More PRML Errata

Yousuke Takada
yousuketakada@gmail.com

November 10, 2016

Preface

This report communicates some more errata for PRML (Bishop, 2006) that are not listed in the official errata document (Svensén and Bishop, 2011)[1] at the time of this writing. When specifying the location of an error, I follow the notational conventions adopted by Svensén and Bishop (2011). I have also included in this report some suggestions for improving the readability.

Corrections

Page 51

Equation (1.98): Following the notation (1.93) for the entropy, we should write the left hand side of (1.98) as $\mathrm{H}[X]$ instead of $\mathrm{H}[p]$. As suggested in the first paragraph on Page 703, if we see $\mathrm{H}[\cdot]$ as a functional, we could write $\mathrm{H}[p]$. However, it is probably better to maintain the notational consistency here.

Page 56

Equation (1.116): In general, we cannot interpret λ_i in Jensen's inequality (1.115) as a probability distribution over a discrete variable x such that $\lambda_i \equiv p(x = x_i)$ because, since (1.115) holds for any $\{x_i\}$, we can take $x_i = x_j$ and $\lambda_i \neq \lambda_j$ where $i \neq j$, assigning different probabilities for the same value of x. Instead, we should rather introduce another set of underlying random variables $\mathbf{z}$ such that $\lambda_i \equiv p(\mathbf{z} = \mathbf{z}_i)$ and a function $\xi(\mathbf{z})$ of the random variables $\mathbf{z}$ such that $x_i \equiv \xi(\mathbf{z}_i)$, giving a more general result

$$f\left(\mathbb{E}_{\mathbf{z}}\left[\xi(\mathbf{z})\right]\right) \leqslant \mathbb{E}_{\mathbf{z}}\left[f\left(\xi(\mathbf{z})\right)\right] \tag{1}$$

where $f(\cdot)$ is a convex function but $\xi(\cdot)$ can be any. Moreover, if $f(\cdot)$ is strictly convex, the equality in (1) holds if and only if $\xi(\cdot)$ is constant. See Figure 1 for an intuitive, physical "proof" for the inequality (1). Since the random variables $\mathbf{z}$ as well as their probability $p(\mathbf{z})$ can be chosen arbitrarily, it makes sense to write $\mathbf{z}$ implicit in (1), giving a simpler form of Jensen's inequality

$$f\left(\mathbb{E}\left[\xi\right]\right) \leqslant \mathbb{E}\left[f\left(\xi\right)\right]. \tag{2}$$

For continuous random variables, we have

$$f\left(\int \xi(\mathbf{x}) p(\mathbf{x})\,\mathrm{d}\mathbf{x}\right) \leqslant \int f\left(\xi(\mathbf{x})\right) p(\mathbf{x})\,\mathrm{d}\mathbf{x} \tag{3}$$

[1] The last line but one of the bibliographic information page of the copy of PRML I have reads "9 8 7 (corrected at 6th printing 2007)." So I refer to Version 2 of the errata.

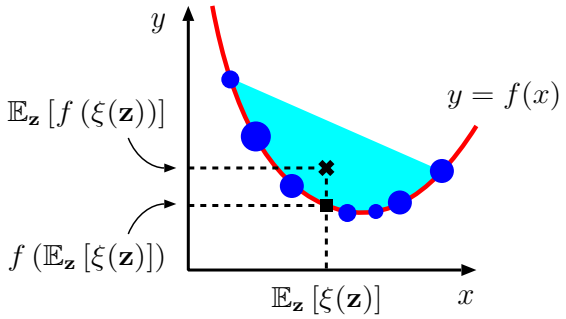

Figure 1 A physical "proof" for Jensen's inequality (MacKay, 2003). Suppose that we have a set of point masses $m_i = p(\mathbf{z} = \mathbf{z}_i)$, denoted by filled blue circles (●) with areas proportional to m_i, and place them at the corresponding locations $(x, y) = (\xi(\mathbf{z}_i), f(\xi(\mathbf{z}_i)))$ on a convex curve $y = f(x)$. The center of gravity of those masses, which is $(\mathbb{E}_{\mathbf{z}}\left[\xi(\mathbf{z})\right], \mathbb{E}_{\mathbf{z}}\left[f\left(\xi(\mathbf{z})\right)\right])$, denoted by a cross sign (×), must lie above the convex curve and thus right above the point $(\mathbb{E}_{\mathbf{z}}\left[\xi(\mathbf{z})\right], f\left(\mathbb{E}_{\mathbf{z}}\left[\xi(\mathbf{z})\right]\right))$ on the curve, denoted by a filled square (■), showing Jensen's inequality (1). One can also see that, if $f(\cdot)$ is strictly convex, the equality in (1) implies that $\xi(\cdot)$ is constant (it is trivial to show that the converse is true).

where we have used $\mathbf{x}$ to denote the underlying random variables for which we take the expectations. Making use of (3) and identifying $f(\xi) = -\ln \xi$ and $\xi(\mathbf{x}) = q(\mathbf{x})/p(\mathbf{x})$, one can show that the Kullback-Leibler divergence (1.113) satisfies $\mathrm{KL}\left(p \| q\right) \geqslant 0$ with equality if and only if $p(\mathbf{x}) = q(\mathbf{x})$.

Page 80

Equation (2.52): We usually take eigenvectors $\mathbf{u}_i$ to be the columns of $\mathbf{U}$ as in (C.37). If we follow this convention, Equation (2.52) and the following text should read

$$\mathbf{y} = \mathbf{U}^{\mathrm{T}}(\mathbf{x} - \boldsymbol{\mu}) \tag{4}$$

where $\mathbf{U}$ is a matrix whose columns are given by $\mathbf{u}_i$ so that $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_D)$. From (2.46) it follows that $\mathbf{U}$ is an *orthogonal* matrix, i.e., it satisfies $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}$ and hence also $\mathbf{U}\mathbf{U}^{\mathrm{T}} = \mathbf{I}$ where $\mathbf{I}$ is the identity matrix.

Page 81

Equations (2.53) and (2.54): If we write the change of variable from $\mathbf{x}$ to $\mathbf{y}$ as (4) instead of (2.52), the Jacobian matrix $\mathbf{J} = (J_{ij})$ is simply given by $\mathbf{U}$. Equation (2.53) should read

$$J_{ij} = \frac{\partial x_i}{\partial y_j} = U_{ij} \tag{5}$$

where U_{ij} is the (ij)-th element of $\mathbf{U}$. The square of the determinant of the Jacobian matrix (2.54) can then be evaluated as

$$|\mathbf{J}|^2 = |\mathbf{U}|^2 = \left|\mathbf{U}^{\mathrm{T}}\right| |\mathbf{U}| = \left|\mathbf{U}^{\mathrm{T}}\mathbf{U}\right| = |\mathbf{I}| = 1. \tag{6}$$

Page 81

Line −1: Since the Jacobian matrix $\mathbf{J}$ is only assumed to be orthogonal here, the determinant of $\mathbf{J}$ can be either positive or negative so that we should write $|\mathbf{J}| = \pm 1$ instead of $|\mathbf{J}| = 1$.

Page 82

Equation (2.56): We should take the absolute value of the determinant for the same reason given above; the factor $|\mathbf{J}|$ should read $|\det(\mathbf{J})|$. Note that we cannot write $||\mathbf{J}||$ to mean $|\det(\mathbf{J})|$ because

it is confusingly similar to the matrix norm $\|\mathbf{J}\|$, which usually refers to the largest singular value of $\mathbf{J}$ (Golub and Van Loan, 2013). This notational inconsistency has been caused by the abuse of the notation $|\cdot|$ for both the absolute value and the matrix determinant. If we always use $\det(\cdot)$ for the determinant, confusion will not arise and the notation be consistent. An alternative solution to this problem would be to explicitly define

$$|\mathbf{A}| \equiv |\det(\mathbf{A})| \tag{7}$$

for any square matrix $\mathbf{A}$ so that we have $|\mathbf{J}| = 1$ and (2.56) holds as is. Note also that this notation (7) is mostly consistent in other part of PRML because we have $|\mathbf{A}| = \det(\mathbf{A})$ for any positive semidefinite matrix $\mathbf{A}$ and most matrices for which we take determinants are positive (semi)definite in PRML.

Page 102

Equation (2.155): Although an interpretation for the parameters of the gamma distribution (2.146) has been given, no such an interpretation for the parameters of the Wishart distribution (2.155) is given here nor in Exercise 2.45. In order to give one, let us consider a simple Bayesian inference problem in which, given a set of N observations $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ for a zero-mean Gaussian random variable, we infer the covariance matrix $\mathbf{\Sigma}$ or, equivalently, the precision matrix $\mathbf{\Lambda} \equiv \mathbf{\Sigma}^{-1}$. The likelihood $p(\mathbf{X}|\mathbf{\Lambda})$ in terms of the precision $\mathbf{\Lambda}$ is given by

$$p(\mathbf{X}|\mathbf{\Lambda}) = \prod_{n=1}^{N} p(\mathbf{x}_n|\mathbf{\Lambda}) = \prod_{n=1}^{N} \mathcal{N}\left(\mathbf{x}_n \middle| \mathbf{0}, \mathbf{\Lambda}^{-1}\right). \tag{8}$$

If we choose the prior $p(\mathbf{\Lambda})$ over $\mathbf{\Lambda}$ to be a Wishart distribution so that

$$p(\mathbf{\Lambda}) = \mathcal{W}(\mathbf{\Lambda}|\mathbf{W}_0, \nu_0) \tag{9}$$

our analysis can be simplified because it is the conjugate prior. In fact, the posterior $p(\mathbf{\Lambda}|\mathbf{X})$ is given by

$$\begin{aligned} p(\mathbf{\Lambda}|\mathbf{X}) &\propto p(\mathbf{X}|\mathbf{\Lambda})\, p(\mathbf{\Lambda}) && (10)\\ &\propto |\mathbf{\Lambda}|^{N/2} \exp\left\{-\frac{1}{2}\sum_{n=1}^{N} \mathbf{x}_n^{\mathrm{T}} \mathbf{\Lambda} \mathbf{x}_n\right\} |\mathbf{\Lambda}|^{(\nu_0 - D - 1)/2} \exp\left\{-\frac{1}{2}\mathrm{Tr}\left(\mathbf{W}_0^{-1}\mathbf{\Lambda}\right)\right\} && (11)\\ &= |\mathbf{\Lambda}|^{(\nu_N - D - 1)/2} \exp\left\{-\frac{1}{2}\mathrm{Tr}\left(\mathbf{W}_N^{-1}\mathbf{\Lambda}\right)\right\} && (12) \end{aligned}$$

where

$$\begin{aligned} \nu_N &= \nu_0 + N && (13)\\ \mathbf{W}_N^{-1} &= \mathbf{W}_0^{-1} + \sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^{\mathrm{T}}. && (14) \end{aligned}$$

Reinstating the normalization constant, we indeed see that the posterior becomes again a Wishart distribution of the form

$$p(\mathbf{\Lambda}|\mathbf{X}) = \mathcal{W}(\mathbf{\Lambda}|\mathbf{W}_N, \nu_N). \tag{15}$$

This result suggests us how we can interpret the parameters of the Wishart distribution (2.155), namely the scale matrix $\mathbf{W}$ and the number of degrees of freedom ν. Since observing N data points increases the number of degrees of freedom ν by N, we can interpret ν_0 in the prior (9) as the number of "effective" prior observations. The N observations also contribute $N\mathbf{\Sigma}_{\mathrm{ML}}$ to the inverse of the scale matrix $\mathbf{W}$ where $\mathbf{\Sigma}_{\mathrm{ML}}$ is the maximum likelihood estimate for the covariance of the observations given by

$$\mathbf{\Sigma}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^{\mathrm{T}}. \tag{16}$$

This suggests an interpretation of $\mathbf{W}$ in terms of the "covariance" parameter

$$\boldsymbol{\Sigma} \equiv (\nu \mathbf{W})^{-1} . \tag{17}$$

More specifically, we can interpret $\boldsymbol{\Sigma}_0 = (\nu_0 \mathbf{W}_0)^{-1}$ as the covariance of the ν_0 "effective" prior observations. Note that this interpretation is in accordance with another observation that the expectation of $\boldsymbol{\Lambda}$ with respect to the prior (9) is indeed given by $\mathbb{E}[\boldsymbol{\Lambda}] = \nu_0 \mathbf{W}_0 = \boldsymbol{\Sigma}_0^{-1}$ where we have used (B.80).

Page 102

Equation (2.157): Again, no interpretation is given for the parameters of the Gaussian-Wishart distribution (2.157) nor for those of the Gaussian-gamma distribution (2.154). Since the Gaussian-gamma can be obtained as a special case of the Gaussian-Wishart where the dimension is one so that $D = 1$, we shall make an interpretation only for the parameters of the Gaussian-Wishart here. Let us consider a problem of inferring the mean $\boldsymbol{\mu}$ and the precision $\boldsymbol{\Lambda}$ given the Gaussian likelihood

$$p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{n=1}^{N} \mathcal{N}\left(\mathbf{x}_n \middle| \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}\right) \tag{18}$$

and the Gaussian-Wishart prior

$$p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \mathcal{N}\left(\boldsymbol{\mu} \middle| \boldsymbol{\mu}_0, (\beta_0 \boldsymbol{\Lambda})^{-1}\right) \mathcal{W}(\boldsymbol{\Lambda}|\mathbf{W}_0, \nu_0) . \tag{19}$$

At this moment, we introduce notations for the maximum likelihood estimates for the mean and the covariance given the N observations $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, i.e.,

$$\boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n, \qquad \boldsymbol{\Sigma}_{\mathrm{ML}} = \frac{1}{N} \sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}} \tag{20}$$

respectively. Evaluating the posterior, we have

$$\begin{aligned} p(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\mathbf{X}) &\propto p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Lambda})\, p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) && (21) \\ &\propto |\boldsymbol{\Lambda}|^{N/2} \exp\left\{-\frac{1}{2} \sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Lambda} (\mathbf{x}_n - \boldsymbol{\mu})\right\} \\ &\quad \times |\boldsymbol{\Lambda}|^{(\nu_0 - D)/2} \exp\left\{-\frac{1}{2} \mathrm{Tr}\left(\left\{\mathbf{W}_0^{-1} + \beta_0 (\boldsymbol{\mu} - \boldsymbol{\mu}_0)(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^{\mathrm{T}}\right\} \boldsymbol{\Lambda}\right)\right\} && (22) \\ &= |\boldsymbol{\Lambda}|^{(\nu_N - D)/2} \exp\left\{-\frac{1}{2} \mathrm{Tr}\left(\left\{\mathbf{W}_N^{-1} + \beta_N (\boldsymbol{\mu} - \boldsymbol{\mu}_N)(\boldsymbol{\mu} - \boldsymbol{\mu}_N)^{\mathrm{T}}\right\} \boldsymbol{\Lambda}\right)\right\} && (23) \end{aligned}$$

where

$$\begin{aligned} \beta_N &= \beta_0 + N && (24) \\ \beta_N \boldsymbol{\mu}_N &= \beta_0 \boldsymbol{\mu}_0 + N \boldsymbol{\mu}_{\mathrm{ML}} && (25) \\ \nu_N &= \nu_0 + N && (26) \\ \mathbf{W}_N^{-1} &= \mathbf{W}_0^{-1} + N \boldsymbol{\Sigma}_{\mathrm{ML}} + \frac{\beta_0 N}{\beta_0 + N} (\boldsymbol{\mu}_{\mathrm{ML}} - \boldsymbol{\mu}_0)(\boldsymbol{\mu}_{\mathrm{ML}} - \boldsymbol{\mu}_0)^{\mathrm{T}} . && (27) \end{aligned}$$

Thus, we find that the posterior is again a Gaussian-Wishart of the form

$$p(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\mathbf{X}) = \mathcal{N}\left(\boldsymbol{\mu} \middle| \boldsymbol{\mu}_N, (\beta_N \boldsymbol{\Lambda})^{-1}\right) \mathcal{W}(\boldsymbol{\Lambda}|\mathbf{W}_N, \nu_N) . \tag{28}$$

From this result, we see that the parameters β_0 and $\boldsymbol{\mu}_0$ for the mean $\boldsymbol{\mu}$ are basically (but not completely) separate from the parameters ν_0 and $\mathbf{W}_0$ for the precision $\boldsymbol{\Lambda}$. We can interpret β_0 as the number

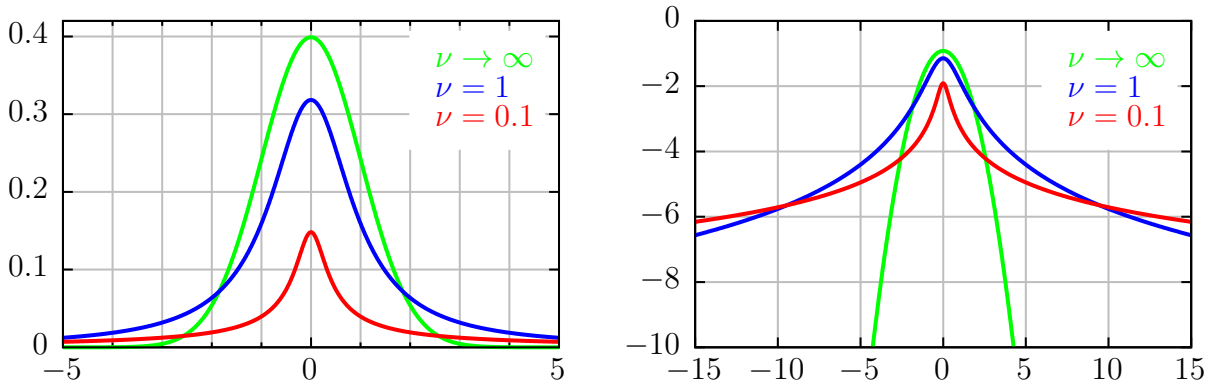

Figure 2 Plot of Student's t-distribution density functions $\mathrm{St}\left(x|\mu,\lambda,\nu\right)$ (left) and corresponding log density functions $\ln \mathrm{St}\left(x|\mu,\lambda,\nu\right)$ (right) for various values of ν where we have fixed $\mu = 0$ and $\lambda = 1$.

of "effective" prior observations for the mean and $\boldsymbol{\mu}_0$ as the mean of the β_0 prior observations. The interpretation of ν_0 and $\mathbf{W}_0$ is similar to the one we have made in the previous erratum except that we have in (27) a term due to the uncertainty in the mean, that is, the difference between the maximum likelihood mean $\boldsymbol{\mu}_{\mathrm{ML}}$ and the prior mean $\boldsymbol{\mu}_0$. Note that a similar result is obtained in Section 10.2.1 for a Bayesian mixture of Gaussians model in which we assume a Gaussian-Wishart prior for each Gaussian component.

Page 102

Line –2: "Gamma" should read "gamma" (without capitalization).

Page 103

Figure 2.15: The tails of the Student's t-distributions are too high; one can easily see that, if compared to the corresponding Gaussian distribution labeled $\nu \to \infty$, the t-distributions are not correctly normalized. Figure 2 gives the correct plot.

Page 104

The text after Equation (2.160): The Gaussian $\mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda}\right)$ should read $\mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda}^{-1}\right)$.

Page 141

Equation (3.13): The use of the gradient operator ∇ is not consistent here. As in (2.224), the gradient of a scalar function is usually defined as a column vector of derivatives so that (3.13) should read[2]

$$\nabla \ln p\left(\mathbf{t}|\mathbf{w},\beta\right) = \beta \sum_{n=1}^{N} \left\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}\left(\mathbf{x}_n\right)\right\} \boldsymbol{\phi}\left(\mathbf{x}_n\right). \tag{29}$$

Moreover, I would like to also suggest that we should give a definition for the gradient before we use it or, at least, in an appendix. Although Appendix C defines the vector derivative $\frac{\partial}{\partial \mathbf{x}}$, which is used interchangeably with the gradient $\nabla_{\mathbf{x}}$ throughout PRML, there is no mention of the gradient. We shall come back to this issue later in this report.

[2] Note that we use a different typeface for the data vector $\mathbf{t} = \left(t_1,\ldots,t_N\right)^{\mathrm{T}}$. See "Mathematical notation" on page xi of PRML.

Page 142

Equation (3.14): The left hand side should be a zero vector $\mathbf{0}$ instead of a scalar zero 0. Thus, Equation (3.14) should read

$$\mathbf{0} = \sum_{n=1}^{N} t_n \boldsymbol{\phi}\left(\mathbf{x}_n\right) - \left(\sum_{n=1}^{N} \boldsymbol{\phi}\left(\mathbf{x}_n\right) \boldsymbol{\phi}\left(\mathbf{x}_n\right)^{\mathrm{T}}\right) \mathbf{w} \tag{30}$$

where we have used the gradient of the form (29) instead of (3.13).

Page 146

Equation (3.31): The left hand side should be $\mathbf{y}\left(\mathbf{x}, \mathbf{W}\right)$ instead of $\mathbf{y}\left(\mathbf{x}, \mathbf{w}\right)$.

Page 166

The second paragraph, Line 1: "Gamma" should read "gamma" (without capitalization).

Pages 168–169, and 177

Equations (3.88), (3.93), and (3.117) as well as the text before (3.93): The derivative operators should be partial differentials. For example, Equation (3.117) should read

$$\frac{\partial}{\partial \alpha} \ln \left|\mathbf{A}\right| = \mathrm{Tr}\left(\mathbf{A}^{-1} \frac{\partial}{\partial \alpha} \mathbf{A}\right). \tag{31}$$

Page 207

Equation (4.92): The gradient and the Hessian in the right hand side, which are in general functions of the parameter $\mathbf{w}$, must be evaluated at the previous estimate $\mathbf{w}^{\text{old}}$ for the parameter. Thus, Equation (4.92) should read

$$\mathbf{w}^{\text{new}} = \mathbf{w}^{\text{old}} - \left[\mathbf{H}\left(\mathbf{w}^{\text{old}}\right)\right]^{-1} \nabla E\left(\mathbf{w}^{\text{old}}\right) \tag{32}$$

where $\mathbf{H}\left(\mathbf{w}\right) \equiv \nabla\nabla E\left(\mathbf{w}\right)$ is the Hessian matrix whose elements comprise the second derivatives of $E\left(\mathbf{w}\right)$ with respect to the components of $\mathbf{w}$.

Page 210

Equation (4.110) and the preceding text: The left hand side of (4.110) is obtained by taking the gradient of $\nabla_{\mathbf{w}_j} E$ given in (4.109) with respect to $\mathbf{w}_k$ and corresponds to the (k, j)-th block of the Hessian, *not* the (j, k)-th. Thus, Equation (4.110) should read

$$\nabla_{\mathbf{w}_k} \nabla_{\mathbf{w}_j} E\left(\mathbf{w}_1, \ldots, \mathbf{w}_K\right) = \sum_{n=1}^{N} y_{nj}\left(I_{kj} - y_{nk}\right) \boldsymbol{\phi}_n \boldsymbol{\phi}_n^{\mathrm{T}}. \tag{33}$$

To be clear, we have used the following notation. If we group all the parameters $\mathbf{w}_1, \ldots, \mathbf{w}_K$ into a column vector

$$\mathbf{w} = \begin{pmatrix} \mathbf{w}_1 \\ \vdots \\ \mathbf{w}_K \end{pmatrix} \tag{34}$$

the gradient and the Hessian of the error function $E\left(\mathbf{w}\right)$ with respect to $\mathbf{w}$ are given by

$$\nabla_{\mathbf{w}} E = \begin{pmatrix} \nabla_{\mathbf{w}_1} E \\ \vdots \\ \nabla_{\mathbf{w}_K} E \end{pmatrix}, \qquad \nabla_{\mathbf{w}} \nabla_{\mathbf{w}} E = \begin{pmatrix} \nabla_{\mathbf{w}_1} \nabla_{\mathbf{w}_1} E & \cdots & \nabla_{\mathbf{w}_1} \nabla_{\mathbf{w}_K} E \\ \vdots & \ddots & \vdots \\ \nabla_{\mathbf{w}_K} \nabla_{\mathbf{w}_1} E & \cdots & \nabla_{\mathbf{w}_K} \nabla_{\mathbf{w}_K} E \end{pmatrix} \tag{35}$$

respectively.

Page 239

Figure 5.6: The eigenvectors $\mathbf{u}_1$ and $\mathbf{u}_2$ should be unit vectors. Thus, their orientations should be shown as in Figure 2.7. Or, the scaled vectors should be labeled as $\lambda_1^{-1/2}\mathbf{u}_1$ and $\lambda_2^{-1/2}\mathbf{u}_2$.

Page 251

The second paragraph: The outer product approximation to the Hessian of the form (5.84) is usually referred to as the *Gauss-Newton* approximation (Press et al., 1992), which not only eliminates the computation of second derivatives but also guarantees that the Hessian thus approximated is positive (semi)definite, whereas the *Levenberg-Marquardt* method (Press et al., 1992) is a method that improves the numerical stability of (Gauss-)Newton type iterations by correcting the Hessian matrix so as to be more diagonal dominant. Let us now compare the two types of approximation to the Hessian, i.e., Gauss-Newton and Levenberg-Marquardt, more specifically in the following. We first observe that the Gauss-Newton approximation to the Hessian given in the right hand side of (5.84) can be written succinctly in terms of matrix product as

$$\mathbf{H}_{\text{GN}} = \mathbf{J}^{\text{T}}\mathbf{J} \tag{36}$$

where $\mathbf{J} = (\nabla a_1, \ldots, \nabla a_N)^{\text{T}}$ is the Jacobian of the activations $a_1, \ldots, a_N$ with respect to the parameters (weights and biases). The Levenberg-Marquardt approximation to the above Hessian typically takes the form

$$\mathbf{H}_{\text{LM}} = \mathbf{J}^{\text{T}}\mathbf{J} + \lambda\mathbf{I} \tag{37}$$

or

$$\mathbf{H}_{\text{LM}} = \mathbf{J}^{\text{T}}\mathbf{J} + \lambda\,\text{diag}\left(\mathbf{J}^{\text{T}}\mathbf{J}\right) \tag{38}$$

where we have introduced an adjustable damping factor $\lambda \geqslant 0$ (which will be adjusted through the iterations) and defined that, for a square matrix $\mathbf{A} = (A_{ij})$, $\text{diag}(\mathbf{A})$ is a diagonal matrix obtained by setting the non-diagonal elements equal to zero so that $\text{diag}(\mathbf{A}) = \text{diag}(A_{ii})$.

Page 275

The text after Equation (5.154): The identity matrix $\mathbf{I}$ should multiply $\sigma_k^2(\mathbf{x}_n)$.

Page 277

Equation (5.160): The factor L should multiply $\sigma_k^2(\mathbf{x})$ because we have

$$s^2(\mathbf{x}) = \mathbb{E}\left[\text{Tr}\left\{(\mathbf{t} - \mathbb{E}\left[\mathbf{t}|\mathbf{x}\right])\,(\mathbf{t} - \mathbb{E}\left[\mathbf{t}|\mathbf{x}\right])^{\text{T}}\right\}\Big|\mathbf{x}\right] \tag{39}$$

$$= \sum_{k=1}^{K} \pi_k(\mathbf{x})\,\text{Tr}\left\{\sigma_k^2(\mathbf{x})\mathbf{I} + (\boldsymbol{\mu}_k(\mathbf{x}) - \mathbb{E}\left[\mathbf{t}|\mathbf{x}\right])\,(\boldsymbol{\mu}_k(\mathbf{x}) - \mathbb{E}\left[\mathbf{t}|\mathbf{x}\right])^{\text{T}}\right\} \tag{40}$$

$$= \sum_{k=1}^{K} \pi_k(\mathbf{x})\left\{L\sigma_k^2(\mathbf{x}) + \|\boldsymbol{\mu}_k(\mathbf{x}) - \mathbb{E}\left[\mathbf{t}|\mathbf{x}\right]\|^2\right\} \tag{41}$$

where L is the dimensionality of $\mathbf{t}$.

Page 295

Line 1: The vector $\mathbf{x}$ should be a column vector so that $\mathbf{x} = (x_1, x_2)^{\text{T}}$.

Page 318

Equations (6.93) and (6.94) as well as the text before (6.93): The text and the equations should read: We can evaluate the derivative of $a_n^\star$ with respect to θ_j by differentiating the relation (6.84) with respect to θ_j to give

$$\frac{\partial \mathbf{a}_N^\star}{\partial \theta_j} = \frac{\partial \mathbf{C}_N}{\partial \theta_j}\left(\mathbf{t}_N - \boldsymbol{\sigma}_N\right) - \mathbf{C}_N \mathbf{W}_N \frac{\partial \mathbf{a}_N^\star}{\partial \theta_j} \tag{42}$$

where the derivatives are Jacobians defined by (C.16) for a vector and analogously by (116) for a matrix. Rearranging (42) then gives

$$\frac{\partial \mathbf{a}_N^\star}{\partial \theta_j} = \left(\mathbf{I} + \mathbf{C}_N \mathbf{W}_N\right)^{-1} \frac{\partial \mathbf{C}_N}{\partial \theta_j}\left(\mathbf{t}_N - \boldsymbol{\sigma}_N\right). \tag{43}$$

Page 355

Equation (7.117): The typeface of the vector $\mathbf{y}$ in (7.117) should be that in (7.110), i.e., $\mathbf{y}$.

Page 414

Figure 8.53, Line 6: The term "max-product" should be "max-sum."

Page 425

Equation (9.3): The right hand side should be a zero vector $\mathbf{0}$ instead of a scalar zero 0.

Page 432

The text after Equation (9.13): It should be noted for clarity that, as the prior $p(\mathbf{z})$ over $\mathbf{z}$ is a multinomial distribution (9.10), the posterior $p\left(\mathbf{z}|\mathbf{x}\right)$ over $\mathbf{z}$ given $\mathbf{x}$ is again a multinomial of the form

$$p\left(\mathbf{z}|\mathbf{x}\right) = \prod_{k=1}^{K} \gamma_k^{z_k} \tag{44}$$

where we have written $\gamma_k \equiv \gamma(z_k)$. This can be directly confirmed by inspecting the functional form of the joint distribution

$$p\left(\mathbf{z}\right) p\left(\mathbf{x}|\mathbf{z}\right) = \prod_{k=1}^{K} \left\{\pi_k \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)\right\}^{z_k}. \tag{45}$$

The above observation helps the reader to understand that evaluating the responsibilities $\gamma(z_k)$ indeed corresponds to the E step of the EM algorithm.

Page 434

Equation (9.15): Although the official errata (Svensén and Bishop, 2011) states that σ_j in the right hand side should be raised to a power of D, the whole right hand side should be raised to D so that Equation (9.15) should read

$$\mathcal{N}\left(\mathbf{x}_n \middle| \mathbf{x}_n, \sigma_j^2 \mathbf{I}\right) = \frac{1}{\left(2\pi\sigma_j^2\right)^{D/2}}. \tag{46}$$

Page 435

Equation (9.16): The right hand side should be a zero vector $\mathbf{0}$.

Page 465

Equations (10.6) and (10.7): In PRML, Equation (10.6) will be later recognized as "a negative Kullback-Leibler divergence between $q_j(\mathbf{Z}_j)$ and $\widetilde{p}(\mathbf{X}, \mathbf{Z}_j)$" (Page 465, Line −2). However, there is no point in taking the Kullback-Leibler divergence between two probability distributions over different sets of random variables; such a quantity is undefined. Moreover, the discussion here seems to be somewhat redundant. Without introducing an intermediate quantity like $\widetilde{p}(\mathbf{X}, \mathbf{Z}_j)$, we can rewrite (10.6) and (10.7) directly in terms of $q_j^\star(\mathbf{Z}_j)$. Specifically, writing down the terms dependent on one of the factors $q_j(\mathbf{Z}_j)$, we obtain the lower bound $\mathcal{L}(q)$ in the form

$$\mathcal{L}(q) = \int q_j(\mathbf{Z}_j)\, \mathbb{E}_{\mathbf{Z}\setminus\mathbf{Z}_j}[\ln p(\mathbf{X}, \mathbf{Z})]\, \mathrm{d}\mathbf{Z}_j - \int q_j(\mathbf{Z}_j) \ln q_j(\mathbf{Z}_j)\, \mathrm{d}\mathbf{Z}_j + \text{const} \tag{47}$$

$$= -\,\mathrm{KL}\left(q_j \middle\| q_j^\star\right) + \text{const} \tag{48}$$

where we have assumed that the expectation $\mathbb{E}_{\mathbf{Z}\setminus\mathbf{Z}_j}[\cdot]$ is taken with respect to $\mathbf{Z}$ but $\mathbf{Z}_j$ so that

$$\mathbb{E}_{\mathbf{Z}\setminus\mathbf{Z}_j}[\ln p(\mathbf{X}, \mathbf{Z})] = \int \cdots \int \ln p(\mathbf{X}, \mathbf{Z}) \prod_{i \neq j} q_i(\mathbf{Z}_i)\, \mathrm{d}\mathbf{Z}_i \tag{49}$$

and defined a new distribution $q_j^\star(\mathbf{Z}_j)$ over $\mathbf{Z}_j$ by the relation

$$\ln q_j^\star(\mathbf{Z}_j) = \mathbb{E}_{\mathbf{Z}\setminus\mathbf{Z}_j}[\ln p(\mathbf{X}, \mathbf{Z})] + \text{const}. \tag{50}$$

It directly follows from (48) that, since the lower bound $\mathcal{L}(q)$ is the negative Kullback-Leibler divergence between $q_j(\mathbf{Z}_j)$ and $q_j^\star(\mathbf{Z}_j)$ up to some additive constant, the maximum of $\mathcal{L}(q)$ occurs when $q_j(\mathbf{Z}_j) = q_j^\star(\mathbf{Z}_j)$.

Page 465

The text before Equation (10.8): The latent variable $\mathbf{z}_i$ should read $\mathbf{Z}_i$.

Page 465

Line −1: If we adopt the representation (48), the probability $\widetilde{p}(\mathbf{X}, \mathbf{Z}_j)$ should read $q_j^\star(\mathbf{Z}_j)$.

Page 466

Line 1: Again, $\widetilde{p}(\mathbf{X}, \mathbf{Z}_j)$ should read $q_j^\star(\mathbf{Z}_j)$. The sentence "Thus we obtain..." should read, e.g., "Thus we see that we have already obtained a general expression for the optimal solution in (50)."

Page 468

The text after Equation (10.16): The constant term in (10.16) is the *negative* entropy of $p(\mathbf{Z})$.

Page 478

Equation (10.63): The additive constant $+1$ on the right hand side should be omitted so that (10.63) should read

$$\nu_k = \nu_0 + N_k. \tag{51}$$

A quick check for the correctness of the re-estimation equations would be to consider the limit of $N \to 0$, in which the effective number of observations N_k also goes to zero and the re-estimation equations should reduce to identities. Equation (10.63) does not reduces to $\nu_k = \nu_0$, failing the test. Note that the solution for Exercise 10.13 given by Svensén and Bishop (2009) correctly derives the result (51).

Page 489

Equation (10.107): The expectations $\mathbb{E}_{\alpha}\left[\ln q(\mathbf{w})\right]_{\mathbf{w}}$ and $\mathbb{E}\left[\ln q(\alpha)\right]$ should read $\mathbb{E}_{\mathbf{w}}\left[\ln q(\mathbf{w})\right]$ and $\mathbb{E}_{\alpha}\left[\ln q(\alpha)\right]$, respectively, where the expectation $\mathbb{E}_{\mathbf{Z}}[\cdot]$ is taken over $q\left(\mathbf{Z}\right)$.

Page 489

Equations (10.108) through (10.112): The expectations are notationally inconsistent with (1.36); they should be of the forms shown in (10.107) or the ones corrected as above.

Page 490

The third paragraph, Line 2: A comma (,) should be inserted after the ellipsis (. . .).

Page 496

Equation (10.140): In order to be consistent with the mathematical notations in PRML, the differential operator d in (10.140) should be upright d. Moreover, the derivative of x with respect to x^2 should be written with parentheses as $\frac{\mathrm{d}x}{\mathrm{d}(x^2)}$, instead of $\frac{\mathrm{d}x}{\mathrm{d}x^2}$, to avoid ambiguity.

Page 501

The text after Equation (10.162): We have that the variational parameter $\lambda(\xi)$ is a monotonic function of ξ for $\xi \geqslant 0$, but not that its derivative $\lambda'(\xi)$ is.

Page 503

The text after Equation (10.168): A period (.) should be appended at the end of the sentence that follows (10.168).

Page 512

Equation (10.222): The factor $\left(2\pi v_n\right)^{D/2}$ in the denominator of the right hand side should be omitted because it has been already included in the Gaussian in (10.213).

Page 513

Equations (10.223) and (10.224): The quantities v^{new} and $\mathbf{m}^{\text{new}}$ in (10.223) and (10.224) are different from those in (10.217) and (10.218).[3] Thus, we should introduce different notations, say, v and $\mathbf{m}$, with appropriate definitions. Specifically, one can rewrite the approximation to the model evidence in the form

$$p(\mathcal{D}) \simeq \left(2\pi v\right)^{D/2} \exp\left(B/2\right) \prod_{n=1}^{N} \left\{ s_n \left(2\pi v_n\right)^{-D/2} \right\} \tag{52}$$

[3] See Svensén and Bishop (2011) for the errata for Equations (10.217) and (10.218).

where

$$B = \frac{\mathbf{m}^{\mathrm{T}}\mathbf{m}}{v} - \sum_{n=1}^{N} \frac{\mathbf{m}_n^{\mathrm{T}}\mathbf{m}_n}{v_n} \tag{53}$$

$$v^{-1} = \sum_{n=1}^{N} v_n^{-1} \tag{54}$$

$$v^{-1}\mathbf{m} = \sum_{n=1}^{N} v_n^{-1}\mathbf{m}_n. \tag{55}$$

Page 515

Equations (10.228) and (10.229): Although Svensén and Bishop (2011) correct (10.228) so that $q^{\backslash b}(\mathbf{x})$ is a normalized distribution, we do not need the normalization of $q^{\backslash b}(\mathbf{x})$ here and, even with this normalization, we cannot ensure that $\hat{p}(\mathbf{x})$ given by (10.229) is normalized. Similarly to (10.195), we can proceed with the unnormalized $q^{\backslash b}(\mathbf{x})$ given by the original (10.228) and, rather than correcting (10.228), we should correct (10.229) so that

$$\hat{p}(\mathbf{x}) \propto q^{\backslash b}(\mathbf{x}) f_b(x_2, x_3) = \ldots \tag{56}$$

implying that $\hat{p}(\mathbf{x})$ is a normalized distribution.

Page 515

The text after Equation (10.229): The new distribution $q^{\text{new}}(\mathbf{z})$ should read $q^{\text{new}}(\mathbf{x})$.

Page 516

Equation (10.240): The subscript k of the product $\prod_k \ldots$ should read $k \neq j$ because we have already removed the term $\widetilde{f}_j(\boldsymbol{\theta}_j)$.

Page 554

Equation (11.72), Line −2: The expectation in the last line but one of (11.72) is taken over the probability $p_G(\mathbf{z})$. This is probably better expressed in words, rather than the unclear notation like $\mathbb{E}_{G(\mathbf{z})}[\cdot]$. Specifically, the expectation should read

$$\mathbb{E}_{\mathbf{z}}\left[\exp\left(-E(\mathbf{z}) + G(\mathbf{z})\right)\right] \tag{57}$$

where we have written the argument $\mathbf{z}$ for $E(\mathbf{z})$ and $G(\mathbf{z})$ for clarity; and the text following (11.72) should read "where $\mathbb{E}_{\mathbf{z}}[\cdot]$ is taken over $p_G(\mathbf{z})$ and $\{\mathbf{z}^{(l)}\}$ are..."

Page 556

Exercise 11.7: The interval should be $[-\pi/2, \pi/2]$ instead of $[0, 1]$.

Page 557

Exercise 11.14, Line 2: The variance should be σ_i^2 instead of σ_i.

Page 564

The text after Equation (12.12): The derivative we consider here is that with respect to b_j (*not* that with respect to b_i).

Page 564

The text after Equation (12.15): The zero should be a zero vector so that we have $\mathbf{u}_j = \mathbf{0}$.

Page 575

The third paragraph, Line 5: The zero vector should be a row vector instead of a column vector so that we have $\mathrm{v}^{\mathrm{T}}\mathbf{U} = \mathbf{0}^{\mathrm{T}}$. Or, the both sides are transposed to give $\mathbf{U}^{\mathrm{T}}\mathbf{v} = \mathbf{0}$.

Page 578

Equation (12.53): As stated in the text preceding (12.53), we should substitute $\boldsymbol{\mu} = \bar{\mathbf{x}}$ into (12.53).

Page 578

The text before Equation (12.56): For the maximization with respect to $\mathbf{W}$, we use (C.25) and (C.27) instead of (C.24).

Page 579

Line 5: The eigendecomposition requires $O(D^3)$ computations (in the plural form).

Page 599

Exercise 12.1, Line −1: The quantity λ_{M+1} is an eigenvalue (not an eigenvector).

Page 602

Exercise 12.25, Line 2: The latent space distribution should read $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})$.

Page 610

The first paragraph, Line −5: The text "our predictions for $\mathbf{x}_{n+1}$ depends on..." should read: "our predictions for $\mathbf{x}_{n+1}$ depend on..." (Remove s from the verb).

Page 620

The second paragraph and the following (unlabeled) equation: The last sentence before the equation and the equation should each terminate with a period (.).

Page 621

Figures 13.12 and 13.13: It should be clarified that, similarly to $\alpha(z_{nk})$ and $\beta(z_{nk})$, the notation $p(\mathbf{x}_n|z_{nk})$ is used to denote the value of $p(\mathbf{x}_n|\mathbf{z}_n)$ when $z_{nk} = 1$.

Page 622

The second paragraph, Line −1: "we see" should be omitted.

Page 623

The first paragraph, Line −2: z_{nk} should read $z_{n-1,k}$.

Page 631

Equation (13.73): The equation should read

$$\sum_{r=1}^{R} \ln \left\{ \frac{p\left(\mathbf{X}_r|\boldsymbol{\theta}_{m_r}\right) p(m_r)}{\sum_{l=1}^{M} p\left(\mathbf{X}_r|\boldsymbol{\theta}_l\right) p(l)} \right\}. \tag{58}$$

Page 637

Equations (13.81), (13.82), and (13.83): The distribution (13.81) over $\mathbf{w}$ should read

$$p(\mathbf{w}) = \mathcal{N}\left(\mathbf{w}|\mathbf{0}, \boldsymbol{\Gamma}\right) \tag{59}$$

and so on.

Page 638

The first paragraph, Line 2: "conditional on" should read "conditioned on."

Page 641

Equation (13.104) and the preceding text: The form of the Gaussian is unclear. Since a multivariate Gaussian is usually defined over a column vector, we should construct a column vector from the concerned random variables to clearly define the mean and the covariance. Specifically, (13.104) and the preceding text should read for example: ..., we see that $\xi(\mathbf{z}_{n-1}, \mathbf{z}_n)$ is a Gaussian of the form

$$\xi(\mathbf{z}_{n-1}, \mathbf{z}_n) = \mathcal{N}\left(\begin{pmatrix} \mathbf{z}_{n-1} \\ \mathbf{z}_n \end{pmatrix} \middle| \begin{pmatrix} \hat{\boldsymbol{\mu}}_{n-1} \\ \hat{\boldsymbol{\mu}}_n \end{pmatrix}, \begin{pmatrix} \hat{\mathbf{V}}_{n-1} & \hat{\mathbf{V}}_{n-1,n} \\ \hat{\mathbf{V}}_{n-1,n}^{\mathrm{T}} & \hat{\mathbf{V}}_n \end{pmatrix} \right) \tag{60}$$

where the mean $\hat{\boldsymbol{\mu}}_n$ and the covariance $\hat{\mathbf{V}}_n$ of $\mathbf{z}_n$ are given by (13.100) and (13.101), respectively; and the covariance $\hat{\mathbf{V}}_{n-1,n}$ between $\mathbf{z}_{n-1}$ and $\mathbf{z}_n$ is given by

$$\hat{\mathbf{V}}_{n-1,n} = \mathrm{cov}\left[\mathbf{z}_{n-1}, \mathbf{z}_n\right] = \mathbf{J}_{n-1} \hat{\mathbf{V}}_n. \tag{61}$$

Pages 642 and 643

Equation (13.109) and the following equations: If we follow the notation in Chapter 9, the typeface of the Q function should be $\mathcal{Q}$.

Page 642

Equation (13.109): If we follow the notation for a conditional expectation given by (1.37), (13.109) should read

$$\mathcal{Q}\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{\mathrm{old}}\right) = \mathbb{E}_{\mathbf{Z}}\left[\ln p\left(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}\right) \middle| \mathbf{X}, \boldsymbol{\theta}^{\mathrm{old}}\right] \tag{62}$$

$$= \int \mathrm{d}\mathbf{Z}\; p\left(\mathbf{Z} \middle| \mathbf{X}, \boldsymbol{\theta}^{\mathrm{old}}\right) \ln p\left(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}\right) \tag{63}$$

which corresponds to (9.30).

Page 643

Equation (13.111): $\mathbf{V}_0^{\mathrm{new}}$ should read $\mathbf{P}_0^{\mathrm{new}}$. Svensén and Bishop (2011) have failed to mention (13.111).

Page 643

Equation (13.114): The size of the opening curly brace "{" should match that of the closing curly brace "}."

Page 647

Figure 13.23, Line −1: $p(\mathbf{x}_{n+1}|\mathbf{z}_{n+1}^{(l)})$ should read $p(\mathbf{x}_{n+1}|z_{n+1}^{(l)})$.

Page 649

Exercise 13.14, Line 1: (8.67) should be (8.64).

Page 658

Figure 14.1, the equation below: The subscript of the summation in the right hand side should read $m = 1$.

Page 668

Equation (14.37): The arguments of the probability are notationally inconsistent with those of (14.34), (14.35), and (14.36). Specifically, the conditioning on $\boldsymbol{\phi}_n$ should read that on t_n and the probability $p(k|\dots)$ be the value of $p(\mathbf{z}_n|\dots)$ when $z_{nk} = 1$, which we write $p(z_{nk} = 1|\dots)$. Moreover, strictly speaking, the old parameters $\pi_k, \mathbf{w}_k, \beta$ should read $\pi_k^{\text{old}}, \mathbf{w}_k^{\text{old}}, \beta^{\text{old}} \in \boldsymbol{\theta}^{\text{old}}$. In order to solve these problems, we should rewrite Equation (14.37) as, for example,

$$\gamma_{nk} = \mathbb{E}\left[z_{nk}\middle|t_n, \boldsymbol{\theta}^{\text{old}}\right] \tag{64}$$

where we have written the conditioning in the expectation explicitly and the expectation is given by

$$\mathbb{E}\left[z_{nk}|t_n, \boldsymbol{\theta}\right] = p\left(z_{nk} = 1|t_n, \boldsymbol{\theta}\right) = \frac{\pi_k \mathcal{N}\left(t_n\middle|\mathbf{w}_k^{\mathrm{T}}\boldsymbol{\phi}_n, \beta^{-1}\right)}{\sum_j \pi_j \mathcal{N}\left(t_n\middle|\mathbf{w}_j^{\mathrm{T}}\boldsymbol{\phi}_n, \beta^{-1}\right)}. \tag{65}$$

Page 668

The unlabeled equation between (14.37) and (14.38): If we write the implicit conditioning in the expectation explicitly (similarly to the above equations), the unlabeled equation should read

$$\mathcal{Q}\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}\right) = \mathbb{E}_{\mathbf{Z}}\left[\ln p(\mathbf{t}, \mathbf{Z}|\boldsymbol{\theta})\middle|\mathbf{t}, \boldsymbol{\theta}^{\text{old}}\right] \tag{66}$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma_{nk}\left\{\ln \pi_k + \ln \mathcal{N}\left(t_n\middle|\mathbf{w}_k^{\mathrm{T}}\boldsymbol{\phi}_n, \beta^{-1}\right)\right\}. \tag{67}$$

Page 669

Equations (14.40) and (14.41): The left hand sides should both read a zero vector $\mathbf{0}$.

Page 669

Equation (14.41): $\boldsymbol{\Phi}$ is undefined. The text following (14.41) should read for example: where $\mathbf{R}_k = \mathrm{diag}(\gamma_{nk})$ is a diagonal matrix of size $N \times N$ and $\boldsymbol{\Phi} = (\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_N)^{\mathrm{T}}$ is an $N \times M$ matrix. Here, N is the size of the data set and M is the dimensionality of the feature vectors $\boldsymbol{\phi}_n$.

Page 669

Equation (14.43): "+const" should be added to the right hand side.

Page 671

The text after Equation (14.46): The text should read: where we have omitted the dependence on $\{\boldsymbol{\phi}_n\}$ and defined $y_{nk} = \ldots$ Or, $\boldsymbol{\phi}$ should have been omitted from the left hand side of (14.45) in the first place.

Page 671

Equation (14.48): The notation should be corrected similarly to (64) and (65).

Page 671

Equation (14.49): The notation should be corrected similarly to (66).

Page 672

Equation (14.52): The negation should be removed so that the Hessian is given by $\mathbf{H}_k \equiv \nabla_k \nabla_k \mathcal{Q}$ where

$$\nabla_k \nabla_k \mathcal{Q} = -\sum_{n=1}^{N} \gamma_{nk} y_{nk}(1 - y_{nk}) \boldsymbol{\phi}_n \boldsymbol{\phi}_n^{\mathrm{T}}. \tag{68}$$

Page 674

Exercise 14.1, Line 1: "of" should be inserted after "set."

Page 686

Line −3: The comma in the first inline math should be removed so that the product should read: $m \times (m-1) \times \cdots \times 2 \times 1$.

Page 687

Equation (B.25): The differential operator d should be upright d.

Page 688

Line 1: "Gamma" should read "gamma" (without capitalization).

Page 689

Line 1: "positive-definite" should read "positive definite" (without hyphenation).

Page 689

Equation (B.49): $\mathbf{x}$ in the right hand side should read $\mathbf{x}_a$.

Page 690

Equation (B.54): When we identify the functional form of a multinomial distribution, e.g., in the posterior distribution (44) for the Gaussian mixture model (45) of Section 9.2, it is helpful to know that the distribution can also be expressed in terms of unnormalized probabilities $\widetilde{\mu}_k \geqslant 0$, i.e.,

$$\operatorname{Mult}\left(\mathbf{x}|\widetilde{\boldsymbol{\mu}}\right) = C(\widetilde{\boldsymbol{\mu}}) \prod_{k=1}^{K} \widetilde{\mu}_k^{x_k} \tag{69}$$

so that

$$\ln \operatorname{Mult}\left(\mathbf{x}|\widetilde{\boldsymbol{\mu}}\right) = \sum_{k=1}^{K} x_k \ln \widetilde{\mu}_k + \text{const} \tag{70}$$

where the normalization constant is given by $C(\widetilde{\boldsymbol{\mu}})^{-1} = \sum_k \widetilde{\mu}_k$ so that the normalized probabilities can be found as

$$\mu_k = p(x_k = 1) = \frac{\widetilde{\mu}_k}{\sum_j \widetilde{\mu}_j}. \tag{71}$$

Page 692

Equation (B.68): This form of multivariate Student's t-distribution is derived in Section 2.3.7 by marginalizing over the gamma distributed (scalar) variable η in (2.161), but *not* by marginalizing over the $D \times D$ precision matrix $\boldsymbol{\Lambda}$ that is governed by the Wishart distribution $\mathcal{W}\left(\boldsymbol{\Lambda}|\mathbf{W}, \nu\right)$ where $\mathbf{W} \succ 0$ and $\nu > D - 1$, which results in a marginal distribution of the form

$$p\left(\mathbf{x}|\boldsymbol{\mu}, \mathbf{W}, \nu\right) = \int \mathcal{N}\left(\mathbf{x}\middle|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}\right) \mathcal{W}\left(\boldsymbol{\Lambda}|\mathbf{W}, \nu\right) \mathrm{d}\boldsymbol{\Lambda}. \tag{72}$$

The above marginal distribution (72) is indeed equivalent to (B.68) with some reparameterization. However, this result is not so obvious that I would like to show it here. Note that such marginalization is also required to derive a mixture of Student's t-distributions given by (10.81) in Exercise 10.19. The key idea is that the integrand in the right hand side of (72) can be identified as an unnormalized Wishart distribution and the marginalization can be done in a symbolic manner. More specifically, we have

$$\begin{aligned} p\left(\mathbf{x}|\boldsymbol{\mu}, \mathbf{W}, \nu\right) = \int \mathrm{d}\boldsymbol{\Lambda}\, & \frac{|\boldsymbol{\Lambda}|^{1/2}}{(2\pi)^{D/2}} \exp\left\{-\frac{1}{2}\left(\mathbf{x} - \boldsymbol{\mu}\right)^{\mathrm{T}} \boldsymbol{\Lambda}\left(\mathbf{x} - \boldsymbol{\mu}\right)\right\} \\ & \times B\left(\mathbf{W}, \nu\right) |\boldsymbol{\Lambda}|^{(\nu - D - 1)/2} \exp\left\{-\frac{1}{2} \operatorname{Tr}\left(\mathbf{W}^{-1}\boldsymbol{\Lambda}\right)\right\} \qquad (73) \\ = & \frac{2^{(\nu+1)D/2} \Gamma_D\left(\frac{\nu+1}{2}\right) \left|\mathbf{W}^{-1} + \left(\mathbf{x} - \boldsymbol{\mu}\right)\left(\mathbf{x} - \boldsymbol{\mu}\right)^{\mathrm{T}}\right|^{-(\nu+1)/2}}{(2\pi)^{D/2}\, 2^{\nu D/2} \Gamma_D\left(\frac{\nu}{2}\right) |\mathbf{W}|^{\nu/2}} \qquad (74) \end{aligned}$$

where we have used the multivariate gamma function (Olver et al., 2016) given by[4]

$$\Gamma_D\left(\frac{\nu}{2}\right) = \pi^{D(D-1)/4} \prod_{i=1}^{D} \Gamma\left(\frac{\nu + 1 - i}{2}\right) \tag{76}$$

[4]The multivariate gamma function $\Gamma_D(\cdot)$ is defined by

$$\Gamma_D(a) \equiv \int_{\mathbf{X} \succ 0} |\mathbf{X}|^{a - (D+1)/2} \exp(-\operatorname{Tr}(\mathbf{X}))\, \mathrm{d}\mathbf{X} \tag{75}$$

where the integration is taken over the space of symmetric positive-definite matrices (Olver et al., 2016). One can see that, when $D = 1$, the multivariate gamma function $\Gamma_D(\cdot)$ reduces to the (univariate) gamma function $\Gamma(\cdot)$ defined by (1.141).

so that we can write the normalization constant (B.79) as

$$B(\mathbf{W},\nu)^{-1} = 2^{\nu D/2}\,|\mathbf{W}|^{\nu/2}\,\Gamma_D\left(\frac{\nu}{2}\right). \tag{77}$$

Finally, we obtain

$$p\left(\mathbf{x}|\boldsymbol{\mu},\mathbf{W},\nu\right) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu+1-D}{2}\right)}\frac{|\mathbf{W}|^{1/2}}{\pi^{D/2}}\left[1+(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\,\mathbf{W}\,(\mathbf{x}-\boldsymbol{\mu})\right]^{-(\nu+1)/2} \tag{78}$$

where we have used (76) and (C.15). Thus, we see that the marginal distribution of the form (72) is indeed a multivariate Student's t-distribution of the form (B.68) or (2.162) and they are related by

$$p\left(\mathbf{x}|\boldsymbol{\mu},\mathbf{W},\nu\right) = \mathrm{St}\left(\mathbf{x}|\boldsymbol{\mu},(\nu+1-D)\mathbf{W},\nu+1-D\right). \tag{79}$$

Having observed that the two marginal distributions (2.161) and (72), involving marginalization over η and $\boldsymbol{\Lambda}$, respectively, both result in the same multivariate Student's t-distribution (with some reparameterization), one might ask which is better than the other. I would argue that fewer latent variables are always better than more if the resulting model is the same, i.e., (2.161) is better than (72), because fewer latent variables mean less computational complexity and a tighter bound on the (marginal) likelihood when we employ the EM algorithm (see Chapter 9) or variational methods (see Chapter 10).

It is also worth noting here that, if we adopt an isotropic scale matrix $\mathbf{W} = \widetilde{W}\mathbf{I}$ where $\widetilde{W} > 0$, which is common in practice, the resulting multivariate Student's t-distribution (79) is again isotropic and the same marginal distribution can be obtained by marginalizing over a univariate Wishart (gamma) prior, i.e.,

$$\int \mathcal{N}\left(\mathbf{x}\middle|\boldsymbol{\mu},\lambda^{-1}\mathbf{I}\right)\mathcal{W}\left(\lambda\middle|\widetilde{W},\widetilde{\nu}\right)\mathrm{d}\lambda = \mathrm{St}\left(\mathbf{x}\middle|\boldsymbol{\mu},\widetilde{\nu}\widetilde{W}\mathbf{I},\widetilde{\nu}\right) \tag{80}$$

where $\widetilde{\nu} = \nu + 1 - D > 0$. Let $\widetilde{\sigma}^2 = \left(\widetilde{\nu}\widetilde{W}\right)^{-1}$ be the "covariance" parameter (17) of the univariate Wishart prior $\mathcal{W}\left(\lambda\middle|\widetilde{W},\widetilde{\nu}\right)$. The "covariance" parameter of the corresponding multivariate Wishart prior $\mathcal{W}\left(\boldsymbol{\Lambda}|\mathbf{W},\nu\right)$ is *not* equal to $\widetilde{\sigma}^2\mathbf{I}$ but given by downscaling it by a factor of $\widetilde{\nu}/(\widetilde{\nu}-1+D)$ so that

$$\boldsymbol{\Sigma} = (\nu\mathbf{W})^{-1} = \frac{\widetilde{\nu}}{\widetilde{\nu}-1+D}\widetilde{\sigma}^2\mathbf{I} \tag{81}$$

which implies $\boldsymbol{\Sigma} = (\widetilde{\sigma}^2/D)\,\mathbf{I}$ for $\widetilde{\nu} = 1$ and $\boldsymbol{\Sigma} \to \widetilde{\sigma}^2\mathbf{I}$ for $\widetilde{\nu} \to \infty$.

Page 693

Equations (B.78) through (B.82): Some appropriate citation is needed for the Wishart distribution because it has been introduced in Section 2.3.6 without any proof for the normalization constant as well as other statistics.

Page 693

Line −1: $b = 1/2W$ should read $b = 1/(2W)$ for clarity.

Page 696

Equation (C.5): Replacing $\mathbf{B}^{\mathrm{T}}$ with $\mathbf{A}$, we obtain a more general identity

$$\left(\mathbf{P}^{-1}+\mathbf{A}\mathbf{R}^{-1}\mathbf{B}\right)^{-1}\mathbf{A}\mathbf{R}^{-1} = \mathbf{P}\mathbf{A}\left(\mathbf{B}\mathbf{P}\mathbf{A}+\mathbf{R}\right)^{-1} \tag{82}$$

which is necessary to show the *push-through identity* (C.6) and also the determinant identity (C.14). As suggested in the text, the above identity (82) can be directly verified by right multiplying both sides by $(\mathbf{BPA}+\mathbf{R})$. However, I would prefer to prove the general push-through identity (82) together with the Woodbury identity (C.7) in terms of the inverse of a partitioned matrix, which we have already seen in Section 2.3.1. To this end, we first introduce a square matrix $\mathbf{M}$ that is partitioned into four submatrices so that

$$\mathbf{M}=\begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{pmatrix} \tag{83}$$

where $\mathbf{A}$ and $\mathbf{D}$ are square (but not necessarily the same dimension) and then note that $\mathbf{M}$ can be block diagonalized as

$$\begin{pmatrix}\mathbf{I} & \mathbf{O}\\ -\mathbf{CA}^{-1} & \mathbf{I}\end{pmatrix}\begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{pmatrix}\begin{pmatrix}\mathbf{I} & -\mathbf{A}^{-1}\mathbf{B}\\ \mathbf{O} & \mathbf{I}\end{pmatrix}=\begin{pmatrix}\mathbf{A} & \mathbf{O}\\ \mathbf{O} & \mathbf{M}/\mathbf{A}\end{pmatrix} \tag{84}$$

or

$$\begin{pmatrix}\mathbf{I} & -\mathbf{BD}^{-1}\\ \mathbf{O} & \mathbf{I}\end{pmatrix}\begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{pmatrix}\begin{pmatrix}\mathbf{I} & \mathbf{O}\\ -\mathbf{D}^{-1}\mathbf{C} & \mathbf{I}\end{pmatrix}=\begin{pmatrix}\mathbf{M}/\mathbf{D} & \mathbf{O}\\ \mathbf{O} & \mathbf{D}\end{pmatrix} \tag{85}$$

if $\mathbf{A}$ or $\mathbf{D}$ is nonsingular, respectively, where we have written the Schur complement of $\mathbf{M}$ with respect to $\mathbf{A}$ or $\mathbf{D}$ as

$$\mathbf{M}/\mathbf{A}\equiv\mathbf{D}-\mathbf{CA}^{-1}\mathbf{B} \tag{86}$$

or

$$\mathbf{M}/\mathbf{D}\equiv\mathbf{A}-\mathbf{BD}^{-1}\mathbf{C} \tag{87}$$

respectively.[5] The above block diagonalization identities (84) and (85) yield two versions of the inverse partitioned matrix $\mathbf{M}^{-1}$, i.e.,

$$\begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{pmatrix}^{-1}=\begin{pmatrix}\mathbf{I} & -\mathbf{A}^{-1}\mathbf{B}\\ \mathbf{O} & \mathbf{I}\end{pmatrix}\begin{pmatrix}\mathbf{A}^{-1} & \mathbf{O}\\ \mathbf{O} & (\mathbf{M}/\mathbf{A})^{-1}\end{pmatrix}\begin{pmatrix}\mathbf{I} & \mathbf{O}\\ -\mathbf{CA}^{-1} & \mathbf{I}\end{pmatrix} \tag{90}$$

$$=\begin{pmatrix}\mathbf{A}^{-1}+\mathbf{A}^{-1}\mathbf{B}\,(\mathbf{M}/\mathbf{A})^{-1}\,\mathbf{CA}^{-1} & -\mathbf{A}^{-1}\mathbf{B}\,(\mathbf{M}/\mathbf{A})^{-1}\\ -(\mathbf{M}/\mathbf{A})^{-1}\,\mathbf{CA}^{-1} & (\mathbf{M}/\mathbf{A})^{-1}\end{pmatrix} \tag{91}$$

and

$$\begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{pmatrix}^{-1}=\begin{pmatrix}\mathbf{I} & \mathbf{O}\\ -\mathbf{D}^{-1}\mathbf{C} & \mathbf{I}\end{pmatrix}\begin{pmatrix}(\mathbf{M}/\mathbf{D})^{-1} & \mathbf{O}\\ \mathbf{O} & \mathbf{D}^{-1}\end{pmatrix}\begin{pmatrix}\mathbf{I} & -\mathbf{BD}^{-1}\\ \mathbf{O} & \mathbf{I}\end{pmatrix} \tag{92}$$

$$=\begin{pmatrix}(\mathbf{M}/\mathbf{D})^{-1} & -(\mathbf{M}/\mathbf{D})^{-1}\,\mathbf{BD}^{-1}\\ -\mathbf{D}^{-1}\mathbf{C}\,(\mathbf{M}/\mathbf{D})^{-1} & \mathbf{D}^{-1}+\mathbf{D}^{-1}\mathbf{C}\,(\mathbf{M}/\mathbf{D})^{-1}\,\mathbf{BD}^{-1}\end{pmatrix} \tag{93}$$

respectively. Equating the right hand sides, we have, e.g.,

$$(\mathbf{M}/\mathbf{D})^{-1}=\mathbf{A}^{-1}+\mathbf{A}^{-1}\mathbf{B}\,(\mathbf{M}/\mathbf{A})^{-1}\,\mathbf{CA}^{-1} \tag{94}$$

and

$$-(\mathbf{M}/\mathbf{A})^{-1}\,\mathbf{CA}^{-1}=-\mathbf{D}^{-1}\mathbf{C}\,(\mathbf{M}/\mathbf{D})^{-1}\,. \tag{95}$$

[5] Note that the notation for the Schur complement is chosen to suggest that it has a flavor of division (Minka, 2000). In fact, taking the determinant on both sides of (84) and (85), we have from the definition of the determinant (C.10) that

$$\det(\mathbf{M})=\det(\mathbf{A})\det(\mathbf{M}/\mathbf{A}) \tag{88}$$

and

$$\det(\mathbf{M})=\det(\mathbf{D})\det(\mathbf{M}/\mathbf{D}) \tag{89}$$

respectively.

Substituting (86) and (87) into both sides and replacing $\mathbf{D}$ with $-\mathbf{D}$, we finally have

$$\left(\mathbf{A}+\mathbf{B}\mathbf{D}^{-1}\mathbf{C}\right)^{-1}=\mathbf{A}^{-1}-\mathbf{A}^{-1}\mathbf{B}\left(\mathbf{D}+\mathbf{C}\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\mathbf{C}\mathbf{A}^{-1} \tag{96}$$

and

$$\left(\mathbf{D}+\mathbf{C}\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\mathbf{C}\mathbf{A}^{-1}=\mathbf{D}^{-1}\mathbf{C}\left(\mathbf{A}+\mathbf{B}\mathbf{D}^{-1}\mathbf{C}\right)^{-1} \tag{97}$$

which are equivalent to (C.7) and (82), respectively.

Page 697

Equation (C.17): It is clear that the definition (C.17) of the derivative of a scalar with respect to a vector contradicts (C.16) and (C.18). The vector derivative of the form (C.17) is called the *gradient* whereas (C.18) is called the *Jacobian* (Minka, 2000). We should use a different notation, say, ∇ for the gradient to avoid ambiguity. More specifically, given a vector function $\mathbf{y}(\mathbf{x})=(y_1(\mathbf{x}),\ldots,y_M(\mathbf{x}))^{\mathrm{T}}$ where $\mathbf{x}=(x_1,\ldots,x_D)^{\mathrm{T}}$, we write the gradient of $\mathbf{y}(\mathbf{x})$ with respect to $\mathbf{x}$ as

$$\nabla_{\mathbf{x}}\mathbf{y}=\left(\frac{\partial y_j}{\partial x_i}\right)=\begin{pmatrix}\frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_M}{\partial x_1}\\ \vdots & \ddots & \vdots\\ \frac{\partial y_1}{\partial x_D} & \cdots & \frac{\partial y_M}{\partial x_D}\end{pmatrix}. \tag{98}$$

As a special case, we see that the gradient of a scalar function $y(\mathbf{x})$ with respect to a column vector $\mathbf{x}$ is again a column vector of the same dimension, corresponding to the right hand side of (C.17), i.e.,

$$\nabla_{\mathbf{x}}y=\left(\frac{\partial y}{\partial x_i}\right)=\begin{pmatrix}\frac{\partial y}{\partial x_1}\\ \vdots\\ \frac{\partial y}{\partial x_D}\end{pmatrix}. \tag{99}$$

Note also that the right hand side of the definition of the gradient (98) is identical to the transpose of the Jacobian matrix $\partial\mathbf{y}/\partial\mathbf{x}=(\partial y_i/\partial x_j)$ so that $\nabla_{\mathbf{x}}\mathbf{y}=(\partial\mathbf{y}/\partial\mathbf{x})^{\mathrm{T}}$, as a consequence of which the chain rule for the gradient is such that the intermediate gradients are built up "towards the left," i.e.,

$$\nabla_{\mathbf{x}}\mathbf{z}(\mathbf{y})=\left(\frac{\partial\mathbf{z}}{\partial\mathbf{y}}\frac{\partial\mathbf{y}}{\partial\mathbf{x}}\right)^{\mathrm{T}}=\nabla_{\mathbf{x}}\mathbf{y}\nabla_{\mathbf{y}}\mathbf{z}. \tag{100}$$

Since the chain rule (100) is handy when we compute the gradients of composite functions (see below), I would suggest that it should also be pointed out in the "(Vector and) Matrix Derivatives" section of Appendix C. At this point, one might wonder why we use the two different forms of vector derivative that are identical up to the transposed layout, i.e., the gradient $\nabla_{\mathbf{x}}\mathbf{y}$ and the Jacobian $\partial\mathbf{y}/\partial\mathbf{x}$. As Minka (2000) points out, Jacobians are useful in calculus while gradients are useful in optimization. For instance, we can write down the Taylor series expansion (up to the second order) of a scalar function $f(\mathbf{x})$ compactly in terms of the gradients as

$$f(\mathbf{x}+\Delta\mathbf{x})\simeq f(\mathbf{x})+\mathbf{g}^{\mathrm{T}}\Delta\mathbf{x}+\frac{1}{2}\Delta\mathbf{x}^{\mathrm{T}}\mathbf{H}\Delta\mathbf{x} \tag{101}$$

where $\mathbf{g}$ and $\mathbf{H}$ are the gradient vector and the Hessian matrix of $f(\mathbf{x})$, respectively, so that

$$\mathbf{g}\equiv\nabla f(\mathbf{x})=\begin{pmatrix}\frac{\partial f}{\partial x_1}\\ \vdots\\ \frac{\partial f}{\partial x_D}\end{pmatrix},\quad \mathbf{H}\equiv\nabla\nabla f(\mathbf{x})=\begin{pmatrix}\frac{\partial^2 f}{\partial x_1\partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_1\partial x_D}\\ \vdots & \ddots & \vdots\\ \frac{\partial^2 f}{\partial x_D\partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_D\partial x_D}\end{pmatrix}. \tag{102}$$

Page 697

Equation (C.19): Following the gradient notation (98), (C.19) should read

$$\nabla\left\{\mathbf{x}^{\mathrm{T}}\mathbf{a}\right\} = \nabla\left\{\mathbf{a}^{\mathrm{T}}\mathbf{x}\right\} = \mathbf{a} \tag{103}$$

where we have omitted the subscript $\mathbf{x}$ in what should be $\nabla_{\mathbf{x}}$. Some other useful identities I would suggest to include are

$$\nabla\left\{\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}\right\} = \nabla\,\mathrm{Tr}\left(\mathbf{x}\mathbf{x}^{\mathrm{T}}\mathbf{A}\right) = \left(\mathbf{A}^{\mathrm{T}} + \mathbf{A}\right)\mathbf{x} \tag{104}$$

$$\nabla\left\{\mathbf{B}\mathbf{x}\right\} = \mathbf{B}^{\mathrm{T}} \tag{105}$$

$$\nabla\left\{\varphi\mathbf{y}\right\} = \nabla\varphi\mathbf{y}^{\mathrm{T}} + \varphi\nabla\mathbf{y} \tag{106}$$

where matrices $\mathbf{A}$ and $\mathbf{B}$ are constants. Note that $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$ in (104) is a quadratic form and thus the square matrix $\mathbf{A}$ is usually taken to be symmetric so that $\mathbf{A} = \mathbf{A}^{\mathrm{T}}$, in which case we have

$$\nabla\left\{\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}\right\} = 2\mathbf{A}\mathbf{x}. \tag{107}$$

Substituting $\mathbf{A} = \mathbf{I}$ gives

$$\nabla\left\|\mathbf{x}\right\|^2 = 2\mathbf{x} \tag{108}$$

where $\left\|\mathbf{x}\right\| = \sqrt{\mathbf{x}^{\mathrm{T}}\mathbf{x}}$ is the norm of $\mathbf{x}$. We make use of the above identity (108) when, e.g., we take the gradient of a sum-of-squares error function of the form (3.12), which can be expressed in terms of the design matrix $\mathbf{\Phi}$ given by (3.16) as

$$E(\mathbf{w}) = \frac{1}{2}\left\|\mathbf{t} - \mathbf{\Phi}\mathbf{w}\right\|^2. \tag{109}$$

Taking the gradient of (109) with respect to $\mathbf{w}$, we have

$$\nabla_{\mathbf{w}}E(\mathbf{w}) = -\mathbf{\Phi}^{\mathrm{T}}\left(\mathbf{t} - \mathbf{\Phi}\mathbf{w}\right) \tag{110}$$

where we have used the identity (108) together with the chain rule (100) and the identity (105). The same result can also be obtained by first expanding the square norm in (109) and then differentiating it using the gradient identities given above. We use (106) when, e.g., we evaluate the Hessian (5.83) of a nonlinear sum-of-squares error function such as (5.82), which takes the form

$$J = \frac{1}{2}\sum_{n=1}^{N}\varepsilon_n^2 = \frac{1}{2}\left\|\boldsymbol{\varepsilon}\right\|^2 \tag{111}$$

where we have written $\boldsymbol{\varepsilon} = \left(\varepsilon_1, \ldots, \varepsilon_N\right)^{\mathrm{T}}$. The gradient and the Hessian of J are evaluated as

$$\nabla J = \sum_n \varepsilon_n \nabla\varepsilon_n = \left(\nabla\boldsymbol{\varepsilon}\right)\boldsymbol{\varepsilon} \tag{112}$$

$$\nabla\nabla J = \sum_n \nabla\varepsilon_n\left(\nabla\varepsilon_n\right)^{\mathrm{T}} + \sum_n \varepsilon_n\nabla\nabla\varepsilon_n \tag{113}$$

$$= \nabla\boldsymbol{\varepsilon}\left(\nabla\boldsymbol{\varepsilon}\right)^{\mathrm{T}} + \sum_n \varepsilon_n\nabla\nabla\varepsilon_n. \tag{114}$$

The second form of the Hessian, which, however, does not necessarily result in efficient implementation (neither does that of the gradient), can be directly obtained by using the identity

$$\nabla\left\{\mathbf{R}\boldsymbol{\varphi}\right\} = \nabla\boldsymbol{\varphi}\mathbf{R}^{\mathrm{T}} + \sum_{m=1}^{M}\varphi_m\nabla\mathbf{r}_m \tag{115}$$

where $\mathbf{R} = \left(\mathbf{r}_1, \ldots, \mathbf{r}_M\right)$ and $\boldsymbol{\varphi} = \left(\varphi_1, \ldots, \varphi_M\right)^{\mathrm{T}}$. One can see that (105) and (106) are special cases of (115).

Page 698

Equation (C.20): Although the Jacobian of a vector with respect to a vector is defined in (C.18), the Jacobian of a matrix with respect to a scalar has not been defined. The Jacobian $\partial \mathbf{A}/\partial x$ of a matrix $\mathbf{A} = (A_{ij})$ with respect to a scalar x is a matrix with the same dimensionality as $\mathbf{A}$ and is given by

$$\frac{\partial \mathbf{A}}{\partial x} = \left(\frac{\partial A_{ij}}{\partial x} \right) \tag{116}$$

which is analogous to (C.18) in that the partial derivatives are laid out according to the numerator, i.e., $\mathbf{A}$ in (116). On the other hand, the gradient (98) is such that the derivatives are laid out according to the denominator. We should also point out that, in a similar analogy, we write the gradient $\nabla_{\mathbf{A}} y$ of a scalar y with respect to a matrix $\mathbf{A}$ as

$$\nabla_{\mathbf{A}} y = \left(\frac{\partial y}{\partial A_{ij}} \right). \tag{117}$$

Page 698

Equation (C.22): For this identity to be well-defined, it is necessary that we have $\det(\mathbf{A}) > 0$. We should make this assumption clear. Or, if we adopt the notation (7) for $|\mathbf{A}|$, which I would recommend, we see that (C.22) holds for any nonsingular $\mathbf{A}$ such that $\det(\mathbf{A}) \neq 0$. The section named "Eigenvector Equation" of Appendix C gives us a hint for a proof of (C.22) where $\mathbf{A}$ is assumed to be symmetric positive definite so that $\mathbf{A} \succ 0$. Although the restricted proof outlined in PRML is indeed highly instructive, we need a more general proof because we make use of this identity, e.g., in Exercise 2.34 without the assumptions required by the restricted proof. To this end, we first show the following identity for any square matrix $\mathbf{A}$

$$\frac{\partial}{\partial x} \det(\mathbf{A}) = \mathrm{Tr}\left(\mathbf{A}^{\dagger} \frac{\partial \mathbf{A}}{\partial x} \right) \tag{118}$$

where $\mathbf{A}^{\dagger}$ is the adjugate matrix of $\mathbf{A}$. The (ij)-th element $A^{\dagger}_{ij}$ of the adjugate matrix $\mathbf{A}^{\dagger}$ is given by

$$A^{\dagger}_{ij} = (-1)^{i+j} \det\left(\mathbf{A}^{(ji)} \right) \tag{119}$$

where $\mathbf{A}^{(ij)}$ is a matrix obtained by removing the i-th row and the j-th column of $\mathbf{A}$. From the identity

$$\mathbf{A}\mathbf{A}^{\dagger} = \mathbf{A}^{\dagger}\mathbf{A} = \det(\mathbf{A})\mathbf{I} \tag{120}$$

we can write the inverse matrix $\mathbf{A}^{-1}$ in terms of the adjugate matrix $\mathbf{A}^{\dagger}$ so that

$$\mathbf{A}^{-1} = \frac{\mathbf{A}^{\dagger}}{\det(\mathbf{A})} \tag{121}$$

if $\mathbf{A}$ is nonsingular so that $\det(\mathbf{A}) \neq 0$. Note also that the above identity (120) implies

$$\det(\mathbf{A}) = \sum_k A_{ik} A^{\dagger}_{ki} = \sum_k A^{\dagger}_{jk} A_{kj} \tag{122}$$

for any i and j. Substituting this identity (122) into the left hand side of (118) and noting that, from the definition (119) of the adjugate matrix, $A^{\dagger}_{ji}$ is independent of A_{ik} nor A_{kj} for any k, we have

$$\frac{\partial}{\partial x} \det(\mathbf{A}) = \sum_{ij} \left\{ \frac{\partial}{\partial A_{ij}} \sum_k A_{ik} A^{\dagger}_{ki} \right\} \frac{\partial A_{ij}}{\partial x} = \sum_{ij} \left\{ \frac{\partial}{\partial A_{ij}} \sum_k A^{\dagger}_{jk} A_{kj} \right\} \frac{\partial A_{ij}}{\partial x} \tag{123}$$

$$= \sum_{ij} A^{\dagger}_{ji} \frac{\partial A_{ij}}{\partial x} \tag{124}$$

$$= \mathrm{Tr}\left(\frac{\partial \mathbf{A}}{\partial x} \mathbf{A}^{\dagger} \right) = \mathrm{Tr}\left(\mathbf{A}^{\dagger} \frac{\partial \mathbf{A}}{\partial x} \right) \tag{125}$$

which proves the identity (118). Making use of (118) together with (121), we can now evaluate the right hand side of (C.22) as

$$\frac{\partial}{\partial x} \ln |\mathbf{A}| = \frac{1}{\det(\mathbf{A})} \frac{\partial}{\partial x} \det(\mathbf{A}) \tag{126}$$

$$= \mathrm{Tr}\left(\mathbf{A}^{-1} \frac{\partial \mathbf{A}}{\partial x}\right) \tag{127}$$

if $\mathbf{A}$ is nonsingular so that $\det(\mathbf{A}) \neq 0$ where we have used the notation (7) for $|\mathbf{A}|$.

Page 698

Equations (C.24), (C.25), (C.26), (C.27), and (C.28): Since these derivatives are gradients of a scalar with respect to a matrix $\mathbf{A}$, the operator $\frac{\partial}{\partial \mathbf{A}}$ should read $\nabla_{\mathbf{A}}$ if we adopt the notation (117). For example, (C.28) should read

$$\nabla_{\mathbf{A}} \ln |\mathbf{A}| = \mathbf{A}^{-\mathrm{T}} \tag{128}$$

where we have used (C.4) and defined

$$\mathbf{A}^{-\mathrm{T}} \equiv \left(\mathbf{A}^{-1}\right)^{\mathrm{T}} = \left(\mathbf{A}^{\mathrm{T}}\right)^{-1}. \tag{129}$$

In addition to the above mentioned identities, I would suggest to include the following

$$\nabla_{\mathbf{A}} \mathrm{Tr}\left(\mathbf{A}\mathbf{B}\mathbf{A}^{\mathrm{T}}\mathbf{C}\right) = \mathbf{C}^{\mathrm{T}}\mathbf{A}\mathbf{B}^{\mathrm{T}} + \mathbf{C}\mathbf{A}\mathbf{B} \tag{130}$$

$$\nabla_{\mathbf{A}} \mathrm{Tr}\left(\mathbf{A}^{-1}\mathbf{B}\right) = -\mathbf{A}^{-\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{A}^{-\mathrm{T}}. \tag{131}$$

We use the identities (130) and (131), e.g., when we show (13.113) in Exercise 13.33 and (2.122) in Exercise 2.34, respectively. It should also be noted that (C.27) is a special case of (130).

Page 700

The second paragraph, Line −1: The determinant of the orthogonal matrix $\mathbf{U}$ can be either positive or negative so that we should write $\det(\mathbf{U}) = \pm 1$, which is equivalent to $|\mathbf{U}| = 1$ if we adopt the notation (7). Although it is possible to take $\mathbf{U}$ such that $\det(\mathbf{U}) = 1$ (one can flip the sign of $\det(\mathbf{U})$ by, say, flipping the sign of any one of the eigenvectors $\{\mathbf{u}_i\}$), there is no point in doing so in practice theoretically nor numerically. In fact, most software implementations of eigenvalue decomposition only guarantee that $\mathbf{U}$ is orthogonal, i.e., $\det(\mathbf{U}) = \pm 1$. In a special case of $\mathbf{A}$ being symmetric positive (semi)definite or $\mathbf{A} \succeq 0$, we can identify the eigenvalue decomposition (C.43) with the *singular value decomposition* or SVD (Press et al., 1992; Golub and Van Loan, 2013)

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}} \tag{132}$$

where we have $\mathbf{U} = \mathbf{V}$ and also identify the eigenvalues $\{\lambda_i\}$ with the singular values $\{\sigma_i\}$ so that we can use an SVD routine for eigenvalue decomposition. Note that the singular values are usually arranged in descending order in the diagonal matrix $\mathbf{\Sigma} = \mathrm{diag}(\sigma_i)$ so that $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant 0$.

Page 700

The text following (C.41): The multiplication by $\mathbf{U}$ can be interpreted as a rotation, a reflection, or a combination of the two.

Page 705

Equation (D.8): It would be helpful if we make it clear that the left hand side of (D.8) corresponds to the functional derivative so that we should modify (D.8) as

$$\frac{\delta F}{\delta y(x)} \equiv \frac{\partial G}{\partial y} - \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial G}{\partial y'}\right) = 0. \tag{133}$$

Page 705

The last paragraph: Despite the statement, it is not that straightforward to extend the results obtained here to higher dimensions. Although such an extension is not required in PRML, it is useful when we analyze a particular type of constrained optimization problem commonly found in computer vision applications such as *optical flow* (Horn and Schunck, 1981). Here, I would like to consider an extension of the calculus of variations to a system of D-dimensional Cartesian coordinates $\mathbf{x} = (x_1, \ldots, x_D)^{\mathrm{T}} \in \mathbb{R}^D$ and find the form of the functional derivative as well as a more general boundary condition for such a derivative to be well-defined. To this end, we first review some identities concerning the *divergence* (Feynman et al., 1964). The divergence of a vector field $\mathbf{p}(\mathbf{x}) = (p_1(\mathbf{x}), \ldots, p_D(\mathbf{x}))^{\mathrm{T}} \in \mathbb{R}^D$ is a scalar field of the form

$$\operatorname{div} \mathbf{p} = \sum_i \frac{\partial p_i}{\partial x_i} \equiv \nabla \cdot \mathbf{p} \tag{134}$$

where we have omitted the coordinates $\mathbf{x}$ in the function arguments to keep the notation uncluttered. For a differentiable vector field $\mathbf{p}(\mathbf{x})$ defined on some volume $\Omega \subset \mathbb{R}^D$, the *divergence theorem* (Feynman et al., 1964) states that

$$\int_\Omega \operatorname{div} \mathbf{p} \, \mathrm{d}V = \oint_{\partial\Omega} \mathbf{p} \cdot \mathbf{n} \, \mathrm{d}S \tag{135}$$

where the left hand side is the volume integral over the volume Ω; the right hand side is the surface integral over its boundary $\partial\Omega$; and $\mathbf{n}(\mathbf{x})$ is the outward unit normal vector of $\partial\Omega$. Assuming that the coordinates $\mathbf{x} = (x_1, \ldots, x_D)^{\mathrm{T}}$ are Cartesian, we can write the volume element as $\mathrm{d}V = \mathrm{d}x_1 \cdots \mathrm{d}x_D \equiv \mathrm{d}\mathbf{x}$ and the inner product as $\mathbf{p} \cdot \mathbf{n} = \mathbf{p}^{\mathrm{T}}\mathbf{n}$. Making use of the divergence theorem (135) together with the following identity

$$\operatorname{div}(\varphi \mathbf{p}) = \nabla\varphi^{\mathrm{T}}\mathbf{p} + \varphi \operatorname{div} \mathbf{p} \tag{136}$$

we obtain a multidimensional version of the "integration by parts" formula

$$\int_\Omega \nabla\varphi^{\mathrm{T}}\mathbf{p} \, \mathrm{d}\mathbf{x} = \oint_{\partial\Omega} \varphi\mathbf{p}^{\mathrm{T}}\mathbf{n} \, \mathrm{d}S - \int_\Omega \varphi \operatorname{div} \mathbf{p} \, \mathrm{d}\mathbf{x}. \tag{137}$$

Let us now consider a functional of the form

$$E[u(\mathbf{x})] = \int_\Omega L(\mathbf{x}, u(\mathbf{x}), \nabla u(\mathbf{x})) \, \mathrm{d}\mathbf{x} \tag{138}$$

where $u(\mathbf{x}) \in \mathbb{R}$ is a function (scalar field) defined over some volume $\Omega \subset \mathbb{R}^D$ and $L(\mathbf{x}, f, \mathbf{g}) \in \mathbb{R}$ is a function of $\mathbf{x} \in \Omega$, $f \in \mathbb{R}$, and $\mathbf{g} \in \mathbb{R}^D$. Thus, the functional $E[u(\mathbf{x})] \in \mathbb{R}$ maps $u(\mathbf{x})$ to a real number. As in the ordinary calculus, we can define the derivative of a functional according to the *calculus of variations* (Feynman et al., 1964; Bishop, 2006). In order to find the form of the functional derivative, we consider how $E[u(\mathbf{x})]$ varies upon a small change $\epsilon\eta(\mathbf{x})$ in $u(\mathbf{x})$ where $\eta(\mathbf{x})$ is the "direction" of the change and ϵ is some small constant. The first-order variation of $E[u(\mathbf{x})]$ in the direction of $\eta(\mathbf{x})$ can be evaluated as

$$\delta E[u; \eta] \equiv \lim_{\epsilon \to 0} \frac{1}{\epsilon} \{E[u + \epsilon\eta] - E[u]\} \tag{139}$$

$$= \lim_{\epsilon \to 0} \frac{1}{\epsilon} \int_\Omega \{L(\mathbf{x}, u + \epsilon\eta, \nabla(u + \epsilon\eta)) - L(\mathbf{x}, u, \nabla u)\} \, \mathrm{d}\mathbf{x} \tag{140}$$

$$= \int_\Omega \left\{ \eta \frac{\partial L}{\partial f} + \nabla\eta^{\mathrm{T}} \nabla_{\mathbf{g}} L \right\} \mathrm{d}\mathbf{x} \tag{141}$$

where we have assumed that $L(\mathbf{x}, f, \mathbf{g})$ is differentiable with respect to both f and $\mathbf{g}$; and we have written

$$\frac{\partial L}{\partial f} \equiv \frac{\partial}{\partial f} L\left(\mathbf{x}, u, \nabla u\right), \qquad \nabla_{\mathbf{g}} L \equiv \nabla_{\mathbf{g}} L\left(\mathbf{x}, u, \nabla u\right). \tag{142}$$

By making use of the multidimensional integration by parts (137), we can integrate the second term in the right hand side of (141), giving

$$\delta E[u;\eta] = \int_{\Omega} \eta \left\{ \frac{\partial L}{\partial f} - \operatorname{div}\left(\nabla_{\mathbf{g}} L\right) \right\} \mathrm{d}\mathbf{x} + \oint_{\partial\Omega} \eta \nabla_{\mathbf{g}} L^{\mathrm{T}} \mathbf{n} \, \mathrm{d}S. \tag{143}$$

In order for the functional derivative to be well-defined, we assume the surface integral term in the variation (143) to vanish so that we have the following boundary condition

$$\oint_{\partial\Omega} \eta \nabla_{\mathbf{g}} L^{\mathrm{T}} \mathbf{n} \, \mathrm{d}S = 0. \tag{144}$$

The boundary condition (144) holds if

$$\eta(\mathbf{x}) = 0 \tag{145}$$

or

$$\nabla_{\mathbf{g}} L^{\mathrm{T}} \mathbf{n}(\mathbf{x}) = 0 \tag{146}$$

for all $\mathbf{x} \in \partial\Omega$. The first condition (145) holds if we assume the *Dirichlet boundary condition* for $u(\mathbf{x})$

$$u(\mathbf{x}) = u_0(\mathbf{x}) \tag{147}$$

where $\mathbf{x} \in \partial\Omega$, i.e., $u(\mathbf{x})$ is assumed to be fixed to some value $u_0(\mathbf{x})$ at the boundary $\partial\Omega$ and so is $u(\mathbf{x}) + \epsilon\eta(\mathbf{x})$ in (139), implying (145). Another common boundary condition for $u(\mathbf{x})$ is the *Neumann boundary condition*

$$\nabla u(\mathbf{x})^{\mathrm{T}} \mathbf{n}(\mathbf{x}) = 0 \tag{148}$$

where $\mathbf{x} \in \partial\Omega$. The Neumann boundary condition (148) is implied by the second condition (146) for the optical-flow energy functional as we shall see shortly. Having assumed that the boundary condition (144) holds, we can write the first order variation (143) in the form

$$\delta E[u;\eta] = \int_{\Omega} \eta \frac{\partial E}{\partial u(\mathbf{x})} \mathrm{d}\mathbf{x} \tag{149}$$

where we have written

$$\frac{\partial E}{\partial u(\mathbf{x})} \equiv \frac{\partial L}{\partial f} - \operatorname{div}\left(\nabla_{\mathbf{g}} L\right). \tag{150}$$

The volume integral in the right hand side of (149) can be seen as the inner product between $\eta(\mathbf{x})$ and $\partial E/\partial u(\mathbf{x})$, from which we conclude that the quantity $\partial E/\partial u(\mathbf{x})$ is what should be called the functional derivative.[6] A stationary point of a functional $E[u(\mathbf{x})]$ is a function $u(\mathbf{x})$ such that the variation $\delta E[u;\eta]$ vanishes in any direction $\eta(\mathbf{x})$ and thus satisfies the *Euler-Lagrange equation* given by

$$\frac{\partial E}{\partial u(\mathbf{x})} = 0. \tag{151}$$

Finally, we present an application of the multidimensional calculus of variations to a dense motion analysis technique called optical flow in the following. Suppose that, given a pair of (grayscale)

[6] Here we use a notation for the functional derivative that is different from the one used in PRML. The notation $\partial E/\partial u(\mathbf{x})$ employed here is more like an ordinary derivative and can be extended to the case of a vector field $\mathbf{u}(\mathbf{x})$ analogously to the gradient as we shall see in (156).

images $I_0(\mathbf{x})$ and $I_1(\mathbf{x})$ where $\mathbf{x} \in \mathbb{R}^2$ that are taken at some discrete time steps $t = 0$ and $t = 1$, respectively, we wish to find a motion vector field from $I_0(\mathbf{x})$ to $I_1(\mathbf{x})$

$$\mathbf{u}(\mathbf{x}) = \begin{pmatrix} u(\mathbf{x}) \\ v(\mathbf{x}) \end{pmatrix} \tag{152}$$

defined over $\mathbf{x} \in \Omega \subset \mathbb{R}^2$. Horn and Schunck (1981) sought for $\mathbf{u}(\mathbf{x})$ that minimizes an energy functional that takes essentially the same form as

$$J[\mathbf{u}(\mathbf{x})] = J_{\text{data}}[\mathbf{u}(\mathbf{x})] + \alpha J_{\text{smooth}}[\mathbf{u}(\mathbf{x})] \tag{153}$$

where

$$J_{\text{data}}[\mathbf{u}(\mathbf{x})] = \frac{1}{2} \int_\Omega \left(I_1(\mathbf{x} + \mathbf{u}(\mathbf{x})) - I_0(\mathbf{x}) \right)^2 \mathrm{d}\mathbf{x} \tag{154}$$

$$J_{\text{smooth}}[\mathbf{u}(\mathbf{x})] = \frac{1}{2} \int_\Omega \left(\|\nabla u(\mathbf{x})\|^2 + \|\nabla v(\mathbf{x})\|^2 \right) \mathrm{d}\mathbf{x}. \tag{155}$$

Here, the domain Ω is assumed to be continuous and is typically rectangular. We call the first term $J_{\text{data}}[\mathbf{u}(\mathbf{x})]$ in (153) the data-fidelity term; the second term $J_{\text{smooth}}[\mathbf{u}(\mathbf{x})]$ the smoothness (regularization) term; and the coefficient α the regularization parameter. According to the multidimensional calculus of variations, a stationary point of the optical-flow energy functional (153) satisfies Euler-Lagrange equations of the form

$$\nabla_{\mathbf{u}(\mathbf{x})} J \equiv \begin{pmatrix} \partial J / \partial u(\mathbf{x}) \\ \partial J / \partial v(\mathbf{x}) \end{pmatrix} = \nabla_{\mathbf{u}} \left\{ \frac{\varepsilon\left(\mathbf{x}, \mathbf{u}(\mathbf{x})\right)^2}{2} \right\} - \alpha \begin{pmatrix} \operatorname{div}\left(\nabla u(\mathbf{x})\right) \\ \operatorname{div}\left(\nabla v(\mathbf{x})\right) \end{pmatrix} = \mathbf{0} \tag{156}$$

where we have written

$$\varepsilon\left(\mathbf{x}, \mathbf{u}\right) = I_1(\mathbf{x} + \mathbf{u}) - I_0(\mathbf{x}). \tag{157}$$

For the functional derivatives $\partial J / \partial u(\mathbf{x})$ and $\partial J / \partial v(\mathbf{x})$ to be well-defined, let us assume the boundary condition given by (146) for each functional derivative, which implies the Neumann boundary condition for $\mathbf{u}(\mathbf{x})$, i.e.,

$$\nabla u(\mathbf{x})^{\mathrm{T}} \mathbf{n}(\mathbf{x}) = 0, \qquad \nabla v(\mathbf{x})^{\mathrm{T}} \mathbf{n}(\mathbf{x}) = 0 \tag{158}$$

for all $\mathbf{x} \in \partial\Omega$ where $\partial\Omega$ is the boundary of Ω and $\mathbf{n}(\mathbf{x})$ is the outward unit normal vector of $\partial\Omega$. Thus, solving the above Euler-Lagrange equations (156) with the Neumann boundary condition (158), we obtain the desired motion vector field $\mathbf{u}(\mathbf{x})$. The Euler-Lagrange equations given by (156) are *elliptic partial differential equations* (elliptic PDEs) and can be solved numerically by a type of relaxation method such as the Gauss-Seidel method or the (weighted) Jacobi method or by a more efficient *multigrid* technique (Press et al., 1992; Briggs et al., 2000).

Page 716

Column 1, Entry −1: "The Feynman Lectures of Physics" should read "The Feynman Lectures on Physics."

Page 717

Column 2, Entry 7: "John Hopkins University Press" should read "The Johns Hopkins University Press."

References

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer. 1, 23

Briggs, W. L., V. E. Henson, and S. F. McCormick (2000). *A Multigrid Tutorial* (Second ed.). SIAM. 25

Feynman, R. P., R. B. Leighton, and M. Sands (1964). *The Feynman Lectures on Physics*, Volume 2. Addison-Wesley. 23

Golub, G. H. and C. F. Van Loan (2013). *Matrix Computations* (Fourth ed.). The Johns Hopkins University Press. 3, 22

Horn, B. K. P. and B. G. Schunck (1981). Determining optical flow. *Artificial Intelligence 17*(1), 185–203. 23, 25

MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press. 2

Minka, T. P. (2000). Old and new matrix algebra useful for statistics. https://research.microsoft.com/en-us/um/people/minka/papers/matrix/. 18, 19

Olver, F. W. J., A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, and B. V. Saunders (2016). NIST digital library of mathematical functions. http://dlmf.nist.gov/. 16

Press, W. M., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery (1992). *Numerical Recipes in C: The Art of Scientific Computing* (Second ed.). Cambridge University Press. 7, 22, 25

Svensén, M. and C. M. Bishop (2009). Pattern recognition and machine learning: Solutions to the exercises (web edition). https://www.microsoft.com/en-us/research/people/cmbishop/#prml-book. 9

Svensén, M. and C. M. Bishop (2011). Pattern recognition and machine learning: Errata and additional comments (version 2). https://www.microsoft.com/en-us/research/people/cmbishop/#prml-book. 1, 8, 10, 11, 13