Gradient Descent（梯度下降）— 线性回归 Week03 重点

适用：考试手算 / 理解为什么用 GD 而不是 Normal Equation

> **一句话：**Gradient Descent 用“沿着 loss $J(\theta)$ 最快下降的方向”迭代更新参数 θ，把误差越压越小。

1. 线性回归模型 & 目标

- 预测：$\hat{y} = X\theta$（X 是 design matrix，θ 是参数向量）
- 常用 cost（MSE 版本）：

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(\hat{y}^{(i)} - y^{(i)})^2$$

  这里 m 是样本数；前面的 $\frac{1}{2}$ 方便求导（把 2 抵消掉）。

2. 核心更新规则（必须会写）

- 总公式：

$$\theta := \theta - \alpha \nabla_\theta J(\theta)$$

  α = learning rate（学习率/步长）；∇J = 梯度（对每个参数的偏导组成的向量）。
- **为什么是减号？** 梯度方向是“上升最快”，我们要让 J 下降，所以走反方向：$-\nabla J$。

3. 线性回归下的梯度（向量形式）

- 对 MSE：

$$\nabla_\theta J(\theta) = \frac{1}{m} X^T(X\theta - y)$$

- 所以更新就是：

$$\theta := \theta - \alpha \frac{1}{m} X^T(X\theta - y)$$

4. Batch / SGD / Mini-batch（区别一句话背下来）

- **Batch GD**：每次用全部 m 个样本算梯度（稳定，但一次更新成本高）。
- **Stochastic GD (SGD)**：每次用 1 个样本更新（快但抖动大）。
- **Mini-batch GD**：每次用一小批样本（最常用，速度与稳定折中）。

5. 学习率 α（很爱考）

- α 太大：可能 overshoot（跨过最低点）→ 不收敛 / 发散。
- α 太小：会收敛但非常慢。

6. 和 Normal Equation 的关系（Week03 教授重点）

- **Normal equation（闭式解）**：

$$\theta = (X^T X)^{-1} X^T y$$

  一步到位，但需要矩阵可逆/求逆可能贵。
- **Gradient Descent（迭代解）**：不需要显式求逆，通过迭代逼近最优 θ。

7. 小手算模板（你可以按这个步骤写题）

1. 写出 $\hat{y} = X\theta$，再写 $J(\theta)$（通常 MSE）。
2. 写梯度：$\nabla J(\theta) = \frac{1}{m} X^T (X\theta - y)$。
3. 代入更新：$\theta \leftarrow \theta - \alpha \nabla J(\theta)$。
4. 重复 1~2 次（题目一般只让你更新几步）。

如果你把题目给的具体 $X, y, \alpha, \theta^{(0)}$ 发我，我可以帮你检查每一步矩阵维度与计算是否对（不直接代你交作业）。