Elements of Statistical Learning
Solution Sheet 3 (Rev 3, 29th April 2021)

1. *Show that S_λ (defined by equation (23), slide 27 in Lecture 11) is symmetric and positive semi-definite.* On page 27 of the notes in Lecture 11, we have

$$S_\lambda = N(N^T N + \lambda\Omega_n)^{-1} N^T. \tag{1}$$

Let $A = N^T N + \lambda\Omega_n$. From the form of Ω_n in equation (19) on page 26 we can see that it is symmetric. Hence, A is symmetric and (we assume) invertible, hence A^{-1} is symmetric (you can find lots of proofs of these, or use the fact that it can be diagonalized). Hence:

$$S_\lambda^T = \{NA^{-1}N^T\}^T = N\{A^{-1}\}^T N^T = NA^{-1}N^T = S_\lambda, \tag{2}$$

and hence S_λ is symmetric.

To show that S_λ is positive semi-definite we only need to show Ω_n is positive semi-definite (because then we're multiplying by positive λ, then adding $N^T N$, which is positive semi-definite and then, A, that we defined above, is also positive semi-definite and so is A^{-1} (as eigenvalues of A are all positive [due to invertibility] and thus so will those of A^{-1} being the reciprocal). It is then easy to show $NA^{-1}N^T$ is positive semi-definite.

So, to show Ω_n is positive semi-definite, we can use a similar argument to slide 25 (but that involved specific θs) and show for arbitrary vector a

$$\begin{aligned} a^T\Omega_n a &= \sum_{i=1}^n a_i \sum_{k=1}^n \Omega_{i,k} a_k & (3)\\ &= \int \sum_{i=1}^n a_i N_i''(t) \sum_{k=1}^n a_k N_k''(t)\, dt & (4)\\ &= \int r^2(t)\, dt \geq 0, & (5) \end{aligned}$$

where $r(t) = \sum_{i=1}^n a_i N_i''(t)$.

2. *Show how the criterion*

$$\mathrm{RSS}(\theta, \lambda) = (y - N\theta)^T (y - N\theta) + \lambda\theta^T \Omega_n \theta, \tag{6}$$

can give solution

$$\hat{\theta} = (N^T N + \lambda\Omega_n)^{-1} N^T y. \tag{7}$$

by using the ridge regression machinery — hint: using a reparametrisation of ridge. This was from slide 25 of Lecture 11. Assume all of the inverses you need.

Consider the parametrisation $\beta = \Omega_n^{1/2}\theta$. Then

$$\begin{aligned} \mathrm{RSS}(\theta, \lambda) &= (y - N\theta)^T (y - N\theta) + \lambda\theta^T \Omega_n \theta & (8)\\ &= (y - N\Omega^{-1/2}\beta)^T (y - N\Omega^{-1/2}\beta) + \lambda\beta^T\beta. & (9) \end{aligned}$$

Now let $X = N\Omega^{-1/2}$, then $X^TX = \Omega^{-1/2}N^TN\Omega^{-1/2}$. The ridge regression estimator is, and can be rewritten as

$$\begin{aligned}
\hat{\beta}^R &= (X^TX + \lambda I)^{-1}X^TY && (10)\\
&= (\Omega^{-1/2}N^TN\Omega^{-1/2} + \lambda I)^{-1}\Omega^{-1/2}N^TY && (11)\\
&= \{\Omega^{-1/2}(N^TN + \lambda\Omega)\Omega^{-1/2}\}^{-1}\Omega^{-1/2}N^TY && (12)\\
&= \Omega^{1/2}(N^TN + \lambda\Omega)^{-1}\Omega^{1/2}\Omega^{-1/2}N^TY && (13)\\
&= \Omega^{1/2}(N^TN + \lambda\Omega)^{-1}N^TY && (14)\\
&= \Omega^{1/2}\hat{\theta}, && (15)
\end{aligned}$$

so $\hat{\theta} = (N^TN + \lambda\Omega)^{-1}N^TY$.

3. *In kernel density estimation, the kernel $K(x)$ is chosen to satisfy the following properties (i) $K(x) \geq 0$; (ii) $\int K(x)\,dx = 1$ and $\int xK(x)\,dx = 0$. Explain why it would not be desirable to choose a kernel that did not satisfy all of those properties.* If the kernel could be negative, this means that it is possible that the kernel density estimate $\hat{f}(x)$ could also be negative, which is not usually what one wants when estimating a density $f(x)$ which is always non-negative (however, it should be pointed out that some orthogonal series density estimators can be negative, but usually slightly). Property (ii) forces the kernel density estimator to integrate to one, again mimicking the true density — if we don't have this property, the estimator might not integrate to 1. For (iii) we usually assume the 'uncertainty' around X_i is symmetric, and hence we need a kernel that puts equal mass to the left and right of each X_i. However, non-symmetric kernels do exist.

4. *Show that the kernel density estimate $\hat{f}_{n,h,K}(x)$ defined on slide 6 of Lecture 12 is a density (i.e. integrates to one).* We will do this by directly integrating $\hat{f}$:

$$\int \hat{f}_{n,h,K}(x)\,dx = (nh)^{-1}\sum_{i=1}^{n}\int_{-\infty}^{\infty} K\{(X_i - x)/h\}\,dx \tag{16}$$

Let $y = (X_i - x)/h$, then $dy = -h^{-1}\,dx$ and we have to change the order of the integration limits, so

$$\int \hat{f}_{n,h,K}(x)\,dx = (nh)^{-1}h\sum_{i=1}^{n}\int_{-\infty}^{\infty} K(y)\,dy = n^{-1}n = 1, \tag{17}$$

as $\int K(y)\,dy = 1$, by definition of the kernel.

5. *Let $\hat{f}(x)$ be a kernel density estimator for density $f(x)$ with kernel K. Use methods similar to those on slides 24, 25 and 26 from Lecture 12 to show that*

$$\text{bias}\{\hat{f}'(x)\} = h^2C_3f'''(x) + \mathcal{O}(h^3). \tag{18}$$

[Assume f is a density that is three times differentiable on $\mathbb{R}$ and the kernel K satisfies $\lim_{x\to\pm\infty} f(x)K(x) = 0$ and that K has at least one derivative.]

To find the bias of $f'(x)$ we first consider its expectation.

$$\begin{aligned}
\mathbb{E}\{\hat{f}'(x)\} &= \mathbb{E}\left[(nh)^{-1}\sum_{i=1}^{n}\frac{\partial}{\partial x}K\{(X_i-x)/h\}\right] && (19)\\
&= \mathbb{E}\left[(nh)^{-1}\sum_{i=1}^{n}K'\{(X_i-x)/h\}\times -h^{-1}\right] && (20)\\
&= -(nh^2)^{-1}\sum_{i=1}^{n}\mathbb{E}\left[K'\{(X_i-x)/h\}\right] && (21)
\end{aligned}$$

Let us now work out the expectation, noting that $X_i \sim f$:

$$\begin{aligned}
\mathbb{E}\left[K'\{(X_i-x)/h\}\right] &= \int_{-\infty}^{\infty}K'\{(y-x)/h\}f(y)\,dy && (22)\\
&= h\left[f(y)K\{(y-x)/h\}\right]_{-\infty}^{\infty} && (23)\\
&\quad -h\int_{-\infty}^{\infty}f'(y)K\{(y-x)/h\}\,dy && (24)
\end{aligned}$$

The first term is zero because $\lim_{x\to\pm\infty} f(x)K(x) = 0$ from the assumption in the question. A Taylor series expansion of $f'(x+\delta)$ for δ small is

$$f'(x+\delta) = f'(x) + \delta f''(x) + \delta^2 f'''(x)/2 + \mathcal{O}(\delta^3). \quad (25)$$

Now let us substitute $v = (y-x)/h$ in (24), so $dv = h^{-1}dy$. So

$$\begin{aligned}
(24) &= -h^2\int_{-\infty}^{\infty}f'(x+hv)K(v)\,dv && (26)\\
&= -h^2\left[f'(x)\cancelto{1}{\int K(v)\,dv} + f''(x)h\cancelto{0}{\int vK(v)\,dv} + \frac{1}{2}f'''(x)h^2\cancelto{C_3^*}{\int v^2K(v)\,dv}\right] + \mathcal{O}(h^5)\\
&= -h^2f'(x) - h^4C_3f'''(x) - \mathcal{O}(h^5), && (27)
\end{aligned}$$

and we absorb the constant $1/2$ into $C_3 = C_3^*/2$. Now substitute (27) into (21)

$$\begin{aligned}
\mathbb{E}\{\hat{f}'(x)\} &= (nh^2)^{-1}n\left\{h^2f'(x) + h^4C_3f'''(x) + \mathcal{O}(h^5)\right\} && (28)\\
&= f'(x) + h^2C_3f'''(x) + \mathcal{O}(h^3). && (29)
\end{aligned}$$

6. *Define the (basis) functions $\Psi_k(x) = \exp(2\pi ikx)$, for $x \in [0,1]$ for $k \in \mathbb{Z}$. Show that the set $\{\Psi_k(x)\}_{k\in\mathbb{Z}}$ is orthonormal.*

Let's look at their inner product for k, j:

$$\begin{aligned}
<\Psi_k, \Psi_j> &= \int_0^1 \Psi_k(x)\overline{\Psi_j(x)}\,dx && (30)\\
&= \int_0^1 \exp\{2\pi i(k-j)x\}\,dx && (31)\\
&= \int_0^1 \exp(2\pi i\ell x)\,dx, && (32)
\end{aligned}$$

where $\ell = k - j$. Here is a not very rigorous argument. For $\ell \neq 0$:

$$< \Psi_k, \Psi_j >= \left[(2\pi i\ell)^{-1} \exp(2\pi i\ell x)\right]_0^1. \tag{33}$$

And, in substituting the $0, 1$ into the expression gives $\exp(0) = 1$ for the 0, and

$$\exp(2\pi i\ell) = \cos(2\pi\ell) + i\sin(2\pi\ell) = 1. \tag{34}$$

Hence, the inner product is zero for $\ell \neq 0$. For $\ell = 0$ we have $< \Psi_k, \Psi_j >= \int_0^1 dx = 1$. Hence the $\{\Psi_j(x)\}$ are orthonormal.

7. *Let $\phi(x)$ be the usual probability density function of the standard normal distribution. Define the function $\psi(x) = \phi''(x)$, the second derivative of the density. Show that $\psi(x)$ satisfies the key wavelet property of $\int \psi(x)\, dx = 0$.* We know that

$$\phi(x) = (2\pi)^{-1/2} \exp(-x^2/2), \tag{35}$$

and hence

$$\phi'(x) = (2\pi)^{-1/2} \times (-x) \times \exp(-x^2/2) = -x\phi(x), \tag{36}$$

Then we can write

$$\int_{-\infty}^{\infty} \psi(x)\, dx = \int_{-\infty}^{\infty} \phi''(x)\, dx = \left[\phi'(x)\right]_{-\infty}^{\infty} = \left[-x\phi(x)\right]_{-\infty}^{\infty} = 0, \tag{37}$$

as required.

8. *Let $f(x), g(x)$ be two functions with orthogonal series expansions of $f(x) = \sum_\nu f_\nu \xi_\nu(x)$ and $g(x) = \sum_\nu g_\nu \xi_\nu(x)$, where $\{\xi_\nu\}$ is some orthogonal basis for the space of functions we're considering. Show Parseval's relation ... where $F = \{f_\nu\}_\nu$ and similarly for G and $< f, g >= \int f(x)\overline{g(x)}\, dx$ and $< F, G >= \sum_\nu f_\nu \overline{g_\nu}$.*

Then

$$\begin{aligned}
< f, g > &= \int f(x)\overline{g(x)}\, dx && (38)\\
&= \int \sum_\nu f_\nu \xi_\nu(x) \sum_\mu \overline{g_\mu}\overline{\xi_\mu(x)}\, dx && (39)\\
&= \sum_\nu \sum_\mu f_\nu \overline{g_\mu} \int \xi_\nu(x)\overline{\xi_\mu(x)}\, dx && (40)\\
&= \sum_\nu \sum_\mu f_\nu \overline{g_\mu} \delta_{\nu,\mu} && (41)\\
&= \sum_\nu f_\nu \overline{g_\nu} && (42)\\
&= < F, G >, && (43)
\end{aligned}$$

where $F = \{f_\nu\}_\nu$ is the coefficient set of all the f_ν, and similarly for G. This is Parseval's relation. Plancherel's theorem is

$$||f||^2 =< f, f >=< F, F >= ||F||_\nu^2, \tag{44}$$

where the first norm is a norm on functions and the second is a norm on the sequence space. These results work, e.g. for Fourier and wavelet systems (and many others) and mean, e.g. that the energy of a function across time is equivalent to the energy across frequency (where energy is a loose wording for the norm squared).

9. *Given a function, $f(t)$, we can compute the continuous wavelet transform (CWT) by*

$$\gamma(s, \tau) = \int f(t)\psi^*_{s,\tau}(t)\, dt, \tag{45}$$

where $*$ denotes complex conjugation and the function $f(t)$ can be reconstructed from the CWT by

$$f(t) = \int\int \gamma(s, \tau)\psi_{s,\tau}(t)\, d\tau ds, \tag{46}$$

where the wavelets are generated by a mother wavelet by

$$\psi_{s,\tau}(t) = s^{-1/2}\psi\left(\frac{t-\tau}{s}\right), \tag{47}$$

where s is the scale factor and τ is the translation or location factor. The rth moment of the wavelet is defined by

$$M_r = \int t^r \psi^*(t)\, dt. \tag{48}$$

A wavelet with p vanishing moments means that $M_r = 0$ for $r = 0, \ldots, p$.

Suppose our wavelet has p vanishing moments, and that $f(t)$ is $(p+1)$-times differentiable. By using a Taylor expansion for $f(t)$ around $t = 0$, show that the wavelet coefficients at scale s are $\gamma(s, 0) = C^* M_{p+1} s^{p+3/2}$, where C^* is some constant.

The Taylor expansion of $f(t)$ around $t = 0$ is

$$\begin{aligned} f(t) &= f(0) + tf^{(1)}(0) + t^2 f^{(2)}(0)/2! + \ldots t^p f^{(p)}(0)/p! + C^* t^{p+1} && (49)\\ &= \sum_{r=0}^{p} f^{(r)}(0)t^r/r! + C^* t^{p+1}, && (50) \end{aligned}$$

using the Lagrange form of the remainder, where C^* is some constant.

Inserting this into the formula for the CWT gives:

$$\gamma(s, 0) = s^{-1/2}\left[\sum_{r=0}^{p} f^{(r)}(0)\int \frac{t^r}{r!}\psi^*\left(\frac{t}{s}\right) dt + C^* \int t^{p+1}\psi^*(t/s)\, dt\right]. \tag{51}$$

Substitute $u = t/s$, $sdu = dt$ gives

$$\begin{aligned} \text{sec} &= s^{-1/2} C^* \int (us)^{p+1} \psi^*(u)\, sdu && (52) \\ &= s^{1/2} s^{p+1} C^* \int u^{p+1} \psi^*(u)\, du && (53) \\ &= s^{p+3/2} C^* M_{p+1}, && (54) \end{aligned}$$

for the second term in (51). Then

$$\begin{aligned} \text{fir} &= s^{1/2} \sum_{r=0}^{p} f^{(r)}(0) \frac{s^r}{r!} \int u^r \psi^*(u)\, du && (55) \\ &= s^{1/2} \sum_{r=0}^{p} f^{(r)}(0) M_r s^r / r!. && (56) \end{aligned}$$

If $M_r = 0$ for $r = 0, \ldots, p$, then fir $= 0$ and $\gamma(s, 0) = C^* M_{p+1} s^{p+3/2}$.