Mini-Batch Gradient Descent (MBGD) – Simplified Theory

1. Introduction

- Gradient Descent (GD) has three main types:
  I. **Batch GD:** Uses the entire dataset for each parameter update.
  II. **Stochastic GD:** Uses only one sample for each update.
  III. **Mini-Batch GD: Compromise between the two**, using **small random batches of data** in each step.

2. What is Mini-Batch Gradient Descent?

- In MBGD, the dataset is **split into smaller batches (b)** of fixed size (e.g., 16, 32, 64).
- The gradient is computed using only **one batch at a time**, not the full dataset or single sample.

Parameter update rule:

$$\theta_j = \theta_j - \alpha \frac{1}{b} \sum_{i=1}^{b} (y_i - \hat{y}_i) x_{ij}$$

Where:

- b = batch size (subset of data).
- α = learning rate.

3. Why Use Mini-Batch GD?

- **Combines benefits of BGD and SGD:**
  I. **Faster** than Batch GD (more frequent updates).
  II. **Less noisy** than SGD (averages over multiple samples).
- Makes use of **vectorized operations** → runs efficiently on GPUs.
- Helps models converge **more smoothly** while still allowing some noise to escape local minima.

4. Algorithm Steps

1. **Shuffle the dataset** to ensure randomness.
2. **Divide data into mini-batches** of fixed size bb.
3. For each mini-batch:

I. Compute gradient using that batch only.
II. Update parameters with the computed gradient.

4. Repeat for all mini-batches → this completes **one epoch**.
5. Continue multiple epochs until convergence.

---

5. Advantages of MBGD

- **Balanced speed:** Faster than BGD, more stable than SGD.
- **Efficient computation:** Uses matrix operations, ideal for GPUs.
- **Noise helps** avoid local minima but is **less chaotic** than SGD.
- Works well with **very large datasets**.

---

6. Disadvantages of MBGD

- Choosing **batch size** can affect performance:
    1. **Small batch:** More noise (closer to SGD).
    2. **Large batch:** Slower, closer to Batch GD.
- Requires tuning of **learning rate** and **batch size** for optimal results.

---

7. Practical Tips

- Common batch sizes: **16, 32, 64, 128** (powers of 2 for computational efficiency).
- Combine MBGD with:
    1. **Learning rate scheduling.**
    2. **Momentum or Adam optimizer** for better convergence.

➢ Key Takeaways

Method	Gradient Calculation	Update Frequency	Noise Level	Speed
Batch GD	All samples	1 per epoch	Low	Slow
SGD	1 sample	n per epoch	High	Fast
Mini-Batch GD	**Subset of samples (b)**	**n/b per epoch**	**Medium**	**Balanced**
