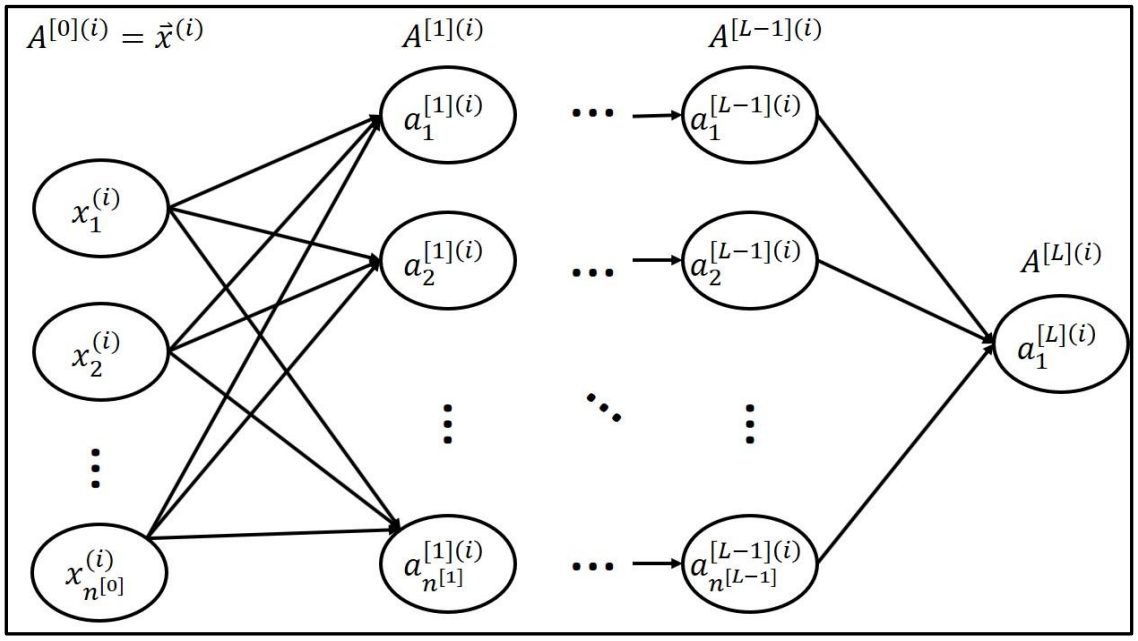
Derivation of He and Xavier Weight Initializations of an L-Layer Neural Network

1. Set-up

The following diagram depicts the structure of a one output L-layer neural network for training sample i, where $i \in \{1,2,\dots,m\}$:



Variables:

- $x_j^{(i)}$ → jth input feature for training sample i, where $j \in \{1,2,\dots,n^{[0]}\}$
- $\vec{x}^{(i)} = \begin{bmatrix} x_1^{(i)} \\ x_2^{(i)} \\ \vdots \\ x_{n^{[0]}}^{(i)} \end{bmatrix}$
- $X = [\vec{x}^{(1)} \quad \vec{x}^{(2)} \quad \dots \quad \vec{x}^{(m)}]$
- $a_j^{[\ell](i)}$ → jth activation in layer ℓ for training sample i, where $j \in \{1,2,\dots,n^{[\ell]}\}$ and $\ell \in \{0,2,\dots,L\}$, with $a_j^{[0](i)} = x_j^{(i)}$. In this model, we assume that $n^{[L]} = 1$.
- $\vec{a}^{[\ell](i)} = \begin{bmatrix} a_1^{[\ell](i)} \\ a_2^{[\ell](i)} \\ \vdots \\ a_{n^{[\ell]}}^{[\ell](i)} \end{bmatrix}$
- $A^{[\ell]} = [\vec{a}^{[\ell](1)} \quad \vec{a}^{[\ell](2)} \quad \dots \quad \vec{a}^{[\ell](m)}]$

- $y^{(i)}$ → True output for training sample i
- $\vec{y} = \left[y^{(1)}, y^{(2)}, \dots, y^{(m)}\right]$
- $w_{j,k}^{[\ell]}$ → Weight coefficient of activation $a_k^{[\ell-1](i)}$ for $i \in \{1,2,\dots,m\}$, where $\ell \in \{1,2,\dots,\ell\}$
- $\vec{w}_j^{[\ell]} = \begin{bmatrix} w_{j,1}^{[\ell]} \\ w_{j,2}^{[\ell]} \\ \vdots \\ w_{j,n^{[\ell-1]}}^{[\ell]} \end{bmatrix}$
- $W^{[\ell]} = \begin{bmatrix} \vec{w}_1^{[\ell]T} \\ \vec{w}_2^{[\ell]T} \\ \vdots \\ \vec{w}_{n^{[\ell]}}^{[\ell]T} \end{bmatrix}$
- $b_j^{[\ell]}$ → Bias for activation $a_j^{[\ell](i)}$ for $i \in \{1,2,\dots,m\}$
- $\vec{b}^{[\ell]} = \begin{bmatrix} b_1^{[\ell]} \\ b_2^{[\ell]} \\ \vdots \\ b_{n^{[\ell]}}^{[\ell]} \end{bmatrix}$
- $z_j^{[\ell](i)} = \vec{w}_j^{[\ell]T} \vec{a}^{[\ell-1](i)} + b_j^{[\ell]}$
- $\vec{z}^{[\ell](i)} = W^{[\ell]} \vec{a}^{[\ell-1](i)} + \vec{b}^{[\ell]}$
- $\mathbf{1}_{p\times q}$ → Matrix of ones with p rows and q columns
- $Z^{[\ell]} = W^{[\ell]} A^{[\ell-1]} + \vec{b}^{[\ell]} \mathbf{1}_{1\times m}$
- $g^{[\ell]}\left(z_j^{[\ell](i)}\right) = a_j^{[\ell](i)}$ → Activation function for members in layer ℓ, where $\ell \in \{1,2,\dots,L\}$
- $\vec{g}^{[\ell]}\left(\vec{z}^{[\ell](i)}\right) = \begin{bmatrix} g^{[\ell]}\left(z_1^{[\ell](i)}\right) \\ g^{[\ell]}\left(z_2^{[\ell](i)}\right) \\ \vdots \\ g^{[\ell]}\left(z_{n^{[\ell]}}^{[\ell](i)}\right) \end{bmatrix} = \vec{a}^{[\ell](i)}$
- $G^{[\ell]}\left(Z^{[\ell]}\right) = \left[\vec{g}^{[\ell]}\left(\vec{z}^{[\ell](i)}\right) \quad \vec{g}^{[\ell]}\left(\vec{z}^{[\ell](2)}\right) \quad \dots \quad \vec{g}^{[\ell]}\left(\vec{z}^{[\ell](m)}\right)\right] = A^{[\ell]}$
- $\mathcal{L}\left(a_1^{[L](i)}, y^{(i)}\right)$ → Loss function for training sample i
- $\mathcal{J}\left(A^{[L]}, \vec{y}\right) = \frac{1}{m}\sum_{i=1}^{m} \mathcal{L}\left(a_1^{[L](i)}, y^{(i)}\right)$

The goal is to find the values in $W^{[\ell]}$ and $\vec{b}^{[\ell]}$, $\ell \in \{1,2,\dots,m\}$, that minimize the value of $\mathcal{J}\left(A^{[L]}, \vec{y}\right)$

We assume that the functions $\mathcal{L}(.,.)$ and $g^{[\ell]}(.)$, for $\ell \in \{1,2,\dots,L\}$ are designed/chosen such that:

- $\frac{\partial^2 \mathcal{J}(A^{[L]},\vec{y})}{\partial {w_{j,k}^{[\ell]}}^2} > 0$ for $j \in \{1,2,\dots,n^{[\ell]}\}$, $k \in \{1,2,\dots,n^{[\ell-1]}\}$, and $\ell \in \{1,2,\dots,L\}$
- $\frac{\partial \mathcal{J}^2(A^{[L]},\vec{y})}{\partial {b_j^{[\ell]}}^2} > 0$ for $j \in \{1,2,\dots,n^{[\ell]}\}$ and $\ell \in \{1,2,\dots,L\}$

In other words, we assume that $\mathcal{L}(.,.)$ and $g^{[\ell]}(.)$ are at least twice differentiable and result in making $\mathcal{J}(.,.)$ strictly convex with respect to the parameters we wish to estimate.

Using the gradient descent methodology, via multivariate calculus and linear algebra, we obtain the following results:

- $\frac{\partial \mathcal{J}(A^{[L]\{n-1\}},\vec{y})}{\partial W^{[\ell]\{n-1\}}} = \frac{1}{m} \times \left(\frac{\partial \mathcal{L}(A^{[L]\{n-1\}},\vec{y})}{\partial A^{[\ell]\{n-1\}}} \circ G^{[\ell]\prime}\left(Z^{[\ell]\{n-1\}}\right) \right) \times A^{[\ell-1]\{n-1\}^T}$
- $\frac{\partial \mathcal{J}(A^{[L]\{n-1\}},\vec{y})}{\partial \vec{b}^{[\ell]\{n-1\}}} = \frac{1}{m} \times \left(\frac{\partial \mathcal{L}(A^{[L]\{n-1\}},\vec{y})}{\partial A^{[\ell]\{n-1\}}} \circ G^{[\ell]\prime}\left(Z^{[\ell]\{n-1\}}\right) \right) \times \mathbf{1}_{m\times 1}$
- $\frac{\partial \mathcal{L}(A^{[L]\{n-1\}},\vec{y})}{\partial A^{[\ell]\{n-1\}}} = W^{[\ell+1]\{n-1\}^T} \times \frac{\partial \mathcal{L}(A^{[L]\{n-1\}},\vec{y})}{\partial Z^{[\ell+1]\{n-1\}}}$
- $\frac{\partial \mathcal{L}(A^{[L]\{n-1\}},\vec{y})}{\partial Z^{[\ell+1]}} = \frac{\partial \mathcal{L}(A^{[L]\{n-1\}},\vec{y})}{\partial A^{[\ell+1]\{n-1\}}} \circ G^{[\ell+1]\prime}\left(Z^{[\ell+1]\{n-1\}}\right)$

The focus of this paper is to determine distributional behaviours for $W^{[\ell]\{0\}}$ for any ℓ, that would promote stability of the calculations (i.e., neither too big nor too small activation calculations) for two expressions of $G^{[\ell-1]}\left(Z^{[\ell-1]\{0\}}\right)$:

- $G^{[\ell-1]}\left(Z^{[\ell-1]\{0\}}\right)$ is a matrix of ReLU functions ($f(x) = \max(0,x)$)
- $G^{[\ell-1]}\left(Z^{[\ell-1]\{0\}}\right)$ is a matrix of symmetric functions centred around zero and have a finite area under the curve ($f(-x) = \text{f(x)}$ and $\int_{-\infty}^{0} f(x)dx = \int_{0}^{\infty} f(x)dx$, $\left|\int_{-\infty}^{\infty} f(x)dx\right| < \infty$)

We derive the distributional behaviours of $W^{[\ell]\{0\}}$ that promote stability for the two steps of gradient descent: forward propagation and backward propagation. We note that it is possible that the distribution of $W^{[\ell]\{0\}}$ under the two case are not identical (which is the case). If that is the case (which it is), we will conduct an analysis to check if choosing either approach will still result in stability when conducting forward and backward propagation.

For any $2 \le \ell \le L$ (we are assuming that $L \ge 2$), we set $W^{[\ell]\{0\}}$ as follows:

- All elements of $W^{[\ell]\{0\}}$ are random normally distributed variables that are independent and identically (iid) and mean zero and a variance yet to be determined
- All elements of $A^{[\ell-1]\{0\}}$ are random variables that are iid
- All elements of $W^{[\ell]\{0\}}$and $A^{[\ell-1]\{0\}}$ are independent of each other
- $\vec{b}^{[\ell]\{0\}}$ is a vector of zeroes
- $\mathcal{L}\left(A^{[L]\{0\}},\vec{y}\right)$ is chosen such that all elements in $\frac{\partial \mathcal{L}(A^{[L]\{0\}},\vec{y})}{\partial A^{[\ell]\{0\}}}$ and $G^{[\ell]\prime}\left(Z^{[\ell]\{0\}}\right)$ are independent of each other

2. Forward Propagation Case

Going to the element-wise level, we have

$$z_j^{[\ell](i)\{0\}} = \vec{w}_j^{[\ell]\{0\}T}\vec{a}^{[\ell-1](i)\{0\}} + b_j^{[\ell]\{0\}} = \sum_{k=1}^{n^{[\ell-1]}} w_{j,k}^{[\ell]\{0\}} a_k^{[\ell-1](i)\{0\}}$$

Thus, taking the variance of $z_j^{[\ell](i)\{0\}}$, adopting the commonly used notations of $Var(.)$ and $E[.]$ to respectively denote variance and expectations, we proceed through the following derivations to arrive at an insightful result:

$$\begin{aligned}
Var\left(z_j^{[\ell](i)\{0\}}\right) &= Var\left(\sum_{k=1}^{n^{[\ell-1]}} w_{j,k}^{[\ell]\{0\}} a_k^{[\ell-1](i)\{0\}}\right) \\
&= \sum_{k=1}^{n^{[\ell-1]}} Var\left(w_{j,k}^{[\ell]\{0\}} a_k^{[\ell-1](i)\{0\}}\right) \\
&= \sum_{k=1}^{n^{[\ell-1]}} Var\left(w_{j,1}^{[\ell]\{0\}} a_1^{[\ell-1](i)\{0\}}\right) \\
&= n^{[\ell-1]} Var\left(w_{j,1}^{[\ell]\{0\}} a_1^{[\ell-1](i)\{0\}}\right) \\
&= n^{[\ell-1]}\left(E\left[\left(w_{j,1}^{[\ell]\{0\}} a_1^{[\ell-1](i)\{0\}}\right)^2\right] - E\left[w_{j,1}^{[\ell]\{0\}} a_1^{[\ell-1](i)\{0\}}\right]^2\right) \\
&= n^{[\ell-1]}\left(E\left[\left(w_{j,1}^{[\ell]\{0\}}\right)^2\right] E\left[\left(a_1^{[\ell-1](i)\{0\}}\right)^2\right] - \left(E\left[w_{j,1}^{[\ell]\{0\}}\right] E\left[a_1^{[\ell-1](i)\{0\}}\right]\right)^2\right) \\
&= n^{[\ell-1]}\left(E\left[\left(w_{j,1}^{[\ell]\{0\}}\right)^2\right] E\left[\left(a_1^{[\ell-1](i)\{0\}}\right)^2\right] - \left(0 \times E\left[a_1^{[\ell-1](i)\{0\}}\right]\right)^2\right) \\
&= n^{[\ell-1]} E\left[\left(w_{j,1}^{[\ell]\{0\}}\right)^2\right] E\left[\left(a_1^{[\ell-1](i)\{0\}}\right)^2\right] \\
&= n^{[\ell-1]} Var\left(w_{j,1}^{[\ell]\{0\}}\right) E\left[\left(a_1^{[\ell-1](i)\{0\}}\right)^2\right] \\
&= n^{[\ell-1]} Var\left(w_{j,1}^{[\ell]\{0\}}\right) E\left[\left(g^{[\ell-1]}\left(z_1^{[\ell-1](i)\{0\}}\right)\right)^2\right]
\end{aligned}$$

We now examine $g^{[\ell-1]}(.)$ under the two cases:

2.1 $g^{[\ell-1]}(x) = \max(0, x)$:

Note that we have

$$g^{[\ell-1]}\left(z_1^{[\ell-1](i)\{0\}}\right) = \begin{cases} z_1^{[\ell-1](i)\{0\}} & if\ z_1^{[\ell-1](i)\{0\}} > 0 \\ 0 & if\ z_1^{[\ell-1](i)\{0\}} \leq 0 \end{cases}$$

Now, since $w_{j,k}^{[\ell-1]\{0\}}$ is assumed to be normally distributed with mean zero and each weight is independent from other weights, then, if we are given the values of $a_k^{[\ell-2](i)\{0\}}$, the expression $\sum_{k=1}^{n^{[\ell-2]}} w_{j,k}^{[\ell-1]\{0\}} a_k^{[\ell-2](i)\{0\}}$ is also normally distributed with mean zero and symmetric (this can be proved via moment generating functions or convolutions or random variables).

Thus, we have the following result:

$$\Pr\left\{\sum_{k=1}^{n^{[\ell-2]}} w_{j,k}^{[\ell-1]\{0\}} a_k^{[\ell-2](i)\{0\}} > 0 | \left\{a_k^{[\ell-2](i)\{0\}} | k \in \{1,2,..,n^{[\ell-1]}\}\right\}\right\} = \frac{1}{2}$$

Let $\mathcal{A}_{n^{[\ell-1]}} = \left\{a_k^{[\ell-2](i)\{0\}} | k \in \{1,2,..,n^{[\ell-1]}\}\right\}$ and $F_{\mathcal{A}_{n^{[\ell-1]}}}$ denote the cumulative distribution function (cdf) of $\mathcal{A}_{n^{[\ell-1]}}$. Thus, we can proceed as follows arriving at a useful insight:

$$\begin{aligned}\Pr\left\{z_1^{[\ell-1](i)\{0\}} > 0\right\} &= \int_{\forall \mathcal{A}_{n^{[\ell-1]}}} \Pr\left\{\sum_{k=1}^{n^{[\ell-2]}} w_{j,k}^{[\ell-1]\{0\}} a_k^{[\ell-2](i)\{0\}} > 0 | \mathcal{A}_{n^{[\ell-1]}}\right\} dF_{\mathcal{A}_{n^{[\ell-1]}}} \\ &= \int_{\forall \mathcal{A}_{n^{[\ell-1]}}} \frac{1}{2} dF_{\mathcal{A}_{n^{[\ell-1]}}} = \frac{1}{2} \int_{\forall \mathcal{A}_{n^{[\ell-1]}}} dF_{\mathcal{A}_{n^{[\ell-1]}}} \\ &= \frac{1}{2} \times 1 \\ &= \frac{1}{2}\end{aligned}$$

Thus, letting $F_{z_1^{[\ell-1](i)\{0\}}}$ denote the cdf of $F_{z_1^{[\ell-1](i)\{0\}}}$, we have the following derivations:

$$g^{[\ell-1]}\left(z_1^{[\ell-1](i)\{0\}}\right) = \begin{cases} z_1^{[\ell-1](i)\{0\}} \; if \; z_1^{[\ell-1](i)\{0\}} > 0 \; with \; probability \frac{1}{2} \\ 0 \qquad\qquad if \; z_1^{[\ell-1](i)\{0\}} \leq 0 \; with \; probability \frac{1}{2} \end{cases}$$

$$\begin{aligned}E\left[g^{[\ell-1]}\left(z_1^{[\ell-1](i)\{0\}}\right)\right] &= \int_0^{\infty} z_1^{[\ell-1](i)\{0\}} dF_{z_1^{[\ell-1](i)\{0\}}} = \frac{1}{2} \times \int_{-\infty}^{\infty} z_1^{[\ell-1](i)\{0\}} dF_{z_1^{[\ell-1](i)\{0\}}} = \frac{1}{2} \times 0 \\ &= 0\end{aligned}$$

$$\begin{aligned}E\left[\left(g^{[\ell-1]}\left(z_1^{[\ell-1](i)\{0\}}\right)\right)^2\right] &= \int_0^{\infty} \left(z_1^{[\ell-1](i)\{0\}}\right)^2 dF_{z_1^{[\ell-1](i)\{0\}}} \\ &= \frac{1}{2} \int_{-\infty}^{\infty} \left(z_1^{[\ell-1](i)\{0\}}\right)^2 dF_{z_1^{[\ell-1](i)\{0\}}} \\ &= \frac{1}{2} E\left[\left(z_1^{[\ell-1](i)\{0\}}\right)^2\right] \\ &= \frac{1}{2} Var\left(z_1^{[\ell-1](i)\{0\}}\right)\end{aligned}$$

Therefore, we have the following recursive result:

$$Var\left(z_j^{[\ell](i)\{0\}}\right) = \frac{1}{2} n^{[\ell-1]} \, Var\left(w_{j,1}^{[\ell]\{0\}}\right) Var\left(z_1^{[\ell-1](i)\{0\}}\right)$$

Thus, we have the following relationship:

$$Var\left(z_j^{[\ell_{end}](i)\{0\}}\right) = Var\left(z_1^{[\ell_{start}](i)\{0\}}\right) \prod_{\ell=\ell_{start}}^{\ell_{end}} \frac{1}{2} n^{[\ell-1]} Var\left(w_{j,1}^{[\ell]\{0\}}\right)$$

To promote stability, we would like $Var\left(z_j^{[\ell_{end}](i)\{0\}}\right) = Var\left(z_1^{[\ell_{start}](i)\{0\}}\right)$ (neither having a variance that explodes nor that disappears) and thus, we would like the following expression to hold

$\prod_{\ell=\ell_{start}}^{\ell_{end}} \frac{1}{2} n^{[\ell-1]} Var\left(w_{j,1}^{[\ell]\{0\}}\right) = 1.$

One way to ensure this is to have $\frac{1}{2} n^{[\ell-1]} Var\left(w_{j,1}^{[\ell]\{0\}}\right) = 1$ for each ℓ. Thus, we have the following formula:

$Var\left(w_{j,1}^{[\ell]\{0\}}\right) = \frac{2}{n^{[\ell-1]}}$.

Therefore, under this set-up, all element of $W^{[\ell]\{0\}}$ are normally distributed (and independent of each other) with mean zero and variance $\frac{2}{n^{[\ell-1]}}$.

2.2 $g^{[\ell-1]}(x) = x$:

Note that we have

$$g^{[\ell-1]}\left(z_1^{[\ell-1](i)\{0\}}\right) = z_1^{[\ell-1](i)\{0\}} = \sum_{k=1}^{n^{[\ell-2]}} w_{1,k}^{[\ell-1]\{0\}} a_k^{[\ell-2](i)\{0\}} = \sum_{k=1}^{n^{[\ell-2]}} w_{1,k}^{[\ell-1]\{0\}} z_k^{[\ell-2](i)\{0\}}$$

Thus,

$$E\left[g^{[\ell-1]}\left(z_1^{[\ell-1](i)\{0\}}\right)\right] = E\left[\sum_{k=1}^{n^{[\ell-2]}} w_{1,k}^{[\ell-1]\{0\}} z_k^{[\ell-2](i)\{0\}}\right] = \sum_{k=1}^{n^{[\ell-2]}} E\left[w_{1,k}^{[\ell-1]\{0\}}\right] E\left[z_k^{[\ell-2](i)\{0\}}\right]$$

Since, $E\left[w_{1,k}^{[\ell-1]\{0\}}\right]$, therefore $E\left[g^{[\ell-1]}\left(z_1^{[\ell-1](i)\{0\}}\right)\right] = 0$.

Thus, $E\left[\left(g^{[\ell-1]}\left(z_1^{[\ell-1](i)\{0\}}\right)\right)^2\right] = Var\left(g^{[\ell-1]}\left(z_1^{[\ell-1](i)\{0\}}\right)\right) = Var\left(z_1^{[\ell-1](i)\{0\}}\right)$

Therefore, we have the following recursive result:

$Var\left(z_j^{[\ell](i)\{0\}}\right) = n^{[\ell-1]}\, Var\left(w_{j,1}^{[\ell]\{0\}}\right) Var\left(z_1^{[\ell-1](i)\{0\}}\right)$

Thus, we have the following relationship:

$$Var\left(z_j^{[\ell_{end}](i)\{0\}}\right) = Var\left(z_1^{[\ell_{start}](i)\{0\}}\right) \prod_{\ell=\ell_{start}}^{\ell_{end}} n^{[\ell-1]} Var\left(w_{j,1}^{[\ell]\{0\}}\right)$$

To promote stability, we would like $Var\left(z_j^{[\ell_{end}](i)\{0\}}\right) = Var\left(z_1^{[\ell_{start}](i)\{0\}}\right)$ (neither having a variance that explodes nor that disappears) and thus, we would like the following expression to hold

$\prod_{\ell=\ell_{start}}^{\ell_{end}} n^{[\ell-1]} Var\left(w_{j,1}^{[\ell]\{0\}}\right) = 1.$

One way to ensure this is to have $n^{[\ell-1]} Var\left(w_{j,1}^{[\ell]\{0\}}\right) = 1$ for each ℓ. Thus, we have the following formula:
$Var\left(w_{j,1}^{[\ell]\{0\}}\right) = \frac{1}{n^{[\ell-1]}}$.

Therefore, under this set-up, all element of $W^{[\ell]\{0\}}$ are normally distributed (and independent of each other) with mean zero and variance $\frac{1}{n^{[\ell-1]}}$.

3. Backward Propagation Case

Going to the element-wise level, we have the following expressions (note that we switch the weight indices j and k since $\frac{\partial \mathcal{L}(A^{[L]\{0\}}, \vec{y})}{\partial z_j^{[\ell](i)\{0\}}}$ has $j \in \{1,2,\dots,n^{[\ell]}\}$ values):

$$\frac{\partial \mathcal{L}(A^{[L]\{0\}}, \vec{y})}{\partial a_j^{[\ell-1](i)\{0\}}} = \sum_{k=1}^{n^{[\ell]}} w_{k,j}^{[\ell]\{0\}} \frac{\partial \mathcal{L}(A^{[L]\{0\}}, \vec{y})}{\partial z_j^{[\ell](i)\{0\}}}$$

$$\frac{\partial \mathcal{L}(A^{[L]\{0\}}, \vec{y})}{\partial z_j^{[\ell](i)\{0\}}} = \frac{\partial \mathcal{L}(A^{[L]\{0\}}, \vec{y})}{\partial a_j^{[\ell](i)\{0\}}} \times g^{[\ell]\prime}\left(z_j^{[\ell](i)\{0\}}\right)$$

Thus, taking the variance of $\frac{\partial \mathcal{L}(A^{[L]\{0\}}, \vec{y})}{\partial a_j^{[\ell-1](i)\{0\}}}$, adopting the commonly used notations of $Var(.)$ and $E[.]$ to respectively denote variance and expectations, we proceed through the following derivations to arrive at insightful results:

$$E\left[\frac{\partial \mathcal{L}(A^{[L]\{0\}}, \vec{y})}{\partial a_j^{[\ell-1](i)\{0\}}}\right] = E\left[\sum_{k=1}^{n^{[\ell]}} w_{k,j}^{[\ell]\{0\}} \frac{\partial \mathcal{L}(A^{[L]\{0\}}, \vec{y})}{\partial z_j^{[\ell](i)\{0\}}}\right] = \sum_{k=1}^{n^{[\ell]}} E\left[w_{k,j}^{[\ell]\{0\}} \frac{\partial \mathcal{L}(A^{[L]\{0\}}, \vec{y})}{\partial z_j^{[\ell](i)\{0\}}}\right]$$

$$= \sum_{k=1}^{n^{[\ell]}} E\left[w_{k,j}^{[\ell]\{0\}}\right] \times E\left[\frac{\partial \mathcal{L}(A^{[L]\{0\}}, \vec{y})}{\partial z_j^{[\ell](i)\{0\}}}\right] = \sum_{k=1}^{n^{[\ell]}} 0 \times E\left[\frac{\partial \mathcal{L}(A^{[L]\{0\}}, \vec{y})}{\partial z_j^{[\ell](i)\{0\}}}\right] = 0$$

$$Var\left(\frac{\partial \mathcal{L}(A^{[L]\{0\}}, \vec{y})}{\partial a_j^{[\ell-1](i)\{0\}}}\right) = Var\left(\sum_{k=1}^{n^{[\ell]}} w_{k,j}^{[\ell]\{0\}} \frac{\partial \mathcal{L}(A^{[L]\{0\}}, \vec{y})}{\partial z_j^{[\ell](i)\{0\}}}\right)$$

$$= \sum_{k=1}^{n^{[\ell]}} Var\left(w_{k,j}^{[\ell]\{0\}} \frac{\partial \mathcal{L}(A^{[L]\{0\}}, \vec{y})}{\partial z_j^{[\ell](i)\{0\}}}\right)$$

$$= \sum_{k=1}^{n^{[\ell]}} Var\left(w_{1,j}^{[\ell]\{0\}} \frac{\partial \mathcal{L}(A^{[L]\{0\}}, \vec{y})}{\partial z_1^{[\ell](i)\{0\}}}\right)$$

$$= n^{[\ell]} Var\left(w_{1,j}^{[\ell]\{0\}} \frac{\partial \mathcal{L}\left(A^{[L]\{0\}}, \vec{y}\right)}{\partial z_1^{[\ell](i)\{0\}}}\right)$$

$$= n^{[\ell]} \left(E\left[\left(w_{1,j}^{[\ell]\{0\}} \frac{\partial \mathcal{L}\left(A^{[L]\{0\}}, \vec{y}\right)}{\partial z_1^{[\ell](i)\{0\}}}\right)^2\right] - E\left[w_{1,j}^{[\ell]\{0\}} \frac{\partial \mathcal{L}\left(A^{[L]\{0\}}, \vec{y}\right)}{\partial z_1^{[\ell](i)\{0\}}}\right]^2\right)$$

$$= n^{[\ell]} \left(E\left[\left(w_{1,j}^{[\ell]\{0\}}\right)^2\right] E\left[\left(\frac{\partial \mathcal{L}\left(A^{[L]\{0\}}, \vec{y}\right)}{\partial z_1^{[\ell](i)\{0\}}}\right)^2\right] - \left(E\left[w_{1,j}^{[\ell]\{0\}}\right] E\left[\frac{\partial \mathcal{L}\left(A^{[L]\{0\}}, \vec{y}\right)}{\partial z_1^{[\ell](i)\{0\}}}\right]\right)^2\right)$$

$$= n^{[\ell]} \left(Var\left(w_{1,j}^{[\ell]\{0\}}\right) E\left[\left(\frac{\partial \mathcal{L}\left(A^{[L]\{0\}}, \vec{y}\right)}{\partial z_1^{[\ell](i)\{0\}}}\right)^2\right] - \left(E\left[w_{1,j}^{[\ell]\{0\}}\right] E\left[\frac{\partial \mathcal{L}\left(A^{[L]\{0\}}, \vec{y}\right)}{\partial z_1^{[\ell](i)\{0\}}}\right]\right)^2\right)$$

$$= n^{[\ell]} \left(Var\left(w_{1,j}^{[\ell]\{0\}}\right) E\left[\left(\frac{\partial \mathcal{L}\left(A^{[L]\{0\}}, \vec{y}\right)}{\partial z_1^{[\ell](i)\{0\}}}\right)^2\right] - \left(0 \times E\left[\frac{\partial \mathcal{L}\left(A^{[L]\{0\}}, \vec{y}\right)}{\partial z_1^{[\ell](i)\{0\}}}\right]\right)^2\right)$$

$$= n^{[\ell]}\, Var\left(w_{1,j}^{[\ell]\{0\}}\right) E\left[\left(\frac{\partial \mathcal{L}\left(A^{[L]\{0\}}, \vec{y}\right)}{\partial z_1^{[\ell](i)\{0\}}}\right)^2\right]$$

$$= n^{[\ell]}\, Var\left(w_{1,j}^{[\ell]\{0\}}\right) E\left[\left(\frac{\partial \mathcal{L}\left(A^{[L]\{0\}}, \vec{y}\right)}{\partial a_j^{[\ell](i)\{0\}}} \times g^{[\ell]\prime}\left(z_j^{[\ell](i)\{0\}}\right)\right)^2\right]$$

$$= n^{[\ell]}\, Var\left(w_{1,j}^{[\ell]\{0\}}\right) \times E\left[\left(\frac{\partial \mathcal{L}\left(A^{[L]\{0\}}, \vec{y}\right)}{\partial a_j^{[\ell](i)\{0\}}}\right)^2\right] \times E\left[\left(g^{[\ell]\prime}\left(z_j^{[\ell](i)\{0\}}\right)\right)^2\right]$$

$$= n^{[\ell]}\, Var\left(w_{1,j}^{[\ell]\{0\}}\right) \times \left(Var\left(\frac{\partial \mathcal{L}\left(A^{[L]\{0\}}, \vec{y}\right)}{\partial a_j^{[\ell](i)\{0\}}}\right) + \left(E\left[\frac{\partial \mathcal{L}\left(A^{[L]\{0\}}, \vec{y}\right)}{\partial a_j^{[\ell](i)\{0\}}}\right]\right)^2\right) \times E\left[\left(g^{[\ell]\prime}\left(z_j^{[\ell](i)\{0\}}\right)\right)^2\right]$$

$$= n^{[\ell]}\, Var\left(w_{1,j}^{[\ell]\{0\}}\right) \times \left(Var\left(\frac{\partial \mathcal{L}\left(A^{[L]\{0\}}, \vec{y}\right)}{\partial a_j^{[\ell](i)\{0\}}}\right) + (0)^2\right) \times E\left[\left(g^{[\ell]\prime}\left(z_j^{[\ell](i)\{0\}}\right)\right)^2\right]$$

$$= n^{[\ell]}\, Var\left(w_{1,j}^{[\ell]\{0\}}\right) \times \left(Var\left(\frac{\partial \mathcal{L}\left(A^{[L]\{0\}}, \vec{y}\right)}{\partial a_j^{[\ell](i)\{0\}}}\right) + 0\right) \times E\left[\left(g^{[\ell]\prime}\left(z_j^{[\ell](i)\{0\}}\right)\right)^2\right]$$

$$= n^{[\ell]}\, Var\left(w_{1,j}^{[\ell]\{0\}}\right) \times Var\left(\frac{\partial \mathcal{L}\left(A^{[L]\{0\}}, \vec{y}\right)}{\partial a_j^{[\ell](i)\{0\}}}\right) \times E\left[\left(g^{[\ell]\prime}\left(z_j^{[\ell](i)\{0\}}\right)\right)^2\right]$$

We now examine $g^{[\ell]\prime}(.)$ under the two cases:

3.1 $g^{[\ell]\prime}(x) = \frac{\partial \max(0,x)}{\partial x}$

Note that we have

$$g^{[\ell]'}\left(z_1^{[\ell](i)\{0\}}\right) = \begin{cases} 1 & if\ z_1^{[\ell](i)\{0\}} > 0 \\ 0 & if\ z_1^{[\ell](i)\{0\}} \leq 0 \end{cases}$$

Thus,

$$\begin{aligned} E\left[\left(g^{[\ell]'}\left(z_j^{[\ell](i)\{0\}}\right)\right)^2\right] &= 1^2 \times \Pr\left\{z_1^{[\ell](i)\{0\}} > 0\right\} + 0^2 \times \Pr\left\{z_1^{[\ell](i)\{0\}} \leq 0\right\} \\ &= \Pr\left\{z_1^{[\ell](i)\{0\}} > 0\right\} + 0 \\ &= \Pr\left\{z_1^{[\ell](i)\{0\}} > 0\right\} \end{aligned}$$

Using the result we obtained from subsection 2.1, we showed that $\Pr\left\{z_1^{[\ell](i)\{0\}} > 0\right\} = \frac{1}{2}$.

Thus, $E\left[\left(g^{[\ell]'}\left(z_j^{[\ell](i)\{0\}}\right)\right)^2\right] = \frac{1}{2}$.

Therefore, we have the following recursive result:

$$Var\left(\frac{\partial \mathcal{L}\left(A^{[L]\{0\}}, \vec{y}\right)}{\partial a_j^{[\ell-1](i)\{0\}}}\right) = n^{[\ell]}\, Var\left(w_{1,j}^{[\ell]\{0\}}\right) \times Var\left(\frac{\partial \mathcal{L}\left(A^{[L]\{0\}}, \vec{y}\right)}{\partial a_j^{[\ell](i)\{0\}}}\right) \times \frac{1}{2}$$

Thus, we have the following relationship:

$$Var\left(\frac{\partial \mathcal{L}\left(A^{[L]\{0\}}, \vec{y}\right)}{\partial a_j^{[\ell_{\text{start}}](i)\{0\}}}\right) = Var\left(\frac{\partial \mathcal{L}\left(A^{[L]\{0\}}, \vec{y}\right)}{\partial a_j^{[\ell_{\text{end}}](i)\{0\}}}\right) \prod_{\ell=\ell_{start}+1}^{\ell_{end}} \frac{1}{2} n^{[\ell]} Var\left(w_{j,1}^{[\ell]\{0\}}\right)$$

To promote stability, we would like $Var\left(\frac{\partial \mathcal{L}\left(A^{[L]\{0\}}, \vec{y}\right)}{\partial a_j^{[\ell_{\text{start}}](i)\{0\}}}\right) = Var\left(\frac{\partial \mathcal{L}\left(A^{[L]\{0\}}, \vec{y}\right)}{\partial a_j^{[\ell_{\text{end}}](i)\{0\}}}\right)$ (neither having a variance that explodes nor that disappears) and thus, we would like the following expression to hold

$\prod_{\ell=\ell_{start}+1}^{\ell_{end}} \frac{1}{2} n^{[\ell]} Var\left(w_{j,1}^{[\ell]\{0\}}\right) = 1.$

One way to ensure this is to have $\frac{1}{2} n^{[\ell]} Var\left(w_{j,1}^{[\ell]\{0\}}\right) = 1$ for each ℓ. Thus, we have the following formula:

$Var\left(w_{j,1}^{[\ell]\{0\}}\right) = \frac{2}{n^{[\ell]}}$.

3.2 $g^{[\ell]'}(x) = \frac{\partial x}{\partial x}$:

Note that we have

$$g^{[\ell-1]'}\left(z_1^{[\ell-1](i)\{0\}}\right) = \frac{\partial z_1^{[\ell-1](i)\{0\}}}{\partial z_1^{[\ell-1](i)\{0\}}} = 1$$

Thus, $E\left[\left(g^{[\ell]'}\left(z_j^{[\ell](i)\{0\}}\right)\right)^2\right] = E[(1)^2] = E[1] = 1.$

Therefore, we have the following recursive result:

$$Var\left(\frac{\partial \mathcal{L}\left(A^{[L]\{0\}},\vec{y}\right)}{\partial a_j^{[\ell-1](i)\{0\}}}\right) = n^{[\ell]}\, Var\left(w_{1,j}^{[\ell]\{0\}}\right) \times Var\left(\frac{\partial \mathcal{L}\left(A^{[L]\{0\}},\vec{y}\right)}{\partial a_j^{[\ell](i)\{0\}}}\right) \times 1$$

Thus, we have the following relationship:

$$Var\left(\frac{\partial \mathcal{L}\left(A^{[L]\{0\}},\vec{y}\right)}{\partial a_j^{[\ell_{\text{start}}](i)\{0\}}}\right) = Var\left(\frac{\partial \mathcal{L}\left(A^{[L]\{0\}},\vec{y}\right)}{\partial a_j^{[\ell_{\text{end}}](i)\{0\}}}\right) \prod_{\ell=\ell_{start}+1}^{\ell_{end}} n^{[\ell]} Var\left(w_{j,1}^{[\ell]\{0\}}\right)$$

To promote stability, we would like $Var\left(\frac{\partial \mathcal{L}\left(A^{[L]\{0\}},\vec{y}\right)}{\partial a_j^{[\ell_{\text{start}}](i)\{0\}}}\right) = Var\left(\frac{\partial \mathcal{L}\left(A^{[L]\{0\}},\vec{y}\right)}{\partial a_j^{[\ell_{\text{end}}](i)\{0\}}}\right)$ (neither having a variance that explodes nor that disappears) and thus, we would like the following expression to hold

$\prod_{\ell=\ell_{start}+1}^{\ell_{end}} n^{[\ell]} Var\left(w_{j,1}^{[\ell]\{0\}}\right) = 1.$

One way to ensure this is to have $n^{[\ell]} Var\left(w_{j,1}^{[\ell]\{0\}}\right) = 1$ for each ℓ. Thus, we have the following formula:

$Var\left(w_{j,1}^{[\ell]\{0\}}\right) = \frac{1}{n^{[\ell]}}$.

4. Analyzing the Differences between the Two Cases

2.1 ReLU

We inter-substitute $Var\left(w_{j,1}^{[\ell]\{0\}}\right)$ into the recursive variance formulas derived in the forward and backward propagation case sections and examine how stable the variance formulas are. If they are stable, then the two derivations of $Var\left(w_{j,1}^{[\ell]\{0\}}\right)$ are compatible with each other:

In the forward propagation case, we have the following variance formula we aim to stabilize:

$$Var\left(z_j^{[\ell_{end}](i)\{0\}}\right) = Var\left(z_1^{[\ell_{start}](i)\{0\}}\right) \prod_{\ell=\ell_{start}}^{\ell_{end}} \frac{1}{2} n^{[\ell-1]} Var\left(w_{j,1}^{[\ell]\{0\}}\right)$$

If we use $Var\left(w_{j,1}^{[\ell]\{0\}}\right) = \frac{2}{n^{[\ell]}}$ from the backward propagation case (instead of the forward propagation case $Var\left(w_{j,1}^{[\ell]\{0\}}\right) = \frac{2}{n^{[\ell-1]}}$), we have the following result:

$$\begin{aligned} Var\left(z_j^{[\ell_{end}](i)\{0\}}\right) &= Var\left(z_1^{[\ell_{start}](i)\{0\}}\right) \prod_{\ell=\ell_{start}}^{\ell_{end}} \frac{1}{2} n^{[\ell-1]} \times \left(\frac{2}{n^{[\ell]}}\right) \\ &= Var\left(z_1^{[\ell_{start}](i)\{0\}}\right) \prod_{\ell=\ell_{start}}^{\ell_{end}} \left(\frac{n^{[\ell-1]}}{n^{[\ell]}}\right) \\ &= Var\left(z_1^{[\ell_{start}](i)\{0\}}\right) \left(\frac{n^{[\ell_{start}-1]}}{n^{[\ell_{end}]}}\right) \end{aligned}$$

This expression can be deemed stable in the sense that it does not yield an explosion or disappearance of variance value as the number of layers increases. Therefore, the weight distribution stable in the backward propagation case is also stable when conducting forward propagation.

We now examine the opposite case:
In the backward propagation case, we have the following variance formula we aim to stabilize:

$$Var\left(\frac{\partial \mathcal{L}\left(A^{[L]\{0\}},\vec{y}\right)}{\partial a_j^{[\ell_{\text{start}}](i)\{0\}}}\right) = Var\left(\frac{\partial \mathcal{L}\left(A^{[L]\{0\}},\vec{y}\right)}{\partial a_j^{[\ell_{\text{end}}](i)\{0\}}}\right) \prod_{\ell=\ell_{start}+1}^{\ell_{end}} \frac{1}{2} n^{[\ell]} Var\left(w_{j,1}^{[\ell]\{0\}}\right)$$

If we use $Var\left(w_{j,1}^{[\ell]\{0\}}\right) = \frac{2}{n^{[\ell-1]}}$ from the forward propagation case (instead of the backward propagation case $Var\left(w_{j,1}^{[\ell]\{0\}}\right) = \frac{2}{n^{[\ell]}}$), we have the following result:

$$\begin{aligned} Var\left(\frac{\partial \mathcal{L}\left(A^{[L]\{0\}},\vec{y}\right)}{\partial a_j^{[\ell_{\text{start}}](i)\{0\}}}\right) &= Var\left(\frac{\partial \mathcal{L}\left(A^{[L]\{0\}},\vec{y}\right)}{\partial a_j^{[\ell_{\text{end}}](i)\{0\}}}\right) \prod_{\ell=\ell_{start}+1}^{\ell_{end}} \frac{1}{2} n^{[\ell]} \times \left(\frac{2}{n^{[\ell-1]}}\right) \\ &= Var\left(\frac{\partial \mathcal{L}\left(A^{[L]\{0\}},\vec{y}\right)}{\partial a_j^{[\ell_{\text{end}}](i)\{0\}}}\right) \prod_{\ell=\ell_{start}+1}^{\ell_{end}} \left(\frac{n^{[\ell]}}{n^{[\ell-1]}}\right) \\ &= Var\left(z_1^{[\ell_{start}](i)\{0\}}\right)\left(\frac{n^{[\ell_{end}]}}{n^{[\ell_{start}-1]}}\right) \end{aligned}$$

This expression can be deemed stable in the sense that it does not yield an explosion or disappearance of variance value as the number of layers increases. Therefore, the weight distribution stable in the forward propagation case is also stable when conducting backward propagation.

3.1 Linear

We inter-substitute $Var\left(w_{j,1}^{[\ell]\{0\}}\right)$ into the recursive variance formulas derived in the forward and backward propagation case sections and examine how stable the variance formulas are. If they are stable, then the two derivations of $Var\left(w_{j,1}^{[\ell]\{0\}}\right)$ are compatible with each other:

In the forward propagation case, we have the following variance formula we aim to stabilize:

$$Var\left(z_j^{[\ell_{end}](i)\{0\}}\right) = Var\left(z_1^{[\ell_{start}](i)\{0\}}\right) \prod_{\ell=\ell_{start}}^{\ell_{end}} n^{[\ell-1]} Var\left(w_{j,1}^{[\ell]\{0\}}\right)$$

If we use $Var\left(w_{j,1}^{[\ell]\{0\}}\right) = \frac{1}{n^{[\ell]}}$ from the backward propagation case (instead of the forward case $Var\left(w_{j,1}^{[\ell]\{0\}}\right) = \frac{1}{n^{[\ell-1]}}$), we have the following result:

$$Var\left(z_j^{[\ell_{end}](i)\{0\}}\right) = Var\left(z_1^{[\ell_{start}](i)\{0\}}\right) \prod_{\ell=\ell_{start}}^{\ell_{end}} n^{[\ell-1]} \times \left(\frac{1}{n^{[\ell]}}\right)$$

$$= Var\left(z_1^{[\ell_{start}](i)\{0\}}\right) \prod_{\ell=\ell_{start}}^{\ell_{end}} \left(\frac{n^{[\ell-1]}}{n^{[\ell]}}\right)$$
$$= Var\left(z_1^{[\ell_{start}](i)\{0\}}\right)\left(\frac{n^{[\ell_{start}-1]}}{n^{[\ell_{end}]}}\right)$$

This expression can be deemed stable in the sense that it does not yield an explosion or disappearance of variance value as the number of layers increases. Therefore, the weight distribution stable in the backward propagation case is also stable when conducting forward propagation.

We now examine the opposite case:
In the backward propagation case, we have the following variance formula we aim to stabilize:

$$Var\left(\frac{\partial \mathcal{L}\left(A^{[L]\{0\}}, \vec{y}\right)}{\partial a_j^{[\ell_{\text{start}}](i)\{0\}}}\right) = Var\left(\frac{\partial \mathcal{L}\left(A^{[L]\{0\}}, \vec{y}\right)}{\partial a_j^{[\ell_{\text{end}}](i)\{0\}}}\right) \prod_{\ell=\ell_{start}+1}^{\ell_{end}} n^{[\ell]} Var\left(w_{j,1}^{[\ell]\{0\}}\right)$$

If we use $Var\left(w_{j,1}^{[\ell]\{0\}}\right) = \frac{1}{n^{[\ell-1]}}$ from the forward propagation case (instead of the backward propagation case $Var\left(w_{j,1}^{[\ell]\{0\}}\right) = \frac{1}{n^{[\ell]}}$), we have the following result:

$$Var\left(\frac{\partial \mathcal{L}\left(A^{[L]\{0\}}, \vec{y}\right)}{\partial a_j^{[\ell_{\text{start}}](i)\{0\}}}\right) = Var\left(\frac{\partial \mathcal{L}\left(A^{[L]\{0\}}, \vec{y}\right)}{\partial a_j^{[\ell_{\text{end}}](i)\{0\}}}\right) \prod_{\ell=\ell_{start}+1}^{\ell_{end}} n^{[\ell]} \times \left(\frac{1}{n^{[\ell-1]}}\right)$$
$$= Var\left(\frac{\partial \mathcal{L}\left(A^{[L]\{0\}}, \vec{y}\right)}{\partial a_j^{[\ell_{\text{end}}](i)\{0\}}}\right) \prod_{\ell=\ell_{start}+1}^{\ell_{end}} \left(\frac{n^{[\ell]}}{n^{[\ell-1]}}\right)$$
$$= Var\left(z_1^{[\ell_{start}](i)\{0\}}\right)\left(\frac{n^{[\ell_{end}]}}{n^{[\ell_{start}-1]}}\right)$$

This expression can be deemed stable in the sense that it does not yield an explosion or disappearance of variance value as the number of layers increases. Therefore, the weight distribution stable in the forward propagation case is also stable when conducting backward propagation.