Appendix A: Basic vector and matrix operations

9.6 Vector operations

Vector addition: The addition of two N dimensional vectors

$$\mathbf{a}_{N\times 1} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{bmatrix} \quad \text{and} \quad \mathbf{b}_{N\times 1} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_N \end{bmatrix}, \tag{9.30}$$

is defined as the entry-wise sum of $\mathbf{a}$ and $\mathbf{b}$, resulting in a vector of the same dimension denoted by

$$\mathbf{a} + \mathbf{b} = \begin{bmatrix} a_1 + b_1 \\ a_2 + b_2 \\ \vdots \\ a_N + b_N \end{bmatrix}. \tag{9.31}$$

Subtraction of two vectors is defined in a similar fashion.

Vector multiplication by a scalar: Multiplying a vector $\mathbf{a}$ by a scalar γ returns a vector of the same dimension whose every element is multiplied by γ

$$\gamma\mathbf{a} = \begin{bmatrix} \gamma a_1 \\ \gamma a_2 \\ \vdots \\ \gamma a_N \end{bmatrix}. \tag{9.32}$$

Vector transpose: The transpose of a *column vector* $\mathbf{a}$ (with vertically stored elements) is a *row vector* with the same elements which are now stored horizontally, denoted by

$$\mathbf{a}^T = \begin{bmatrix} a_1 & a_2 & \cdots & a_N \end{bmatrix}. \tag{9.33}$$

Similarly the transpose of a row vector is a column vector, and we have in general that $\left(\mathbf{a}^T\right)^T = \mathbf{a}$.

Inner product of two vectors: The inner product (or dot product) of two vectors $\mathbf{a}$ and $\mathbf{b}$ (of the same dimensions) is simply the sum of their component-wise product, written as

$$\mathbf{a}^T\mathbf{b} = \sum_{n=1}^{N} a_n b_n. \tag{9.34}$$

The inner product of $\mathbf{a}$ and $\mathbf{b}$ is also often written as $\langle \mathbf{a}, \mathbf{b} \rangle$.

Inner product rule and correlation: the inner product rule between two vectors provides a measurement of

$$\mathbf{a}^T\mathbf{b} = \|\mathbf{a}\|_2 \|\mathbf{b}\|_2 \cos(\theta) \tag{9.35}$$

where θ is the angle between the vectors $\mathbf{a}$ and $\mathbf{b}$. Dividing both vectors by their length gives the *correlation* between the two vectors

$$\frac{\mathbf{a}^T\mathbf{b}}{\|\mathbf{a}\|_2 \|\mathbf{b}\|_2} = \cos(\theta) \tag{9.36}$$

which ranges between -1 and 1 (when the vectors point in completely opposite or parallel directions respectively).

Outer product of two vectors: The outer product of two vectors $\mathbf{a}_{N\times 1}$ and $\mathbf{b}_{M\times 1}$ is an $N \times M$ matrix $\mathbf{C}$ defined as

$$\mathbf{C} = \begin{bmatrix} a_1b_1 & a_1b_2 & \cdots & a_1b_M \\ a_2b_1 & a_2b_2 & \cdots & a_2b_M \\ \vdots & \vdots & \ddots & \vdots \\ a_Nb_1 & a_Nb_2 & \cdots & a_Nb_M \end{bmatrix}. \tag{9.37}$$

The outer product of $\mathbf{a}$ and $\mathbf{b}$ is also written as $\mathbf{a}\mathbf{b}^T$. Unlike the inner product, the outer product does not hold the commutative property meaning that the outer product of $\mathbf{a}$ and $\mathbf{b}$ does *not* necessarily equal the outer product of $\mathbf{b}$ and $\mathbf{a}$.

9.7 Matrix operations

Matrix addition: The addition of two $N \times M$ matrices

$$\mathbf{A}_{N\times M} = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,M} \\ a_{2,1} & a_{2,2} & \cdots & a_{2.M} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N,1} & a_{N,2} & \cdots & a_{N,M} \end{bmatrix} \quad \text{and} \quad \mathbf{B}_{N\times M} = \begin{bmatrix} b_{1,1} & b_{1,2} & \cdots & b_{1,M} \\ b_{2,1} & b_{2,2} & \cdots & b_{2.M} \\ \vdots & \vdots & \ddots & \vdots \\ b_{N,1} & b_{N,2} & \cdots & b_{N,M} \end{bmatrix}, \quad (9.38)$$

is defined again as the entry-wise sum of $\mathbf{A}$ and $\mathbf{B}$, given as

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} a_{1,1} + b_{1,1} & a_{1,2} + b_{1,2} & \cdots & a_{1,M} + b_{1,M} \\ a_{2,1} + b_{2,1} & a_{2,2} + b_{2,2} & \cdots & a_{2.M} + b_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N,1} + b_{N,1} & a_{N,2} + b_{N,2} & \cdots & a_{N,M} + b_{N,M} \end{bmatrix}. \quad (9.39)$$

Subtraction of two matrices is defined in a similar fashion.

Matrix multiplication by a scalar: Multiplying a matrix $\mathbf{A}$ by a scalar γ returns a matrix of the same dimension whose every element is multiplied by γ

$$\gamma\mathbf{A} = \begin{bmatrix} \gamma a_{1,1} & \gamma a_{1,2} & \cdots & \gamma a_{1,M} \\ \gamma a_{2,1} & \gamma a_{2,2} & \cdots & \gamma a_{2.M} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma a_{N,1} & \gamma a_{N,2} & \cdots & \gamma a_{N,M} \end{bmatrix}. \quad (9.40)$$

Matrix transpose: The transpose of an $N \times M$ matrix $\mathbf{A}$ is formed by putting the transpose of each column of $\mathbf{A}$ into the corresponding row of $\mathbf{A}^T$, giving the $M \times N$ transpose matrix as

$$\mathbf{A}^T = \begin{bmatrix} a_{1,1} & a_{2,1} & \cdots & a_{N,1} \\ a_{1,2} & a_{2,2} & \cdots & a_{N,2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1,M} & a_{2,M} & \cdots & a_{M,N} \end{bmatrix}. \quad (9.41)$$

Again, we have $\left(\mathbf{A}^T\right)^T = \mathbf{A}$.

Matrix multiplication: The product of two matrices $\mathbf{A}_{N\times M}$ and $\mathbf{B}_{M\times P}$ is an $N \times P$ matrix defined via the sum of M outer product matrices, as

$$\mathbf{C} = \mathbf{AB} = \sum_{m=1}^{M} \mathbf{a}_m \mathbf{b}^m, \quad (9.42)$$

where $\mathbf{a}_m$ and $\mathbf{b}^m$ respectively denote the m^{th} column of $\mathbf{A}$ and the m^{th} row of $\mathbf{B}$. The p^{th} column of $\mathbf{C}$ can be found via multiplying $\mathbf{A}$ by the p^{th} column of $\mathbf{B}$

$$\mathbf{c}_p = \mathbf{A}\mathbf{b}_p = \sum_{m=1}^{M} \mathbf{a}_m b_{m,p}. \tag{9.43}$$

The n^{th} row of $\mathbf{C}$ can be found via multiplying the n^{th} row of $\mathbf{A}$ by $\mathbf{B}$

$$\mathbf{c}^n = \mathbf{a}^n\mathbf{B} = \sum_{m=1}^{M} a_{n,m}\mathbf{b}^m. \tag{9.44}$$

The $(n,\, p)^{\text{th}}$ entry of $\mathbf{C}$ is found by multiplying the n^{th} row of $\mathbf{A}$ by the p^{th} column of $\mathbf{B}$

$$c_{n,p} = \mathbf{a}^n\mathbf{b}_p. \tag{9.45}$$

Note that vector inner and outer products are special cases of matrix multiplication.

Entry-wise product: The entry-wise product (or Hadamard product) of two matrices $\mathbf{A}_{N\times M}$ and $\mathbf{B}_{N\times M}$ is defined as

$$\mathbf{A}\circ\mathbf{B} = \begin{bmatrix} a_{1,1}b_{1,1} & a_{1,2}b_{1,2} & \cdots & a_{1,M}b_{1,M} \\ a_{2,1}b_{2,1} & a_{2,2}b_{2,2} & \cdots & a_{2.M}b_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N,1}b_{N,1} & a_{N,2}b_{N,2} & \cdots & a_{N,M}b_{N,M} \end{bmatrix}. \tag{9.46}$$

In other words, the $(n,\, m)^{\text{th}}$ entry of $\mathbf{A}\circ\mathbf{B}$ is simply the product of the $(n,\, m)^{\text{th}}$ entry of $\mathbf{A}$ and the $(n,\, m)^{\text{th}}$ entry of $\mathbf{B}$.

Appendix B: Basics of vector calculus

9.8 Basic definitions

Throughout this section suppose that $g(\mathbf{w})$ is a scalar valued function of the $N \times 1$ vector $\mathbf{w} = \begin{bmatrix} w_1 & w_2 & \cdots & w_N \end{bmatrix}^T$.

A *partial derivative* is the derivative of a multivariable function with respect to one of its variables. For instance, the partial derivative of g with respect to w_i is written as

$$\frac{\partial}{\partial w_i} g(\mathbf{w}). \tag{9.47}$$

The *gradient* of g is then the vector of all partial derivatives denoted as

$$\nabla g(\mathbf{w}) = \begin{bmatrix} \frac{\partial}{\partial w_1} g(\mathbf{w}) \\ \frac{\partial}{\partial w_2} g(\mathbf{w}) \\ \vdots \\ \frac{\partial}{\partial w_N} g(\mathbf{w}) \end{bmatrix}. \tag{9.48}$$

For example, the gradient for the linear function $g_1(\mathbf{w}) = \mathbf{w}^T\mathbf{b}$ and quadratic function $g_2(\mathbf{w}) = \mathbf{w}^T\mathbf{A}\mathbf{w}$ can be computed as $\nabla g_1(\mathbf{w}) = \mathbf{b}$ and $\nabla g_2(\mathbf{w}) = \left(\mathbf{A} + \mathbf{A}^T\right)\mathbf{w}$, respectively.

The *second order partial derivative* of g with respect to variables w_i and w_j is written as

$$\frac{\partial^2}{\partial w_i \partial w_j} g(\mathbf{w}), \tag{9.49}$$

or equivalently as

$$\frac{\partial^2}{\partial w_j \partial w_i} g(\mathbf{w}). \tag{9.50}$$

The *Hessian* of g is then the square symmetric matrix of all second order partial derivatives

of g, denoted as

$$\nabla^2 g\left(\mathbf{w}\right) = \begin{bmatrix} \frac{\partial^2}{\partial w_1 \partial w_1} g\left(\mathbf{w}\right) & \frac{\partial^2}{\partial w_1 \partial w_2} g\left(\mathbf{w}\right) & \cdots & \frac{\partial^2}{\partial w_1 \partial w_N} g\left(\mathbf{w}\right) \\ \frac{\partial^2}{\partial w_2 \partial w_1} g\left(\mathbf{w}\right) & \frac{\partial^2}{\partial w_2 \partial w_2} g\left(\mathbf{w}\right) & \cdots & \frac{\partial^2}{\partial w_2 \partial w_N} g\left(\mathbf{w}\right) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial w_N \partial w_1} g\left(\mathbf{w}\right) & \frac{\partial^2}{\partial w_N \partial w_2} g\left(\mathbf{w}\right) & \cdots & \frac{\partial^2}{\partial w_N \partial w_N} g\left(\mathbf{w}\right) \end{bmatrix}. \tag{9.51}$$

9.9 Commonly used rules for computing derivatives

Here we give five rules commonly used when making gradient and Hessian calculations.

1. The derivative of a sum is the sum of derivatives

If $g\left(w\right)$ is a sum of P functions $g\left(w\right) = \sum_{p=1}^{P} h_p\left(w\right)$ then $\frac{d}{dw} g\left(w\right) = \sum_{p=1}^{P} \frac{d}{dw} h_p\left(w\right)$

2. The chain rule

If g is a composition of functions of the form $g\left(w\right) = h\left(r\left(w\right)\right)$ then the derivative $\frac{d}{dw} g\left(w\right) = \frac{d}{dt} h\left(t\right) \frac{d}{dw} r\left(w\right)$ where t is evaluated at $t = r(w)$.

3. The product rule

If g is a product of functions of the form $g\left(w\right) = h\left(w\right) r\left(w\right)$ then the derivative $\frac{d}{dw} g\left(w\right) = \left(\frac{d}{dw} h\left(w\right)\right) r\left(w\right) + h\left(w\right) \left(\frac{d}{dw} r\left(w\right)\right)$.

4. Various derivative formulae

For example if $g\left(w\right) = w^c$ then $\frac{d}{dw} g\left(w\right) = cw^{c-1}$, or if $g\left(w\right) = \sin\left(w\right)$ then $\frac{d}{dw} g\left(w\right) = \cos\left(w\right)$, or if $g\left(w\right) = c$ where c is some constant then $\frac{d}{dw} g\left(w\right) = 0$, etc.

5. The sizes/shapes of gradients and Hessians

Remember: if $\mathbf{w}$ is an $N \times 1$ column vector then $\nabla g\left(\mathbf{w}\right)$ is also an $N \times 1$ column vector, and $\nabla^2 g\left(\mathbf{w}\right)$ is an $N \times N$ symmetric matrix.

9.10 Examples of gradient and Hessian calculations

Here we show detailed calculations for the gradient and Hessian of various functions employing the definitions and common rules previously stated. We begin by showing several first and second derivative calculations for scalar input functions, and then show the gradient/Hessian calculations for the analogous vector input versions of these functions.

Example 9.5. Practice derivative calculations: scalar input functions

Below we compute the first and second derivatives of three scalar input functions $g(w)$ where w is a scalar input. Note that throughout we will use two notations for a scalar derivative $\frac{d}{dw}g(w)$ and $g'(w)$ interchangeably.

a) $g(w) = \frac{1}{2}qw^2 + rw + d$ where q, r, and d are constants

Using derivative formula and the fact that the derivative of a sum is the sum of derivatives we have

$$g'(w) = qw + r \tag{9.52}$$

and

$$g''(w) = q. \tag{9.53}$$

b) $g(w) = -\cos(2\pi w^2) + w^2$

Using the chain rule on the $\cos(\cdot)$ part, the fact that the derivative of a sum is the sum of derivatives, and derivative formulae we have

$$g'(w) = \sin\left(2\pi w^2\right) 4\pi w + 2w. \tag{9.54}$$

Likewise taking the second derivative we differentiate the above (additionally using the product rules) as

$$g''(w) = \cos\left(2\pi w^2\right) (4\pi w)^2 + \sin\left(2\pi w^2\right) 4\pi + 2. \tag{9.55}$$

c) $g(w) = \sum_{p=1}^{P} \log\left(1 + e^{-a_p w}\right)$ where $a_1 ... a_P$ are constants

Call the p^{th} summand $h_p(w) = \log\left(1 + e^{-a_p w}\right)$. Then using the chain rule since $\frac{d}{dt}\log(t) = \frac{1}{t}$ and $\frac{d}{dw}\left(1 + e^{-a_p w}\right) = -a_p e^{-a_p w}$ together we have $\frac{d}{dw}h_p(w) = \frac{1}{1+e^{-a_p w}}\left(-a_p e^{-a_p w}\right) = -\frac{a_p e^{-a_p w}}{1+e^{-a_p w}} = -\frac{a_p}{e^{a_p w}+e^{a_p w}e^{-a_p w}} = -\frac{a_p}{1+e^{a_p w}}$. Now using this result, and since the derivative of a sum is the

sum of the derivatives and $g(w) = \sum_{p=1}^{P} h_p(w)$, we have $\frac{d}{dw} g(w) = \sum_{p=1}^{P} \frac{d}{dw} h_p(w)$ and so

$$g'(w) = -\sum_{p=1}^{P} \frac{a_p}{1 + e^{a_p w}}. \tag{9.56}$$

To compute the second derivative let us again do so by first differentiating the above summand-by-summand. Denote the p^{th} summand above as $h_p(w) = \frac{a_p}{1+e^{a_p w}}$. To compute its derivative we must apply the product and chain rules once again, we have $h_p'(w) = -\frac{a_p}{(1+e^{a_p w})^2} a_p e^{a_p w} = -\frac{e^{a_p w}}{(1+e^{a_p w})^2} a_p^2$. We can then compute the full second derivative as $g''(w) = -\sum_{p=1}^{P} h_p'(w)$ or likewise

$$g''(w) = \sum_{p=1}^{P} \frac{e^{a_p w}}{(1 + e^{a_p w})^2} a_p^2. \tag{9.57}$$

Example 9.6. Practice derivative calculations: vector input functions

Below we compute the gradients and Hessians of three vector input functions $g(\mathbf{w})$ where $\mathbf{w}$ is an $N \times 1$ dimensional input vector. The functions discussed here are analogous to the scalar functions discussed in the first example.

a) $g(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{Q}\mathbf{w} + \mathbf{r}^T\mathbf{w} + d$, here $\mathbf{Q}$ is an $N \times N$ symmetric matrix, $\mathbf{r}$ is an $N \times 1$ vector, and d is a scalar.

Notice that $g(\mathbf{w})$ here is the vector version of the function shown in **a)** of the first example. We should therefore expect the final shape of the gradient and Hessian to generally match the first and second derivatives we found there.

Writing out g in terms of the individual entries of $\mathbf{w}$ we have $g(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} w_n Q_{nm} w_m + \sum_{n=1}^{N} r_n w_n + d$ then taking the j^{th} partial derivative we have, since the derivative of a sum is the sum of derivatives $\frac{\partial}{\partial w_j} g(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} \frac{\partial}{\partial w_j}(w_n Q_{nm} w_m) + \sum_{n=1}^{N} \frac{\partial}{\partial w_j}(r_n w_n)$ where the d vanishes since it is a constant and $\frac{\partial}{\partial w_j} d = 0$. Now evaluating each derivative we apply the product rule to each $w_n Q_{nm} w_m$ (and remembering that all other terms in w_k where $k \neq j$

are constant and thus vanish when taking the w_j partial derivative) we have

$$\frac{\partial}{\partial w_j} g\left(\mathbf{w}\right) = \frac{1}{2}\left(\sum_{n=1}^{N} w_n Q_{nj} + \sum_{m=1}^{N} Q_{jm} w_m\right) + r_j. \tag{9.58}$$

All together the gradient can then be written compactly as

$$\nabla g\left(\mathbf{w}\right) = \frac{1}{2}\left(\mathbf{Q} + \mathbf{Q}^T\right)\mathbf{w} + \mathbf{r}, \tag{9.59}$$

and because $\mathbf{Q}$ is symmetric this is equivalently

$$\nabla g\left(\mathbf{w}\right) = \mathbf{Q}\mathbf{w} + \mathbf{r}. \tag{9.60}$$

Notice how the gradient here takes precisely the same shape as the corresponding scalar derivative shown in (9.52).

To compute the Hessian we compute mixed partial derivatives of the form $\frac{\partial^2}{\partial w_i \partial w_j} g\left(\mathbf{w}\right)$. To do this efficiently we can take the partial $\frac{\partial}{\partial w_i}$ of equation (9.58) since $\frac{\partial^2}{\partial w_i \partial w_j} g\left(\mathbf{w}\right) = \frac{\partial}{\partial w_i}\left(\frac{\partial}{\partial w_j} g\left(\mathbf{w}\right)\right)$which gives

$$\frac{\partial^2}{\partial w_i \partial w_j} g\left(\mathbf{w}\right) = \frac{1}{2}\left(Q_{ij} + Q_{ji}\right). \tag{9.61}$$

All together we then have that the full Hessian matrix is

$$\nabla^2 g\left(\mathbf{w}\right) = \frac{1}{2}\left(\mathbf{Q} + \mathbf{Q}^T\right), \tag{9.62}$$

and because $\mathbf{Q}$ is symmetric this is equivalently

$$\nabla^2 g\left(\mathbf{w}\right) = \mathbf{Q}. \tag{9.63}$$

Notice how this exactly the vector form of the second derivative given in (9.53).

b) $g\left(\mathbf{w}\right) = -\cos\left(2\pi \mathbf{w}^T\mathbf{w}\right) + \mathbf{w}^T\mathbf{w}$

First, notice that this is the vector input version of **b)** from the first example, therefore we should expect the final shape of the gradient and Hessian to generally match the first and second derivatives we found there.

Writing out g in terms of individual entries of $\mathbf{w}$ we have $g\left(\mathbf{w}\right) = -\cos\left(2\pi\sum_{n=1}^{N} w_n^2\right) + \sum_{n=1}^{N} w_n^2$, now taking the j^{th} partial we have

$$\frac{\partial}{\partial w_j} g\left(\mathbf{w}\right) = \sin\left(2\pi\sum_{n=1}^{N} w_n^2\right) 4\pi w_j + 2w_j. \tag{9.64}$$

From this we can see that the full gradient then takes the form

$$\nabla g\left(\mathbf{w}\right) = \sin\left(2\pi\mathbf{w}^T\mathbf{w}\right)4\pi\mathbf{w} + 2\mathbf{w}. \tag{9.65}$$

This is precisely the analog of the first derivative scalar version of this function shown in equation (9.54) of the previous example.

To compute the second derivatives we can take the partial $\frac{\partial}{\partial w_i}$ of equation (9.64) which gives

$$\frac{\partial^2}{\partial w_i \partial w_j} g\left(\mathbf{w}\right) = \begin{cases} \cos\left(2\pi\sum_{n=1}^{N} w_n^2\right)\left(4\pi\right)^2 w_i w_j + \sin\left(2\pi\sum_{n=1}^{N} w_n^2\right)4\pi + 2 & \text{if } i = j \\ \cos\left(2\pi\sum_{n=1}^{N} w_n^2\right)\left(4\pi\right)^2 w_i w_j & \text{else.} \end{cases} \tag{9.66}$$

All together then denoting $\mathbf{I}_{N\times N}$ the $N \times N$ identity matrix we may write the Hessian as

$$\nabla^2 g\left(\mathbf{w}\right) = \cos\left(2\pi\mathbf{w}^T\mathbf{w}\right)\left(4\pi\right)^2\mathbf{w}\mathbf{w}^T + \left(2 + \sin\left(2\pi\mathbf{w}^T\mathbf{w}\right)4\pi\right)\mathbf{I}_{N\times N}. \tag{9.67}$$

Notice that this is analogous to the second derivative, shown in equation (9.55), of the scalar version of the function.

c) $g\left(\mathbf{w}\right) = \sum_{p=1}^{P} \log\left(1 + e^{-\mathbf{a}_p^T\mathbf{w}}\right)$ where $\mathbf{a}_1 ... \mathbf{a}_P$ are $N \times 1$ vectors

This is the vector-input version of **c)** from the first example, so we should expect similar patterns emerge when computing derivatives here.

Denote by $h_p\left(\mathbf{w}\right) = \log\left(1 + e^{-\mathbf{a}_p^T\mathbf{w}}\right) = \log\left(1 + e^{-\sum_{n=1}^{N} a_{pn}w_n}\right)$ one of the summands of g.

Then using the chain rule twice the j^{th} partial can be written as

$$\frac{\partial}{\partial w_j} h_p\left(\mathbf{w}\right) = \frac{1}{1 + e^{-\mathbf{a}_p^T\mathbf{w}}} e^{-\mathbf{a}_p^T\mathbf{w}}\left(-a_{pj}\right). \tag{9.68}$$

Since $\frac{1}{1+e^{-\mathbf{a}_p^T\mathbf{w}}} e^{-\mathbf{a}_p^T\mathbf{w}} = \frac{1}{e^{\mathbf{a}_p^T\mathbf{w}} + e^{\mathbf{a}_p^T\mathbf{w}} e^{-\mathbf{a}_p^T\mathbf{w}}} = \frac{1}{1+e^{\mathbf{a}_p^T\mathbf{w}}}$ we can rewrite the above more compactly as

$$\frac{\partial}{\partial w_j} h_p\left(\mathbf{w}\right) = -\frac{a_{pj}}{1 + e^{\mathbf{a}_p^T\mathbf{w}}} \tag{9.69}$$

and summing over p gives

$$\frac{\partial}{\partial w_j} g\left(\mathbf{w}\right) = -\sum_{p=1}^{P} \frac{a_{pj}}{1 + e^{\mathbf{a}_p^T\mathbf{w}}} \tag{9.70}$$

The full gradient of g is then given by

$$\nabla g\left(\mathbf{w}\right) = -\sum_{p=1}^{P} \frac{\mathbf{a}_p}{1 + e^{\mathbf{a}_p^T \mathbf{w}}}. \tag{9.71}$$

Notice the similar shape of this gradient compared to the derivative of the scalar form of the function, as shown in equation (9.56).

Computing the second partial derivatives from equation (9.70) we have

$$\frac{\partial^2}{\partial w_i \partial w_j} g\left(\mathbf{w}\right) = \sum_{p=1}^{P} \frac{e^{\mathbf{a}_p^T \mathbf{w}}}{\left(1 + e^{\mathbf{a}_p^T \mathbf{w}}\right)^2} a_{pi} a_{pj}, \tag{9.72}$$

and so we may write the full Hessian compactly as

$$\nabla^2 g\left(\mathbf{w}\right) = \sum_{p=1}^{P} \frac{e^{\mathbf{a}_p^T \mathbf{w}}}{\left(1 + e^{\mathbf{a}_p^T \mathbf{w}}\right)^2} \mathbf{a}_p \mathbf{a}_p^T. \tag{9.73}$$

Notice how this is the analog of the second derivative of the scalar version of the function shown in equation (9.57).