Parameter Initialization

Dr. Alireza Aghamohammadi

May 5, 2024

Problem Statement

- How should we initialize parameters (weights and biases) in deep neural networks?

$$f_i = \beta_i + \Omega_i \cdot \texttt{ReLU}[f_{i-1}]$$

- Suppose biases (β_i) are initialized to zero and weights (Ω_i) follow a normal distribution with mean zero and variance σ^2.
- If the variance σ^2 is very small (e.g., 10^{-5}),
  - The output range of $\texttt{ReLU}[f_{i-1}]$ is less than f_{i-1}.
  - As a result, the output range of f_i is reduced.
  - This reduction also affects the backward pass during parameter updates.
  - Each gradient update involves multiplication by Ω_i^T.
  - This leads to the vanishing gradient problem.
- If the variance σ^2 is very large (e.g., 10^5),
  - The output range of f_i increases.
  - This leads to the exploding gradient problem.

Mathematical Analysis (Part 1)

- Let's consider f' and f with dimensions $D_{h'}$ and D_h respectively. We have:

$$h = \texttt{ReLU}[f]$$
$$f' = \beta_i + \Omega_i \cdot h$$

- Now, suppose we initialize all the biases (β_i) to zero and the elements of the weight (Ω_{ij}) to a normal distribution with a mean of zero and a variance of σ_Ω^2.
- The expected value of f' can be calculated as follows:

$$\begin{aligned}
\mathbb{E}[f'_i] &= \mathbb{E}[\beta_i + \sum_{j=1}^{D_h} \Omega_{ij} \cdot h_j] \\
&= \mathbb{E}[\beta_i] + \sum_{j=1}^{D_h} \mathbb{E}[\Omega_{ij} \cdot h_j] \\
&= \mathbb{E}[\beta_i] + \sum_{j=1}^{D_h} \mathbb{E}[\Omega_{ij}]\mathbb{E}[h_j] \\
&= 0 + \sum_{j=1}^{D_h} 0 \cdot \mathbb{E}[h_j] = 0
\end{aligned}$$

Mathematical Analysis (Part 2)

- Using the previous results, we can calculate the variance $\sigma_{f'}^2$ of f' as follows:

$$
\begin{aligned}
\sigma_{f'}^2 &= \mathbb{E}[f_i'^2] - \mathbb{E}[f_i']^2 \\
&= \mathbb{E}[(\beta_i + \sum_{j=1}^{D_h} \Omega_{ij} \cdot h_j)^2] - 0 \\
&= \mathbb{E}[(\sum_{j=1}^{D_h} \Omega_{ij} \cdot h_j)^2] \\
&= \sum_{j=1}^{D_h} \mathbb{E}[\Omega_{ij}^2]\mathbb{E}[h_j^2] \\
&= \sigma_\Omega^2 \sum_{j=1}^{D_h} \mathbb{E}[h_j^2] \\
&= \sigma_\Omega^2 \sum_{j=1}^{D_h} \frac{\sigma_f^2}{2} = \frac{1}{2} D_h \sigma_\Omega^2 \sigma_f^2
\end{aligned}
$$

He Initialization

- We would like the variances $\sigma_{f'}^2$ and σ_f^2 to be the same. This can be expressed as:

$$\sigma_{f'}^2 = \frac{1}{2} D_h \sigma_\Omega^2 \sigma_f^2$$

- Therefore, the optimal variance for the weights (σ_Ω^2) should be:

$$\sigma_\Omega^2 = \frac{2}{D_h}$$

Effect of He Initialization

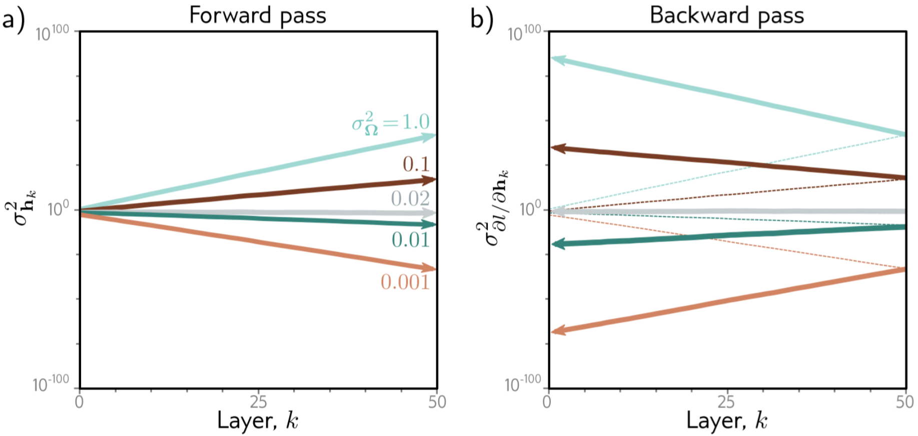

Figure: The effect of He Initialization.[1]

[1] Adopted from the book, Understanding Deep Learning