📘 AdaGrad — Adaptive Gradient Descent

🚀 Core Idea

AdaGrad is an improvement over standard gradient descent. Instead of using a **fixed learning rate (η\eta)**, it uses an **adaptive learning rate** that adjusts for each parameter individually based on the past squared gradients. This helps the algorithm converge faster and more effectively, especially when different parameters have different sensitivities.

🧠 Intuition

- In standard gradient descent:

$$w_t = w_{t-1} - \eta \cdot \frac{\partial L}{\partial w}$$

- In **AdaGrad** (adaptive learning rate):

$$w_t = w_{t-1} - \eta'_t \cdot \frac{\partial L}{\partial w}$$

  Where:

  - η'_t = adaptive learning rate at iteration t

📐 Adaptive Learning Rate Formula

$$\eta'_t = \frac{\eta}{\sqrt{\alpha_t + \epsilon}}$$

Where:

- η = initial learning rate
- α_t = accumulated sum of squared gradients
- ϵ = small constant (e.g., 10^{-8}) to avoid division by zero

---

🧮 Formula for α_t

$$\alpha_t = \sum_{i=1}^{t} \left(\frac{\partial L}{\partial w_i} \right)^2$$

This means:

- α_t keeps **increasing** with every iteration as it accumulates the squared gradients.
- As α_t increases, the adaptive learning rate η_t' **decreases**.

🔁 Behavior Over Time

- In the **early stages**, gradients are small → α_t is small → higher learning rate
- As training progresses, α_t increases → η_t' decreases
- This leads to **smaller and more stable updates**, helping to fine-tune the weights

---

⚠️ Limitation of AdaGrad

- In **deep neural networks**, as iterations increase, α_t may become **very large**
- This causes η_t' to become **extremely small**
- When learning rate becomes too small:

$$w_t \approx w_{t-1}$$

→ **No effective weight updates**, and the model stops learning

---

✅ Summary Table

Concept	Description
Learning Rate	Adaptive (decreases over time)
Learning Rate Formula	$\eta_t' = \frac{\eta}{\sqrt{\alpha_t + \epsilon}}$
α_t Formula	$\alpha_t = \sum_{i=1}^{t} \left(\frac{\partial L}{\partial w_i} \right)^2$
Purpose of ϵ	Avoids division by zero
Advantage	Fast convergence, especially on sparse features
Limitation	Learning rate may become too small over time