Vector/Matrix Calculus

In neural networks, we often encounter problems with analysis of several variables. Vector/Matrix calculus extends calculus of one variable into that of a vector or a matrix of variables.

Vector Gradient: Let $g(\mathbf{w})$ be a differentiable scalar function of m variables, where

$$\mathbf{w} = [w_1, \ldots, w_m]^T$$

Then the vector gradient of $g(\mathbf{w})$ w.r.t. $\mathbf{w}$ is the m-dimensional vector of partial derivatives of g

$$\frac{\partial g}{\partial \mathbf{w}} = \nabla g = \nabla_{\mathbf{w}} g = \begin{pmatrix} \frac{\partial g}{\partial w_1} \\ \vdots \\ \frac{\partial g}{\partial w_m} \end{pmatrix}$$

Similarly we can define second-order gradient or Hessian matrix.

Hessian matrix is defined as

$$\frac{\partial^2 g}{\partial \mathbf{w}^2} = \begin{pmatrix} \frac{\partial^2 g}{\partial w_1^2} & \cdots & \frac{\partial^2 g}{\partial w_1 w_m} \\ \vdots & & \vdots \\ \frac{\partial^2 g}{\partial w_m w_1} & \cdots & \frac{\partial^2 g}{\partial w_m^2} \end{pmatrix}$$

Jacobian matrix: Generalization to the vector valued functions

$$\mathbf{g}(\mathbf{w}) = [g_1(\mathbf{w}), \ldots, g_n(\mathbf{w})]^T$$

leads to a definition of the Jacobian matrix of $\mathbf{g}$ w.r.t. $\mathbf{w}$

$$\frac{\partial \mathbf{g}}{\partial \mathbf{w}} = \begin{pmatrix} \frac{\partial g_1}{\partial w_1} & \cdots & \frac{\partial g_n}{\partial w_1} \\ \vdots & & \vdots \\ \frac{\partial g_1}{\partial w_m} & \cdots & \frac{\partial g_n}{\partial w_m} \end{pmatrix}$$

In this vector convention the columns of the Jacobian matrix are gradients of the corresponding components functions $g_i(\mathbf{w})$ w.r.t. the vector $\mathbf{w}$.

Differentiation Rules:

The differentiation rules are analogous to the case of ordinary functions:

- $$\frac{\partial f(\mathbf{w})g(\mathbf{w})}{\partial \mathbf{w}} = \frac{\partial f(\mathbf{w})}{\partial \mathbf{w}} g(\mathbf{w}) + f(\mathbf{w}) \frac{\partial g(\mathbf{w})}{\partial \mathbf{w}}$$

- $$\frac{\partial f(\mathbf{w})/g(\mathbf{w})}{\partial \mathbf{w}} = \frac{\left[\frac{\partial f(\mathbf{w})}{\partial \mathbf{w}} g(\mathbf{w}) - f(\mathbf{w}) \frac{\partial g(\mathbf{w})}{\partial \mathbf{w}}\right]}{g^2(\mathbf{w})}$$

- $$\frac{\partial f(g(\mathbf{w}))}{\partial \mathbf{w}} = f'(g(\mathbf{w})) \frac{\partial g(\mathbf{w})}{\partial \mathbf{w}}$$

Example: Consider

$$g(\mathbf{w}) = \sum_{i=1}^{m} a_i w_i = \mathbf{a}^T \mathbf{w}$$

where $\mathbf{a}$ is a constant vector. Thus,

$$\frac{\partial g}{\partial \mathbf{w}} = \begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix}$$

or in the vector notation

$$\frac{\partial \mathbf{a}^T \mathbf{w}}{\partial \mathbf{w}} = \mathbf{a}.$$

Example: Let $\mathbf{w} = (w_1, w_2, w_3)^T \in R^3$.

$$g(\mathbf{w}) = 2w_1 + 5w_2 + 12w_3 = \begin{pmatrix} 2 & 5 & 12 \end{pmatrix} \mathbf{w}.$$

Because

$$\frac{\partial g}{\partial w_1} = \frac{\partial}{\partial w_1}(2w_1 + 5w_2 + 12w_3) = 2 + 0 + 0 = 2$$

hence,

$$\frac{\partial g}{\partial \mathbf{w}} = \begin{pmatrix} 2 \\ 5 \\ 12 \end{pmatrix}.$$

Example: Consider

$$g(\mathbf{w}) = \sum_{i=1}^{m} \sum_{j=1}^{m} a_{ij} w_i w_j = \mathbf{w}^T \mathbf{A} \mathbf{w}$$

where $\mathbf{A}$ is a constant square matrix. Thus,

$$\frac{\partial g}{\partial \mathbf{w}} = \begin{pmatrix} \sum_{j=1}^{m} w_j a_{1j} + \sum_{i=1}^{m} w_i a_{i1} \\ \vdots \\ \sum_{j=1}^{m} w_j a_{mj} + \sum_{i=1}^{m} w_i a_{im} \end{pmatrix}$$

and so in the vector notation

$$\frac{\partial \mathbf{w}^T \mathbf{A} \mathbf{w}}{\partial \mathbf{w}} = \mathbf{A}\mathbf{w} + \mathbf{A}^T \mathbf{w}.$$

Hessian of $g(\mathbf{w})$ is

$$\frac{\partial^2 \mathbf{w}^T \mathbf{A} \mathbf{w}}{\partial \mathbf{w}^2} = \begin{pmatrix} 2a_{11} & \cdots & a_{1m} + a_{m1} \\ \vdots & & \vdots \\ a_{m1} + a_{1m} & \cdots & 2a_{mm} \end{pmatrix}$$

which equals to $\mathbf{A} + \mathbf{A}^T$.

Example: Let $\mathbf{w} = (w_1, w_2)^T \in R^2$.

$$g(\mathbf{w}) = 3w_1w_1 + 2w_1w_2 + 6w_2w_1 + 5w_2w_2$$

$$= \begin{pmatrix} w_1 & w_2 \end{pmatrix} \begin{pmatrix} 3 & 2 \\ 6 & 5 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$$

$$\frac{\partial g}{\partial w_1} = \frac{\partial}{\partial w_1}(3w_1w_1 + 2w_1w_2 + 6w_2w_1 + 5w_2w_2)$$

$$= 3 * 2w_1 + 2w_2 + 6w_2 + 0 = 6w_1 + 8w_2$$

$$\frac{\partial g}{\partial w_2} = 0 + 2 * w_1 + 6w_1 + 5 * 2w_2 = 8w_1 + 10w_2$$

and so in the vector notation

$$\frac{\partial \mathbf{w}^T \mathbf{A} \mathbf{w}}{\partial \mathbf{w}} = \begin{pmatrix} 3 & 2 \\ 6 & 5 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} + \begin{pmatrix} 3 & 6 \\ 2 & 5 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$$

$$= \begin{pmatrix} 6 & 8 \\ 8 & 10 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$$

Hessian of $g(\mathbf{w})$ is

$$\frac{\partial^2 \mathbf{w}^T \mathbf{A} \mathbf{w}}{\partial \mathbf{w}^2} = \frac{\partial}{\partial \mathbf{w}}(\frac{\partial \mathbf{w}^T \mathbf{A} \mathbf{w}}{\partial \mathbf{w}})$$

$$= \begin{pmatrix} 2 * 3 & 2 + 6 \\ 6 + 2 & 2 * 5 \end{pmatrix} = \begin{pmatrix} 6 & 8 \\ 8 & 10 \end{pmatrix}.$$

Matrix Gradient: Consider a scalar valued function $g(\mathbf{W})$ of the $n \times m$ matrix $\mathbf{W} = \{w_{ij}\}$ (e.g. determinant of a matrix). The matrix gradient w.r.t $\mathbf{W}$ is a matrix of the same dimension as $\mathbf{W}$ consisting of partial derivatives of $g(\mathbf{W})$ w.r.t. components of $\mathbf{W}$:

$$\frac{\partial g}{\partial \mathbf{W}} = \begin{pmatrix} \frac{\partial g}{\partial w_{11}} & \cdots & \frac{\partial g}{\partial w_{1n}} \\ \vdots & & \vdots \\ \frac{\partial g}{\partial w_{m1}} & \cdots & \frac{\partial g}{\partial w_{mn}} \end{pmatrix}$$

Example: If $g(\mathbf{W}) = \text{tr}(\mathbf{W})$, then $\frac{\partial g}{\partial \mathbf{W}} = \mathbf{I}$
Example: Consider a matrix function

$$g(\mathbf{W}) = \sum_{i=1}^{m} \sum_{j=1}^{m} w_{ij} a_i a_j = \mathbf{a}^T \mathbf{W} \mathbf{a}$$

i.e. assume that $\mathbf{a}$ is a constant vector, whereas $\mathbf{W}$ is a matrix of variables. Taking the gradient w.r.t. to $\mathbf{W}$ yields $\frac{\partial \mathbf{a}^T \mathbf{W} \mathbf{a}}{\partial w_{ij}} = a_i a_j$. Thus, in the matrix form

$$\frac{\partial \mathbf{a}^T \mathbf{W} \mathbf{a}}{\partial \mathbf{W}} = \mathbf{a}\mathbf{a}^T$$

Example:

$$g(\mathbf{W}) = 9w_{11} + 6w_{21} + 6w_{12} + 4w_{22}$$

$$= \begin{pmatrix} 3 & 2 \end{pmatrix} \begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{pmatrix} \begin{pmatrix} 3 \\ 2 \end{pmatrix}$$

Thus we have

$$\frac{\partial g}{\partial w_{11}} = \frac{\partial}{\partial w_{11}}(9w_{11} + 6w_{21} + 6w_{12} + 4w_{22}) = 9$$

$$\frac{\partial g}{\partial w_{12}} = \frac{\partial}{\partial w_{12}}(9w_{11} + 6w_{21} + 6w_{12} + 4w_{22}) = 6$$

$$\frac{\partial g}{\partial w_{21}} = \frac{\partial}{\partial w_{21}}(9w_{11} + 6w_{21} + 6w_{12} + 4w_{22}) = 6$$

$$\frac{\partial g}{\partial w_{22}} = \frac{\partial}{\partial w_{22}}(9w_{11} + 6w_{21} + 6w_{12} + 4w_{22}) = 4$$

hence

$$\frac{\partial g}{\partial \mathbf{W}} = \begin{pmatrix} 9 & 6 \\ 6 & 4 \end{pmatrix} = \begin{pmatrix} 3 \\ 2 \end{pmatrix} \begin{pmatrix} 3 & 2 \end{pmatrix}$$

Example: Let $\mathbf{W}$ be an invertible square matrix of dimension m with determinant, $\det(\mathbf{W})$. Then,

$$\frac{\partial}{\partial \mathbf{W}} \det \mathbf{W} = (\mathbf{W}^T)^{-1} \det \mathbf{W}$$

Recall that

$$\mathbf{W}^{-1} = \frac{1}{\det \mathbf{W}} \texttt{adj}(\mathbf{W}).$$

In the above $\texttt{adj}(\mathbf{W})$ is the *adjoint* of $\mathbf{W}$:

$$\texttt{adj}(\mathbf{W}) = \begin{pmatrix} W_{11} & \cdots & W_{m1} \\ \vdots & & \vdots \\ W_{1m} & \cdots & W_{mm} \end{pmatrix}$$

where W_{ij} is a *cofactor* obtained by multiplying term $(-1)^{i+j}$ by the determinant of a matrix obtained from $\mathbf{W}$ by removing the i^{th} row and the j^{th} column. Recall that the determinant of $\mathbf{W}$ can be also obtained using cofactors:

$$\det \mathbf{W} = \sum_{k=1}^{m} w_{ik} W_{ik}.$$

In the above formula i denotes an arbitrary row. Now taking derivative of $\det(\mathbf{W})$ w.r.t. W_{ij} gives

$$\frac{\partial \det(\mathbf{W})}{\partial w_{ij}} = W_{ij}$$

but from the definition of the matrix gradient it follows that

$$\frac{\partial \det(\mathbf{W})}{\partial \mathbf{W}} = \mathtt{adj}(\mathbf{W})^T$$

Using the formula for the inverse of $\mathbf{W}$.

$$\mathbf{W}^{-1} = \frac{1}{\det \mathbf{W}} \mathtt{adj}(\mathbf{W}).$$

We have

$$\frac{\partial \det(\mathbf{W})}{\partial \mathbf{W}} = (\mathbf{W}^T)^{-1} \det \mathbf{W}$$

<u>Homework</u>: Prove that

$$\frac{\partial \log |\det(\mathbf{W})|}{\partial \mathbf{W}} = \frac{1}{|\det \mathbf{W}|} \frac{\partial |\det \mathbf{W}|}{\partial \mathbf{W}} = (\mathbf{W}^T)^{-1}.$$