Junda An

Problem Set 3

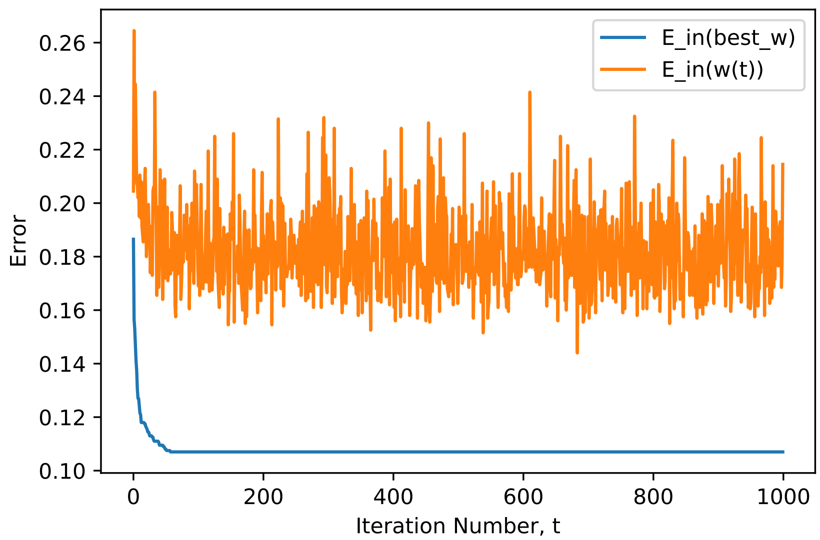
CSCI 420
April 9, 2019

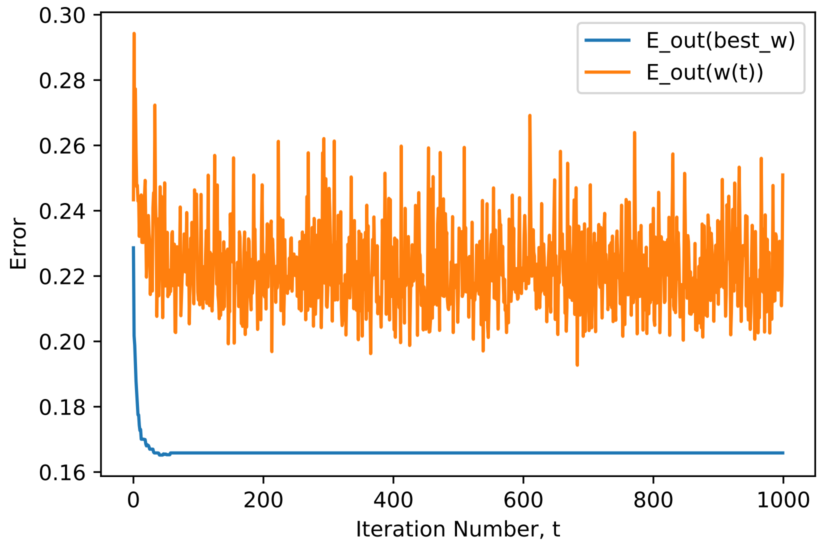
AML EX:3.2; 4.[3,9]

Answer to the problem goes here.

1. EX 3.2 answer here.

   Please see the jupyter notebook named Ex3.2.





2. Ex 4.3 answer here.

   If H is fixed and we increase the complexity of f, the part of the target function 'outside' of the best hypothesis in H will be larger. Therefore, it will be more difficult for our model to fit the target function. Therefore, the deterministic error will go up. According to the bias-variance decomposition in Section 2.3.1, $E_D[E_{out}] = \sigma^2 + bias + var$. In this case, *bias* that is directly influenced by the deterministic error and *var* which is indirectly influenced by the deterministic error will both go up. Therefore, there is more overfitting.

If f is fixed and we decrease the complexity of H, the part of the target function 'outside' of the best hypothesis in H will be larger. Therefore, it will be more difficult for out model to fit the target function. Thus, the deterministic error will go up. According to the bias-variance decomposition in Section 2.3.1, $E_D[E_{out}] = \sigma^2 + bias + var$. In this case, *bias* that is directly influenced by the deterministic error will go up, making the model more likely to overfit. On the other hand, *var* will go down since H becomes less complex, making the model less likely to overfit. In this case, the model is simple. Even if the E_{out} will increase, this increment is due to underfitting. Therefore, *var* generally has greater effect in determining overfitting than *bias*. So, there is less overfitting.

3. Ex 4.9 answer here.

   When K increases, we are using less data for training. Therefore, it is harder for the final hypothesis that we choose to fit the target function. Therefore, E_{out} is increasing. Since E_{val} is an estimate for E_{out}, it is also increasing with K.

   On the other hand, when K increases, we are using more data for validation. The estimate is becoming more reliable. Therefore, E_{val} is getting closer to E_{out} with K increasing.

PRML 1.[2,3,39]; 2.[12]; 3.[4,11]

1. EX 1.2 answer here.

   To find coefficients $\{w_i\}$ that minimizes the error function, we need to take derivative of the error function with respect to each w_i and find w_i that makes the derivative zero.

$$\frac{\partial \tilde{E}(w)}{\partial w_i} = 0$$

$$\Rightarrow \frac{\partial \frac{1}{2}\sum_{n=1}^{N}(y(x_n, w) - t_n)^2 + \frac{\lambda}{2}||w||^2}{\partial w_i} = 0$$

$$\Rightarrow \frac{\partial \frac{1}{2}\sum_{n=1}^{N}(\sum_{j=0}^{M} w_j x_n^j - t_n)^2 + \frac{\lambda}{2}||w||^2}{\partial w_i} = 0 \; (substitute\; y(x_n, w) = \sum_{j=0}^{M} w_j x_n^j \; into\, the\, equation)$$

$$\Rightarrow \frac{\partial \frac{1}{2}\sum_{n=1}^{N}(\sum_{j=0}^{M} w_j x_n^j - t_n)^2 + \frac{\lambda}{2}(w_0^2 + w_1^2 + ... + w_M^2)}{\partial w_i} = 0$$

$$(substitute\ ||w||^2 = w_0^2 + w_1^2 + ... + w_M^2\ into\ the\ equation)$$

$$\Rightarrow \frac{1}{2}\sum_{n=1}^{N} 2(\sum_{j=0}^{M} w_j x_n^j - t_n)x_n^i + \frac{\lambda}{2} 2w_i = 0\ (chain\ rule)$$

$$(since\ w_i\ is\ the\ only\ term\ in\ (w_1, w_2, ..., w_M)\ that\ has\ a\ relationship\ with\ w_i)$$

$$\Rightarrow \sum_{n=1}^{N}(\sum_{j=0}^{M} w_j x_n^j - t_n)x_n^i + \lambda w_i = 0$$

$$\Rightarrow \sum_{n=1}^{N}(\sum_{j=0}^{M} x_n^{i+j} w_j - x_n^i t_n) + \lambda w_i = 0\ (since\ x_n^i\ does\ not\ depend\ on\ j)$$

$$\Rightarrow \sum_{n=1}^{N}(\sum_{j=0}^{M} x_n^{i+j} w_j) + \lambda w_i = \sum_{n=1}^{N} x_n^i t_n$$

$$\Rightarrow \sum_{j=0}^{M}(\sum_{n=1}^{N} x_n^{i+j} w_j) + \lambda w_i = \sum_{n=1}^{N} x_n^i t_n\ (exchange\ the\ summation\ symbol)$$

$$\Rightarrow \sum_{j=0}^{M}(\sum_{n=1}^{N} x_n^{i+j}) w_j + \lambda w_i = \sum_{n=1}^{N} x_n^i t_n\ (since\ w_j\ is\ independent\ of\ n)$$

$$\Rightarrow \sum_{j=0}^{M}(\sum_{n=1}^{N} x_n^{i+j} + \lambda I_{ij}) w_j = \sum_{n=1}^{N} x_n^i t_n\ (let\ I_{ij} = 1,\ if\ i = j.\ Otherwise,\ I_{ij} = 0.)$$

$$(since\ we\ can\ only\ combine \sum_{j=0}^{M}(\sum_{n=1}^{N} x_n^{i+j}) w_j\ and\ \lambda w_i\ when\ i = j)$$

Like what I did in EX 1.1, let $A_{ij} = \sum_{n=1}^{N} x_n^{i+j}$, and $T_i = \sum_{n=1}^{N} x_n^i t_n$. Therefore, $\sum_{j=0}^{M}(A_{ij} + \lambda I_{ij}) w_j = T_i$

2. Ex 1.3 answer here

   Let $p(F = a|B = r)$ denotes the probability of selecting an apple from the red box. Let $p(F = a|B = b)$ denotes the probability of selecting an apple from the blue box. Let $p(F = a|B = g)$ denotes the probability of selecting an apple from the green box. Since probability of selecting any of the items in the box is equal, $p(F = a|B = r) = \frac{number\ of\ Apple\ in\ red\ box}{total\ number\ of\ fruit\ in\ red\ box} = \frac{3}{3+4+3} = \frac{3}{10} = 0.3$. Similarly, $p(F = a|B = b) = \frac{1}{1+1+0} = 0.5$ and $p(F = a|B = g) = \frac{3}{3+3+4} = 0.3$

   Therefore, $p(a) = p(r)p(F = a|B = r) + p(b)p(F = a|B = b) + p(g)p(F = a|B = g) = 0.2 * 0.3 + 0.2 * 0.5 + 0.6 * 0.3 = 0.34$ (according to Bayes' theorem).

   $p(F = o|B = r) = \frac{4}{10} = 0.4$. $p(F = o|B = b) = \frac{1}{2} = 0.5$. $p(F = o|B = g) = \frac{3}{10} = 0.3$. $p(o) = p(r)p(F = o|B = r) + p(b)p(F = o|B = b) + p(g)p(F = o|B = g) = 0.2 * 0.4 + 0.2 * 0.5 + 0.6 * 0.3 = 0.36$ Thus, according to Bayes' theorem, $p(B = g|F = o) = \frac{p(F=o|B=g)p(g)}{p(o)} = \frac{0.3*0.6}{0.36} = 0.5$

3. Ex 1.39 answer here.

   According to the Table 1.3, $p(x = 0, y = 0) = \frac{1}{3}$, $p(x = 1, y = 0) = 0$, $p(x = 0, y = 1) = \frac{1}{3}$, and $p(x = 1, y = 1) = \frac{1}{3}$. Therefore, $p(x = 0) = \frac{1}{3} + \frac{1}{3} = \frac{2}{3}$, and $p(x = 1) = \frac{1}{3}$. $p(y = 0) = \frac{1}{3}$ and $p(y = 1) = \frac{1}{3} + \frac{1}{3} = \frac{2}{3}$. $p(x = 0|y = 0) = \frac{p(x=0,y=0)}{p(y=0)} = \frac{\frac{1}{3}}{\frac{1}{3}} = 1$. $p(x = 1|y = 0) = \frac{p(x=1,y=0)}{p(y=0)} = 0$. $p(x = 0|y = 1) = \frac{p(x=0,y=1)}{p(y=1)} = \frac{1}{2}$. $p(x = 1|y = 1) = \frac{p(x=1,y=1)}{p(y=1)} = \frac{1}{2}$. $p(y = 0|x = 0) = \frac{p(x=0,y=0)}{p(x=0)} = \frac{1}{2}$. $p(y = 1|x = 0) = \frac{p(x=0,y=1)}{p(x=0)} = \frac{1}{2}$. $p(y = 0|x = 1) = \frac{p(x=1,y=0)}{p(x=1)} = 0$. $p(y = 1|x = 1) = \frac{p(x=1,y=1)}{p(x=1)} = 1$.

   (a) $H(x) = -\sum_i p(x_i)ln(p(x_i))$ $(definition\ of\ entropy)$ $= -(p(x = 0)ln(p(x = 0)) + p(x = 1)ln(p(x = 1))) = -(\frac{2}{3}ln(\frac{2}{3}) + \frac{1}{3}ln(\frac{1}{3})) = 0.637$

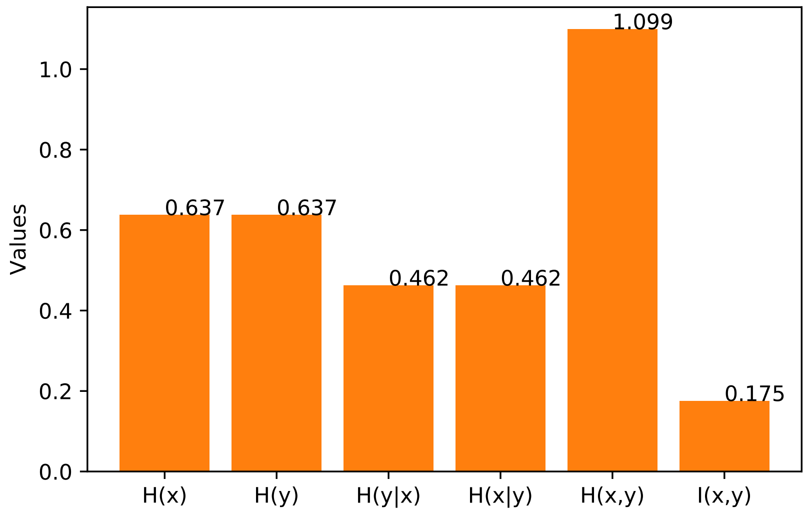
   (b) $H(y) = -\sum_i p(y_i)ln(p(y_i))$ $(definition\ of\ entropy)$ $= -(p(y = 0)ln(p(y = 0)) + p(y = 1)ln(p(y = 1))) = \frac{1}{3}ln(\frac{1}{3}) + \frac{2}{3}ln(\frac{2}{3})) = 0.637$

   (c) $H(y|x) = -\sum_i \sum_j p(x_i, y_j)ln(p(x_i|y_j))$ $(definition\ of\ entropy)$ $= -(p(x = 0, y = 0)ln(p(x = 0|y = 0)) + p(x = 1, y = 0)ln(p(x = 1|y = 0)) + p(x = 0, y = 1)ln(p(x = 0|y = 1))p(x = 1, y = 1)ln(p(x = 1|y = 1))) = -(\frac{1}{3}ln(1) + 0 + \frac{1}{3}ln(\frac{1}{2}) + \frac{1}{3}ln(\frac{1}{2})) = 0.462$

   (d) $H(x|y) = -\sum_i \sum_j p(y_i, x_j)ln(p(y_i|x_j))$ $(definition\ of\ entropy)$ $= -(p(y = 0, x = 0)ln(p(y = 0|x = 0)) + p(y = 1, x = 0)ln(p(y = 1|x = 0)) + p(y = 0, x = 1)ln(p(y = 0|x = 1))p(y = 1, x = 1)ln(p(y = 1|x = 1))) = -(\frac{1}{3}ln(\frac{1}{2}) + \frac{1}{3}ln(\frac{1}{2}) + 0 + \frac{1}{3}ln(1)) = 0.462$

   (e) $H(x, y) = -\sum_i \sum_j p(x_i, y_j)ln(p(x_i, y_j))$ $(definition\ of\ entropy)$ $= -(p(x = 0, y = 0)ln(p(x = 0, y = 0)) + p(x = 1, y = 0)ln(p(x = 1, y = 0)) + p(x = 0, y = 1)ln(p(x = 0, y = 1))p(x = 1, y = 1)ln(p(x = 1, y = 1))) = -(\frac{1}{3}ln(\frac{1}{3}) + 0 + \frac{1}{3}ln(\frac{1}{3}) + \frac{1}{3}ln(\frac{1}{3})) = 1.099$

(f) $I(x, y) = H(x) - H(x|y)$ (*according to the equation given in the hint*) $= 0.637 - 0.462 = 0.175$



From the diagram, we can see $H(x) = H(y)$, $H(x|y) = H(y|x)$, and $I(x, y) = H(x) - H(x|y) = H(y) - H(y|x)$

4. Ex 2.12 answer here

   $\int_a^b U(x|a, b)dx = \int_a^b \frac{1}{b-a}dx$ (*substitute* $U(x|a, b) = \frac{1}{b-a}$) $= \frac{1}{b-a}x|_a^b = \frac{b}{b-a} - \frac{a}{b-a} = 1$. Therefore, this distribution is normalized.

   According to Eq(1.34), $mean = E(x) = \int_a^b U(x|a, b)xdx$ (*substitute* $U(x|a, b) = \frac{1}{b-a}$) $= \int_a^b \frac{x}{b-a}dx = \frac{1}{2(b-a)}x^2|_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{(b+a)(b-a)}{2(b-a)} = \frac{b+a}{2}$.

   $E(x^2) = \int_a^b U(x|a, b)x^2dx$ (*substitute* $U(x|a, b) = \frac{1}{b-a}$) $= \int_a^b \frac{x^2}{b-a}dx = \frac{x^3}{3(b-a)}|_a^b = \frac{b^3-a^3}{3(b-a)} = \frac{(b-a)(b^2+ab+a^2)}{3(b-a)} = \frac{b^2+ab+a^2}{3}$. Then, according to Eq(1.40), $var(x) = E(x^2) - E(x)^2 = \frac{b^2+ab+a^2}{3} - (\frac{b+a}{2})^2 = \frac{a^2+b^2-2ab}{12} = \frac{(b-a)^2}{12}$.

5. Ex 3.4 answer here

   Let $\tilde{y_n}$ denotes the output produced by the input, x_n with noise. Let y_n denotes the output produced by the input, x_n without noise. Therefore, according to Eq (3.105), $y(x_n, w) = y_n = w_0 + \sum_{i=1}^D w_i x_{ni}$. Similarly, $\tilde{y}(x_n, w) = \tilde{y_n} = w_0 + \sum_{i=1}^D w_i(x_{ni} + \epsilon_{ni}) = w_0 + \sum_{i=1}^D (w_i x_{ni} + w_i\epsilon_{ni}) = w_0 + \sum_{i=1}^D w_i x_{ni} + \sum_{i=1}^D w_i\epsilon_{ni} = y_n + \sum_{i=1}^D w_i\epsilon_{ni}$ (substitute $y_n = w_0 + \sum_{i=1}^D w_i x_{ni}$), where ϵ_{ni} $N(0, \sigma^2)$.

   According to Eq. (3.106), $E_D(w) = \frac{1}{2}\sum_{n=1}^N (y_n - t_n)^2 = \frac{1}{2}\sum_{n=1}^N (y_n^2 - 2y_nt_n + t_n^2)$

$$\begin{aligned}
\tilde{E}_D(w) &= \frac{1}{2}\sum_{n=1}^{N}(\tilde{y}(x_n,w)-t_n)^2 \ (According\ to\ Eq.(3.106))\\
&= \frac{1}{2}\sum_{n=1}^{N}(\tilde{y}_n^2-2\tilde{y}_nt_n+t_n^2)\\
&= \frac{1}{2}\sum_{n=1}^{N}((y_n+\sum_{i=1}^{D}w_i\epsilon_{ni})^2-2(y_n+\sum_{i=1}^{D}w_i\epsilon_{ni})t_n+t_n^2) \ (substitute\ \tilde{y}_n=y_n+\sum_{i=1}^{D}w_i\epsilon_{ni})\\
&= \frac{1}{2}\sum_{n=1}^{N}(y_n^2+2y_n\sum_{i=1}^{D}w_i\epsilon_{ni}+(\sum_{i=1}^{D}w_i\epsilon_{ni})^2-2y_nt_n-2t_n\sum_{i=1}^{D}w_i\epsilon_{ni}+t_n^2)\\
&= \frac{1}{2}\sum_{n=1}^{N}(y_n^2-2y_nt_n+t_n^2+2(y_n-t_n)\sum_{i=1}^{D}w_i\epsilon_{ni}+(\sum_{i=1}^{D}w_i\epsilon_{ni})^2) \ (rearrange\ terms)\\
&= E_D(w)+\frac{1}{2}\sum_{n=1}^{N}(2(y_n-t_n)\sum_{i=1}^{D}w_i\epsilon_{ni}+(\sum_{i=1}^{D}w_i\epsilon_{ni})^2)\\
&(substitute\ E_D(w)=\frac{1}{2}\sum_{n=1}^{N}(y_n^2-2y_nt_n+t_n^2))
\end{aligned}$$

Then, the error averaged over the noise distribution is: $E_\epsilon(\tilde{E}_D(w)) = E_\epsilon(E_D(w) + \frac{1}{2}\sum_{n=1}^{N}(2(y_n-t_n)\sum_{i=1}^{D}w_i\epsilon_{ni}+(\sum_{i=1}^{D}w_i\epsilon_{ni})^2)) = E_D(w)+\frac{1}{2}\sum_{n=1}^{N}2(y_n-t_n)E_\epsilon(\sum_{i=1}^{D}w_i\epsilon_{ni})+\frac{1}{2}\sum_{n=1}^{N}E_\epsilon((\sum_{i=1}^{D}w_i\epsilon_{ni})^2))$ (since n is independent of ϵ).

$E_\epsilon(\sum_{i=1}^{D}w_i\epsilon_{ni}) = \sum_{i=1}^{D}w_iE_\epsilon(\epsilon_{ni})\ (since\ i\ is\ independent\ of\ \epsilon) = 0$ (since $E(\epsilon_i)=0$)

$$\begin{aligned}
E_\epsilon((\sum_{i=1}^{D}w_i\epsilon_{ni})^2) &= E_\epsilon(\sum_{i=1}^{D}\sum_{j=1}^{D}w_iw_j\epsilon_{ni}\epsilon_{nj}) \ (expanding\ the\ square)\\
&= \sum_{i=1}^{D}\sum_{j=1}^{D}w_iw_jE_\epsilon(\epsilon_{ni}\epsilon_{nj}) \ (since\ w,\ i,\ j\ are\ independent\ of \epsilon)\\
&= \sum_{i=1}^{D}\sum_{j=1}^{D}w_iw_j\delta_{ij}\sigma^2 \ (since\ E_\epsilon(\epsilon_i\epsilon_j)=\delta_{ij}\sigma^2)\\
&= \sigma^2\sum_{i=1}^{D}\sum_{j=1}^{D}w_iw_j\delta_{ij} \ (since\ \sigma\ is\ independent\ of\ i,j)\\
&= \sigma^2\sum_{i=1}^{D}w_i^2 \ (definition\ of\ \delta_{ij})
\end{aligned}$$

According to the calculation above, we can get $E_\epsilon(\tilde{E}_D(w)) = E_D(w) + \frac{1}{2}\sum_{n=1}^{N}(0 + \sigma^2\sum_{i=1}^{D} w_i^2) = E_D(w) + \frac{\sigma^2}{2}\sum_{i=1}^{D} w_i^2$

6. Ex 3.11 answer here

According to Eq. (3.59), $\sigma_N^2(x) = \frac{1}{\beta} + \phi^T(x)S_N\phi(x)$. Therefore, $\sigma_{N+1}^2(x) = \frac{1}{\beta} + \phi^T(x)S_{N+1}\phi(x)$.

Then, $\sigma_N^2(x) - \sigma_{N+1}^2(x) = \frac{1}{\beta} + \phi^T(x)S_N\phi(x) - (\frac{1}{\beta} + \phi^T(x)S_{N+1}\phi(x)) = \phi^T(x)S_N\phi(x) - \phi^T(x)S_{N+1}\phi(x) = \phi^T(x)(S_N - S_{N+1})\phi(x)$

According to hint given by Professor, $S_{N+1}^{-1} = S_N^{-1} + \beta\phi_{N+1}\phi_{N+1}^T$. Therefore,

$$\begin{aligned}
S_{N+1} &= (S_N^{-1})^{-1} \\
&= (S_N^{-1} + \beta\phi_{N+1}\phi_{N+1}^T)^{-1}\ (substitute\ S_{N+1}^{-1} = S_N^{-1} + \beta\phi_{N+1}\phi_{N+1}^T) \\
&= (S_N^{-1} + \sqrt{\beta}\phi_{N+1}\sqrt{\beta}\phi_{N+1}^T)^{-1} \\
&= (S_N^{-1})^{-1} - \frac{((S_N^{-1})^{-1}\sqrt{\beta}\phi_{N+1})(\sqrt{\beta}\phi_{N+1}^T(S_N^{-1})^{-1})}{1 + \sqrt{\beta}\phi_{N+1}^T(S_N^{-1})^{-1}\sqrt{\beta}\phi_{N+1}}\ (According\ to\ Eq.\ (3.110)) \\
&= S_N - \frac{\beta(S_N\phi_{N+1})(\phi_{N+1}^T S_N)}{1 + \beta\phi_{N+1}^T S_N\phi_{N+1}}
\end{aligned}$$

Therefore, according to the calculation above,

$$\begin{aligned}
\sigma_N^2(x) - \sigma_{N+1}^2(x) &= \phi^T(x)(S_N - S_{N+1})\phi(x) \\
&= \phi^T(x)(S_N - (S_N - \frac{\beta(S_N\phi_{N+1}(x))(\phi_{N+1}^T(x)S_N)}{1 + \beta\phi^T(x)_{N+1}S_N\phi_{N+1}(x)}))\phi(x) \\
&(plug\ in\ S_N = S_N - \frac{\beta(S_N\phi_{N+1})(\phi_{N+1}^T S_N)}{1 + \beta\phi_{N+1}^T S_N\phi_{N+1}}) \\
&= \phi^T(x)\frac{\beta(S_N\phi_{N+1}(x))(\phi_{N+1}^T(x)S_N)}{1 + \beta\phi^T(x)_{N+1}S_N\phi_{N+1}(x)}\phi(x) \\
&= \frac{(\phi^T(x)S_N\phi_{N+1}(x))(\phi_{N+1}^T(x)S_N\phi(x))}{1/\beta + \phi^T(x)_{N+1}S_N\phi_{N+1}(x)}\ (associative\ rule) \\
&= \frac{(\phi^T(x)S_N\phi_{N+1}(x))(\phi_N^T(x)S_N\phi_{N+1}(x))^T}{1/\beta + \phi^T(x)_{N+1}S_N\phi_{N+1}(x)}
\end{aligned}$$

Since S_N is positive definite, $\phi^T(x)_{N+1}S_N\phi_{N+1}(x) > 0$. Since $ww^T \geq 0$ for every vector w, $(\phi^T(x)S_N\phi_{N+1}(x))(\phi_N^T(x)S_N\phi_{N+1}(x))^T \geq 0$. Since $\beta > 0$, $\sigma_N^2(x) - \sigma_{N+1}^2(x) = \frac{(\phi^T(x)S_N\phi_{N+1}(x))(\phi_N^T(x)S_N\phi_{N+1}(x))^T}{1/\beta + \phi^T(x)_{N+1}S_N\phi_{N+1}(x)} \geq 0$. Therefore, $\sigma_N^2(x) \geq \sigma_{N+1}^2(x)$.