Think-Help：没错，这是 AI 给出的例子，希望你也能善用 AI，加油!

Mini-Batch 反向传播（多维输入 + 多个样本）

假设我们训练一个**线性层**（或者说全连接层）：

$$\hat{Y} = XW + b$$

其中：

- $X \in \mathbb{R}^{6\times 3}$：6 个样本、每个样本 3 个特征
- $W \in \mathbb{R}^{3\times 2}$：权重矩阵，输入维 3，输出维 2
- $b \in \mathbb{R}^{1\times 2}$：偏置向量
- $\hat{Y}, Y \in \mathbb{R}^{6\times 2}$：预测值与真实标签
- 损失：均方误差（MSE）平均形式

这里的b仍是1*2，没有问题。
b在计算时会自动广播（扩展为6*2）
这里也可以看出不受样本数量N的影响

$$L = \frac{1}{6}\sum_{i=1}^{6} \| \hat{Y}_i - Y_i \|^2$$

Step 1：前向传播

1 矩阵乘法：

$$Z = XW + b$$

这里的 b会**广播**到每个样本（6 行）。

2 输出预测：

$$\hat{Y} = Z$$

（假设输出层无激活函数——典型回归场景）

Step 2：计算损失

$$L = \frac{1}{6}\sum_{i=1}^{6}\sum_{j=1}^{2}(\hat{Y}_{ij} - Y_{ij})^2$$

Step 3：反向传播求梯度

对每个参数求导，按矩阵规则写出：

(1) 对输出求导

$$\frac{\partial L}{\partial \hat{Y}} = \frac{2}{6}(\hat{Y} - Y) = \frac{1}{3}(\hat{Y} - Y)$$

形状：6×2

(2) 对权重求导

$$\frac{\partial L}{\partial W} = X^T \frac{\partial L}{\partial \hat{Y}}$$

维度:

$(3 \times 6) \times (6 \times 2) = (3 \times 2)$，
完全与 W一致。

矩阵乘法自动帮你把所有样本的梯度累加/求和，
结果仍是固定形状的矩阵。

(3) 对偏置求导

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{6} \frac{\partial L}{\partial \hat{Y}_i}$$

即在样本维上求和，维度（1×2），与 b一致。

W矩阵自动计算——固定形状（输入维度*输出维度）
b手动求和计算——固定形状（1*输出维度）

⚙️ Step 4：参数更新

$$W \leftarrow W - \eta \frac{\partial L}{\partial W}$$

$$b \leftarrow b - \eta \frac{\partial L}{\partial b}$$

其中学习率 η是标量，比如 0.01。

🧠 Step 5：思考与直觉

项目	含义	维度变化
X	输入数据（6 样本 × 3 特征）	6×3
W	权重矩阵	3×2
XW	线性映射	6×2
b	偏置（广播）	$1 \times 2 \to 6 \times 2$
$\hat{Y}$	输出预测	6×2
$\frac{\partial L}{\partial \hat{Y}}$	误差梯度	6×2
$\frac{\partial L}{\partial W}$	累积梯度	3×2
$\frac{\partial L}{\partial b}$	样本求和	1×2

✅ **关键理解：**

无论样本多少、输入多大，最终梯度的形状总与参数一致。
因为在反向传播中，对样本维（batch 维）进行了**求平均**或**求和**，
所以不会让 W, b变“更大”。