AdaGrad (Adaptive Gradient Algorithm)

Parameter-Specific Learning Rate Adaptation

🎯 Core Concept

Adaptively adjusts learning rate
for each parameter based on historical gradients

Update Rule

$$\theta = \theta - \eta/\sqrt{(G + \varepsilon)} \odot \nabla L(\theta)$$

G: Cumulative sum of squared gradients
⊙: Element-wise multiplication

Adaptation Mechanism

Infrequent Parameters
Larger updates

Frequent Parameters
Smaller updates

✨ Benefits

- ✓ No manual learning rate tuning needed
- ✓ Advantageous for sparse data

⚠ Main Drawbacks

- ! Learning rate monotonically decreases over time
- ! Aggressive decay can cause premature training stop

ε Epsilon Parameter

Prevents division by zero (typically 1e-8)

💡 Best Use Cases

- Sparse Data Processing
- Natural Language Processing