summarizing the **theory, visuals, and formulas** behind **Mini-Batch Stochastic Gradient Descent with Momentum**, and how it reduces noise during optimization.

📘 Background: Why We Need Momentum in Mini-Batch SGD

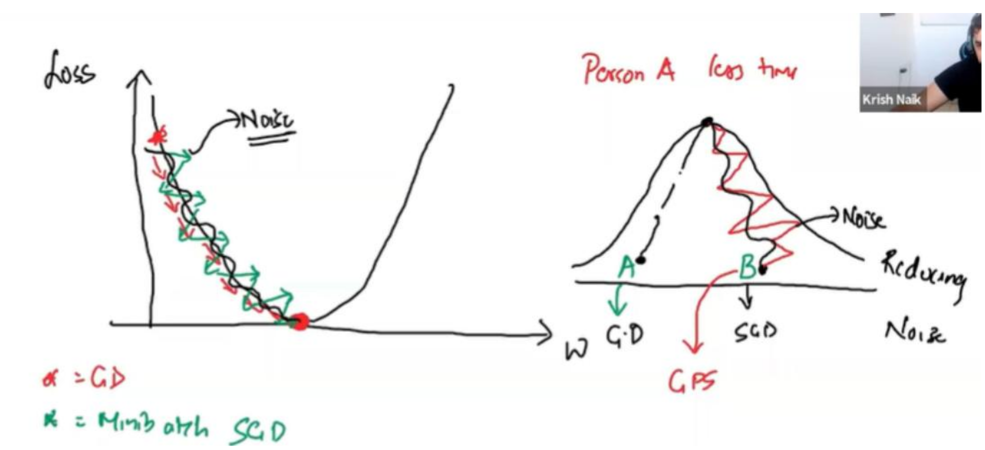
- **Mini-Batch SGD** introduces **noise** due to randomness in small batches.
- This leads to **fluctuating updates**, causing instability and slow convergence.

📈 Left Plot (Image 1):

- **Red path (α)**: Gradient Descent (GD) — stable but slower.
- **Green path (x)**: Mini-Batch SGD — faster but noisy.

🏔️ Right Plot (Image 1):

- Two persons climbing a hill:
    - **Person A (GD)** — slow, stable path.
    - **Person B (SGD)** — faster, but noisy.
    - If we give GPS (guidance = **momentum**), Person B will reach faster and smoother.



🔁 SGD with Momentum

⚙️ Formula:

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

But instead of using raw gradients, we add **momentum** to smooth the path.

📐 Concept of Exponential Weighted Average

🔢 Key Formula:

$$V_t = \beta V_{t-1} + (1 - \beta) a_t$$

- a_t = current value (e.g., gradient)
- $\beta \in [0, 1)$ = smoothing factor (common: 0.9 or 0.95)
- It gives **more weight to recent history**, smoothing noisy updates.

Example:

$$V_2 = \beta a_1 + (1 - \beta) a_2$$

$$V_3 = \beta V_2 + (1 - \beta) a_3$$

🔄 Exponential Weighted Average in SGD

For weights and bias:

$$w_t = w_{t-1} - \eta \cdot V_{dw} \quad , \quad b_t = b_{t-1} - \eta \cdot V_{db}$$

Where:

$$V_{dw_t} = \beta V_{dw_{t-1}} + (1 - \beta) \cdot \frac{\partial L}{\partial w_{t-1}}$$

$$V_{db_t} = \beta V_{db_{t-1}} + (1 - \beta) \cdot \frac{\partial L}{\partial b_{t-1}}$$

⚙️ Implementation Steps

🧮 Initialization:

- Set $V_{dw} = 0, V_{db} = 0$

🔁 On every iteration (inside epoch):

1. Compute gradients: dw, db
2. Update:

$$V_{dw} = \beta V_{dw} + (1 - \beta) \cdot dw$$

$$V_{db} = \beta V_{db} + (1 - \beta) \cdot db$$

3. Apply to weights:

$$w = w - \eta \cdot V_{dw} \quad , \quad b = b - \eta \cdot V_{db}$$

📌 Summary

Concept	Description
Problem	Mini-batch SGD is fast but noisy
Solution	Use **momentum** to smooth updates
Key Technique	**Exponential Weighted Average**
Hyperparameter	β=0.9\beta = 0.9 or 0.950.95
Result	Faster and smoother convergence, better minima