B: Basics of vector calculus

B.1 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

$$\frac{\partial}{\partial w_i} g(\mathbf{w}). \tag{B.1}$$

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

$$\nabla g(\mathbf{w}) = \begin{bmatrix} \frac{\partial}{\partial w_1} g(\mathbf{w}) \\ \frac{\partial}{\partial w_2} g(\mathbf{w}) \\ \vdots \\ \frac{\partial}{\partial w_N} g(\mathbf{w}) \end{bmatrix}. \tag{B.2}$$

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

$$\frac{\partial^2}{\partial w_i \partial w_j} g(\mathbf{w}), \tag{B.3}$$

or equivalently as

$$\frac{\partial^2}{\partial w_j \partial w_i} g(\mathbf{w}). \tag{B.4}$$

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

$$\nabla^2 g(\mathbf{w}) = \begin{bmatrix} \frac{\partial^2}{\partial w_1 \partial w_1} g(\mathbf{w}) & \frac{\partial^2}{\partial w_1 \partial w_2} g(\mathbf{w}) & \cdots & \frac{\partial^2}{\partial w_1 \partial w_N} g(\mathbf{w}) \\ \frac{\partial^2}{\partial w_2 \partial w_1} g(\mathbf{w}) & \frac{\partial^2}{\partial w_2 \partial w_2} g(\mathbf{w}) & \cdots & \frac{\partial^2}{\partial w_2 \partial w_N} g(\mathbf{w}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial w_N \partial w_1} g(\mathbf{w}) & \frac{\partial^2}{\partial w_N \partial w_2} g(\mathbf{w}) & \cdots & \frac{\partial^2}{\partial w_N \partial w_N} g(\mathbf{w}) \end{bmatrix}. \tag{B.5}$$

B.2 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 $g(w)$ is a sum of P functions $g(w) = \sum_{p=1}^{P} h_p(w)$, then $\frac{d}{dw}g(w) = \sum_{p=1}^{P} \frac{d}{dw}h_p(w)$.

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

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

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

5. The sizes/shapes of gradients and Hessians
Remember: if $\mathbf{w}$ is an $N \times 1$ column vector then $\nabla g(\mathbf{w})$ is also an $N \times 1$ column vector, and $\nabla^2 g(\mathbf{w})$ is an $N \times N$ symmetric matrix.

B.3 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 B.1 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 formulae and the fact that the derivative of a sum is the sum of derivatives, we have

$$g'(w) = qw + r \tag{B.6}$$

and

$$g''(w) = q. \tag{B.7}$$

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

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

$$g'(w) = \sin\left(2\pi w^2\right) 4\pi w + 2w. \tag{B.8}$$

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

$$g''(w) = \cos\left(2\pi w^2\right) (4\pi w)^2 + \sin\left(2\pi w^2\right) 4\pi + 2. \tag{B.9}$$

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

Call the pth 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

$$g'(w) = -\sum_{p=1}^{P} \frac{a_p}{1 + e^{a_p w}}. \tag{B.10}$$

To compute the second derivative let us again do so by first differentiating the above summand-by-summand. Denote the pth 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

$$g''(w) = \sum_{p=1}^{P} \frac{e^{a_p w}}{(1 + e^{a_p w})^2} a_p^2. \tag{B.11}$$

Example B.2 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.

Note 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 jth 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 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), and we have

$$\frac{\partial}{\partial w_j} g(\mathbf{w}) = \frac{1}{2}\left(\sum_{n=1}^{N} w_n Q_{nj} + \sum_{m=1}^{N} Q_{jm} w_m\right) + r_j. \tag{B.12}$$

All together the gradient can then be written compactly as

$$\nabla g(\mathbf{w}) = \frac{1}{2}\left(\mathbf{Q} + \mathbf{Q}^T\right)\mathbf{w} + \mathbf{r}, \tag{B.13}$$

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

$$\nabla g(\mathbf{w}) = \mathbf{Q}\mathbf{w} + \mathbf{r}. \tag{B.14}$$

Note how the gradient here takes precisely the same shape as the corresponding scalar derivative shown in (B.6).

To compute the Hessian we compute mixed partial derivatives of the form $\frac{\partial^2}{\partial w_i \partial w_j} g(\mathbf{w})$. To do this efficiently we can take the partial $\frac{\partial}{\partial w_i}$ of Equation (B.12), since $\frac{\partial^2}{\partial w_i \partial w_j} g(\mathbf{w}) = \frac{\partial}{\partial w_i}\left(\frac{\partial}{\partial w_j} g(\mathbf{w})\right)$, which gives

$$\frac{\partial^2}{\partial w_i \partial w_j} g(\mathbf{w}) = \frac{1}{2}\left(Q_{ij} + Q_{ji}\right). \tag{B.15}$$

All together we then have that the full Hessian matrix is

$$\nabla^2 g(\mathbf{w}) = \frac{1}{2}\left(\mathbf{Q} + \mathbf{Q}^T\right), \tag{B.16}$$

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

$$\nabla^2 g(\mathbf{w}) = \mathbf{Q}. \tag{B.17}$$

Note how this is exactly the vector form of the second derivative given in (B.7).

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

First, note 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 $g(\mathbf{w}) = -\cos\left(2\pi\sum_{n=1}^{N} w_n^2\right) + \sum_{n=1}^{N} w_n^2$, now taking the jth partial we have

$$\frac{\partial}{\partial w_j} g(\mathbf{w}) = \sin\left(2\pi\sum_{n=1}^{N} w_n^2\right) 4\pi w_j + 2w_j. \tag{B.18}$$

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

$$\nabla g(\mathbf{w}) = \sin\left(2\pi\,\mathbf{w}^T\mathbf{w}\right) 4\pi\,\mathbf{w} + 2\mathbf{w}. \tag{B.19}$$

This is precisely the analog of the first derivative of the scalar version of this function shown in Equation (B.8) of the previous example.

To compute the second derivatives we can take the partial $\frac{\partial}{\partial w_i}$ of Equation (B.18), which gives

$$\frac{\partial^2}{\partial w_i \partial w_j} g(\mathbf{w}) = \begin{cases} \cos\left(2\pi\sum_{n=1}^{N} w_n^2\right)(4\pi)^2 w_i w_j + \sin\left(2\pi\sum_{n=1}^{N} w_n^2\right) 4\pi + 2 & \text{if } i = j \\ \cos\left(2\pi\sum_{n=1}^{N} w_n^2\right)(4\pi)^2 w_i w_j & \text{else.} \end{cases} \tag{B.20}$$

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

$$\nabla^2 g(\mathbf{w}) = \cos\left(2\pi\,\mathbf{w}^T\mathbf{w}\right)(4\pi)^2\,\mathbf{w}\mathbf{w}^T + \left(2 + \sin\left(2\pi\,\mathbf{w}^T\mathbf{w}\right)4\pi\right)\mathbf{I}_{N\times N}. \tag{B.21}$$

Note that this is analogous to the second derivative, shown in Equation (B.9), of the scalar version of the function.

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

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

Denote by $h_p(\mathbf{w}) = \log\left(1 + e^{-\mathbf{a}_p^T\mathbf{w}}\right) = \log\left(1 + e^{-\sum_{n=1}^{N} a_{pn} w_n}\right)$ one of the summands of g. Then using the chain rule, twice the jth partial can be written as

$$\frac{\partial}{\partial w_j} h_p(\mathbf{w}) = \frac{1}{1 + e^{-\mathbf{a}_p^T\mathbf{w}}} e^{-\mathbf{a}_p^T\mathbf{w}} \left(-a_{pj}\right). \tag{B.22}$$

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

$$\frac{\partial}{\partial w_j} h_p(\mathbf{w}) = -\frac{a_{pj}}{1 + e^{\mathbf{a}_p^T\mathbf{w}}} \tag{B.23}$$

and summing over p gives

$$\frac{\partial}{\partial w_j} g(\mathbf{w}) = -\sum_{p=1}^{P} \frac{a_{pj}}{1 + e^{\mathbf{a}_p^T \mathbf{w}}}. \tag{B.24}$$

The full gradient of g is then given by

$$\nabla g(\mathbf{w}) = -\sum_{p=1}^{P} \frac{\mathbf{a}_p}{1 + e^{\mathbf{a}_p^T \mathbf{w}}}. \tag{B.25}$$

Note the similar shape of this gradient compared to the derivative of the scalar form of the function, as shown in Equation (B.10).

Computing the second partial derivatives from equation (B.24), we have

$$\frac{\partial^2}{\partial w_i \partial w_j} g(\mathbf{w}) = \sum_{p=1}^{P} \frac{e^{\mathbf{a}_p^T \mathbf{w}}}{\left(1 + e^{\mathbf{a}_p^T \mathbf{w}}\right)^2} a_{pi} a_{pj}, \tag{B.26}$$

and so we may write the full Hessian compactly as

$$\nabla^2 g(\mathbf{w}) = \sum_{p=1}^{P} \frac{e^{\mathbf{a}_p^T \mathbf{w}}}{\left(1 + e^{\mathbf{a}_p^T \mathbf{w}}\right)^2} \mathbf{a}_p \mathbf{a}_p^T. \tag{B.27}$$

Note how this is the analog of the second derivative of the scalar version of the function shown in Equation (B.11).