ECE1513
Tutorial 2:
Selected Exercises from Chapter 3

TA: Faye Pourghasem, fateme.pourghasem@mail.utoronto.ca

September 16, 2023

1 Example 3.1

(⋆) **WWW** Show that the 'tanh' function and the logistic sigmoid function (3.6) are related by

$$\tanh(a) = 2\sigma(2a) - 1. \tag{3.100}$$

Hence show that a general linear combination of logistic sigmoid functions of the form

$$y(x, \mathbf{w}) = w_0 + \sum_{j=1}^{M} w_j \sigma\left(\frac{x - \mu_j}{s}\right) \tag{3.101}$$

is equivalent to a linear combination of 'tanh' functions of the form

$$y(x, \mathbf{u}) = u_0 + \sum_{j=1}^{M} u_j \tanh\left(\frac{x - \mu_j}{s}\right) \tag{3.102}$$

and find expressions to relate the new parameters $\{u_1, \ldots, u_M\}$ to the original parameters $\{w_1, \ldots, w_M\}$.

Solution

3.1 **NOTE**: In the 1st printing of PRML, there is a 2 missing in the denominator of the argument to the 'tanh' function in equation (3.102).

Using (3.6), we have

$$\begin{aligned}
2\sigma(2a) - 1 &= \frac{2}{1 + e^{-2a}} - 1 \\
&= \frac{2}{1 + e^{-2a}} - \frac{1 + e^{-2a}}{1 + e^{-2a}} \\
&= \frac{1 - e^{-2a}}{1 + e^{-2a}} \\
&= \frac{e^{a} - e^{-a}}{e^{a} + e^{-a}} \\
&= \tanh(a)
\end{aligned}$$

If we now take $a_j = (x - \mu_j)/2s$, we can rewrite (3.101) as

$$\begin{aligned} y(\mathbf{x}, \mathbf{w}) &= w_0 + \sum_{j=1}^{M} w_j \sigma(2a_j) \\ &= w_0 + \sum_{j=1}^{M} \frac{w_j}{2} \left(2\sigma(2a_j) - 1 + 1\right) \\ &= u_0 + \sum_{j=1}^{M} u_j \tanh(a_j), \end{aligned}$$

where $u_j = w_j/2$, for $j = 1, \ldots, M$, and $u_0 = w_0 + \sum_{j=1}^{M} w_j/2$.

2 Example 3.4

(⋆) **WWW** Consider a linear model of the form

$$y(x, \mathbf{w}) = w_0 + \sum_{i=1}^{D} w_i x_i \tag{3.105}$$

together with a sum-of-squares error function of the form

$$E_D(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 . \tag{3.106}$$

Now suppose that Gaussian noise ϵ_i with zero mean and variance σ^2 is added independently to each of the input variables x_i. By making use of $\mathbb{E}[\epsilon_i] = 0$ and $\mathbb{E}[\epsilon_i \epsilon_j] = \delta_{ij}\sigma^2$, show that minimizing E_D averaged over the noise distribution is equivalent to minimizing the sum-of-squares error for noise-free input variables with the addition of a weight-decay regularization term, in which the bias parameter w_0 is omitted from the regularizer.

Solution

3.4 Let

$$\begin{aligned}\widetilde{y}_n &= w_0 + \sum_{i=1}^{D} w_i(x_{ni} + \epsilon_{ni}) \\ &= y_n + \sum_{i=1}^{D} w_i \epsilon_{ni}\end{aligned}$$

where $y_n = y(x_n, \mathbf{w})$ and $\epsilon_{ni} \sim \mathcal{N}(0, \sigma^2)$ and we have used (3.105). From (3.106) we then define

$$\begin{aligned}\widetilde{E} &= \frac{1}{2}\sum_{n=1}^{N}\{\widetilde{y}_n - t_n\}^2 \\ &= \frac{1}{2}\sum_{n=1}^{N}\{\widetilde{y}_n^2 - 2\widetilde{y}_n t_n + t_n^2\} \\ &= \frac{1}{2}\sum_{n=1}^{N}\left\{y_n^2 + 2y_n\sum_{i=1}^{D} w_i\epsilon_{ni} + \left(\sum_{i=1}^{D} w_i\epsilon_{ni}\right)^2 \right. \\ &\qquad \left. -2t_n y_n - 2t_n\sum_{i=1}^{D} w_i\epsilon_{ni} + t_n^2\right\}.\end{aligned}$$

If we take the expectation of $\widetilde{E}$ under the distribution of ϵ_{ni}, we see that the second and fifth terms disappear, since $\mathbb{E}[\epsilon_{ni}] = 0$, while for the third term we get

$$\mathbb{E}\left[\left(\sum_{i=1}^{D} w_i\epsilon_{ni}\right)^2\right] = \sum_{i=1}^{D} w_i^2\sigma^2$$

since the ϵ_{ni} are all independent with variance σ^2.

From this and (3.106) we see that

$$\mathbb{E}\left[\widetilde{E}\right] = E_D + \frac{1}{2}\sum_{i=1}^{D} w_i^2\sigma^2,$$

as required.

3 Example 3.6

(⋆) WWW Consider a linear basis function regression model for a multivariate target variable $\mathbf{t}$ having a Gaussian distribution of the form

$$p(\mathbf{t}|\mathbf{W}, \mathbf{\Sigma}) = \mathcal{N}(\mathbf{t}|\mathbf{y}(\mathbf{x}, \mathbf{W}), \mathbf{\Sigma}) \tag{3.107}$$

where

$$\mathbf{y}(\mathbf{x}, \mathbf{W}) = \mathbf{W}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}) \tag{3.108}$$

together with a training data set comprising input basis vectors $\boldsymbol{\phi}(\mathbf{x}_n)$ and corresponding target vectors $\mathbf{t}_n$, with $n = 1, \ldots, N$. Show that the maximum likelihood solution $\mathbf{W}_{\mathrm{ML}}$ for the parameter matrix $\mathbf{W}$ has the property that each column is given by an expression of the form (3.15), which was the solution for an isotropic noise distribution. Note that this is independent of the covariance matrix $\mathbf{\Sigma}$. Show that the maximum likelihood solution for $\mathbf{\Sigma}$ is given by

$$\mathbf{\Sigma} = \frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{t}_n - \mathbf{W}_{\mathrm{ML}}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\right)\left(\mathbf{t}_n - \mathbf{W}_{\mathrm{ML}}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\right)^{\mathrm{T}}. \tag{3.109}$$

Solution

3.6 We first write down the log likelihood function which is given by

$$\ln L(\mathbf{W}, \mathbf{\Sigma}) = -\frac{N}{2}\ln|\mathbf{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{t}_n - \mathbf{W}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n))^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{t}_n - \mathbf{W}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)).$$

First of all we set the derivative with respect to $\mathbf{W}$ equal to zero, giving

$$0 = -\sum_{n=1}^{N}\mathbf{\Sigma}^{-1}(\mathbf{t}_n - \mathbf{W}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n))\boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}}.$$

Multiplying through by $\mathbf{\Sigma}$ and introducing the design matrix $\mathbf{\Phi}$ and the target data matrix $\mathbf{T}$ we have

$$\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{W} = \mathbf{\Phi}^{\mathrm{T}}\mathbf{T}$$

Solving for $\mathbf{W}$ then gives (3.15) as required.

The maximum likelihood solution for $\mathbf{\Sigma}$ is easily found by appealing to the standard result from Chapter 2 giving

$$\mathbf{\Sigma} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{t}_n - \mathbf{W}_{\mathrm{ML}}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n))(\mathbf{t}_n - \mathbf{W}_{\mathrm{ML}}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n))^{\mathrm{T}}.$$

as required. Since we are finding a joint maximum with respect to both $\mathbf{W}$ and $\mathbf{\Sigma}$ we see that it is $\mathbf{W}_{\mathrm{ML}}$ which appears in this expression, as in the standard result for an unconditional Gaussian distribution.

4 Example 3.11

($\star\star$) 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)

$$\left(\mathbf{M}+\mathbf{v}\mathbf{v}^{\mathrm{T}}\right)^{-1}=\mathbf{M}^{-1}-\frac{\left(\mathbf{M}^{-1}\mathbf{v}\right)\left(\mathbf{v}^{\mathrm{T}}\mathbf{M}^{-1}\right)}{1+\mathbf{v}^{\mathrm{T}}\mathbf{M}^{-1}\mathbf{v}} \tag{3.110}$$

to show that the uncertainty $\sigma_N^2(\mathbf{x})$ associated with the linear regression function given by (3.59) satisfies

$$\sigma_{N+1}^2(\mathbf{x}) \leqslant \sigma_N^2(\mathbf{x}). \tag{3.111}$$

Solution

3.11 From (3.59) we have

$$\sigma_{N+1}^2(\mathbf{x})=\frac{1}{\beta}+\boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{S}_{N+1}\boldsymbol{\phi}(\mathbf{x}) \tag{133}$$

where $\mathbf{S}_{N+1}$ is given by (131). From (131) and (3.110) we get

$$\begin{aligned}\mathbf{S}_{N+1} &= \left(\mathbf{S}_N^{-1}+\beta\boldsymbol{\phi}_{N+1}\boldsymbol{\phi}_{N+1}^{\mathrm{T}}\right)^{-1}\\ &= \mathbf{S}_N-\frac{\left(\mathbf{S}_N\boldsymbol{\phi}_{N+1}\beta^{1/2}\right)\left(\beta^{1/2}\boldsymbol{\phi}_{N+1}^{\mathrm{T}}\mathbf{S}_N\right)}{1+\beta\boldsymbol{\phi}_{N+1}^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}_{N+1}}\\ &= \mathbf{S}_N-\frac{\beta\mathbf{S}_N\boldsymbol{\phi}_{N+1}\boldsymbol{\phi}_{N+1}^{\mathrm{T}}\mathbf{S}_N}{1+\beta\boldsymbol{\phi}_{N+1}^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}_{N+1}}.\end{aligned}$$

Using this and (3.59), we can rewrite (133) as

$$\begin{aligned}\sigma_{N+1}^2(\mathbf{x}) &= \frac{1}{\beta}+\boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\left(\mathbf{S}_N-\frac{\beta\mathbf{S}_N\boldsymbol{\phi}_{N+1}\boldsymbol{\phi}_{N+1}^{\mathrm{T}}\mathbf{S}_N}{1+\beta\boldsymbol{\phi}_{N+1}^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}_{N+1}}\right)\boldsymbol{\phi}(\mathbf{x})\\ &= \sigma_N^2(\mathbf{x})-\frac{\beta\boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}_{N+1}\boldsymbol{\phi}_{N+1}^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x})}{1+\beta\boldsymbol{\phi}_{N+1}^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}_{N+1}}.\end{aligned} \tag{134}$$

Since $\mathbf{S}_N$ is positive definite, the numerator and denominator of the second term in (134) will be non-negative and positive, respectively, and hence $\sigma_{N+1}^2(\mathbf{x}) \leqslant \sigma_N^2(\mathbf{x})$.

5 Example 3.15

(⋆) WWW Consider a linear basis function model for regression in which the parameters α and β are set using the evidence framework. Show that the function $E(\mathbf{m}_N)$ defined by (3.82) satisfies the relation $2E(\mathbf{m}_N) = N$.

Solution

3.15 This is easily shown by substituting the re-estimation formulae (3.92) and (3.95) into (3.82), giving

$$\begin{aligned} E(\mathbf{m}_N) &= \frac{\beta}{2}\|\mathbf{t} - \mathbf{\Phi}\mathbf{m}_N\|^2 + \frac{\alpha}{2}\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N \\ &= \frac{N-\gamma}{2} + \frac{\gamma}{2} = \frac{N}{2} \,. \end{aligned}$$

6 Example 3.18

(⋆⋆) **WWW** By completing the square over $\mathbf{w}$, show that the error function (3.79) in Bayesian linear regression can be written in the form (3.80).

Solution

3.18 We can rewrite (3.79)

$$\begin{aligned}
&\frac{\beta}{2}\|\mathbf{t}-\mathbf{\Phi w}\|^2+\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\\
&\quad=\frac{\beta}{2}\left(\mathbf{t}^{\mathrm{T}}\mathbf{t}-2\mathbf{t}^{\mathrm{T}}\mathbf{\Phi w}+\mathbf{w}^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi w}\right)+\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\\
&\quad=\frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t}-2\beta\mathbf{t}^{\mathrm{T}}\mathbf{\Phi w}+\mathbf{w}^{\mathrm{T}}\mathbf{A w}\right)
\end{aligned}$$

where, in the last line, we have used (3.81). We now use the tricks of adding $\mathbf{0}=\mathbf{m}_{\mathrm{N}}^{\mathrm{T}}\mathbf{A}\mathbf{m}_{\mathrm{N}}-\mathbf{m}_{\mathrm{N}}^{\mathrm{T}}\mathbf{A}\mathbf{m}_{\mathrm{N}}$ and using $\mathbf{I}=\mathbf{A}^{-1}\mathbf{A}$, combined with (3.84), as follows:

$$\begin{aligned}
&\frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t}-2\beta\mathbf{t}^{\mathrm{T}}\mathbf{\Phi w}+\mathbf{w}^{\mathrm{T}}\mathbf{A w}\right)\\
&\quad=\frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t}-2\beta\mathbf{t}^{\mathrm{T}}\mathbf{\Phi}\mathbf{A}^{-1}\mathbf{A w}+\mathbf{w}^{\mathrm{T}}\mathbf{A w}\right)\\
&\quad=\frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t}-2\mathbf{m}_N^{\mathrm{T}}\mathbf{A w}+\mathbf{w}^{\mathrm{T}}\mathbf{A w}+\mathbf{m}_N^{\mathrm{T}}\mathbf{A}\mathbf{m}_N-\mathbf{m}_N^{\mathrm{T}}\mathbf{A}\mathbf{m}_N\right)\\
&\quad=\frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t}-\mathbf{m}_N^{\mathrm{T}}\mathbf{A}\mathbf{m}_N\right)+\frac{1}{2}(\mathbf{w}-\mathbf{m}_N)^{\mathrm{T}}\mathbf{A}(\mathbf{w}-\mathbf{m}_N).
\end{aligned}$$

Here the last term equals term the last term of (3.80) and so it remains to show that the first term equals the r.h.s. of (3.82). To do this, we use the same tricks again:

$$\begin{aligned}
\frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t}-\mathbf{m}_N^{\mathrm{T}}\mathbf{A}\mathbf{m}_N\right)&=\frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t}-2\mathbf{m}_N^{\mathrm{T}}\mathbf{A}\mathbf{m}_N+\mathbf{m}_N^{\mathrm{T}}\mathbf{A}\mathbf{m}_N\right)\\
&=\frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t}-2\mathbf{m}_N^{\mathrm{T}}\mathbf{A}\mathbf{A}^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{t}\beta+\mathbf{m}_N^{\mathrm{T}}\left(\alpha\mathbf{I}+\beta\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right)\mathbf{m}_N\right)\\
&=\frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t}-2\mathbf{m}_N^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}\mathbf{t}\beta+\beta\mathbf{m}_N^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{m}_N+\alpha\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N\right)\\
&=\frac{1}{2}\left(\beta(\mathbf{t}-\mathbf{\Phi}\mathbf{m}_N)^{\mathrm{T}}(\mathbf{t}-\mathbf{\Phi}\mathbf{m}_N)+\alpha\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N\right)\\
&=\frac{\beta}{2}\|\mathbf{t}-\mathbf{\Phi}\mathbf{m}_N\|^2+\frac{\alpha}{2}\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N
\end{aligned}$$

7 Example 3.20

(⋆⋆) WWW Starting from (3.86) verify all of the steps needed to show that maximization of the log marginal likelihood function (3.86) with respect to α leads to the re-estimation equation (3.92).

Solution

3.20 We only need to consider the terms of (3.86) that depend on α, which are the first, third and fourth terms.

Following the sequence of steps in Section 3.5.2, we start with the last of these terms,

$$-\frac{1}{2}\ln|\mathbf{A}|.$$

From (3.81), (3.87) and the fact that that eigenvectors $\mathbf{u}_i$ are orthonormal (see also Appendix C), we find that the eigenvectors of $\mathbf{A}$ to be $\alpha+\lambda_i$. We can then use (C.47) and the properties of the logarithm to take us from the left to the right side of (3.88).

The derivatives for the first and third term of (3.86) are more easily obtained using standard derivatives and (3.82), yielding

$$\frac{1}{2}\left(\frac{M}{\alpha}+\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N\right).$$

We combine these results into (3.89), from which we get (3.92) via (3.90). The expression for γ in (3.91) is obtained from (3.90) by substituting

$$\sum_i^M \frac{\lambda_i+\alpha}{\lambda_i+\alpha}$$

for M and re-arranging.

8 Example 3.23

(⋆⋆) WWW Show that the marginal probability of the data, in other words the model evidence, for the model described in Exercise 3.12 is given by

$$p(\mathbf{t}) = \frac{1}{(2\pi)^{N/2}} \frac{b_0^{a_0}}{b_N^{a_N}} \frac{\Gamma(a_N)}{\Gamma(a_0)} \frac{|\mathbf{S}_N|^{1/2}}{|\mathbf{S}_0|^{1/2}} \tag{3.118}$$

by first marginalizing with respect to $\mathbf{w}$ and then with respect to β.

Solution

3.23 From (3.10), (3.112) and the properties of the Gaussian and Gamma distributions (see Appendix B), we get

$$
\begin{aligned}
p(\mathbf{t}) &= \iint p(\mathbf{t}|\mathbf{w},\beta)p(\mathbf{w}|\beta)\,\mathrm{d}\mathbf{w}p(\beta)\,\mathrm{d}\beta \\
&= \iint \left(\frac{\beta}{2\pi}\right)^{N/2} \exp\left\{-\frac{\beta}{2}(\mathbf{t}-\mathbf{\Phi}\mathbf{w})^{\mathrm{T}}(\mathbf{t}-\mathbf{\Phi}\mathbf{w})\right\} \\
&\qquad \left(\frac{\beta}{2\pi}\right)^{M/2} |\mathbf{S}_0|^{-1/2} \exp\left\{-\frac{\beta}{2}(\mathbf{w}-\mathbf{m}_0)^{\mathrm{T}}\mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0)\right\}\,\mathrm{d}\mathbf{w} \\
&\qquad \Gamma(a_0)^{-1} b_0^{a_0} \beta^{a_0-1} \exp(-b_0\beta)\,\mathrm{d}\beta \\
&= \frac{b_0^{a_0}}{((2\pi)^{M+N}|\mathbf{S}_0|)^{1/2}} \iint \exp\left\{-\frac{\beta}{2}(\mathbf{t}-\mathbf{\Phi}\mathbf{w})^{\mathrm{T}}(\mathbf{t}-\mathbf{\Phi}\mathbf{w})\right\} \\
&\qquad \exp\left\{-\frac{\beta}{2}(\mathbf{w}-\mathbf{m}_0)^{\mathrm{T}}\mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0)\right\}\,\mathrm{d}\mathbf{w} \\
&\qquad \beta^{a_0-1}\beta^{N/2}\beta^{M/2} \exp(-b_0\beta)\,\mathrm{d}\beta \\
&= \frac{b_0^{a_0}}{((2\pi)^{M+N}|\mathbf{S}_0|)^{1/2}} \iint \exp\left\{-\frac{\beta}{2}(\mathbf{w}-\mathbf{m}_N)^{\mathrm{T}}\mathbf{S}_N^{-1}(\mathbf{w}-\mathbf{m}_N)\right\}\,\mathrm{d}\mathbf{w} \\
&\qquad \exp\left\{-\frac{\beta}{2}\left(\mathbf{t}^{\mathrm{T}}\mathbf{t} + \mathbf{m}_0^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{m}_0 - \mathbf{m}_N^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{m}_N\right)\right\} \\
&\qquad \beta^{a_N-1}\beta^{M/2} \exp(-b_0\beta)\,\mathrm{d}\beta
\end{aligned}
$$

where we have completed the square for the quadratic form in $\mathbf{w}$, using

$$
\begin{aligned}
\mathbf{m}_N &= \mathbf{S}_N\left[\mathbf{S}_0^{-1}\mathbf{m}_0 + \mathbf{\Phi}^{\mathrm{T}}\mathbf{t}\right] \\
\mathbf{S}_N^{-1} &= \beta\left(\mathbf{S}_0^{-1} + \mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right) \\
a_N &= a_0 + \frac{N}{2} \\
b_N &= b_0 + \frac{1}{2}\left(\mathbf{m}_0^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{m}_0 - \mathbf{m}_N^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{m}_N + \sum_{n=1}^{N} t_n^2\right).
\end{aligned}
$$

Now we are ready to do the integration, first over $\mathbf{w}$ and then β, and re-arrange the

terms to obtain the desired result

$$
\begin{aligned}
p(\mathbf{t}) &= \frac{b_0^{a_0}}{\left((2\pi)^{M+N}|\mathbf{S}_0|\right)^{1/2}}(2\pi)^{M/2}|\mathbf{S}_N|^{1/2}\int \beta^{a_N-1}\exp(-b_N\beta)\,\mathrm{d}\beta \\
&= \frac{1}{(2\pi)^{N/2}}\frac{|\mathbf{S}_N|^{1/2}}{|\mathbf{S}_0|^{1/2}}\frac{b_0^{a_0}}{b_N^{a_N}}\frac{\Gamma(a_N)}{\Gamma(a_0)}.
\end{aligned}
$$