Let's calculate and prove how Gradient Descent (GD) **with** and **without momentum** behaves mathematically using an example.

We will:

1. Define a simple quadratic cost function $J(\theta) = \frac{1}{2}\theta^2$.
2. Calculate parameter updates step-by-step for both methods.
3. Compare results to demonstrate the effect of momentum.

---

## Example Setup

1. **Cost Function:**

$$J(\theta) = \frac{1}{2}\theta^2$$

   - Gradient: $\nabla J(\theta) = \theta$.

2. **Initial Parameters:**
   - Initial value of $\theta_0 = 2$.
3. **Learning Rate:**
   - $\eta = 0.1$.
4. **Momentum Coefficient:**
   - $\beta = 0.9$ (for GD with momentum).
5. **Iterations:**
   - Perform updates for 5 steps to observe the difference.

## Gradient Descent (Without Momentum)

The update rule for basic GD is:

$$\theta_{t+1} = \theta_t - \eta \nabla J(\theta_t)$$

**Step-by-step Calculations:**

- **Step 0:** Start with $\theta_0 = 2$.
- **Step 1:**

$$\theta_1 = \theta_0 - \eta \cdot \nabla J(\theta_0)$$

$$\theta_1 = 2 - 0.1 \cdot 2 = 1.8$$

- **Step 2:**

$$\theta_2 = \theta_1 - \eta \cdot \nabla J(\theta_1)$$

$$\theta_2 = 1.8 - 0.1 \cdot 1.8 = 1.62$$

- **Step 3:**

$$\theta_3 = \theta_2 - \eta \cdot \nabla J(\theta_2)$$

$$\theta_3 = 1.62 - 0.1 \cdot 1.62 = 1.458$$

Similarly, we can compute:

- $\theta_4 = 1.3122$,
- $\theta_5 = 1.18098$.

## Gradient Descent (With Momentum)

The update rules with momentum are:

$$v_{t+1} = \beta v_t + (1 - \beta)\nabla J(\theta_t)$$

$$\theta_{t+1} = \theta_t - \eta v_{t+1}$$

**Step-by-step Calculations:**

- **Initialization:**
  - $\theta_0 = 2,$
  - $v_0 = 0.$
- **Step 1:**

$$v_1 = \beta v_0 + (1 - \beta)\nabla J(\theta_0)$$

$$v_1 = 0.9 \cdot 0 + 0.1 \cdot 2 = 0.2$$

$$\theta_1 = \theta_0 - \eta v_1$$

$$\theta_1 = 2 - 0.1 \cdot 0.2 = 1.98$$

- **Step 2:**

$$v_2 = \beta v_1 + (1 - \beta)\nabla J(\theta_1)$$

$$v_2 = 0.9 \cdot 0.2 + 0.1 \cdot 1.98 = 0.398$$

$$\theta_2 = \theta_1 - \eta v_2$$

$$\theta_2 = 1.98 - 0.1 \cdot 0.398 = 1.9402$$

- **Step 3:**

$$v_3 = \beta v_2 + (1 - \beta)\nabla J(\theta_2)$$

$$v_3 = 0.9 \cdot 0.398 + 0.1 \cdot 1.9402 = 0.55222$$

$$\theta_3 = \theta_2 - \eta v_3$$

$$\theta_3 = 1.9402 - 0.1 \cdot 0.55222 = 1.88498$$

Similarly, we can compute:

- $v_4 = 0.674998, \theta_4 = 1.81748,$
- $v_5 = 0.769824, \theta_5 = 1.74050.$

---

## Comparison of Updates

| Iteration $t$ | $\theta_t$ (GD) | $\theta_t$ (GD with Momentum) |
|---|---|---|
| 0 | 2.0000 | 2.0000 |
| 1 | 1.8000 | 1.9800 |
| 2 | 1.6200 | 1.9402 |
| 3 | 1.4580 | 1.8849 |
| 4 | 1.3122 | 1.8175 |

| Iteration $t$ | $\theta_t$ (GD) | $\theta_t$ (GD with Momentum) |
|---|---|---|
| 5 | 1.1810 | 1.7405 |

---

## Observations:

1. **GD Without Momentum:**
    - The updates decrease uniformly and slowly converge toward the minimum.
    - It moves in small, steady steps.
2. **GD With Momentum:**
    - The updates are faster because of the accumulated velocity.
    - The values "overshoot" slightly in the beginning but settle faster than basic GD.
    - It avoids oscillations and accelerates convergence along consistent gradient directions.

---

## Conclusion:

- **Momentum accelerates convergence** by incorporating past gradient information into the current update.
- Momentum can cause slight overshooting initially, but it helps the optimizer settle into the global minimum more quickly.