Chapters 4 and 5

Abhimanyu Talwar

November 21, 2018

ESL Problem 4.3
NOTE: This problem was solved with the help of hints from Teaching Fellows of the course STAT 195 at Harvard University (Fall 2018).

Let Σ and $\hat{\Sigma}$ represent the variance-covariance matrices for X and $\hat{Y}$ respectively. Let μ_k and $\hat{\mu}_k$ represent the means for class k for X and $\hat{Y}$ respectively. Let B represent the matrix given by $(X^TX)^{-1}X^TY$, where we have the relation $\hat{Y} = XB$. Let Y_k represent the K^{th} column of the matrix Y. Let g_i represent the class label of the i^{th} training sample. We can then write:

$$\hat{\mu}_k = \sum_{g_i=k} \frac{\hat{y}_i}{N_k} \tag{1}$$

$$= \sum_{g_i=k} \frac{B^T x_i}{N_k} \tag{2}$$

$$= \frac{B^T X^T Y_k}{N_k} \tag{3}$$

Similarly for covariance matrix $\hat{\Sigma}$:

$$\hat{\Sigma} = \frac{1}{N-K} \sum_{k=1}^{K} \sum_{g_i=k} (\hat{y}_i - \hat{\mu}_k)(\hat{y}_i - \hat{\mu}_k)^T \tag{4}$$

$$= \frac{1}{N-K} \sum_{k=1}^{K} \sum_{g_i=k} (B^T x_i - \hat{\mu}_k)(B^T x_i - \hat{\mu}_k)^T \tag{5}$$

$$= \frac{1}{N-K} \sum_{k=1}^{K} \sum_{g_i=k} (B^T x_i - B^T \mu_k)(B^T x_i - B^T \mu_k)^T \tag{6}$$

$$= B^T \Sigma B \tag{7}$$

I also note that Σ may be written as:

$$\Sigma = \frac{1}{N-K} (X^T - X^T Y D^{-1} Y^T)(X^T - X^T Y D_{-1} Y^T)^T \tag{8}$$

$$= \frac{1}{N-K} X^T (I - Y D^{-1} Y^T)^2 X \tag{9}$$

$$= \frac{1}{N-K} X^T (I - Y D^{-1} Y^T) X \tag{10}$$

$$\tag{11}$$

Here D is the diagonal matrix:

$$D = \begin{bmatrix} n_1 & 0 & \cdots & 0 \\ 0 & n_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & n_k \end{bmatrix} \tag{12}$$

The discriminant function using $\hat{Y}$ may be written as:

$$\delta_k = \hat{Y} \hat{\Sigma}^{-1} \hat{\mu}_k - \frac{1}{2} \hat{\mu}_k^T \hat{\Sigma}^{-1} \hat{\mu}_k + \log \pi_k \tag{13}$$

Eq. 13 has three sub-expressions and I'll consider each of them one by one, to eventually prove that using the discriminant function in Eq. 13 is equivalent to LDA using X.

(a) $\hat{Y}\hat{\Sigma}^{-1}\hat{\mu}_k$

Expanding this sub-expression, we can write:

$$\hat{Y}\hat{\Sigma}^{-1}\hat{\mu}_k = (XB)(B^T\Sigma B)^{-1}\left(\frac{B^TX^TY_k}{N_k}\right) \tag{14}$$

If we write the discriminant function for all classes at once, we could have written:

$$\hat{Y}\hat{\Sigma}^{-1}\hat{\mu} = (XB)(B^T\Sigma B)^{-1}\left(B^TX^TYD^{-1}\right) \tag{15}$$

Now $B^T\Sigma B$ can be written using Eq. 11 and using the expansion for B, as:

$$\begin{aligned}
B^T\Sigma B &= \frac{1}{N-K}B^TX^T(I - YD^{-1}Y^T)XB && (16)\\
&= \frac{1}{N-K}B^TX^TXB - B^TX^TYD^{-1}Y^TXB && (17)\\
&= \frac{1}{N-K}Y^TX(X^TX)^{-1}X^TX(X^TX)^{-1}X^TY - B^TX^TYD^{-1}Y^TXB && (18)\\
&= \frac{1}{N-K}B^TX^TY - B^TX^TYD^{-1}Y^TXB && (19)\\
&= \frac{1}{N-K}(Q - QD^{-1}Q) && (20)
\end{aligned}$$

Here $Q = B^TX^TY$ and we can easily show that $Q = Q^T$. Now consider the expression $B(B^T\Sigma B)^{-1}B^TX^TY$. This can be written as:

$$\begin{aligned}
B(B^T\Sigma B)^{-1}B^TX^TY &= B\left(\frac{1}{N-K}Q(I - D^{-1}Q)\right)^{-1}B^TX^TY && (21)\\
&= \Sigma^{-1}\Sigma(N-K)B(I - D^{-1}Q)^{-1}Q^{-1}B^TX^TY && (22)\\
&= \Sigma^{-1}(N-K)\Sigma B(I - D^{-1}Q)^{-1}Q^{-1}Q && (23)\\
&= \Sigma^{-1}X^T(I - YD^{-1}Y^T)XB(I - D^{-1}Q)^{-1} && (24)\\
&= \Sigma^{-1}(X^TXB - X^TYD^{-1}Y^TXB)(I - D^{-1}Q)^{-1} && (25)\\
&= \Sigma^{-1}(X^TY - X^YD^{-1}Q)(I - D^{-1}Q)^{-1} && (26)\\
&= \Sigma^{-1}X^TY(I - D^{-1}Q)(I - D^{-1}Q)^{-1} && (27)\\
&= \Sigma^{-1}X^TY && (28)\\
& && (29)
\end{aligned}$$

Now using the result from Eq. 29 in Eq. 15, we can write:

$$\begin{aligned}
\hat{Y}\hat{\Sigma}^{-1}\hat{\mu} &= X\Sigma^{-1}X^TYD^{-1} && (30)\\
&= X\Sigma^{-1}\mu && (31)
\end{aligned}$$

Eq. 31 proves that for this first sub-expression of the entire discriminant expression, using LDA on $\hat{Y}$ is equivalent to using LDA on X.

(b) $\hat{\mu}_k^T\hat{\Sigma}^{-1}\hat{\mu}_k$

Now if we are to write the discriminant function for all K classes at once, this expression can be written as $\hat{\mu}_k^T\hat{\Sigma}^{-1}\hat{\mu}$ where $\hat{\mu}$ is a $K \times K$ matrix, the k^{th} column of which represents $\hat{\mu}_k$. So now we can write:

$$\begin{aligned}
\hat{\mu}_k^T\hat{\Sigma}^{-1}\hat{\mu} &= (B^T\mu_k)^T\hat{\Sigma}^{-1}\left(B^TX^TYD^{-1}\right) && (32)\\
&= \mu_k^TB\hat{\Sigma}^{-1}\left(B^TX^TYD^{-1}\right) && (33)\\
&= \mu_k^TB(B^T\Sigma B)^{-1}B^TX^TYD^{-1} && (34)
\end{aligned}$$

Now using the result from Eq. 29 in Eq. 34, we can write:

$$\begin{aligned}
\hat{\mu}_k^T\hat{\Sigma}^{-1}\hat{\mu} &= \mu_k^T\Sigma^{-1}X^TYD^{-1} && (35)\\
&= \mu_k^T\Sigma^{-1}\mu && (36)
\end{aligned}$$

Eq. 36 proves that for the second sub-expression of the entire discriminant function, using LDA on $\hat{Y}$ is equivalent to using LDA on X.

(c) $\log \pi_k$
This is straightforward because the sample probability π_k is not impacted by our transformation of X, and so this term is the same whether we do LDA on X or on $\hat{Y}$.

Hence we have proved that doing LDA on $\hat{Y}$ is equivalent to doing LDA on X.

ESL Problem 4.5
The log-likelihod function can be written as:

$$\ell(\beta) = \sum_{i=1}^{N} \left(y_i \beta^T x_i - \log(1 + \exp \beta^T x_i) \right) \tag{37}$$

Now since the data is separable around x_0, let me define:

$$y_i = \begin{cases} 0 & x_i \leq x_0 \\ 1 & x_i > x_0 \end{cases} \tag{38}$$

Then Eq. 37 can be re-written as:

$$\begin{aligned} \ell(\beta) = & \sum_{\substack{i=1 \\ x_i \leq x_0}}^{N} -\log(1 + \exp(\beta_0 + \beta_1 x_i)) \\ & + \sum_{\substack{i=1 \\ x_i > x_0}}^{N} [\beta_0 + \beta_1 x_i - \log(1 + \exp(\beta_0 + \beta_1 x_i))] \\ = & \sum_{\substack{i=1 \\ x_i \leq x_0}}^{N} -\log(1 + \exp(\beta_0 + \beta_1 x_0 + \beta_1 (x_i - x_0))) \\ & + \sum_{\substack{i=1 \\ x_i > x_0}}^{N} [\beta_0 + \beta_1 x_0 + \beta_1 (x_i - x_0) - \log(1 + \exp(\beta_0 + \beta_1 x_0 + \beta_1 (x_i - x_0)))] \end{aligned} \tag{39}$$

Now we are trying to maximize $\ell(\beta)$. If we let $\beta_0 = -\beta_1 x_0$ and let β_1 go towards positive inifinity, we can show that it will maximize $\ell(\beta)$. Intuitively what this is pointing towards is a decision boundary which is vertical at $x = x_0$.
To look at this more formally, consider $\ell(\beta)$ when we let $\beta_0 = -\beta_1 x_0$:

$$\ell(\beta) = \sum_{\substack{i=1 \\ x_i \leq x_0}}^{N} -\log(1 + \exp(\beta_1 (x_i - x_0))) + \sum_{\substack{i=1 \\ x_i > x_0}}^{N} [\beta_1 (x_i - x_0) - \log(1 + \exp(\beta_1 (x_i - x_0)))] \tag{40}$$

Now as we let $\beta_1 \to \infty$, the first sub-expression in Eq. 40, $[-\log(1 + \exp(\beta_1 (x_i - x_0)))] \to 0$. This is because $-log(1 + z)$ attains its maximum value at $z = 0$, and when we let $\beta_1 \to \infty$, the term $\exp(\beta_1 (x_i - x_0)) \to 0$ when $x_i < x_0$.
As for the second sub-expression, it is of the form $g(z) = z - \log(1 + exp(z))$. The graph of $g(z)$ looks like Fig. and it attains its maximum value at infinity. So here also if we let $\beta_1 \to \infty$, we maximize this sub-expression.

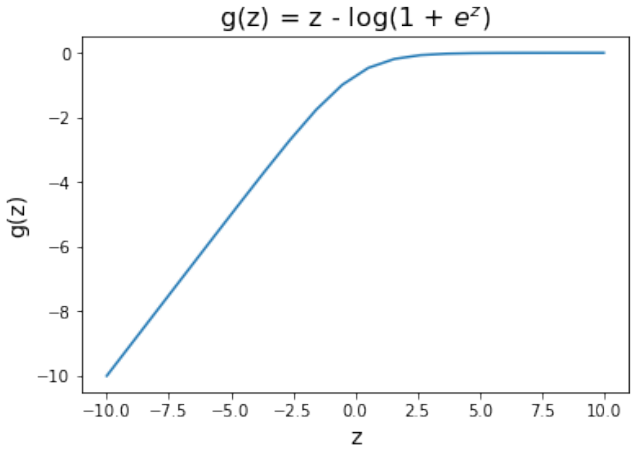

Figure 1: ESL Problem 4.5

Hence, the solution for a single feature separable Logistic Regression can be characterized by $\beta_1 \to \infty$ and $\beta_0 \to -[sign(x_0)]\infty$.

Generalizing this to $\mathbb{R}^p$ for two classes, the two classes will be separable by a single hyperplane, and generalizing it to multiple classes, each pair of classes will be separable by a different hyperplane (so for K classes there will be $K-1$ hyperplanes).

ESL Problem 4.7

When we enforce the constraint $\|\beta\| = 1$, the expression $x_i^T\beta + \beta_0$ equals the *signed* distance of the point x_i from a hyperplane given by $x^T\beta + \beta_0 = 0$. So if try to minimize $D(\beta, \beta_0)$, that optimization will coerce the signed distance of x_i to take a high positive value when $y_i = +1$ and a very negative value when $y_i = -1$, and thus minimizing this criterion will find the separating hyperplane between two classes. However note that unlike the Optimal Separating Hyperplane Criterion, it will not try to maximize distance of each individual point from the separating hyperplane - it is only looking at the sum of distance, whereas the Optimal Separating Hyperplane had the following *pointwise* constraint:

$$y_i(x_i^T\beta + \beta_0) \geq M, i = 1, \cdots, N \tag{41}$$

To illustrate this distinction with help of an example, in Figure 2, the Optimal Separating Hyperplane will be the black dashed line, whereas our criterion will likely give us the red colored line as the separating hyperplane.

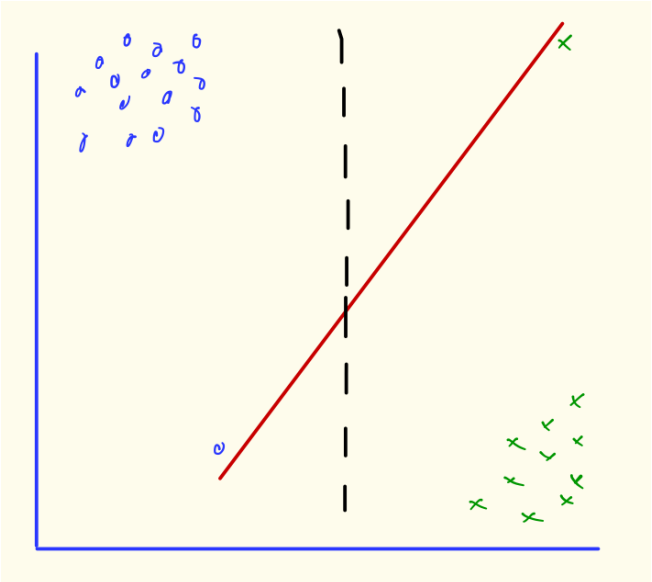

Figure 2: ESL Problem 4.7

ESL Problem 5.1

Let me write $f(x)$ as a linear combination of Basis functions given by Eq. 5.3 of The Elements of Statistical Learning. We have:

$$f(x) = \sum_{m=1}^{6} \beta_m h_m(x) \tag{42}$$

Now I will show that $f(x)$ is continuous at knots ξ_1 and ξ_2 and that its first and second derivatives are also continuous at the knots. Assume that $\xi_1 < \xi_2$. I will show continuity of $f(x)$, its first derivative and its second derivative at $x = \xi_1$, and one can similarly prove these results at $x = \xi_2$. So at $x = \xi_1$:

(a) **Continuity of $f(x)$**
Let $h > 0$. Then:

$$\begin{aligned} f(\xi_1 - h) &= \beta_1 + \beta_2(\xi_1 - h) + \beta_3(\xi_1 - h)^2 + \beta_4(\xi_1 - h)^3 \\ &\quad + \beta_5(\xi_1 - h - \xi_1)_+^3 + \beta_6(\xi_1 - h - \xi_2)_+^3 \\ &= \beta_1 + \beta_2(\xi_1 - h) + \beta_3(\xi_1 - h)^2 + \beta_4(\xi_1 - h)^3 + 0 + 0 \\ \implies lim_{h\to 0} f(\xi_1 - h) &= \beta_1 + \beta_2\xi_1 + \beta_3\xi_1^2 + \beta_4\xi_1^3 \end{aligned} \tag{43}$$

Eq. 43 gives us the left limit at $x = \xi_1$. Now looking at the right limit:

$$\begin{aligned} f(\xi_1 + h) &= \beta_1 + \beta_2(\xi_1 + h) + \beta_3(\xi_1 + h)^2 + \beta_4(\xi_1 + h)^3 \\ &\quad + \beta_5(\xi_1 + h - \xi_1)_+^3 + \beta_6(\xi_1 + h - \xi_2)_+^3 \\ \implies lim_{h\to 0} f(\xi_1 + h) &= \beta_1 + \beta_2\xi_1 + \beta_3\xi_1^2 + \beta_4\xi_1^3 \end{aligned} \tag{44}$$

Eq. 44 gives the right limit at $x = \xi_1$. It follows from the fact that $lim_{h\to 0}\beta_5(h)_+^3 = 0$ and also that $lim_{h\to 0}\beta_6(h + (\xi_1 - \xi 2))_+^3 = 0$ (because $\xi_1 < \xi_2$). **We can observe that the left limit (Eq. 43) and the right limit (Eq. 44) are indeed equal at $x = \xi_1$, and both are also equal to $f(\xi_1)$, which proves that $f(x)$ is continuous at $x = \xi_1$.**

(b) **Continuity of $f'(x)$**
The left first derivative at $x = \xi_1$ is given by:

$$f'_-(\xi_1) = lim_{h\to 0} \frac{f(\xi_1) - f(\xi_1 - h)}{h} \tag{45}$$

Now using the definition of $f(\xi_1 - h)$ from Eq. 43 and using $f(\xi_1) = \beta_1 + \beta_2\xi_1 + \beta_3\xi_1^2 + \beta_4\xi_1^3$, we get the following (here $O(h^2)$ represents terms containing h raised to the power ≥ 2). Taking limit with respect to h, we get the left side derivative:

$$\begin{aligned} f(\xi_1) - f(\xi_1 - h) &= \beta_2 h + 2\beta_3\xi_1 h + 3\beta_4\xi_1^2 h + O(h^2) \\ \implies f'_-(\xi_1) &= \beta_2 + 2\beta_3\xi_1 + 3\beta_4\xi_1^2 \end{aligned} \tag{46}$$

For the right side derivative, we have:

$$f'_+(\xi_1) = lim_{h\to 0} \frac{f(\xi_1 + h) - f(\xi_1)}{h} \tag{47}$$

Similar to the left side derivative, we can derive the right side derivative as follows:

$$\begin{aligned} f(\xi_1 + h) - f(\xi_1) &= \beta_2 h + 2\beta_3\xi_1 h + 3\beta_4\xi_1^2 h \\ &\quad + \beta_5(\xi_1 + h - \xi_1)_+^3 + \beta_6(\xi_1 + h - \xi_2)_+^3 + O(h^2) \\ \implies f'_+(\xi_1) &= \beta_2 + 2\beta_3\xi_1 + 3\beta_4\xi_1^2 \end{aligned} \tag{48}$$

We can observe that the left side first derivative (Eq. 46) and the right side first derivative (Eq. 48) are indeed equal, and both also equal the derivative *at* $x = \xi_1$, which proves that the first derivative is continuous at $x = \xi_1$.

(c) **Continuity of $f''(x)$**
One can show in a manner similar to Part (b) above, that $f''_-(\xi_1) = f''_+(\xi_1) = f''(\xi_1) = 6\beta_4\xi_1^2$.

Hence, we have shown that a function formed using this basis is indeed continuous at the knots and its first and second derivatives are also continuous at the knots.

ESL Problem 5.4

Since Natural Cubic Spline enforce linearity in both boundary regions, the coefficients of x^2 and x^3 in those regions should be zero. I first consider the left boundary region, where $x < \xi_1$. In this region, $(x - \xi_k)_+ = 0$ for $k \in 1, 2, \cdots, K$. Therefore the coefficient of x^2 is β_2 and coefficient of x^3 is β_3. By the constraints imposed by Natural Cubic Spline, we get:

$$\beta_2 = 0 \tag{49}$$

$$\beta_3 = 0 \tag{50}$$

Now looking at the right boundary region, where $x > \xi_1$, we have that $(x - \xi_k)_+ \neq 0$ for $k \in 1, 2, \cdots, K$. In this region the prediction function may be written as:

$$f(x) = \sum_{j=0}^{3} \beta_j x^j + \sum_{k=1}^{K} \theta_k (x - \xi_k)^3 \tag{51}$$

$$= \beta_0 + \beta_1 x + \sum_{k=1}^{K} \theta_k \left(x^3 - \xi_k^3 - 3x^2\xi_k + 3x\xi_k^2\right) \tag{52}$$

Here Eq. 52 follows from the fact that $\beta_2 = \beta_3 = 0$ as shown in Eqs. 49 and 50. Now looking at Eq. 52, the coefficient of x^3 should be zero, and so we get:

$$\sum_{k=1}^{K} \theta_k = 0 \tag{53}$$

And the coefficient of x^2 should also be zero. We get:

$$-\sum_{k=1}^{K} 3\theta_k \xi_k = 0 \tag{54}$$

$$\implies \sum_{k=1}^{K} \theta_k \xi_k = 0 \tag{55}$$

Now I will prove that the basis of a natural cubic spline is given by Eqs. 5.4 and 5.5 from The Elements of Statistical Learning. Starting with a general cubic spline, we can express it in terms of its basis as follows:

$$f(x) = \sum_{j=0}^{3} \beta_j x^j + \sum_{k=1}^{K} \theta_k (x - \xi_k)_+^3 \tag{56}$$

$$= \beta_0 + \beta_1 x + g(x) \tag{57}$$

Here I have used the fact that $\beta_2 = \beta_3 = 0$, and now I will focus on the second expression in Eq. 57, i.e. $g(x)$. Now from Eq.53, we get:

$$\theta_K = -\sum_{k=1}^{K-1} \theta_k \tag{58}$$

Substituting in $g(x)$, we get:

$$g(x) = \sum_{k=1}^{K-1} \theta_k \left((x - \xi_k)_+^3 - (x - \xi_K)_+^3\right) \tag{59}$$

Now from Eq. 55, we can derive:

$$\theta_{K-1} = \sum_{k=1}^{K-2} \frac{\xi_k - \xi_K}{\xi_K - \xi_{K-1}} \tag{60}$$

Substituting in Eq 59, we get:

$$g(x) = \sum_{k=1}^{K-2} \theta_k \left((x - \xi_k)_+^3 - (x - \xi_K)_+^3\right) + \sum_{k=1}^{K-2} \theta_k \frac{\xi_k - \xi_K}{\xi_K - \xi_{K-1}} \left((x - \xi_{K-1})_+^3 - (x - \xi_K)_+^3\right) \tag{61}$$

$$= \sum_{k=1}^{K-2} \theta_k \left(\xi_k - \xi_K\right) \left(d_k(x) - d_{K-1}(x)\right) \tag{62}$$

$$= \sum_{k=1}^{K-2} \phi_k \left(d_k(x) - d_{K-1}(x)\right) \tag{63}$$

Here the terms ϕ_k and $d_k(x)$ in Eq. 63 are written as:

$$\phi_k = \theta_k \left(\xi_k - \xi_K\right) \tag{64}$$

$$d_k(x) = \frac{(x-\xi_k)_+^3 - (x-\xi_K)_+^3}{\xi_K - \xi_k} \tag{65}$$

So now we can rewrite Eq. 57 for a Natural Cubic Spline as:

$$f(x) = \beta_0 + \beta_1 x + \sum_{k=1}^{K-2} \phi_k \left(d_k(x) - d_{K-1}(x)\right) \tag{66}$$

Hence proved that the basis for a Natural Cubic Spline is indeed given by Eqs. 5.4 and 5.5 from The Elements of Statistical Learning.

ESL Problem 5.7

(a) As g is a natural cubic spline, it is linear outside the region bound by knots x_1 and x_N, and its second and higher order derivatives outside this region are 0. Further, since $x = a$ and $x = b$ lie outside the region bounded by (x_1, x_N), we have $g''(a) = g''(b) = 0$. Now using integration by parts, we have:

$$\int_a^b g''(x)h''(x)dx = g''(x)h'(x)\Big|_a^b - \int_a^b g'''(x)h''(x)dx \tag{67}$$

$$= 0 - \int_a^b g'''(x)h''(x)dx \tag{68}$$

$$= -g'''(x)h''(x)\Big|_a^b + \int_a^b g''''(x)h'(x)dx \tag{69}$$

Here, Eq. 67 follows from the fact that $g''(a) = g''(b) = 0$. Eq. 68 is again an application of integration by parts. Now since $g(x)$ is a cubic function, we know that its fourth derivative $g''''(x) = 0$. Therefore the second sub-expression in Eq. 69 is 0. For the first sub-expression we know that $g'''(x) = 0 \forall x \notin (x_1, x_N)$, and therefore we can rewrite the integral evaluation as follows:

$$\int_a^b g''(x)h''(x)dx = -g'''(x)h''(x)\Big|_{x_1}^{x_N} \tag{70}$$

$$= -g'''(x)h''(x)\Big|_{x_1}^{x_2} \cdots - g'''(x)h''(x)\Big|_{x_{N-1}}^{x_N} \tag{71}$$

$$= -\sum_{j=1}^{N-1} \left(g'''(x_{j+1}^-)h''(x_{j+1}) - g'''(x_j^+)h''(x_j)\right) \tag{72}$$

$$= -\sum_{j=1}^{N-1} g'''(x_j^+)\{h''(x_{j+1}) - h''(x_j)\} \tag{73}$$

$$= 0 \tag{74}$$

Here Eq. 73 follows from the fact that *inside* the region between two knots, the third derivative of $g(x)$ is the same everywhere. And the final step Eq. 74 follows from the fact that $h(x) = 0$ at all the knots $x_1, \cdots, x_N$ (because both g and $\tilde{g}$ interpolate the N pairs.)

(b) Using Part (a), we can write:

$$\int_a^b g''(x)h''(x)dx = 0 \tag{75}$$

$$\implies \int_a^b g''(x)\left(\tilde{g}''(x) - g''(x)\right)dx = 0 \tag{76}$$

$$\implies \int_a^b g''(x)^2 = \int_a^b g''(x)\tilde{g}''(x) \tag{77}$$

$$\implies \int_a^b g''(x)^2 \leq \left(\int_a^b g''(x)^2\right)^{1/2} \left(\int_a^b \tilde{g}''(x)^2\right)^{1/2} \tag{78}$$

Here Eq. 78 follows from the Cauchy-Schwarz inequality. Now if $g''(x)$ is not zero everywhere in the interval (a, b), then we can conclude that:

$$\int_a^b g''(x)^2 \leq \int_a^b \tilde{g}''(x)^2 \tag{79}$$

Now equality holds for Eq. 79 when $g''(x) = \tilde{g}''(x) \forall x \in (a, b)$. This implies that $h''(x) = 0 \forall x \in (a, b)$ or in other words $h(x)$ is linear in that interval. However, we know that $h(x) = 0$ at all the knots. This implies that for equality to hold in Eq. 79, $h(x) = 0 \forall x \in (a, b)$.

(c) Let $\tilde{g}(x)$ be a function which minimizes this objective function. Let $g(x)$ be a natural cubic spline with knots at $x_1, \cdots, x_N$ and which satisfies $g(x_i) = \tilde{g}(x_i) \forall i \in 1, \cdots, N$. Then the least squared part of the objective function is the same for $g(x)$ and $\tilde{g}(x)$. Now from Part (b), we know that:

$$\int_a^b g''(x)^2 \leq \int_a^b \tilde{g}''(x)^2 \tag{80}$$

$$\implies \lambda \int_a^b g''(x)^2 \leq \lambda \int_a^b \tilde{g}''(x)^2 \tag{81}$$

Here Eq. 81 is true $\forall \lambda > 0$. So now we've found a natural cubic spline function $g(x)$ which optimizes our objective better than $\tilde{g}(x)$. But we know that $\tilde{g}(x)$ is a minimizer, hence it must also be a natural cubic spline.