Chapter 5

Basis Expansion and Regularization

Overview

$$\text{RSS}(f, \lambda) = \sum_{i=1}^{N} (y_i - f(x_i))^2 + \lambda \int f''(x)^2 dx.$$

From the solution to Problem 5.7 we know that the minimizer to the above objective function is cubic splines with $x_i, i = 1, 2, \ldots, N$ as the knots, meaning that we can rewrite $f(x)$ as $\sum_{j=1}^{N} N_j(x)\theta_j$ and have

$$\begin{aligned}
\text{RSS} =& \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{N} N_j(x_i)\theta_j \right)^2 + \lambda \int \left(\sum_{j=1}^{N} N_j''(x)\theta_j \right)^2 dx \\
=& ||y - N^T\theta||^2 + \lambda \int \sum_{j=1}^{N} \sum_{k=1}^{N} N_j''(x) N_k''(x) \theta_j \theta_k dx \\
=& ||y - N^T\theta||^2 + \lambda \theta^T \Omega_N \theta.
\end{aligned}$$

The solution to the above problem is $\hat{y} = S_\lambda y$ where

$$\begin{aligned}
S_\lambda =& N(N^T N + \lambda \Omega_N)^{-1} N^T \\
=& (I + \lambda (N^T)^{-1} \Omega_N N^{-1})^{-1} = (I + \lambda K)^{-1} \\
=& \sum_{i=1}^{N} \frac{1}{1 + \lambda d_k} u_k u_k^T \quad \left(\begin{array}{c} d_k \text{ and } u_k \text{ are eigenvalue-eigenvector of matrix } K; \\ d_k \text{ represents the amount of penalty for } u_k. \end{array} \right) \\
=& U^T (I + \lambda D)^{-1} U \quad (\Rightarrow S_\lambda S_\lambda \preceq S_\lambda \Rightarrow S_\lambda \text{ and } K \text{ have the same eigenvector.})
\end{aligned}$$

Problem 5.2

(a) Since $B_{i,1}(x) = \mathbf{1}(\tau_i \leq x < \tau_{i+1})$ the base case is verified. Let us assume that $B_{i,m-1}(x) = 0$ for $x \notin [\tau_i, \tau_{i+m-1}]$ for any i. From the recursive definition, we have

$$B_{i,m}(x) = \frac{x - \tau_i}{\tau_{i+m-1} - \tau_i} B_{i,m-1}(x) + \frac{\tau_{i+m} - x}{\tau_{i+m} - \tau_{i+1}} B_{i+1,m-1}(x),$$

which is zero if $x \notin [\tau_i, \tau_{i+m-1}] \cup [\tau_{i+1}, \tau_{i+m}] = [\tau_i, \tau_{i+m}]$, which completes the proof.

(b) The base case starts with $m = 2$ in this case. Let us assume that $B_{i,m-1}(x) > 0$ for $x \in (\tau_i, \tau_{i+m-1})$ for any i. From the recursive definition, we have

$$B_{i,m}(x) = \frac{x-\tau_i}{\tau_{i+m-1}-\tau_i} B_{i,m-1}(x) + \frac{\tau_{i+m}-x}{\tau_{i+m}-\tau_{i+1}} B_{i+1,m-1}(x),$$

Since the coefficient of $B_{i,m-1}(x)$ is strictly larger than 0 for $x > \tau_i$ and the coefficient of $B_{i,m}(x)$ is strictly larger than 0 for $x < \tau_{i+m}$, we know that $B_{i,m}(x) > 0$ for $x \in (\tau_i, \tau_{i+m-1}) \cup (\tau_{i+1}, \tau_{i+m}) = (\tau_i, \tau_{i+m})$, which completes the proof.

(c) The base case holds for $m = 1$. Let us assume $\sum_{i=1}^{K+M} B_{i,m-1}(x) = 1, \forall x \in [\xi_0, \xi_{K+1}]$. Then we have

$$\begin{aligned}
\sum_{i=1}^{K+M} B_{i,m}(x) &= \sum_{i=1}^{K+M} \left(\frac{x-\tau_i}{\tau_{i+m-1}-\tau_i} B_{i,m-1}(x) + \frac{\tau_{i+m}-x}{\tau_{i+m}-\tau_{i+1}} B_{i+1,m-1}(x) \right) \\
&= \sum_{i=2}^{K+M-1} \left(\frac{x-\tau_i}{\tau_{i+m-1}-\tau_i} + \frac{\tau_{i+m-1}-x}{\tau_{i+m-1}-\tau_i} \right) B_{i,m-1}(x) \\
&= \sum_{i=2}^{K+M-1} B_{i,m-1}(x) = 1.
\end{aligned}$$

(d) It is easy to see that $B_{i,1}$ is a piecewise polynomial of order 1 (degree 0), with breaks at knots $\xi_1, \xi_2, \ldots, \xi_K$. Let us assume that $B_{i,m-1}$ is a piecewise polynomial of order $m-1$ (degree $m-2$), then by the recursive definition of the basis function, it is easy to see that $B_{i,m}$ is a piecewise polynomial of order m with breaks at the knots $\xi_i, \ldots, x_K$.

(e) Let us define $B_1(x) = \mathbf{1}(0 < x < 1)$, and recursively define

$$B_m(x) = \frac{x}{m-1} B_{m-1}(x) + \frac{m-x}{m-1} B_{m-1}(x-1).$$

Then it suffices to show that $B_m(x) = B_1(x) * B_{m-1}(x)$ for any m. Let us make the induction hypothesis that $B_m(x) = B_1(x) * B_{m-1}(x) = \int_0^1 B_{m-1}(x-t)dt$, then we have

$$\begin{aligned}
B_1(x) * B_m(x) &= \int_t \frac{x-t}{m-1} B_{m-1}(x-t) B_1(t) + \frac{m-x+t}{m-1} B_{m-1}(x-t-1) B_1(t) dt \\
&= \int_0^1 \frac{x-t}{m-1} B_{m-1}(x-t) dt + \int_0^1 \frac{m-x+t}{m-1} B_{m-1}(x-t-1) dt
\end{aligned}$$

to be finished [1]

Problem 5.4

Starting from cubic spline equation:

$$f(X) = \sum_{j=0}^{3} \beta_j X^j + \sum_{k=1}^{K} \theta_k (X - \xi_k)_+^3 . \quad (5.1)$$

Since $f''(X)=0$ for $X<\xi_1$, it should be clear that $\beta_2=0$ and $\beta_3=0$. For $X>\xi_1$

$$\begin{aligned} f(X) =&\beta_0+\beta_1 X+\sum_{k=1}^{K}\theta_k(X-\xi_k)^3 \\ =&\beta_0+\beta_1 X+\sum_{k=1}^{K}\theta_k\left(X^3-\xi_k^3+3\xi_k^2X-3\xi_kX^2\right). \end{aligned}$$

Since $f''(X)=0$ for $X>\xi_K$, we know that $\sum_{k=1}^{K}\theta_k=0$ and $\sum_{k=1}^{K}\xi_k\theta_k=0$.

Combining $\sum_{k=1}^{K}\theta_k=0$ and $\sum_{k=1}^{K}\xi_k\theta_k=0$, we have $\theta_K=-\sum_{k=1}^{K-1}\theta_k$ and $\theta_{K-1}=-\sum_{k=1}^{K-2}\frac{\xi_k-\xi_K}{\xi_{K-1}-\xi_K}\theta_k$, which, by plugging into Equation (5.1), completes the proof.

Problem 5.6

The truncated power basis for cubic spline with two knots (ξ_1,ξ_2) are restated below (Equation (5.3) in book)

$$\begin{aligned} h_1(X)=1, h_3(X)=X^2, h_5(X)=(X-\xi_1)_+^3,\\ h_2(X)=X, h_4(X)=X^3, h_6(X)=(X-\xi_2)_+^3. \end{aligned}$$

For periodic function, we know that X, X^2 are not free to choose, as it depends on the point from the other end. Therefore we are left with the following basis

$$h_1(X)=1, h_4(X)=X^3, h_5(X)=(X-\xi_1)_+^3, h_6(X)=(X-\xi_2)_+^3.$$

Furthermore, once the coefficient of $h_1(X), h_4(X), h_5(X)$ are fixed, then there is only one possible choice for the coefficient of $h_6(X)$, in order to make ends meet. More precisely, we need to have

$$\begin{aligned} &\alpha_4h_4(\xi_{K+1})+\alpha_5h_5(\xi_{K+1})+\alpha_6h_6(\xi_{K+1})=0\\ \Rightarrow&\alpha_4\xi_{K+1}^3+\alpha_5(\xi_{K+1}-\xi_1)^3+\alpha_6(\xi_{K+1}-\xi_2)^3=0\\ \Rightarrow&\alpha_6=-\frac{\xi_{K+1}^3}{(\xi_{K+1}-\xi_2)^3}\alpha_4-\frac{(\xi_{K+1}-\xi_1)^3}{(\xi_{K+1}-\xi_2)^3}\alpha_5, \end{aligned}$$

and thus we can modify the basis to

$$h_1(X)=1, \tilde{h}_4(X)=X^3-\frac{\xi_{K+1}^3}{(\xi_{K+1}-\xi_2)^3}(X-\xi_2)_+^3, \tilde{h}_5(X)=(X-\xi_1)_+^3-\frac{(\xi_{K+1}-\xi_1)^3}{(\xi_{K+1}-\xi_2)^3}(X-\xi_2)_+^3.$$

Indeed it is easy to check that $\tilde{h}_4(0)=\tilde{h}_4(\xi_{K+1})$ and $\tilde{h}_5(0)=\tilde{h}_5(\xi_{K+1})$. It is then straightforward to generalize this result and obtain a truncate power basis for periodic cubic spline with any number of knots $\xi_1,\xi_2,\dots,\xi_K$ with boundary of ξ_0 and ξ_{K+1} as

$$\begin{aligned} &h_k(X)=(X-\xi_k)^3-\frac{(\xi_{K+1}-\xi_k)^3}{(\xi_{K+1}-\xi_K)^3}(X-\xi_K)_+^3, \text{ for } k=0,2,\dots,K-1 \text{ and}\\ &h_K(X)=1. \end{aligned}$$

Problem 5.7

(a) Using integration by parts, we obtain

$$\int_a^b g''(x)h''(x)dx = \int_a^b g''(x)dh'(x) = g''(x)h'(x)|_a^b - \int_a^b h'(x)g'''(x)dx.$$

Since $g(x)$ is a natural spline, we know that $g''(x)$ is a continuous piecewise linear function whose value is 0 on the edge, and $g'''(x)$ is a discontinuous piecewise constant function. Therefore $g''(a) = g''(b) = 0$, and

$$\begin{aligned}\int_a^b g''(x)h''(x)dx = -\int_a^b h'(x)g'''(x)dx = -\sum_{j=1}^{N-1}\int_{x_j}^{x_{j+1}} g'''(x)dh(x) = -\sum_{j=1}^{N-1} g'''(x_j^+)\int_{x_j}^{x_{j+1}} dh(x) \\ = -\sum_{j=1}^{N-1} g'''(x_j^+)(h(x_{j+1}) - h(x_j)),\end{aligned}$$

which is 0 given that both g and $\tilde{g}$ interpolate the N training data pairs.

(b)

$$\begin{aligned}\int_a^b \tilde{g}''(t)^2dt &= \int_a^b (h''(t) + g''(t))^2dt = \int_a^b (h''(t)^2 + g''(t)^2 + 2g''(t)h''(t))dt \\ &= \int_a^b h''(t)^2dt + \int_a^b g''(t)^2dt \geq \int_a^b g''(t)^2dt.\end{aligned}$$

(c)

$$\min_f \left[\sum_{i=1}^N (y_i - f(x_i))^2 + \lambda \int_a^b f''(t)^2dt\right].$$

Since natural spline with N knots at $x_1, x_2, \ldots, x_N$ has N degrees of freedom, for any set of points $(x_i, w_i), i = 1, 2, \ldots, N$, we can find a unique function in the space spanned by the natural splines that interpolate those points. The results from (a) and (b) further say that any curve that pass through $(x_i, w_i), i = 1, 2, \ldots, N$ has a norm no smaller than the unique function in the space spanned by the natural splines that interpolate those points. Therefore, the minimizer must be a cubic spline with knots at each of the x_i.

Problem 5.9

$$\begin{aligned}S_\lambda =& N(N^TN + \lambda\Omega_N)^{-1}N^T \\ =& ((N^T)^{-1}N^TNN^{-1} + \lambda(N^T)^{-1}\Omega_N N^{-1})^{-1} \\ =& (I + \lambda K)^{-1},\end{aligned}$$

where $K = (N^T)^{-1}\Omega_N N^{-1}$.

Problem 5.12

Let us rewrite the weighted RSS in vector form:

$$\begin{aligned}\text{RSS}(f,\lambda) =& \sum_{i=1}^{N} w_i (y_i - f(x_i))^2 + \lambda \int f''(t)^2 dt \\ & (y - N\theta)^T W (y - N\theta) + \theta^T \Omega_N \theta.\end{aligned}$$

By enforcing the derivative to be zero, we can obtain the minimizer of the above equation as

$$\theta = (N^T W N + \lambda\Omega)^{-1} N^T W y,$$

which leads to

$$\begin{aligned}\hat{y} =& N(N^T W N + \lambda\Omega)^{-1} N^T W y \\ & (W + \lambda K)^{-1} W y \\ & (W + \lambda U D U^T)^{-1} W y,\end{aligned} \tag{5.2}$$

where W is a diagonal matrix with the i^{th} entry being w_i.

Problem 5.13

Let us first establish a connection between N-fold cross validation with the solution to Problem 5.12. With weight matrix $W = I$, the solution to Problem 5.12 reduces to the normal smoothing spline problem, for which we have

$$\begin{aligned}\hat{y} &= (I + \lambda K)^{-1} y \\ \Rightarrow \hat{y} + \lambda K \hat{y} &= y.\end{aligned} \tag{5.3}$$

Consider a weight matrix W where the diagonal is 1 except for the i^{th} term, denoted as $W^{(-i)}$, then this effectively results in a solution that leaves out the effect of the i^{th} observation from the data. Resume from Equation (5.2), we obtain

$$\begin{aligned}\hat{y}^{(-i)} &= \left(W^{(-i)} + \lambda K\right)^{-1} W^{(-i)} y \\ \Rightarrow W^{(-i)} \hat{y}^{(-i)} + \lambda K \hat{y}^{(-i)} &= y^{(-i)} \\ \Rightarrow \hat{y}^{(-i)} - e_i \hat{y}_i^{(-i)} + \lambda K \hat{y}^{(-i)} &= y^{(-i)}.\end{aligned} \tag{5.4}$$

By subtracting Equation (5.4) from Equation (5.3), we finish the proof:

$$
\begin{aligned}
e_i\hat{y}_i^{(-i)} + \hat{y} - \hat{y}^{(-i)} + \lambda K\left(\hat{y} - \hat{y}^{(-i)}\right) &= y - y^{(-i)} \\
e_i\hat{y}_i^{(-i)} + (I + \lambda K)\left(\hat{y} - \hat{y}^{(-i)}\right) &= e_i y_i \\
\hat{y} - \hat{y}^{(-i)} &= (I + \lambda K)^{-1} e_i \left(y_i - \hat{y}_i^{(-i)}\right) \\
\hat{y} - \hat{y}^{(-i)} &= S_\lambda e_i \left(y_i - \hat{y}_i^{(-i)}\right) \\
\hat{y}_i - \hat{y}_i^{(-i)} &= S_{ii} \left(y_i - \hat{y}_i^{(-i)}\right) \\
-(y_i - \hat{y}_i) + \left(y_i - \hat{y}_i^{(-i)}\right) &= S_{ii} \left(y_i - \hat{y}_i^{(-i)}\right) \\
y_i - \hat{y}_i^{(-i)} &= \frac{1}{1 - S_{ii}}(y_i - \hat{y}_i).
\end{aligned}
$$

Problem 5.14 - unfinished

$$
f(x) = \beta_0 + \beta^T x + \sum_{j=1}^{N} \alpha_j h_j(x).
$$

$$
h_j(x) = \frac{1}{2}||x - x_j||^2 \log ||x - x_j||^2..
$$

$$
\begin{aligned}
\frac{\partial h_j(x)}{\partial x_1} =& \frac{1}{2}\left(\log ||x - x_j||^2 + 1\right)\frac{\partial ||x - x_j||^2}{\partial x_1} = \left(\log ||x - x_j||^2 + 1\right)(x_1 - x_{j1}) \\
\frac{\partial^2 h_j(x)}{\partial x_1^2} =& \log ||x - x_j||^2 + 1 + 2(x_1 - x_{j1})^2 \frac{1}{||x - x_j||^2} \\
\frac{\partial^2 h_j(x)}{\partial x_1 \partial x_2} =& 2(x_1 - x_{j1})(x_2 - x_{j2})\frac{1}{||x - x_j||^2} \\
\left(\frac{\partial^2 f(x)}{\partial x_1 \partial x_2}\right)^2 =& \sum_{j=1}^{N}\sum_{k=1}^{N} \alpha_j \alpha_k \left(\frac{2(x_1 - x_{1j})(x_2 - x_{2j})}{||x - x_j||^2}\right)\left(\frac{2(x_1 - x_{1k})(x_2 - x_{2k})}{||x - x_k||^2}\right) \\
\left(\frac{\partial^2 f(x)}{\partial x_1^2}\right)^2 =& \sum_{j=1}^{N}\sum_{k=1}^{N} \alpha_j \alpha_k \left(\log ||x - x_j||^2 + 1 + \frac{2(x_1 - x_{j1})^2}{||x - x_j||^2}\right)\left(\log ||x - x_k||^2 + 1 + \frac{2(x_1 - x_{k1})^2}{||x - x_k||^2}\right)
\end{aligned}
$$

Problem 5.15

Mercer's theorem says the following:

- Any PD kernel $K(\cdot, \cdot)$ can be expressed as an eigen-expansion of

$$
K(x, y) = \sum_{i=1}^{\infty} c_i \phi_i(x)\phi_i(y),
$$

which can be viewed as an eigen-decomposition of the positive definite matrix.

- Define the space of functions $\mathcal{H}_K$ generated by the linear span of $\{K(\cdot, y), y \in \mathbb{R}^d\}$, then elements of $\mathcal{H}_K$ have an expansion in terms of the eigen-functions $\phi_i(\cdot)$:

$$f(x) = \sum_{i=1}^{\infty} c_i \phi_i(x), \text{ with}$$
$$||f||^2_{\mathcal{H}_K} \triangleq \sum_{i=1}^{\infty} c_i^2/\gamma_i < \infty.$$

Since $\phi_i(\cdot), i = 1, 2, \dots$ are eigen-functions, we have

$$\begin{aligned}||f||^2_{\mathcal{H}_K} &= \left\langle \sum_{i=1}^{\infty} c_i\phi_i(\cdot), \sum_{i=1}^{\infty} c_i\phi_i(\cdot) \right\rangle_{\mathcal{H}_K} = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} c_i c_j \left\langle \phi_i(\cdot), \phi_j(\cdot)\right\rangle_{\mathcal{H}_K} \\ &= \sum_{i=1}^{\infty} c_i^2 \left\langle \phi_i(\cdot), \phi_i(\cdot)\right\rangle_{\mathcal{H}_K} = \sum_{i=1}^{\infty} c_i^2/\gamma_i \\ &\Rightarrow \left\langle \phi_i(\cdot), \phi_i(\cdot)\right\rangle_{\mathcal{H}_K} = 1/\gamma_i.\end{aligned}$$

(a)

$$\begin{aligned}\langle K(\cdot, x_i), f\rangle_{\mathcal{H}_K} &= \left\langle \sum_j \gamma_j\phi_j(x_i)\phi_j(\cdot), \sum_l c_l\phi_l(\cdot) \right\rangle_{\mathcal{H}_K} \\ &= \sum_j \gamma_j \phi_j(x_i) \sum_l c_l \left\langle \phi_j(\cdot), \phi_l(\cdot)\right\rangle_{\mathcal{H}_K} \\ &= \sum_j \gamma_j\phi_j(x_i) c_j 1/\gamma_i = f(x_i)\end{aligned}$$

(b)

$$\begin{aligned}\langle K(\cdot, x_i), K(\cdot, x_j)\rangle &= \left\langle \sum_l \gamma_l\phi_l(x_i)\phi_l(\cdot), \sum_k \gamma_k\phi_k(x_j)\phi_k(\cdot) \right\rangle_{\mathcal{H}_K} \\ &= \sum_l\sum_k \gamma_l\gamma_k\phi_l(x_i)\phi_k(x_j) \left\langle \phi_l(\cdot), \phi_k(\cdot)\right\rangle_{\mathcal{H}_K} \\ &= \sum_l \gamma_l\phi_l(x_i)\phi_l(x_j) = K(x_i, x_j)\end{aligned}$$

(c)

$$J(g) = ||g||_{\mathcal{H}_K} = \left\langle \sum_{i=1}^{N} \alpha_i K(\cdot, x_i), \sum_{i=1}^{N} \alpha_i K(\cdot, x_i) \right\rangle_{\mathcal{H}_K} = \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i\alpha_j K(x_i, x_j).$$

(d) Since $\rho(x) \in \mathcal{H}_K$ is orthogonal in $\mathcal{H}_K$ to each of $K(x, x_i), i = 1, 2, \dots, N$, we have

$$\langle \rho(\cdot), K(\cdot, x_i)\rangle_{\mathcal{H}_K} = 0 \Rightarrow \rho(x_i) = 0.$$

Then we have

$$
\begin{aligned}
&\sum_{i=1}^{N} L(y_i, \tilde{g}(x_i)) + \lambda J(\tilde{g}) \\
=&\sum_{i=1}^{N} L(y_i, g(x_i) + \rho(x_i)) + \lambda J(g) + \lambda J(\rho) \\
=&\sum_{i=1}^{N} L(y_i, g(x_i)) + \lambda J(g) + \lambda J(\rho) \\
\geq&\sum_{i=1}^{N} L(y_i, g(x_i)) + \lambda J(g).
\end{aligned}
$$

Problem 5.16

$$
\begin{aligned}
&\sum_{i=1}^{N}\left(y_i - \sum_{i=1}^{M}\beta_m h_m(x_i)\right)^2 + \lambda\sum_{m=1}^{M}\beta_m^2 \\
=&\sum_{i=1}^{N}\left(y_i - h(x)^T\beta\right)^2 + \lambda\beta^T\beta \\
=&\sum_{i=1}^{N}\left(y_i - \phi(x)^T D_\gamma^{\frac{1}{2}} V^T\beta\right)^2 + \lambda\beta^T\beta \\
\left(\substack{\text{change of coordination}\\ b=V^T\beta}\right) =&\sum_{i=1}^{N}\left(y_i - \phi(x)^T D_\gamma^{\frac{1}{2}} b\right)^2 + \lambda b^T b \\
\left(c = D_\gamma^{\frac{1}{2}} b\right) =&\sum_{i=1}^{N}\left(y_i - \phi(x)^T c\right)^2 + \lambda c^T D_\gamma^{-1} c \\
=&\sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{M} c_j\phi_j(x_i)\right)^2 + \lambda\sum_{j=1}^{M}\frac{c_j^2}{\gamma_j}.
\end{aligned}
$$

Unfinished Problems

Problem 5.1

- Problem 5.3
- Problem 5.8
- Problem 5.10
- Problem 5.11
- Problem 5.14
- Problem 5.16
- Problem 5.17
- Problem 5.18
- Problem 5.19