

## Lecture 12 奇异值的变分性质、经典角、仿射空间 — 2025.12.2

教授：邵美悦

Scribe: 路人甲

## 目录

<b>1 奇异值的变分性质</b>	<b>1</b>
1.1 Hermite 矩阵的变分性质 (续)	1
1.2 奇异值的变分性质	2
1.3 Eckart Young 定理	6
<b>2 经典角</b>	<b>7</b>
2.1 经典角	7
2.2 经典角在收敛性证明中的应用	9
<b>3 仿射空间</b>	<b>10</b>

## 1 奇异值的变分性质

## 1.1 Hermite 矩阵的变分性质 (续)

在上节课的最后，我们提到了两个不等式：

$$1. \text{ Weyl 不等式: } |\lambda_i(\mathbf{A} + \mathbf{E}) - \lambda_i(\mathbf{A})| \leq \|\mathbf{E}\|_2 \Rightarrow \|\mathbf{\Lambda}(\mathbf{A} + \mathbf{E}) - \mathbf{\Lambda}(\mathbf{A})\|_2 \leq \|\mathbf{E}\|_2;$$

$$2. \text{ Hoffman-Wielandt 不等式: } \|\mathbf{\Lambda}(\mathbf{A} + \mathbf{E}) - \mathbf{\Lambda}(\mathbf{A})\|_F \leq \|\mathbf{E}\|_F.$$

我们可以注意到，上面两个不等式唯一的区别就是使用的范数——它们使用的范数分别是 2-范数和 Frobenius 范数——都是酉不变范数。一个自然的猜想是，对于更一般的酉不变范数，是否有类似的结论呢？也就是说，对于一般的酉不变范数  $\|\cdot\|_{\text{ui}}$  (ui: unitary invariant)，是否有

$$\|\mathbf{\Lambda}(\mathbf{A} + \mathbf{E}) - \mathbf{\Lambda}(\mathbf{A})\|_{\text{ui}} \leq \|\mathbf{E}\|_{\text{ui}}$$

事实上这个结论是成立的，但是证明起来比较复杂，邵老师在课上也没有讲，因此我们这里就不展开了。

这些特征值不等式在扰动分析中有着重要的应用。例如，我们常常需要对矩阵进行酉相似变换。从纯数学的角度来看，酉相似变换不会改变矩阵的特征值；但在实际计算中我们算出的酉矩阵往往带有一定的误差，从而导致计算出的特征值也会有误差。这也就是说，假设  $\varepsilon = \|\mathbf{X}^* \mathbf{X} - \mathbf{I}\|_2$  很小，我们希望知道  $\lambda_i(\mathbf{X}^* \mathbf{A} \mathbf{X})$  和  $\lambda_i(\mathbf{A})$  会差多少。我们也就是要估计  $|\lambda_i(\mathbf{X}^* \mathbf{A} \mathbf{X}) - \lambda_i(\mathbf{A})|$  的大小。

我们需要建立已知式  $\mathbf{X}^* \mathbf{X} - \mathbf{I}$  与未知式的关系。为方便起见，不妨设  $\lambda_i(\mathbf{A}) = \theta_i$ ，于是

$$\lambda_i(\mathbf{X}^* \mathbf{A} \mathbf{X}) - \theta_i = \lambda_i(\mathbf{X}^* \mathbf{A} \mathbf{X} - \theta_i \mathbf{I}) = \lambda_i(\mathbf{X}^* (\mathbf{A} - \theta_i \mathbf{I}) \mathbf{X} + \theta_i (\mathbf{X}^* \mathbf{X} - \mathbf{I})).$$

现在我们来看看我们这个操作是否有意义。右端项  $\mathbf{X}^* \mathbf{X} - \mathbf{I}$  是一个小矩阵，这是已知信息；左端的  $\mathbf{X}^* (\mathbf{A} - \theta_i \mathbf{I}) \mathbf{X}$  我们虽然不大清楚它的性质，但是  $\mathbf{A} - \theta_i \mathbf{I}$  的特征值我们是清楚的：它的第  $i$  个特征值  $\lambda_i(\mathbf{A} - \theta_i \mathbf{I}) = 0$ 。这是一个重要的发现，重要性在于，由于  $\mathbf{X}^* (\mathbf{A} - \theta_i \mathbf{I}) \mathbf{X}$  是对  $\mathbf{A} - \theta_i \mathbf{I}$  的合同变换，我们知道合同变换虽然会改变特征值，但并不改变惯性指数（正特征值、负特征值、零特征值的个数）。因此，我们一定有  $\lambda_i(\mathbf{X}^* (\mathbf{A} - \theta_i \mathbf{I}) \mathbf{X}) = 0$ 。

整理一下，我们现在需要估计一个矩阵的第  $i$  个特征值，而这个矩阵可以拆分为两个矩阵之和，其中一个矩阵的第  $i$  个特征值为零，另一个矩阵是已知的小矩阵。那么，我们该如何估计这个和矩阵的特征值呢？这正是 Weyl 不等式的用武之地。根据 Weyl 不等式，我们有

$$|\lambda_i(\mathbf{X}^* (\mathbf{A} - \theta_i \mathbf{I}) \mathbf{X} + \theta_i (\mathbf{X}^* \mathbf{X} - \mathbf{I})) - \lambda_i(\mathbf{X}^* (\mathbf{A} - \theta_i \mathbf{I}) \mathbf{X})| \leq \|\theta_i (\mathbf{X}^* \mathbf{X} - \mathbf{I})\|_2 = |\theta_i| \varepsilon.$$

又由  $\lambda_i(\mathbf{X}^* (\mathbf{A} - \theta_i \mathbf{I}) \mathbf{X}) = 0$ ，我们最终得到

$$|\lambda_i(\mathbf{X}^* (\mathbf{A} - \theta_i \mathbf{I}) \mathbf{X} + \theta_i (\mathbf{X}^* \mathbf{X} - \mathbf{I}))| \leq |\theta_i| \varepsilon,$$

亦即

$$|\lambda_i(\mathbf{X}^* \mathbf{A} \mathbf{X}) - \lambda_i(\mathbf{A})| \leq |\lambda_i(\mathbf{A})| \cdot \|\mathbf{X}^* \mathbf{X} - \mathbf{I}\|_2.$$

这个结论告诉我们，当我们对矩阵进行一个近似酉相似变换时，特征值的误差是可控的，并且误差的上界与原矩阵的特征值大小成正比。

## 1.2 奇异值的变分性质

特征值是方阵独有的，实特征值更是 Hermite 矩阵专属的。一般的矩阵没有特征值，只有奇异值。那么，对于奇异值，我们是否也有类似的变分性质呢？我们知道，矩阵  $\mathbf{A}$  的奇异值就是矩阵  $\mathbf{A}^* \mathbf{A}$  的特征值。因此，Hermite 矩阵的变分性质可以自然地推广到一般矩阵的奇异值上。

**Rayleigh 商** 设  $\mathbf{A} \in \mathbb{C}^{m \times n}$ . 由  $\sigma_i(\mathbf{A}) = \sqrt{\lambda_{n+1-i}(\mathbf{A}^* \mathbf{A})}$  可知, 奇异值的 Rayleigh 商定理为

$$\sigma_1^2(\mathbf{A}) = \max_{\mathbf{x} \in \mathbb{C}^n \setminus \{0\}} \frac{\mathbf{x}^* \mathbf{A}^* \mathbf{A} \mathbf{x}}{\mathbf{x}^* \mathbf{x}} = \max_{\mathbf{x} \in \mathbb{C}^n \setminus \{0\}} \frac{\|\mathbf{A} \mathbf{x}\|_2^2}{\|\mathbf{x}\|_2^2},$$

即

$$\sigma_1(\mathbf{A}) = \max_{\mathbf{x} \in \mathbb{C}^n \setminus \{0\}} \frac{\|\mathbf{A} \mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \|\mathbf{A}\|_2.$$

这个结论是我们所熟知的。再看

$$\sigma_n(\mathbf{A}) = \min_{\mathbf{x} \in \mathbb{C}^n \setminus \{0\}} \frac{\|\mathbf{A} \mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \min_{\|\mathbf{x}\|_2=1} \|\mathbf{A} \mathbf{x}\|_2.$$

以上两个结论告诉我们, 线性变换  $\mathbf{A}$  可以把一个单位球映成一个椭球, 而这个椭球的长轴和短轴长度分别是  $\sigma_1(\mathbf{A})$  和  $\sigma_n(\mathbf{A})$ , 其中长轴和短轴方向分别是  $\mathbf{u}_1$  和  $\mathbf{u}_n$  的方向。如下图所示。

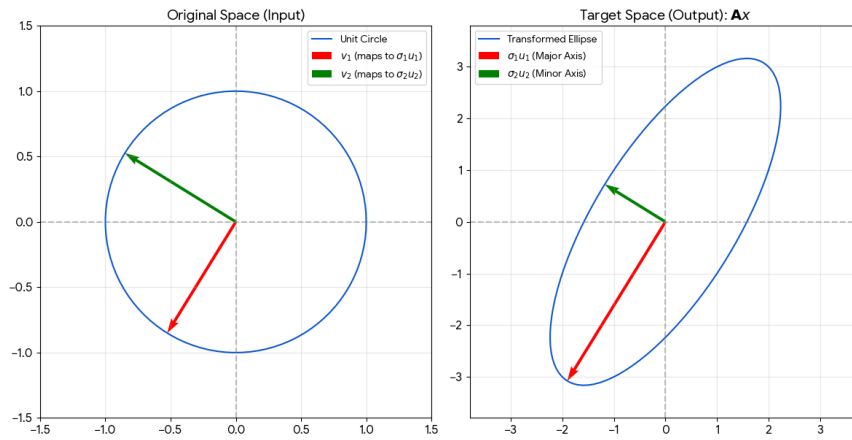


图 1:  $2 \times 2$  线性变换  $\mathbf{A}$  下的单位球与椭球

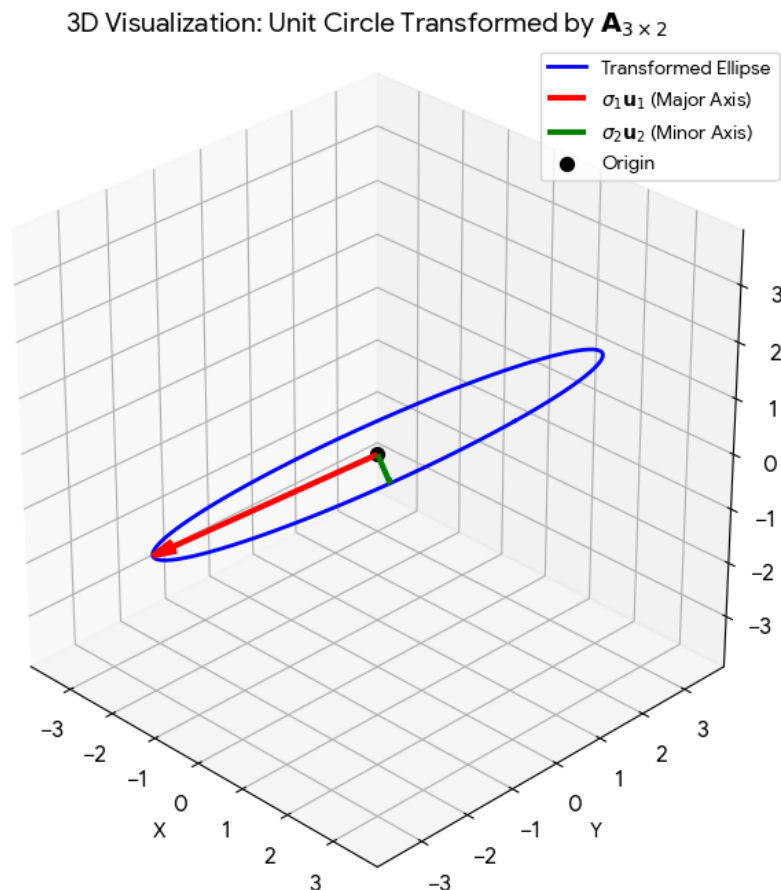


图 2:  $3 \times 2$  线性变换  $\mathbf{A}$  作用在单位球上得到的椭球

对于长方形矩阵，我们一定会有奇异值是零。在这些奇异值对应的右奇异向量方向上，线性变换  $\mathbf{A}$  会把向量映射为零向量。因此，虽然最后我们得到的椭球是在高维空间中的，但它本质上还是一个低维的椭球。

对于最大奇异值，我们还有另一种形式对它进行描述。任意给定一个输入方向  $\mathbf{x} \in \mathbb{C}^n \setminus \{\mathbf{0}\}$ ，且  $\|\mathbf{x}\|_2 = 1$ ，线性变换  $\mathbf{A}$  会把它映射为  $\mathbf{Ax}$ 。由于最大奇异值是映射  $\mathbf{A}$  的最大缩放因子，如果我们取任意一个单位向量  $\mathbf{y} \in \mathbb{C}^m \setminus \{\mathbf{0}\}$ ，让它与  $\mathbf{Ax}$  作内积，那么内积的最大模长应当就是  $\sigma_1(\mathbf{A})$ 。于是我们有

$$\sigma_1(\mathbf{A}) = \max_{\substack{\mathbf{x} \in \mathbb{C}^n \setminus \{\mathbf{0}\} \\ \mathbf{y} \in \mathbb{C}^m \setminus \{\mathbf{0}\}}} \frac{|\mathbf{y}^* \mathbf{Ax}|}{\|\mathbf{x}\|_2 \|\mathbf{y}\|_2}.$$

这被称为奇异值的双线性形式 (bilinear form)。

或许有同学会尝试作如下推广：

$$\sigma_n(\mathbf{A}) = \min_{\substack{\mathbf{x} \in \mathbb{C}^n \setminus \{\mathbf{0}\} \\ \mathbf{y} \in \mathbb{C}^m \setminus \{\mathbf{0}\}}} \frac{|\mathbf{y}^* \mathbf{Ax}|}{\|\mathbf{x}\|_2 \|\mathbf{y}\|_2}.$$

请注意，这个结论不对！右端的最小值是零，因为我们总能找到一个  $\mathbf{y}$  与  $\mathbf{Ax}$  正交，从而使得内积为零。而左端是不一定为零的——万一  $\mathbf{A}$  列满秩呢？

**Courant–Fischer min-max 定理** 我们还是从 Hermite 矩阵的 Courant–Fischer min-max 定理出发，来推导奇异值的 Courant–Fischer min-max 定理。设  $\mathbf{A} \in \mathbb{C}^{m \times n}$ ，则

$$\begin{aligned}\sigma_i(\mathbf{A})^2 &= \lambda_{n+1-i}(\mathbf{A}^* \mathbf{A}) = \max_{\dim V=i} \min_{\mathbf{x} \in V \setminus \{\mathbf{0}\}} R(\mathbf{A}^* \mathbf{A}, \mathbf{x}) = \max_{\dim V=i} \min_{\mathbf{x} \in V \setminus \{\mathbf{0}\}} \frac{\|\mathbf{Ax}\|_2^2}{\|\mathbf{x}\|_2^2} \\ &= \min_{\dim U=n-i+1} \max_{\mathbf{x} \in U \setminus \{\mathbf{0}\}} R(\mathbf{A}^* \mathbf{A}, \mathbf{x}) = \min_{\dim U=n-i+1} \max_{\mathbf{x} \in U \setminus \{\mathbf{0}\}} \frac{\|\mathbf{Ax}\|_2^2}{\|\mathbf{x}\|_2^2}.\end{aligned}$$

亦即

$$\begin{aligned}\sigma_i(\mathbf{A}) &= \max_{\dim V=i} \min_{\mathbf{x} \in V \setminus \{\mathbf{0}\}} \frac{\|\mathbf{Ax}\|_2}{\|\mathbf{x}\|_2}, \\ &= \min_{\dim U=n-i+1} \max_{\mathbf{x} \in U \setminus \{\mathbf{0}\}} \frac{\|\mathbf{Ax}\|_2}{\|\mathbf{x}\|_2}.\end{aligned}$$

这就是奇异值的 Courant–Fischer min-max 定理。形式上，它和 Hermite 矩阵的 Courant–Fischer min-max 定理非常相似，只不过由于奇异值和特征值的排列顺序相反，所以最大值和最小值的位置也对调了。它的重要性在于：它帮助我们成功地探测到了夹在最大奇异值和最小奇异值之间的其他奇异值。

这个定理理解起来也是没有任何难度的。我们可以把它想象成：在所有的  $i$  维子空间中，找一个子空间  $V$ ，使得在这个子空间中，线性变换  $\mathbf{A}$  的最小缩放因子最大；或者，在所有的  $n-i+1$  维子空间中，找一个子空间  $U$ ，使得在这个子空间中，线性变换  $\mathbf{A}$  的最大缩放因子最小。

**Weyl 不等式** Hermite 矩阵的 Weyl 不等式同样可以推广到奇异值上。在这里，我们想要的结论是：设  $\mathbf{A}, \mathbf{E} \in \mathbb{C}^{m \times n}$ ，则

$$|\sigma_i(\mathbf{A} + \mathbf{E}) - \sigma_i(\mathbf{A})| \leq \|\mathbf{E}\|_2.$$

如果像前面一样，利用 Hermite 矩阵的 Weyl 不等式来证明这个结论，那么我们需要处理  $(\mathbf{A} + \mathbf{E})^*(\mathbf{A} + \mathbf{E}) - \mathbf{A}^* \mathbf{A}$ ，二阶项比较多，会比较麻烦。这里我们采用另一种思路来证明这个结论。

在讲奇异值的时候，我们提到过一个重要的结论：矩阵

$$\tilde{\mathbf{A}} = \begin{bmatrix} \mathbf{0} & \mathbf{A} \\ \mathbf{A}^* & \mathbf{0} \end{bmatrix}$$

的特征值是  $\pm\sigma_1(\mathbf{A}), \dots, \pm\sigma_n(\mathbf{A}), 0, \dots, 0$ 。不妨设  $\tilde{\mathbf{E}}$  也按照同样的方式构造，我们对  $\tilde{\mathbf{A}}$  利用 Courant–Fischer min-max 定理有

$$\begin{aligned}|\lambda(\tilde{\mathbf{A}} + \tilde{\mathbf{E}}) - \lambda(\tilde{\mathbf{A}})| &\leq \|\tilde{\mathbf{E}}\|_2 = \|\mathbf{E}\|_2 \\ \Rightarrow |\sigma_i(\mathbf{A} + \mathbf{E}) - \sigma_i(\mathbf{A})| &\leq \|\mathbf{E}\|_2.\end{aligned}$$

**Hoffman–Wielandt 不等式** 同样地, Hermite 矩阵的 Hoffman–Wielandt 不等式也可以推广到奇异值上。设  $\mathbf{A}, \mathbf{E} \in \mathbb{C}^{m \times n}$ , 考虑

$$\tilde{\mathbf{A}} = \begin{bmatrix} \mathbf{0} & \mathbf{A} \\ \mathbf{A}^* & \mathbf{0} \end{bmatrix}, \quad \tilde{\mathbf{E}} = \begin{bmatrix} \mathbf{0} & \mathbf{E} \\ \mathbf{E}^* & \mathbf{0} \end{bmatrix}.$$

于是,

$$\begin{aligned} 2 \sum_{i=1}^n |\sigma_i(\mathbf{A} + \mathbf{E}) - \sigma_i(\mathbf{A})|^2 &= \sum_k |\lambda_k(\tilde{\mathbf{A}} + \tilde{\mathbf{E}}) - \lambda_k(\tilde{\mathbf{A}})|^2 \\ &\leq \left\| \begin{bmatrix} \mathbf{0} & \mathbf{E} \\ \mathbf{E}^* & \mathbf{0} \end{bmatrix} \right\|_{\text{F}}^2 \\ &= 2 \|\mathbf{E}\|_{\text{F}}^2 \end{aligned}$$

第一个等号是因为  $\tilde{\mathbf{A}}$  和  $\tilde{\mathbf{A}} + \tilde{\mathbf{E}}$  的特征值正负成对出现, 第二个不等号利用了 Hermite 矩阵的 Hoffman–Wielandt 不等式。最终我们得到

$$\sum_{i=1}^n |\sigma_i(\mathbf{A} + \mathbf{E}) - \sigma_i(\mathbf{A})|^2 \leq \|\mathbf{E}\|_{\text{F}}^2.$$

对于一般的酉不变范数  $\|\cdot\|_{\text{ui}}$ , 类似地, 我们也可以得到

$$\|\Sigma(\mathbf{A} + \mathbf{E}) - \Sigma(\mathbf{A})\|_{\text{ui}} \leq \|\mathbf{E}\|_{\text{ui}}.$$

这些不等式共同说明了奇异值对矩阵扰动的稳定性。这也不奇怪, 奇异值本质上是特征值的一种推广, 而特征值问题是很良态的。

### 1.3 Eckart Young 定理

对于一个高维数据集, 我们常常希望用一个低秩矩阵来近似它, 通过研究低秩的矩阵, 我们可以提取出数据的主要特征, 去除噪声和冗余信息。这在数据压缩、图像处理、机器学习等领域都有着广泛的应用。那么, 如何找到一个最佳的低秩近似矩阵呢? 这正是 Eckart Young 定理所要解决的问题。我们知道, 若  $\mathbf{A} = \sum_{i=1}^r \mathbf{u}_i \sigma_i \mathbf{v}_i^*$  是  $\mathbf{A}$  的奇异值分解, 一个低秩逼近矩阵是:  $\mathbf{B} = \sum_{i=1}^k \mathbf{u}_i \sigma_i \mathbf{v}_i^*$ . 这是很容易想到的。Eckart Young 定理告诉我们, 这个低秩矩阵  $\mathbf{B}$  实际上是所有秩为  $k$  的矩阵中, 最接近  $\mathbf{A}$  的矩阵。

**定理 1.1** (Eckart Young 定理). 设  $\mathbf{A} = \sum_{i=1}^r \mathbf{u}_i \sigma_i \mathbf{v}_i^*$ , 其中  $r = \text{rank}(\mathbf{A})$ , 则对任意  $k < r$ , 有

$$\min_{\text{rank}(\mathbf{B})=k} \|\mathbf{A} - \mathbf{B}\|_{\text{ui}} = \sigma_{k+1}(\mathbf{A}),$$

其中  $\|\cdot\|_{\text{ui}}$  是酉不变范数。

证明. 由不等式

$$\|\mathbf{A} - \mathbf{B}\|_{\text{ui}} \geq \|\Sigma(\mathbf{A}) - \Sigma(\mathbf{B})\|_{\text{ui}} \geq \sigma_{k+1}(\mathbf{A} - \mathbf{B}),$$

知

$$\min_{\text{rank}(\mathbf{B})=k} \|\mathbf{A} - \mathbf{B}\|_{\text{ui}} \geq \sigma_{k+1}(\mathbf{A}).$$

另一方面, 取

$$\mathbf{B} = \sum_{i=1}^k \mathbf{u}_i \sigma_i \mathbf{v}_i^*,$$

则  $\text{rank}(\mathbf{B}) = k$ , 且

$$\|\mathbf{A} - \mathbf{B}\|_{\text{ui}} = \left\| \begin{bmatrix} \sigma_{k+1} & & & \\ & \ddots & & \\ & & \sigma_r & \\ & & & 0 \\ & & & & \ddots \\ & & & & & 0 \end{bmatrix} \right\|_{\text{ui}} = \sigma_{k+1}(\mathbf{A}).$$

因此, 结论成立。 □

我们还可以证明, 在  $\|\cdot\|_2$  和  $\|\cdot\|_F$  下, Eckart Young 定理的最优解都是唯一的, 这是 Weyl 不等式和 Hoffman–Wielandt 不等式的直接推论。但在一般的酉不变范数下, 最优解不一定是唯一的。

Eckart Young 定理告诉我们, 线性变换  $\mathbf{A}$  在哪些方向上的作用是主要的、哪些方向上是次要的: 各个方向的重要性由奇异值的大小来决定。通过保留最大的  $k$  个奇异值及其对应的奇异向量, 我们可以得到一个秩为  $k$  的矩阵, 这个秩  $k$  矩阵正是对原矩阵的最佳近似。这也是主成分分析 (PCA) 的数学基础。

## 2 经典角

本小节限于有限维实内积空间, 即欧氏空间。

### 2.1 经典角

在前一节中, 我们通过变分原理深入研究了奇异值的极值特征。这告诉我们, 一个矩阵的奇异值刻画了线性变换在不同方向上的伸缩能力。现在, 我们进一步探讨两个子空间之间的关系。让我们暂时忘掉线性变换, 回归到一个更基础的几何问题: 如何度量两个子空间在空间中的接近程度? 例

如，给定  $\mathbb{R}^n$  的两个子空间  $\mathcal{U}$  和  $\mathcal{V}$ ，我们能否像定义两条直线的夹角一样，精确地定义这两个子空间的夹角？

任何一个子空间都可以由其一组标准正交基的列向量张成。设  $\mathbf{U} \in \mathbb{R}^{n \times p}$ ,  $\mathbf{V} \in \mathbb{R}^{n \times q}$  的列向量分别是  $\mathcal{U}, \mathcal{V}$  的标准正交基。那么，研究这两个子空间的关系，就自然转化为研究这两个列空间的关系。

思路到这里似乎陷入了停滞。我们不妨回顾一下我们是如何定义两个向量的夹角的。设有两个单位向量  $\mathbf{u}, \mathbf{v}$ ，那么，这两个向量的夹角定义为

$$\angle(\mathbf{u}, \mathbf{v}) = \arccos(\langle \mathbf{u}, \mathbf{v} \rangle).$$

再来看直线和平面的夹角。由立体几何知识我们知道，直线和平面的夹角可以通过直线与平面法向量的夹角来定义——它是直线与平面法向量夹角的补角。我们还知道，线面角是所有线线角中的最小角。

再看平面和平面的夹角。平面和平面的夹角就是法向量的夹角。二面角是所有线面角中的最大角。

这些夹角的定义都离不开“最大”和“最小”这两个关键词。因此，对于一般的子空间  $\mathcal{U}$  和  $\mathcal{V}$ ，我们似乎也要去捕捉它们之间的某种“最大角”和“最小角”，来定义它们之间的接近程度。

考虑所有可能的单位向量对  $\mathbf{u} \in \mathcal{U}$  和  $\mathbf{v} \in \mathcal{V}$ ，它们的夹角的余弦是  $\mathbf{u}^T \mathbf{v}$ 。那么，这些余弦的极值会出现在哪里呢？我们似乎又没了思路。再来回看一下上面三个角——它们都是通过逐层递进的极值来定义的。那么，我们是否也可以通过**逐层最优的对齐**来定义子空间之间的夹角呢？

1. 首先，在两个子空间中分别寻找一对单位向量，使得它们的夹角  $\theta_1$  最小。这对  $\mathbf{u}_1, \mathbf{v}_1$  代表了两个子空间最对齐的方向。
2. 接着，在与已有向量正交的约束下，再次寻找一对单位向量，使其夹角  $\theta_2$  最小。这对  $\mathbf{u}_2, \mathbf{v}_2$  代表了在剩余空间中最对齐的方向。
3. 重复这个过程，直到无法再找到新的向量对为止。由于每次选取都要求与之前的所有方向正交，我们最终能得到  $\min\{p, q\}$  个角度。

这些按递增顺序排列的角度  $0 \leq \theta_1 \leq \theta_2 \leq \dots \leq \theta_{\min\{p, q\}} \leq \frac{\pi}{2}$  就被称为经典角。

这样找角度的方式比较复杂，我们能否找到一种更简洁的数学表达式呢？眼力比较好的读者可能已经注意到了，奇异值也是通过**递归对齐**的方式来定义的。根据 Courant–Fischer min-max 定理，

$$\sigma_i(\mathbf{A}) = \max_{\dim V=i} \min_{\mathbf{x} \in V \setminus \{\mathbf{0}\}} \frac{\|\mathbf{A}\mathbf{x}\|_2}{\|\mathbf{x}\|_2}.$$

我们先在全空间中找到一个方向，使得线性变换  $\mathbf{A}$  在这个方向上的伸缩因子最大，记这个伸缩因子为  $\sigma_1(\mathbf{A})$ ；然后在与这个方向正交的子空间中，找到一个新的方向，使得线性变换  $\mathbf{A}$  在这个新



方向上的伸缩因子最大，记伸缩因子为  $\sigma_2(\mathbf{A})$ ；如此反复，直到找到所有的奇异值。奇异值和经典角在定义的思想上是统一的，它们也必然有着千丝万缕的联系。

事实上，假如我们设  $\mathbf{U} \in \mathbb{R}^{n \times p}$  和  $\mathbf{V} \in \mathbb{R}^{n \times q}$  分别是子空间  $\mathcal{U}$  和  $\mathcal{V}$  的标准正交基。我们考虑矩阵  $\mathbf{G} = \mathbf{U}^T \mathbf{V}$ ，它的元素是不同基向量之间的内积。由 Courant–Fischer min-max 定理， $\cos \theta_1, \dots, \cos \theta_{\min\{p,q\}}$  正好是矩阵  $\mathbf{G}$  的奇异值！

**定义 2.1** (经典角). 若  $\mathcal{X}$  和  $\mathcal{Y}$  是欧氏空间  $\mathcal{V}$  的非零子空间，且  $\{\mathbf{x}_1, \dots, \mathbf{x}_m\}$  和  $\{\mathbf{y}_1, \dots, \mathbf{y}_n\}$  分别是  $\mathcal{X}$  和  $\mathcal{Y}$  的标准正交基。定义以  $\langle \mathbf{x}_i, \mathbf{y}_j \rangle$  为  $(i, j)$  元素的矩阵  $\mathbf{A} \in \mathbb{R}^{m \times n}$ ，其奇异值按升序排列为

$$\sigma_1 \leq \sigma_2 \leq \dots \leq \sigma_{\min\{m,n\}},$$

则称  $\angle(\mathcal{X}, \mathcal{Y}) = \arccos \sigma_k$  为  $\mathcal{X}$  和  $\mathcal{Y}$  之间的第  $k$  个经典角。

带着这个定义，我们再来回看我们之前提到的线线角、线面角和二面角。直线与直线、直线与平面都只有一个经典角，而平面与平面则有两个经典角，最大的那个经典角正是我们之前定义的二面角。

一般地，若  $\dim \mathcal{X} = \dim \mathcal{Y} = p$ ，我们能得到  $p$  个经典角。这些角完整地描述了两个子空间在旋转意义上的差异。若所有经典角都为零，则两个子空间重合；若所有经典角都为  $\frac{\pi}{2}$ ，则子空间正交。

若  $p = \dim \mathcal{X} < \dim \mathcal{Y} = q$ ，我们能够得到  $p$  个经典角。这些角描述了较低维的子空间  $\mathcal{X}$  与较高维的子空间  $\mathcal{Y}$  中的一个  $p$  维子空间之间的对齐情况。这个  $p$  维子空间正是  $\mathcal{Y}$  中与  $\mathcal{X}$  最接近的子空间。事实上，经典角的递归构造过程实际上在  $\mathcal{Y}$  中自然选择了一个  $p$  维子空间  $\mathcal{Y}_p$ ，使得  $\mathcal{X}$  与  $\mathcal{Y}_p$  有着最优的对齐关系。这个子空间由递归定义中得到的向量  $\{\mathbf{y}_1, \dots, \mathbf{y}_p\}$  张成，而其在  $\mathcal{Y}$  中的正交补空间  $\mathcal{Y}_p^\perp$  也会与  $\mathcal{X}$  正交。

经典角还有如下几条性质。由于其证明比较简单，这里就不展开了。

1.  $\mathcal{X}$  与  $\mathcal{Y}$  之间的经典角与  $\mathcal{X}$  和  $\mathcal{Y}$  的标准正交基的取法无关。
2.  $\angle_k(\mathcal{X}, \mathcal{Y}) = \angle_k(\mathcal{Y}, \mathcal{X})$ .
3.  $\angle_k(\mathcal{X}, \mathcal{X}^\perp) = \frac{\pi}{2}$ .
4. 若  $\mathcal{X} \subset \mathcal{Y}$ ，则  $\angle_k(\mathcal{X}, \mathcal{Y}) = 0$ .

## 2.2 经典角在收敛性证明中的应用

经典角在数值线性代数中有着重要的应用，特别是在迭代方法的收敛性分析中。许多迭代方法都是通过不断调整子空间来逼近目标解，而经典角则提供了一种度量这些子空间之间接近程度的工具。例如，在求解特征值问题的乘幂法中，我们可以通过分析迭代过程中子空间与目标特征向量所

在子空间之间的经典角来评估收敛速度。经典角越小，说明子空间越接近目标子空间，迭代过程的收敛性也就越好。

具体地，我们需要证明  $\lim_{k \rightarrow \infty} \angle(\mathbf{A}^k \mathbf{x}_0, \mathbf{q}_1) = 0$ ，其中  $\mathbf{q}_1$  是方阵  $\mathbf{A}$  的主特征向量（最大特征值对应的特征向量）。为简单起见，我们不妨设  $\mathbf{A}$  可对角化，且特征值按模长递减排列： $|\lambda_1| > |\lambda_2| \geq \dots \geq |\lambda_n|$ ，对应的特征向量基为  $\{\mathbf{q}_1, \dots, \mathbf{q}_n\}$ 。设初始向量  $\mathbf{x}_0$  在该基下的展开为

$$\mathbf{x}_0 = \sum_{i=1}^n \mathbf{q}_i \alpha_i = \mathbf{q}_1 \alpha_1 + \mathbf{p}_2 \beta_1, \quad \alpha_1 \neq 0.$$

经过  $k$  步迭代后，有

$$\mathbf{x}_k = \mathbf{q}_1 \lambda_1^k \alpha_1 + \mathbf{p}_2 \beta_k = \lambda_1^k (\mathbf{q}_1 \alpha_1 + \mathbf{p}_2 \frac{\beta_k}{\lambda_1^k}).$$

因此

$$\tan \theta_k = \left| \frac{\beta_k}{\lambda_1^k \alpha_1} \right| \rightarrow 0, \quad \text{as } k \rightarrow \infty.$$

因此乘幂法最后能够收敛到主特征向量所在的方向上。我们只要在每一步做归一化操作，防止向量范数爆炸，让最后的向量保持单位长度，就能得到主特征向量的一个很好的近似。

### 3 仿射空间

我们知道，任何一个向量空间都必须包含零向量，并且对向量的加法和数乘运算封闭。然而，在实际应用中，我们经常会遇到一些不满足这些条件的集合。例如，平面、直线以及更一般的、我们接下来要介绍的仿射空间，它们并不包含零向量，但在几何和工程问题中却非常重要。为了更好地理解这些几何对象，我们需要引入仿射空间的概念。

那么，我们该如何给它们下定义呢？一种很好的定义方式是通过平移来描述。我们可以把上面这些对象看成是某个向量空间经过平移得到的结果。具体来说，设  $\mathcal{V}$  是一个向量空间， $\mathbf{a} \in \mathcal{V}$  是一个固定的点，那么集合

$$\mathcal{A} = \{\mathbf{a} + \mathbf{v} : \mathbf{v} \in \mathcal{V}\}$$

就是一个仿射空间。这个集合中的每个元素都可以看作是向量空间  $\mathcal{V}$  中的一个向量  $\mathbf{v}$  加上一个固定点  $\mathbf{a}$  的结果。通过这种方式，我们可以将向量空间的结构与平移操作结合起来，从而得到更丰富的几何对象。

上面这个定义中，我们涉及两个概念：点和向量。那么，什么叫做点，什么叫做向量呢？从代数的角度来看，点和向量没有本质区别——它们本质上都是向量。然而，在几何的语境下，点和向量有着不同的含义。

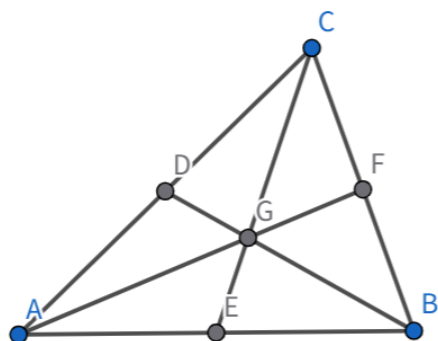


图 3: 点与向量

如上图, 我们模糊地知道  $A, B, \dots, G$  都是点, 而  $\overrightarrow{AB}, \overrightarrow{BC}, \dots$  都是向量。这些点和向量其实只需要三个基本点  $A, B, C$  就能完整地描述:

$$\text{点: } D = \frac{1}{2}A + \frac{1}{2}C$$

$$\text{点: } E = \frac{1}{2}A + \frac{1}{2}B$$

$$\text{点: } F = \frac{1}{2}B + \frac{1}{2}C$$

$$\text{点: } G = \frac{1}{3}A + \frac{1}{3}B + \frac{1}{3}C$$

$$\text{向量: } \overrightarrow{AB} = B - A$$

$$\text{向量: } \overrightarrow{BC} = C - B$$

$$\text{向量: } \overrightarrow{AC} = C - A$$

可以看到, 点和向量都是  $A, B, C$  的线性组合——点的系数和为 1, 而向量的系数和为 0. 这就是点和向量的本质区别。下面, 我们给第一种线性组合一个专门的名字, 它叫做仿射组合。

**定义 3.1** (仿射组合). 设  $x_1, x_2, \dots, x_k \in \mathbb{R}^n$ , 若存在  $\alpha_1, \alpha_2, \dots, \alpha_k \in \mathbb{R}$ , 使得

$$\sum_{i=1}^k \alpha_i = 1,$$

则称向量

$$y = \sum_{i=1}^k \alpha_i x_i$$

为向量  $x_1, x_2, \dots, x_k$  的仿射组合。

由此可见，点的仿射组合仍然是点。点加上向量仍然是点——这是因为向量可以看成是点的系数为零的线性组合，从而不影响系数和为一的性质。在接下来的讨论中，我们只允许出现以下几种运算：

1. 点与点之间的仿射组合——生成点；
2. 点与向量之间的加法——生成点；
3. 向量与向量之间的线性组合——生成向量。

这几种运算都有着鲜明的几何意义。通过这些运算，我们可以构造出各种各样的几何对象，例如直线、平面以及更一般的仿射空间。这些运算也就告诉我们：仿射空间不是向量的集合，而是点的集合——因为仿射空间的定义是点加上向量，得到的还是点。

仿射组合中有一类特殊的组合，它叫做凸组合。

**定义 3.2** (凸组合). 设  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k \in \mathbb{R}^n$ ，若存在  $\alpha_1, \alpha_2, \dots, \alpha_k \in \mathbb{R}$ ，使得

$$\sum_{i=1}^k \alpha_i = 1, \quad \alpha_i \geq 0, \quad i = 1, 2, \dots, k,$$

则称向量

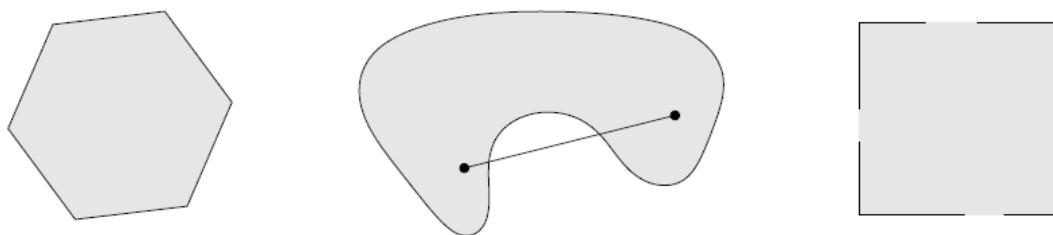
$$\mathbf{y} = \sum_{i=1}^k \alpha_i \mathbf{x}_i$$

为向量  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k$  的凸组合。

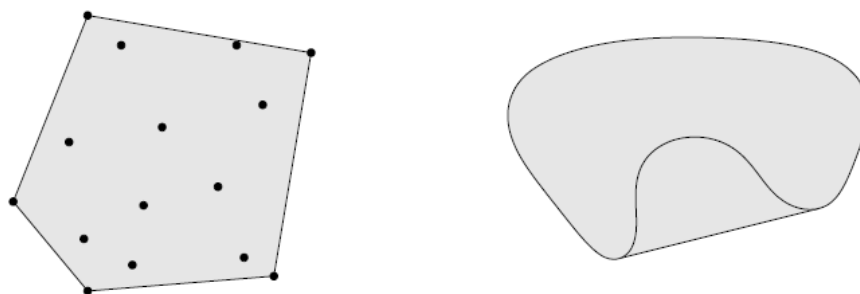
凸组合在仿射组合的基础上，进一步要求系数非负。从几何角度来看，若干个点的凸组合得到的点，必然位于它们的某种意义上的“内部”。那么这个“内部”是相对于什么而言的呢？

**定义 3.3** (凸集). 若  $\Omega \subset \mathbb{R}^n$ ，且对任何  $\mathbf{x}, \mathbf{y} \in \Omega$  以及  $\alpha \in [0, 1]$  都有  $\alpha\mathbf{x} + (1 - \alpha)\mathbf{y} \in \Omega$ ，则称集合  $\Omega$  为凸集。

这个定义告诉我们，凸集中的任意两点之间的线段都完全包含在该集合内。换句话说，若我们从集合中的一个点出发，朝着另一个点前进，那么在前进的过程中，我们始终不会离开这个集合。如果用更数学化的语言来说，凸集是对凸组合封闭的集合。



**Figure 2.2** Some simple convex and nonconvex sets. *Left.* The hexagon, which includes its boundary (shown darker), is convex. *Middle.* The kidney shaped set is not convex, since the line segment between the two points in the set shown as dots is not contained in the set. *Right.* The square contains some boundary points but not others, and is not convex.



**Figure 2.3** The convex hulls of two sets in  $\mathbb{R}^2$ . *Left.* The convex hull of a set of fifteen points (shown as dots) is the pentagon (shown shaded). *Right.* The convex hull of the kidney shaped set in figure 2.2 is the shaded set.

图 4: 凸集与非凸集

凸

不是 凸集!

— Prof. 5eb0fe

**定义 3.4 (凸包).** 对于  $\mathcal{X} \subset \mathbb{R}^n$ , 包含  $\mathcal{X}$  的最小凸集称为  $\mathcal{X}$  的凸包 (*convex hull*), 记为  $\text{Conv}(\mathcal{X})$ 。

根据定义, 我们知道若干个点的凸组合, 一定位于这些点的凸包内。换句话说, 凸包是由这些点通过凸组合所能生成的所有点的集合。

凸组合的概念可以推广到无穷级数、积分，乃至概率分布等更广泛的情形中。假设  $(\theta_i)_{i \in \mathbb{N}}$  满足

$$\theta_i \geq 0, \quad \sum_{i=1}^{\infty} \theta_i = 1,$$

设  $\mathbf{x}_1, \mathbf{x}_2, \dots \in C \subset \mathbb{R}^n$ ，且  $C$  是凸集，则若级数  $\sum_{i=1}^{\infty} \theta_i \mathbf{x}_i$  存在，有

$$\sum_{i=1}^{\infty} \theta_i \mathbf{x}_i \in C.$$

更一般地，设  $p: \mathbb{R}^n \rightarrow \mathbb{R}$  满足  $p(\mathbf{x}) \geq 0$  对任意的  $\mathbf{x} \in C$  恒成立，且  $\int_C p(\mathbf{x}) d\mathbf{x} = 1$ ，其中  $C \subset \mathbb{R}^n$  是一个凸集，则若积分  $\int_C p(\mathbf{x}) \mathbf{x} d\mathbf{x}$  存在，有

$$\int_C p(\mathbf{x}) \mathbf{x} d\mathbf{x} \in C.$$

最一般的情形是：设  $C \subset \mathbb{R}^n$  是一个凸集， $\mathbf{x}$  是一个随机向量且  $\mathbf{x} \in C$ , a.e.，则有  $\mathbb{E}[\mathbf{x}] \in C$ 。事实上，这种情形包含了以上所有情况。例如，若  $\mathbf{x}$  服从离散分布，且  $\mathbb{P}(\mathbf{x} = \mathbf{x}_i) = \theta_i$ ，则

$$\mathbb{E}[\mathbf{x}] = \sum_{i=1}^{\infty} \theta_i \mathbf{x}_i \in C.$$

这就回到了前面的级数情形。

我们再来解释一个概念。什么叫做顶点 (vertex) 呢？

**定义 3.5 (顶点).** 设  $\Omega$  是一个凸集， $\mathbf{v} \in \Omega$  不能表示成其他点的凸组合，则称  $\mathbf{v}$  为  $\Omega$  的一个顶点。

顶点是凸集中的“极端”点，不能通过其他点的凸组合来表示。换句话说，顶点是构成凸集的基本元素。对于一个多边形来说，其顶点就是多边形的角点；对于一个多面体来说，其顶点就是多面体的顶点；对于一个圆盘来说，其顶点就是圆周上的点。

定义完仿射组合和凸组合后，我们再来回看仿射空间的定义。定义告诉我们，如果我们在仿射空间  $\mathcal{A}$  中取定一个点  $\mathbf{u}$ ，任取  $\mathbf{v} \in \mathcal{A}$ ，集合  $\{\mathbf{v} - \mathbf{u} : \mathbf{v} \in \mathcal{A}\}$  构成了一个向量空间。而向量空间是对加法和数乘封闭的，因此对任意  $\mathbf{x}, \mathbf{y} \in \mathcal{A}$  以及  $\alpha \in \mathbb{R}$ ，都有

$$\alpha(\mathbf{x} - \mathbf{u}) + (1 - \alpha)(\mathbf{y} - \mathbf{u}) + \mathbf{u} = \alpha\mathbf{x} + (1 - \alpha)\mathbf{y} \in \mathcal{A}.$$

换句话说，仿射空间对仿射组合封闭。那么，反过来呢？如果一个集合对仿射组合封闭，它会是一个仿射空间吗？我们不妨设  $\mathcal{A}$  对仿射组合封闭，即对任意  $\mathbf{x}, \mathbf{y} \in \mathcal{A}$  以及  $\alpha \in \mathbb{R}$  都有  $\alpha\mathbf{x} + (1 - \alpha)\mathbf{y} \in \mathcal{A}$ 。两边同时减去  $\mathbf{y}$  并整理，有

$$\alpha(\mathbf{x} - \mathbf{y}) \in \{\mathbf{v} - \mathbf{y} : \mathbf{v} \in \mathcal{A}\}.$$

现在我们只需证明集合  $\{\mathbf{v} - \mathbf{y} : \mathbf{v} \in \mathcal{A}\}$  是一个向量空间即可。对数乘封闭是显然的，我们只需证它对加法封闭。设  $\mathbf{a}, \mathbf{b} \in \{\mathbf{v} - \mathbf{y} : \mathbf{v} \in \mathcal{A}\}$ ，则存在  $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{A}$ ，使得

$$\mathbf{a} = \mathbf{x}_1 - \mathbf{y}, \quad \mathbf{b} = \mathbf{x}_2 - \mathbf{y}.$$

则对任意  $\alpha, \beta \in \mathbb{R}$ ，有

$$\begin{aligned} \alpha \mathbf{a} + \beta \mathbf{b} &= \alpha(\mathbf{x}_1 - \mathbf{y}) + \beta(\mathbf{x}_2 - \mathbf{y}) = \alpha \mathbf{x}_1 + \beta \mathbf{x}_2 - (\alpha + \beta)\mathbf{y} \\ &= (\alpha + \beta) \left( \frac{\alpha}{\alpha + \beta} \mathbf{x}_1 + \frac{\beta}{\alpha + \beta} \mathbf{x}_2 - \mathbf{y} \right) \in \{\mathbf{v} - \mathbf{y} : \mathbf{v} \in \mathcal{A}\}. \end{aligned}$$

因此，集合  $\{\mathbf{v} - \mathbf{y} : \mathbf{v} \in \mathcal{A}\}$  对加法封闭。综上所述，我们得到了仿射空间的等价定义：

**定义 3.6** (仿射空间). 若  $\mathcal{A}$  是  $\mathbb{R}^n$  的非空子集，且对任意  $\mathbf{x}, \mathbf{y} \in \mathcal{A}$  以及  $\alpha \in \mathbb{R}$  都有  $\alpha \mathbf{x} + (1 - \alpha)\mathbf{y} \in \mathcal{A}$ ，则称  $\mathcal{A}$  为仿射空间。