derivation.
explain only what i asked , now cover what happen in stochastic gradient descent, and mini batch gradient descent

Below is the **strictly mathematical and technical derivation** of **Multiple Linear Regression using Gradient Descent**, followed by **Stochastic Gradient Descent (SGD)** and **Mini-Batch Gradient Descent**, exactly as requested.

---

1. Problem Definition

Let

$$\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$$

where

$$\mathbf{x}_i \in \mathbb{R}^d, \quad y_i \in \mathbb{R}$$

Define

$$X \in \mathbb{R}^{n \times d}, \quad \mathbf{y} \in \mathbb{R}^n$$

Let

$$\boldsymbol{\theta} \in \mathbb{R}^d$$

Assumptions:

$$y_i = \mathbf{x}_i^\top \boldsymbol{\theta} + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i] = 0$$

Objective:

$$\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$$

where

$$J(\boldsymbol{\theta}) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \boldsymbol{\theta})^2$$

---

2. Mathematical Setup

Define:

$$\hat{y}_i = h_{\boldsymbol{\theta}}(\mathbf{x}_i)$$

$$h_{\boldsymbol{\theta}}(\mathbf{x}_i) = \mathbf{x}_i^\top \boldsymbol{\theta}$$

Vector form:

$$\hat{\mathbf{y}} = X\boldsymbol{\theta}$$

Residual vector:

$$\mathbf{r} = \mathbf{y} - X\boldsymbol{\theta}$$

Cost function:

$$J(\boldsymbol{\theta}) = \frac{1}{2n}\mathbf{r}^\top \mathbf{r}$$

---

3. Derivation of the Core Equations

3.1 Hypothesis Model

$$h_{\boldsymbol{\theta}}(X) = X\boldsymbol{\theta}$$

3.2 Cost Function Expansion

$$J(\boldsymbol{\theta}) = \frac{1}{2n}(\mathbf{y} - X\boldsymbol{\theta})^\top(\mathbf{y} - X\boldsymbol{\theta})$$

Expand:

$$J(\boldsymbol{\theta}) = \frac{1}{2n}\left[\mathbf{y}^\top\mathbf{y} - 2\mathbf{y}^\top X\boldsymbol{\theta} + \boldsymbol{\theta}^\top X^\top X\boldsymbol{\theta}\right]$$

3.3 Partial Derivatives

For parameter θ_j:

$$\frac{\partial J}{\partial \theta_j} = \frac{1}{2n}\frac{\partial}{\partial \theta_j}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^\top\boldsymbol{\theta})^2$$

$$= \frac{1}{n} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \boldsymbol{\theta}) \, (-x_{ij})$$

$$= -\frac{1}{n} \sum_{i=1}^{n} x_{ij} (y_i - \mathbf{x}_i^\top \boldsymbol{\theta})$$

3.4 Gradient Vector

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial J}{\partial \theta_1} \\ \vdots \\ \frac{\partial J}{\partial \theta_d} \end{bmatrix}$$

Vector form:

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = -\frac{1}{n} X^\top (\mathbf{y} - X\boldsymbol{\theta})$$

---

4. Solving for Optimal Parameters

4.1 Stationary Condition

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = 0$$

$$X^\top X \boldsymbol{\theta} = X^\top \mathbf{y}$$

(Closed-form exists but **not used** here.)

4.2 Gradient Descent Update Rule

Let learning rate $\alpha > 0$.

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \alpha \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}^{(k)})$$

Substitute gradient:

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \frac{\alpha}{n} X^\top (\mathbf{y} - X\boldsymbol{\theta}^{(k)})$$

---

5. Stochastic Gradient Descent (SGD)

At iteration k, sample index $i_k \in \{1, \dots, n\}$.

5.1 Stochastic Loss

$$J_{i_k}(\boldsymbol{\theta}) = \frac{1}{2}(y_{i_k} - \mathbf{x}_{i_k}^\top \boldsymbol{\theta})^2$$

5.2 Stochastic Gradient

$$\nabla_{\boldsymbol{\theta}} J_{i_k}(\boldsymbol{\theta}) = -\mathbf{x}_{i_k}(y_{i_k} - \mathbf{x}_{i_k}^\top \boldsymbol{\theta})$$

5.3 SGD Update Rule

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \alpha \mathbf{x}_{i_k}(y_{i_k} - \mathbf{x}_{i_k}^\top \boldsymbol{\theta}^{(k)})$$

---

6. Mini-Batch Gradient Descent

Let mini-batch $B_k \subset \{1, \dots, n\}$, $|B_k| = m$.

6.1 Mini-Batch Loss

$$J_{B_k}(\boldsymbol{\theta}) = \frac{1}{2m} \sum_{i \in B_k} (y_i - \mathbf{x}_i^\top \boldsymbol{\theta})^2$$

6.2 Mini-Batch Gradient

$$\nabla_{\boldsymbol{\theta}} J_{B_k}(\boldsymbol{\theta}) = -\frac{1}{m} \sum_{i \in B_k} \mathbf{x}_i (y_i - \mathbf{x}_i^\top \boldsymbol{\theta})$$

Matrix form:

$$\nabla_{\boldsymbol{\theta}} J_{B_k} = -\frac{1}{m} X_{B_k}^\top (\mathbf{y}_{B_k} - X_{B_k} \boldsymbol{\theta})$$

6.3 Mini-Batch Update Rule

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \frac{\alpha}{m} X_{B_k}^\top (\mathbf{y}_{B_k} - X_{B_k} \boldsymbol{\theta}^{(k)})$$

---

7. Final Formula Summary

Batch Gradient Descent

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \frac{\alpha}{n} X^\top (\mathbf{y} - X \boldsymbol{\theta}^{(k)})$$

Stochastic Gradient Descent

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \alpha \mathbf{x}_i (y_i - \mathbf{x}_i^\top \boldsymbol{\theta}^{(k)})$$

Mini-Batch Gradient Descent

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \frac{\alpha}{m} X_B^\top (\mathbf{y}_B - X_B \boldsymbol{\theta}^{(k)})$$

---

8. Computational Form

Implementable Update Rules

- Batch GD:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \frac{\alpha}{n} X^\top (\mathbf{y} - X\boldsymbol{\theta})$$

- SGD:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \mathbf{x}_i (y_i - \mathbf{x}_i^\top \boldsymbol{\theta})$$

- Mini-Batch GD:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \frac{\alpha}{m} X_B^\top (\mathbf{y}_B - X_B \boldsymbol{\theta})$$

If you want, next we can **formally prove convergence conditions**, **analyze Hessian properties**, or **derive momentum / Adam from this base**—strictly mathematically.