

## Chapter 6 Kernel Methods

Yue Yu

### 6.1 Dual Representations

(6.3)

By setting the gradient of  $J(\mathbf{w})$  with respect to  $\mathbf{w}$  to  $\mathbf{0}$ , it is easy to see that

$$\begin{aligned}\mathbf{w} &= \arg_{\mathbf{w}} \nabla_{\mathbf{w}} J(\mathbf{w}) = \mathbf{0} \\ &= \arg_{\mathbf{w}} \left( \sum_{n=1}^N \{ \mathbf{w}^T \phi(\mathbf{x}_n) - t_n \} + \lambda \mathbf{w} = \mathbf{0} \right) \\ &= -\frac{1}{\lambda} \sum_{n=1}^N \{ \mathbf{w}^T \phi(\mathbf{x}_n) - t_n \} \phi(\mathbf{x}_n) \\ &= \Phi^T \mathbf{a}\end{aligned}$$

where

$$\mathbf{a} = -\frac{1}{\lambda} (\Phi \mathbf{w} - \mathbf{t}).$$

(6.5)

$$\begin{aligned}J(\mathbf{a}) &= \frac{1}{2} \sum_{n=1}^N \{ \mathbf{w}^T \phi(\mathbf{x}_n) - t_n \}^2 + \frac{\lambda}{2} \mathbf{w}^T \mathbf{w} \\ &= \frac{1}{2} \sum_{n=1}^N \{ \phi(\mathbf{x}_n)^T \mathbf{w} - t_n \}^2 + \frac{\lambda}{2} \mathbf{w}^T \mathbf{w} \\ &= \frac{1}{2} \sum_{n=1}^N \{ \mathbf{w}^T \phi(\mathbf{x}_n) \phi(\mathbf{x}_n)^T \mathbf{w} - 2 \mathbf{w}^T \phi(\mathbf{x}_n) t_n + t_n^2 \} + \frac{\lambda}{2} \mathbf{w}^T \mathbf{w} \\ &= \frac{1}{2} \mathbf{w}^T \left( \sum_{n=1}^N \phi(\mathbf{x}_n) \phi(\mathbf{x}_n)^T \right) \mathbf{w} - \mathbf{w}^T \sum_{n=1}^N \phi(\mathbf{x}_n) t_n + \frac{1}{2} \sum_{n=1}^N t_n^2 + \frac{\lambda}{2} \mathbf{w}^T \mathbf{w} \\ &= \frac{1}{2} \mathbf{a}^T \Phi \Phi^T \Phi \Phi^T \mathbf{a} - \mathbf{a}^T \Phi \Phi^T \mathbf{t} + \frac{1}{2} \mathbf{t}^T \mathbf{t} + \frac{\lambda}{2} \mathbf{a}^T \mathbf{K} \mathbf{a}.\end{aligned}$$

(6.8)

From the derivation of (6.3), we have

$$\begin{aligned}\mathbf{a} &= -\frac{1}{\lambda} (\Phi \mathbf{w} - \mathbf{t}) \\ &= -\frac{1}{\lambda} (\Phi \Phi^T \mathbf{a} - \mathbf{t}) \\ &= -\frac{1}{\lambda} (\mathbf{K} \mathbf{a} - \mathbf{t}).\end{aligned}$$

Solving for  $\mathbf{a}$ , we obtain

$$\mathbf{a} = (\mathbf{K} + \lambda \mathbf{I}_N)^{-1} \mathbf{t}.$$

(6.9)

$$y(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) = \mathbf{a}^T \Phi \phi(\mathbf{x}) = (\Phi \phi(\mathbf{x}))^T \mathbf{a} = \mathbf{k}(\mathbf{x})^T (\mathbf{K} + \lambda \mathbf{I}_N)^{-1} \mathbf{t}.$$

## 6.2 Constructing Kernels

(6.13)

To verify (6.13) is a valid kernel, we see that for any vector  $\mathbf{u} \in \mathbb{R}^N$ ,

$$\mathbf{u}^T \mathbf{K} \mathbf{u} = \mathbf{u}^T (c \mathbf{K}_1) \mathbf{u} = c (\mathbf{u}^T \mathbf{K}_1 \mathbf{u}) \geq 0,$$

which implies that  $\mathbf{K} \succeq 0$ .

(6.14)

To verify (6.14) is a valid kernel, we first notice that  $k(\mathbf{x}, \mathbf{x}')$  is symmetric by

$$k(\mathbf{x}', \mathbf{x}) = ck_1(\mathbf{x}', \mathbf{x}) = ck_1(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}').$$

Also, we see that for any vector  $\mathbf{u} \in \mathbb{R}^N$ ,

$$\begin{aligned} \mathbf{u}^T \mathbf{K} \mathbf{u} &= \sum_{i=1}^N \sum_{j=1}^N k(\mathbf{x}_i, \mathbf{x}_j) u_i u_j \\ &= \sum_{i=1}^N \sum_{j=1}^N k_1(\mathbf{x}_i, \mathbf{x}_j) (f(\mathbf{x}_i) u_i) (f(\mathbf{x}_j) u_j) \\ &= \sum_{i=1}^N \sum_{j=1}^N k_1(\mathbf{x}_i, \mathbf{x}_j) v_i v_j \\ &\geq 0 \end{aligned}$$

where we denote  $v_i = f(\mathbf{x}_i) u_i$ , implying  $\mathbf{K} \succeq 0$ . Therefore, (6.14) is a valid kernel.

(6.15)

To verify (6.15) is a valid kernel, let  $q(k_1(\mathbf{x}, \mathbf{x}')) = \sum_m c_m k_1(\mathbf{x}, \mathbf{x}')^m$  where  $c_m \geq 0$ . By repeatedly applying (6.18), we see that  $k_1(\mathbf{x}, \mathbf{x}')^m$  is a valid kernel. Then, according to (6.13),  $c_m k_1(\mathbf{x}, \mathbf{x}')^m$  is also a valid kernel. Finally, using (6.17), we can conclude that  $q(k_1(\mathbf{x}, \mathbf{x}')) = \sum_m c_m k_1(\mathbf{x}, \mathbf{x}')^m$  is a valid kernel.

(6.16)

To verify (6.16) is a valid kernel, we first notice that  $k(\mathbf{x}, \mathbf{x}')$  is symmetric by

$$k(\mathbf{x}', \mathbf{x}) = \exp(k_1(\mathbf{x}', \mathbf{x})) = \exp(k_1(\mathbf{x}, \mathbf{x}')) = k(\mathbf{x}, \mathbf{x}').$$

Also, we see that for any vector  $\mathbf{u} \in \mathbb{R}^N$ ,

$$\begin{aligned} \mathbf{u}^T \mathbf{K} \mathbf{u} &= \sum_{i=1}^N \sum_{j=1}^N k(\mathbf{x}_i, \mathbf{x}_j) u_i u_j \\ &= \sum_{i=1}^N \sum_{j=1}^N \exp(k_1(\mathbf{x}_i, \mathbf{x}_j)) u_i u_j. \end{aligned}$$

Applying the Maclaurin series for the exponential term, we have

$$\mathbf{u}^T \mathbf{K} \mathbf{u} = \sum_{i=1}^N \sum_{j=1}^N \left( \sum_{n=0}^{\infty} \frac{k_1(\mathbf{x}_i, \mathbf{x}_j)^n}{n!} \right) u_i u_j \geq 0$$

where we made use of (6.15), implying  $\mathbf{K} \succeq 0$ . Therefore, (6.16) is a valid kernel.

**(6.17)**

To verify (6.17) is a valid kernel, we first notice that  $k(\mathbf{x}, \mathbf{x}')$  is symmetric by

$$k(\mathbf{x}', \mathbf{x}) = k_1(\mathbf{x}', \mathbf{x}) + k_2(\mathbf{x}', \mathbf{x}) = k_1(\mathbf{x}, \mathbf{x}') + k_2(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}').$$

Also, we see that for any vector  $\mathbf{u} \in \mathbb{R}^N$ ,

$$\begin{aligned} \mathbf{u}^T \mathbf{K} \mathbf{u} &= \sum_{i=1}^N \sum_{j=1}^N k(\mathbf{x}_i, \mathbf{x}_j) u_i u_j \\ &= \sum_{i=1}^N \sum_{j=1}^N (k_1(\mathbf{x}_i, \mathbf{x}_j) + k_2(\mathbf{x}_i, \mathbf{x}_j)) u_i u_j \\ &= \sum_{i=1}^N \sum_{j=1}^N k_1(\mathbf{x}_i, \mathbf{x}_j) u_i u_j + \sum_{i=1}^N \sum_{j=1}^N k_2(\mathbf{x}_i, \mathbf{x}_j) u_i u_j \\ &\geq 0, \end{aligned}$$

which implies that  $\mathbf{K} \succeq 0$ . Therefore, (6.17) is a valid kernel.

**(6.18)**

To verify (6.18) is a valid kernel, we first notice that  $k(\mathbf{x}, \mathbf{x}')$  is symmetric by

$$k(\mathbf{x}', \mathbf{x}) = k_1(\mathbf{x}', \mathbf{x}) k_2(\mathbf{x}', \mathbf{x}) = k_1(\mathbf{x}, \mathbf{x}') k_2(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}').$$

Also, we see that for any vector  $\mathbf{u} \in \mathbb{R}^N$ ,

$$\begin{aligned} \mathbf{u}^T \mathbf{K} \mathbf{u} &= \sum_{i=1}^N \sum_{j=1}^N k(\mathbf{x}_i, \mathbf{x}_j) u_i u_j \\ &= \sum_{i=1}^N \sum_{j=1}^N k_1(\mathbf{x}_i, \mathbf{x}_j) k_2(\mathbf{x}_i, \mathbf{x}_j) u_i u_j \\ &= \sum_{i=1}^N \sum_{j=1}^N \phi_1(\mathbf{x}_i)^T \phi_1(\mathbf{x}_j) \phi_2(\mathbf{x}_i)^T \phi_2(\mathbf{x}_j) u_i u_j \\ &= \sum_{i=1}^N \sum_{j=1}^N \left( \sum_{l=1}^M \phi_{1l}(\mathbf{x}_i) \phi_{1l}(\mathbf{x}_j) \right) \left( \sum_{m=1}^M \phi_{2m}(\mathbf{x}_i) \phi_{2m}(\mathbf{x}_j) \right) u_i u_j \\ &= \sum_{l=1}^M \sum_{m=1}^M \left( \sum_{i=1}^N \phi_{1l}(\mathbf{x}_i) \phi_{2m}(\mathbf{x}_i) u_i \right)^2 \\ &\geq 0, \end{aligned}$$

which implies that  $\mathbf{K} \succeq 0$ . Therefore, (6.18) is a valid kernel.

(6.20)

To verify (6.20) is a valid kernel, recall that any symmetric matrix  $\mathbf{S}$  can be diagonalized by  $\mathbf{S} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$  where the columns of  $\mathbf{Q}$  are orthogonal eigenvectors. Hence,

$$k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T \mathbf{A} \mathbf{x}' = \mathbf{x}^T \mathbf{Q} \mathbf{\Lambda} \mathbf{Q}^T \mathbf{x}' = (\mathbf{Q}^T \mathbf{x})^T \mathbf{\Lambda} (\mathbf{Q}^T \mathbf{x}').$$

Denoting  $\mathbf{v} = \mathbf{Q}^T \mathbf{x}$  and  $\mathbf{v}' = \mathbf{Q}^T \mathbf{x}'$ ,

$$k(\mathbf{x}, \mathbf{x}') = \mathbf{v}^T \mathbf{\Lambda} \mathbf{v}' = \sum_{i=1}^N \Lambda_{ii} v_i v'_i = \sum_{i=1}^N (\sqrt{\Lambda_{ii}} v_i) (\sqrt{\Lambda_{ii}} v'_i) = \phi(\mathbf{x})^T \phi(\mathbf{x}') = k'(\phi(\mathbf{x}), \phi(\mathbf{x}'))$$

where  $k'$  is the linear kernel, and the square roots exist because the eigenvalues for a positive semidefinite matrix are non-negative. Therefore, according to (6.19),  $k(\mathbf{x}, \mathbf{x}')$  is a valid kernel.

(6.21)

$$\begin{aligned} k(\mathbf{x}, \mathbf{x}') &= k_a(\mathbf{x}_a, \mathbf{x}'_a) + k_b(\mathbf{x}_b, \mathbf{x}'_b) \\ &= \sum_{s=1}^S \phi_{as}(\mathbf{x}_a) \phi_{as}(\mathbf{x}'_a) + \sum_{t=1}^T \phi_{bt}(\mathbf{x}_b) \phi_{bt}(\mathbf{x}'_b) \\ &= \sum_{i=1}^{S+T} \phi_i(\mathbf{x}) \phi_i(\mathbf{x}'), \end{aligned}$$

which is a valid kernel.

(6.22) – To be updated

$$\begin{aligned} k(\mathbf{x}, \mathbf{x}') &= k_a(\mathbf{x}_a, \mathbf{x}'_a) k_b(\mathbf{x}_b, \mathbf{x}'_b) \\ &= \sum_{s=1}^S \phi_{as}(\mathbf{x}_a) \phi_{as}(\mathbf{x}'_a) \sum_{t=1}^T \phi_{bt}(\mathbf{x}_b) \phi_{bt}(\mathbf{x}'_b) \\ &= \sum_{s=1}^S \sum_{t=1}^T \{ \phi_{as}(\mathbf{x}_a) \phi_{bt}(\mathbf{x}_b) \} \{ \phi_{as}(\mathbf{x}'_a) \phi_{bt}(\mathbf{x}'_b) \} \end{aligned}$$

## 6.3 Radial Basis Function Networks

Skipped reading.

## 6.4 Gaussian Processes

(6.52)

$$\mathbb{E}[\mathbf{y}] = \mathbb{E}[\mathbf{\Phi} \mathbf{w}] = \mathbf{\Phi} \mathbb{E}[\mathbf{w}] = \mathbf{0}.$$

(6.53)

$$\text{cov}[\mathbf{y}] = \mathbb{E}[\mathbf{y} \mathbf{y}^T] - \mathbb{E}[\mathbf{y}] \mathbb{E}[\mathbf{y}^T] = \mathbb{E}[\mathbf{y} \mathbf{y}^T] = \mathbb{E}[\mathbf{\Phi} \mathbf{w} \mathbf{w}^T \mathbf{\Phi}^T] = \mathbf{\Phi} \mathbb{E}[\mathbf{w} \mathbf{w}^T] \mathbf{\Phi}^T = \mathbf{\Phi} \text{cov}[\mathbf{w}] \mathbf{\Phi}^T = \frac{1}{\alpha} \mathbf{\Phi} \mathbf{\Phi}^T = \mathbf{K}$$

where

$$K_{nm} = k(\mathbf{x}_n, \mathbf{x}_m) = \frac{1}{\alpha} \phi(\mathbf{x}_n)^T \phi(\mathbf{x}_m).$$

(6.61)

We use the property that if  $p(\mathbf{x})$  and  $p(\mathbf{y}|\mathbf{x})$  are Gaussian,  $p(\mathbf{y})$  is also a Gaussian. Plugging the equations

$$\begin{aligned}\boldsymbol{\mu} &= \mathbf{0} \\ \boldsymbol{\Lambda}^{-1} &= \mathbf{K} \\ \mathbf{A}\mathbf{x} + \mathbf{b} &= \mathbf{y} \\ \mathbf{L}^{-1} &= \beta^{-1}\mathbf{I}_N\end{aligned}$$

into (2.115), we obtain

$$p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{y})p(\mathbf{y})d\mathbf{y} = \mathcal{N}(\mathbf{0}, \mathbf{C})$$

where

$$\mathbf{C} = \mathbf{K} + \beta^{-1}\mathbf{I}_N.$$

(6.66) – (6.67)

Using the property that if two sets of variables are jointly Gaussian, the conditional distribution of one on the other is also Gaussian, we have

$$\begin{aligned}\mu_{t_{N+1}|\mathbf{t}} &= \mu_{t_{N+1}} + \boldsymbol{\Sigma}_{t_{N+1}\mathbf{t}}\boldsymbol{\Sigma}_{\mathbf{t}\mathbf{t}}^{-1}(\mathbf{t} - \boldsymbol{\mu}_{\mathbf{t}}) \\ &= 0 + \mathbf{k}^T\mathbf{C}^{-1}(\mathbf{t} - \mathbf{0}) \\ &= \mathbf{k}^T\mathbf{C}_N^{-1}\mathbf{t} \\ \sigma_{t_{N+1}|\mathbf{t}}^2 &= \sigma_{t_{N+1}t_{N+1}}^2 - \boldsymbol{\Sigma}_{t_{N+1}\mathbf{t}}\boldsymbol{\Sigma}_{\mathbf{t}\mathbf{t}}^{-1}\boldsymbol{\Sigma}_{\mathbf{t}t_{N+1}} \\ &= c - \mathbf{k}^T\mathbf{C}_N^{-1}\mathbf{k}.\end{aligned}$$

(6.79)

$$\begin{aligned}p(\mathbf{t}_N|\mathbf{a}_N) &= \prod_{n=1}^N \sigma(a_n)^{t_n} (1 - \sigma(a_n))^{1-t_n} \\ &= \prod_{n=1}^N \left( \frac{1}{1 + \exp(-a_n)} \right)^{t_n} \left( \frac{\exp(-a_n)}{1 + \exp(-a_n)} \right)^{1-t_n} \\ &= \prod_{n=1}^N \left( \frac{1}{1 + \exp(-a_n)} \frac{1 + \exp(-a_n)}{\exp(-a_n)} \right)^{t_n} \exp(-a_n) \\ &= \prod_{n=1}^N e^{a_n t_n} \sigma(-a_n).\end{aligned}$$

(6.81)

Here,  $-\boldsymbol{\sigma}_N$  is derived from

$$\begin{aligned}\nabla \left( -\sum_{n=1}^N \ln(1 + e^{a_n}) \right) &= \left[ \frac{\partial f}{\partial a_1}, \dots, \frac{\partial f}{\partial a_N} \right]^T \\ &= [\sigma(a_1), \dots, \sigma(a_N)]^T \\ &= -\boldsymbol{\sigma}_N.\end{aligned}$$

(6.82)

Noticing that

$$\frac{d\sigma(a_i)}{da_j} = \begin{cases} \sigma(a_i)(1 - \sigma(a_i)) & \text{if } i = j \\ 0 & \text{if } i \neq j, \end{cases}$$

and  $\mathbf{C}_N$  is symmetric, it is easy to see that

$$\nabla \nabla \Psi(\mathbf{a}_N) = -\mathbf{W}_N - \mathbf{C}_N^{-1}.$$

For any diagonal element, we have

$$\sigma(a_i)(1 - \sigma(a_i)) = -\left(\sigma(a_i) - \frac{1}{2}\right)^2 + \frac{1}{4} \leq \frac{1}{4},$$

which implies that  $\mathbf{W}_N$  is positive definite.

(6.83)

According to the Newton-Raphson formula, we have

$$\begin{aligned} \mathbf{a}_N^{\text{new}} &= \mathbf{a}_N - \mathbf{H}^{-1} \nabla \Psi(\mathbf{a}_N) \\ &= \mathbf{a}_N - (-\mathbf{W}_N - \mathbf{C}_N^{-1})^{-1} \{\mathbf{t}_N - \boldsymbol{\sigma}_N - \mathbf{C}_N^{-1} \mathbf{a}_N\} \\ &= \mathbf{a}_N + (\mathbf{W}_N + \mathbf{C}_N^{-1})^{-1} \{\mathbf{t}_N - \boldsymbol{\sigma}_N - \mathbf{C}_N^{-1} \mathbf{a}_N\} \\ &= (\mathbf{W}_N + \mathbf{C}_N^{-1})^{-1} \{(\mathbf{W}_N + \mathbf{C}_N^{-1}) \mathbf{a}_N + \mathbf{t}_N - \boldsymbol{\sigma}_N - \mathbf{C}_N^{-1} \mathbf{a}_N\} \\ &= \mathbf{C}_N (\mathbf{I} + \mathbf{W}_N \mathbf{C}_N)^{-1} \{\mathbf{t}_N - \boldsymbol{\sigma}_N + \mathbf{W}_N \mathbf{a}_N\}. \end{aligned}$$

(6.84)

At the mode, the gradient  $\nabla \Psi(\mathbf{a}_N)$  vanishes, that is

$$\nabla \Psi(\mathbf{a}_N) = \mathbf{t}_N - \boldsymbol{\sigma}_N - \mathbf{C}_N^{-1} \mathbf{a}_N = \mathbf{0}.$$

Solving for  $\mathbf{a}_N$ , we obtain

$$\mathbf{a}_N = \mathbf{C}_N (\mathbf{t}_N - \boldsymbol{\sigma}_N).$$

Therefore,

$$\mathbf{a}_N^* = \mathbf{a}_N - \mathbf{H}^{-1} \nabla \Psi(\mathbf{a}_N) = \mathbf{a}_N = \mathbf{C}_N (\mathbf{t}_N - \boldsymbol{\sigma}_N).$$

(6.87) – (6.88)

Let

$$\begin{aligned} p(\mathbf{a}_N | \mathbf{t}_N) &= \mathcal{N}(\mathbf{a}_N | \mathbf{a}_N^*, \mathbf{H}^{-1}) = \mathcal{N}(\mathbf{a}_N | \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \\ p(a_{N+1} | \mathbf{a}_N) &= \mathcal{N}(a_{N+1} | \mathbf{k}^T \mathbf{C}_N^{-1} \mathbf{a}_N, c - \mathbf{k}^T \mathbf{C}_N^{-1} \mathbf{k}) = \mathcal{N}(a_{N+1} | \mathbf{A} \mathbf{a}_N + \mathbf{b}, \mathbf{L}^{-1}), \end{aligned}$$

we have

$$\begin{aligned} \boldsymbol{\mu} &= \mathbf{a}_N^* \\ \boldsymbol{\Lambda}^{-1} &= \mathbf{H}^{-1} \\ \mathbf{A} &= \mathbf{k}^T \mathbf{C}_N^{-1} \\ \mathbf{b} &= \mathbf{0} \\ \mathbf{L}^{-1} &= c - \mathbf{k}^T \mathbf{C}_N^{-1} \mathbf{k}. \end{aligned}$$

According to (2.115), we obtain

$$\begin{aligned}
\mathbb{E}[a_{N+1}|\mathbf{t}_N] &= \mathbf{A}\boldsymbol{\mu} + \mathbf{b} \\
&= \mathbf{k}^T \mathbf{C}_N^{-1} \mathbf{a}_N^* + \mathbf{0} \\
&= \mathbf{k}^T \mathbf{C}_N^{-1} \mathbf{C}_N (\mathbf{t}_N - \boldsymbol{\sigma}_N) \\
&= \mathbf{k}^T (\mathbf{t}_N - \boldsymbol{\sigma}_N) \\
\text{var}[a_{N+1}|\mathbf{t}_N] &= \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T \\
&= c - \mathbf{k}^T \mathbf{C}_N^{-1} \mathbf{k} + \mathbf{k}^T \mathbf{C}_N^{-1} \mathbf{H}^{-1} (\mathbf{k}^T \mathbf{C}_N^{-1})^T \\
&= c - \mathbf{k}^T \mathbf{C}_N^{-1} \mathbf{k} + \mathbf{k}^T \mathbf{C}_N^{-1} (\mathbf{W}_N + \mathbf{C}_N^{-1})^{-1} \mathbf{C}_N^{-1} \mathbf{k} \\
&= c - \mathbf{k}^T \mathbf{C}_N^{-1} (\mathbf{W}_N + \mathbf{C}_N^{-1})^{-1} (\mathbf{W}_N + \mathbf{C}_N^{-1}) \mathbf{k} + \mathbf{k}^T \mathbf{C}_N^{-1} (\mathbf{W}_N + \mathbf{C}_N^{-1})^{-1} \mathbf{C}_N^{-1} \mathbf{k} \\
&= c - \mathbf{k}^T \mathbf{C}_N^{-1} (\mathbf{W}_N + \mathbf{C}_N^{-1})^{-1} ((\mathbf{W}_N + \mathbf{C}_N^{-1}) - \mathbf{C}_N^{-1}) \mathbf{k} \\
&= c - \mathbf{k}^T \mathbf{C}_N^{-1} (\mathbf{W}_N + \mathbf{C}_N^{-1})^{-1} \mathbf{W}_N \mathbf{k} \\
&= c - \mathbf{k}^T (\mathbf{W}_N^{-1} (\mathbf{W}_N + \mathbf{C}_N^{-1}) \mathbf{C}_N)^{-1} \mathbf{k} \\
&= c - \mathbf{k}^T ((\mathbf{I} + \mathbf{W}_N^{-1} \mathbf{C}_N^{-1}) \mathbf{C}_N)^{-1} \mathbf{k} \\
&= c - \mathbf{k}^T (\mathbf{W}_N^{-1} + \mathbf{C}_N)^{-1} \mathbf{k}.
\end{aligned}$$

**(6.90)**

Identifying  $f(\mathbf{a}_N) = p(\mathbf{t}_N|\mathbf{a}_N)p(\mathbf{a}_N|\boldsymbol{\theta})$  and  $Z = p(\mathbf{t}_N|\boldsymbol{\theta})$  and applying the Laplace approximation, we obtain

$$\begin{aligned}
\ln p(\mathbf{t}_N|\boldsymbol{\theta}) &= \ln \left( f(\mathbf{a}_N^*) \frac{(2\pi)^{N/2}}{|\mathbf{H}|^{1/2}} \right) \\
&= \ln(p(\mathbf{t}_N|\mathbf{a}_N)p(\mathbf{a}_N|\boldsymbol{\theta})) - \frac{1}{2} \ln |\mathbf{W}_N + \mathbf{C}_N^{-1}| + \frac{N}{2} \ln(2\pi) \\
&= \Psi(\mathbf{a}_N^*) - \frac{1}{2} \ln |\mathbf{W}_N + \mathbf{C}_N^{-1}| + \frac{N}{2} \ln(2\pi)
\end{aligned}$$

where we made use of (4.135), (6.80) and (6.85).

**(6.91) – (6.94) – To be updated**