CSE 847 Home Assignment 2

Submitted by: Ritam Guha (MSU ID: guharita)

Date: February 28, 2021

1 Linear Algebra II

1. (20 points) Compute (by hand) the eigenvalues and the eigenvectors of the following matrix:

$$A = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Reponse:

In order to find the eigen values, we need to find the solutions to the following equation:

$$|A - \lambda \mathbb{I}| = 0$$

$$\begin{vmatrix} 2-\lambda & 1 & 0 \\ 1 & 2-\lambda & 0 \\ 0 & 0 & 1-\lambda \end{vmatrix} = 0$$

$\Rightarrow (2-\lambda)(2-\lambda)(1-\lambda) - 1(1-\lambda) = 0$

$\Rightarrow (1-\lambda)(\lambda^2 - 4\lambda + 4 - 1)0$

$\Rightarrow (1-\lambda)(\lambda^2 - 4\lambda + 3) = 0$

$\Rightarrow (1-\lambda)(\lambda - 1)(\lambda - 3) = 0$

$\Rightarrow (\lambda - 1)^2(\lambda - 3) = 0$

This equation has one root at $\lambda = 3$ and two roots at $\lambda = 1$.

The corresponding eigen vectors can be found using the following equation:

$$Av = \lambda v$$

We can represent v as a general vector of the form: $\begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}$

So, Av becomes: $\begin{bmatrix} 2v_1 + v_2 \\ v_1 + 2v_2 \\ v_3 \end{bmatrix}$

<u>Now for $\lambda = 3$</u>,

$2v_1 + v_2 = 3v_1 \Rightarrow v_1 = v_2$

$v_1 + 2v_2 = 3v_2 \Rightarrow v_1 = v_2$

$v_3 = 3v_3 \Rightarrow v_3 = 0$

So, the eigen vector becomes any vector of the form: $\begin{bmatrix} x \\ x \\ 0 \end{bmatrix}$ where x is an arbitrary constant.

<u>Now for $\lambda = 1$</u>,

$2v_1 + v_2 = v_1 \Rightarrow v_1 = -v_2$

$v_1 + 2v_2 = v_2 \Rightarrow v_1 = -v_2$

$v_3 = v_3$

So, the eigen vector becomes any vector of the form: $\begin{bmatrix} x \\ -x \\ y \end{bmatrix}$ where x and y are arbitrary constants.

2. Given the three vectors $v_1 = (2, 0, -1), v_2 = (0, -1, 0)$ and $v_3 = (2, 0, 4)$ in $\mathbb{R}^3$.

   - (10 points) Show that they form an orthogonal set under the standard Euclidean inner product for $\mathbb{R}^3$ but not an orthonormal set.

Response:

In order to show that the given vectors form an orthogonal set under the standard Euclidean inner product for $\mathbb{R}^3$, we need to show that:

$$v_1^T v_2 = 0$$
$$v_2^T v_3 = 0$$
$$v_3^T v_1 = 0$$

$\underline{v_1^T v_2}$: $\begin{bmatrix} 2 & 0 & -1 \end{bmatrix} \begin{bmatrix} 0 \\ -1 \\ 0 \end{bmatrix} = 0$

$\underline{v_2^T v_3}$: $\begin{bmatrix} 0 & -1 & 0 \end{bmatrix} \begin{bmatrix} 2 \\ 0 \\ 4 \end{bmatrix} = 0$

$\underline{v_3^T v_1}$: $\begin{bmatrix} 2 & 0 & 4 \end{bmatrix} \begin{bmatrix} 2 \\ 0 \\ -1 \end{bmatrix} = 0$

So, from the previous computations, we can see that the given vectors form an orthogonal set under inner product in $\mathbb{R}^3$.

But in order to form an orthonormal set, $l2$-norm of each of the vectors should be equal to 1. If we calculate the $l2$-norm values of the vectors, we will see:

$\|v_1\|_2 = \sqrt{5}$

$\|v_2\|_2 = 1$

$\|v_3\|_2 = \sqrt{20} = 2\sqrt{5}$

The $l2$-norm values of v_1 and v_3 are not equal to 1. So, although they form an orthogonal set, they are not orthonormal.

- (10 points) Turn them into a set of vectors that will form an orthonormal set of vectors under the standard Euclidean inner product for $\mathbb{R}^3$.

Response:

In order to convert the vectors to form an orthonormal set, the vectors should be normalized. So, the new set of vectors will be:

$v_{1n} = \frac{v_1}{\|v_1\|_2}$; $v_{2n} = \frac{v_2}{\|v_2\|_2}$; $v_{3n} = \frac{v_3}{\|v_3\|_2}$

$v_{1n} = (\frac{2}{\sqrt{5}}, 0, \frac{-1}{\sqrt{5}})$; $v_{2n} = (0, -1, 0)$; $v_{3n} = (\frac{1}{\sqrt{5}}, 0, \frac{2}{\sqrt{5}})$.

3. (10 points) Suppose that A is an $n \times m$ matrix with linearly independent columns. Show that $A^T A$ is an invertible matrix.

   Response:

   A is an $n \times m$ matrix with linearly independent columns. So, A^T is an $m \times n$ matrix with linearly independent rows. The multiplication of A^T and A provides a square matrix of size $m \times m$ with m independent columns and rows which means that the resultant matrix is a full-rank matrix. A full-rank square matrix is non-singular. Therefore, the final resultant matrix is an invertible matrix.

4. (10 points) Suppose that A is an $n \times m$ matrix with linearly independent columns. Let $\bar{x}$ be a least squares solution to the system of equations $Ax = b$ (the solution of $\min_x \|Ax - b\|_2^2$). Show that $\bar{x}$ is the **unique** solution to the associated normal system $A^T A\bar{x} = A^T b$.

   Response: A is an $n \times m$ matrix with linearly independent columns. So, A is a full-rank matrix with linearly independent columns (Here we have assumed that $m < n$). So, $A^T A$ is an $m \times m$ matrix whose rank is m and hence it is invertible.

   Suppose there exists another solution to system of normal equations apart from $\bar{x}$ (say y). Satisfying the system of equations with y, we get:

   $A^T Ay = A^T b$

   $\Rightarrow y = (A^T A)^{-1} A^T b$ (As we know that $A^T A$ is an invertible matrix)

   $\Rightarrow y = (A^T A)^{-1} A^T A\bar{x}$ (from the solution of $Ax = b$)

   $\Rightarrow y = (A^T A)^{-1}(A^T A)\bar{x}$

$\Rightarrow y = \mathbb{I}\bar{x}$

$\Rightarrow y = \bar{x}$

So, from these steps, we can see that if there exists any other solution for $A^T Ax = A^T b$, it should be equal to $\bar{x}$. So, $\bar{x}$ is a unique solution for the associated normal system.

2 Linear Regression I

Questions in the textbook Pattern Recognition and Machine Learning:

1. (10 points) Page 174, Question 3.2

> Show that the matrix
>
> $$\mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T$$
>
> takes any vector v and projects it onto the space spanned by the columns of $\mathbf{\Phi}$. Use this result to show that the least-squares solution (3.15):
>
> $$\mathbf{w}_{\mathrm{ML}} = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{t} \tag{3.15}$$
>
> corresponds to an orthogonal projection of the vector $\mathbf{t}$ onto the manifold $\mathcal{S}$ as shown in Figure 3.2.

Response:

The given matrix is: $P = \mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T$

Any vector v can be represented as combination two components: one along the space spanned by the columns of $\mathbf{\Phi}$ (v_1) and another one orthogonal to it (v_2) as:

$v = v_1 + v_2$

v_1 can be represented as $\mathbf{\Phi}q$ as it is a linear combination of the columns of $\mathbf{\Phi}$.

$Pv = \mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T v_1 + \mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T v_2$

$= \mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{\Phi}q + \mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}(\mathbf{\Phi}^T v_2)$

$= \mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}(\mathbf{\Phi}^T\mathbf{\Phi})q + \mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}(0)$ [as v_2 is orthogonal to the space spanned by the columns of $\mathbf{\Phi}$]

$= \mathbf{\Phi} q$

Thus we can see that the inner product of P and v gives us only that component of v which is along the space spanned by the columns of $\mathbf{\Phi}$. It is the definition of projection. So, P is the projection matrix. [Proved - 1]

From the least-square solution, we get: $w = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T t$. So, the predictions then become: $\mathbf{\Phi} w = \mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T t$ which is basically projection t onto the space spanned by the columns of $\mathbf{\Phi}$. So, $\mathbf{\Phi} w = Pt$. We have to prove that this projection is an orthogonal projection onto the manifold S. Any vector in the manifold can be represented as a linear combination of the columns of $\mathbf{\Phi}$ as $\mathbf{\Phi} q$. The difference between the vector t and its projection Pt is: $(t - Pt)$. Mathematically we need to show: $(t - Pt)^T\mathbf{\Phi} q = 0$.

Proof:

$(t - Pt)^T\mathbf{\Phi} q$

$= (t^T - t^T P^T)\mathbf{\Phi} q$

$= t^T\mathbf{\Phi} q - t^T P^T\mathbf{\Phi} q$

$= t^T\mathbf{\Phi} q - t^T P^T\mathbf{\Phi} q$

$= t^T\mathbf{\Phi} q - t^T(\mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T)^T\mathbf{\Phi} q$ [substituting P]

$= t^T\mathbf{\Phi} q - t^T(\mathbf{\Phi}((\mathbf{\Phi}^T\mathbf{\Phi})^{-1})^T\mathbf{\Phi}^T)\mathbf{\Phi} q$

$= t^T\mathbf{\Phi} q - t^T(\mathbf{\Phi}((\mathbf{\Phi}^T\mathbf{\Phi})^T)^{-1}\mathbf{\Phi}^T)\mathbf{\Phi} q$ $[(A^{-1})^T = (A^T)^{-1}]$

$= t^T\mathbf{\Phi} q - t^T(\mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T)\mathbf{\Phi} q$

$= t^T\mathbf{\Phi} q - t^T(\mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}(\mathbf{\Phi}^T\mathbf{\Phi}))q$ [$\mathbf{\Phi}^T\mathbf{\Phi}$ must be invertible]

$= t^T\mathbf{\Phi} q - t^T\mathbf{\Phi} q$ $[(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}(\mathbf{\Phi}^T\mathbf{\Phi}) = \mathbb{I}]$

$= 0$

In this way, it is shown that the least-squares solution corresponds to an orthogonal projection of the vector t onto the manifold S. [Proved - 2]

2. (10 points) Page 175, Question 3.7

By using the technique of completing the square, verify the result (3.49) for the posterior distribution of the parameters $\mathbf{w}$ in the linear basis function model in which $\mathbf{m}_N$ and $\mathbf{S}_N$ are defined by (3.50) and (3.51) respectively.

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N) \tag{3.49}$$

$$\mathbf{m}_N = \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\mathbf{\Phi}^T\mathbf{t}) \tag{3.50}$$

$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta\mathbf{\Phi}^T\mathbf{\Phi} \tag{3.51}$$

Response:

The posterior distribution is given by:

$P(w|t) = \mathcal{N}(w|m_N, S_N)$

where $m_N = S_N(S_0^{-1}m_0 + \beta\mathbf{\Phi}^T t)$ and $S_N = S_0^{-1} + \beta\mathbf{\Phi}^T t$.

In this problem, we have to verify this expression for posterior.

The posterior distribution is proportional to the product of likelihood $(P(t|w))$ and prior distribution $(P(w))$. The expressions for both likelihood and prior are provided below:

$P(t|w) = \prod_{n=1}^{N} \mathcal{N}(t_n|w^t\mathbf{\Phi}, \beta^{-1})$

$P(w) = \mathcal{N}(w|m_0, S_0)$

so, from the relationship among prior, posterior and likelihood, we have:

$P(w|t) \propto P(t|w)P(w)$

We have to check the equivalence of the multiplication of these two terms and a Gaussian Distribution to obtain the expressions for m_N and S_N. If we consider the posterior, its general form can be represented as:

$\mathcal{N}(w|m_N, S_N) = (2\pi)^{-D/2}|S_n|^{-1/2}exp\{-\frac{1}{2}(w - m_N)^T S_N^{-1}(w - m_N)\}$

The exponential term in the expression can be represented as:

$$-\tfrac{1}{2}(w-m_N)^T S_N^{-1}(w-m_N)$$
$$=-\tfrac{1}{2}(w^T-m_N^T)S_N^{-1}(w-m_N)$$
$$=-\tfrac{1}{2}(w^T S_N^{-1}-m_N^T S_N^{-1})(w-m_N)$$
$$=-\tfrac{1}{2}(w^T S_N^{-1}w-w^T S_N^{-1}m_N-m_N^T S_N^{-1}w+m_N^T S_N^{-1}m_N)$$
$$=-\tfrac{1}{2}(w^T S_N^{-1}w-2m_N^T S_N^{-1}w+m_N^T S_N^{-1}m_N) \quad \text{[As } m_N^T S_N^{-1}w \text{ is a scalar]}$$

Similarly, the exponential term of the product of prior and likelihood can be represented as:

$$-\tfrac{\beta}{2}\sum_{n=1}^{N}\{t_n-w^T\boldsymbol{\phi}(x_n)\}^2-\frac{1}{2}(w-m_0)^T S_0^{-1}(w-m_0)$$

Let's consider the first part,

$$-\tfrac{\beta}{2}\sum_{n=1}^{N}\{t_n-w^T\boldsymbol{\phi}(x_n)\}^2$$
$$=-\tfrac{\beta}{2}\sum_{n=1}^{N}(t_n^2-2t_n\boldsymbol{\phi}^T(x_n)w+w^T\boldsymbol{\phi}(x_n)\boldsymbol{\phi}^T(x_n)w)$$
$$=-\tfrac{1}{2}(\sum_{n=1}^{N}w^T\beta\boldsymbol{\phi}(x_n)\boldsymbol{\phi}^T(x_n)w-\sum_{n=1}^{N}2\beta t_n\boldsymbol{\phi}^T(x_n)w+\sum_{n=1}^{N}\beta t_n^2)$$

For the second part,

$$-\tfrac{1}{2}(w-m_0)^T S_0^{-1}(w-m_0)$$
$$=-\tfrac{1}{2}(w^T-m_0^T)S_0^{-1}(w-m_0)$$
$$=-\tfrac{1}{2}(w^T S_0^{-1}-m_0^T S_0^{-1})(w-m_0)$$
$$=-\tfrac{1}{2}(w^T S_0^{-1}-m_0^T S_0^{-1})(w-m_0)$$
$$=-\tfrac{1}{2}(w^T S_0^{-1}w-m_0^T S_0^{-1}w-w^T S_0^{-1}m_0+m_0^T S_0^{-1}m_0)$$
$$=-\tfrac{1}{2}(w^T S_0^{-1}w-2m_0^T S_0^{-1}w+m_0^T S_0^{-1}m_0)$$

Combining these two expressions, we get:

$$-\frac{1}{2}\{w^T(S_0^{-1}+\sum_{n=1}^{N}\beta\boldsymbol{\phi}(x_n)\boldsymbol{\phi}^T(x_n))w-(2m_0^T S_0^{-1}+\sum_{n=1}^{N}2\beta t_n\boldsymbol{\phi}^T(x_n))w+(m_0^T S_0^{-1}m_0+\sum_{n=1}^{N}\beta t_n^2)\}$$

Comparing this final expression with the expression of the posterior, we get:

$S_N^{-1} = S_0^{-1} + \sum_{n=1}^{N} \beta \boldsymbol{\phi}(x_n)\boldsymbol{\phi}^T(x_n)$

$= S_0^{-1} + \beta \boldsymbol{\Phi}^T \boldsymbol{\Phi}$

and

$m_N^T S_N^{-1} = m_0^T S_0^{-1} + \beta t_n \boldsymbol{\phi}^T(x_n)$

$\Rightarrow m_N = S_N(S_0^{-1} m_0 + \beta \boldsymbol{\Phi}^T t)$

Hence, the implication is proved.

3. (10 points) Page 175, Question 3.10

By making use of the result (2.115) to evaluate the integral in (3.57), verify that the predictive distribution for the Bayesian linear regression model is given by (3.58) in which the input-dependent variance is given by (3.59).

$$p(t|\mathbf{t}, \alpha, \beta) = \int p(t|\mathbf{w}, \beta) p(\mathbf{w}|\mathbf{t}, \alpha, \beta) d\mathbf{w} \tag{3.57}$$

$$p(t|\mathbf{x}, \mathbf{t}, \alpha, \beta) = \mathcal{N}(t|\mathbf{m}_N^T \boldsymbol{\phi}(\mathbf{x}), \sigma_N^2(\mathbf{x})) \tag{3.58}$$

$$\sigma_N^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^T \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}). \tag{3.59}$$

Marginal and Conditional Gaussians Given a marginal Gaussian distribution for $\mathbf{x}$ and a conditional Gaussian distri- bution for y given $\mathbf{x}$ in the form

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \tag{2.113}$$

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}) \tag{2.114}$$

the marginal distribution of y and the conditional distribution of x given y are given by

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T) \tag{2.115}$$

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{\mathbf{A}^T\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma}) \tag{2.116}$$

where

$$\Sigma = (\boldsymbol{\Lambda} + \mathbf{A}^T\mathbf{L}\mathbf{A})^{-1} \tag{2.117}$$

Response:

Predictive distribution of Bayesian Linear Regression attempts to provide the probability of a new target variable t for a given input based on the training dataset which can be represented as:

$p(t|\mathbf{t}, \alpha, \beta) = \int p(t|\mathbf{w}, \beta)p(\mathbf{w}|\mathbf{t}, \alpha, \beta)d\mathbf{w}$

Using (3.3), (3.8) and (3.49), we can express marginal distribution $p(\mathbf{w}|\mathbf{t})$ and the conditional distribution $p(t|\mathbf{x}, \mathbf{w}, \beta)$ as:

$$y(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x})$$

$$p(t|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(t|y(\mathbf{x}, \mathbf{w}), \beta^{-1})$$

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$$

Comparing these two representations with equations 2.113 and 2.114, we get:

$\boldsymbol{\mu} = \mathbf{m}_N$

$\boldsymbol{\Lambda}^{-1} = \mathbf{S}_N$

$\mathbf{A} = \boldsymbol{\phi}(x)^T$

$\mathbf{L}^{-1} = \beta^{-1}$

From these expressions, we can derive the following things:

$\mathbf{A}\boldsymbol{\mu} + \mathbf{b} = \boldsymbol{\phi}(x)^T\mathbf{m}_N$

$\mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{A}^T = \boldsymbol{\phi}(x)^T\mathbf{S}_N\boldsymbol{\phi}(x)$

So, according to equation 2.113 the marginal distribution $p(t|\mathbf{x}, \mathbf{t}, \alpha, \beta)$ can be represented as:

$p(t|\mathbf{x}, \mathbf{t}, \alpha, \beta) = \mathcal{N}(t|\boldsymbol{\phi}(x)^T\mathbf{m}_N, \beta^{-1} + \boldsymbol{\phi}(x)^T\mathbf{S}_N\boldsymbol{\phi}(x))$

which is the same as mentioned in equations 3.58 and 3.59 because $\boldsymbol{\phi}(x)^T\mathbf{m}_N = \mathbf{m}_N^T\boldsymbol{\phi}(x)$ as it is a scalar.

4. (10 points) Page 175, Question 3.11

> We have seen that, as the size of a data set increases, the uncertainty associated with the posterior distribution over model parameters decreases. Make use of the matrix identity (Appendix C)
>
> $$(\mathbf{M} + \mathbf{v}\mathbf{v}^T)^{-1} = \mathbf{M}^{-1} - \frac{(\mathbf{M}^{-1}\mathbf{v})(\mathbf{v}^T\mathbf{M}^{-1})}{1 + \mathbf{v}^T\mathbf{M}^{-1}\mathbf{v}} \tag{3.110}$$
>
> to show that the uncertainty $\sigma_N^2(\mathbf{x})$ associated with the linear regression function given by (3.59) satisfies
>
> $$\sigma_{N+1}^2(\mathbf{x}) \leq \sigma_N^2(\mathbf{x})$$

Response:

From equation 3.59, the expression for $\sigma_N^2(\mathbf{x})$ can be represented as:

$\sigma_N^2(\mathbf{x}) = \dfrac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^T\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x})$

$\sigma_{N+1}^2(\mathbf{x})$ then becomes: $\dfrac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^T\mathbf{S}_{N+1}\boldsymbol{\phi}(\mathbf{x})$

We want to find the difference between $\sigma_N^2(\mathbf{x})$ and $\sigma_{N+1}^2(\mathbf{x})$:

$\sigma_N^2(\mathbf{x})$ - $\sigma_{N+1}^2(\mathbf{x})$

$= \boldsymbol{\phi}(\mathbf{x})^T\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}) - \boldsymbol{\phi}(\mathbf{x})^T\mathbf{S}_{N+1}\boldsymbol{\phi}(\mathbf{x})$

$= \boldsymbol{\phi}(\mathbf{x})^T(\mathbf{S}_N - \mathbf{S}_{N+1})\boldsymbol{\phi}(\mathbf{x})$

where

$\mathbf{S}_{N+1}^{-1} = \mathbf{S}_N^{-1} + \beta\boldsymbol{\phi}_{N+1}(x)\boldsymbol{\phi}_{N+1}(x)^T$

$\mathbf{S}_{N+1} = (\mathbf{S}_N^{-1} + \beta\boldsymbol{\phi}_{N+1}(x)\boldsymbol{\phi}_{N+1}(x)^T)^{-1}$

Using the matrix identity form, we can express the following terms:

$\mathbf{M} = \mathbf{S}_N^{-1}$

$\mathbf{v} = \sqrt{\beta}\boldsymbol{\phi}_{N+1}(x)$

So, $\mathbf{S}_{N+1}$ becomes:

$$(\mathbf{S}_N^{-1})^{-1} - \frac{((\mathbf{S}_N^{-1})^{-1}\sqrt{\beta}\boldsymbol{\phi}_{N+1}(x))((\sqrt{\beta}\boldsymbol{\phi}_{N+1}(x))^T(\mathbf{S}_N^{-1})^{-1})}{1 + (\sqrt{\beta}\boldsymbol{\phi}_{N+1}(x))^T(\mathbf{S}_N^{-1})^{-1}\sqrt{\beta}\boldsymbol{\phi}_{N+1}(x)}$$

$$= \mathbf{S}_N - \frac{\mathbf{S}_N\sqrt{\beta}\boldsymbol{\phi}_{N+1}(x)\sqrt{\beta}\boldsymbol{\phi}_{N+1}(x)^T\mathbf{S}_N}{1 + \sqrt{\beta}\boldsymbol{\phi}_{N+1}(x)^T\mathbf{S}_N\sqrt{\beta}\boldsymbol{\phi}_{N+1}(x)}$$

$\mathbf{S}_N - \mathbf{S}_{N+1}$

$$= \frac{\mathbf{S}_N\sqrt{\beta}\boldsymbol{\phi}_{N+1}(x)\sqrt{\beta}\boldsymbol{\phi}_{N+1}(x)^T\mathbf{S}_N}{1 + \sqrt{\beta}\boldsymbol{\phi}_{N+1}(x)^T\mathbf{S}_N\sqrt{\beta}\boldsymbol{\phi}_{N+1}(x)}$$

After plugging this value into the expression for $\sigma_N^2(\mathbf{x})$ - $\sigma_{N+1}^2(\mathbf{x})$, it becomes:

$$\boldsymbol{\phi}(\mathbf{x})^T\frac{\mathbf{S}_N\sqrt{\beta}\boldsymbol{\phi}_{N+1}(x)\sqrt{\beta}\boldsymbol{\phi}_{N+1}(x)^T\mathbf{S}_N}{1 + \sqrt{\beta}\boldsymbol{\phi}_{N+1}(x)^T\mathbf{S}_N\sqrt{\beta}\boldsymbol{\phi}_{N+1}(x)}\boldsymbol{\phi}(\mathbf{x})$$

$$= \frac{\beta(\boldsymbol{\phi}_{N+1}(x)^T\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}))^2}{1 + \sqrt{\beta}\boldsymbol{\phi}_{N+1}(x)^T\mathbf{S}_N\sqrt{\beta}\boldsymbol{\phi}_{N+1}(x)} \geq 0$$

$\sigma_N^2(\mathbf{x}) - \sigma_{N+1}^2(\mathbf{x}) \geq 0$

$\Rightarrow \sigma_N^2(\mathbf{x}) \geq \sigma_{N+1}^2(\mathbf{x})$ [Proved]