Derivation of He initialization

Neural networks for people who get confused easily

James McCammon

June 11, 2024

1 Introduction

We define the following indices:

- i is the index of the neurons in the current layer (the rows represent the weights of each neuron in the current layer).
- j is the index of the neurons in the *previous* layer, which are the inputs to this layer (each column represents an input from the previous layer, somewhat equivalent to a feature input).

Thus, row 1 represents all of the weights for the first neuron. Column 1 represents how each neuron in the current layer weighs or processes the first "feature" or input (i.e., the first neuron in the previous layer's connection to each neuron in the current layer).

Assume weights are independent Assume h is not

2 Detailed Explanation of He Initialization

Introduction: In He initialization, the aim is to set up initial weights for a neural network in such a way that the variance of the outputs from each neuron approximates unity, preventing the gradient vanishing or exploding problems.

2.1 Derivation of the mean

With the He initialization, we assume our biases are set to zero and we're drawing our weights from a normal distribution with mean $\mu = 0$ and a variance σ^2 that needs to be calculated. First, let's start by finding the expected value of f given our weights have $\mu = 0$.

$$\mathbb{E}[f_i'] = \mathbb{E}\left[\beta_i + \sum_{j=1}^{D_h} \Omega_{i,j} h_i\right] \qquad \text{(expectation of a sum)} \tag{1}$$

$$= \mathbb{E}[\beta_i] + \mathbb{E}\left[\sum_{j=1}^{D_h} \Omega_{i,j} h_j\right] \qquad \text{(sum of expectations)} \tag{2}$$

$$= 0 + \mathbb{E}\left[\sum_{j=1}^{D_h} \Omega_{i,j} h_j\right] \qquad (\mathbb{E}[\beta_i] = 0) \tag{3}$$

$$= \sum_{j=1}^{D_h} \mathbb{E}[\Omega_{i,j} h_j] \qquad \text{(sum of expectations)} \tag{4}$$

$$= \sum_{j=1}^{D_h} \mathbb{E}[\Omega_{i,j}]\mathbb{E}[h_j] \qquad \text{(independence)} \tag{5}$$

$$= \sum_{j=1}^{D_h} 0 \cdot \mathbb{E}[h_j] \tag{6}$$

$$\mathbb{E}[f_i'] = 0 \tag{7}$$

Explanation: Here, in both Steps 2 and 4 we used the fact that the expectation of a sum is the sum of the expectations. That is, $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$. In Step 3 we used the fact that $\mathbb{E}[\beta_i] = 0$. This is by assumption because we have decided to set our biases to 0 initially. In Step 5 we used the fact that $\Omega_{i,j}$ and h_j are independent and for independent expectations $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$. Finally, in Step 6 we have used the fact that $\mathbb{E}[\Omega_{i,j}] = 0$. Again, this by assumption since we said that our weights are drawn for a normal distribution with mean 0.

2.2 Derivation of a useful variance identity

There is one identify we'll need to derive the variance for the He initialization. Let's derive that for practice. We start with the definition of the variance, σ^2 which is the expected value of the squared difference between the random variable X and its expected value $\mathbb{E}[X]$. This makes sense, the variance is just how far our random variable is deviating from its expected value. Squaring this deviation – as opposed to taking the absolute value for example – ensure the variance is both non-negative and that larger variances are penalized more. While we're here, a frequent confusion is between X and x. Sometimes you might see something like $\mathbb{P}(X = x)$. X is the random variable which can take a specific value, x is the specific value the random variable is taking. Here we're using X because the variance is a property of the random variable itself, not a property of a specific value.

$$
\begin{aligned}
\sigma^2 &= \mathbb{E}[(X - \mathbb{E}[X])^2] && && (1)\\
&= \mathbb{E}[(X^2 - 2X \cdot \mathbb{E}[X] + \mathbb{E}[X]^2)] && \text{(expanding the square)} && (2)\\
&= \mathbb{E}[X^2] - \mathbb{E}[2X \cdot \mathbb{E}[X]] + \mathbb{E}[\mathbb{E}[X]^2] && \text{(sum of expectations)} && (3)\\
&= \mathbb{E}[X^2] - 2\mathbb{E}[X]\mathbb{E}[X] + \mathbb{E}[X]^2 && (\mathbb{E}[X] \text{ is a constant}) && (4)\\
&= \mathbb{E}[X^2] - 2(\mathbb{E}[X])^2 + (\mathbb{E}[X])^2 && \text{(simplification)} && (5)\\
&= \mathbb{E}[X^2] - \mathbb{E}[X]^2 && \text{(simplification)} && (6)
\end{aligned}
$$

Explanation: In Equation 1 we start with the definition of the variance. Moving to Equation 2 we expands the square inside the expectation. In Equation 3 we once again use the fact that the expectation of the sum is the sum of the expectations. The logic from Equation 3 to Equation 4 is more subtle. While X is a random variable, $\mathbb{E}[X]$ is just a number. It's the mean for whatever distribution is chosen for X. This means we get to pull the inner expectation outside of the outer expectation as we would any other constant. The general rule is $\mathbb{E}[\alpha X] = \alpha\mathbb{E}[X]$, where in this case α is the inner $\mathbb{E}[X]$ in Equation 3. Of course, we can also move the 2 outside the outter expectation since it too is a constant. In Equation 5 we're just combining the two $\mathbb{E}[X]$ that are multiplied in the middle term of Equation 4. Equation 6 is simple subtraction, we have a $-2\mathbb{E}[X]^2$ term and a $+\mathbb{E}[X]^2$ term, which leaves us with a single $-\mathbb{E}[X]^2$.

2.3 Derivation of a Useful Fact

We start by expanding the square of the sum of terms.

2.3.1 Expanding the Square

$$
\begin{aligned}
\left(\sum_{j=1}^{D_h} \Omega_{i,j} h_i\right)^2 &= (\Omega_{i,1}h_1 + \Omega_{i,2}h_2 + \cdots + \Omega_{i,D_h}h_{D_h})^2\\
&= (\Omega_{i,1}h_1)^2 + \Omega_{i,1}h_1 \cdot \Omega_{i,2}h_2 + \cdots + \Omega_{i,1}h_1 \cdot \Omega_{i,D_h}h_{D_h}\\
&\qquad + \Omega_{i,2}h_2 \cdot \Omega_{i,1}h_1 + (\Omega_{i,2}h_2)^2 + \cdots + (\Omega_{i,D_h}h_{D_h})^2
\end{aligned}
\tag{1}
$$

2.3.2 Analysis of Expectations

Taking Expectations: Now let's think about what would happen if we took the expectation of this large function. Recall that the expectation of a sum is the sum of expectations. Therefore, we end up with D_h^2 expectations.

Considering Non-Squared Terms: Consider a non-squared term such as $\Omega_{i,1}h_1 \cdot \Omega_{i,2}h_2$. First, note that we can gather like terms and rewrite this as $\Omega_{i,1}\Omega_{i,2} \cdot h_1 h_2$. After applying the expectation, we have:

$$\mathbb{E}[\Omega_{i,1}\Omega_{i,2} \cdot h_1 h_2]$$

Recall that Ω and h are assumed to be independent, and using the independence of expectations rule, we obtain:

$$\mathbb{E}[\Omega_{i,1}\Omega_{i,2} \cdot h_1 h_2] = \mathbb{E}[\Omega_{i,1}\Omega_{i,2}] \cdot \mathbb{E}[h_1 h_2]$$

But our weights are *also* independent from one another at initialization and so we can again apply our rule to obtain:

$$\mathbb{E}[\Omega_{i,1}\Omega_{i,2} \cdot h_1 h_2] = \mathbb{E}[\Omega_{i,1}] \cdot \mathbb{E}[\Omega_{i,2}] \cdot \mathbb{E}[h_1 h_2]$$

Now, recall that $\mathbb{E}[\Omega_{i,j}] = 0$ because of our assumption that our weights are drawn from a distribution with mean 0. This means that the entire expression is zero. This logic can be applied to all non-squared terms, which go to 0, while we keep the squared terms, highlighted in yellow below.

$$\begin{aligned}
\mathbb{E}[(\Omega_{i,1}h_1)^2] &+ \mathbb{E}[\cancelto{0}{\Omega_{i,1}h_1 \cdot \Omega_{i,2}h_2}] + \mathbb{E}[\cancelto{0}{\Omega_{i,1}h_1 \cdot \Omega_{i,D_h}h_{D_h}}] \\
&+ \mathbb{E}[\cancelto{0}{\Omega_{i,2}h_2 \cdot \Omega_{i,1}h_1}] + \mathbb{E}[(\Omega_{i,2}h_2)^2] + \cdots + \mathbb{E}[(\Omega_{i,D_h}h_{D_h})^2]
\end{aligned}$$

Now let's return the original formulation of our equation using this new information. We now understand that: Recall that we start by expanding the expected value of the squared sum of weights and activations:

$$\mathbb{E}\left[\left(\sum_{j=1}^{D_h} \Omega_{i,j}h_i\right)^2\right] = \mathbb{E}\left[(\Omega_{i,1}h_1 + \Omega_{i,2}h_2 + \cdots + \Omega_{i,D_h}h_{D_h})^2\right]$$

Expanding the square:

$$\mathbb{E}\left[(\Omega_{i,1}h_1)^2 + \Omega_{i,1}h_1 \cdot \Omega_{i,2}h_2 + \cdots + \Omega_{i,1}h_1 \cdot \Omega_{i,D_h}h_{D_h} + \Omega_{i,2}h_2 \cdot \Omega_{i,1}h_1 + (\Omega_{i,2}h_2)^2 + \cdots + (\Omega_{i,D_h}h_{D_h})^2\right]$$

Now using the fact that the expectation of sums is the sum of expectations.

$$\begin{aligned}
&\mathbb{E}\left[(\Omega_{i,1}h_1)^2\right] + \mathbb{E}\left[\Omega_{i,1}h_1 \cdot \Omega_{i,2}h_2\right] + \cdots + \mathbb{E}\left[\Omega_{i,1}h_1 \cdot \Omega_{i,D_h}h_{D_h}\right] \\
&\quad + \mathbb{E}\left[\Omega_{i,2}h_2 \cdot \Omega_{i,1}h_1\right] + \mathbb{E}\left[(\Omega_{i,2}h_2)^2\right] + \cdots + \mathbb{E}\left[(\Omega_{i,D_h}h_{D_h})^2\right]
\end{aligned}$$

Now we apply our finding that only the squared terms remain while the rest go to 0.

$$\mathbb{E}[(\Omega_{i,1}h_1)^2] + \mathbb{E}[(\Omega_{i,2}h_2)^2] + \cdots + \mathbb{E}[(\Omega_{i,D_h}h_{D_h})^2]$$

We can distribute the square of each term to the individual terms as $\mathbb{E}[(\Omega_{i,j}h_j)^2] = \mathbb{E}[\Omega_{i,j}^2 h_j^2]$.

$$\mathbb{E}\left[\Omega_{i,1}^2 h_1^2\right] + \mathbb{E}\left[\Omega_{i,2}^2 h_2^2\right] + \cdots + \mathbb{E}\left[\Omega_{i,D_h}^2 h_{D_h}^2\right]$$

We can again use independence by remembering that $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ if X and Y are independent and define $X = \Omega_{i,j}^2$ and $Y = h_j^2$.

$$\mathbb{E}\left[\Omega_{i,1}^2\right]\mathbb{E}\left[h_1^2\right] + \mathbb{E}\left[\Omega_{i,2}^2\right]\mathbb{E}\left[h_2^2\right] + \cdots + \mathbb{E}\left[\Omega_{i,D_h}^2\right]\mathbb{E}\left[h_{D_h}^2\right]$$

Now let's recombine the terms back into a sum

$$\sum_{j=1}^{D_h} \mathbb{E}[\Omega_{i,j}^2]\mathbb{E}[h_j^2]$$

So we have that:

$$\mathbb{E}\left[\left(\sum_{j=1}^{D_h} \Omega_{i,j}h_i\right)^2\right] = \sum_{j=1}^{D_h} \mathbb{E}[\Omega_{i,j}^2]\mathbb{E}[h_j^2]$$

Now let's use this fact:

2.4 Derivation of the He initialization variance

$$\begin{aligned}
\sigma_{f'}^2 &= \mathbb{E}[f'^2] - \mathbb{E}[f']^2 && \text{(make use of Equation 6 in Section 2.2)} \quad (1)\\
&= \mathbb{E}[f'^2] - 0^2 && (\mathbb{E}[f_i'] = 0 \text{ from Equation 7 in Section 2.1)} \quad (2)\\
&= \mathbb{E}\left[\left(\beta_i + \sum_{j=1}^{D_h} \Omega_{i,j}h_i\right)^2\right] && \text{(by definition of } f'\text{)} \quad (3)\\
&= \mathbb{E}\left[\left(\sum_{j=1}^{D_h} \Omega_{i,j}h_i\right)^2\right] && (\beta_i = 0 \text{ by construction)} \quad (4)\\
&= \sum_{j=1}^{D_h} \mathbb{E}[\Omega_{i,j}^2]\mathbb{E}[h_j^2] && \text{(from Section 2.3)} \quad (5)\\
& && (6)
\end{aligned}$$

Now let's make use of the variance identity again. We derived that $\sigma_X^2 = \mathbb{E}[X^2] - \mathbb{E}[X]^2$. Let's set $X = \Omega_{i,j}$. Then we have $\sigma_\Omega^2 = \mathbb{E}[\Omega_{i,j}^2] - \mathbb{E}[\Omega_{i,j}]^2$. But we previously found that $\mathbb{E}[\Omega_{i,j}] = 0$ since we're assuming it came from a distribution with mean 0. This simplifies the equation to $\sigma_\Omega^2 = \mathbb{E}[\Omega_{i,j}^2]$. Let's use this equivalence.

$$= \sum_{j=1}^{D_h} \sigma_\Omega^2 \mathbb{E}[h_j^2] \tag{1}$$

The variance no longer depends on j so we can pull it outside the sum.

$$= \sigma_\Omega^2 \sum_{j=1}^{D_h} \mathbb{E}[h_j^2] \tag{1}$$

Now we must think about $\mathbb{E}[h_j^2]$. By definition we have:

$$\mathbb{E}[h_j^2] = \mathbb{E}[\max(0, f_j]$$

where recall that f_j is the pre-activation for the jth neuron and h_j is activation function.

$$\begin{aligned} \mathbb{E}[\max(0, f_j)] &= \mathbb{E}\left[\frac{1}{2}(f_j \mid f_j \geq 0) + \frac{1}{2}(f_j \mid f_j < 0)\right] \\ \mathbb{E}[\max(0, f_j)] &= \frac{1}{2}\mathbb{E}[f_j \mid f_j \geq 0] + \frac{1}{2}\mathbb{E}[f_j \mid f_j < 0] \end{aligned}$$

And by construction of ReLU, $f_j = 0$ when it's negative:

$$\begin{aligned} \mathbb{E}[h_j^2] = \mathbb{E}[\max(0, f_j)] &= \frac{1}{2}\mathbb{E}[f_j \mid f_j \geq 0] + \cancelto{0}{\frac{1}{2}\mathbb{E}[f_j \mid f_j < 0]} \\ \mathbb{E}[h_j^2] &= \frac{1}{2}\mathbb{E}[f_j^2] \end{aligned}$$