🧠 Stochastic Gradient Descent (SGD)

Definition:
Stochastic Gradient Descent is a variant of gradient descent where **weights are updated using only one randomly selected training sample at a time**, rather than the entire dataset.

Update rule:

$$w = w - \eta \cdot \nabla L(x_i, y_i)$$

- (x_i, y_i) = one random sample
- η = learning rate
- ∇L = gradient of the loss function

✅ Pros of SGD

1. **Faster Updates per Step**
   - Only one sample is used → much faster computation per iteration.
2. **Works Well with Large Datasets**
   - You don't need to load the entire dataset into memory.
3. **Can Escape Local Minima**
   - The randomness in updates helps jump out of local minima and saddle points.
4. **Suitable for Online Learning**
   - Can learn from streaming data — ideal for real-time updates.

❌ Cons of SGD

1. **High Variance in Updates**
   - Weight updates are noisy, which can make the loss function zigzag and unstable.
2. **Harder to Converge**

- May not settle near the exact minimum due to fluctuations.

3. **Sensitive to Learning Rate**
    - A bad learning rate can cause divergence or very slow progress.
4. **May Require More Epochs**
    - Because of noisy updates, it might take longer (more epochs) to reach a good solution.

Summary Table

Aspect	**Explanation**
Speed	Faster per update (one sample at a time)
Noise	Updates are noisy and fluctuate
Memory	Very memory-efficient
Convergence	Less stable, may oscillate near minima
Best Use	Large-scale or online learning scenarios