Problem 0.

Let $\mathcal{L}_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ be our learning sample, where $y_i \in \mathbb{R}$, $x_i \in X$, for some some non-empty set X. Let $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ be a cost function, and $\mathcal{H}$ be a RKHS with reproducing kernel $K(\cdot, \cdot)$ on $X \times X$. We are looking for a solution to the problem

$$f^*(x) = \arg\min_{f \in \mathcal{H}} \sum_{i=1}^{n} \ell(f(x_i), y_i) + \Omega(||f||_{\mathcal{H}}) \,, \tag{1}$$

where $\Omega : \mathbb{R}_+ \to \mathbb{R}$ is a strictly increasing function, and $|| \cdot ||_{\mathcal{H}}$ the norm induced by the dot product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ defined on $\mathcal{H}$.

(a) Argue with some reasonably amount of details why $||f||_{\mathcal{H}} < \infty$ ensures that f is a relatively smooth function.

(b) Our goal is now to derive a general expression for the solution to the problem (1). Let $f \in \mathcal{H}$. Using the reproducing property, show that

$$f(x_i) = \sum_{j=1}^{n} \alpha_j K(x_j, x_i) \,,$$

for some coefficients α_j.

(c) Conclude that necessarily the solution $f^*(x)$ to (1) must satisfy

$$f^*(x) = \sum_{j=1}^{n} \alpha_j K(x, x_i) \,, \quad \forall x \in X \,,$$

for some coefficients α_j.

(d) We are now looking for a regression function $f \in \mathcal{H}$ which minimizes the penalized sum of squares

$$C(f) = \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda ||f||_{\mathcal{H}}^2 \,.$$

It directly follows from question *(c)* that the solution can be written $f^*(x) = \sum_{i=1}^{n} \alpha_i K(x, x_i)$.

(i) Show that the problem of solving $f^* = \arg\min C(f)$ is equivalent to

$$\alpha^* = \arg\min_{\alpha} \, (y - K\alpha)^t (y - K\alpha) + \lambda \alpha^t K \alpha \,,$$

where $\alpha^t = (\alpha_1, \ldots, \alpha_n)$, $y^t = (y_1, \ldots, y_n)$, and some matrix K whose entries you will derive.

(ii) Derive the optimal solution α^* to the problem derived in *(d)*.

(e) Consider now the problem of binary classification, with $y_i \in \{-1, 1\}$. Which loss ℓ and penalty Ω would you use to turn (1) into a kernel SVM problem? Derive the primal and dual optimisation problem of the kernel SVM, and give an expression for the final classifier.

Problem 1.

(i) Take some $X \subset \mathbb{R}^p$. Show that for $f : X \to \mathbb{R}$, the function $K(x, y) = f(x)f(y)$ is a valid kernel.

(ii) Put $X = [-2, 2]^2$, and consider the set of functions on $[-2, 2]$ defined by the kernel $\mathcal{K}(x, y) = 1 + xy \exp(x + y)$.

(a) Argue that $\mathcal{K}$ is a legitimate kernel function.

(b) Show that $g(x) = 1$ and $h(x) = x \exp(x)$ both belong to the RKHS $\mathcal{H}$ with kernel $\mathcal{K}$.

(c) Determine whether or not g and h are orthonormal. If they are not, find an orthonormal basis for the span of $\{g, h\}$ in the RKHS with kernel $\mathcal{K}$.

Problem 2.
We consider real functions on the compact interval $X = [-\pi, \pi]$ with periodic boundary conditions. A Fourier series expansion yields the representation

$$f(x) = \sum_{l=-\infty}^{+\infty} f_l \, e^{ilx} = \sum_{l=-\infty}^{+\infty} f_l \, \psi_l(x) \, .$$

where we put $\psi_l(x) := \exp(ilx)$, Since $f(x)$ is real, the Fourier coefficients satisfy $f_{-l} = \bar{f}_l$, where $\bar{z}$ denotes the complex conjugate of z. Consider a Kernel which takes a single argument corresponding to the difference of the inputs, $\mathcal{K}(x, y) = K(x - y)$, with Fourier representation,

$$K(x) = \sum_{l=-\infty}^{+\infty} k_l \, \psi_l(x) \, , \tag{2}$$

where the coefficients satisfy $k_{-l} = k_l$ and $\bar{k}_l = k_l$, assuming K to be a symmetric real function. Let $\mathcal{H}$ be the set of functions of the form

$$\mathcal{H} = \left\{ f : X \to \mathbb{R} \;\mid\; f(x) = \sum_l f_l \, \psi_l(x) \right\} \, ,$$

endowed with the dot product

$$\langle f, \, g \rangle_{\mathcal{H}} := \sum_l \frac{f_l \, \bar{g}_l}{k_l} \, . \tag{3}$$

It can be shown that $\mathcal{H}$ is an RKHS associated with $\mathcal{K}$, provided $||f||_{\mathcal{H}} < \infty$, where $|| \cdot ||_{\mathcal{H}}$ is the norm induced by the dot product.

(i) Verify that the reproducibility property for f holds,

$$f(x) = \langle f, \, \mathcal{K}(\cdot, x) \rangle_{\mathcal{H}} \, ,$$

(ii) Check that the reproducibility property holds as well for the kernel itself,

$$\langle \mathcal{K}(\cdot, x),\, \mathcal{K}(\cdot, y)\rangle_{\mathcal{H}} = \mathcal{K}(x, y)\,.$$

(iii) You decide to perform a kernel ridge regression with the kernel (2).

(a) Write down the penalized sum of squares objective function you want to minimize.

(b) Using the representer theorem, provide a general expression of the minimizer.

(c) Using the definition of the dot product (3), explain why the penalty term derived in question *(ii)(a)* favours smooth solutions.

Problem 3. *Gaussian and Laplace kernels*

(i) Consider the Gaussian kernel on $X = \mathbb{R}$,

$$\mathcal{K}(x, y) = K(x - y) = \exp\left(-\frac{1}{2}(x - y)^2\right)\,.$$

We define an RKHS with inner product

$$\langle f, g\rangle_{\mathcal{H}} = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \frac{\hat{f}(\omega)\overline{\hat{g}(\omega)}}{\hat{\kappa}(\omega)} d\omega\,,$$

where $\hat{f}$ denotes the Fourier transform of f,

$$\hat{f}(\omega) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} f(x) e^{-i\omega x} dx\,, \quad \text{and} \quad f(x) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \hat{f}(\omega) e^{i\omega x} d\omega\,.$$

Given a function $f(x) = \exp\left(-ax^2\right) \in \mathcal{H}$, with $a > 0$, what is the minimum a for which $||f||_{\mathcal{H}} < \infty$?

Hint: You may use the known results that $e^{-x^2/2}$ has Fourier transform $e^{-\omega^2/2}$, and that $f(ax)$ has Fourier transform $a\hat{f}(\omega/a)$.

(ii) Define the Laplace kernel on $\mathbb{R}$,

$$\mathcal{K}(x, y) = K(x - y) = \exp\left(-\frac{1}{2}|x - y|\right)\,,$$

with Fourier transform

$$\hat{\kappa}(\omega) = \frac{2}{1 + \omega^2}\,.$$

Given the inner product in question *(i)*, comment on the smoothing penalty enforced by the RKHS norm $||f||_{\mathcal{H}}$ for the Gaussian kernel, versus that with the Laplace kernel.

Problem 4. *Kernel SVM*
Consider a binary classification problem, with the following training dataset, with input variable $x \in \mathbb{R}$,

x_i	1	2	3	5
y_i	1	-1	1	-1

(i) Is there a linear classifier based only on x with zero training error?

(ii) Is there a kernel SVM classifier based on the kernel $\mathcal{K}(x, y) = (1 + xy)^2$ with zero training error?

(iii) Is there a kernel SVM classifier based on the kernel $\mathcal{K}(x, y) = \exp(-2(x - y)^2)$ with zero training error?

Problem 5. Consider the space of functions

$$\mathcal{H} := \{f : [0,\ 1] \to \mathbb{R} \,|\, \text{absolutely continuous}, f(0) = 0,\ f' \in L_2[0, 1]\},$$

where $L_2[0, 1]$ denotes the space of square integrable functions on the interval $[0, 1]$. The space $\mathcal{H}$ is endowed with the bilinear form

$$\langle f, g \rangle_{\mathcal{H}} := \int_0^1 f'(x) g'(x) dx\,.$$

Show that $\mathcal{H}$ is an RKHS with reproducing kernel $K(x, y) = \min(x, y)$.

You do not need to show that $\mathcal{H}$ is complete, but you need to show everything else; in particular that the bilinear form $\langle f, g \rangle_{\mathcal{H}}$ is an inner product on $\mathcal{H}$.

Problem 6. Let $\mathcal{H}$ be an RKHS with reproducing kernel K. Solve the kernel logistic regression problem

$$\min_{f \in \mathcal{H}} \sum_{i=1}^{n} \log\left(1 + e^{-y_i f(x_i)}\right) + \frac{\lambda}{2} ||f||^2_{\mathcal{H}}$$

using a Newton procedure. Show that each iteration of the algorithm corresponds to a new weighted kernel ridge regression problem, that you will make explicit.