1. 场景

考虑一个多层感知机（MLP）的中间层：

$$z = Wx + b, \quad y = \sigma(z)$$

- $x \in \mathbb{R}^n$：前一层输出（本层输入）
- $W \in \mathbb{R}^{m \times n}$：权重矩阵
- $b \in \mathbb{R}^m$：偏置
- $z \in \mathbb{R}^m$：线性变换结果
- $y \in \mathbb{R}^m$：激活函数输出
- 激活函数 σ 可以是 sigmoid、ReLU 等，逐元素作用。

目标：计算 $\frac{\partial L}{\partial W}$ 以进行梯度下降，其中 L 是损失函数（如交叉熵）。

---

2. 损失函数示例

最后一层输出经过 softmax 得到预测概率 $\hat{y}$：

$$\hat{y}_i = \frac{e^{z_i}}{\sum_j e^{z_j}}$$

损失使用交叉熵（cross-entropy）：

$$L = -\sum_i t_i \ln(\hat{y}_i)$$

- t_i 为真实标签（one-hot编码）
- 反向传播需要求梯度：$\partial L / \partial W$

---

3. 理论上的三维张量

对中间层 $z = Wx + b$，严格地说：

$$\frac{\partial y_i}{\partial W_{j,k}}$$

- $i = 1..m$ → 输出索引
- $j = 1..m$ → W 行索引
- $k = 1..n$ → W 列索引

这形成 **一个 $m \times m \times n$ 的三维张量。**

直观理解：

- 每个输出 y_i 对应一页 "梯度矩阵"
- 每页大小 $m \times n$，存储它对 W 所有元素的偏导
- 大部分元素为零，因为每个 y_i 只依赖 W 的对应行（线性关系）

---

4. 链式法则分量形式

对任意输出 y_i 与 W 元素 $W_{j,k}$：

$$\frac{\partial y_i}{\partial W_{j,k}} = \sum_{r=1}^{m} \frac{\partial y_i}{\partial z_r} \frac{\partial z_r}{\partial W_{j,k}}$$

4.1 计算 $\partial z_r / \partial W_{j,k}$

$$z_r = \sum_{t=1}^{n} W_{r,t} x_t + b_r$$

$$\frac{\partial z_r}{\partial W_{j,k}} = \begin{cases} x_k & r = j \\ 0 & r \neq j \end{cases} = \delta_{rj} x_k$$

- 这是一个稀疏张量，只有对应行的元素非零

4.2 代入求和

$$\frac{\partial y_i}{\partial W_{j,k}} = \sum_{r=1}^{m} \frac{\partial y_i}{\partial z_r} \delta_{rj} x_k = \frac{\partial y_i}{\partial z_j} x_k$$

解释：

- 对每个输出 y_i，它对 W 的梯度只取决于第 j 行的 z
- 对逐元素激活函数（sigmoid、ReLU 等），$\frac{\partial y_i}{\partial z_j} = 0$ 如果 $i \neq j$，否则是 $\sigma'(z_i)$
- 这就是为什么三维张量"几乎稀疏"，大部分元素为 0

---

5. 矩阵形式：外积表达

将所有输出 i 和输入列 k 堆叠：

$$\frac{\partial y}{\partial W} = \begin{bmatrix} \frac{\partial y_1}{\partial W_{1,:}} \\ \frac{\partial y_2}{\partial W_{2,:}} \\ \vdots \\ \frac{\partial y_m}{\partial W_{m,:}} \end{bmatrix} = \begin{bmatrix} \frac{\partial y_1}{\partial z_1} x^\top \\ \frac{\partial y_2}{\partial z_2} x^\top \\ \vdots \\ \frac{\partial y_m}{\partial z_m} x^\top \end{bmatrix} = (\frac{\partial y}{\partial z}) x^\top$$

- $\frac{\partial y}{\partial z}$：长度 m 的列向量，每个元素是激活函数导数 $\sigma'(z_i)$
- $x^\top$：行向量
- 外积得到 $m \times n$ 矩阵 → 反向传播梯度矩阵

---

6. “三维张量 → 矩阵”的坍缩过程

形象化理解:

1. **原始三维张量** (m, m, n)
   - 第一维：输出索引 i
   - 第二维：W 行 j
   - 第三维：W 列 k
   - 非零元素只在 i=j 行
2. **链式法则求和（收缩）**：

$$\frac{\partial L}{\partial W_{j,k}} = \sum_i \frac{\partial L}{\partial y_i} \frac{\partial y_i}{\partial W_{j,k}}$$

   - Kronecker delta 选择 i=j
   - 结果：每行只保留非零部分
3. **外积形式**:
   - 每行梯度 = 对应输出导数 × 输入向量
   - 堆叠所有行 → $\partial y / \partial W = (\partial y / \partial z) x^\top$

---

7. 数值示例

假设:

- $m = 2$, $n = 3$
- 输入 $x = [1, 2, 3]^\top$
- z = [0.1, -1.0]
- sigmoid 导数：$\sigma'(0.1) \approx 0.2494$, $\sigma'(-1.0) \approx 0.1966$

外积:

$$\frac{\partial y}{\partial W} = \begin{bmatrix} 0.2494 \\ 0.1966 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \end{bmatrix} = \begin{bmatrix} 0.2494 & 0.4988 & 0.7482 \\ 0.1966 & 0.3932 & 0.5898 \end{bmatrix}$$

每行对应输出对 W 的梯度。

---

8. 链式法则与矩阵求导的核心逻辑

1. **链式法则**保证梯度可以按函数嵌套分解:

$$\frac{\partial L}{\partial W} = \frac{\partial L}{\partial y} \frac{\partial y}{\partial z} \frac{\partial z}{\partial W}$$

2. **线性变换** $z = Wx + b$:
   - $\frac{\partial z}{\partial W}$ 的非零元素是 x
   - 分量乘法 + 求和 → 直接形成外积

3. **稀疏性 + 求和 = 坍缩**：
   - 原本三维张量不必实际存储
   - 外积形式即可完全表示梯度矩阵

---

9. 关键总结

- 对中间层线性变换 $z = Wx + b$：
  - 严格 $\partial y/\partial W$ 是三维张量
  - 线性 + 逐元素激活 → 大部分元素为零
  - 链式法则收缩后 → 外积形式
  - 公式：

$$\boxed{\frac{\partial y}{\partial W} = (\frac{\partial y}{\partial z})x^\top}$$

- **理解要点**：
  1. 每个输出的梯度只依赖对应行的 W
  2. 外积表达式 = 梯度向量 × 输入向量
  3. 三维张量的"坍缩"实际上是链式法则 + 稀疏性的自然结果

曾神考我的问题的关键在于下面这个式子从第二行到第三行的推导：

$$\begin{aligned} &\frac{dy}{dW} \\ =&\frac{dy}{dz}\frac{dz}{dW} \\ =&\frac{dy}{dz}x^T \end{aligned}$$

下面附上曾神的两种证明过程：

思路一：使用爱因斯坦求和约定

$$\frac{d}{dW}L \qquad L = f(y) \qquad y = Wx.$$

$$\frac{d}{dA_{pq}}L = \frac{d}{dy_i}L \quad \frac{d}{dA_{pq}}y_i$$

$$\frac{d}{dA_{pq}}y_i = \frac{\partial(A_{ij}x_j)}{\partial A_{pq}}. = \delta_{ip}\delta_{jq}x_j = \delta_{ip}x_q.$$

$$\frac{d}{dA_{pq}}L = \frac{d}{dy_i}L\cdot\delta_{ip}x_q = \frac{d}{dy_p}L\cdot x^q \Longrightarrow \frac{d}{dA}L = \frac{d}{dy}L\cdot X^T$$

证明:

1. **定义:** $L = f(y)\, y = Wx$
2. **链式法则:** 根据链式法则，我们可以将 L 对 W 的导数分解为: $\frac{dL}{dW} = \frac{dL}{dy_i}\frac{dy_i}{dW}$

   补充：这里使用了爱因斯坦求和约定，对重复索引 i 进行求和。我在N-20230217.pdf中学过了这一块，但是现在忘了。

3. **计算 $\frac{dL}{dW_{pq}}$:** $\frac{dL}{dW_{pq}} = \frac{dL}{dy_i}\frac{dy_i}{dW_{pq}}$
4. **计算 $\frac{dy_i}{dW_{pq}}$:** $y_i = (Wx)_i = W_{ij}x_j$ (同样使用爱因斯坦求和约定) 所以，当 W 矩阵的元素被表示为 A_{ij} 时: $y_i = A_{ij}x_j$ $\frac{dy_i}{dW_{pq}} = \frac{\partial(A_{ij}x_j)}{\partial W_{pq}} = \delta_{ip}\delta_{jq}x_j = \delta_{ip}x_q$

   补充：这里的 δ_{ip} 和 δ_{jq} 是 Kronecker delta 符号。当 $i = p$ 且 $j = q$ 时为 1，否则为 0。所以 $\frac{\partial(A_{ij}x_j)}{\partial W_{pq}}$ 实际上是在询问当 $i = p$ 时 $A_{pj}x_j$ 对 W_{pq} 的偏导，结果就是 x_q。

5. **代回链式法则:** $\frac{dL}{dW_{pq}} = \frac{dL}{dy_i}(\delta_{ip}x_q) = \frac{dL}{dy_p}x_q$

   补充：根据 Kronecker delta 的性质，$\frac{dL}{dy_i}\delta_{ip}$ 会使得所有 $i \neq p$ 的项为零，只保留 $i = p$ 的项，所以结果是 $\frac{dL}{dy_p}$。

6. **将结果写成矩阵形式:** 观察结果 $\frac{dL}{dW_{pq}} = \frac{dL}{dy_p}x_q$，这表示 L 对 W 矩阵的第 (p, q) 个元素的导数。 将其组织成矩阵形式，$\frac{dL}{dW}$ 的第 (p, q) 个元素就是 $\frac{dL}{dy_p}x_q$。 这意味着 $\frac{dL}{dW} = (\frac{dL}{dy})(x^T)$。 所以，$\frac{dL}{dW} = (\frac{dL}{dy})x^T$

补充:

- 在一般情况下，如果 L 是一个标量， y 是一个向量 ($N \times 1$)， W 是一个矩阵 ($M \times N$)， x 是一个向量 ($N \times 1$)。那么 $\frac{dL}{dW}$ 应该是一个 $M \times N$ 的矩阵。
- 当计算 $\frac{dL}{dy_i}$ 时，得到的是一个 $1 \times N$ 的行向量 (或转置后是 $N \times 1$ 的列向量)。
- 当计算 $\frac{dy_i}{dW_{pq}}$ 时，得到的是一个标量 x_q (当 $i = p$ 时)。
- 最终 $\frac{dL}{dW_{pq}} = \frac{dL}{dy_p}x_q$。如果我们将 $\frac{dL}{dy}$ 看作一个行向量， x^T 看作一个行向量，那么它们的"外积"会形成一个矩阵。
- 曾神提到"线性函数导致求和的式子里出现了两个kronecker符号，就那个δ 导致最终结果只有1维了"。这可能是由于在将最终的矩阵形式写出来时，对于特定情况或误解了矩阵乘法的维度。在正确的矩阵求导中，如果 L 是标量， W 是矩阵，那么导数 $\frac{dL}{dW}$ 应该是一个与 W 同维度的矩阵。
- 曾神还提到"在一般的情形中，虽然维度消失了，但由于不能得知函数的表达式，所以写出来不会发生任何变化"。这也许暗示了在某些中间步骤中，为了简化表示，导致了维度的隐含变化（？）
- "对于线性函数，用Kronecker乘积展开它，会出现一个单位矩阵，这个单位矩阵可以与向量合并"，说明了 Kronecker 乘积在处理线性函数和维度变化中的作用。

思路二：使用向量化和 Kronecker 乘积

$$\frac{\partial y}{\partial \mathrm{vec}(W)} = \frac{\partial y}{\partial \vec{z}} \frac{\partial \vec{z}}{\partial \mathrm{vec}(W)}$$

$$= \vec{V} \begin{bmatrix} \frac{\partial z_1}{\partial \mathrm{vec}(W)} \\ \vdots \\ \frac{\partial z_n}{\partial \mathrm{vec}(W)} \end{bmatrix}$$

$$= \boxed{1 \times n}\ \boxed{n \times nm}$$

$$= \boxed{1 \times nm}$$

$y = f(\vec{z})$　y: ~~scalar~~ ~~vector~~ scalar

$\vec{z} = g(W)$.　$\vec{z}$: vector ($n \times 1$)

W: matrix ($n \times m$)

$$\frac{\partial y}{\partial \vec{z}} = \left[\frac{\partial y}{\partial z_1}, \dots, \frac{\partial y}{\partial z_n}\right] \text{ 记为 } \vec{V}.$$

$$\frac{\partial \vec{z}}{\partial \mathrm{vec}(W)} = \begin{bmatrix} \frac{\partial z_1}{\partial \mathrm{vec}(W)} \\ \vdots \\ \frac{\partial z_n}{\partial \mathrm{vec}(W)} \end{bmatrix}$$

$$\frac{\partial a \to m\text{维}}{\partial b \to n\text{维}} = \begin{bmatrix} \ \end{bmatrix} \ m \times n\text{维}$$

如果 $\vec{z} = W\vec{x}$.　$\vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix}$

$$= (x^T \otimes I_n)\, \mathrm{vec}(W) \qquad \frac{\partial \vec{z}}{\partial \mathrm{vec}(W)} = x^T \otimes I_n$$

代回上式: $\frac{\partial y}{\partial \mathrm{vec}(W)} = \vec{V} \cdot (x^T \otimes I_n)$　scalar $x_1, \dots, x_m$

$$= \vec{V}\, [x_1 I_n \ x_2 I_n \ \dots \ x_m I_n]$$

$$= [x_1 \vec{V} \ x_2 \vec{V} \ \dots \ x_m \vec{V}]$$

$$= (\mathrm{vec}(\vec{V}\vec{x}))^T.$$

$$\Rightarrow \frac{\partial y}{\partial W} = \vec{V}\vec{x} = \frac{\partial y}{\partial \vec{z}} \vec{x}$$

证明:

1. **定义:** $y = f(Z)\ Z = Wx$
2. **链式法则 (向量化形式):** $\frac{\partial y}{\partial \mathrm{vec}(W)} = \frac{\partial y}{\partial Z} \frac{\partial Z}{\partial \mathrm{vec}(W)}$

   补充：这里 $\mathrm{vec}(W)$ 是将矩阵 W 按列堆叠成一个列向量。$\frac{\partial y}{\partial Z}$ 会是一个行向量 (如果 y 是标量，Z 是列向量)。

3. **计算 $\frac{\partial y}{\partial Z}$:** 如果 y 是标量，Z 是 $m \times 1$ 的向量，那么 $\frac{\partial y}{\partial Z}$ 是一个 $1 \times m$ 的行向量：
   $\frac{\partial y}{\partial Z} = \left[\frac{\partial y}{\partial z_1}, \frac{\partial y}{\partial z_2}, \dots, \frac{\partial y}{\partial z_m}\right]$
4. **计算 $\frac{\partial Z}{\partial \mathrm{vec}(W)}$:** $Z = Wx$ 我们将 Z 的每个元素展开，假设 W 是 $m \times n$ 矩阵，x 是 $n \times 1$ 向量。 $z_i = \sum_{j=1}^{n} W_{ij} x_j$

   现在考虑 $\mathrm{vec}(W)$。假设 W 有 $m \times n$ 个元素。 $\frac{\partial Z}{\partial \mathrm{vec}(W)}$ 将是一个大的矩阵，它的每一行对应 Z 的一个元素，每一列对应 $\mathrm{vec}(W)$ 的一个元素。

我们知道 $\mathrm{vec}(ABC) = (C^T \otimes A)\mathrm{vec}(B)$。这里 $Z = Wx = WI_n x$ (将 x 视为 $I_n x$ 方便应用规则)。 或者更直接地， $Z = Wx$ 可以看作是一个线性变换。 $\mathrm{vec}(Z) = \mathrm{vec}(Wx)$.

为了计算 $\frac{\partial Z}{\partial \mathrm{vec}(W)}$，我们也可以使用以下性质： $\frac{\partial (Ax)}{\partial x} = A^T$ (当 A 为常数矩阵时) $\frac{\partial (x^T A)}{\partial x} = A$

但这里是对 W 求导，且 W 是矩阵。 从笔记中可以看出，为了处理 $Z = Wx$，曾神使用了 Kronecker 乘积和向量化： $Z = Wx$ $\mathrm{vec}(Z) = \mathrm{vec}(Wx)$ 可以写成 $\mathrm{vec}(Z) = (x^T \otimes I_m)\mathrm{vec}(W)$。 因此， $\frac{\partial \mathrm{vec}(Z)}{\partial \mathrm{vec}(W)} = (x^T \otimes I_m)$

补充：这里 I_m 是 $m \times m$ 的单位矩阵。$x^T \otimes I_m$ 的维度是 $m \times (mn)$。

5. **代回链式法则:** $\frac{\partial y}{\partial \mathrm{vec}(W)} = \frac{\partial y}{\partial Z}(x^T \otimes I_m)$

   补充：$\frac{\partial y}{\partial Z}$ 是一个 $1 \times m$ 的行向量， $(x^T \otimes I_m)$ 是一个 $m \times (mn)$ 的矩阵。最终结果是一个 $1 \times (mn)$ 的行向量，这正是 $\frac{\partial y}{\partial \mathrm{vec}(W)}$ 的形式。

6. **将结果还原为矩阵形式 (如果需要):** 如果 $\frac{\partial y}{\partial \mathrm{vec}(W)}$ 是一个 $1 \times (mn)$ 的行向量，我们可以通过逆向量化操作将其还原为 $m \times n$ 的矩阵 $\frac{\partial y}{\partial W}$。 通过一些线性代数的性质，可以证明 $\frac{\partial y}{\partial W} = \frac{\partial y}{\partial Z} x^T$。 这与思路一的结果一致。

   笔记的推导过程： $\frac{\partial y}{\partial \mathrm{vec}(W)} = \frac{\partial y}{\partial Z}\frac{\partial Z}{\partial \mathrm{vec}(W)}$ 其中 $\frac{\partial y}{\partial Z} = \vec{v}$ (行向量)。 $\frac{\partial Z}{\partial \mathrm{vec}(W)} = (x^T \otimes I_n)$ (笔记中写成 I_n 但根据维度应该是 I_m) $\frac{\partial y}{\partial \mathrm{vec}(W)} = \vec{v}(x^T \otimes I_m) = [\vec{v}x_1, \vec{v}x_2, \ldots, \vec{v}x_n]$ 这应该是将 $\vec{v}$ 与 x 的每个元素相乘，然后堆叠起来。

   笔记的最后一步： $\Rightarrow \frac{\partial y}{\partial W} = \vec{V}\vec{X} = \frac{\partial y}{\partial Z}\vec{X}$ 这里 $\vec{V}$ 是 $\frac{\partial y}{\partial Z}$， $\vec{X}$ 是 x^T。 所以 $\frac{\partial y}{\partial W} = \frac{\partial y}{\partial Z}x^T$。

补充:

- 第二种思路使用向量化和 Kronecker 乘积，更加形式化和严谨地处理了矩阵求导，对我而言比较难理解。
- “第三行相比于第二行，可以认为是提前把这个矩阵乘法算了一部分”。这是对核心问题的更严谨的解释。在推导中，通过对 $Z = Wx$ 进行向量化处理，其实已经隐含地进行了部分矩阵乘法运算。
- “所以维度消失的现象体现在了外部”，这也许意味着在向量化过程中，虽然中间表示的维度增加了 (如 $\mathrm{vec}(W)$)，但最终的导数结果在还原为矩阵形式时，其维度与原始矩阵 W 保持一致。

总结

相同结论： $\frac{dL}{dW} = (\frac{dL}{dy})x^T$。

- **思路一 (爱因斯坦求和约定):** 侧重于分量级别的推导，通过 Kronecker delta 处理索引，最终将结果组合成矩阵形式。这种方法比较物理，应该是在张量分析中常用的方法，也是我比较好接受的。
- **思路二 (向量化和 Kronecker 乘积):** 侧重于整体矩阵和向量的运算，通过向量化将矩阵转换为向量，再利用 Kronecker 乘积的性质进行求导。这种方法据说在机器学习和统计学中对矩阵微积分的推导非常有用。