Homework 9 - Solution

1. Let $x_1, \dots, x_n \in \mathbb{R}^+$ and $\lambda_1, \dots, \lambda_n \in \mathbb{R}$. We have

$$\sum_{1 \le i,j \le n} \lambda_i \lambda_j \min(x_i, x_j) = \sum_{1 \le i,j \le n} \lambda_i \lambda_j \int_0^\infty \mathbf{1}\{s \le \min(x_i, x_j)\} \mathrm{d}s$$
$$= \int_0^\infty \sum_{1 \le i,j \le n} \lambda_i \lambda_j \mathbf{1}\{s \le x_i\} \mathbf{1}\{s \le x_j\} \mathrm{d}s$$
$$= \int_0^\infty \left(\sum_{i=1}^n \lambda_i \mathbf{1}\{s \le x_i\} \right)^2 \mathrm{d}s \ge 0.$$

2.
   - Let $\mathcal{X} = \mathbb{R}$ and $k(x, x') = xx'$.
   - Let $\mathcal{X} = \mathbb{R}$ and $k(x, x') = \sqrt{|x + x'|}$. The function k is not a kernel. Indeed, let $x_1 = 1$, $x_2 \ge 0$, $\lambda_1 = 1$, $\lambda_2 = -1$. Then, for k to be a kernel, we should have

$$\lambda_1^2 k(x_1, x_1) + 2\lambda_1 \lambda_2 k(x_1, x_2) + \lambda_2^2 k(x_2, x_2) \ge 0.$$

   However, this is equal to $\sqrt{2} - 2\sqrt{1 + x_2} + \sqrt{2x_2}$, which is negative by concavity of the square root function.

3. (a) We have

$$\mathrm{Err}(B) = \sum_{i=1}^{n} \|\tilde{\Phi}_B(\mathbf{x_i}) - \Phi(\mathbf{x_i})\|_{\mathcal{H}}^2$$

$$= \sum_{i=1}^{n} \|\Phi(\mathbf{x_i})\|_{\mathcal{H}}^2 + \|\tilde{\Phi}_B(\mathbf{x_i})\|_{\mathcal{H}}^2 - 2\langle \tilde{\Phi}_B(\mathbf{x_i}), \Phi(\mathbf{x_i})\rangle_{\mathcal{H}}$$

$$= \sum_{i=1}^{n} k(\mathbf{x_i}, \mathbf{x_i}) + \|\sum_{j=1}^{m} B_{ij}\Phi(\mathbf{x_j})\|_{\mathcal{H}}^2 - 2\langle \sum_{j=1}^{m} B_{ij}\Phi(\mathbf{x_j}), \Phi(\mathbf{x_i})\rangle_{\mathcal{H}}$$

$$= \sum_{i=1}^{n} k(\mathbf{x_i}, \mathbf{x_i}) + \sum_{1\le j,j'\le m} B_{ij}B_{ij'}\langle \Phi(\mathbf{x_j}), \Phi(\mathbf{x_{j'}})\rangle_{\mathcal{H}} - 2\sum_{j=1}^{m} B_{ij}k(\mathbf{x_j}, \mathbf{x_i})$$

$$= \sum_{i=1}^{n} k(\mathbf{x_i}, \mathbf{x_i}) + \sum_{1\le j,j'\le m} B_{ij}B_{ij'}k(\mathbf{x_j}, \mathbf{x_{j'}}) - 2\sum_{j=1}^{m} B_{ij}k(\mathbf{x_j}, \mathbf{x_i}).$$

(b) The function $B \mapsto \mathrm{Err}(B)$ is a convex quadratic function. Therefore, its minimum is obtained by finding where the gradient vanishes. The derivative of the map $B \mapsto \mathrm{Err}(B)$ with respect to a fixed entry B_{ij} is given by

$$\partial_{B_{ij}}\mathrm{Err}(B) = -2\mathbf{G}_{ij} + 2\sum_{j'=1}^{m} B_{ij'}\mathbf{G}_{jj'} = -2\mathbf{G}_{ij} + 2(B\mathbf{G}^{mm})_{ij},$$

where we use for the last equality that $\mathbf{G}$ is symmetric. If $B = \mathbf{G}^{nm}(\mathbf{G}^{mm})^{-1}$, this expression is equal to

$$-2\mathbf{G}_{ij} + 2\mathbf{G}_{ij}^{nm} = 0.$$

Therefore, this value of B attains the minimal error.

(c) Let $1 \le i, i' \le n$. By definition, $\tilde{\mathbf{G}}_{ii'} = \langle \tilde{\Phi}(\mathbf{x_i}), \tilde{\Phi}(\mathbf{x_{i'}})\rangle_{\mathcal{H}}$. This is equal to

$$\sum_{1\le j,j'\le m} B_{ij}B_{i'j'}\langle \Phi(\mathbf{x_j}), \Phi(\mathbf{x_{j'}})\rangle_{\mathcal{H}} = \sum_{1\le j,j'\le m} B_{ij}B_{i'j'}k(\mathbf{x_j}, \mathbf{x_{j'}})$$
$$= (B\mathbf{G}^{mm}B^{\top})_{ii'}.$$

For $B = \mathbf{G}^{nm}(\mathbf{G}^{mm})^{-1}$, the matrix $B\mathbf{G}^{mm}B^\top$ is equal to

$$\begin{aligned} B\mathbf{G}^{mm}B^\top &= \mathbf{G}^{nm}(\mathbf{G}^{mm})^{-1}G^{mm}(\mathbf{G}^{mm})^{-1}(\mathbf{G}^{nm})^\top \\ &= \mathbf{G}^{nm}(\mathbf{G}^{mm})^{-1}(\mathbf{G}^{nm})^\top, \end{aligned}$$

where we use that $\mathbf{G}^{mm}$ is symmetric.

(d) Let $U = \mathbf{G}^{nm}$, $V = (\mathbf{G}^{mm})^{-1}/(\lambda n)$ and $W = (\mathbf{G}^{nm})^\top$. Then, we have $\tilde{\mathbf{G}}/(\lambda n) = UVW$, and the solution of kernel ridge regression with feature map $\tilde{\Phi}_B$ is given by

$$\hat{a} = (\tilde{\mathbf{G}} + \lambda n \mathrm{Id}_n)^{-1}\mathbf{Y} = \frac{1}{\lambda n}\left(UVW + \mathrm{Id}_n\right)^{-1}\mathbf{Y}.$$

Also, by using the Sherman-Woodbury-Morrison formula, we have

$$(\mathrm{Id}_n + UVW)^{-1}\mathbf{Y} = \mathbf{Y} - U(V^{-1} + WU)^{-1}W\mathbf{Y}.$$

To compute the last term, we first inverse the matrix $(V^{-1} + WU)$ ($O(m^3)$ operations), then compute the product $W\mathbf{Y}$ ($O(nm)$ operations), compute the product $(V^{-1} + WU)^{-1}W\mathbf{Y}$ ($O(m^3)$ operations), and then eventually multiply by U ($O(nm^2)$ operations). As $m \leq n$, the total number of operations needed is $O(nm^2)$.

Also, storing $\tilde{\mathbf{G}}$ requires to store $nm + m^2 \leq 2nm$ number, which is much smaller n^2 if $m \ll n$.