Vanishing and Exploding Gradients — Summary

Vanishing and exploding gradients are two major problems that occur during training of deep neural networks, especially when using gradient-based optimization.

---

1. What Are Gradients?

During backpropagation, each layer receives a gradient (error signal) that tells it how much to change its weights.

This gradient passes backward through the network:

Loss → dZ[L] → dZ[L-1] → ... → dZ[1]

These gradients are formed by repeatedly multiplying:

- weight matrices
- activation derivatives

across many layers.

---

2. Vanishing Gradient Problem

Definition

Vanishing gradients occur when gradients become **extremely small** as they are backpropagated through deep networks.

This causes **early layers to stop learning**.

Why It Happens

1. **Activation functions like sigmoid/tanh have very small derivatives**
    - Sigmoid derivative ≤ 0.25
    - Repeated multiplication drives gradient toward 0.
2. **Weights initialized too small**
    - Numbers < 1 multiplied many times → shrink to zero.

Effects

- Slow or no learning
- Early layers do not update
- Deep networks fail to train properly
- Loss decreases extremely slowly

---

3. Exploding Gradient Problem

Definition

Exploding gradients occur when gradients become **very large** during backprop, causing unstable updates.

Why It Happens

1. **Weights initialized too large**
2. **Repeating multiplication of values > 1 through many layers**

Effects

- Loss becomes NaN or diverges
- Weights grow uncontrollably

- Model becomes unstable
- Training crashes or oscillates

---

4. How Deep Learning Solves These Problems

To reduce vanishing gradients:

- Use **ReLU** or variants (large, stable derivative)
- Use **He initialization** for ReLU networks
- Use **Batch Normalization** to stabilize activations
- Use **Residual connections (ResNet)** to allow gradients to flow
- Avoid deep stacks of sigmoid/tanh

To reduce exploding gradients:

- Use **gradient clipping**
- Use **Xavier/He initialization**
- Apply **BatchNorm**
- Use stable optimizers (Adam, RMSProp)

---

5. Short Definitions to Memorize

✔

Vanishing Gradient

Gradients become extremely small as they propagate backward, causing early layers to learn very slowly or not at all.

✔

Exploding Gradient

Gradients grow excessively large during backpropagation, causing unstable updates and divergence.