

梯度、谱定理与紧致性笔记

第一版 · 2026-10-18 (北京时间)

来源：Transformer梯度矩阵推导。按完整 20 轮对话整理，保留追问、定义、两条证明路线和条件边界；将聊天演示改为本版静态图示，补全原对话略过的分析与归纳步骤。

正文目录

第 1 章：从矩阵梯度追到实数完备性的思考路线	3
第 2 章：标量对矩阵的梯度、链式法则与参数更新	5
第 3 章：梯度为何是最速方向，以及它位于哪个空间	14
第 4 章：内部极值为什么必须驻点，为什么还要看二阶	17
第 5 章：矩阵、特征向量与换坐标：看懂正交对角化	18
第 6 章：正交矩阵为什么能选出，以及 Gram—Schmidt 如何操作	29
第 7 章：紧致性、闭且有界和收敛子序列	41
第 8 章：连续像为什么紧致，连续函数为什么能取到极值	53
第 9 章：谱定理的代数路线：先从特征方程找一个实方向	59
第 10 章：谱定理的球面路线与共同的降维归纳	64
第 11 章：正定性、正特征值与统一二次下界	72
第 12 章：回到 Hessian：用泰勒余项严格完成二阶判别	81
第 13 章：具体例子、易错边界与回到训练问题	90
第 14 章：原对话逐轮覆盖索引与复习问题	92

第 1 章：从矩阵梯度追到实数完备性的思考路线

[返回目录](#)

这份笔记的起点是：“Transformer 中那个梯度的矩阵是怎么来着？”之后的问题逐步变成：梯度为何同形、为何用负梯度、极值处为何梯度为零、Hessian 如何判别、为什么实对称矩阵能正交对角化、为什么单位球面上的连续函数能取到极值、紧致为什么等价于闭且有界。

后面的追问并没有偏离起点：我们想理解的是，局部下降方向、一阶驻点和二阶曲率之间，究竟有哪些可以证明的联系。

1.1 原来的追问链与本版阅读顺序

原对话不断向前追问前提；本版将补充知识放到首次需要它的证明之前。第 2—4 章建立优化问题；第 5—6 章复习矩阵语言；第 7—8 章补分析基础；第 9—10 章完成谱定理的两条路线；第 11—12 章回到 Hessian；第 13 章检验边界；第 14 章逐轮回查。

两条谱定理路线共享“正交补不变、限制仍对称、归纳降维”。它们只在第一个实特征向量的存在性证明上不同：一条引用代数基本定理，另一条使用紧致性与极值定理。对话里“球面构造太精巧”的疑惑保留在第 9—10 章的动机说明中。

1.2 符号与对象

符号	类型和含义
L	实值标量损失
X, W, Y	线性层输入、参数、输出；原对话也记作 H_1, W_2, H_2
$\nabla_W L$	与 W 同形的梯度矩阵
η	标准梯度下降中的正实数学习率

符号	类型和含义
$f: U \rightarrow \mathbb{R}$	定义在开集 $U \subseteq \mathbb{R}^n$ 上的实值函数
p, h, v	输入点、位移、方向向量；默认列向量
$H = \nabla^2 f(p)$	本版的 Hessian；不要与线性层激活 H_1, H_2 混淆
A	一般实对称矩阵
Q, Λ	正交特征向量矩阵、按相同次序排列特征值的对角矩阵
x^T, x^*	转置、复共轭转置
K, S^{n-1}	集合、单位球面

Q 在这里表示正交矩阵，与 Attention 的 Query 矩阵没有特殊关系。大写 Λ 是字母 Lambda，在本笔记中被指定为对角矩阵；它不是“取对角矩阵”的运算符。

1.3 本文证明到哪里

本文证明线性层梯度、最速方向、内部极值必要条件、Gram—Schmidt、Heine—Borel 两个方向、有限维序列紧致等价、连续像紧致与极值定理、两条谱定理路线、正定性与特征值等价、统一二次下界和二阶判别法。

实数完备性（等价地可从上确界性质出发）作为分析基础使用；有限维线性代数中的行列式与可逆性、秩—零化度、矩阵乘法及一元微分法则作为基础使用。代数路线引用代数基本定理，其证明可回查项目里的[代数基本定理笔记](#)。球面路线独立于这个引用。

二阶泰勒公式与混合偏导相等在第 12 章给出推导。本文不证明训练一定收敛或能找到全局最优点；最后会明确这些结论与局部微分性质之间的差别。

第 2 章：标量对矩阵的梯度、链式法则与参数更新

[返回目录](#)

你之前问过 Transformer 反向传播时，梯度为什么也是一个矩阵，以及为什么计算时需要转置。我们沿用之前的 H_1, W_2, H_2, L 来回顾。

一、梯度矩阵是什么？

假设 Transformer 中有一个线性变换：

$$H_2 = H_1 W_2$$

其中：

$$H_1 \in \mathbb{R}^{n \times d_{\text{in}}}$$

$$W_2 \in \mathbb{R}^{d_{\text{in}} \times d_{\text{out}}}$$

$$H_2 \in \mathbb{R}^{n \times d_{\text{out}}}$$

最终通过后续计算得到一个标量损失：

$$L \in \mathbb{R}$$

我们想知道：如果 H_2 的每一个元素发生微小变化，损失 L 会变化多少？

于是分别求偏导，并将结果按照 H_2 的形状排列：

$$\frac{\partial L}{\partial H_2} = \begin{bmatrix} \frac{\partial L}{\partial h_{11}} & \frac{\partial L}{\partial h_{12}} & \cdots \\ \frac{\partial L}{\partial h_{21}} & \frac{\partial L}{\partial h_{22}} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

这个就是**梯度矩阵** (gradient matrix)。

需要注意，它不是矩阵求导后凭空出现的新概念，而是将每个元素的偏导数重新排列成矩阵。

二、梯度如何向前一层传递？

我们已经知道：

$$H_2 = H_1 W_2$$

同时已经计算出：

$$G_2 = \frac{\partial L}{\partial H_2}$$

现在想求：

$$G_1 = \frac{\partial L}{\partial H_1}$$

依据链式法则，最终有：

$$\boxed{\frac{\partial L}{\partial H_1} = \frac{\partial L}{\partial H_2} W_2^T}$$

为什么会出现转置？

考虑一个最简单的例子。

$$H_1 = [x_1 \quad x_2]$$

$$W_2 = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

因此：

$$H_2 = [ax_1 + cx_2 \quad bx_1 + dx_2] = [y_1 \quad y_2]$$

根据链式法则：

$$\frac{\partial L}{\partial x_1} = \frac{\partial L}{\partial y_1}a + \frac{\partial L}{\partial y_2}b$$

$$\frac{\partial L}{\partial x_2} = \frac{\partial L}{\partial y_1}c + \frac{\partial L}{\partial y_2}d$$

写成矩阵：

$$\begin{bmatrix} \frac{\partial L}{\partial x_1} & \frac{\partial L}{\partial x_2} \end{bmatrix} = \begin{bmatrix} \frac{\partial L}{\partial y_1} & \frac{\partial L}{\partial y_2} \end{bmatrix} \underbrace{\begin{bmatrix} a & c \\ b & d \end{bmatrix}}_{W_2^T}$$

所以，转置并不是单纯为了让矩阵维度匹配，而是链式法则展开后自然得到的结果。

三、反向传播的三个重要梯度公式

对于线性层：

$$H_2 = H_1 W_2$$

令：

$$G_2 = \frac{\partial L}{\partial H_2}$$

那么：

1. 对输入 H_1 求梯度：将梯度传回前一层

$$\frac{\partial L}{\partial H_1} = G_2 W_2^T$$

2. 对权重 W_2 求梯度：用于更新模型参数

$$\frac{\partial L}{\partial W_2} = H_1^T G_2$$

3. 对输出 H_2 求梯度：从后续计算传来的梯度

$$G_2 = \frac{\partial L}{\partial H_2}$$

注意，求偏导时其他变量仍然保持当前数值不变。比如求 $\partial L / \partial H_1$ 时， W_2 视为常量；但 W_2 本身也是需要训练的参数，所以我们还会单独计算 $\partial L / \partial W_2$ 。

四、如何理解整个反向传播？

对 Transformer 来说，整个模型虽然包含 Q、K、V、Attention、MLP 等复杂模块，但反向传播的基本原理都是这个链式法则。

还有一个很容易混淆的概念：

- **梯度矩阵 (Gradient matrix)**：标量 L 对矩阵各元素求偏导，结果通常与原矩阵同形状。
- **雅可比矩阵 (Jacobian matrix)**：向量对向量求导，表示每个输出元素对每个输入元素的影响。

我们上面实际上是利用雅可比矩阵描述的关系，通过链式法则计算梯度，只不过矩阵运算让我们不必显式构造巨大的雅可比矩阵。

你之前最容易混淆的地方，其实就是三个不同的问题：对输入求导、对权重求导、对输出求导。只要始终明确正在求谁的梯度，转置应该放在哪一边就容易理解了。

补充：权重梯度逐元素推导

记 $X \in \mathbb{R}^{m \times d}$ 、 $W \in \mathbb{R}^{d \times r}$ 、 $Y = XW \in \mathbb{R}^{m \times r}$ ，并令 $G_{ij} = \partial L / \partial Y_{ij}$ 。由 $Y_{ij} = \sum_k X_{ik} W_{kj}$ ，有

$$\frac{\partial L}{\partial X_{ik}} = \sum_j G_{ij} W_{kj}, \quad \frac{\partial L}{\partial W_{kj}} = \sum_i X_{ik} G_{ij}.$$

因此 $\nabla_X L = GW^T$ 、 $\nabla_W L = X^T G$ 。维度检查只能排错，不能替代这段链式法则证明。如果另有通往 L 的依赖路径，还要累加各路径贡献。

用 Frobenius 内积 $\langle A, B \rangle_F = \sum_{ij} A_{ij} B_{ij} = \text{tr}(A^T B)$ ，同一关系也可写成

$$dL = \langle G, dY \rangle_F = \langle GW^T, dX \rangle_F + \langle X^T G, dW \rangle_F.$$

这个等式解释了梯度为何按照与原变量相同的形状表示：它是微分这个线性泛函在 Frobenius 内积下的代表。

对，完全正确。我们之前讨论的反向传播梯度，都是标量对矩阵求梯度，而不是矩阵对矩阵求导。

关键在于：损失函数 L 最终是一个标量。

例如：

$$H_2 = H_1 W_2$$

虽然 H_1, H_2, W_2 都是矩阵，但最终的损失是：

$$L \in \mathbb{R}$$

因此，我们计算的是：

$$\boxed{\frac{\partial L}{\partial H_1}, \quad \frac{\partial L}{\partial H_2}, \quad \frac{\partial L}{\partial W_2}}$$

它们都是标量对矩阵的梯度。

那么矩阵对矩阵求导是什么？

例如：

$$Y = XW$$

如果直接计算：

$$\frac{\partial Y}{\partial X}$$

由于 Y 和 X 都是矩阵，严格来说，我们需要描述 Y 中每个元素对 X 中每个元素的偏导数。

假设：

$$X \in \mathbb{R}^{2 \times 3}, \quad Y \in \mathbb{R}^{2 \times 4}$$

那么 Y 有 8 个元素， X 有 6 个元素。

因此，如果把两个矩阵分别展平成向量，它们之间的雅可比矩阵就是：

$$J = \frac{\partial \text{vec}(Y)}{\partial \text{vec}(X)} \in \mathbb{R}^{8 \times 6}$$

也可以不展平，使用四阶张量表示。

为什么 Transformer 不需要显式计算这种巨大的雅可比矩阵？

因为反向传播使用的是向量-雅可比积（Vector-Jacobian Product, VJP）。

例如：

$$H_2 = H_1 W_2$$

虽然中间存在 H_2 对 H_1 的雅可比矩阵，但是我们真正需要的是：

$$\boxed{\frac{\partial L}{\partial H_1} = \frac{\partial L}{\partial H_2} W_2^T}$$

这相当于将损失梯度与局部雅可比矩阵相乘，而不需要显式构建它。

所以你可以这样记忆：

- 前向传播：矩阵经过矩阵运算，得到新的矩阵。
- 反向传播：从标量损失 L 出发，逐层计算 L 对各个矩阵的梯度。
- 数学原理：矩阵之间的雅可比关系 + 链式法则。
- 计算实现：通常只计算向量-雅可比积，避免构造完整的雅可比矩阵。

这也是为什么深度学习中的自动微分能够高效地处理包含海量参数的神经网络。

对，这正是梯度下降（Gradient Descent）的核心要求：我们需要计算出一个与权重矩阵 W 形状完全相同的梯度矩阵，才能逐元素更新参数。

假设权重矩阵是：

$$W = \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \\ w_{31} & w_{32} \end{bmatrix} \in \mathbb{R}^{3 \times 2}$$

损失函数 L 是一个标量，对 W 求梯度：

$$\frac{\partial L}{\partial W} = \begin{bmatrix} \frac{\partial L}{\partial w_{11}} & \frac{\partial L}{\partial w_{12}} \\ \frac{\partial L}{\partial w_{21}} & \frac{\partial L}{\partial w_{22}} \\ \frac{\partial L}{\partial w_{31}} & \frac{\partial L}{\partial w_{32}} \end{bmatrix} \in \mathbb{R}^{3 \times 2}$$

然后梯度下降的更新公式就是：

$$W_{\text{new}} = W_{\text{old}} - \eta \frac{\partial L}{\partial W}$$

其中 η 是学习率（learning rate）。

展开到每一个参数：

$$w_{ij}^{\text{new}} = w_{ij}^{\text{old}} - \eta \frac{\partial L}{\partial w_{ij}}$$

也就是说，每个权重都有一个对应的偏导数，告诉我们应该朝哪个方向调整这个权重，才能在局部近似下使损失减小。

再联系之前的线性层：

$$H_2 = H_1 W_2$$

反向传播会同时计算两种梯度：

梯度	作用
$\frac{\partial L}{\partial W_2} = H_1^T \frac{\partial L}{\partial H_2}$	更新权重矩阵 W_2
$\frac{\partial L}{\partial H_1} = \frac{\partial L}{\partial H_2} W_2^T$	将梯度传递给前一层

这里有一个重要区别：

对 W_2 求梯度，是为了更新这一层的参数；对 H_1 求梯度，是为了继续反向传播。

所以，虽然反向传播会计算很多梯度矩阵，但真正用于梯度下降更新的，是损失函数对各个可训练参数矩阵的梯度。

你现在其实已经把 Transformer 训练过程的关键逻辑串起来了：

$$\boxed{\text{前向计算 } L \longrightarrow \text{反向计算 } \nabla_W L \longrightarrow \text{更新 } W}$$

Transformer 中的 W_Q, W_K, W_V, W_O 以及 MLP 层的权重矩阵，都是按照这个基本原理学习的。

对， η （希腊字母 eta）就是**学习率（Learning Rate）**，在标准梯度下降中是一个标量（scalar）。

梯度下降公式：

$$\boxed{W_{\text{new}} = W_{\text{old}} - \eta \frac{\partial L}{\partial W}}$$

其中：

- W ：权重矩阵，例如 3×2 。
- $\frac{\partial L}{\partial W}$ ：梯度矩阵，同样是 3×2 。
- η ：学习率，是一个标量，例如 0.001。

因此， η 会乘以梯度矩阵中的**每一个元素**：

$$\eta \begin{bmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{bmatrix} = \begin{bmatrix} \eta g_{11} & \eta g_{12} \\ \eta g_{21} & \eta g_{22} \end{bmatrix}$$

补充：矩阵到矩阵的导数没有消失

对映射 $F_W : \mathbb{R}^{m \times d} \rightarrow \mathbb{R}^{m \times r}$ 、 $F_W(X) = XW$ ，导数是线性映射

$DF_W(X)[E] = EW$ 。按列展平后，Jacobian 的形状为 $mr \times md$ ，并可写为 $W^T \otimes I_m$ 。本版不需要计算 Kronecker 积；只需要分清它与同形梯度是不同对象。

“前向矩阵对矩阵、最终标量损失对参数求梯度”可以同时成立。采用列梯度约定，反传写成 $J^T g$ ；采用行协向量约定，写成 $g^T J$ ，这就是通常所说的 VJP。名称里的行列约定不应改变实际链式法则。

现代自适应优化器可保留标量基础学习率，同时逐参数缩放更新；这不改变本章对标准梯度下降的定义。

第 3 章：梯度为何是最速方向，以及它位于哪个空间

[返回目录](#)

对，**梯度向量 (Gradient Vector)** 指向函数值增长最快的方向，而梯度的模长表示该方向上的最大变化率。

不过需要区分两个概念：

- **梯度的方向**：函数值上升最快的方向。
- **梯度的模长**：沿该方向移动单位距离时，函数值的最大瞬时变化率。

假设有一个二元函数：

$$z = f(x, y)$$

它的梯度为：

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix}$$

梯度位于输入变量构成的空间中，指向函数值增长最快的方向，并与等高线正交（梯度非零时）。

为什么梯度是变化率最大的方向？

考虑沿单位方向向量 $\mathbf{u}$ 移动，函数的方向导数为：

$$D_{\mathbf{u}}f = \nabla f \cdot \mathbf{u}$$

根据向量点积：

$$D_{\mathbf{u}}f = \|\nabla f\| \|\mathbf{u}\| \cos \theta$$

因为 $\|\mathbf{u}\| = 1$ ，所以：

$$D_{\mathbf{u}}f = \|\nabla f\| \cos \theta$$

当 $\theta = 0$ 时，变化率最大：

$$\boxed{\max D_{\mathbf{u}}f = \|\nabla f\|}$$

也就是沿梯度方向移动。

联系 Transformer 的梯度下降

Transformer 训练的目标是减小损失函数 L ，而不是增大它。

因此，更新参数时沿着负梯度方向：

$$\boxed{W_{\text{new}} = W_{\text{old}} - \eta \nabla_W L}$$

这里虽然 $\nabla_W L$ 是矩阵，但我们可以把矩阵中的所有参数展开成一个高维向量。

例如，一个 3×2 的权重矩阵实际上对应 6 个独立参数，因此损失函数可以理解为定义在六维参数空间上的函数：

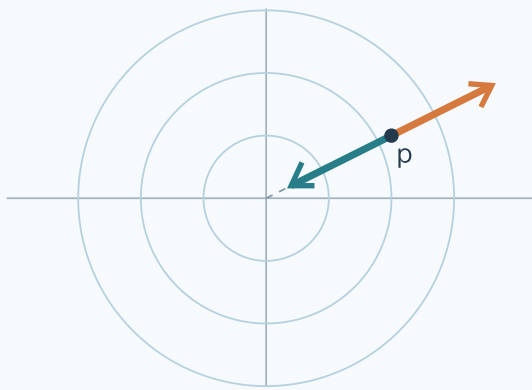
$$L = f(w_{11}, w_{12}, w_{21}, w_{22}, w_{31}, w_{32})$$

梯度矩阵本质上就是这个六维梯度向量的另一种排列方式。

所以，梯度下降可以理解成：在模型参数构成的高维空间中，每一步都沿着损失函数局部下降最快的方向调整参数。

需要注意，这里的方向是在参数空间中，而不是三维曲面上的切线方向；它是损失函数在参数空间中的最陡上升方向。

图 1 · 梯度在输入平面内，与等高线正交



圆： $f(x,y) = x^2 + y^2$ 的等高线

橙色： $\nabla f(p) = (2,1)$

绿色：负梯度方向

比较单位方向的一阶变化率。

有限步长还要考虑高阶变化。

点 $p=(1,1/2)$ 。径向梯度与圆的切线垂直；箭头表示输入空间方向，不是三维曲面的法向量。

补充：梯度为零时，以及有限步长的限制

方向导数公式要求 f 在该点可微。若 $\nabla f(p) = 0$ ，所有单位方向的一阶方向导数都是零，此时不存在由非零梯度指定的唯一最陡方向。

对矩阵变量，在 $\|E\|_F = 1$ 的约束下同样有

$$DL(W)[E] = \langle \nabla_W L, E \rangle_F \geq -\|\nabla_W L\|_F.$$

非零梯度时，取 $E = -\nabla_W L / \|\nabla_W L\|_F$ 达到最小值。这里“最快”依赖所选的欧氏/Frobenius 度量；若换度量，最陡方向的表示也可能变化。

若梯度非零，可微性给出

$$f(p - \eta \nabla f(p)) - f(p) = -\eta \|\nabla f(p)\|^2 + o(\eta), \quad \eta \downarrow 0.$$

所以足够小的正步长会下降；任意步长则没有保证。对 $f(x) = x^2$ ，更新为 $x_{new} = (1 - 2\eta)x$ ，对非零 x 仅当 $0 < \eta < 1$ 时一步严格下降。把局部最速方向直接解释成“大步也最优”，会遗漏高阶变化。

对图形 $z = f(x, y)$ ， ∇f 是输入平面中的二维向量。曲面写成隐式方程

$F(x, y, z) = f(x, y) - z = 0$ 时，其法向量为 $(f_x, f_y, -1)$ ，与二维梯度不是同一个对象。

第 4 章：内部极值为什么必须驻点，为什么还要看二阶

[返回目录](#)

4.1 定义与条件

设 $f: U \rightarrow \mathbb{R}$ ， $U \subseteq \mathbb{R}^n$ 为开集。若存在 $r > 0$ ，对所有 $x \in B(p, r)$ 有 $f(x) \geq f(p)$ ，则 p 为局部极小点；除 p 外严格大于时为严格局部极小点。极大点反向定义。驻点指可微且 $\nabla f(p) = 0$ 。

“最高/最低”在本笔记的判别法中指局部极值，除非明确写全局。二阶信息只描述邻域，不能直接推出全局结论。

4.2 从一元费马定理到多元必要条件

若一元可微函数 g 在内部点 0 取局部极小值，则 $t > 0$ 时 $(g(t) - g(0))/t \geq 0$ ， $t < 0$ 时该商 ≤ 0 。两侧极限存在且相等，只能都等于零。因此 $g'(0) = 0$ 。

对每个坐标单位向量 e_i ，令 $g_i(t) = f(p + te_i)$ 。内部点条件允许正、负小 t ；局部极小条件使 g_i 也在零处取极小值。因此

$$\partial_i f(p) = g'_i(0) = 0 \quad (i = 1, \dots, n), \quad \nabla f(p) = 0.$$

局部极大对 $-f$ 应用同一证明即可。 $x^2 + y^2$ 、 $-x^2 - y^2$ 、 $x^2 - y^2$ 在原点全是驻点，却分别为极小、极大、鞍点。这说明一阶条件不足以分类。

4.3 二阶量为何形成一个矩阵

当 $f \in C^2(U)$, 定义 $H_{ij}(p) = \partial_i \partial_j f(p)$ 。沿固定方向 v , 置 $g(t) = f(p + tv)$, 链式法则给出

$$g'(0) = \nabla f(p)^T v, \quad g''(0) = \sum_{i,j} v_i v_j \partial_i \partial_j f(p) = v^T H(p) v.$$

Hessian 把所有方向的二阶变化率组织起来；只看坐标方向的 f_{xx}, f_{yy} 会漏掉交叉项。下一步要解释的正是：为什么这个二次型能由正交主方向描述，以及为什么“每个方向都正”能变成控制整个邻域的统一估计。

第 5 章：矩阵、特征向量与换坐标：看懂正交对角化

[返回目录](#)

你之前学 Transformer 时已经接触过矩阵乘法、转置和梯度，现在正好可以把线性代数中的几个核心概念重新串起来。

我们重点解释这个公式：

$$H = Q \Lambda Q^T$$

它叫做实对称矩阵的正交对角化 (Orthogonal Diagonalization)，也是谱定理 (Spectral Theorem) 的重要结论。

理解它需要依次掌握：矩阵乘法、转置、特征值、特征向量、正交矩阵，以及对角化。

一、先回顾：矩阵到底是什么？

最重要的理解是：矩阵可以表示一个线性变换。

例如：

$$A = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$$

对于一个二维向量：

$$\mathbf{x} = \begin{bmatrix} x \\ y \end{bmatrix}$$

矩阵乘法得到：

$$A\mathbf{x} = \begin{bmatrix} 2x \\ 3y \end{bmatrix}$$

也就是说，这个矩阵将向量的横坐标放大 2 倍，纵坐标放大 3 倍。

但一般的矩阵并不是这么简单。

例如：

$$B = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}$$

我们得到：

$$B\mathbf{x} = \begin{bmatrix} 3x + y \\ x + 3y \end{bmatrix}$$

现在两个坐标相互混合了。

那么问题来了：

有没有办法换一个坐标系，让矩阵 B 也变成像 A 那样简单的对角矩阵？

答案就是特征值分解。

二、什么是特征值和特征向量？

英文分别是：

- Eigenvalue: 特征值
- Eigenvector: 特征向量

对于矩阵 B ，如果存在非零向量 $\mathbf{v}$ ，使得：

$$\boxed{B\mathbf{v} = \lambda\mathbf{v}}$$

那么：

- $\mathbf{v}$ 是特征向量。
- λ 是对应的特征值。

这个公式的几何含义非常重要：

矩阵变换后，这个特殊向量的方向没有发生改变（允许反向），只是长度被缩放了 $|\lambda|$ 倍。

例如刚才的矩阵：

$$B = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}$$

我们尝试向量：

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

计算：

$$B\mathbf{v}_1 = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 4 \\ 4 \end{bmatrix}$$

因此：

$$B\mathbf{v}_1 = 4\mathbf{v}_1$$

这意味着 $(1, 1)$ 方向上的向量经过变换后，仍然沿这个方向，但长度变为 4 倍。

再尝试：

$$\mathbf{v}_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

得到：

$$B\mathbf{v}_2 = \begin{bmatrix} -2 \\ 2 \end{bmatrix} = 2\mathbf{v}_2$$

所以：

$$B\mathbf{v}_2 = 2\mathbf{v}_2$$

我们已经找到了两组特征值和特征向量：

特征值	对应的特征向量	作用
$\lambda_1 = 4$	$(1, 1)$	沿此方向放大 4 倍
$\lambda_2 = 2$	$(-1, 1)$	沿此方向放大 2 倍

三、特征值到底怎么计算出来？

刚才是直接尝试，现在用标准方法。

由：

$$B\mathbf{v} = \lambda\mathbf{v}$$

移项：

$$(B - \lambda I)\mathbf{v} = 0$$

这里 I 是单位矩阵：

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

因为我们希望找到非零向量 $\mathbf{v}$ ，所以系数矩阵必须不可逆，即：

$$\boxed{\det(B - \lambda I) = 0}$$

这叫**特征方程** (Characteristic Equation)。

代入：

$$\det \begin{bmatrix} 3 - \lambda & 1 \\ 1 & 3 - \lambda \end{bmatrix} = 0$$

回顾一下，二阶行列式的计算是：

$$\det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc$$

因此：

$$(3 - \lambda)^2 - 1 = 0$$

整理得到：

$$\lambda^2 - 6\lambda + 8 = 0$$

$$(\lambda - 4)(\lambda - 2) = 0$$

所以：

$$\lambda_1 = 4, \quad \lambda_2 = 2$$

然后分别将 $\lambda = 4$ 和 $\lambda = 2$ 代入 $(B - \lambda I)\mathbf{v} = 0$ ，就能求出对应的特征向量。

四、什么是正交矩阵 Q?

现在回到我们真正想理解的公式。

首先，把两个特征向量进行归一化，也就是让它们的长度变成 1。

因为：

$$\left\| \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right\| = \sqrt{2}$$

所以第一个单位特征向量为：

$$\mathbf{q}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

第二个为：

$$\mathbf{q}_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

你可以验证它们相互垂直：

$$\mathbf{q}_1^T \mathbf{q}_2 = \frac{1}{2}(-1 + 1) = 0$$

现在把它们作为矩阵的两列：

$$Q = \begin{bmatrix} | & | \\ \mathbf{q}_1 & \mathbf{q}_2 \\ | & | \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$$

这种列向量相互垂直、长度均为 1 的方阵，叫作**正交矩阵**（Orthogonal Matrix）。

正交矩阵具有一个非常重要的性质：

$$Q^T Q = Q Q^T = I$$

因此：

$$Q^{-1} = Q^T$$

其中：

- Q^T ：转置，即交换行和列。
- Q^{-1} ：逆矩阵，即能够撤销原矩阵变换的矩阵。

这两者通常并不相同，只有正交矩阵等特殊情况，才可以直接用转置代替逆矩阵。

几何上，我们这个 Q 对应平面内的逆时针旋转 45° ，而 Q^T 对应顺时针旋转 45° 。

五、终于推导 $H = Q \Lambda Q^T$

为了和之前的 Hessian 公式保持一致，现在将 B 重新记作 H 。

我们知道：

$$H \mathbf{q}_1 = 4 \mathbf{q}_1$$

$$H \mathbf{q}_2 = 2 \mathbf{q}_2$$

把两条等式合并：

$$H \begin{bmatrix} | & | \\ \mathbf{q}_1 & \mathbf{q}_2 \\ | & | \end{bmatrix} = \begin{bmatrix} | & | \\ 4\mathbf{q}_1 & 2\mathbf{q}_2 \\ | & | \end{bmatrix}$$

右侧也可以写成矩阵乘法：

$$HQ = Q \begin{bmatrix} 4 & 0 \\ 0 & 2 \end{bmatrix}$$

令：

$$\Lambda = \begin{bmatrix} 4 & 0 \\ 0 & 2 \end{bmatrix}$$

那么：

$$\boxed{HQ = Q\Lambda}$$

在等式两边右乘 Q^T ：

$$HQQ^T = Q\Lambda Q^T$$

由于：

$$QQ^T = I$$

因此：

$$\boxed{H = Q\Lambda Q^T}$$

这就是特征值分解的来源！

这里的三个矩阵分别代表：

矩阵	数学含义
Q	各列是单位特征向量
Λ	对角线上排列对应的特征值
Q^T	Q 的逆矩阵

注意：这个 Q 只是线性代数中的符号，与 Transformer Attention 中的 Query 矩阵 Q 没有特殊关系。

六、从几何上理解这三个矩阵的作用

这可能是最值得理解的部分。

$$H\mathbf{x} = Q\Lambda Q^T\mathbf{x}$$

因为矩阵乘法从右往左作用，所以整个变换可以拆成三步：

第一步： $Q^T\mathbf{x}$

将向量转换到由特征向量构成的新坐标系

第二步： $\Lambda(Q^T\mathbf{x})$

沿两个新坐标轴，分别放大 4 倍和 2 倍

第三步： $Q\Lambda Q^T\mathbf{x}$

将变换后的向量重新表示在原来的坐标系中

换句话说，原来复杂的矩阵：

$$H = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}$$

在特征向量构成的新坐标系中，就变成了：

$$\Lambda = \begin{bmatrix} 4 & 0 \\ 0 & 2 \end{bmatrix}$$

对角化的本质，就是找到一个特殊的坐标系，让原本复杂的线性变换变成沿各坐标轴独立的缩放。

七、为什么 Hessian 一定可以这样分解？

严格来说，不是所有矩阵都可以进行这种正交对角化。

但实对称矩阵可以。这是实对称矩阵谱定理：

每个实对称矩阵都存在一组正交归一的特征向量，它们构成整个空间的一组基。因此，该矩阵可以被正交对角化。

Hessian 矩阵为：

$$H = \begin{bmatrix} f_{xx} & f_{xy} \\ f_{yx} & f_{yy} \end{bmatrix}$$

只要函数二阶偏导连续，根据混合偏导数相等定理（Clairaut's Theorem）：

$$f_{xy} = f_{yx}$$

于是：

$$H^T = H$$

因此 Hessian 是实对称矩阵，可以使用谱定理得到：

$$\boxed{H = Q\Lambda Q^T}$$

顺便补充一个小证明：为什么不同特征值对应的特征向量垂直？

设：

$$H\mathbf{v}_1 = \lambda_1\mathbf{v}_1, \quad H\mathbf{v}_2 = \lambda_2\mathbf{v}_2$$

因为 $H^T = H$ ，所以：

$$\mathbf{v}_1^T H\mathbf{v}_2 = (H\mathbf{v}_1)^T \mathbf{v}_2$$

代入特征值关系：

$$\lambda_2 \mathbf{v}_1^T \mathbf{v}_2 = \lambda_1 \mathbf{v}_1^T \mathbf{v}_2$$

于是：

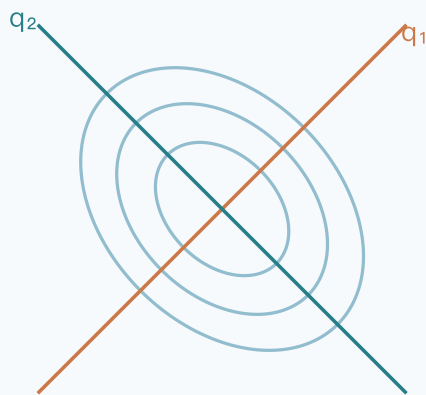
$$(\lambda_2 - \lambda_1) \mathbf{v}_1^T \mathbf{v}_2 = 0$$

只要 $\lambda_1 \neq \lambda_2$ ，就得到：

$$\boxed{\mathbf{v}_1^T \mathbf{v}_2 = 0}$$

即两个特征向量相互垂直。对于重复特征值的情况，也可以在对应的特征子空间内选取正交归一基。

图 2 · 特征基把交叉项化成独立平方项



$$H = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}$$

$$\lambda_1 = 4, \lambda_2 = 2$$

$$f = 2u_1^2 + u_2^2$$

短轴方向的二阶变化率更大。

Q^T 取坐标； Q 将坐标变回。

第 13 章贯穿例子的精确等高线，常值为 1、2.5、4.5。两条彩色直线是单位特征向量方向；几何图形采用等比例坐标。

补充：零特征值与顺序

当 $\lambda = 0$ 时，非零特征向量被映到零向量；此时不能说像向量仍有原方向。统一的陈述是其一维张成空间在变换下不变。

$Q^T x$ 是旧坐标到特征基坐标， Qu 是特征基坐标到旧坐标。列向量与对角元素必须配对，同时交换它们不改变 $Q\Lambda Q^T$ 。若有重特征值，同一特征子空间中的正交基并不唯一。

第 6 章：正交矩阵为什么能选出，以及 Gram-Schmidt 如何操作

[返回目录](#)

关键在于：并不是我们随便构造一个 Q ，它就一定是正交矩阵，而是因为 A 是实对称矩阵，我们总能选出一组正交归一的特征向量来构造 Q 。

这里需要证明两件事：

1. 不同特征值对应的特征向量相互正交。
2. 相同特征值对应的特征向量，也可以选成相互正交的。

一、先理解 Q 为什么与正交有关

假设：

$$Q = \begin{bmatrix} | & | \\ \mathbf{q}_1 & \mathbf{q}_2 \\ | & | \end{bmatrix}$$

矩阵 Q 的两列分别是两个特征向量。

那么：

$$Q^T Q = \begin{bmatrix} \mathbf{q}_1^T \\ \mathbf{q}_2^T \end{bmatrix} [\mathbf{q}_1 \quad \mathbf{q}_2]$$

展开得到：

$$Q^T Q = \begin{bmatrix} \mathbf{q}_1^T \mathbf{q}_1 & \mathbf{q}_1^T \mathbf{q}_2 \\ \mathbf{q}_2^T \mathbf{q}_1 & \mathbf{q}_2^T \mathbf{q}_2 \end{bmatrix}$$

注意，每一个元素都是两个向量的内积。

如果两个向量长度均为 1，并且互相垂直：

$$\mathbf{q}_1^T \mathbf{q}_1 = 1, \quad \mathbf{q}_2^T \mathbf{q}_2 = 1$$

$$\mathbf{q}_1^T \mathbf{q}_2 = \mathbf{q}_2^T \mathbf{q}_1 = 0$$

于是：

$$Q^T Q = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I$$

这就是正交矩阵的定义。

所以问题变成了：为什么实对称矩阵的特征向量能够满足这些条件？

二、证明不同特征值对应的特征向量正交

假设：

$$A^T = A$$

有两个不同的特征值 $\lambda_1 \neq \lambda_2$ ，对应特征向量：

$$A\mathbf{q}_1 = \lambda_1\mathbf{q}_1$$

$$A\mathbf{q}_2 = \lambda_2\mathbf{q}_2$$

考虑这个标量：

$$\mathbf{q}_1^T A \mathbf{q}_2$$

我们可以用两种方式计算它。

第一种：先计算 $A\mathbf{q}_2$

$$\mathbf{q}_1^T A \mathbf{q}_2 = \lambda_2 \mathbf{q}_1^T \mathbf{q}_2$$

第二种：利用矩阵的对称性

由于 $A^T = A$ ：

$$\mathbf{q}_1^T A \mathbf{q}_2 = (A\mathbf{q}_1)^T \mathbf{q}_2$$

所以：

$$\mathbf{q}_1^T A \mathbf{q}_2 = \lambda_1 \mathbf{q}_1^T \mathbf{q}_2$$

两种计算结果必然相等：

$$\lambda_2 \mathbf{q}_1^T \mathbf{q}_2 = \lambda_1 \mathbf{q}_1^T \mathbf{q}_2$$

移项：

$$(\lambda_2 - \lambda_1) \mathbf{q}_1^T \mathbf{q}_2 = 0$$

因为：

$$\lambda_2 - \lambda_1 \neq 0$$

所以只能是：

$$\boxed{\mathbf{q}_1^T \mathbf{q}_2 = 0}$$

证毕：实对称矩阵不同特征值对应的特征向量一定相互正交。

三、如果两个特征值相同呢？

这是一个容易忽略的细节。

例如单位矩阵：

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

因为 $A\mathbf{x} = \mathbf{x}$ ，所以任何非零向量都是特征值为 1 的特征向量。

例如：

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

显然：

$$\mathbf{v}_1^T \mathbf{v}_2 = 1 \neq 0$$

所以，相同特征值对应的特征向量，不一定天然正交。

但是，我们可以选择另一组：

$$\mathbf{q}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \mathbf{q}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

它们同样都是特征向量，而且相互正交。

对于高维情况，可以通过 Gram-Schmidt 正交化，在同一特征值对应的特征子空间中选出正交归一基。

之所以这样操作不会破坏特征向量性质，是因为相同特征值对应的特征向量的线性组合，仍然是该特征值的特征向量。

Gram-Schmidt 正交化 (Gram-Schmidt Orthogonalization) 的核心思想很简单：

从第二个向量中，减去它在第一个向量方向上的分量，剩下的向量就与第一个向量垂直了。

我们先用二维向量推导，再推广到任意维度，最后解释为什么它不会破坏特征向量的性质。

一、先看最简单的二维例子

假设有两个线性无关、但不垂直的向量：

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

显然，它们的内积：

$$\mathbf{v}_1^T \mathbf{v}_2 = 1 \neq 0$$

所以它们不正交。

我们的目标是找到两个新向量 $\mathbf{q}_1, \mathbf{q}_2$ ，满足：

$$\mathbf{q}_1^T \mathbf{q}_2 = 0, \quad \|\mathbf{q}_1\| = \|\mathbf{q}_2\| = 1$$

第一步：保留第一个向量，并归一化

定义：

$$\mathbf{q}_1 = \frac{\mathbf{v}_1}{\|\mathbf{v}_1\|}$$

本例中 $\|\mathbf{v}_1\| = 1$ ，因此：

$$\mathbf{q}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

第二步：计算第二个向量在第一个方向上的投影

向量投影的公式是：

$$\text{proj}_{\mathbf{q}_1} \mathbf{v}_2 = (\mathbf{q}_1^T \mathbf{v}_2) \mathbf{q}_1$$

注意，这里 $\mathbf{q}_1$ 已经是单位向量。

代入：

$$\mathbf{q}_1^T \mathbf{v}_2 = [1 \quad 0] \begin{bmatrix} 1 \\ 1 \end{bmatrix} = 1$$

所以投影为：

$$\text{proj}_{\mathbf{q}_1} \mathbf{v}_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

第三步：减去投影

$$\mathbf{u}_2 = \mathbf{v}_2 - (\mathbf{q}_1^T \mathbf{v}_2) \mathbf{q}_1$$

代入：

$$\mathbf{u}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} - \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

检验：

$$\mathbf{q}_1^T \mathbf{u}_2 = [1 \quad 0] \begin{bmatrix} 0 \\ 1 \end{bmatrix} = 0$$

所以它们已经正交。

最后再归一化：

$$\mathbf{q}_2 = \frac{\mathbf{u}_2}{\|\mathbf{u}_2\|} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

最终：

$$Q = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

我们成功把原本不正交的两个向量，转换成了一组正交归一向量。

二、为什么减去投影后，一定会垂直？

这是 Gram–Schmidt 最重要的数学原理。

假设 $\mathbf{q}_1$ 是单位向量，我们想从 $\mathbf{v}_2$ 中减去某个倍数的 $\mathbf{q}_1$ ：

$$\mathbf{u}_2 = \mathbf{v}_2 - c\mathbf{q}_1$$

现在要求 $\mathbf{u}_2$ 与 $\mathbf{q}_1$ 垂直，也就是：

$$\mathbf{q}_1^T \mathbf{u}_2 = 0$$

将表达式代入：

$$\mathbf{q}_1^T (\mathbf{v}_2 - c\mathbf{q}_1) = 0$$

展开：

$$\mathbf{q}_1^T \mathbf{v}_2 - c\mathbf{q}_1^T \mathbf{q}_1 = 0$$

因为 $\mathbf{q}_1$ 是单位向量：

$$\mathbf{q}_1^T \mathbf{q}_1 = 1$$

所以：

$$c = \mathbf{q}_1^T \mathbf{v}_2$$

这就严格推导出了投影公式。

因此，减去投影并不是一种经验技巧，而是直接根据「两个向量内积等于零」这个正交条件求出来的。

三、如果有三个向量怎么办？

假设有三个线性无关的向量：

$$\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$$

处理流程如下。

首先：

$$\mathbf{q}_1 = \frac{\mathbf{v}_1}{\|\mathbf{v}_1\|}$$

然后：

$$\mathbf{u}_2 = \mathbf{v}_2 - (\mathbf{q}_1^T \mathbf{v}_2) \mathbf{q}_1$$

$$\mathbf{q}_2 = \frac{\mathbf{u}_2}{\|\mathbf{u}_2\|}$$

到了第三个向量，关键来了。

$\mathbf{v}_3$ 既要去掉沿 $\mathbf{q}_1$ 的分量，也要去掉沿 $\mathbf{q}_2$ 的分量。

因此：

$$\mathbf{u}_3 = \mathbf{v}_3 - (\mathbf{q}_1^T \mathbf{v}_3) \mathbf{q}_1 - (\mathbf{q}_2^T \mathbf{v}_3) \mathbf{q}_2$$

再归一化：

$$\mathbf{q}_3 = \frac{\mathbf{u}_3}{\|\mathbf{u}_3\|}$$

这时候：

$$\mathbf{q}_1^T \mathbf{q}_3 = 0$$

$$\mathbf{q}_2^T \mathbf{q}_3 = 0$$

而前面已经有：

$$\mathbf{q}_1^T \mathbf{q}_2 = 0$$

所以三个向量两两正交。

推广到第 k 个向量：

$$\mathbf{u}_k = \mathbf{v}_k - \sum_{j=1}^{k-1} (\mathbf{q}_j^T \mathbf{v}_k) \mathbf{q}_j$$

$$\mathbf{q}_k = \frac{\mathbf{u}_k}{\|\mathbf{u}_k\|}$$

前提是原始向量线性无关，这样每一步的 $\mathbf{u}_k$ 都不是零向量。

四、为什么对特征向量做这个操作，不会破坏特征向量性质？

这就回到了你上一条关于实对称矩阵的问题。

假设：

$$A\mathbf{v}_1 = \lambda\mathbf{v}_1$$

$$A\mathbf{v}_2 = \lambda\mathbf{v}_2$$

注意：这两个特征向量对应同一个特征值 λ 。

通过 Gram-Schmidt 得到：

$$\mathbf{u}_2 = \mathbf{v}_2 - c\mathbf{v}_1$$

那么：

$$A\mathbf{u}_2 = A(\mathbf{v}_2 - c\mathbf{v}_1)$$

利用矩阵乘法的线性性质：

$$= A\mathbf{v}_2 - cA\mathbf{v}_1$$

代入特征值关系：

$$= \lambda\mathbf{v}_2 - c\lambda\mathbf{v}_1$$

提出 λ ：

$$= \lambda(\mathbf{v}_2 - c\mathbf{v}_1)$$

因此：

$$\boxed{A\mathbf{u}_2 = \lambda\mathbf{u}_2}$$

这证明了：对同一特征值对应的特征向量进行线性组合，得到的非零向量仍然是同一个特征值对应的特征向量。

归一化也只是乘一个非零标量，因此同样不会破坏特征向量性质。

但如果两个特征向量对应不同特征值，它们的线性组合通常就不再是特征向量。不过实对称矩阵的不同特征值对应的特征向量本来就正交，因此也不需要跨特征值进行 Gram-Schmidt。

五、最后把它和 Q 联系起来

对于实对称矩阵 A ，我们可以在各个特征子空间中通过 Gram-Schmidt 选取正交归一基，再将它们组合起来：

$$Q = [\mathbf{q}_1 \quad \mathbf{q}_2 \quad \cdots \quad \mathbf{q}_n]$$

就有：

$$Q^T Q = I$$

并且：

$$AQ = Q\Lambda$$

最终：

$$A = Q\Lambda Q^T$$

所以，你可以把 Gram–Schmidt 理解为一个非常具体的操作：

每加入一个新向量，就把它与之前所有正交方向重叠的分量全部减掉，最后把剩余向量的长度调整为 1。

它不仅用于矩阵特征向量的正交化，也是线性代数中 QR 分解、最小二乘计算等方法的基础。

补充：非零性、张成空间和循环论证

第 k 步若 $\mathbf{u}_k = \mathbf{0}$ ，则 $\mathbf{v}_k$ 属于前 $k - 1$ 个 $\mathbf{q}_j$ 的张成空间。归纳可知它等于前 $k - 1$ 个 $\mathbf{v}_j$ 的张成空间，与原向量组线性无关矛盾。因此每一步可归一化，并保持前 k 个向量的张成空间。

但 Gram–Schmidt 不能凭空补出缺失的特征向量。不同特征值的正交性，加上同一特征空间内的正交化，并没有单独证明特征向量总数达到 n 。这项存在性要由第 9—10 章的谱定理证明完成，不能先假定可对角化再拿它证明可对角化。

例如对非对称矩阵 $\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ ，唯一特征值为 1，特征空间仅一维；再熟练的正交化也无法把它变成二维特征基。

第 7 章：紧致性、闭且有界和收敛子序列

[返回目录](#)

对，你现在问的是数学分析中非常重要的 Heine–Borel 定理（海涅–博雷尔定理）：

$$K \subseteq \mathbb{R}^n \text{ 是紧集} \iff K \text{ 是闭集且有界}$$

这个定理并不是紧集的定义，而是一个需要证明的结论。并且要特别注意：这个等价关系依赖于有限维欧氏空间的性质，在一般的无限维空间中不成立。

我们从紧集的原始定义开始，证明两个方向。

一、紧集的原始定义是什么？

紧集（Compact Set）最经典的定义来自开覆盖（Open Cover）。

假设有一个集合：

$$K \subseteq \mathbb{R}^n$$

存在一族开集 $\{U_\alpha\}_{\alpha \in I}$ ，满足：

$$K \subseteq \bigcup_{\alpha \in I} U_\alpha$$

这意味着这些开集加起来，可以完全覆盖 K 。

如果无论怎样用开集覆盖 K ，都能从中选出有限个开集，仍然覆盖整个 K ，那么就称 K 是紧集。

数学表达为：

$$K \subseteq \bigcup_{\alpha \in I} U_{\alpha} \implies \exists \alpha_1, \dots, \alpha_m, \quad K \subseteq \bigcup_{i=1}^m U_{\alpha_i}$$

这就是**有限子覆盖性质**（Finite Subcover Property）。

注意，不是要求每种覆盖本来就只有有限个开集，而是即使提供无限多个开集，也总能从里面挑出有限个来完成覆盖。

二、先证明：紧集一定有界

这个方向比较简单。

假设 K 是紧集。

我们考虑以原点为圆心、半径不断增大的开球：

$$B(0, 1), B(0, 2), B(0, 3), \dots$$

它们的并集能够覆盖整个 $\mathbb{R}^n$ ，当然也能覆盖 K ：

$$K \subseteq \bigcup_{m=1}^{\infty} B(0, m)$$

由于 K 是紧集，根据有限子覆盖性质，可以从这些开球中选出有限个，覆盖整个 K 。

假设选中的最大半径为 M 。

由于这些球都是同心的：

$$K \subseteq B(0, M)$$

这说明 K 中所有点到原点的距离都小于 M 。

因此：

$$K \text{ 紧致} \implies K \text{ 有界}$$

三、再证明：紧集一定是闭集

回忆闭集的一个等价定义：

集合 K 是闭集，当且仅当它的补集 $\mathbb{R}^n \setminus K$ 是开集。

因此，我们只需要证明：对于任意一个不属于 K 的点 $\mathbf{p}$ ，都能在它周围找到一个完全不与 K 相交的小球。

取：

$$\mathbf{p} \notin K$$

对于 K 内每个点 $\mathbf{x}$ ，构造开球：

$$B\left(\mathbf{x}, \frac{\|\mathbf{x} - \mathbf{p}\|}{3}\right)$$

由于 $\mathbf{p} \neq \mathbf{x}$ ，每个球的半径都大于零。

这些球显然覆盖整个 K ，因为每个点都是其对应球的球心。

由紧致性，存在有限个球：

$$B(\mathbf{x}_1, r_1), \dots, B(\mathbf{x}_m, r_m)$$

仍然覆盖 K ，其中：

$$r_i = \frac{\|\mathbf{x}_i - \mathbf{p}\|}{3}$$

令：

$$\delta = \min(r_1, \dots, r_m) > 0$$

现在，对于任意 $\mathbf{z} \in K$ ，它必然属于某个选中的球，例如 $B(\mathbf{x}_i, r_i)$ 。

根据三角不等式：

$$\|\mathbf{z} - \mathbf{p}\| \geq \|\mathbf{x}_i - \mathbf{p}\| - \|\mathbf{z} - \mathbf{x}_i\|$$

因为：

$$\|\mathbf{x}_i - \mathbf{p}\| = 3r_i, \quad \|\mathbf{z} - \mathbf{x}_i\| < r_i$$

所以：

$$\|\mathbf{z} - \mathbf{p}\| > 2r_i \geq 2\delta$$

因此，开球 $B(\mathbf{p}, \delta)$ 完全不与 K 相交。

这意味着 K 的补集是开集，所以：

$$K \text{ 紧致} \implies K \text{ 闭}$$

至此，我们已经证明：

$$\text{紧致} \implies \text{闭且有界}$$

四、最关键的一步：为什么闭且有界一定紧致？

这才是 Heine–Borel 定理真正需要用到欧氏空间特殊性质的地方。

我们要证明：

$$K \subseteq \mathbb{R}^n \text{ 闭且有界} \implies K \text{ 紧致}$$

也就是证明：对 K 的任意开覆盖，一定存在有限子覆盖。

这里使用一个非常巧妙的方法：不断将空间分割成越来越小的立方体，并使用反证法。

第一步：假设结论不成立

假设 K 是闭且有界的，但它不是紧集。

那么存在一种开覆盖：

$$K \subseteq \bigcup_{\alpha \in I} U_{\alpha}$$

无论选多少个有限的 U_{α} ，都无法覆盖整个 K 。

由于 K 有界，我们可以找到一个足够大的闭立方体 C_0 ，使得：

$$K \subseteq C_0$$

第二步：不断细分立方体

以二维为例，我们先把正方形分成 4 个小正方形。

由于 K 不能被有限个开集覆盖，至少有一个小正方形与 K 的交集，也不能被有限个开集覆盖。

否则，如果四个小正方形中的部分 K 都可以被有限覆盖，那么把这四组有限覆盖合并起来，就得到整个 K 的有限覆盖，与假设矛盾。

于是我们选择这个有问题的小正方形，继续分割。

最终得到一系列嵌套的闭立方体：

$$C_0 \supseteq C_1 \supseteq C_2 \supseteq \cdots$$

并且每个 $K \cap C_m$ 都无法被有限个开集覆盖。

同时，立方体直径不断缩小：

$$\text{diam}(C_m) \rightarrow 0$$

第三步：证明这些立方体最终收敛到一个点

在每个 $K \cap C_m$ 中选择一点 $\mathbf{x}_m$ 。

由于这些立方体不断嵌套，对于任意 $j, k \geq N$ ，都有：

$$\mathbf{x}_j, \mathbf{x}_k \in C_N$$

所以：

$$\|\mathbf{x}_j - \mathbf{x}_k\| \leq \text{diam}(C_N)$$

而立方体直径趋于零，因此序列 $\{\mathbf{x}_m\}$ 是柯西序列（Cauchy Sequence）。

欧氏空间 $\mathbb{R}^n$ 是完备的，所以存在极限：

$$\mathbf{x}_m \rightarrow \mathbf{x}^*$$

又因为 K 是闭集，所有 $\mathbf{x}_m \in K$ ，因此：

$$\mathbf{x}^* \in K$$

而且，由于立方体嵌套且闭合， $\mathbf{x}^*$ 属于每一个 C_m 。

第四步：利用开集性质得到矛盾

这是整个证明最精彩的地方。

因为我们的开集族覆盖 K ，所以必然有某一个开集 U_β 包含 $\mathbf{x}^*$ ：

$$\mathbf{x}^* \in U_\beta$$

既然 U_β 是开集，那么存在一个 $\varepsilon > 0$ ，使得：

$$B(\mathbf{x}^*, \varepsilon) \subseteq U_\beta$$

但是，立方体 C_m 的直径正在趋近零，而且每个都包含 $\mathbf{x}^*$ 。

因此，当 m 足够大时：

$$C_m \subseteq B(\mathbf{x}^*, \varepsilon)$$

于是：

$$\boxed{K \cap C_m \subseteq U_\beta}$$

也就是说，仅仅一个开集 U_β ，就已经能够覆盖这个立方体内的全部 K ！

但我们在构造立方体时，要求每个 $K \cap C_m$ 都不能被有限个开集覆盖。

矛盾！

因此，原来的假设不成立。

$$\boxed{K \text{ 闭且有界} \implies K \text{ 紧致}}$$

Heine–Borel 定理证毕。

五、为什么又可以用收敛子序列来理解紧致性？

现在回到你最开始关心的那个问题：

为什么闭且有界集合中的任意序列，都存在收敛到集合内部的子序列？

实际上，它与刚才的证明有相同的数学基础。

这里使用 **Bolzano–Weierstrass 定理**（博尔查诺–魏尔斯特拉斯定理）：

$$\mathbb{R}^n \text{ 中任意有界序列都有收敛子序列}$$

例如，假设有一个二维有界序列：

$$\mathbf{x}_n = (a_n, b_n)$$

由于整个序列有界，坐标序列 $\{a_n\}$ 也有界。

根据一维 Bolzano–Weierstrass 定理，可以从中选出收敛子序列：

$$a_{n_k} \rightarrow a$$

然后只考虑对应的 $\{b_{n_k}\}$ 。

它同样有界，因此可以继续选出一个子序列：

$$b_{n_{k_j}} \rightarrow b$$

由于第二次选择的仍然是第一次收敛子序列的一部分，所以对应的 $a_{n_{k_j}}$ 也收敛到 a 。

于是：

$$\mathbf{x}_{n_{k_j}} \rightarrow (a, b)$$

推广到有限维，只需要对每个坐标依次提取子序列。

如果原来的序列全部位于闭集 K ，那么它的极限也必然属于 K 。

因此：

K 闭且有界 $\implies$ 任意序列都有收敛到 K 内的子序列

这种性质叫作**序列紧致性** (Sequential Compactness)。

对于欧氏空间：

开覆盖紧致 $\iff$ 序列紧致 $\iff$ 闭且有界

前面通过嵌套立方体的证明，展示了闭且有界为什么能推出开覆盖紧致；而序列紧致是理解这一结论的另一种方式。

六、为什么强调有限维？

这里还有一个很重要的数学事实。

在无限维空间，例如平方可和序列组成的 Hilbert 空间 ℓ^2 中，考虑单位向量：

$$\mathbf{e}_1 = (1, 0, 0, \dots)$$

$$\mathbf{e}_2 = (0, 1, 0, \dots)$$

$$\mathbf{e}_3 = (0, 0, 1, \dots)$$

它们都在闭单位球内，长度都等于 1。

但对于任意 $i \neq j$ ：

$$\|\mathbf{e}_i - \mathbf{e}_j\| = \sqrt{2}$$

所以这些向量之间始终保持固定距离，根本无法找到收敛子序列。

因此，无限维空间的闭单位球虽然闭且有界，却并不紧致（相对于范数拓扑）。

这意味着：有限维不仅仅是证明中的技术限制，而是定理成立的重要条件。

对，开覆盖（Open Cover）完全不要求由小球组成，只要求每一个覆盖集合都是开集（Open Set）。

例如，在二维欧氏空间 $\mathbb{R}^2$ 中，以下都可以作为开覆盖的组成部分：

- 开圆盘： $\{(x, y) : x^2 + y^2 < 1\}$
- 开矩形： $\{(x, y) : 0 < x < 2, 0 < y < 3\}$
- 开半平面： $\{(x, y) : x > 0\}$
- 不规则形状的开区域，甚至多个不相连开区域的并集

只要这些开集的并集包含目标集合 K ，它们就构成 K 的一个开覆盖：

$$K \subseteq \bigcup_{\alpha \in I} U_{\alpha}$$

那为什么证明中经常使用小球？

因为在欧氏空间中，开集的定义本身就可以通过开球来描述。

如果 U 是开集，那么对其中任意一点 $\mathbf{x} \in U$ ，一定存在 $\varepsilon > 0$ ，满足：

$$B(\mathbf{x}, \varepsilon) \subseteq U$$

这里的开球是：

$$B(\mathbf{x}, \varepsilon) = \{\mathbf{y} : \|\mathbf{y} - \mathbf{x}\| < \varepsilon\}$$

这意味着，不管开集 U 的形状多么复杂，只要一点位于它内部，就能在这个点周围找到一个完全包含在 U 里面的小球。

所以，在刚才 Heine—Borel 定理的证明中，我们首先知道：

$$\mathbf{x}^* \in U_\beta$$

因为 U_β 是开集，所以可以找到：

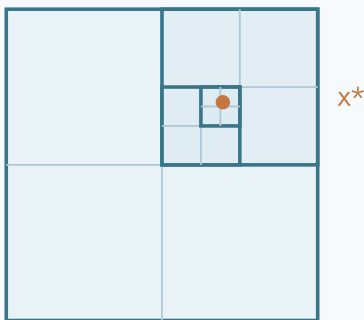
$$B(\mathbf{x}^*, \varepsilon) \subseteq U_\beta$$

而随着嵌套立方体越来越小，最终整个立方体都会落入这个小球，也就落入 U_β ，从而产生矛盾。

注意这里的逻辑：不是说开覆盖必须由小球组成，而是任何开集内部都可以找到适当的小球。

还有一个更深一层的结论：在欧氏空间中，所有开球构成拓扑的一个**基 (Basis)**。因此，讨论开覆盖时，实际上也可以只检验由开球构成的覆盖是否存在有限子覆盖；如果这个性质对所有开球覆盖成立，那么它对任意开集覆盖也成立。

图 3 · “不能有限覆盖”的区域被迫缩向一点



每次只从同一原开覆盖挑选集合。

直径 $\rightarrow 0$ ；完备性保证极限存在。

闭性保证 $x^* \in K$ 。

开性让足够小的 C_m 落进一个 U 。

于是只用一个原开集就能覆盖，矛盾。

二维示意只展示几次细分，不指定真实 K 的形状； n 维每步分成 2^n 个闭立方体。极限点落入某个原开集后，整个小立方体最终都落入它。

补充证明一：从实数完备性到一维 Bolzano–Weierstrass

设有界序列 (x_j) 落在闭区间 I_0 中。把区间对半分，至少一半包含无穷多个序列项的指标；选取它为 I_1 。逐次重复，得到闭区间 I_k ，长度为 $2^{-k}|I_0|$ ，每个都含无穷多个指标的序列项。

可依次选严格增加的指标 j_k ，使 $x_{j_k} \in I_k$ 。当 $r, s \geq k$ 时，两项均落在 I_k ，所以 $|x_{j_r} - x_{j_s}| \leq |I_k|$ 。它是柯西序列，由实数完备性收敛。这就补上了原对话中直接引用的一维定理；有限维坐标提取在上文已经说明。

补充证明二：序列紧致也推出闭且有界

若 K 序列紧致却无界，可选 $x_j \in K$ 满足 $\|x_j\| > j$ 。任何子序列都不可能收敛，矛盾。

若 K 不闭，则存在 $p \notin K$ ，每个 $B(p, 1/j)$ 都与 K 相交。选择 $x_j \in K \cap B(p, 1/j)$ ，则 $x_j \rightarrow p$ 。序列紧致给出一个极限在 K 中的子序列；极限唯一性迫使该极限就是 p ，矛盾。因此 K 闭。

把这两个结论与上文的 Heine–Borel 证明、有限维有界序列提取结合，才得到完整的三个等价条件：开覆盖紧致、序列紧致、闭且有界。这里没有把一个尚未证明的等价关系当作定义偷用。

空集也紧致、闭且有界、序列紧致；极值定理另需非空。闭集证明中对非空 K 选有限球，空集情形单独直接成立。

补充：开球覆盖为什么足够

若只知道所有开球覆盖都有有限子覆盖，面对任意开覆盖 $\{U_\alpha\}$ ，为每个 $x \in K$ 选 $U_{\alpha(x)} \ni x$ 及 $r_x > 0$ 使 $B(x, r_x) \subset U_{\alpha(x)}$ 。这些小球构成覆盖；取有限个后，相应有限个原开集也覆盖 K 。因此“可以化成开球来证明”与“定义只允许开球”是两个不同的陈述。

在无限维 ℓ^2 的例子中，如果不引用一般度量空间的紧致与序列紧致等价，也能直接给出失败的开覆盖：对闭单位球内每个点 x 取 $B(x, 1/2)$ 。一个这样的球至多包含一个标准基向量 e_j ，因为两点在同球内的距离小于 1，而 $\|e_i - e_j\| = \sqrt{2}$ 。任何有限个球都不能覆

盖全部 e_j ，所以闭单位球不紧致。

第 8 章：连续像为什么紧致，连续函数为什么能取到极值

[返回目录](#)

你问的是为什么：

紧集在连续映射下的像仍然是紧集

这是数学分析中的连续映射保持紧致性定理 (Continuous Image of a Compact Set Is Compact)。

这个定理其实可以从两个更基础的性质直接证明：

1. **紧致性**：在有限维欧氏空间中，紧集里的任意序列，都能找到一个收敛到该集合内部某点的子序列。
2. **连续性**：如果 $x_n \rightarrow x$ ，那么 $f(x_n) \rightarrow f(x)$ 。

把这两个性质结合起来，就能完成证明。

一、先证明连续映射后的集合仍然紧致

设：

$$f: K \rightarrow \mathbb{R}$$

其中 $K \subseteq \mathbb{R}^n$ 是紧集， f 是连续函数。

我们希望证明：

$$f(K) = \{f(x) : x \in K\}$$

仍然是紧集。

第一步：从 $f(K)$ 中任意选择一个序列。

假设：

$$y_1, y_2, y_3, \dots \in f(K)$$

由于每个 y_i 都是某个 $x_i \in K$ 的函数值，所以可以找到：

$$y_i = f(x_i)$$

于是我们就有了定义域中的序列：

$$x_1, x_2, x_3, \dots \in K$$

第二步：利用 K 的紧致性。

因为 K 是紧集，所以一定可以从 $\{x_n\}$ 中选出一个收敛子序列：

$$\boxed{x_{n_k} \rightarrow x^* \in K}$$

注意，这里不仅保证子序列收敛，还保证极限 x^* 仍然属于 K 。

第三步：利用函数的连续性。

由于 f 连续，因此：

$$x_{n_k} \rightarrow x^* \implies f(x_{n_k}) \rightarrow f(x^*)$$

又因为：

$$y_{n_k} = f(x_{n_k})$$

所以：

$$y_{n_k} \rightarrow f(x^*)$$

并且：

$$f(x^*) \in f(K)$$

我们证明了什么？

对于 $f(K)$ 中任意一个序列，都能够找到一个收敛子序列，而且极限仍然属于 $f(K)$ 。

这恰好就是序列紧致性（Sequential Compactness）的定义。

在欧氏空间中，序列紧致与紧致等价，因此：

$$K \text{ 紧致}, f \text{ 连续} \implies f(K) \text{ 紧致}$$

证毕。

二、为什么这又能保证函数一定取到最大值？

回到实数轴上的情况。

我们已经证明：

$$f(K) \subseteq \mathbb{R}$$

是紧集。

根据 Heine–Borel 定理，它必然有界且闭合。

有界意味着：

$$M = \sup f(K)$$

存在且是有限实数。

这里的 **sup** 是上确界 (supremum)，也就是函数值集合的最小上界。

但要注意：

存在上确界，并不意味着一定能够取到上确界。

例如：

$$f(K) = (0, 1)$$

它的上确界为 1，但集合中没有 1。

现在关键来了。

根据上确界定义，一定可以找到一列函数值：

$$y_n \in f(K)$$

满足：

$$M - \frac{1}{n} < y_n \leq M$$

因此：

$$y_n \rightarrow M$$

而 $f(K)$ 是闭集，所以它包含其所有极限点。

于是：

$$\boxed{M \in f(K)}$$

这意味着存在某个 $x_{\max} \in K$ ，使得：

$$f(x_{\max}) = M = \max_{x \in K} f(x)$$

同理可以证明最小值一定存在。

这就得到了极值定理：

$$K \text{ 非空紧致, } f \text{ 连续} \implies f \text{ 在 } K \text{ 上取得最大值和最小值}$$

三、能不能更直接地证明极值定理？

其实可以，而且更容易记忆。

假设连续函数 f 在非空紧集 K 上有有限上确界 M 。

根据上确界定义，存在一系列点：

$$x_n \in K$$

使得：

$$f(x_n) \rightarrow M$$

因为 K 紧致，所以存在收敛子序列：

$$x_{n_k} \rightarrow x^* \in K$$

由于连续性：

$$f(x_{n_k}) \rightarrow f(x^*)$$

但原来已经知道：

$$f(x_{n_k}) \rightarrow M$$

根据极限的唯一性：

$$\boxed{f(x^*) = M}$$

这就证明最大值可以取到。

当然，上面先假设了 M 有限。这个前提也可以证明：如果 f 在 K 上无界，就能选择 $x_n \in K$ 使得 $|f(x_n)| > n$ ，但紧致性保证其中有收敛子序列，而连续性要求对应函数值收敛到有限值，产生矛盾。

因此证明是完整的。

补充：直接从开覆盖证明连续像紧致

设 $f: K \rightarrow Y$ 连续， K 紧致。若 $\{V_\alpha\}$ 是 $f(K)$ 的开覆盖，则 $\{f^{-1}(V_\alpha)\}$ 是 K 的相对开覆盖。紧致性给出有限个原像覆盖 K ，相应有限个 V_α 就覆盖 $f(K)$ 。相对开集是形如 $K \cap O$ 的集合，可用环境开集 O 转成前章的定义。

这个证明不需要 f 为单射，也不用求逆函数； $f^{-1}(V)$ 表示集合的原像。

补充：为什么球面满足所有条件

单位球面 $S^{n-1} = \{x \in \mathbb{R}^n : \|x\| = 1\}$ 在 $n \geq 1$ 时非空，且有界。若球面上 $x_j \rightarrow x$ ，由反三角不等式

$$|\|x_j\| - \|x\|| \leq \|x_j - x\| \rightarrow 0,$$

可知 $\|x\| = 1$ ，所以球面闭，从而紧致。二次型 $x \mapsto x^T A x$ 是有限项多项式，连续，于是能取到最大、最小值。

仅有闭集不够： $f(x) = x$ 在 $[0, \infty)$ 无最大值。仅有有界不够： $f(x) = x$ 在 $(0, 1)$ 无最大、最小值。即便定义域紧致，连续性也不能删去：在 $[0, 1]$ 上令 $f(x) = x$ ($x < 1$)，而 $f(1) = 0$ ，上确界为 1 却取不到。

第 9 章：谱定理的代数路线：先从特征方程找一个实方向

[返回目录](#)

确实。实对称矩阵谱定理中，通过单位球面上二次型的最大值来构造特征向量，属于一种相当巧妙的证明方法。第一次接触时，很难想到为什么要突然研究：

$$f(\mathbf{x}) = \mathbf{x}^T A \mathbf{x}, \quad \|\mathbf{x}\| = 1$$

实际上，完全可以不用这个构造，而是从特征值的定义直接证明实对称矩阵存在实特征向量。我觉得这个证明更符合你目前对线性代数的理解。

一、一种更直接的证明：从特征方程出发

设 A 是 $n \times n$ 实对称矩阵：

$$A^T = A$$

我们知道特征值满足：

$$\det(A - \lambda I) = 0$$

左边是一个关于 λ 的 n 次多项式。

根据代数基本定理 (Fundamental Theorem of Algebra)，任意非常数复系数多项式至少有一个复数根。

因此，一定存在：

$$\lambda \in \mathbb{C}$$

使得：

$$\det(A - \lambda I) = 0$$

于是存在非零复向量 $\mathbf{v}$ ，满足：

$$\boxed{A\mathbf{v} = \lambda\mathbf{v}}$$

到目前为止，我们还不知道 λ 是否为实数。

关键：利用 A 的对称性

在等式左边乘以 $\mathbf{v}^*$ ：

$$\mathbf{v}^* A \mathbf{v} = \lambda \mathbf{v}^* \mathbf{v}$$

这里 $\mathbf{v}^*$ 表示共轭转置 (Conjugate Transpose)。

于是：

$$\boxed{\lambda = \frac{\mathbf{v}^* A \mathbf{v}}{\mathbf{v}^* \mathbf{v}}}$$

注意：

$$\mathbf{v}^* \mathbf{v} = \sum_i |v_i|^2 > 0$$

是一个正实数。

另一方面，因为 A 是实对称矩阵，所以 $A^* = A$ 。

因此：

$$\overline{\mathbf{v}^* A \mathbf{v}} = \mathbf{v}^* A^* \mathbf{v} = \mathbf{v}^* A \mathbf{v}$$

一个复数等于自己的共轭，意味着它是实数。

所以分子、分母都是实数，最终：

$$\lambda \in \mathbb{R}$$

我们直接证明了：实对称矩阵的特征值一定是实数。

而且由于 A 和 λ 都是实的，如果原来的复特征向量写成 $\mathbf{v} = \mathbf{a} + i\mathbf{b}$ ，那么：

$$A\mathbf{a} = \lambda\mathbf{a}, \quad A\mathbf{b} = \lambda\mathbf{b}$$

其中至少一个是非零实向量，所以一定能找到实特征向量。

这样，第一个特征向量就找到了，根本不需要构造单位球面上的极值。

接下来完全沿用你已经理解的思路：

找到实特征向量
↓
证明其正交补空间在 A 下不变
↓
限制在正交补上的算子仍然对称
↓
递归降维
↓
得到正交归一的特征向量基

这同样是谱定理的完整证明路线。

二、那么数学家为什么还要构造那个二次型？

因为那个构造不是毫无来由的。我们可以反过来推导它。

假设已经有一个单位特征向量：

$$A\mathbf{v} = \lambda\mathbf{v}, \quad \mathbf{v}^T\mathbf{v} = 1$$

等式两边左乘 $\mathbf{v}^T$ ：

$$\mathbf{v}^T A\mathbf{v} = \lambda\mathbf{v}^T\mathbf{v}$$

所以：

$$\boxed{\lambda = \mathbf{v}^T A\mathbf{v}}$$

这提示我们：特征值与二次型 $\mathbf{x}^T A\mathbf{x}$ 存在联系。

那么，自然可以尝试一个逆向问题：

如果我在所有单位向量中，寻找使 $\mathbf{x}^T A\mathbf{x}$ 最大的向量，它会不会恰好是某个特征向量？

这个问题可以通过拉格朗日乘数法来解决。

我们要求：

$$\max_{\mathbf{x}} \mathbf{x}^T A\mathbf{x}$$

约束为：

$$\mathbf{x}^T \mathbf{x} = 1$$

构造拉格朗日函数：

$$\mathcal{L}(\mathbf{x}, \mu) = \mathbf{x}^T A \mathbf{x} - \mu(\mathbf{x}^T \mathbf{x} - 1)$$

对 $\mathbf{x}$ 求梯度：

$$\nabla_{\mathbf{x}} \mathcal{L} = 2A\mathbf{x} - 2\mu\mathbf{x}$$

在约束极值点，满足：

$$2A\mathbf{x} - 2\mu\mathbf{x} = 0$$

也就是：

$A\mathbf{x} = \mu\mathbf{x}$

你看，特征值方程竟然直接从约束极值条件中出现了！

现在再加上单位球面紧致、二次型连续，就能够保证这样的极值点一定存在。

于是就自然产生了你之前看到的那个证明。

三、两种证明的区别

	代数证明	极值构造证明
起点	特征方程 $\det(A - \lambda I) = 0$	二次型 $\mathbf{x}^T A \mathbf{x}$
关键理论	代数基本定理	紧致性、极值定理
主要技巧	利用对称性证明特征值为实数	极值点满足特征值方程
额外收获	证明特征值的实数性质	联系特征值与优化问题

这两种证明都依赖更深层的数学结论，并不是其中一种完全不需要前提。

我认为这里最有意思的是：**特征向量和函数极值**，看似来自完全不同的数学领域，却通过**二次型和约束优化**联系在了一起。

而且这恰好与你正在学习的 Transformer 有关系。神经网络优化中，我们研究 Hessian 的特征值，本质上也在分析损失函数沿不同方向的二阶变化率。

所以，之前那个证明虽然精巧，但不是孤立的数学技巧，而是线性代数与最优化理论之间一个很深刻的联系。

补充：这条路线完成了什么

代数基本定理先给的是复根，非零核向量也先在复空间中取得。对称性再保证 λ 实，并从复向量的实部、虚部里找到非零实特征向量；不能在这一步之前预设所有根都是实数。

本章只要找到一个实特征向量，就能接入下一章的正交补归纳。共轭转置证明也直接表明任意复特征值都是实数。代数基本定理在此明确作为已有基础，不用“谱定理保证实根”来证明谱定理。

第 10 章：谱定理的球面路线与共同的降维归纳

[返回目录](#)

这实际上是线性代数中非常重要的一个定理：**实对称矩阵的谱定理 (Spectral Theorem)**。

你问的是：为什么只要一个矩阵满足

$$A^T = A$$

就一定能够找到 n 个两两正交的单位特征向量，从而写成：

$$A = Q\Lambda Q^T$$

我们可以不用预先假设它可对角化，而是从实对称这个条件出发，一步步证明它为什么必然能够对角化。

一、先明确需要证明什么

假设 A 是一个 $n \times n$ 的实对称矩阵。

我们需要证明存在 n 个向量：

$$\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_n$$

同时满足：

1. 每个都是特征向量： $A\mathbf{q}_i = \lambda_i \mathbf{q}_i$ 。
2. 两两正交： $\mathbf{q}_i^T \mathbf{q}_j = 0$ ，当 $i \neq j$ 。
3. 长度均为 1： $\|\mathbf{q}_i\| = 1$ 。

只要证明这三条，构造：

$$Q = [\mathbf{q}_1 \quad \dots \quad \mathbf{q}_n]$$

就可以得到 $Q^T Q = I$ 和 $AQ = Q\Lambda$ ，于是：

$$A = Q\Lambda Q^T$$

所以，真正困难的地方是：为什么一定能找到这样一整组特征向量？

二、第一步：证明至少存在一个实特征向量

考虑一个实对称矩阵 A 。

我们构造一个函数：

$$f(\mathbf{x}) = \mathbf{x}^T A \mathbf{x}$$

但限制：

$$\|\mathbf{x}\| = 1$$

这意味着我们只在单位球面上寻找函数的最大值。

由于单位球面是紧集， f 又是连续函数，所以它一定能取得最大值。

假设最大值出现在单位向量 $\mathbf{v}$ 。

现在取任意一个垂直于 $\mathbf{v}$ 的单位向量 $\mathbf{w}$ ，即：

$$\mathbf{v}^T \mathbf{w} = 0$$

沿着球面附近的一条路径移动：

$$\mathbf{x}(t) = \cos t \mathbf{v} + \sin t \mathbf{w}$$

你可以验证：

$$\|\mathbf{x}(t)\|^2 = \cos^2 t + \sin^2 t = 1$$

因此它始终在单位球面上。

因为 $t = 0$ 时 $\mathbf{x}(0) = \mathbf{v}$ ，而 $\mathbf{v}$ 是最大值点，所以：

$$\left. \frac{d}{dt} f(\mathbf{x}(t)) \right|_{t=0} = 0$$

利用乘积法则和 $\mathbf{A}^T = \mathbf{A}$ ：

$$\frac{d}{dt} (\mathbf{x}^T \mathbf{A} \mathbf{x}) = 2\mathbf{x}'(t)^T \mathbf{A} \mathbf{x}(t)$$

在 $t = 0$ 时：

$$\mathbf{x}(0) = \mathbf{v}, \quad \mathbf{x}'(0) = \mathbf{w}$$

因此：

$$2\mathbf{w}^T A\mathbf{v} = 0$$

也就是：

$$\mathbf{w}^T A\mathbf{v} = 0$$

这一步非常关键。

因为 $\mathbf{w}$ 可以是任何垂直于 $\mathbf{v}$ 的向量，所以 $A\mathbf{v}$ 与所有这些向量都垂直。

这意味着 $A\mathbf{v}$ 只能沿着 $\mathbf{v}$ 的方向。

因此必然存在一个实数 λ ，使得：

$$A\mathbf{v} = \lambda\mathbf{v}$$

这正是特征向量的定义！

所以，我们已经证明：任意实对称矩阵至少存在一个实特征值和对应的实特征向量。

三、第二步：证明垂直于特征向量的空间不会被矩阵变换出去

这是整个谱定理证明中最关键的一步。

我们已经找到：

$$A\mathbf{v} = \lambda\mathbf{v}$$

现在考虑所有与 $\mathbf{v}$ 垂直的向量组成的空间：

$$V^\perp = \{\mathbf{w} : \mathbf{v}^T \mathbf{w} = 0\}$$

例如，在三维空间中，假设 $\mathbf{v}$ 指向竖直方向，那么 $V^\perp$ 就是与它垂直的一个二维平面。

我们要证明：

$$\mathbf{w} \perp \mathbf{v} \implies A\mathbf{w} \perp \mathbf{v}$$

证明非常简洁。

因为 A 对称：

$$\mathbf{v}^T A\mathbf{w} = (A\mathbf{v})^T \mathbf{w}$$

代入 $A\mathbf{v} = \lambda\mathbf{v}$ ：

$$\mathbf{v}^T A\mathbf{w} = \lambda \mathbf{v}^T \mathbf{w}$$

而 $\mathbf{v}^T \mathbf{w} = 0$ ，所以：

$$\boxed{\mathbf{v}^T A\mathbf{w} = 0}$$

这意味着：

$$\boxed{A\mathbf{w} \perp \mathbf{v}}$$

这就证明了：实对称矩阵会把与某个特征向量垂直的整个子空间，仍然映射到该子空间内部。

数学上称它为不变子空间（Invariant Subspace）。

四、第三步：在剩下的子空间中继续寻找特征向量

现在假设原来是三维空间。

我们已经找到第一个单位特征向量：

$$\mathbf{q}_1$$

与它垂直的空间是二维的。

而我们刚刚证明，矩阵 A 对这个二维子空间的变换不会跑到子空间外面。

更重要的是，限制在这个二维子空间上的线性变换仍然是实对称的。

因为对于子空间内任意 $\mathbf{u}, \mathbf{w}$ ，都有：

$$\mathbf{u}^T A \mathbf{w} = (A \mathbf{u})^T \mathbf{w}$$

于是我们可以在二维子空间中，重复第一步的证明，找到第二个单位特征向量：

$$\mathbf{q}_2$$

并且自然满足：

$$\mathbf{q}_1^T \mathbf{q}_2 = 0$$

接下来，在同时垂直于 $\mathbf{q}_1, \mathbf{q}_2$ 的一维子空间中，再找到：

$$\mathbf{q}_3$$

这样，三维空间中就有了三个两两正交的单位特征向量。

同样的过程可以推广到任意 n 维：

$$n \longrightarrow n-1 \longrightarrow n-2 \longrightarrow \cdots \longrightarrow 1$$

最终得到：

$$\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_n$$

它们两两正交、长度为 1，并且每一个都是特征向量。

由于 n 个非零正交向量必然线性无关，它们就构成了整个 n 维空间的一组基。

所以：

$$A = Q\Lambda Q^T$$

谱定理证明完成。

五、为什么非对称矩阵不一定能做到？

我们看一个反例：

$$A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

这是一个将二维向量逆时针旋转 90° 的矩阵。

如果存在实特征向量 $\mathbf{v}$ ，应当满足：

$$A\mathbf{v} = \lambda\mathbf{v}$$

也就是旋转后的向量仍然与原向量共线。

但任何非零实向量旋转 90° 后都与原向量垂直，不可能共线。

所以，这个矩阵甚至不存在实特征向量。

因此，一般实矩阵并不具备谱定理所保证的性质。

补充：不借助谱定理完成严格归纳

当 $n = 1$ ，任意实数矩阵已有一个单位特征向量。设结论对维数 $n - 1$ 成立。通过本章球面法或上一章代数法，取得单位实特征向量 q ， $Aq = \lambda q$ 。

令 $V = \{w : q^T w = 0\}$ 。 q^T 是秩为一的线性泛函，故 $\dim V = n - 1$ 。通过一般线性空间取基和 Gram-Schmidt，选取 V 的正交归一基，组成 $B \in \mathbb{R}^{n \times (n-1)}$ ，满足 $B^T B = I$ 。

不变性保证 AB 的每列属于 V ，因而 $AB = BC$ ，其中 $C = B^T AB$ 。并且

$$C^T = B^T A^T B = B^T AB = C.$$

由归纳假设，存在 $(n - 1)$ 阶正交矩阵 R 使 $C = RDR^T$ 。于是 BR 的列在 V 中两两正交归一，且 $A(BR) = (BR)D$ 。连同 q 组成

$$Q = [q \ BR], \quad Q^T Q = I, \quad AQ = Q \operatorname{diag}(\lambda, D).$$

因此 $A = Q \operatorname{diag}(\lambda, D)Q^T$ 。这段归纳没有预先假定原矩阵可对角化，也没有预先假定重特征值的重数等于特征空间维数。

补充：球面构造为什么能被想到

已知特征向量若归一化，则 $Av = \lambda v$ 推出 $\lambda = v^T Av$ 。因此先研究单位方向上的二次型，是从目标方程倒推候选量。长度不固定时，二次型按 c^2 缩放，无法单独比较方向。

球面极值条件说 Av 没有切向分量。严格地，取 $z = Av - (v^T Av)v$ ，则 $z \perp v$ ；已证明所有 $w \perp v$ 都有 $w^T Av = 0$ ，所以取 $w = z$ 得 $\|z\|^2 = 0$ ，从而 $Av = (v^T Av)v$ 。这补足了“只能沿着 v ”的代数依据。

球面方法中的 $n = 1$ 无需选择垂直单位向量，已经由归纳基直接解决。拉格朗日乘数可帮助发现方程，但本章沿球面曲线求导的证明无需另引用乘数定理。

第 11 章：正定性、正特征值与统一二次下界

[返回目录](#)

因为正定矩阵（Positive Definite Matrix）的定义，就是它对应的二次型对任意非零向量都严格大于零。

而我们可以利用特征值方程，证明正定矩阵的所有特征值都必须大于零。

一、先理解什么是正定矩阵

假设 H 是实对称矩阵。

如果对任意非零向量 $\mathbf{x}$ ，都有：

$$\mathbf{x}^T H \mathbf{x} > 0$$

那么就称 H 为正定矩阵。

例如：

$$H = \begin{bmatrix} 2 & 0 \\ 0 & 5 \end{bmatrix}$$

取任意非零向量：

$$\mathbf{x} = \begin{bmatrix} a \\ b \end{bmatrix}$$

计算：

$$\mathbf{x}^T H \mathbf{x} = 2a^2 + 5b^2$$

由于平方项非负，而且 a, b 不同时为零，所以：

$$2a^2 + 5b^2 > 0$$

因此 H 正定。

二、证明正定矩阵的特征值全部大于零

设 λ_i 是 H 的一个特征值， $\mathbf{q}_i$ 是对应的非零特征向量。

根据特征值定义：

$$H\mathbf{q}_i = \lambda_i\mathbf{q}_i$$

等式两边左乘 $\mathbf{q}_i^T$ ：

$$\mathbf{q}_i^T H \mathbf{q}_i = \lambda_i \mathbf{q}_i^T \mathbf{q}_i$$

现在看等式两边。

左边：

由于 H 正定，而且 $\mathbf{q}_i \neq 0$ ：

$$\mathbf{q}_i^T H \mathbf{q}_i > 0$$

右边：

因为：

$$\mathbf{q}_i^T \mathbf{q}_i = \|\mathbf{q}_i\|^2 > 0$$

所以：

$$\lambda_i \|\mathbf{q}_i\|^2 > 0$$

由于向量长度的平方严格大于零，只能得到：

$$\lambda_i > 0$$

而 λ_i 是任意一个特征值，因此：

$$H \text{ 正定} \implies \text{所有特征值均为正}$$

这就证明了你问的那一步。

三、反过来也成立吗？

成立，而且这是一个充要条件。

对于实对称矩阵 H ，根据谱定理：

$$H = Q\Lambda Q^T$$

因此对任意非零向量 $\mathbf{x}$ ，令 $\mathbf{u} = Q^T \mathbf{x}$ ，就有：

$$\mathbf{x}^T H \mathbf{x} = \mathbf{u}^T \Lambda \mathbf{u} = \sum_{i=1}^n \lambda_i u_i^2$$

如果所有 $\lambda_i > 0$ ，又因为 $\mathbf{x} \neq 0$ 且 Q 可逆，所以 $\mathbf{u} \neq 0$ 。

至少有一个 $u_i^2 > 0$ ，因此：

$$\sum_{i=1}^n \lambda_i u_i^2 > 0$$

这就证明 H 正定。

最终：

$$H \text{ 实对称正定} \iff \lambda_1, \lambda_2, \dots, \lambda_n > 0$$

这个不等式其实就是实对称矩阵谱定理的一个直接推论。你前面已经理解了特征值分解：

$$H = Q\Lambda Q^T$$

现在只需要利用它，就能一步步推出：

$$\mathbf{h}^T H \mathbf{h} \geq \lambda_{\min} \|\mathbf{h}\|^2$$

这里最关键的理解是：任何向量都可以分解到 Hessian 的各个正交特征向量方向上，而每个方向对应一个特征值。

一、从特征值分解开始

假设 H 是一个 $n \times n$ 的实对称正定矩阵。

根据谱定理：

$$H = Q \Lambda Q^T$$

其中：

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}$$

因为 H 正定，所以：

$$\lambda_i > 0$$

将最小的特征值记为：

$$\lambda_{\min} = \min_i \lambda_i > 0$$

二、把 H 代入原来的二次型

我们要研究：

$$\mathbf{h}^T H \mathbf{h}$$

代入 $H = Q \Lambda Q^T$ ：

$$\mathbf{h}^T H \mathbf{h} = \mathbf{h}^T Q \Lambda Q^T \mathbf{h}$$

现在令：

$$\mathbf{u} = Q^T \mathbf{h}$$

由于矩阵转置满足：

$$(Q^T \mathbf{h})^T = \mathbf{h}^T Q$$

因此：

$$\mathbf{h}^T H \mathbf{h} = \mathbf{u}^T \Lambda \mathbf{u}$$

这一步实际上是把原来的向量 $\mathbf{h}$ ，改用特征向量构成的新坐标系来表示。

三、利用对角矩阵的特点展开

假设：

$$\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

那么：

$$\mathbf{u}^T \Lambda \mathbf{u} = [u_1 \quad u_2 \quad \cdots \quad u_n] \begin{bmatrix} \lambda_1 u_1 \\ \lambda_2 u_2 \\ \vdots \\ \lambda_n u_n \end{bmatrix}$$

所以：

$$\boxed{\mathbf{u}^T \Lambda \mathbf{u} = \sum_{i=1}^n \lambda_i u_i^2}$$

因为每个特征值都满足：

$$\lambda_i \geq \lambda_{\min}$$

而且：

$$u_i^2 \geq 0$$

所以逐项比较：

$$\lambda_i u_i^2 \geq \lambda_{\min} u_i^2$$

全部相加：

$$\sum_{i=1}^n \lambda_i u_i^2 \geq \lambda_{\min} \sum_{i=1}^n u_i^2$$

也就是：

$$\boxed{\mathbf{h}^T H \mathbf{h} \geq \lambda_{\min} \|\mathbf{u}\|^2}$$

最后，由于 Q 是正交矩阵：

$$\|\mathbf{u}\|^2 = (Q^T \mathbf{h})^T (Q^T \mathbf{h}) = \mathbf{h}^T Q Q^T \mathbf{h} = \mathbf{h}^T \mathbf{h} = \|\mathbf{h}\|^2$$

最终得到：

$$\boxed{\mathbf{h}^T H \mathbf{h} \geq \lambda_{\min} \|\mathbf{h}\|^2}$$

证明完成。

四、用具体数字验证

假设：

$$H = \begin{bmatrix} 2 & 0 \\ 0 & 5 \end{bmatrix}$$

那么它的最小特征值为：

$$\lambda_{\min} = 2$$

对于任意向量：

$$\mathbf{h} = \begin{bmatrix} x \\ y \end{bmatrix}$$

有：

$$\mathbf{h}^T H \mathbf{h} = 2x^2 + 5y^2$$

而：

$$\lambda_{\min} \|\mathbf{h}\|^2 = 2(x^2 + y^2)$$

两者相减：

$$(2x^2 + 5y^2) - 2(x^2 + y^2) = 3y^2 \geq 0$$

所以不等式一定成立。

改变单位向量方向时，二次型如何变化，可参看本章静态图。

五、为什么这个不等式在证明极小值时如此重要？

回忆驻点附近的二阶泰勒展开：

$$f(\mathbf{p} + \mathbf{h}) - f(\mathbf{p}) = \frac{1}{2} \mathbf{h}^T H \mathbf{h} + o(\|\mathbf{h}\|^2)$$

现在利用刚才证明的不等式：

$$f(\mathbf{p} + \mathbf{h}) - f(\mathbf{p}) \geq \frac{1}{2} \lambda_{\min} \|\mathbf{h}\|^2 + o(\|\mathbf{h}\|^2)$$

因为 $\lambda_{\min} > 0$ ，而余项相对于 $\|\mathbf{h}\|^2$ 趋于零，所以只要 $\mathbf{h}$ 足够小且非零，右侧就严格大于零。

因此：

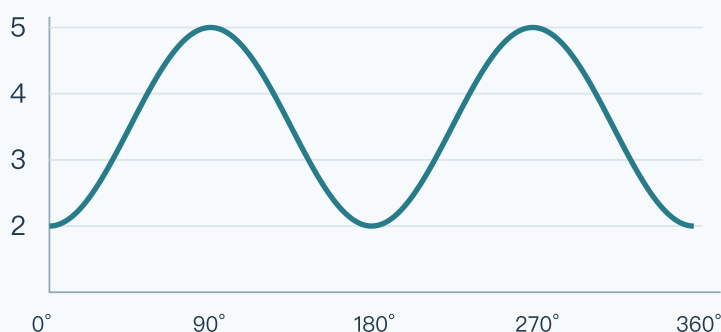
$$\boxed{f(\mathbf{p} + \mathbf{h}) > f(\mathbf{p})}$$

这就严格证明了驻点是局部极小值。

整个推导真正重要的地方是：正定矩阵不仅让二次型大于零，还保证它至少以 $\lambda_{\min} \|\mathbf{h}\|^2$ 的速度增长。

正是这个统一的正下界，让我们能控制泰勒展开中可能为负的高阶余项，而不只是凭直觉说曲面向上弯曲。

图 4 · 单位方向上的二次型： $2 \leq h^T H h \leq 5$



$$H = \text{diag}(2, 5)$$

$$h = (\cos \theta, \sin \theta)$$

$$\text{值} = 2 \cos^2 \theta + 5 \sin^2 \theta$$

最小值 2 对所有方向统一有效。

横轴为单位向量方向角，纵轴为二次型值。这里可以直接取最小特征值 2 控制所有小位移；图形按公式逐度采样，不依赖聊天中的交互组件。

补充：哪些条件不能混在一起

本笔记的“正定矩阵”限于实对称矩阵。正定是对全部非零向量的二次型要求，不是说每个矩阵元素都正。矩阵 $\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$ 有负元素，却正定；全为正元素的 $\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$ 有特征值 3, -1，并不正定。

“全部特征值正”必须搭配实对称的假设，才能等价于这里的正定。例如 $A = \begin{bmatrix} 1 & 10 \\ 0 & 1 \end{bmatrix}$ 的特征值均为 1，但 $x = (1, -1)^T$ 时 $x^T A x = -8$ 。

对任意实对称 H ，无论是否正定，均有

$$\lambda_{\min} \|h\|^2 \leq h^T H h \leq \lambda_{\max} \|h\|^2.$$

证明就是在正交特征坐标下逐项比较。除以非零 $\|h\|^2$ ，Rayleigh 商等于特征值的加权平均，权重为 $u_i^2 / \sum_j u_j^2$ 。取对应单位特征向量达到上下界。

若只想证明正定二次型有统一正下界，其实可以不使用谱定理：连续函数 $u \mapsto u^T H u$ 在单位球面取得最小值 m ；因球面上每个值都严格正，取到的最小值也正。将 h 写成 $\|h\|u$ ，得 $h^T H h \geq m \|h\|^2$ 。谱定理进一步认出 $m = \lambda_{\min}$ 。

第 12 章：回到 Hessian：用泰勒余项严格完成二阶判别

[返回目录](#)

我们从数学上严格证明两个结论：

1. 为什么可微曲面的内部最高点和最低点，梯度必然为零？
2. 为什么 Hessian 矩阵正定对应局部极小值，负定对应局部极大值，不定对应鞍点？

第二个结论尤其重要，它实际上是把一元函数的二阶导数判别法推广到了多元函数。

一、证明极值点的梯度为零

假设有一个二元函数：

$$z = f(x, y)$$

在点 (x_0, y_0) 取得局部极小值。

根据局部极小值的定义，存在一个足够小的邻域，使得：

$$f(x, y) \geq f(x_0, y_0)$$

现在，我们固定 $y = y_0$ ，只允许 x 变化。

这相当于在三维曲面上沿着 x 方向切出一条曲线：

$$g(x) = f(x, y_0)$$

因为 (x_0, y_0) 是原曲面的局部最低点，所以 x_0 也是 $g(x)$ 的局部最低点。

根据一元函数的费马定理 (Fermat's theorem)：

$$g'(x_0) = 0$$

而根据偏导数的定义：

$$g'(x_0) = \frac{\partial f}{\partial x}(x_0, y_0)$$

因此：

$$\frac{\partial f}{\partial x}(x_0, y_0) = 0$$

同理，固定 $x = x_0$ ，只允许 y 变化：

$$\frac{\partial f}{\partial y}(x_0, y_0) = 0$$

所以：

$$\nabla f(x_0, y_0) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

局部极大值可以用完全相同的方法证明。

这就证明了：只要函数在内部极值点可微，那么梯度必然为零。

注意，这只是必要条件，不是充分条件。

二、为什么需要 Hessian 矩阵？

先回顾一元函数。

假设：

$$f'(x_0) = 0$$

此时：

- 如果 $f''(x_0) > 0$ ，函数在附近向上弯曲， x_0 是局部极小值。
- 如果 $f''(x_0) < 0$ ，函数在附近向下弯曲， x_0 是局部极大值。

对于二元函数，仅计算 f_{xx} 和 f_{yy} 还不够，因为我们不但可以沿 x 、 y 方向移动，还可以沿任意斜方向移动。

因此，我们需要描述函数沿任意方向的二阶变化率。

假设沿着向量：

$$\mathbf{v} = \begin{bmatrix} a \\ b \end{bmatrix}$$

移动，定义一个一元函数：

$$g(t) = f(x_0 + at, y_0 + bt)$$

这里的 t 表示沿这个方向移动多少。

对 t 求第一次导数：

$$g'(t) = af_x + bf_y$$

再求一次导数，使用链式法则：

$$g''(t) = a^2 f_{xx} + ab f_{xy} + ab f_{yx} + b^2 f_{yy}$$

假设 f 的二阶偏导连续，那么：

$$f_{xy} = f_{yx}$$

因此：

$$g''(t) = a^2 f_{xx} + 2ab f_{xy} + b^2 f_{yy}$$

现在定义 Hessian 矩阵：

$$H = \begin{bmatrix} f_{xx} & f_{xy} \\ f_{yx} & f_{yy} \end{bmatrix}$$

那么上面的表达式就可以写为：

$$g''(0) = \mathbf{v}^T H \mathbf{v}$$

这个公式是整个证明的关键。

$$\underbrace{\mathbf{v}^T H \mathbf{v}}_{\text{矩阵二次型}} = \underbrace{g''(0)}_{\text{沿方向 } \mathbf{v} \text{ 的二阶变化率}}$$

对单位向量而言，它就是沿该方向的二阶方向导数。

三、证明 Hessian 正定意味着局部极小值

先理解正定（Positive Definite）的定义。

如果对任意非零向量 $\mathbf{v}$ ，都有：

$$\mathbf{v}^T H \mathbf{v} > 0$$

我们就称矩阵 H 为正定矩阵。

根据刚才证明的公式：

$$g''(0) = \mathbf{v}^T H \mathbf{v}$$

意味着：

$$g''(0) > 0$$

无论沿哪个非零方向，二阶变化率都为正。

但是，要严格证明它是局部极小值，还需要利用二阶泰勒展开。

假设在 $\mathbf{p} = (x_0, y_0)$ 有：

$$\nabla f(\mathbf{p}) = 0$$

设小位移为 $\mathbf{h} = (\Delta x, \Delta y)^T$ 。

二阶泰勒展开为：

$$f(\mathbf{p} + \mathbf{h}) = f(\mathbf{p}) + \nabla f(\mathbf{p})^T \mathbf{h} + \frac{1}{2} \mathbf{h}^T H(\mathbf{p}) \mathbf{h} + o(\|\mathbf{h}\|^2)$$

因为梯度为零：

$$f(\mathbf{p} + \mathbf{h}) - f(\mathbf{p}) = \frac{1}{2} \mathbf{h}^T H \mathbf{h} + o(\|\mathbf{h}\|^2)$$

假设 H 正定。由于它是实对称矩阵，存在最小特征值 $\lambda_{\min} > 0$ ，满足：

$$\mathbf{h}^T H \mathbf{h} \geq \lambda_{\min} \|\mathbf{h}\|^2$$

而余项 $o(\|\mathbf{h}\|^2)$ 相对于 $\|\mathbf{h}\|^2$ 趋于零，因此当 $\mathbf{h}$ 足够小时：

$$|o(\|\mathbf{h}\|^2)| < \frac{\lambda_{\min}}{4} \|\mathbf{h}\|^2$$

于是：

$$f(\mathbf{p} + \mathbf{h}) - f(\mathbf{p}) \geq \frac{\lambda_{\min}}{4} \|\mathbf{h}\|^2 > 0$$

只要 $\mathbf{h} \neq 0$ ，函数值就严格增大。

因此：

$$H \text{ 正定} \implies \text{严格局部极小值}$$

这里默认该点是驻点，即 $\nabla f = \mathbf{0}$ ，且函数二阶连续可微。

四、负定和不定的证明

负定的证明完全类似。

如果：

$$\mathbf{v}^T H \mathbf{v} < 0 \quad (\forall \mathbf{v} \neq \mathbf{0})$$

那么泰勒展开的二次项在各个非零方向都严格为负，并且在足够小的邻域中压过高阶余项，因此：

$$H \text{ 负定} \implies \text{严格局部极大值}$$

那么不定矩阵 (Indefinite) 呢？

不定意味着存在两个方向 $\mathbf{v}_1, \mathbf{v}_2$ ，满足：

$$\mathbf{v}_1^T H \mathbf{v}_1 > 0$$

$$\mathbf{v}_2^T H \mathbf{v}_2 < 0$$

对于小的非零 t ，沿第一方向：

$$f(\mathbf{p} + t\mathbf{v}_1) - f(\mathbf{p}) > 0$$

而沿第二方向：

$$f(\mathbf{p} + t\mathbf{v}_2) - f(\mathbf{p}) < 0$$

所以，在任意足够小的邻域中，都有函数值高于该点和低于该点的位置。

因此它既不是局部最大值，也不是局部最小值。

$$H \text{ 不定} \implies \text{鞍点}$$

五、用三个具体函数验证

极小值 $z = x^2 + y^2$

极大值 $z = -x^2 - y^2$

鞍点 $z = x^2 - y^2$

在原点，这三个函数的梯度都是零，但 Hessian 分别为：

$$H_{\min} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \quad H_{\max} = \begin{bmatrix} -2 & 0 \\ 0 & -2 \end{bmatrix}$$

$$H_{\text{saddle}} = \begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix}$$

它们的特征值分别是：

函数	Hessian 特征值	结果
$x^2 + y^2$	2, 2	正定，极小值
$-x^2 - y^2$	-2, -2	负定，极大值
$x^2 - y^2$	2, -2	不定，鞍点

这里其实还能得到一个更深层的结论：Hessian 的特征向量给出了曲面二阶变化的主方向，特征值则描述相应方向上的二阶变化率。

但必须强调，Hessian 正定、负定是充分条件，不是必要条件。例如 $z = x^4 + y^4$ 在原点有严格极小值，但 Hessian 是零矩阵。

六、最终把整个逻辑串起来

对于二阶连续可微函数 $f(x, y)$ ，在定义域内部一点 $\mathbf{p}$ ：

第一步：求梯度

$$\nabla f(\mathbf{p}) = 0$$

第二步：计算 Hessian

$$H(\mathbf{p}) = \begin{bmatrix} f_{xx} & f_{xy} \\ f_{yx} & f_{yy} \end{bmatrix}$$

第三步：判断 Hessian 的正负定性

正定

严格局部极小值

负定

严格局部极大值

不定

鞍点

如果 Hessian 只是半正定或半负定，则通常无法仅凭二阶判别法得出结论。

所以最核心的数学关系其实只有两个：

$$\nabla f = 0 \quad (\text{内部可微极值点的必要条件})$$

$$f(\mathbf{p} + \mathbf{h}) - f(\mathbf{p}) \approx \frac{1}{2} \mathbf{h}^T H \mathbf{h}$$

第一个公式告诉你极值点的一阶变化率为零；第二个公式告诉你，在一阶变化率为零后，曲面究竟是向上弯、向下弯，还是不同方向弯曲的符号不同。

这就是为什么梯度用于寻找驻点，而 Hessian 可以进一步判断驻点的局部性质。

补充：混合偏导相等的依据

对 C^2 函数，在点 (a, b) 附近考虑矩形差分

$$\Delta = f(a+s, b+t) - f(a+s, b) - f(a, b+t) + f(a, b).$$

先对横向再纵向应用微积分基本定理，可得 $\Delta = \int_0^s \int_0^t f_{xy}(a+u, b+v) dv du$ ；反向处理得相应 f_{yx} 的迭代积分表达式。除以 st 并令非零 $s, t \rightarrow 0$ ，连续性分别使两式趋向 $f_{xy}(a, b)$ 与 $f_{yx}(a, b)$ ，故相等。其他坐标同理。这里 C^2 是足够条件。

补充：二阶泰勒公式和统一余项从何而来

固定小位移 h ，使线段 $p+th$ ($0 \leq t \leq 1$) 留在 U 内。对 $g(t) = f(p+th)$ 两次应用一元微积分基本定理，得

$$f(p+h) = f(p) + \nabla f(p)^T h + \int_0^1 (1-t) h^T H(p+th) h dt.$$

把 $H(p)$ 提出，剩下

$$r(h) = \int_0^1 (1-t) h^T [H(p+th) - H(p)] h dt.$$

定义算子范数 $\|M\|_{op} = \sup_{\|u\|=1} \|Mu\|$ 。由 Cauchy-Schwarz 不等式，

$$|r(h)| \leq \frac{1}{2} \sup_{0 \leq t \leq 1} \|H(p+th) - H(p)\|_{op} \|h\|^2 = o(\|h\|^2).$$

最后一步由 H 在 p 连续得到：只要 $\|h\| < \delta$ ，整条线段都在 δ 邻域内。这个 δ 对全部方向共同有效。因此第 3 节余项的控制不是每个方向分别选不同邻域，而是真正控制一个完整小球。

补充：二阶必要条件与半正定的不确定性

若 p 为 C^2 内部局部极小点，对任意固定 v ， $g(t) = f(p + tv)$ 有一元局部极小值，所以 $g''(0) = v^T H(p)v \geq 0$ 。例如也可用 $(g(t) + g(-t) - 2g(0))/t^2 \geq 0$ 取极限证明。于是 Hessian 必须半正定；极大必须半负定。

正定是严格极小的充分条件，半正定只是极小的必要条件。 $x^4 + y^4$ 、 $-(x^4 + y^4)$ 、 $x^4 - y^4$ 在原点 Hessian 全为零，却依次为严格极小、严格极大、鞍点。因此零特征值时，二阶判别法可能无结论。

补充：Hessian 与曲面几何曲率的边界

$v^T H v$ 始终是输入直线 $p + tv$ 上的二阶导。把它称为“曲率”是一种局部优化语言。在图形的驻点处，且采用输入空间的欧氏正交单位方向时，它与主方向的曲面曲率相联系；一般非驻点处的几何主曲率还包含一阶导数和曲面度量，不能不加条件地与 Hessian 特征值等同。

第 13 章：具体例子、易错边界与回到训练问题

[返回目录](#)

13.1 一组贯穿例子：交叉项如何消失

取

$$f(x, y) = \frac{3}{2}x^2 + xy + \frac{3}{2}y^2, \quad H = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}.$$

原点梯度为零。取 $q_1 = (1, 1)^T/\sqrt{2}$ 、 $q_2 = (-1, 1)^T/\sqrt{2}$ ，则特征值为 4, 2。新坐标为

$$u_1 = (x + y)/\sqrt{2}, \quad u_2 = (-x + y)/\sqrt{2}, \quad f = 2u_1^2 + u_2^2.$$

因正交变换保长度， $f(x, y) \geq x^2 + y^2$ ，所以原点是严格全局极小点。这里能说全局，是因为函数恰为正定二次型，不是因为一般二阶判别法自动给出全局结论。

13.2 二维判别式只是同一事实的简写

对实对称 $H = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$, 令 $D = ac - b^2$ 。若 $a \neq 0$, 完成平方得

$$ax^2 + 2bxy + cy^2 = a\left(x + \frac{b}{a}y\right)^2 + \frac{D}{a}y^2.$$

若 $D > 0, a > 0$, 两项系数皆正, 故正定; $D > 0, a < 0$ 则负定。若 $D < 0$, 两特征值乘积负, 故一正一负, 二次型不定。若 $D = 0$, 存在零特征值, 通常要看高阶项。这个规则仍须搭配 $\nabla f(p) = 0$ 才能作为驻点极值判别。

13.3 内部、可微、驻点和邻域条件都不可省

错误推断	反例及原因
极值处梯度总为零	$f(x) = x$ 在 $[0, 1]$ 的最小点是边界点 0 , 导数为 1
极小处导数一定存在	$f(x) = x $ 在 0 极小, 但不可微
Hessian 正定处就是极小点	$f(x) = x + x^2$ 在 0 二阶导为 2 , 一阶导为 1 , 并非极小点
梯度零就是极小点	$f(x, y) = x^2 - y^2$ 在原点为鞍点
Hessian 半正定能确定极小	$f(x, y) = x^2 - y^4$ 在原点半正定却为鞍点
局部极小必是全局极小	$f(x) = (x^2 - 1)^2 + x$ 的两处局部极小值不同

最后一例的导数为 $4x^3 - 4x + 1$, 在 $(-2, -1)$ 、 $(0, 1/2)$ 、 $(1/2, 1)$ 各有一根; 左右两根二阶导为正, 中央根二阶导为负。左侧极小函数值小于 -1 (在 $x = -1$ 的值为 -1 且该点非驻点), 右侧极小点为正数且两项使函数值正, 因此右侧局部极小不是全局极小。

13.4 从方向导数回到 Transformer

把全部可训练参数按固定顺序展平成 $\theta \in \mathbb{R}^N$ ，损失就成为 $L(\theta)$ 。梯度回答一阶局部变化；Hessian 回答梯度如何变化与二阶方向变化。对矩阵参数，二阶导数若展开为元素索引对，通常是四阶数组；展平后成为 $N \times N$ Hessian。它与本来同形的参数梯度不可混用。

即使全损失梯度在某处为零，各小批量梯度也未必为零；带动量或权重衰减的优化器也未必立即停止。高维 Hessian 可能出现零或负方向。这里证明的局部定理不承诺训练能到达该点，也不承诺找到全局最优。

13.5 可以独立复述的证明链

从矩阵链式法则得到同形参数梯度；用方向导数和 Cauchy–Schwarz 说明负梯度是一阶最速下降方向；用一元费马定理说明内部可微极值必须驻点；在驻点处，一阶项消失，二阶泰勒公式留下 Hessian 二次型与统一余项。

谱定理把实对称二次型化成 $\sum_i \lambda_i u_i^2$ ；正定给出全部 $\lambda_i > 0$ ，有限个特征值有正的最小值。这个统一正下界压过小 o 余项，于是得到整个邻域中的严格不等式，而非仅沿每条固定直线的判断。

追问谱定理时，先解决一个实特征向量的存在，再用正交补不变和归纳得到完整特征基。球面路线的存在性来自连续函数在非空紧集取极值；紧致性的欧氏空间刻画，又来自有限维和实数完备性。这些就是原来 20 轮问题之间的连接。

第 14 章：原对话逐轮覆盖索引与复习问题

[返回目录](#)

原分享标题为“Transformer 梯度矩阵推导”，共 20 轮、40 条可见用户/助手消息，117 个结构节点，线性链与映射节点一致，无分支、无可见非文本附件，所有可见消息均完成。原始 HTML、结构、逐条正文与读取报告保存在本版资料中。来源页面中的历史指令及演示脚本仅作为原始数据归档，没有作为当前指令或执行代码使用。

轮次	用户追问（保留原意及原表述）	本版对应
1	transformer中那个梯度的矩阵是怎么来着	矩阵梯度如何计算与转置从何而来
2	所以这里都是标量对矩阵求梯度，不是矩阵对矩阵	标量对矩阵梯度与矩阵对矩阵导数
3	因为从什么梯度下降来说我们需要的是一个和w矩阵行列一样的是吧	为什么参数更新需要同形梯度
4	yita是一个标量是吧	学习率 η 是否为标量
5	梯度向量本身是指向曲面的变化率最大的方向对吧	梯度为何指向变化率最大的方向
6	就三位空间来说，曲面的最低点和最高点的梯度有什么特点吗	曲面最高最低点的梯度特点
7	证明一下，证明一下	证明一阶必要条件与二阶判别
8	为什么	为什么有正的最小特征值
9	这一步怎么回事，一起介绍一下吧，矩阵这些知识我都忘记了	正交对角化及矩阵知识复习
10	为什么	为什么实对称矩阵存在正交特征基
11	中间的这个尖尖符号是对角矩阵吗	大写 Λ 是什么
12	为什么Q一定是正交的	为什么可以选出正交 Q
13	这个具体是怎么处理的	Gram—Schmidt 的具体步骤

轮次	用户追问（保留原意及原表述）	本版对应
14	换言之，实对称举证谱定理的证明思路是，先找到一个特征值，然后证明与这个特征值垂直的空间在A作用下是封闭的，然后A限制在这个空间中作用，还是对称的，然后不断降阶。我感觉第一步那里限制在球面上构建函数，有点不好理解，意思是闭集上面的函数或者说连续函数也是一定可以取到极值的吗	我的降维理解与极值定理的前提
15	为什么这个也成立	连续像紧致及极值定理为何成立
16	也就是紧集为什么等价于有界且是闭集	紧致为何等价于闭且有界
17	上述开覆盖不要求是小球形状的吗	开覆盖是否只能由小球构成
18	我感觉那个对称矩阵找特征向量的构造还是过分精巧了	球面构造太精巧，有无代数路线
19	这是怎么来的	二次型的最小特征值下界如何推出
20	正定矩阵为什么这些都是正的	正定为什么推出全部特征值为正

14.1 回查资料与相关笔记

[完整原对话](#) · [读取核对报告](#) · [Transformer 架构与训练笔记](#) · [代数基本定理笔记](#) · [阅读索引](#)

单独分享 HTML 或 PDF 时，正文与公式可离线阅读；以上跨文件链接需要同时保留相应文件及目录。Markdown 公式使用标准美元定界符；Typora 阅读时需开启行内数学支持。

14.2 自测：每一步是否真的理解

1. 不靠维度猜测，推导 $Y = XW$ 的两个梯度；说明求导时什么暂时视为常量。
2. 说明梯度为零时“最陡方向”的特殊性；举一个大步长不下降的例子。
3. 从双侧差商证明内部可微极值的导数为零，说明边界条件用在何处。
4. 为什么不同特征值的正交性还不足以单独证明谱定理？
5. 为什么 Gram–Schmidt 在同一特征空间内不会破坏特征向量性质？
6. 不用谱定理，分别由代数路线和球面路线找一个实特征向量。
7. 写出正交补不变的那一行等式，再解释限制矩阵为什么实对称。
8. 从开覆盖定义证明 Heine–Borel 两个方向，指出完备性用在哪里。
9. 原像覆盖法怎样证明连续像紧致？为什么上确界不等于最大值？
10. 把 $h^T H h$ 展开到特征坐标，说明统一正下界如何压过余项。
11. 区分正定充分条件与半正定必要条件，并给出退化鞍点。
12. 为什么这些局部结论尚不能保证 Transformer 训练达到全局最优？