Batch vs Mini-Batch vs Stochastic Gradient Descent

Balancing Speed, Stability, and Generalization

Batch GD

Entire Dataset

- ✓ Accurate gradient
- ✓ Stable convergence
- ✗ Slow speed
- ✗ Sharp minima

Mini-Batch GD

Subset of Data

- ✓ Balances speed and stability
- ✓ GPU parallel processing
- ✓ Better generalization
- ✓ Practical choice

Stochastic GD

Single Sample

- ✓ Fast updates
- ✓ Flat minima
- ✗ Noisy gradient
- ✗ Unstable convergence

📊 Typical Mini-Batch Sizes

32 | 64 | 128 | 256

Larger Batches

Gradients: → More stable
Hardware: → More efficient
Minima: → Sharp

Smaller Batches

Gradients: → Noisier
Generalization: → Better
Minima: → Flat

📐 Gradient Calculation Examples

Understanding How Each Method Computes Gradients

Example Setup

Dataset: $N = 1000$ samples | **Loss Function:** $L(\theta) = \frac{1}{2}(y - \hat{y})^2$ | **Learning Rate:** $\alpha = 0.01$

1 Batch Gradient Descent

Gradient:

$$\nabla L(\theta) = (1/N)\ \Sigma_{i=1}^{N}\ \nabla L(\theta, x_i, y_i)$$

Step 1: Calculate loss for all 1000 samples

Step 2: Compute average gradient: $\nabla L = (1/1000)\ \Sigma_{i=1}^{1000}\ \nabla L_i$

Step 3: Update parameters: $\theta \leftarrow \theta - 0.01 \times \nabla L$

💡 **Key Feature:** 1 epoch = 1 parameter update (using all 1000 samples)

2 Mini-Batch Gradient Descent

Gradient (batch size B=100):

$$\nabla L(\theta) = (1/B)\ \Sigma_{i=1}^{B}\ \nabla L(\theta, x_i, y_i)$$

Batch 1: Use samples 1-100 → Compute ∇L_1 → $\theta \leftarrow \theta - 0.01 \times \nabla L_1$

Batch 2: Use samples 101-200 → Compute ∇L_2 → $\theta \leftarrow \theta - 0.01 \times \nabla L_2$

⋮ *Continue...*

Batch 10: Use samples 901-1000 → Compute ∇L_{10} → $\theta \leftarrow \theta - 0.01 \times \nabla L_{10}$

💡 **Key Feature:** 1 epoch = 10 parameter updates (1000/100 = 10 batches)

3 Stochastic Gradient Descent

Gradient (single sample):
$\nabla L(\theta) = \nabla L(\theta, x_i, y_i)$

Sample 1: Use $x_1, y_1 \rightarrow$ Compute $\nabla L_1 \rightarrow \theta \leftarrow \theta - 0.01 \times \nabla L_1$

Sample 2: Use $x_2, y_2 \rightarrow$ Compute $\nabla L_2 \rightarrow \theta \leftarrow \theta - 0.01 \times \nabla L_2$

⋮ *Continue...*

Sample 1000: Use $x_{1000}, y_{1000} \rightarrow$ Compute $\nabla L_{1000} \rightarrow \theta \leftarrow \theta - 0.01 \times \nabla L_{1000}$

💡 **Key Feature:** 1 epoch = 1000 parameter updates (update after each sample)

⚖️ Method Comparison Summary

Method	Sample Size	Updates per Epoch	Computation Time	Gradient Accuracy
Batch GD	N = 1000	1	Slowest	Very accurate
Mini-Batch GD	**B = 100**	**10**	**Balanced**	**Sufficiently accurate**
Stochastic GD	B = 1	1000	Fast updates	Noisy