Statistical Learning HW 1

Johnson Lin

October 28, 2019

1 Problem 1

ESL Ex. 5.1

Equation 5.3 is

$$h_1(X) = 1, h_3(X) = X^2, h_5(X) = (X - \xi_1)^3_+,$$

$$h_2(X) = X, h_4(X) = X^3, h_6(X) = (X - \xi_2)^3_+.$$

We can then write $f(x)$ as a linear combination of these basis functions:

$$f(x) = \sum_{m=1}^{6} \beta_m h_m(x)$$

We need to show that $f(x)$ is continuous at the two knots, ξ_1 and ξ_2. We also need to show this for $f'(x)$ and $f''(x)$. First we look at the continuity of $f(x)$ at $x = \xi_1$

$$f(\xi_1 - h) = \beta_1 + \beta_2(\xi_1 - h) + \beta_3(\xi_1 - h)^2 + \beta_4(\xi_1 - h)^3 + \beta_5(\xi_1 - h - \xi_1)^3 + \beta_6(\xi_1 - h - \xi_2)^3_+$$

As $h \to 0$,

$$= \beta_1 + \beta_2\xi_1 + \beta_3\xi_1^2 + \beta_4\xi_1^3 + 0 + 0$$

This gives us the left limit. Now doing the same thing for the right limit:

$$f(\xi_1 + h) = \beta_1 + \beta_2(\xi_1 + h) + \beta_3(\xi_1 + h)^2 + \beta_4(\xi_1 + h)^3 + \beta_5(\xi_1 + h - \xi_1)^3 + \beta_6(\xi_1 + h - \xi_2)^3_+$$

$$= \beta_1 + \beta_2\xi_1 + \beta_3\xi_1^2 + \beta_4\xi_1^3 + 0 + 0$$

Both limits are equal, so $f(x)$ is continuous in this case. Now we look at continuity of $f'(x)$ starting with the left side:

$$f'(\xi_1) = lim_{h \to 0} \frac{f(\xi_1) - f(\xi_1 - h)}{h}$$

$$(\beta_1 + \beta_2\xi_1 + \beta_3\xi_1^2 + \beta_4\xi_1^3) - (\beta_1 + \beta_2(\xi_1 - h) + \beta_3(\xi_1 - h)^2 + \beta_4(\xi_1 - h)^3 + \beta_5(\xi_1 - h - \xi_1)^3 + \beta_6(\xi_1 - h - \xi_2)^3_+)$$

$$= \beta_2 h + 2\beta_3\xi_1 h - \beta_3 h^2 + 3\beta_4\xi_1^2 h - 3\beta_4\xi_1 h^2 + \beta_4 h^3 + \beta_5 h^3 + \beta_6(\xi_1 - h - \xi_2)^3_+$$

Because we are dividing by h and then looking at $h \to 0$, we only care about terms that have an order of 1. Anything greater will be reduced to 0.

$$= \beta_2 h + 2\beta_3\xi_1 h + 3\beta_4\xi_1^2 h$$

$$= \beta_2 + 2\beta_3\xi_1 + 3\beta_4\xi_1^2$$

Now we look at the right side,

$$f'(\xi_1) = lim_{h\to 0}\frac{f(\xi_1 + h) - f(\xi_1}{h}$$

$$(\beta_1+\beta_2(\xi_1+h)+\beta_3(\xi_1+h)^2+\beta_4(\xi_1+h)^3+\beta_5(\xi_1+h-\xi_1)^3+\beta_6(\xi_1+h-\xi_2)^3_+)-(\beta_1+\beta_2\xi_1+\beta_3\xi_1^2+\beta_4\xi_1^3)$$

$$= \beta_2 h + 2\beta_3\xi_1 h + \beta_3 h^2 + 3\beta_4\xi_1^2 h + 3\beta_4\xi_1 h^2 + \beta_4 h^3 + \beta_5 h^3 + \beta_6(\xi_1 + h - \xi_2)^3_+$$

$$= \beta_2 h + 2\beta_3\xi_1 h + 3\beta_4\xi_1^2 h$$

$$= \beta_2 + 2\beta_3\xi_1 + 3\beta_4\xi_1^2$$

Thus, proving the continuity for $f'(x)$.

To look at $f''(\xi_1)$, we can use the same terms we generated from above, but take the derivative again with respect to ξ_1. Because we have shown that $f'(x)$ is equal for both sides, we know that $f''(x)$ is also the same and that it is:

$$f'(\xi_1) = \beta_2 + 2\beta_3\xi_1 + 3\beta_4\xi_1^2$$

$$f''(\xi_1) = 2\beta_3 + 6\beta_4\xi_1$$

It follows very similarly if $x = \xi_2$. Thus, we have shown continuity and proven it is a basis for a cubic spline at those two knots.

2 Problem 2

ESL Ex. 5.4 Equations 5.4 and 5.5 are respectively

$$N_1(X) = 1, N_2(X) = X, N_{k+2}(X) = d_k(X) - d_{K-1}(X),$$

where

$$d_k(X) = \frac{(X - \xi_k)^3_+ - (X - \xi_K)^3_+}{\xi_K - \xi_k}$$

Natural cubic splines enforce linearity in both boundary regions. This means that the coefficients of x^2 and x^3 must be zero. In the left boundary region, this is

$$\beta_2 = 0, \beta_3 = 0$$

In the right boundary region, the prediction function is written as

$$f(x)\sum_{j=0}^{3}\beta_j x^j + \sum_{k=1}^{K}\sigma_k(x - \xi_k)^3$$

$$= \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{k=1}^{K} \sigma_k (x^3 - \xi_k^3 - 3x^2 \xi_k + 3x\xi_k^2)$$

Since $\beta_2, \beta_3 = 0$,

$$= \beta_0 + \beta_1 x + \sum_{k=1}^{K} \sigma_k (x^3 - \xi_k^3 - 3x^2 \xi_k + 3x\xi_k^2)$$

Since we know that the coefficient of x^3 should be zero, we know

$$\sigma_{k=1}^{K} \sigma_k = 0$$

The coefficient of x^2 should also be zero, so

$$-\sigma_{k=1}^{K} 3\sigma_k \xi_k = 0$$

which implies that

$$\sigma_{k=1}^{K} \sigma_k \xi_k = 0$$

Thus we have proven that the linear constraints on the coefficients follow. Now we need to prove that the basis of a natural cubic spline is given by Equation 5.4 and 5.5. A generic cubic spline can be expressed as

$$f(x) = \sum_{j=0}^{3} \beta_j x^j + \sum_{k=1}^{K} \sigma_k (x - \xi_k)_+^3$$

$$= \beta_0 + \beta_1 x + g(x)$$

We know from our linear constraints that

$$\sigma_K = -\sum_{k-1}^{K-1} \sigma_k$$

Substituting in $g(x)$,

$$g(x) = \sum_{k=1}^{K-1} \sigma_k ((x - \xi_k)_+^3 - (x - \xi_K)_+^3)$$

$$\sigma_{K-1} = \sum_{k=1}^{K-2} \frac{\xi_k - \xi_K}{\xi_K - \xi_{K-1}}$$

So now we get,

$$g(x) = \sum_{k=1}^{K-2} \sigma_k ((x - \xi_k)_+^3 - (x - \xi_K)_+^3) + \sum_{k=1}^{K-2} \sigma_k \frac{\xi_k - \xi_K}{\xi_K - \xi_{K-1}} ((x - \xi_{K-1})_+^3 - (x - \xi_K)_+^3)$$

$$= \sum_{k=1}^{K-2} \sigma_k (\xi_k - \xi_k)(d_k(x) - d_{K-1}(x))$$

$$= \sum_{k=1}^{K-2} \phi_k(d_k(x) - d_{K-1}(x))$$

where

$$\phi_k = \sigma_k(\xi_k - \xi_K)$$

and $d_k(x)$ is given by Equation 5.5 stated previously. So now we can write the equation as a Natural Cubic Spline, finishing the proof:

$$f(x) = \beta_0 + \beta_1 x + \sum_{k=1}^{K-2} \phi_k(d_k(x) - d_{K-1}(x))$$

3 Problem 3

ESL Ex. 5.7

a. g is a natural cubic spline, so it is linear outside the region bound by the knots x_1 and x_N. This means any derivative with order higher than one outside the region is 0. Since a and b are outside this region, then we know $g''(a) = g''(b) = 0$. Now we have

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

$$= 0 - \int_a^b g'''(x)h''(x)dx$$

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

Because $g(x)$ is a cubic function, the fourth derivative $g''''(x) = 0$. So what we're left with is

$$= -g'''(x)h''(x)|_a^b$$

which we can rewrite as

$$= -g'''(x)h''(x)|_{x_1}^{x_N}$$

$$= -g'''(x)h''(x)|_{x_1}^{x_2} - \cdots - g'''(x)h''(x)|_{x_{N-1}}^{x_N}$$

$$= -\sum_{j=1}^{N-1} (g'''(x_{j+1}^-)h''(x_{j+1}) - g'''(x_j^+)h''(x_j))$$

Because the third derivative of $g(x)$ is the same for the region inside, between the two knots,

$$= -\sum_{j=1}^{N-1} g'''(x_j^+)(h''(x_{j+1}) - h''(x_j))$$

And because $h(x) = 0$ at all knots,

$$= 0$$

b. From part a,

$$\int_a^b g''(x)h''(x)dx = 0$$

$$\int_a^b g''(x)(\tilde{g}''(x) - g''(x))dx = 0$$

$$\int_a^b g''(x)^2 = \int_a^b g''(x)\tilde{g}''(x)$$

Using the Cauchy-Schwarz inequality,

$$\int_a^b g''(x)^2 \leq (\int_a^b g''(x)^2)^{1/2}(\int_a^b \tilde{g}''(x)^2)^{1/2}$$

If $g''(x) = \tilde{g}''(x)$ then we get,

$$\int_a^b g''(x)^2 \leq \int_a^b \tilde{g}''(x)^2$$

For this inequality to hold our first assumption must be true. If $g''(x) = \tilde{g}''(x)$ then this implies that $h''(x) = 0$ in $[a, b]$.

c. Let $\tilde{g}(x)$ be the function that minimizes the penalized least squares problem. $g(x)$ is a natural cubic spline with knots at $x_1, \ldots, x_N$ and that it satisfies $g(x_i) = \tilde{g}(x_i)$ for all i. From part b, we know

$$\int_a^b g''(x)^2 \leq \int_a^b \tilde{g}''(x)^2$$

$$\lambda \int_a^b g''(x)^2 \leq \lambda \int_a^b \tilde{g}''(x)^2$$

which holds true as long as $\lambda > 0$. So our natural cubic spline function $g(x)$ minimizes the equation better than $\tilde{g}(x)$. Since we know that $\tilde{g}(x)$ is also a minimizer, it must also be a natural cubic spline.

4 Problem 4

ESL Ex. 5.13 Equation 5.26 is

$$CV(\hat{f}_\lambda) = \sum_{i=1}^{N}(y_i - \hat{f}_\lambda^{-1}(x_i))^2$$

The equation to minimize the normal smoothing spline problem is

$$\hat{y} = (I + \lambda K)^{-1}y$$

However, in this case, we have replaced $\hat{y}$ with $\hat{f}_\lambda(x_0)$.

$$\hat{f}_\lambda(x_0) = (I + \lambda K)^{-1}y$$

$$\hat{f}_\lambda(x_0) + \lambda K \hat{f}_\lambda(x_0) = y$$

Consider a weight matrix W where all of the diagonals are 1 except for the i^{th} term, W^{-i}. This creates a solution that essentially disregards the i^{th} term from the data. So

$$\hat{f}_\lambda(x_0)^{-i} = (W^{-1} + \lambda K)^{-1} W^{-i} y$$

$$W^{-i}\hat{f}_\lambda(x_0)^{-i} + \lambda K \hat{f}_\lambda(x_0)^{-i} = y^{-i}$$

$$\hat{f}_\lambda(x_0)^{-i} - e_i \hat{f}_\lambda(x_0)^{-i} + \lambda K \hat{f}_\lambda(x_0)^{-i} = y^{-i}$$

By subtracting the two equations, we get:

$$e_i \hat{f}_\lambda(x_0)^{-i} + \hat{f}_\lambda(x_0) - \hat{f}_\lambda(x_0)^{-i} + \lambda K(\hat{f}_\lambda(x_0) - \hat{f}_\lambda(x_0)^{-i}) = y - y^{-i}$$

$$e_i \hat{f}_\lambda(x_0)^{-i} + (I + \lambda K)(\hat{f}_\lambda(x_0) - \hat{f}_\lambda(x_0)^{-i}) = e_i y_i$$

$$(\hat{f}_\lambda(x_0) - \hat{f}_\lambda(x_0)^{-i}) = (I + \lambda K)^{-1} e_i (y_i - \hat{f}_\lambda(x_0)^{-i})$$

$$\hat{f}_\lambda(x_0) - \hat{f}_\lambda(x_0)^{-i} = S_\lambda e_i (y_i - \hat{f}_\lambda(x_0)^{-i})$$

$$\hat{f}_\lambda(x_0) - \hat{f}_\lambda(x_0)^{-i} = S_{ii}(y_i - \hat{f}_\lambda(x_0)^{-i})$$

$$-(y_i - \hat{f}_\lambda(x_0)) + (y_i - \hat{f}_\lambda(x_0)^{-i}) = S_{ii}(y_i - \hat{f}_\lambda(x_0)^{-i})$$

$$(y_i - \hat{f}_\lambda(x_0)^{-i}) = \frac{1}{1 - S_{ii}}(y_i - \hat{f}_\lambda(x_0))$$

5 Problem 5

ESL Ex. 5.15

a. Using Mercer's theorem, we know that any PD Kernel K can be expressed as an eigen-expansion of

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

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

$$f(x) = \sum_{i=1}^{\infty} c_i \phi_i(x)$$

$$||f||^2_{\mathcal{H}_K} \triangleq \sum_{i=1}^{\infty} c_i^2 / \gamma_i < \infty$$

Since $\phi_i(\cdot) = 1, 2, \ldots$ are all eigen-functions,

$$||f||^2_{\mathcal{H}_K} = \langle \sum_i c_i \phi_i(\cdot), \sum_i c_i \phi_i(\cdot) \rangle$$

$$= \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} c_i c_j \langle \phi_i(\cdot), \phi_j(\cdot) \rangle_{\mathcal{H}_K}$$

$$= \sum_{i=1}^{\infty} c_i^2 \langle \phi_i(\cdot), \phi_i(\cdot) \rangle_{\mathcal{H}_K}$$

$$= \sum_{i=1}^{\infty} c_i^2 / gamma_i$$

which implies

$$\langle \phi_i(\cdot), \phi_i(\cdot) \rangle_{\mathcal{H}_K} = 1/\gamma_i$$

So

$$\langle K(\cdot, x_i), f \rangle_{\mathcal{H}_K} = \langle \sum_j \gamma_j \phi_j(x_i) \phi_j(\cdot), \sum_l c_l \phi_l(\cdot) \rangle_{\mathcal{H}_K}$$

$$= \sum_j \gamma_j \phi_j(x_i) \sum_l c_l \langle \phi_j(\cdot), \phi_l(\cdot) \rangle_{\mathcal{H}_K}$$

$$= \sum_j \gamma_j \phi_j(x_i) c_j (1/\gamma_i) = f(x_i)$$

b. Both parts b and c follow easily after everything established in part a.

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

$$= \sum_l \sum_k \gamma_l \gamma_k \phi_l(x_i) \phi_k(x_j) \langle \phi_l(\cdot), \phi_k(\cdot) \rangle_{\mathcal{H}_K}$$

$$= \sum_l \gamma_l \phi_l(x_i) \phi_l(x_j) = K(x_i, x_j)$$

c.

$$J(g) = ||g||_{\mathcal{H}_k}$$

$$= \langle \sum_{i=1}^{N} \alpha_i K(\cdot, x_i), \sum_{i=1}^{N} \alpha_i K(\cdot, x_i \rangle_{\mathcal{H}_K}$$

$$= \sum_{i=1}^{N} \sum_{i=1}^{N} K(x_i, x_j) a_i a_j$$

d. We have $\rho(x)$ is orthogonal to $\mathcal{H}_K$. This means that we have

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

which implies that

$$\rho(x_i) = 0$$

So we have

$$\sum_{i=1}^{N} L(y_i, \tilde{g}(x_i) + \lambda J(\tilde{g})$$

$$= \sum_{i=1}^{N} L(y_i, g(x_i) + \rho(x_i)) + \lambda J(g) + \lambda J(\rho)$$

$$= \sum_{i=1}^{N} L(y_i, g(x_i) + \rho(x_i)) + \lambda J(g) + \lambda J(\rho)$$

Since $\lambda > 0$,

$$\sum_{i=1}^{N} L(y_i, g(x_i) + \rho(x_i)) + \lambda J(g) + \lambda J(\rho) \geq \sum_{i=1}^{N} L(y_i, g(x_i) + \rho(x_i)) + \lambda J(g)$$

So we have proven that

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

as long as $\rho(x) = 0$.

6 Problem 6

ESL Ex. 6.2 Define $\mathbf{l}(x_0) = [l_1(x_0), l_2(x_0), \ldots, l_N(x_0)]^T$. It is a $N \times 1$ vector. Equation 6.8 gives us

$$\mathbf{l}^T(x_0) = b(x_0)^T (B^T W(x_0) B)^{-1} B^T W(x_0)$$

$$\begin{aligned}\mathbf{l}^T(x_0) B &= [\sum_{i=1}^{N} l_i(x_0), \ldots, \sum_{i=1}^{N} l_i(x_0) x_i^k] \\ &= b^T(x_o) \\ &= [1, \ldots, x_0^k]\end{aligned}$$

So,

$$\sum_{i=1}^{N} l_i(x_0) x_i^j = x_0^j, \quad j = 0, \ldots, k$$

This equation tells us that when $j = 0$, we get

$$\sum_{i=1}^{N} l_i(x_0) = 1$$

If $j > 0$, then we get

$$\sum_{i=1}^{N} l_i(x_0) x_i^j = x_0^j = x_0^l \times x_0^{j-l}$$

$$= \sum_{i=1}^{N} l_i(x_0) x_i^l x_0^{j-l}$$

Given this equation, for any $0 \leq l \leq j$, it follows that

$$\sum_{i=1}^{N} l_i(x_0)(x_i - x_0)^j = 0$$

This suggests that the bias is 0 for any order k.

7 Problem 7

ESL Ex. 6.3 The variance should hold true for any weighting matrix W, so we will look at the simplest case where $W = I$, the identity matrix.

$$y = X\beta + u \qquad \text{with } y, u \in \mathbf{R}^n; X \in \mathbf{R}^{n \times k}; \beta \in \mathbf{R}^k$$

Let $(x_1, x_2, \ldots, x_n)^T =: x \in \mathbf{R}^n$ be some vector,

$$X := [x^0, x^1, \ldots, x^{k-1}]$$

The OLS estimate for the weights are

$$\hat{\beta} := (X^T X)^{-1} X^T y$$

The estimate for y is

$$\hat{y}_t := z^T \hat{\beta}$$

where

$$z := \begin{bmatrix} t^0 \\ t^1 \\ \vdots \\ t^{k-1} \end{bmatrix} \in \mathbf{R}^k$$

Next, we find the expected value of $\hat{y}_t$:

$$\mathbf{E}[\hat{y}_t] = \mathbf{E}[z^T \hat{\beta}]$$

$$= \mathbf{E}[z^T (X^T X)^{-1} X^T y]$$

$$= \mathbf{E}[z^T (X^T X)^{-1} X^T X \beta + z^T (X^T X)^{-1} X^T u]$$

$$= z^T \beta + z^T (X^T X)^{-1} X^T \mathbf{E}[u]$$

$$= z^T \beta$$

Now we know

$$\hat{y}_t - \mathbf{E}[\hat{y}_t] = z^T (X^T X)^{-1} X^T u$$

So now we can find the variance:

$$Var[\hat{y}_t] = \mathbf{E}[(\hat{y}_t - \mathbf{E}[\hat{y}_t])(\hat{y}_t - \mathbf{E}[\hat{y}_t])^T]$$
$$= \mathbf{E}[(z^T (X^T X)^{-1} X^T u)(z^T (X^T X)^{-1} X^T u)^T]$$
$$= \mathbf{E}[(z^T (X^T X)^{-1} X^T u)(u^T X (X^T X)^{-1} z)]$$
$$= z^T (X^T X)^{-1} X^T \mathbf{E}[uu^T] X (X^T X)^{-1} z$$
$$= \sigma^2 z^T (X^T X)^{-1} z$$

Now we increase k to $k+1$:

$$X := [x^0, x^1, \ldots, x^{k-1}, x^k] \in \mathbf{R}^{n \times (k+1)}$$

$$z := \begin{bmatrix} t^0 \\ t^1 \\ \vdots \\ t^{k-1} \\ t^k \end{bmatrix} \in \mathbf{R}^{k+1}$$

$$Var[\hat{y}_t] = \sigma^2 z^T (X^T X)^{-1} z$$

Variance of $\hat{y}_t$ turns into a $(k+1) \times (k+1)$ matrix. We need to compare this to the original variance matrix which is a $k \times k$ matrix. We can write our new X and z as:

$$X := [X, x^k], z := \begin{bmatrix} z \\ t^k \end{bmatrix}$$

$$Var[\hat{y}_t] = \sigma^2 (z^T, t^k) \begin{bmatrix} X^T X & X^T x^k \\ (x^k)^T X & (x^k)^T x^k \end{bmatrix}^{-1} \begin{bmatrix} z \\ t^k \end{bmatrix}$$

Since we are looking for the inverse, we can use the Schur complement

$$Var[\hat{y}_t] = \sigma^2 (z^T, t^k) \begin{bmatrix} (X^T X - X^T x^k ((x^k)^T x^k)^{-1} (x^k)^T X)^{-1} & X^T x^k \\ (X^T x^k)^T & (x^k)^T x^k \end{bmatrix}^{-1} \begin{bmatrix} z \\ t^k \end{bmatrix}$$

$$= \sigma^2 (z^T (X^T X - X^T x^k ((x^k)^T x^k)^{-1} (x^k)^T X)^{-1} z + t^k z^T (X^T x^k) + t^k (X^T x^k)^T z + t^{2k} (x^k)^T x^k)$$

The matrix $X^T x^k ((x^k)^T x^k)^{-1} (x^k)^T X$ can be written as

$$X^T x^k ((x^k)^T x^k)^{-1} (x^k)^T X = ((x^k)^T x^k)^{-1} X^T x^k ((x^k)^T X$$

We can think of $x^k((x^k)^T$ as a rank 1 matrix with the only non-vanishing eigenvalue which is equal to $((x^k)^T x^k$. The matrix $x^k((x^k)^T x^k)^{-1}(x^k)^T$ can be thought of as the projection on the subspace spanned by x^k. This means that

$$X^T X \geq X^T x^k ((x^k)^T x^k)^{-1} (x^k)^T X$$
$$(X^T X)^{-1} \leq (X^T x^k ((x^k)^T x^k)^{-1} (x^k)^T X)^{-1}$$

Thus, we can now calculate every term in our new variance and it follows that the variance increases if polynomial degree increases.

8 Problem 8

ESL Ex. 6.5 Equation 6.19 gives us the local log-likelihood for the J class model:

$$\sum_{i=1}^{N} K_\lambda(x_0, x_i)\{\beta_{g_i,0}(x_0)+\beta_{g_i}(x_0)^T(x_i-x_0)-log[1+\sum_{k=1}^{J-1} exp(B_{k0}(x_0)+b_k(x_0)^T(x_i-x_0))]\}$$

We can write this as

$$l(\beta) = \sum_{i=1}^{N} K_\lambda(x_0, x_i)[\sum_{j=1}^{J-1} y_{ij} logPr(G = j|X = x_i)+(1-\sum_{j=1}^{J-1} y_{ij})logPr(G = J|X = x_i)]$$

$$= \sum_{i=1}^{N} K_\lambda(x_0, x_i)[\sum_{j=1}^{J-1} y_{ij}\beta_{j0}(x_0) - log(1 + \sum_{j=1}^{J-1} exp(\beta_{j0}(x_0))]$$

To maximize this, we need to set the derivative equal to 0:

$$\frac{\partial l(\beta)}{\partial \beta_{j0}(x_0)} = \sum_{i=1}^{N} K_\lambda(x_0, x_i)(y_{ij} - \frac{exp(\beta_{j0}(x_0))}{1 + \sum_{k=1}^{J-1} exp(\beta_{k0}(x_o))}) = 0$$

We can select $\beta_{j0} = 1, \ldots, J-1$ so that

$$= \sum_{i=1}^{N} K_\lambda(x_0, x_i)(y_{ij} - \frac{\sum_{i=1}^{N} K_\lambda(x_0, x_i)y_{ij}}{\sum_{i=1}^{N} K_\lambda})$$

Thus we have shown that this amounts to smoothing the binary response indicators for each class separately, using a Nadaraya-Watson kernel smoother with kernel weights $K_\lambda(x_0, x_i)$