Suggestions, PS 9

1. (More on RHKS). Note that we have two Hilbert spaces here. $L^2(F)$ is the Hilbert space of all square integrable functions, i.e. such that $\int_F f^2(x)dx < \infty$. The inner product on $L^2(F)$ is defined as $\langle f(x), g(x)\rangle_{L^2(F)} = \int_F f(x)g(x)dx$. Note that the inner product in the smaller subspace $\mathcal{H} \subset L^2(F)$ is $\langle f(x), g(x)\rangle_{\mathcal{H}}$. Also, we assume function $f \in \mathcal{H}$ iff $||f||^2_{\mathcal{H}} = \langle f, f\rangle < \infty$. Note that in general the values of γ_k are assumed to go to 0, so that convergence of the norm $||g||^2_{\mathcal{H}} = \sum_k a_k^2/\gamma_k$ of the function $g(x) = \sum_k a_k\phi_k(x)$ requires the coefficients a_k to go to 0 faster than the condition of finiteness of $||g||^2_{L^2(F)} = \sum_k a_k^2$. Consider the example where the functions ϕ_k are just the Fourier series functions $\sin kx$ and $\cos kx$ on the interval $F = [-\pi, \pi] \subset \mathbb{R}$.

If we require the coefficients a_k and b_k to go to 0 rapidly in the Fourier series $g(x) = \sum_k a_k \cos kx + b_k \sin kx$, this (as we have shown) will make g smoother. Thus the condition that $g \in \mathcal{H}$, i.e., that $||g||^2_{\mathcal{H}} = \sum_k a_k^2/\gamma_k < \infty$, is a smoothing condition and essentially requires that our Hilbert space $\mathcal{H}$ be a space of smooth functions on the same domain F as $L^2(F)$.

Note more generally that the requirement in the Lagrangian that $||g||^2_{\mathcal{H}}$ be small is a requirement that that a_k go to 0 faster as $k \to \infty$, the bigger the $1/\gamma_k$ are. Again this becomes a smoothness requirement, since the 'unsmooth' parts of g are the components with high k.

(a) Again show it is closed under addition, etc., and that the inner product defined satisfies the right properties.

(b) For $\mathcal{H}$ to be an RKHS, the linear functional $l(f) = f(\mathbf{x})$ must be bounded for any fixed $\mathbf{x}$ (see notes). Show for fixed $\mathbf{x}$ we need $|f(\mathbf{x})| \le C||f||_{\mathcal{H}}$ (for all f). Show it suffices that

$$\sum_k \gamma_k \phi(\mathbf{x}_1)^2 < A < \infty \quad \text{for all } \mathbf{x}_1 \in F \text{ with some constant } A. \tag{1}$$

How can you simplify condition (1)? Note that if $\sum_k \gamma_k < \infty$ then $\sum_k \gamma_k \phi(\mathbf{x}_1)^2 \le M^2 \sum_k \gamma_k$.

Note also the Schwarz inequality

$$\left|\sum_k a_k b_k\right| \le \sqrt{\sum_k a_k^2 \sum_k b_k^2}\ .$$

Show that then that if $f(\mathbf{x}) = \sum_k c_k \phi_k(\mathbf{x})$, then

$$|f(\mathbf{x})| = \Big|\sum_k c_k\phi_k(\mathbf{x})\Big| = \Big|\sum_k \frac{c_k}{\sqrt{\gamma_k}}\big(\sqrt{\gamma_k}\phi_k(\mathbf{x})\big)\Big| \le \Big(\sum_k \frac{c_k^2}{\gamma_k}\Big)\Big(\sum_k \gamma_k\phi_k^2(\mathbf{x})\Big).$$

(c) How about $K(\mathbf{x},\mathbf{y}) = \sum_k \gamma_k\phi_k(\mathbf{x})\phi_k(\mathbf{y})$ - does this work? Explain carefully. Show that if $f = \sum_k c_k\phi_k(\mathbf{x}) \in \mathcal{H}$, then

$$\langle K(\mathbf{x},\,\cdot\,), f(\,\cdot\,)\rangle_{\mathcal{H}} = \Big\langle \sum_k \gamma_k\phi_k(\mathbf{x})\phi_k(\,\cdot\,), \sum_l c_l\phi_l(\,\cdot\,)\Big\rangle_{\mathcal{H}} = f(\mathbf{x}).$$

2. (Hastie 5.15)

Problem logic: The logic of this problem is a bit more general than the previous one. Here we start with any positive definite function $K(x,y)$ on all of $\mathbb{R}^p$ (i.e., so K is symmetric and positive definite); additionally assume $K(x,y)$ is continuous, making it a Mercer kernel. We can then construct a unique RKHS $\mathcal{H} \equiv \mathcal{H}_K$ of functions based on this kernel, as was done in the notes (this is done on all of $\mathbb{R}^p$ here).

To show that such examples are different than in the last problem consider the case where $K(x,y) = (x\cdot y + 1)^d$; it can be shown (see problem 5.16) that in this case $\mathcal{H}_K$ is the space of all *polynomials* of degree less than d in, $\mathbb{R}^p$, i.e., a finite dimensional space, unlike the previous RKHS spaces.

Now it can be shown (not done here) that for any positive definite kernel function $K(x,y)$, there are functions $\phi_i(x) \in \mathcal{H}$ and $\gamma_i > 0$ such that

$$K(x,y) = \sum_{i=1}^{M} \gamma_i\phi_i(x)\phi_i(y);$$

here M may be either finite or infiinte. If $M = \infty$ then the sum is assumed to converge for *all* $x, y \in \mathbb{R}^p$ (though not necessarily absolutely). Similarly, for any function $f \in \mathcal{H}_K$, you can assume the sum $\sum_{i=1}^{\infty} c_i\phi_i(x) = f(x)$ converges for all x.

From the proof of Theorem 2 of Lecture 11, the functions $K(\,\cdot\,, y)$ span $\mathcal{H}_K$, i.e., every $f(\,\cdot\,) \in \mathcal{H}_K$ can be approximated by finite linear combinations of such functions arbitrarily well. From this show that the functions $\{\phi_i(\,\cdot\,)\}_{i=1}^M$ must span $\mathcal{H}_K$.

Note all inner products in this problem are in $\mathcal{H} \equiv \mathcal{H}_K$, i.e., $\langle a, b\rangle \equiv \langle a, b\rangle_{\mathcal{H}_K}$.

Additionally, show that the functions $\{\phi_i(\,\cdot\,)\}_{i=1}^M$ are orthogonal. Indeed, note

$$\phi_i(x) = \langle K(\,\cdot\,, x), \phi_i(\,\cdot\,)\rangle = \Big\langle \sum_{j=1}^{M} \gamma_j \phi_j(\,\cdot\,)\phi_j(x), \phi_i(\,\cdot\,)\Big\rangle$$

$$= \sum_{j=1}^{M} \gamma_j \Big\langle \phi_j(\,\cdot\,)\phi_j(x), \phi_i(\,\cdot\,)\Big\rangle = \sum_{j=1}^{M} \gamma_j \phi_j(x)\langle \phi_j, \phi_i\rangle, \qquad (1)$$

so that $\langle \phi_j, \phi_i\rangle = \delta_{ij}/\gamma_i$, where $\delta_{ij} = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}$

You can also assume that $|\phi_k(x)| \leq M$ for all k and $x \in R$, for some $M > 0$.

(a) Keep in mind what is the inner product on $\mathcal{H}_K$ (not stated in the text): if $f(\mathbf{x}), g(\mathbf{x}) \in \mathcal{H}_K$ with

$$f(\mathbf{x}) = \sum_{i=1}^{\infty} a_i \phi_i(\mathbf{x}), \quad g(\mathbf{x}) = \sum_{i=1}^{\infty} b_i \phi_i(\mathbf{x}), \quad \text{then} \quad \langle f, g\rangle_{\mathcal{H}} = \sum_{i=1}^{\infty} \frac{a_i b_i}{\gamma_i}.$$

You can write

$$K(\,\cdot\,, x_i) = \sum_{k=1}^{\infty} \{\gamma_k \phi_k(\mathbf{x}_i)\}\phi_k(\,\cdot\,); \quad f(\,\cdot\,) = \sum_{k=1}^{\infty} a_k \phi_k(\,\cdot\,);$$

how does it follow that **(a)** holds just using definitions?
(b) Use a similar representation to that in **(a)**
(d) When is $\rho(\cdot) \in \mathcal{H}$ orthogonal (in $\mathcal{H}$) to $K(\cdot, \mathbf{x}_i)$ for fixed $\mathbf{x}_i$ (i.e. they have dot product 0 in the active variable)? Show $\rho(\mathbf{x}_i) = 0$. How does this affect the Lagrangian sum $\sum_{i=1}^{N} L(\widehat{f}(\mathbf{x}_i), y_i)$? What happens to $||\widehat{f}||^2_{\mathcal{H}}$?

Hastie, problem 5.16

(a) We have a kernel function $K(x, y) = (x \cdot y + 1)^d = \Big(\sum_{i=1}^{p} x_i y_i + 1\Big)^d$, which we know is positive definite. We are also assuming we have a Hilbert space $\mathcal{H}$ of fuctions for which $K(x, y)$ is a reproducing kernel, and we are given that $\mathcal{H}$ consists of all polynomials of degree $\leq d$ in $\mathbb{R}^p$, i.e. in p variables $x_1, \ldots, x_p$. Note the form of $K(x, y)$ determines what the inner product of two functions in $\mathcal{H}_K \equiv \mathcal{H}$ is.

Note that this problem refers exclusively to the Hilbert space discussed in this part of the textbook, namely polynomials of p variables of total degree $\leq d$, whose dimension is $M = \binom{p+d}{d}$ (this is the 'choose' function defining the number of combinations of $p + d$ objects taken d at a time); you can derive or assume this.

We also are given that there are M functions $\{\phi_m(x)\}_{m=1}^M$ that form a basis for $\mathcal{H}$, such that

$$K(x,y) = \sum_{i=1}^{M} \gamma_i \phi_i(x)\phi_i(y); \tag{1a}$$

note that the sum here is finite since our Hilbert space is finite dimensional.

Thus we see that if

$$h_m(x) = \sqrt{\gamma_i}\phi_m(x), \tag{2}$$

then

$$K(x,y) = \sum_{m=1}^{M} h_m(x)h_m(y).$$

Notice that in the above problems the Hilbert space is all square integrable functions on a domain, and so is infinite dimensional. These examples involve only finite dimensional spaces of polynomials, though they have the same structure. In particular there is a unique inner product defined in our Hilbert space though it is finite dimensional.

Using the argument of equation (1), show that

$$\langle \phi_j, \phi_i \rangle_{\mathcal{H}} = \delta_{ij}/\gamma_{ii}, \tag{2a}$$

where $\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$.

So by (2a) show it follows that $h_m(x)$ are an orthonormal basis for $\mathcal{H} \equiv \mathcal{H}_K$ (since the ϕ_m are orthogonal already). Of course they are all polynomials since they are all in the space $\mathcal{H}$.

For this reason the vector $D_\gamma^{1/2}\phi(x) = \begin{bmatrix} \gamma_1^{1/2}\phi_1(x) \\ \gamma_2^{1/2}\phi_2(x) \\ \vdots \\ \gamma_m^{1/2}\phi_m(x) \end{bmatrix}$ has orthonormal functions as entries. Since $h(x) = \begin{bmatrix} h_1(x) \\ h_2(x) \\ \vdots \\ h_m(x) \end{bmatrix}$ is a vector of orthonormal functions, it must be true for any two sets of orthonormal vectors that there is an orthogonal matrix relating them, i.e., that $h_m(x) = \sum_{n=1}^{M} v_{mn}\left(\gamma_n^{1/2}\right)\phi_n(x)$. Letting $V = (v_{mn})$, show that relationship (5.62) holds.

Note: the matrix V is necessary because even though we have defined $h_m(x) = \gamma_m^{1/2}\phi_m(x)$, in fact we may permute (rearrange) the $h_m(x)$ and still obtain the same kernel function $K(x,y) = \sum_m h_m(x)h_m(y)$. The matrix V is needed only if there is such a permutation (i.e. relabeling) of the h_m. In that case V is a permutation matrix (rearranging components of a vector), which is a orthogonal matrix. However, each h_i basis element *must* be a multiple of a *single* ϕ_j (though we can permute the labels); arbitrary linear combinations of ϕ_i will not work to give h_i because of the uniqueness of the form of (1a).

(b) Once we have the basis functions $h_m(x)$ for the space $\mathcal{H}$ of polynomials above, we can transform the problem as in (5.63), into finding β_m's minimizing the Lagrangian (5.63), which you will want to show is related to the above problem.

Show that if we write $f(x) = \sum_{m=1}^{M} \beta_m h_m(x)$, then the Lagrangian in (5.63) is

$$\mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{N}\left(y_i - \sum_{m=1}^{M}\beta_m h_m(x_i)\right) + \lambda\sum_{m=1}^{M}\beta_m^2$$

$$= (\mathbf{y} - \mathbf{f})^2 + \lambda\|f\|_{\mathcal{H}}^2 \qquad (3)$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}; \quad \mathbf{f} = \begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_N) \end{bmatrix}.$$

using the orthonormality of the functions $h_m(x)$ to obtain the $\|f\|_{\mathcal{H}}^2$ term. Thus show the minimizing f will minimize (5.48), i.e. standard regularization problem fitting the dataset in $\mathcal{H}$, so the solution (as derived in class and in the book) must match that in (5.75), as in equation (5.55). Since the function $f(x)$ is the same for both representations, of course it follows that $\widehat{\mathbf{f}}$ is as well (why?).

(c) Note that the minimizer of (1a) is $\widehat{f} = \sum_j \beta_j h_j(x)$, and is also expressible as the second term of (5.76) as shown in the text and class, so they must be the same.

(d) Show there are no differences in the case $N > M$ - namely, the structure of $\widehat{f}(x)$, K and $(K + \lambda I)^{-1}$ work the same way. However, in this case from the viewpoint of (3), note that now the problem has a unique least squares solution even if $\lambda = 0$.