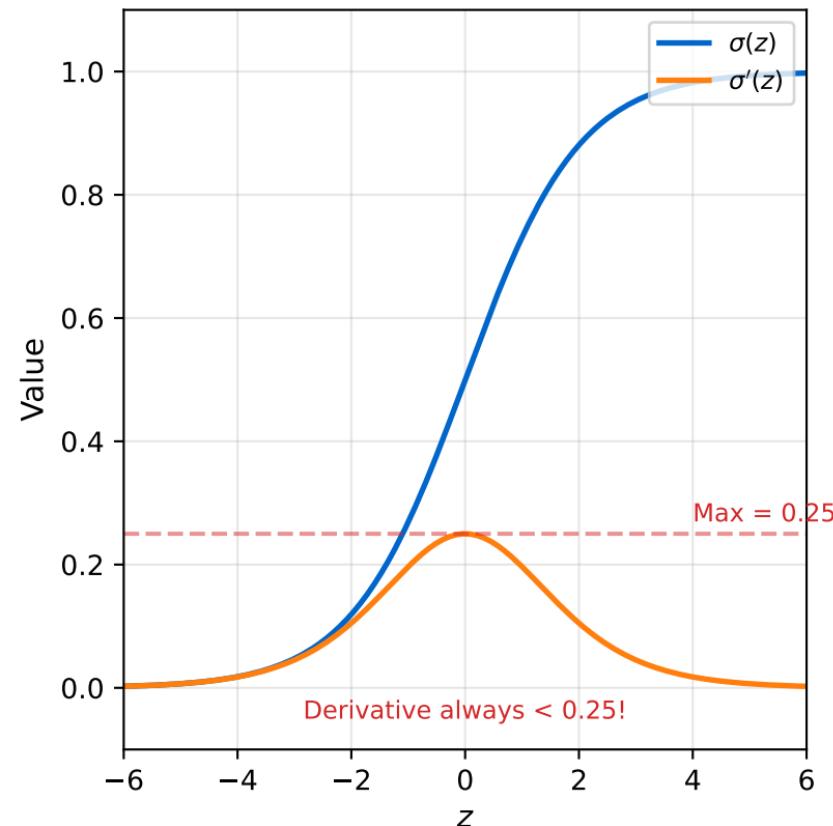
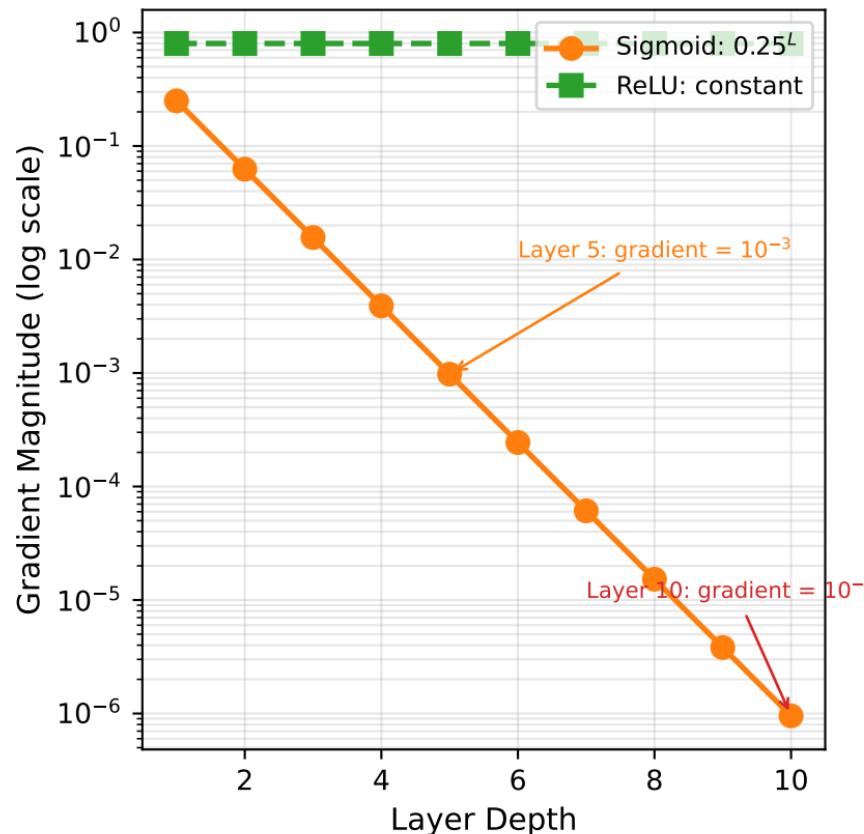


Vanishing Gradients: Why Deep Networks Were Hard to Train

Sigmoid & Its Derivative



Gradient Decay Through Layers



The Vanishing Gradient Problem

Problem: Gradients shrink exponentially with depth

Cause: Sigmoid/tanh derivatives are < 1

Effect: Early layers don't learn (gradients ~ 0)

1. Use ReLU activation (gradient = 1 for $z > 0$)

Solutions:
2. Careful weight initialization (He, Xavier)

3. Batch normalization

4. Residual connections (skip connections)

5. LSTM/GRU for recurrent networks

Key insight: ReLU solved this for feedforward networks!