CS230 Handson session 5: "Xavier initialization"

In this discussion section, you will study the theory behind Xavier Initialization. We will use menti during this section. Please connect on www.menti.com and enter the access code provided by the TA.

1. Why is parameter initialization a critical aspect of the training of a deep neural network?

   It can help the training process by avoiding the saturation of the gradient signal (exploding gradient problem) or the excessive shrinkage of the gradient signal (vanishing gradient problem) in very deep neural networks.

2. How can you stabilize the training process in deep neural networks?

A: 1: Using ReLU/Leaky-ReLU activation functions (instead of sigmoid or tanh)
2: Initialize the weights with some heuristics (Xavier, He or Glorot Initialization)
3: Gradient clipping: We set a threshold value, and if a chosen function of a gradient is larger than this threshold, we set it to another value

3. Let's go over the details of Xavier initialization heuristic. What is the main goal and assumptions made in Xavier initialization? Could you derive it?

Goal: Initialize weights in such a way that you break the **symmetry** problem (thus, we need random generation) and you have a **stable backprop signal** in deep NN. Assuming we will sample the initial weight values from a centered Gaussian, what should be its variance to satisfy the requirements?

Assumptions:
In the forward pass, the output and input of a layer should maintain the same variance (that is: $Var(a^{L-1}) = Var(a^{L})$)
1: Inputs and Weights are random variables centered in 0 (mean 0)
2: Inputs and Weights are identically distributed and independent (iid)
3: Biases are deterministic (always initialized as 0s)
4: tanh() activation function is linearized $\rightarrow$ $Var(a^{[L]}) \approx Var(z^{[L]})$
5: For step 5, see:
https://stats.stackexchange.com/questions/31177/does-the-variance-of-a-sum-equal-the-sum-of-the-variances

$$Var(a^{[L]}) \approx Var(z^{[L]}) = Var\left(z_j^{[L]}\right) = Var\left(\sum_1^{n_{L-1}} W_{jk}^L a_k^{L-1}\right)$$

$$Var\left(\sum_1^{n_{L-1}} W_{jk}^L a_k^{L-1}\right) = \sum_1^{n_{L-1}} Var[W_{jk}^L a_k^{L-1}] = \sum_1^{n_{L-1}} Var[W_{jk}^L] Var[a_k^{L-1}]$$

$$\sum_{1}^{n_{L-1}} Var[W_{jk}^{L}] Var\left[a_{k}^{L-1}\right] = \sum_{1}^{n_{L-1}} Var\left[W^{L}\right] Var\left[a^{L-1}\right] = n_{L-1}\, Var\left[W^{L}\right] Var\left[a^{L-1}\right]$$

Thus $Var(a^{L-1}) = Var(a^{L})$, implies $Var[W^{L}] = \frac{1}{n_{L-1}}$