🧠 Mini-Batch Stochastic Gradient Descent (Mini-Batch SGD)

Definition:
Mini-Batch SGD is a middle ground between **Batch Gradient Descent** and **Stochastic Gradient Descent (SGD)**. Instead of using the whole dataset (like Batch GD) or just one data point (like SGD), it updates the model using **a small subset of data (a mini-batch)** — typically 32, 64, or 128 samples.

Update Rule:

$$w = w - \eta \cdot \nabla L(\text{mini-batch})$$

- Each mini-batch is randomly selected from the dataset.
- The loss and gradients are averaged over the batch.

✅ Pros of Mini-Batch SGD

1. **Balances Speed and Stability**
   - Faster than full batch GD, more stable than SGD.
2. **Hardware Efficiency**
   - Leverages matrix operations and parallel processing on GPUs efficiently.
3. **Less Noisy than SGD**
   - Reduces the variance in gradient estimates, making convergence smoother.
4. **Scalable to Large Datasets**
   - Can train on large datasets without needing to load the entire dataset into memory.
5. **Better Generalization**
   - The slight randomness in batch selection acts as regularization, helping prevent overfitting.

❌ Cons of Mini-Batch SGD

1. **Still Sensitive to Learning Rate**
   - Requires tuning to ensure stable and efficient convergence.
2. **May Oscillate Near Minima**
   - Doesn't always settle perfectly at the minimum due to small gradient noise.
3. **Choosing Batch Size Can Be Tricky**
   - Too small → unstable; too large → slow or memory-heavy.
4. **Requires Shuffling**
   - To avoid bias, the dataset must be shuffled before forming mini-batches each epoch.

Summary Table

Aspect	Explanation
Speed	Faster than batch GD, slower than SGD
Noise	Moderate (less than SGD, more than batch GD)
Efficiency	Optimized for GPUs and vectorized operations
Convergence	Good balance between stability and performance
Best Use	Deep learning models with large datasets