Multi-layer Perceptrons: back-propagation derivation

Harsha Vardhan

February 6, 2022

Abstract

This document contains derivation of the gradients for a 4-layer dense neural-network, using back-propagation (or reverse-mode differentiation). For the implementation of the neural-network see the accompanying notebooks.

1 Network Architecture

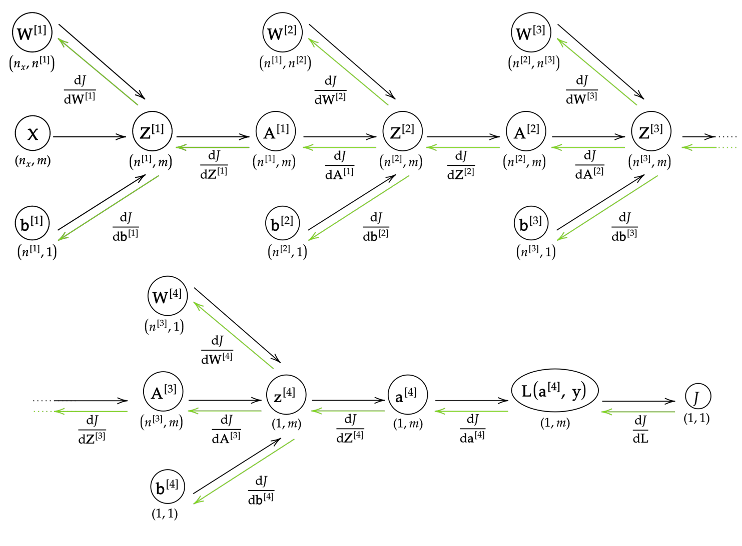

Figure 1: The black-colored arrows represent forward-propagation, and the green-colored arrows represent the gradient-flow.

2 Forward Propagation

The equations for forward propagation are as follows:

$$
\begin{aligned}
\mathbf{Z}^{[l]} &= (\mathbf{W}^{[l]})^\intercal \mathbf{A}^{[l-1]} + \mathbf{b}^{[l]} \vec{1}_{(1,m)} \\
\mathbf{A}^{[l]} &= f_{activation}(\mathbf{Z}^{[l]}) \\
\mathbf{L}(\mathbf{a}^{[4]}, \mathbf{y}) &= -\mathbf{y} \log(\hat{\mathbf{y}}) - (1 - \mathbf{y}) \log(1 - \hat{\mathbf{y}}) \\
J &= \frac{1}{m} \sum_{i=1}^{m} L(\hat{y}^{(i)}, y^{(i)}) = \frac{1}{m} \mathbf{L} \vec{1}_{(m,1)}
\end{aligned}
$$

where, m	is the number of training-examples
$\mathbf{W}^{[l]}$	is a $(n^{[l-1]}, n^{[l]})$ dimensional weight-matrix
$\mathbf{A}^{[l-1]}$	is a $(n^{[l-1]}, m)$ dimenstional activation-vector; and, $\mathbf{A}^{[0]} = \mathbf{X}$
$\mathbf{b}^{[l]}$	is a $(n^{[l]}, 1)$ dimenstional bias-vector
$\vec{1}_{(1,m)}$	is a $(1, m)$ dimensional vector of all 1's. Multiplying this with $\mathbf{b}^{[l]}$ has the same effect as python broadcasting.
$f_{activation}()$	is ReLU for all hidden-layers; is Sigmoid for the output-layer
$\hat{\mathbf{y}}$	$= \mathbf{a}^{[4]}$, the result of the output-layer
$\mathbf{L}$	is the loss function, and is a $(1, m)$ row vector
J	is the cost-function

3 Optimization: gradient-descent

The optimization is performed according to the following equations:

$$
\begin{aligned}
\mathbf{W}^{[l]} &:= \mathbf{W}^{[l]} - \alpha \frac{\mathrm{d}J}{\mathrm{d}\mathbf{W}^{[l]}} \\
\mathbf{b}^{[l]} &:= \mathbf{b}^{[l]} - \alpha \frac{\mathrm{d}J}{\mathrm{d}\mathbf{b}^{[l]}}
\end{aligned}
$$

where, α is the learning-rate/step-size.

3.1 Back-propagation

The gradients in the above equations are derived using back-propagation, as follows:

Since, J is a scalar, we can write $J = \mathrm{tr}(J) = J^\intercal = \mathrm{tr}(J^\intercal)$. And the derivative can be computed as follows:

$$\begin{aligned}\mathrm{d}J &= \mathrm{d}(\mathrm{tr}(J)) \\ &= \mathrm{tr}(\mathrm{d}J)\end{aligned}$$

The objective while computing the above derivative, is to massage the expression to the following form:

$$\mathrm{d}y = \mathrm{tr}(\mathbf{A}\mathrm{d}\mathbf{X})$$

then,

$$\frac{\mathrm{d}y}{\mathrm{d}\mathbf{X}} = \mathbf{A}$$

See [1] and [2] for more information.

3.1.1 Computing $\frac{\mathrm{d}J}{\mathrm{d}\mathbf{L}^\intercal}$

$$\begin{aligned}\mathrm{d}J &= \mathrm{tr}(\mathrm{d}J^\intercal) \\ &= \mathrm{tr}(\mathrm{d}(\frac{1}{m}(\vec{1}_{(m,1)})^\intercal \mathbf{L}^\intercal)) \\ &= \mathrm{tr}(\frac{1}{m}(\vec{1}_{(m,1)})^\intercal \mathrm{d}\mathbf{L}^\intercal) \\ \implies \frac{\mathrm{d}J}{\mathrm{d}\mathbf{L}^\intercal} &= \frac{1}{m}(\vec{1}_{(m,1)})^\intercal\end{aligned}$$

3.1.2 Computing $\frac{\mathrm{d}J}{\mathrm{d}(\mathbf{a}^{[4]})^\intercal}$

$$\begin{aligned}\mathrm{d}J &= \mathrm{tr}(\frac{\mathrm{d}J}{\mathrm{d}\mathbf{L}^\intercal}\frac{\mathrm{d}\mathbf{L}^\intercal}{\mathrm{d}(\mathbf{a}^{[4]})^\intercal}\mathrm{d}(\mathbf{a}^{[4]})^\intercal) \\ &= \mathrm{tr}\left(\frac{1}{m}(\vec{1}_{(m,1)})^\intercal \begin{bmatrix} \frac{\partial L^{(1)}}{\partial a^{[4](1)}} & 0 & \dots & 0 \\ 0 & \frac{\partial L^{(3)}}{\partial a^{[4](2)}} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{\partial L^{(m)}}{\partial a^{[4](m)}} \end{bmatrix} \mathrm{d}(\mathbf{a}^{[4]})^\intercal\right) \\ \implies \frac{\mathrm{d}J}{\mathrm{d}(\mathbf{a}^{[4]})^\intercal} &= \frac{1}{m}\left[\frac{\partial L^{(1)}}{\partial a^{[4](1)}}, \frac{\partial L^{(2)}}{\partial a^{[4](2)}}, \dots, \frac{\partial L^{(m)}}{\partial a^{[4](m)}}\right]\end{aligned}$$

where,

$$\frac{\mathrm{d}\mathbf{L}^\intercal}{\mathrm{d}(\mathbf{a}^{[4]})^\intercal} = \left[\frac{\partial}{\partial a^{[4](1)}}, \frac{\partial}{\partial a^{[4](2)}}, \ldots, \frac{\partial}{\partial a^{[4](m)}}\right] \otimes \begin{bmatrix} L^{(1)} \\ L^{(2)} \\ \vdots \\ L^{(m)} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial L^{(1)}}{\partial a^{[4](1)}} & 0 & \ldots & 0 \\ 0 & \frac{\partial L^{(3)}}{\partial a^{[4](2)}} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \frac{\partial L^{(m)}}{\partial a^{[4](m)}} \end{bmatrix}$$

where, $\otimes$ is the Kronecker-product.

The expansion of the derivative $\frac{\mathrm{d}\mathbf{L}^\intercal}{\mathrm{d}(\mathbf{a}^{[4]})^\intercal}$, w.r.t. to how the Kronecker-product was performed (i.e., denominator-transpose $\otimes$ numerator), is because we are following the numerator layout.

3.1.3 Computing $\frac{\mathrm{d}J}{\mathrm{d}(\mathbf{z}^{[4]})^\intercal}$

$$\mathrm{d}J = \mathrm{tr}(\frac{\mathrm{d}J}{\mathrm{d}(\mathbf{a}^{[4]})^\intercal} \frac{\mathrm{d}(\mathbf{a}^{[4]})^\intercal}{\mathrm{d}(\mathbf{z}^{[4]})^\intercal} \mathrm{d}(\mathbf{z}^{[4]})^\intercal)$$

$$= \mathrm{tr}\left(\frac{\mathrm{d}J}{\mathrm{d}(\mathbf{a}^{[4]})^\intercal} \begin{bmatrix} \frac{\partial a^{[4](1)}}{\partial z^{[4](1)}} & 0 & \ldots & 0 \\ 0 & \frac{\partial a^{[4](2)}}{\partial z^{[4](2)}} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \frac{\partial a^{[4](m)}}{\partial z^{[4](m)}} \end{bmatrix} \mathrm{d}(\mathbf{z}^{[4]})^\intercal\right)$$

$$= \mathrm{tr}\left(\frac{\mathrm{d}J}{\mathrm{d}(\mathbf{a}^{[4]})^\intercal} \circ \mathrm{diag}^{-1}\left(\frac{\mathrm{d}(\mathbf{a}^{[4]})^\intercal}{\mathrm{d}(\mathbf{z}^{[4]})^\intercal}\right)^\intercal \mathrm{d}(\mathbf{z}^{[4]})^\intercal\right)$$

$$\implies \frac{\mathrm{d}J}{\mathrm{d}(\mathbf{z}^{[4]})^\intercal} = \frac{\mathrm{d}J}{\mathrm{d}(\mathbf{a}^{[4]})^\intercal} \circ \mathrm{diag}^{-1}\left(\frac{\mathrm{d}(\mathbf{a}^{[4]})^\intercal}{\mathrm{d}(\mathbf{z}^{[4]})^\intercal}\right)^\intercal$$

where, $\circ$ is the hadamard-product.

where,

$$\frac{\mathrm{d}(\mathbf{a}^{[4]})^\intercal}{\mathrm{d}(\mathbf{z}^{[4]})^\intercal} = \left[\frac{\partial}{\partial z^{[4](1)}}, \frac{\partial}{\partial z^{[4](2)}}, \ldots, \frac{\partial}{\partial z^{[4](m)}}\right] \otimes \begin{bmatrix} a^{[4](1)} \\ a^{[4](2)} \\ \vdots \\ a^{[4](m)} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial a^{[4](1)}}{\partial z^{[4](1)}} & 0 & \ldots & 0 \\ 0 & \frac{\partial a^{[4](2)}}{\partial z^{[4](2)}} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \frac{\partial a^{[4](m)}}{\partial z^{[4](m)}} \end{bmatrix}$$

Notice that the matrix-product of $\frac{\mathrm{d}J}{\mathrm{d}(\mathbf{a}^{[4]})^\intercal}$ [a $(1, m)$-dimensional vector] and $\frac{\mathrm{d}(\mathbf{a}^{[4]})^\intercal}{\mathrm{d}(\mathbf{z}^{[4]})^\intercal}$ [a (m, m)-dimensional diagonal-matrix] can be re-written as the following hadamard-product:

$$\frac{\mathrm{d}J}{\mathrm{d}(\mathbf{a}^{[4]})^\intercal}\frac{\mathrm{d}(\mathbf{a}^{[4]})^\intercal}{\mathrm{d}(\mathbf{z}^{[4]})^\intercal} = \frac{\mathrm{d}J}{\mathrm{d}(\mathbf{a}^{[4]})^\intercal} \circ \mathrm{diag}^{-1}\left(\frac{\mathrm{d}(\mathbf{a}^{[4]})^\intercal}{\mathrm{d}(\mathbf{z}^{[4]})^\intercal}\right)^\intercal$$

Although, the above matrix product is straight-forward to compute, the usefulness of this step will become clear when dealing with derivatives of the form $\frac{\mathrm{d}\mathbf{A}^{[l]}}{\mathrm{d}\mathbf{Z}^{[l]}}$, where both $\mathbf{A}^{[l]}$ and $\mathbf{Z}^{[l]}$ are rank-2 tensors (i.e. 2D matrices), theoretically resulting in a rank-4 tensor.

3.1.4 Computing $\frac{\mathrm{d}J}{\mathrm{d}\mathbf{W}^{[4]}}$, $\frac{\mathrm{d}J}{\mathrm{d}\mathbf{b}^{[4]}}$, and $\frac{\mathrm{d}J}{\mathrm{d}\mathbf{A}^{[3]}}$

$$\mathrm{d}J = \mathrm{tr}(\frac{\mathrm{d}J}{\mathrm{d}(\mathbf{z}^{[4]})^\intercal}\mathrm{d}(\mathbf{z}^{[4]})^\intercal)$$

where, $\mathrm{d}(\mathbf{z}^{[4]})^\intercal$ can be expanded as,

$$\begin{aligned} \mathrm{d}(\mathbf{z}^{[4]})^\intercal &= \mathrm{d}((\mathbf{A}^{[3]})^\intercal\mathbf{W}^{[4]} + (\vec{1}_{(1,m)})^\intercal(\mathbf{b}^{[4]})^\intercal) \\ &= \mathrm{d}(\mathbf{A}^{[3]})^\intercal\mathbf{W}^{[4]} + (\mathbf{A}^{[3]})^\intercal\mathrm{d}(\mathbf{W}^{[4]}) + (\vec{1}_{(1,m)})^\intercal\mathrm{d}(\mathbf{b}^{[4]})^\intercal \end{aligned}$$

when differentiating w.r.t. $\mathbf{W}^{[4]}$, we have $\mathrm{d}(\mathbf{b}^{[4]})^\intercal = \mathrm{d}(\mathbf{A}^{[3]})^\intercal = 0$. So,

$$\mathrm{d}J = \mathrm{tr}(\frac{\mathrm{d}J}{\mathrm{d}(\mathbf{z}^{[4]})^\intercal}(\mathbf{A}^{[3]})^\intercal\mathrm{d}\mathbf{W}^{[4]})$$

$$\implies \frac{\mathrm{d}J}{\mathrm{d}\mathbf{W}^{[4]}} = \frac{\mathrm{d}J}{\mathrm{d}(\mathbf{z}^{[4]})^\intercal}(\mathbf{A}^{[3]})^\intercal$$

when differentiating w.r.t. $\mathbf{b}^{[4]}$, we have $\mathrm{d}\mathbf{W}^{[4]} = \mathrm{d}(\mathbf{A}^{[3]})^\intercal = 0$. So,

$$\begin{aligned}
\mathrm{d}J &= \mathrm{tr}(\frac{\mathrm{d}J}{\mathrm{d}(\mathbf{z}^{[4]})^\intercal}(\vec{1}_{(1,m)})^\intercal \mathrm{d}(\mathbf{b}^{[4]})^\intercal) \\
&= \mathrm{tr}(\mathrm{d}\mathbf{b}^{[4]}\vec{1}_{(1,m)}(\frac{\mathrm{d}J}{\mathrm{d}(\mathbf{z}^{[4]})^\intercal})^\intercal) \\
&= \mathrm{tr}(\vec{1}_{(1,m)}(\frac{\mathrm{d}J}{\mathrm{d}(\mathbf{z}^{[4]})^\intercal})^\intercal \mathrm{d}\mathbf{b}^{[4]}) \\
\implies \frac{\mathrm{d}J}{\mathrm{d}\mathbf{b}^{[4]}} &= \vec{1}_{(1,m)}(\frac{\mathrm{d}J}{\mathrm{d}(\mathbf{z}^{[4]})^\intercal})^\intercal
\end{aligned}$$

when differentiating w.r.t. $\mathbf{A}^{[3]}$, we have $\mathrm{d}\mathbf{W}^{[4]} = \mathrm{d}(\mathbf{b}^{[4]})^\intercal = 0$. So,

$$\begin{aligned}
\mathrm{d}J &= \mathrm{tr}(\frac{\mathrm{d}J}{\mathrm{d}(\mathbf{z}^{[4]})^\intercal}\mathrm{d}(\mathbf{A}^{[3]})^\intercal \mathbf{W}^{[4]}) \\
&= \mathrm{tr}((\mathbf{W}^{[4]})^\intercal \mathrm{d}\mathbf{A}^{[3]}(\frac{\mathrm{d}J}{\mathrm{d}(\mathbf{z}^{[4]})^\intercal})^\intercal) \\
&= \mathrm{tr}((\frac{\mathrm{d}J}{\mathrm{d}(\mathbf{z}^{[4]})^\intercal})^\intercal(\mathbf{W}^{[4]})^\intercal \mathrm{d}\mathbf{A}^{[3]}) \\
\implies \frac{\mathrm{d}J}{\mathrm{d}\mathbf{A}^{[3]}} &= (\frac{\mathrm{d}J}{\mathrm{d}(\mathbf{z}^{[4]})^\intercal})^\intercal(\mathbf{W}^{[4]})^\intercal
\end{aligned}$$

3.1.5 Computing $\frac{\mathrm{d}J}{\mathrm{d}\mathbf{Z}^{[3]}}$

$$\mathrm{d}J = \mathrm{tr}(\frac{\mathrm{d}J}{\mathrm{d}\mathbf{A}^{[3]}}\frac{\mathrm{d}\mathbf{A}^{[3]}}{\mathrm{d}\mathbf{Z}^{[3]}}\mathrm{d}\mathbf{Z}^{[3]})$$

Here, $\frac{\mathrm{d}\mathbf{A}^{[3]}}{\mathrm{d}\mathbf{Z}^{[3]}}$ should be a rank-4 tensor. But, the derivatives corresponding to a single training-example, i.e. $\frac{\mathrm{d}\mathbf{A}^{[3](i)}}{\mathrm{d}\mathbf{Z}^{[3](i)}}$, is a rank-2 tensor (more specifically, a

2D diagnonal-matrix), computed as follows:

$$\frac{\mathrm{d}\mathbf{A}^{[3](i)}}{\mathrm{d}\mathbf{Z}^{[3](i)}} = \left[\frac{\partial}{\partial z_1^{[3](i)}}, \frac{\partial}{\partial z_2^{[3](i)}}, \dots, \frac{\partial}{\partial z_{n^{[3]}}^{[3](i)}}\right] \otimes \begin{bmatrix} a_1^{[3](i)} \\ a_2^{[3](i)} \\ \vdots \\ a_{n^{[3]}}^{[3](i)} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial a_1^{[3](i)}}{\partial z_1^{[3](i)}} & 0 & \dots & 0 \\ 0 & \frac{\partial a_2^{[3](i)}}{\partial z_2^{[3](i)}} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{\partial a_{n^{[3]}}^{[3](i)}}{\partial z_{n^{[3]}}^{[3](i)}} \end{bmatrix}$$

The matrix-product of $\frac{\mathrm{d}J}{\mathrm{d}\mathbf{A}^{[3](i)}}$ [a $(1, n^{[3]})$ dimensional-vector] and $\frac{\mathrm{d}\mathbf{A}^{[3](i)}}{\mathrm{d}\mathbf{Z}^{[3](i)}}$ [a $(n^{[3]}, n^{[3]})$ dimensional diagonal-matrix] can be re-written as the following hadamard-product:

$$\frac{\mathrm{d}J}{\mathrm{d}\mathbf{A}^{[3](i)}} \frac{\mathrm{d}\mathbf{A}^{[3](i)}}{\mathrm{d}\mathbf{Z}^{[3](i)}} = \frac{\mathrm{d}J}{\mathrm{d}\mathbf{A}^{[3](i)}} \circ \mathrm{diag}^{-1}\left(\frac{\mathrm{d}\mathbf{A}^{[3](i)}}{\mathrm{d}\mathbf{Z}^{[3](i)}}\right)^{\intercal}$$

So, the derivative $\frac{\mathrm{d}\mathbf{A}^{[3]}}{\mathrm{d}\mathbf{Z}^{[3]}}$ corresponding to all m-training-examples can be computed, by vertically stacking the hadamard-product for each training-example, as follows:

$$\frac{\mathrm{d}J}{\mathrm{d}\mathbf{A}^{[3]}} \frac{\mathrm{d}\mathbf{A}^{[3]}}{\mathrm{d}\mathbf{Z}^{[3]}} = \frac{\mathrm{d}J}{\mathrm{d}\mathbf{A}^{[3]}} \circ \begin{bmatrix} \mathrm{diag}^{-1}\left(\frac{\mathrm{d}\mathbf{A}^{[3](1)}}{\mathrm{d}\mathbf{Z}^{[3](1)}}\right)^{\intercal} \\ \mathrm{diag}^{-1}\left(\frac{\mathrm{d}\mathbf{A}^{[3](2)}}{\mathrm{d}\mathbf{Z}^{[3](2)}}\right)^{\intercal} \\ \vdots \\ \mathrm{diag}^{-1}\left(\frac{\mathrm{d}\mathbf{A}^{[3](m)}}{\mathrm{d}\mathbf{Z}^{[3](m)}}\right)^{\intercal} \end{bmatrix}$$

3.1.6 Computing $\frac{\mathrm{d}J}{\mathrm{d}\mathbf{W}^{[3]}}$, $\frac{\mathrm{d}J}{\mathrm{d}\mathbf{b}^{[3]}}$, and $\frac{\mathrm{d}J}{\mathrm{d}\mathbf{A}^{[2]}}$

$$\mathrm{d}J = \mathrm{tr}\left(\frac{\mathrm{d}J}{\mathrm{d}\mathbf{Z}^{[3]}} \mathrm{d}\mathbf{Z}^{[3]}\right)$$

where, $\mathrm{d}\mathbf{Z}^{[3]}$ can be expanded as,

$$\mathrm{d}\mathbf{Z}^{[3]} = \mathrm{d}(\mathbf{W}^{[3]})^{\intercal}\mathbf{A}^{[2]} + (\mathbf{W}^{[3]})^{\intercal}\mathrm{d}\mathbf{A}^{[2]} + \mathrm{d}\mathbf{b}^{[3]}\vec{1}_{(1,m)}$$

when differentiating w.r.t. $\mathbf{W}^{[3]}$, we have $\mathrm{d}\mathbf{A}^{[2]} = \mathrm{d}\mathbf{b}^{[3]} = 0$. So,

$$\begin{aligned}
\mathrm{d}J &= \mathrm{tr}\left(\frac{\mathrm{d}J}{\mathrm{d}\mathbf{Z}^{[3]}}\mathrm{d}(\mathbf{W}^{[3]})^\intercal\mathbf{A}^{[2]}\right) \\
&= \mathrm{tr}\left((\mathbf{A}^{[2]})^\intercal\mathrm{d}\mathbf{W}^{[3]}\left(\frac{\mathrm{d}J}{\mathrm{d}\mathbf{Z}^{[3]}}\right)^\intercal\right) \\
&= \mathrm{tr}\left(\left(\frac{\mathrm{d}J}{\mathrm{d}\mathbf{Z}^{[3]}}\right)^\intercal(\mathbf{A}^{[2]})^\intercal\mathrm{d}\mathbf{W}^{[3]}\right) \\
\implies \frac{\mathrm{d}J}{\mathrm{d}\mathbf{W}^{[3]}} &= \left(\frac{\mathrm{d}J}{\mathrm{d}\mathbf{Z}^{[3]}}\right)^\intercal(\mathbf{A}^{[2]})^\intercal
\end{aligned}$$

when differentiating w.r.t. $\mathbf{b}^{[3]}$, we have $\mathrm{d}(\mathbf{W}^{[3]})^\intercal = \mathrm{d}\mathbf{A}^{[2]} = 0$. So,

$$\begin{aligned}
\mathrm{d}J &= \mathrm{tr}\left(\frac{\mathrm{d}J}{\mathrm{d}\mathbf{Z}^{[3]}}\mathrm{d}\mathbf{b}^{[3]}\vec{1}_{(1,m)}\right) \\
&= \mathrm{tr}\left(\vec{1}_{(1,m)}\frac{\mathrm{d}J}{\mathrm{d}\mathbf{Z}^{[3]}}\mathrm{d}\mathbf{b}^{[3]}\right) \\
\implies \frac{\mathrm{d}J}{\mathrm{d}\mathbf{b}^{[3]}} &= \vec{1}_{(1,m)}\frac{\mathrm{d}J}{\mathrm{d}\mathbf{Z}^{[3]}}
\end{aligned}$$

when differentiating w.r.t. $\mathbf{A}^{[2]}$, we have $\mathrm{d}(\mathbf{W}^{[3]})^\intercal = \mathrm{d}\mathbf{b}^{[3]} = 0$. So,

$$\begin{aligned}
\mathrm{d}J &= \mathrm{tr}\left(\frac{\mathrm{d}J}{\mathrm{d}\mathbf{Z}^{[3]}}(\mathbf{W}^{[3]})^\intercal\mathrm{d}\mathbf{A}^{[2]}\right) \\
\implies \frac{\mathrm{d}J}{\mathrm{d}\mathbf{A}^{[2]}} &= \frac{\mathrm{d}J}{\mathrm{d}\mathbf{Z}^{[3]}}(\mathbf{W}^{[3]})^\intercal
\end{aligned}$$

So far, we have performed the following types of back-propagation:

1. Propagating from a layer with $n^{[4]}$-unit to a layer with $n^{[3]}$-units (where, $n^{[4]} = 1, n^{[3]} > 1$); i.e., layer-4's activation to layer-3's activation.

2. Propagating from a layer with $n^{[3]}$-units to a layer with $n^{[2]}$-units (where, $n^{[3]}, n^{[2]} > 1$); i.e., layer-3's activation to layer-2's activation.

Therefore, when propagating from layer-2's activation to layer-1's activation, we can generalize the results from step-2 above, as follows:

$$\begin{aligned}
\frac{\mathrm{d}J}{\mathrm{d}\mathbf{Z}^{[2]}} &= \frac{\mathrm{d}J}{\mathrm{d}\mathbf{A}^{[2]}} \circ \begin{bmatrix} \operatorname{diag}^{-1}\left(\frac{\mathrm{d}\mathbf{A}^{[2](1)}}{\mathrm{d}\mathbf{Z}^{[2](1)}}\right)^{\intercal} \\ \vdots \\ \operatorname{diag}^{-1}\left(\frac{\mathrm{d}\mathbf{A}^{[2](m)}}{\mathrm{d}\mathbf{Z}^{[2](m)}}\right)^{\intercal} \end{bmatrix} \\
\frac{\mathrm{d}J}{\mathrm{d}\mathbf{W}^{[2]}} &= \left(\frac{\mathrm{d}J}{\mathrm{d}\mathbf{Z}^{[2]}}\right)^{\intercal} (\mathbf{A}^{[1]})^{\intercal} \\
\frac{\mathrm{d}J}{\mathrm{d}\mathbf{b}^{[2]}} &= \vec{1}_{(1,m)} \frac{\mathrm{d}J}{\mathrm{d}\mathbf{Z}^{[2]}} \\
\frac{\mathrm{d}J}{\mathrm{d}\mathbf{A}^{[1]}} &= \frac{\mathrm{d}J}{\mathrm{d}\mathbf{Z}^{[2]}} (\mathbf{W}^{[2]})^{\intercal}
\end{aligned}$$

So, given $\frac{\mathrm{d}J}{\mathrm{d}\mathbf{A}^{[l]}}$, for some layer-l, we can derive the following generalizations:

$$\begin{aligned}
\frac{\mathrm{d}J}{\mathrm{d}\mathbf{Z}^{[l]}} &= \frac{\mathrm{d}J}{\mathrm{d}\mathbf{A}^{[l]}} \circ \begin{bmatrix} \operatorname{diag}^{-1}\left(\frac{\mathrm{d}\mathbf{A}^{[l](1)}}{\mathrm{d}\mathbf{Z}^{[l](1)}}\right)^{\intercal} \\ \vdots \\ \operatorname{diag}^{-1}\left(\frac{\mathrm{d}\mathbf{A}^{[l](m)}}{\mathrm{d}\mathbf{Z}^{[l](m)}}\right)^{\intercal} \end{bmatrix} \\
\frac{\mathrm{d}J}{\mathrm{d}\mathbf{W}^{[l]}} &= \left(\frac{\mathrm{d}J}{\mathrm{d}\mathbf{Z}^{[l]}}\right)^{\intercal} (\mathbf{A}^{[l-1]})^{\intercal} \\
\frac{\mathrm{d}J}{\mathrm{d}\mathbf{b}^{[l]}} &= \vec{1}_{(1,m)} \frac{\mathrm{d}J}{\mathrm{d}\mathbf{Z}^{[l]}} \\
\frac{\mathrm{d}J}{\mathrm{d}\mathbf{A}^{[l-1]}} &= \frac{\mathrm{d}J}{\mathrm{d}\mathbf{Z}^{[l]}} (\mathbf{W}^{[l]})^{\intercal}
\end{aligned}$$

And for the layer-L (i.e. the output/last-layer), the derivative $\frac{\mathrm{d}J}{\mathrm{d}\mathbf{A}^{[L]}}$ is obtained by differentiating the cost function.

3.2 Jacobian or Gradient?

In the above derivations, we have used the numerator layout while performing matrix-derivatives. One of the consequences of this decision is that the derivatives that we have computed are in-fact jacobians and not gradients. Fortunately, gradients are just transpose of jacobians. So, based on our derivations the gradients would be the following:

$$
\begin{aligned}
\nabla_{\mathbf{W}^{[l]}}(J) &= \left(\frac{\mathrm{d}J}{\mathrm{d}\mathbf{W}^{[l]}}\right)^{\intercal} &&= \mathbf{A}^{[l-1]}\frac{\mathrm{d}J}{\mathrm{d}\mathbf{Z}^{[l]}} \\
\nabla_{\mathbf{b}^{[l]}}(J) &= \left(\frac{\mathrm{d}J}{\mathrm{d}\mathbf{b}^{[l]}}\right)^{\intercal} &&= \left(\frac{\mathrm{d}J}{\mathrm{d}\mathbf{Z}^{[l]}}\right)^{\intercal}(\vec{1}_{(1,m)})^{\intercal}
\end{aligned}
$$

References

[1] T. Minka, "Old and new matrix algebra useful for statistics," Sep. 1997. [Online]. Available: `https://www.microsoft.com/en-us/research/publication/old-new-matrix-algebra-useful-statistics/`.

[2] J. R. Magnus and H. Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics*, Second. John Wiley, 1999, ISBN: 0471986321 9780471986324 047198633X 9780471986331.