Practice questions and solutions: Advanced Topcics in Machine Learning

Arthur Gretton

December 1, 2021

1 Question 1

Consider an input variable x mapped to the RKHS $\mathcal{H}$, using the featue map $\phi(x)$ with kernel $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$. We introduce the notation

$$\Phi_n := \begin{bmatrix} \phi(x_1) & \dots & \phi(x_n) \end{bmatrix} \qquad \Phi_n^\top f = \begin{bmatrix} \langle \phi(x_1), f \rangle_{\mathcal{H}} \\ \vdots \\ \langle \phi(x_n), f \rangle_{\mathcal{H}} \end{bmatrix}$$

(i.e. Φ_n is a "matrix" where each column is an element in the feature space). We define the kernel matrix K with i, jth entry $k(x_i, x_j)$, and where using the above notation

$$K_n := \Phi_n^\top \Phi_n.$$

1. **[4 points]** Consider a set of m feature mapped points $\phi(z_1) \dots \phi(z_m)$, and define
$$\Phi_m := \begin{bmatrix} \phi(z_1) & \dots & \phi(z_m) \end{bmatrix}, \qquad K_m = \Phi_m^\top \Phi_m.$$
What is the expression for the projection P_m that takes an RKHS function $f \in \mathcal{H}$ and projects it onto a function $f_m := P_m f$ in the span of these points? Assume K_m is full rank and invertible. **Hints:** the projection of f on the span of Φ_m should minimize the squared RKHS norm from f to its projection. For symmetric $B \in \mathbb{R}^{m \times m}$ and $b \in \mathbb{R}^m$,
$$\frac{d}{da} a^\top B a = 2Ba \qquad \frac{d}{da} b^\top a = \frac{d}{da} a^\top b = b.$$

2. **[3 points]** Consider the Gram matrix $K_n := \Phi_n^\top \Phi_n$. What form does this take when we replace each entry $\phi(x_i)$ in Φ_n by its projection $P_m \phi(x_i)$? **Hint:** you'll use the matrix $K_{nm} = \Phi_n^\top \Phi_m$.

3. **[3 points]** Recall the definition of the tensor product, $(a \otimes b)c = \langle b, c \rangle_{\mathcal{H}} a$. Show that the operator
$$\Phi_n \Phi_n^\top = \sum_{i=1}^n \phi(x_i) \otimes \phi(x_i) \tag{1}$$

is a positive operator, meaning

$$\left\langle f, \left[\Phi_n \Phi_n^\top\right] f \right\rangle \geq 0 \quad \forall f \in \mathcal{F}.$$

Show that as a consequence, the eigenvalue decomposition

$$\Phi_n \Phi_n^\top = \sum_i s_i u_i(x) \otimes u_i(x)$$

cannot have negative s_i (note the symmetry of $\Phi_n \Phi_n^\top$).

4. **[5 points]** In ridge regression, we are given pairs $\{(x_i, y_i)\}_{i=1}^n$, and we minimise the regularised squared loss

$$f^* = \arg\min_{f \in \mathcal{H}} \mathcal{L}(f) := \arg\min_{f \in \mathcal{H}} \left[\sum_{i=1}^n \left(y_i - \langle f, \phi(x_i) \rangle_{\mathcal{H}}\right)^2 + \lambda \|f\|_{\mathcal{H}}^2 \right] \quad (2)$$

Show that a solution is

$$f^* := \left(\Phi_n \Phi_n^\top + \lambda I\right)^{-1} \Phi_n \mathbf{y},$$

where $\mathbf{y} = \begin{bmatrix} y_1 & \dots & y_n \end{bmatrix}^\top$, and recalling the definition in (1). For full marks, conmment on the existence of an inverse of $\Phi_n \Phi_n^\top + \lambda I$: you might argue from the eigenvalue decomposition

$$\Phi_n \Phi_n^\top + \lambda I = \sum_i q_i u_i(x) \otimes u_i(x),$$

noting that an inverse of the operator amounts to inverting all the eigenvalues. **Hint:** after expanding out the first term in the loss (2), define $g := \left(\Phi_n \Phi_n^\top + \lambda I\right)^{1/2} f$ where $\left(\Phi_n \Phi_n^\top + \lambda I\right)^{1/2} = \sum_i \sqrt{q_i} u_i(x) \otimes u_i(x)$, thus write $f := \left(\Phi_n \Phi_n^\top + \lambda I\right)^{-1/2} g$; then complete the square. You should not try to take derivatives in the infinite feature space $\mathcal{H}$. **Hint 2:** the next two parts of this question are easier, and don't require you to have solved this part, so if you get stuck, try the later parts.

5. **[2 points]** Show that the above solution can also be written

$$f^* = \Phi_n \left(K_n + \lambda I\right)^{-1} \mathbf{y}. \quad (3)$$

Hint: it will be easier *not* to use the solution of the previous part to prove this - i.e., what can we say about solutions to problems of the form (2)? Full marks will be awarded for a correct proof regardless of which approach was used.

6. **[3 points]** What happens if we substitute the approximation of K_n from the second part of this question into (3): assuming $m \ll n$, is the new solution any more computationally efficient? If not, can you propose a more computationally efficient solution? **Hint:**

$$(A + CBC^\top)^{-1} = A^{-1} - A^{-1}C(B^{-1} + C^\top A^{-1} C)^{-1} C^\top A^{-1}.$$

2 Question 2

We define the eigenexpansion of $k(x, x')$ with respect to a non-negative finite measure $p(x)$ on $\mathcal{X} := \mathbb{R}$,

$$\lambda_i e_i(x) = \int k(x, x') e_i(x') p(x') dx', \qquad \int_{L_2(p)} e_i(x) e_j(x) p(x) dx = \begin{cases} 1 & i = j \\ 0 & i \neq j. \end{cases} \tag{4}$$

We can write

$$k(x, x') = \sum_{\ell=1}^{\infty} \lambda_\ell e_\ell(x) e_\ell(x'), \tag{5}$$

We are given two functions f, g in $L_2(p)$, expanded in terms of the orthonormal system $\{e_\ell\}_{\ell=1}^{\infty}$,

$$f(x) = \sum_{\ell=1}^{\infty} \hat{f}_\ell e_\ell(x) \qquad g(x) = \sum_{\ell=1}^{\infty} \hat{g}_\ell e_\ell(x), \tag{6}$$

The standard dot product in $L_2(p)$ between f, g is

$$\begin{aligned} \langle f, g \rangle_{L_2(\mu)} &= \left\langle \sum_{q=1}^{\infty} \hat{f}_q e_q(x), \sum_{r=1}^{\infty} \hat{g}_r e_r(x) \right\rangle_{L_2(\mu)} \\ &= \sum_{\ell=1}^{\infty} \hat{f}_\ell \hat{g}_\ell. \end{aligned}$$

We can define the dot product in $\mathcal{H}$ to have a roughness penalty, yielding

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{\ell=1}^{\infty} \frac{\hat{f}_\ell \hat{g}_\ell}{\lambda_\ell} \qquad \|f\|_{\mathcal{H}}^2 = \sum_{\ell=1}^{\infty} \frac{\hat{f}_\ell^2}{\lambda_\ell}.$$

1. **[2 points]** We define an operator C_p which takes a function f and maps it to a new function $C_p f$, such that evaluating at x gives

   $$[C_p f](x) = \int k(x, x') f(x') p(x') dx'.$$

   Assume f is in $\mathcal{H}$. Show that

   $$\langle g, C_p f \rangle_{\mathcal{H}} = \int f(x) g(x) p(x) dx.$$

   In other words, C_p defines an uncentered variance operator.

2. **[3 points]** Is $C_p f$ smoother than f, or less smooth? Explain your answer in terms of the representation of f in (6). Assume that there are countably infinitely many $\lambda_\ell > 0$, and that $\sum_{\ell=1}^{\infty} \lambda_\ell^2 < \infty$.

3. **[3 points]** We are given i.i.d. samples $\{x_1, \ldots, x_m\}$ from p. We define the feature map $\phi(x) \in \mathcal{H}$ such that $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$. The empirical *centered* variance operator is

$$\widehat{C}_p = \frac{1}{m} \sum_{i=1}^{m} \phi(x_i) \otimes \phi(x_i) - \hat{\mu}_p \otimes \hat{\mu}_p \qquad \hat{\mu}_p := \frac{1}{m} \sum_{j=1}^{m} \phi(x_j),$$

where $(a \otimes b)c = \langle b, c \rangle_{\mathcal{H}} a$. Given a function $f = \sum_{i=1}^{m} \alpha_i \phi(x_i)$, derive the expression for the empirical variance of f,

$$\widehat{\mathrm{var}}(f) = \left\langle f, \widehat{C}_p f \right\rangle_{\mathcal{H}},$$

in terms of α and the Gram matrix K with entries $K_{ij} = k(x_i, x_j)$ (for ease of use later in the question, it will help to express this in matrix-vector form).

4. **[5 points]** We are given a second density q, and i.i.d samples $\{x_{m+1}, \ldots, x_{m+n}\}$ from q. The kernel Fisher discriminant amounts to solving

$$f^* = \arg\max_f \langle f, [(\hat{\mu}_p - \hat{\mu}_q) \otimes (\hat{\mu}_p - \hat{\mu}_q)] f \rangle_{\mathcal{H}} \quad \text{subject to}$$
$$1 \geq \left\langle f, \left(m\widehat{C}_p + n\widehat{C}_q \right) f \right\rangle_{\mathcal{H}} + \gamma \|f\|_{\mathcal{H}}^2$$

Write the resulting Lagrangian (simply express this in terms of f, $\widehat{C}_p$, etc: you should not introduce kernels at this stage). Show, using an argument that generalizes the proof of the standard representer theorem, that an optimal solution will take the form

$$f(x) = \sum_{i=1}^{m+n} \alpha_i k(x, x_i). \tag{7}$$

5. **[5 points]** Using the Lagrangian in the previous section, show that α is the solution of a generalized eigenvalue problem in terms of the Gram matrix

$$K := \begin{bmatrix} K_{pp} & K_{pq} \\ K_{qp} & K_{qq} \end{bmatrix}$$

across all samples, where K_{pp} is the Gram matrix between samples from p, and K_{pq} is the Gram matrix between samples from p and samples from q. **Hints:** start with the answer from the previous part for the quickest way to obtain the desired result. The primal variables will be α. For concise notation, it might be helpful to define a vector $\hat{\zeta}_p$ such that

$$\langle f, \hat{\mu}_p \rangle_{\mathcal{H}} = \alpha^\top \hat{\zeta}_p.$$

6. **[2 points]** What quantity does the Fisher discriminant approach when γ becomes very large?

3 Question 3

Assume $\mathcal{H}$ is a reproducing kernel Hilbert space with a Gaussian kernel,

$$k(x_i, x_j) = \exp\left(\frac{-\|x_i - x_j\|^2}{\sigma}\right) = \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{H}} \tag{8}$$

We have a sample $(x_i, y_i)_{i=1}^m$ drawn independently and identically from some distribution P_{XY}, where the $y_i \in \mathbb{R}$. Support vector regression finds a function:

$$f(x) = \langle w(\cdot), \phi(x) \rangle_{\mathcal{H}} + b$$

which solves the following problem:

$$\underset{w \in \mathcal{H}, \xi, \xi^* \in \mathbb{R}^m, b \in \mathbb{R}}{\text{minimize}} \frac{1}{2}\|w\|_{\mathcal{H}}^2 + \frac{C}{m}\sum_{i=1}^m (\xi_i + \xi_i^*), \tag{9}$$

$$\text{subject to} \quad (\langle w(\cdot), \phi(x_i) \rangle_{\mathcal{H}} + b) - y_i \leq \epsilon + \xi_i \tag{10}$$

$$y_i - (\langle w(\cdot), \phi(x_i) \rangle_{\mathcal{H}} + b) \leq \epsilon + \xi_i^* \tag{11}$$

$$\xi_i, \xi_i^* \geq 0$$

where $C, \epsilon \in \mathbb{R}^{++}$ are parameters of the algorithm (the notation means that both C and ϵ are strictly greater than zero).

1. **[2 points]** Define strong duality in the general setting of an optimization problem with equality and inequality constraints. Then describe the two conditions that hold for the support vector regression problem which ensure strong duality (hint: the second of these conditions is trivially satisfied here, since there are no equality constraints).

2. **[6 points]** Write the Lagrangian for the SV regression problem. State the KKT conditions as they apply to the problem. What is implied about the maximum of the dual problem when the KKT conditions hold?

3. **[6 points]** Write the Lagrange dual function for this optimization problem. In particular, you should obtain a form

   $$w = \sum_{i=1}^m (\alpha_i^* - \alpha_i)\phi(x_i)$$

   due to the two constraints (10) and (11).

4. **[6 points]** What do the KKT conditions imply about the allowable range of α_i? Where are points with $\alpha_i = 0$ situated relative to the regression function $f(x)$? Where are points for which α_i attains its upper bound? Finally, where are those points with α_i between the lower and upper bound (you do *not* need to obtain the analogous results for α_i^*)?

4 Question 1

1. The projection of f onto the basis set Φ_m minimises

$$a^* := \arg\min_{a\in\mathbb{R}^m} \|f - \Phi_m a\|_{\mathcal{H}}^2 = \langle f, f\rangle_{\mathcal{H}} - 2\langle f, \Phi_m a\rangle_{\mathcal{H}} + a^\top \Phi_m^\top \Phi_m a.$$

Differentiate wrt a and set to zero,

$$\begin{aligned} 0 &= -2\Phi_m^\top f + 2K_m a \\ a &= K_m^{-1}\Phi_m^\top f. \end{aligned}$$

Thus the projection of f is

$$P_m f := \Phi_m K_m^{-1}\Phi_m^\top f.$$

2. The Gram matrix of the projected features has i,jth entry

$$\begin{aligned} \tilde{k}_{ij} &= \langle P_m\phi(x_i), P_m\phi(x_j)\rangle_{\mathcal{H}} \\ &= \left\langle \Phi_m K_m^{-1} \begin{bmatrix} \vdots \\ k(z_q, x_i) \\ \vdots \end{bmatrix}, \Phi_m K_m^{-1} \begin{bmatrix} \vdots \\ k(z_r, x_j) \\ \vdots \end{bmatrix} \right\rangle_{\mathcal{H}} \\ &= \begin{bmatrix} \vdots \\ k(x_q, x_i) \\ \vdots \end{bmatrix}^\top K_m^{-1} K_m K_m^{-1} \begin{bmatrix} \vdots \\ k(x_r, x_j) \\ \vdots \end{bmatrix}, \end{aligned}$$

where $q, r \in \{1, \ldots, m\}$. Thus the solution is

$$\widetilde{K} := K_{nm} K_{mm}^{-1} K_{mn}.$$

3. We have

$$\begin{aligned} \left\langle f, \left[\Phi_n \Phi_n^\top\right] f\right\rangle_{\mathcal{H}} &= \left\langle f, \left[\sum_{i=1}^n \phi(x_i)\otimes\phi(x_i)\right] f\right\rangle_{\mathcal{H}} \\ &= \sum_{i=1}^n \langle f, \phi(x_i)\rangle_{\mathcal{H}}^2 \\ &\geq 0. \end{aligned}$$

Next imagine we had a negative eigenvalue s_i for eigenvector u_i. Choosing $f = u_i$, we have

$$\begin{aligned} \left\langle u_i, \left[\sum_j s_j u_j(x)\otimes u_j(x)\right] u_i\right\rangle_{\mathcal{H}} &= s_i \|u_i\|_{\mathcal{H}}^4 \\ &= s_i < 0 \end{aligned}$$

which contradicts the positivity described above.

4. Our goal is:

$$f^* \quad = \quad \arg\min_{f\in\mathcal{H}} \sum_{i=1}^{n} \left(y_i - \langle f, \phi(x_i)\rangle_{\mathcal{H}}\right)^2 + \lambda \|f\|_{\mathcal{H}}^2$$

Expanding out the above term, we get

$$\begin{aligned}
&\sum_{i=1}^{n} \left(y_i - \langle f, \phi(x_i)\rangle_{\mathcal{H}}\right)^2 + \lambda \|f\|_{\mathcal{H}}^2 \\
&= y^\top y - 2y^\top \mathbf{\Phi}_n^\top f + \sum_{i=1}^{n} \left(\langle f, \phi(x_i)\rangle_{\mathcal{H}}\right)^2 + \lambda \langle f, f\rangle_{\mathcal{H}} \\
&= y^\top y - 2y^\top \mathbf{\Phi}_n^\top f + \lambda \langle f, f\rangle_{\mathcal{H}} + \sum_{i=1}^{n} \langle f, [\phi(x_i) \otimes \phi(x_i)]\, f\rangle_{\mathcal{H}} \\
&= y^\top y - 2y^\top \mathbf{\Phi}_n^\top f + \left\langle f, \left(\mathbf{\Phi}_n \mathbf{\Phi}_n^\top + \lambda I\right) f\right\rangle_{\mathcal{H}} = (*)
\end{aligned}$$

Define $g = \left(\mathbf{\Phi}_n \mathbf{\Phi}_n^\top + \lambda I\right)^{1/2} f$, where the square root is well defined since the operator is positive definite. Even though $\mathbf{\Phi}_n \mathbf{\Phi}_n^\top$ is not invertible for infinite dimensional feature spaces, adding λI ensures we can substitute $f = \left(\mathbf{\Phi}_n \mathbf{\Phi}_n^\top + \lambda I\right)^{-1/2} g$ - the fact that $\mathbf{\Phi}_n \mathbf{\Phi}_n^\top$ is positive means its singular values are positive or zero, and thus $s_i + \lambda > 0$. Then

$$\begin{aligned}
(*) \quad &= \quad y^\top y - 2y^\top \mathbf{\Phi}_n^\top \left(\mathbf{\Phi}_n \mathbf{\Phi}_n^\top + \lambda I\right)^{-1/2} g + \langle g, g\rangle_{\mathcal{H}} \\
&= \quad y^\top y + \left\| \left(\mathbf{\Phi}_n \mathbf{\Phi}_n^\top + \lambda I\right)^{-1/2} \mathbf{\Phi}_n y - g \right\|_{\mathcal{H}}^2 - \left\| y^\top \mathbf{\Phi}_n^\top \left(\mathbf{\Phi}_n \mathbf{\Phi}_n^\top + \lambda I\right)^{-1/2} \right\|_{\mathcal{H}}^2,
\end{aligned}$$

where we complete the square. This is minimized when

$$\begin{aligned}
g^* \quad &= \quad \left(\mathbf{\Phi}_n \mathbf{\Phi}_n^\top + \lambda I\right)^{-1/2} \mathbf{\Phi}_n y \quad \text{or} \\
f^* \quad &= \quad \left(\mathbf{\Phi}_n \mathbf{\Phi}_n^\top + \lambda I\right)^{-1} \mathbf{\Phi}_n y.
\end{aligned}$$

5. The representer theorem tells us

$$f^* = \mathbf{\Phi}_n \alpha^*.$$

Thus we want to solve

$$\begin{aligned}
\sum_{i=1}^{n} \left(y_i - \langle f, \phi(x_i)\rangle_{\mathcal{H}}\right)^2 + \lambda \|f\|_{\mathcal{H}}^2 \quad &= \quad \|y - K_n \alpha\|^2 + \lambda \alpha^\top K_n \alpha \\
&= \quad y^\top y - 2y^\top K_n \alpha + \alpha^\top \left(K_n^2 + \lambda K_n\right) \alpha
\end{aligned}$$

Differentiating wrt α and setting this to zero, we get

$$\alpha^* = (K_n + \lambda I_n)^{-1} y. \tag{12}$$

6. If we substitute $K_n \to K_{nm}K_{mm}^{-1}K_{mn}$ in (12), we get

$$\alpha^* = (K_{nm}K_{mm}^{-1}K_{mn} + \lambda I_n)^{-1}y.$$

However the cost is still the same: the matrix to be inverted is still $n \times n$. Using the hint,

$$a^* = \left[\lambda^{-1}I - \lambda^{-1}K_{nm}\left(K_{nm}^\top K_{nm} + \lambda K_{mm}\right)^{-1}K_{nm}^\top\right]y.$$

In this case, we only need to invert an $m \times m$ matrix. and the overall cost is $O(m^2 n)$, i.e. linear in n.

4.1 Question 2

1. We start with

$$\begin{aligned} C_p f &= \int k(x,x')f(x')p(x')dx' \\ &= \int \left[\sum_{\ell=1}^{\infty} \lambda_\ell e_\ell(x)e_\ell(x')\right]\left[\sum_{j=1}^{\infty} \hat{f}_j e_j(x)\right] p(x')dx' \\ &= \sum_{\ell=1}^{\infty} \lambda_\ell \hat{f}_\ell e_\ell(x) \end{aligned}$$

Thus we have

$$\begin{aligned} \langle g, C_p f\rangle_{\mathcal{H}} &= \sum_{\ell=1}^{\infty} \frac{\hat{g}_\ell\left(\lambda_\ell \hat{f}_\ell\right)}{\lambda_\ell} \\ &= \sum_{\ell=1}^{\infty} \hat{f}_\ell \hat{g}_\ell \\ &= \int f(x)g(x)p(x)dx. \end{aligned}$$

2. Under the stated assumptions, λ_ℓ must decay to zero as $\ell \to \infty$. Therefore $C_p f$ is smoother than f, since it has the effect of transforming $\hat{f}_\ell \to \hat{f}_\ell \lambda_\ell$, hence the high frequency components of f will be shrunk. Aside:

$$\|C_p f\|_{\mathcal{H}}^2 = \sum_\ell \frac{\lambda_\ell^2 \hat{f}_\ell^2}{\lambda_\ell} = \sum_\ell \lambda_\ell \hat{f}_\ell^2$$

vs

$$\|f\|_{\mathcal{H}}^2 = \sum_\ell \frac{\hat{f}_\ell^2}{\lambda_\ell}.$$

Thus for any given ℓ for which $\lambda_\ell < 1$,

$$\frac{\hat{f}_\ell^2}{\lambda_\ell} \geq \lambda_\ell \hat{f}_\ell^2$$

which again yields an insight that the RKHS norm of f will be greater than that of $C_p f$ (meaning f is less smooth).

3. Given

$$\widehat{C}_p = \frac{1}{m}\sum_{i=1}^m \phi(x_i)\otimes\phi(x_i) - \hat{\mu}_p \otimes \hat{\mu}_p \qquad \hat{\mu}_p := \frac{1}{m}\sum_{j=1}^m \phi(x_j),$$

and $f = \sum_{i=1}^m \alpha_i \phi(x_i)$, then

$$\begin{aligned}
\widehat{\mathrm{var}}(f) &= \left\langle f, \widehat{C}_p f \right\rangle_{\mathcal{H}} \\
&= \left\langle f, \left(\frac{1}{m}\sum_{i=1}^m \phi(x_i)\otimes\phi(x_i)\right) f \right\rangle_{\mathcal{H}} - \left(\langle f, \hat{\mu}_p\rangle_{\mathcal{H}}\right)^2 \\
&= \frac{1}{m}\sum_{i=1}^m \left(\sum_{j=1}^m \alpha_j k(x_i x_j)\right)\left(\sum_{q=1}^m \alpha_q k(x_i x_q)\right) - \left(\frac{1}{m}\sum_{i=1}^m\sum_{j=1}^m k(x_i x_j)\alpha_j\right)^2 \\
&= \frac{1}{m}\alpha K^2 \alpha - \frac{1}{m^2}\left(\alpha^\top K 1_m\right)^2 \\
&= \frac{1}{m}\alpha^\top K H K \alpha,
\end{aligned}$$

where here 1_m is the $m \times 1$ vector of ones and $H = I - m^{-1}1_m 1_m^\top$.

4. This result, and that of the next section, are taken from [**?**]. The constrained problem is

$$\begin{aligned}
f^* &= \arg\max_f \langle f, [(\hat{\mu}_p - \hat{\mu}_q)\otimes(\hat{\mu}_p - \hat{\mu}_q)]\, f\rangle_{\mathcal{H}} \\
1 &\geq \left\langle f, \left(m\widehat{C}_p + n\widehat{C}_q\right) f\right\rangle_{\mathcal{H}} + \gamma \left\|f\right\|_{\mathcal{H}}^2 .
\end{aligned}$$

The Lagrangian is

$$f^* = \arg\min_f - \langle f, [(\hat{\mu}_p - \hat{\mu}_q)\otimes(\hat{\mu}_p - \hat{\mu}_q)]\, f\rangle_{\mathcal{H}} + \eta\left(\left\langle f, \left(m\widehat{C}_p + n\widehat{C}_q\right) f\right\rangle_{\mathcal{H}} + \gamma\left\|f\right\|_{\mathcal{H}}^2 - 1\right). \tag{13}$$

where $\eta > 0$ is the Lagrange multiplier. We now show that a solution f^* must take the form

$$f(x) = \sum_{i=1}^{m+n} \alpha_i k(x, x_i). \tag{14}$$

Assume a solution of the form $f = f_\parallel + f_\perp$, where $f_\parallel$ is of the form (14), and $f_\perp$ is perpendicular to the span of the observed sample. Consider each term in turn. The first term is

$$\langle f, \hat{\mu}_p - \hat{\mu}_q \rangle_{\mathcal{H}}^2 .$$

Now $\langle f, \hat{\mu}_p \rangle_{\mathcal{H}} = \langle f_\parallel, \hat{\mu}_p \rangle_{\mathcal{H}}$, and likewise for $\langle f, \hat{\mu}_q \rangle$. Next, consider

$$\left\langle f, \widehat{C}_p f \right\rangle_{\mathcal{H}} = \frac{1}{m} \left\langle f, \left[\sum_{i=1}^m \phi(x_i) \otimes \phi(x_i) \right] f \right\rangle_{\mathcal{H}} - \left(\langle f, \hat{\mu}_p \rangle_{\mathcal{H}} \right)^2 .$$

Note that

$$\left\langle f, \left[\sum_{i=1}^m \phi(x_i) \otimes \phi(x_i) \right] f \right\rangle_{\mathcal{H}} = \sum_{i=1}^m \langle f, \phi(x_i) \rangle_{\mathcal{H}}^2 = \sum_{i=1}^m \langle f_\parallel, \phi(x_i) \rangle_{\mathcal{H}}^2 .$$

Thus, the first two terms in (13) involving f are independent of $f_\perp$. The final term decomposes as $\|f\|^2 = \|f_\parallel\|^2 + \|f_\perp\|^2$, and is minimized when $\|f_\perp\|^2 = 0$.

5. We first write the first term of the Lagrangian in kernel form,

$$\langle f, [(\hat{\mu}_p - \hat{\mu}_q) \otimes (\hat{\mu}_p - \hat{\mu}_q)] f \rangle_{\mathcal{H}} \tag{15}$$
$$= \langle f, \hat{\mu}_p - \hat{\mu}_q \rangle_H^2 \tag{16}$$
$$= \alpha^\top \left(\hat{\zeta}_p - \hat{\zeta}_q \right) \left(\hat{\zeta}_p - \hat{\zeta}_q \right)^\top \alpha, \tag{17}$$

where

$$\hat{\zeta}_p := \frac{1}{m} \begin{bmatrix} K_{pp} \\ K_{qp} \end{bmatrix} 1_m, \qquad \hat{\zeta}_q = \frac{1}{n} \begin{bmatrix} K_{pq} \\ K_{qq} \end{bmatrix} 1_n.$$

The second term is

$$\left\langle f, \left(m\widehat{C}_p + n\widehat{C}_q \right) f \right\rangle_{\mathcal{H}}$$
$$= \left\langle f, \left(\sum_{i=1}^m \phi(x_i) \otimes \phi(x_i) + \sum_{i=m+1}^n \phi(x_i) \otimes \phi(x_i) \right) f \right\rangle_{\mathcal{H}} - \alpha^\top \left(m\hat{\zeta}_p \hat{\zeta}_p^\top + n\hat{\zeta}_q \hat{\zeta}_q^\top \right) \alpha.$$

To express the first part using kernels, we write

$$\left\langle f, \left(\sum_{i=1}^m \phi(x_i) \otimes \phi(x_i) + \sum_{i=m+1}^{m+n} \phi(x_i) \otimes \phi(x_i) \right) f \right\rangle_{\mathcal{H}}$$
$$EITHER = \left\langle f, \left(\sum_{i=1}^{m+n} \phi(x_i) \otimes \phi(x_i) \right) f \right\rangle_{\mathcal{H}}$$
$$OR = \alpha^\top \begin{bmatrix} K_{pp} \\ K_{qp} \end{bmatrix} \begin{bmatrix} K_{pp} & K_{qp}^\top \end{bmatrix} \alpha + \alpha^\top \begin{bmatrix} K_{pq} \\ K_{qq} \end{bmatrix} \begin{bmatrix} K_{pq}^\top & K_{qq} \end{bmatrix} \alpha$$
$$= \alpha^\top K^2 \alpha.$$

(last line from part 3). The third term is

$$\|f\|_{\mathcal{H}}^2 = \alpha^\top K \alpha.$$

The Lagrangian is

$$\begin{aligned}\arg\max_{\alpha} \quad & \alpha^\top \left(\hat{\zeta}_p - \hat{\zeta}_q\right)\left(\hat{\zeta}_p - \hat{\zeta}_q\right)^\top \alpha \\ & + \eta\left[1 - \alpha^\top\left(K^2 - \left(m\hat{\zeta}_p\hat{\zeta}_p^\top + n\hat{\zeta}_q\hat{\zeta}_q^\top\right) + \gamma K\right)\alpha\right]\end{aligned}$$

where $\eta \geq 0$. Differentiating wrt α and setting the resulting expression to zero yields

$$K\begin{bmatrix} m^{-1}1_m \\ -n^{-1}1_n \end{bmatrix}\begin{bmatrix} m^{-1}1_m^\top & -n^{-1}1_n^\top \end{bmatrix} K\alpha = \eta\left(K^2 - \left(m\hat{\zeta}_p\hat{\zeta}_p^\top + n\hat{\zeta}_q\hat{\zeta}_q^\top\right) + \gamma K\right)\alpha,$$

which is a generalized eigenvalue problem.

6. When γ becomes very large, the variance constraint is reduced in relative importance, and the discriminant approaches the squared MMD (up to scaling).

5 Question 3

1. Strong duality: at global optimum, minimum of primal and maximum of dual are equal. Conditions for strong duality:

   (a) Primal problem is convex, i.e. of the form (note affine equality constraint, although this can be ignored here since no equality constraint in problem)

   $$\begin{aligned} \text{minimize} \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) \leq 0 \qquad i = 1, \ldots, n \\ & Ax = b \end{aligned} \tag{18}$$

   for convex $f_0, \ldots, f_m$, *and*

   (b) Slater's condition holds: there exists some *strictly* feasible point[1] $\tilde{x} \in \text{relint}(\mathcal{D})$ such that

   $$f_i(\tilde{x}) < 0 \quad i = 1, \ldots, m \quad A\tilde{x} = b.$$

   Since inequality constraints are affine for this problem, however, then these become trivial, and reduce to the inequality constrints

   $$f_i(x) \leq 0 \quad i = 1, \ldots, n$$

   (and there are no equality constraimts).

[1] We denote by relint($\mathcal{D}$) the relative interior of the set $\mathcal{D}$. This looks like the interior of the set, but is non-empty even when the set is a subspace of a larger space.

2. Recall the optimization problem:

$$\begin{aligned}\underset{w\in\mathcal{H},\xi,\xi^*\in\mathbb{R}^m,b\in\mathbb{R}}{\text{minimize}} \; & \frac{1}{2}\|w\|_{\mathcal{H}}^2 + \frac{C}{m}\sum_{i=1}^{m}(\xi_i+\xi_i^*), \\ \text{subject to } & (\langle w(\cdot),\phi(x_i)\rangle_{\mathcal{H}}+b) - y_i \le \epsilon+\xi_i \\ & y_i - (\langle w(\cdot),\phi(x_i)\rangle_{\mathcal{H}}+b) \le \epsilon+\xi_i^* \\ & \xi_i,\xi_i^* \ge 0\end{aligned} \tag{19}$$

The Lagrangian is:

$$\begin{aligned}L :=& \frac{1}{2}\|w\|_{\mathcal{H}}^2 + \frac{C}{m}\sum_{i=1}^{m}(\xi_i+\xi_i^*) - \sum_{i=1}^{m}(\eta_i\xi_i+\eta_i^*\xi_i^*) \\ & + \sum_{i=1}^{m}\alpha_i(-\epsilon-\xi_i-y_i+\langle w(\cdot),\phi(x_i)\rangle_{\mathcal{H}}+b) \\ & + \sum_{i=1}^{m}\alpha_i^*(-\epsilon-\xi_i^*+y_i-\langle w(\cdot),\phi(x_i)\rangle_{\mathcal{H}}-b).\end{aligned}$$

The KKT conditions: when **strong duality holds** and using notation from (18) (again ignoring absent equality constraint), the KKT conditions are

$$\begin{aligned} f_i(x) &\le 0,\ i=1,\dots,m \\ \lambda_i &\ge 0,\ i=1,\dots,m \\ \lambda_i f_i(x) &= 0,\ i=1,\dots,m \\ \nabla f_0(x)+\sum_{i=1}^{m}\lambda_i\nabla f_i(x) &= 0\end{aligned} \tag{20}$$

These are **necessary and sufficient for optimality** under strong duality. The condition $\lambda_i f_i = 0$ translates to

$$0 = \eta_i\xi_i \tag{21}$$
$$0 = \eta_i^*\xi_i^*$$
$$0 = \alpha_i(-\epsilon-\xi_i-y_i+\langle w(\cdot),\phi(x_i)\rangle_{\mathcal{H}}+b) \tag{22}$$
$$0 = \alpha_i^*(-\epsilon-\xi_i^*+y_i-\langle w(\cdot),\phi(x_i)\rangle_{\mathcal{H}}-b)$$

The dual variables satisfy

$$\alpha_i,\alpha_i^*,\eta_i,\eta_i^* \ge 0. \tag{23}$$

Taking derivatives wrt the primal parameters and setting to zero gives the

remaining conditions:

$$\frac{\partial L}{\partial w} = w(\cdot) + \sum_{i=1}^{m} \alpha_i \phi(x_i) - \sum_{i=1}^{m} \alpha_i^* \phi(x_i) = 0 \tag{24}$$

$$\frac{\partial L}{\partial \xi_i} = \frac{C}{m} - \alpha_i - \eta_i = 0 \tag{25}$$

$$\frac{\partial L}{\partial \xi_i^*} = \frac{C}{m} - \alpha_i^* - \eta_i^* = 0 \tag{26}$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{m} (\alpha_i - \alpha_i^*) = 0 \tag{27}$$

3. We use the minimum Lagrangian wrt the primal parameters, which we can readily compute since we have the point at which the primal derivatives are zero. From (24),

$$w^*(\cdot) = \sum_{i=1}^{m} (\alpha_i^* - \alpha_i) \phi(x_i).$$

Substituting the optimal w^* and the expressions for η_i and η_i^* back into the Lagrangian, we get

$$\begin{aligned}
L :=& \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) k(x_i x_j) + \frac{C}{m} \sum_{i=1}^{m} (\xi_i + \xi_i^*) \\
& - \sum_{i=1}^{m} \left[\xi_i \left(\frac{C}{m} - \alpha_i \right) + \xi_i^* \left(\frac{C}{m} - \alpha_i^* \right) \right] \\
& + \sum_{i=1}^{m} \alpha_i \left[-\epsilon - \xi_i - y_i + \sum_{j=1}^{m} (\alpha_j^* - \alpha_j) k(x_i, x_j) + b \right] \\
& + \sum_{i=1}^{m} \alpha_i^* \left[-\epsilon - \xi_i^* + y_i - \sum_{j=1}^{m} (\alpha_j^* - \alpha_j) k(x_i, x_j) - b \right], \\
=& - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) k(x_i, x_j) \\
& - \epsilon \sum_{i=1}^{m} (\alpha_i + \alpha_i^*) + \sum_{i=1}^{m} y_i (\alpha_i^* - \alpha_i)
\end{aligned}$$

This is the Lagrange dual function. To get the desired solution, it must be maximized wrt α_i, α_i^*.

4. Solution remark: we can simplify the KKT conditions as follows:

$$0 = \eta_i \xi_i = \left(\frac{C}{m} - \alpha_i \right) \xi_i$$
$$0 = \eta_i^* \xi_i^* = \left(\frac{C}{m} - \alpha_i^* \right) \xi_i^*,$$

which removes the requirement to discuss η_i, η_i^* explicitly. Original solution: there are three cases:

(a) When $\eta_i = 0$ then we have $\xi_i \geq 0$ from (21). From (25), $\frac{C}{m} = \alpha_i$ for these points. From (22), since $\alpha_i \neq 0$, we must have

$$\begin{aligned} 0 &= -\epsilon - \xi_i - y_i + \langle w(\cdot), \phi(x_i) \rangle_{\mathcal{H}} + b \\ \epsilon + \xi_i &= -y_i + \langle w(\cdot), \phi(x_i) \rangle_{\mathcal{H}} + b, \end{aligned}$$

and y_i is below the regression function a distance $\epsilon + \xi_i$.

(b) When $\alpha_i = 0$ then $\eta_i = C/m$ by (25), hence we must have $\xi_i = 0$ from (21), and since $\alpha_i = 0$, we have

$$\begin{aligned} -\epsilon - y_i + \langle w(\cdot), \phi(x_i) \rangle_{\mathcal{H}} + b &\leq 0 \\ \langle w(\cdot), \phi(x_i) \rangle_{\mathcal{H}} + b - y_i &\leq \epsilon, \end{aligned}$$

i.e. y_i is not below the regression function by more than ϵ (it might still make a large positive error, however).

(c) When $\alpha_i \neq 0$ and $\eta_i \neq 0$, we have $\xi_i = 0$ from (21) *and,* from (22), we have

$$\begin{aligned} 0 &= -\epsilon - y_i + \langle w(\cdot), \phi(x_i) \rangle_{\mathcal{H}} + b \\ \epsilon &= -y_i + \langle w(\cdot), \phi(x_i) \rangle_{\mathcal{H}} + b, \end{aligned}$$

Thus y_i is below the regression function by a distance exactly ϵ. From (25) and (23), $\alpha_i \in (0, C/m)$; it can't attain the upper bound since $\eta_i \neq 0$.