e_i\phi_i(\mathbf{x})$ with $(c_i)_{i=1}^{\infty}, (d_i)_{i=1}^{\infty}, (e_i)_{i=1}^{\infty} \in \ell^2$ we have,

1. *Positivity:* $\langle f, f\rangle_{\mathcal{H}} = \sum_{i=1}^{\infty} c_i^2/\gamma_i \geq 0$ with equality if and only if $c_i = 0$ for $i \geq 1$, that is, if and only if $f = 0$.
2. *Symmetry:* $\langle f, g\rangle_{\mathcal{H}} = \sum_{i=1}^{\infty} c_i d_i/\gamma_i = \langle g, f\rangle_{\mathcal{H}}$.
3. *Linearity:* $\langle \alpha f + \beta g, h\rangle_{\mathcal{H}} = \sum_{i=1}^{\infty} (\alpha c_i + \beta d_i) e_i/\gamma_i = \alpha\langle f, h\rangle_{\mathcal{H}} + \beta\langle g, h\rangle_{\mathcal{H}}$.

(b) For any $f \in \mathcal{H}$ and $\mathbf{x} \in \mathbb{R}^p$, show that $\langle K(\cdot, \mathbf{x}), f\rangle_{\mathcal{H}} = f(\mathbf{x})$.

Solution: By definition, we have

$$\begin{aligned}
\langle K(\cdot, \mathbf{x}), f\rangle_{\mathcal{H}} &= \left\langle \sum_{j=1}^{\infty} \gamma_j \phi_j(\cdot)\phi_j(\mathbf{x}), \sum_{j=1}^{\infty} c_j \phi_j(\cdot) \right\rangle_{\mathcal{H}} \\
&= \sum_{j=1}^{\infty} \frac{\gamma_j \phi_j(\mathbf{x}) c_j}{\gamma_j} \\
&= f(\mathbf{x}).
\end{aligned}$$

(c) For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$, show that $\langle K(\cdot, \mathbf{x}), K(\cdot, \mathbf{y})\rangle_{\mathcal{H}} = K(\mathbf{x}, \mathbf{y})$.

Solution: We have

$$\begin{aligned}
\langle K(\cdot, \mathbf{x}), K(\cdot, \mathbf{y})\rangle_{\mathcal{H}} &= \left\langle \sum_{j=1}^{\infty} \gamma_j \phi_j(\cdot)\phi_j(\mathbf{x}), \sum_{j=1}^{\infty} \gamma_j \phi_j(\cdot)\phi_j(\mathbf{y}) \right\rangle_{\mathcal{H}} \\
&= \sum_{j=1}^{\infty} \frac{\gamma_j \phi_j(\mathbf{x}) \gamma_j \phi_j(\mathbf{y})}{\gamma_k} = K(\mathbf{x}, \mathbf{y}).
\end{aligned}$$

(d) If $f(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i K(\mathbf{x}, \mathbf{x}_i)$ and $g(\mathbf{x}) = \sum_{j=1}^{k} \beta_j K(\mathbf{x}, \mathbf{x}_j)$, show that

$$\langle f, g\rangle_{\mathcal{H}} = \sum_{i=1}^{m}\sum_{j=1}^{k} \alpha_i \beta_j K(\mathbf{x}_i, \mathbf{x}_j)$$

and in particular,

$$\|f\|_{\mathcal{H}}^2 = \sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j).$$

Solution: Using the result of (c), we have

$$\begin{aligned}
\langle f, g\rangle_{\mathcal{H}} &= \left\langle \sum_{i=1}^{m} \alpha_i K(\cdot, \mathbf{x}_i), \sum_{j=1}^{k} \beta_j K(\cdot, \mathbf{x}_j) \right\rangle_{\mathcal{H}} = \sum_{i=1}^{m}\sum_{j=1}^{k} \alpha_i \beta_j \langle K(\cdot, \mathbf{x}_i), K(\cdot, \mathbf{x}_j)\rangle_{\mathcal{H}} \\
&= \sum_{i=1}^{m}\sum_{j=1}^{k} \alpha_i \beta_j K(\mathbf{x}_i, \mathbf{x}_j)
\end{aligned}$$

and the result for $\|f\|_{\mathcal{H}}^2$ follows from $\|f\|_{\mathcal{H}}^2 = \langle f, f\rangle_{\mathcal{H}}$.