Matrix Analysis–part one

毛海涛　傅宇千　徐晓豪　赵睿思　葛诗然　郑雨嘉　王博涵

更新：2022 年 2 月 4 日

目录

1 课程笔记

1.1 lecture 1: The Column Space of A Contains All Vectors AX

1.1.1 Ax 的理解

$$Ax=\begin{bmatrix}2&1&3\\3&1&4\\5&7&12\end{bmatrix}\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix}=x_1\begin{bmatrix}2\\3\\5\end{bmatrix}+x_2\begin{bmatrix}1\\1\\7\end{bmatrix}+x_3\begin{bmatrix}3\\4\\12\end{bmatrix}=\begin{bmatrix}2\cdot x_1+1\cdot x_2+3\cdot x_3\\3\cdot x_1+1\cdot x_2+4\cdot x_3\\5\cdot x_1+7\cdot x_2+12\cdot x_3\end{bmatrix}$$

Ax 的两种理解方式：

- 中间的等式是对其做 combination of column A，即向量的线性组合
- 右边仅为我们通常的计算过程。

1.1.2 column space 和 row space

当我们取遍 x 时，通过 Ax 计算即可得到 A 的 column space，在 R_3 中根据 A 的不同可以形成 3D, plane 或 line，上例中的 A 的 column space 为 plane（可以直观看出 A 的第三列为前两列的和）。同理 row space 则是矩阵的行空间。

1.1.3 rank

$$column\quad rank=row\quad rank=column/row\quad space\quad dimension=basis\quad number$$

这个等式针对于数值关系。column 和 row 之间的关系可以用转置来理解。

1.1.4 independent column 和 basis

independent column 即是 basis。以 A 的 factorization 为例：

$$A=\begin{bmatrix}2&1&3\\3&1&4\\5&7&12\end{bmatrix}=\begin{bmatrix}2&1\\3&1\\5&7\end{bmatrix}\begin{bmatrix}1&0&1\\0&1&1\end{bmatrix}=C\cdot R$$

从 column 和 row 的角度我们都可以找到其对应的 basis。MAO: 其中 R 并非为标准的 basis，basis 应该相互正交，为了保证正交，一般的矩阵分解都会写成如下的形式 $C\Lambda R$

1.1.5 看第一节和第二节补充

$$A=\begin{bmatrix}2&1&3\\3&1&4\\5&7&12\end{bmatrix}=\begin{bmatrix}2&1\\3&1\\5&7\end{bmatrix}\begin{bmatrix}1&0&1\\0&1&1\end{bmatrix}$$

列空间：由 A 的列向量线性组合张成的空间行空间：由 A 的行向量线性组合张成的空间直观来看，A 的第三列由由前两列线性组合而成，而第一列，第二列之间无法互相转化，因此 A 的列空间有两个基 (basis)，列秩为 2. CR 分解结果从左向右看，是 C 的两个相互独立的列组成了 A 的列，从而可以从 C 的列向量组合出 A 的列向量并张成 A 的列空间；从右向左，由 R 的行可以组合出 A 的所有行，并张成 A 的行空间，且 R 的两个行向量不能互相转化，而由矩阵乘法，矩阵的 shape 必须满足 $(m*n)*(n*p)$, 因此行秩等于列秩，并且矩阵的秩等于其转置的秩。

关于正交补空间: 假设 A 为 $(m*n)$ 的矩阵 (假设 $m<n$), 秩为 r, 则 $C(A)=r, C(A^T)=r$,(A 的列空间秩为 r，A 的转置的列空间的秩也为 r)。设 X 和 Y 为 R^n 的子空间，若对每一对 $x\in X$ 和 $y\in Y$ 都满足 x^Ty=0，则称 X 和 Y 是正交的。因此若 Ax=0，则 A^T 的列空间 $C(A)$ 与 A 的零空间 $N(A)$ 正交, $N(A)$ 的秩为 $n-r$，同理 $N(A^T)$=$m-r$。(有助于理解奇异值分解)

1.1.6 本节习题

1.

$$\begin{bmatrix}2&-2&0\\2&-2&0\\2&-2&0\\2&-2&0\end{bmatrix}\begin{bmatrix}1\\1\\0\end{bmatrix}=\begin{bmatrix}0\\0\\0\\0\end{bmatrix}$$

4.

$$x=\begin{bmatrix}1\\-1\\0\end{bmatrix},y=\begin{bmatrix}1\\1\\-2\end{bmatrix}$$

找到三个 independent 的向量的话在 R^3 中就可以组合为任意向量了，但 A 的 space 并非零空间

8. m=3, n>=3, r=3

18.

$$\begin{bmatrix}C\\C\end{bmatrix}\begin{bmatrix}0&R\end{bmatrix}$$

MAO: 1.4 阅读 TODO

1.2 lecture 3: Orthonormal Columns in Q Give $Q^TQ = I$

1.2.1 正交矩阵 (Orthonormal Matrix) 的定义及性质

Definition: 正交矩阵 (Orthonormal Matrix)Q 满足 $Q^TQ = I$.

Property 1: 当 Q 是方阵时，$QQ^T = I$.

Proof 1：对方阵来说左逆等于右逆，故 $QQ^T = QQ^T)^T = I^T = I$.

Property 2: Q 不改变向量的长度.

Proof 2：$(Qx)^TQx = ||Qx||^2 = x^TQ^TQx = x^Tx = ||x||^2, ||Qx|| = ||x||$

1.2.2 Householder reflection (豪斯霍德转换)

Definition：如果 u 为 n 阶单位向量, 即满足 $u^Tu = I$,I 是单位矩阵。则可以构造豪斯霍德矩阵 $H = I2uu^T$ 满足对称正交矩阵性质。

Proof:

$$H^TH = H^2 = I4uu^T + 4uu^Tuu^T = I$$

Example 1：对于二维平面中某一旋转矩阵 Q_1(rotation matrix)

$$Q_1 = \begin{bmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{bmatrix}, Q_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} cos\theta \\ sin\theta \end{bmatrix}, Q_1 \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} -sin\theta \\ cos\theta \end{bmatrix}$$

Example 2：二维平面中某一反射矩阵 Q_2(reflection matrix) 将向量关于直线 $y = cos\frac{\theta}{2}x$ 进行对称.

$$Q_2 = \begin{bmatrix} cos\theta & sin\theta \\ sin\theta & -cos\theta \end{bmatrix}, Q_2 \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} cos\theta \\ sin\theta \end{bmatrix}, Q_2 \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} sin\theta \\ -cos\theta \end{bmatrix}$$

1.2.3 Hadamard matrix(哈达玛矩阵)

Definition：H_2 是正交矩阵，由 1, 1 构成。当 n 是 2 的整数倍时,H_n 仍为 Hadamard matrix，即由 1,-1 构成的正交矩阵，构造方法如下.

$$H_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, H_{2^{n+1}} = \frac{1}{\sqrt{2^{n+1}}} \begin{bmatrix} H_{2^n} & H_{2^n} \\ H_{2^n} & -H_{2^n} \end{bmatrix}$$

Hadamard 变换可以被视为是由大小为 2 的离散傅里叶变换（DFT）构成的，实际上相当于大小为 2^N 的多维 DFT, 它将任意输入向量分解为 Walsh 函数的叠加。

MAO: 能不能补充一下 hadamard 矩阵的意义？我做一下 wavelet 矩阵的用法XU: OK! 简单补充了两句 MAO: 目前 wavelet 仍然 todo 等我多看过几篇 graphwavelet 的文章再说

1.2.4 正交矩阵的特征向量矩阵

Definition：置换矩阵 P（permutation matrix）是正交矩阵且特征向量相互正交.n 阶置换矩阵的特征向量矩阵是 n 阶傅里叶矩阵. 例如下面置换矩阵 P 的特征向量矩阵是 4 阶傅里叶矩阵 F.

$$Q = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}, F = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & i & i^2 & i^3 \\ 1 & i^2 & i^4 & i^6 \\ 1 & i^3 & i^6 & i^9 \end{bmatrix}$$

1.2.5 课后习题

Problem 1：If u and v are orthogonal unit vectors, show that u + v is orthogonal to u - v. What are the lengths of those vectors ?（如果 u 和 v 是正交单位向量，试证明 u+v 和 u-v 正交. 另外 u+v 和 u-v 这两个向量的模长是多少?）

Proof 1：因为 u 和 v 是单位向量，所以 $||u|| = ||v|| = 1, u^2 = v^2, (u+v)(u-v) = 0$, 故 u+v 和 u-v 正交. 因为 u 和 v 是正交单位向量，所以 $||u+v|| = \sqrt{u^2+2uv+v^2} = \sqrt{1+0+1} = \sqrt{2}$,$||u-v|| = \sqrt{u^2-2uv+v^2} = \sqrt{1-0+1} = \sqrt{2}$.

Problem 2：Draw unit vectors u and v that are not orthogonal. Show that $w = v - u(u^Tv)$ is orthogonal to u (and add w to your picture).（画出一组单位向量 u 和 u 使得他们不正交。同时证明 $w = v - u(u^Tv)$ 与 u 正交（将 w 画到图中）.）

Proof 2:（图画法不唯一，略）. $u^Tw = u^T(v-u(u^Tv)) = u^Tv - I(u^Tv) = 0$, 故 $w = v-u(u^Tv)$ 与 u 正交。

Problem 3：Draw any two vectors u and v out from the origin (0, 0). Complete two more sides to make a parallelogram with diagonals $w = u+v$ and $z = u-v$. Show that $w^Tw + z^Tz$ is equal to $2u^Tu + 2v^Tv$.（画出除了原点 (0,0) 以外的两个向量 u 和 v, 再完成两条边 $w = u+v$ and $z = u-v$, 形成一个带对角线的平行四边形. 证明 $w^Tw + z^Tz$ 等于 $2u^Tu + 2v^Tv$.）

Proof 3：$w^Tw + z^Tz = ||w||^2 + ||z||^2 = ||u+v||^2 + ||u-v||^2 = u^Tu + 2u^Tv + v^Tv + u^Tu - 2u^Tv + v^Tv = 2u^Tu + 2v^Tv$.

Problem 4：Key property of every orthogonal matrix: $||Qx||^2 = ||x||^2$ for every vector x. More than this, show that $(Qx)^T(Qy) = x^Ty$ for every vector x and y. So lengths and angles are not changed by Q. Computations with Q never overflow!（对于正交矩阵 Q 的一个重要的性质是 $||Qx||^2 = ||x||^2$. 基于此，请进一步证明对于一组向量 x 和 y, 满足 $(Qx)^T(Qy) = x^Ty$. 也就是说一组向量 x 和 y 的长度和角度都不会被 Q 改变。也就是说关于 Q 的计算不会溢出!）

Proof 4：由 $||Qx||^2 = ||x||^2, ||Qy||^2 = ||y||^2 ||Q(x+y)||^2 = ||(x+y)||^2$，可得: $||Q(x+y)||^2 = ||(x+y)||^2, (||Qx||)^2 + (||Qy||)^2 + 2(Qx)^T(Qy) = ||x||^2 + ||y||^2 + 2x^Ty, (Qx)^T(Qy) = x^Ty$, 得证。

Problem 5: If Q is orthogonal, how do you know that Q is invertible and Q^{-1} is also orthogonal? If $Q_1^T = Q_1^{-1}$ and $Q_2^T = Q_2^{-1}$, show that Q_1Q_2 is also an orthogonal matrix. （如果 Q 是一个正交矩阵，如何证明 Q 可逆同时 Q^{-1} 也是正交矩阵？如果 $Q_1^T = Q_1^{-1}$ 且 $Q_2^T = Q_2^{-1}$，试证明 Q_1Q_2 也是一个正交矩阵.）

Proof 5: 因为 $Q^TQ = I$, 故存在 $Q^{-1} = Q^T$，所以 Q 可逆。

$(Q^{-1})^T(Q^{-1}) = (Q^T)^TQ^T = QQ^T = (Q^TQ)^T = I^T = I$，故 Q^{-1} 是正交矩阵。

$(Q_1Q_2)^T(Q_1Q_2) = (Q_2)^T(Q_1)^T(Q_1Q_2) = (Q_2)^{-1}(Q_1)^{-1}Q_1Q_2 = I$, 故 Q_1Q_2 也是正交矩阵.

Problem 6: A permutation matrix has the same columns as the identity matrix (in some order).Explain why this permutation matrix and every permutation matrix is orthogonal:

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

has orthonormal columns. So P^TP=?, P^{-1}=?. When a matrix is symmetric or orthogonal, it will have orthogonal eigenvectors. This is the most important source of orthogonal vectors in applied mathematics. (置换矩阵与单位矩阵具有相同的列（以某种顺序）。解释为什么此置换矩阵和每个置换矩阵是正交的。如果

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

有正交列，那么 P^TP 和 P^{-1} 分别是多少？当矩阵对称或正交时，它将具有正交的特征向量。这是应用数学中正交向量的最重要来源。）

Proof 6: 由于一个置换矩阵的列构成了 R^n 空间的正交基。考虑正交性, $P^TP = I, P^{-1} = P^T$.

Problem 7: Four eigenvectors of that matrix P are $x_1 = (1,1,1,1)$, $x_2 = (1,i,i^2,i^3)$, $x_3 = (1,i^2,i^4,i^6)$, and $x_4 = (1,i^3,i^6,i^9)$. Multiply P times each vector to find λ_1, λ_2, λ_3, λ_4. The eigenvectors are the columns of the 4 by 4 Fourier matrix F. Show that

$$Q = \frac{F}{2} = \frac{1}{2}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & i & -1 & -i \\ 1 & i^2 & 1 & -1 \\ 1 & i^3 & -1 & i \end{bmatrix}$$

has orthonormal columns: $\bar{Q}^TQ = I$.

（矩阵 P 的四个特征向量是 $x_1 = (1,1,1,1)$, $x_2 = (1,i,i^2,i^3)$, $x_3 = (1,i^2,i^4,i^6)$, 和 $x_4 = (1,i^3,i^6,i^9)$. 我们计算 P 乘上每一个向量来得到 λ_1, λ_2, λ_3, λ_4. 这些特征向量是 4×4 的傅里叶

矩阵 F. 试证明：

$$Q=\frac{F}{2}=\frac{1}{2}\begin{bmatrix}1&1&1&1\\1&i&-1&-i\\1&i^2&1&-1\\1&i^3&-1&i\end{bmatrix}$$

有正交列向量 Q 满足 $\bar{Q}^TQ=I$.)

Proof 7：

$$Q=\begin{bmatrix}1&1&1&1\\1&i&-1&-i\\1&i^2&1&-1\\1&i^3&-1&i\end{bmatrix}=\begin{bmatrix}1&1&1&1\\1&i&i^2&i^3\\1&i^2&i^4&i^6\\1&i^3&i^6&i^9\end{bmatrix}$$

由正交矩阵的特征向量互相正交，可知 $\bar{Q}^TQ=I$.

Problem 8：Haar wavelets are orthogonal vectors (columns of W) using only 1, -1, and 0. n=4:

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

Find W^TW and W^{-1} and the eight Haar wavelets for n = 8.

Proof 8：

n=4 时,

$$W^TW=\begin{bmatrix}4&0&0&0\\0&4&0&0\\0&0&2&0\\0&0&0&2\end{bmatrix}$$

$$W^{-1}=(W^TW)^{-1}W^T=\begin{bmatrix}\frac{1}{4}&0&0&0\\0&\frac{1}{4}&0&0\\0&0&\frac{1}{2}&0\\0&0&0&\frac{1}{2}\end{bmatrix}\begin{bmatrix}1&1&1&-1\\1&1&-1&0\\1&-1&0&0\\1&0&1&-1\end{bmatrix}=\begin{bmatrix}\frac{1}{4}&\frac{1}{4}&\frac{1}{4}&-\frac{1}{4}\\\frac{1}{4}&\frac{1}{4}&-\frac{1}{4}&0\\\frac{1}{2}&-\frac{1}{2}&0&0\\\frac{1}{2}&0&\frac{1}{2}&-\frac{1}{2}\end{bmatrix}$$

n=8 时，

$$W_8 = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & -1 & 0 & 0 & 0 \\ 1 & 1 & -1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & -1 & 0 & 0 & -1 & 0 & 0 \\ 1 & -1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & -1 & 0 & 1 & 0 & 0 & -1 & 0 \\ 1 & -1 & 0 & -1 & 0 & 0 & 0 & 1 \\ 1 & -1 & 0 & -1 & 0 & 0 & 0 & -1 \end{bmatrix}$$

MAO: Q: 为什么要寻找正交矩阵?

A: 正交矩阵可以对坐标系 basis 进行变化，且不会在计算中出现 overflow 问题（每行的模长都是 1）

Q: 如何寻找正交矩阵?

A: 实对称矩阵的特征向量，或者正交矩阵的特征向量

Eigenvalue of $S^T = S$ or $Q^T Q = I$

S 的所有特征值和特征向量都是实数，并且 $S = Q\Lambda Q^T$, Notice that, no inverse matrix here! 但 Q 矩阵的特征向量存在于复数域当中，特征向量做乘法的时候要加上共轭。即：$\bar{col_1} \cdot col_2 = 0$

Wavelet Matrix:

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

Wavelet 作为一组正交基底，第一组为对均值的描述，从第二组开始是 difference in different scale, which is a kind of signal prototype. XU: wavelet matrix 如果作用在一个时序信息或者图像上的话，感觉可以用于不同细粒度下的特征聚合。

1.3 Eigenvalue and Eigenvectors

1.3.1 特征值和特征矩阵的限制、性质

特征值的性质：

- 特征值和特征向量是定义在方阵上的性质（很受限）(Also, an orthogonal matrix can only be defined in square matrix. There are (rectangular) matrices that their columns are orthogonal to each other but not with unitary norm. Unfortunately, there is no name for them.)

- 矩阵的迹等于特征值的和 **目前缺少直观**
- 矩阵的行列式等于特征值的乘积 **目前缺少直观** (The volume of a high-dimensional shape is equal to its determinant. So perhaps one could imagine the volume as the product of eigenvalues, with the combination of the geometric intuition of eigenvalues.)
- 不同特征值对应的特征向量彼此正交。那相同特征值呢? (Does it really matter? I'm not sure.)

现实中如何使用特征值

- 对称矩阵的特征值都是实数（特征向量也都是实数吗?）**但不一定相互正交**
- 特征向量只有在原始矩阵满足 $AA^T = A^TA$, 才满足相互正交的条件
- $e^{At}x = e^{\lambda t}x$ connection with the Fourier transformation

tips: 反对称矩阵对应的特征值是纯复数，一般的矩阵可以分解为对称矩阵 + 反对称矩阵

使用误区

- $A + B$ 的特征值不等于 A 的特征值加 B 的特征值。
- 乘法亦然。主要原因在于不同矩阵的特征向量是不同的，特征值之间不存在这样一一对应的关系
- 重复的特征值不一定对应不同的特征向量

1.3.2 Another way to think

新的概念：相似矩阵（构建特征值和原始矩阵之间的关联性）B is similar to A. As:

$$B = M^{-1}AM \tag{1}$$

And similar matrix means the same eigenvalues.

The eigenvalue matrix also has the similar relation as

$$M^{-1}AMy = \lambda y A(My) = M\lambda y A(My) = \lambda(My) \tag{2}$$

where My is the new eigenvector basis.

再来一个例子吧！证明：$\lambda(AB) = \lambda(BA)$
假设 $M(AB)M^{-1} = BA$
带入 $M = B$ 即可得到结果

MAO: 略去了特征值矩阵和原始矩阵相似的部分

1.3.3 Questions

Q1: 什么叫做矩阵的线性变换：

A1: 参考自 线性代数的本质 (3)--矩阵与线性变换. 矩阵乘法可以被是做一种基向量的线性组合。线性变换具有良好的性质:（1）网格线平行且等距（2）原点不动（3）变化后直线仍然是直线从前面的章节，我们知道，Ax 得到的向量是 A 矩阵列向量的线性求和。换句话来说，我们可以把 Ax 看作 $A \cdot Ix$. 其中 I 矩阵为 x 的原始基底，然后我们可以通过矩阵结合律，转化为 $(AI) \cdot x$。即我们将原始的 $[1, 0, 0, 0]$ 基底转化为一个新的基底,A 矩阵的 col_1，但在每个分量上的长度没有改变。**通常来说，基底的变化会使得原始矩阵的方向和大小都产生变化，除了特征向量**

Q2: Does the origin matrix A come first or the corresponding vector comes up first?

A2: 参考自 矩阵的特征：特征值，特征向量，行列式，trace. 我们是通过矩阵去寻找特征向量。我们将做出以下的解释来帮助大家理。首先，特征向量的英文是 eigenvector, 或者也被成为本征向量，其哲学意义应该是自身的，固有的，也就是对于 A 矩阵独有的一些特性。对于每个矩阵，我们都可以找到这样的向量，这样的向量可以轻易的表达矩阵的变化。**对于一个固定的矩阵 A，一般的向量被 transform 之后都会改变方向，但是对于特征矩阵，则只会改变模长**

$$Av = \lambda v \tag{3}$$

Q3: 特征向量该如何使用?

A3: 简化计算，主要应用于 spectral theorem. 简化计算的过程.

$$A^2 v = A(Av) = A(\lambda v) = \lambda(Av) = \lambda(\lambda v) = (\lambda)^2 v \tag{4}$$

于是乎，所有的向量在和 A 矩阵相乘时，都可以用特征向量简化计算, **有效的简化计算**。In other word, The matrix A controls a system of linear differential equations $\frac{du}{dt} = Au$. power λ^n are changed to exponentials $e^{\lambda t}$

$$\begin{gathered}
w = t_1 v_1 + t_2 v_2 \\
A^2 w = A(Aw) = A\left(A\left(t_1 v_1 + t_2 v_2\right)\right) = A\left(t_1 \lambda_1 v_1 + t_2 \lambda_2 v_2\right) = t_1 \lambda_1^2 v + t_2 \lambda_2^2 v \\
A^{100} w = t_1 \lambda_1^{100} v + t_2 \lambda_2^{100} v
\end{gathered}$$

1.4 Symmetric Positive Definite Matrices

1.4.1 正定矩阵的判定条件

先验条件，必须是对称矩阵

- 所有特征值都是正数
- $\forall x \neq 0$, **能量函数** $x^T S x > 0$
- $S = A^T A$ ，A 为列无关的特征矩阵。
- leading determinants 大于 0
- elimination 矩阵的对角线上的值都大于 0

1.4.2 半正定矩阵的判定条件

半正定矩阵是正定矩阵和非正定矩阵的临界值，也有不错的性质

- 所有特征值都是正数或 0
- $\forall x \neq 0$, **能量函数** $x^T Sx \geq 0$
- $S = A^T A$ ，A 为一个矩阵中线性无关的部分。
- leading determinants 大于等于 0
- elimination 矩阵的对角线上的值都大于等于 0

1.4.3 能量函数和优化

其中能量函数是一切的核心

例：

$$\begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 3 & 4 \\ 5 & 6 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \tag{5}$$

能量函数即为 $S(x, y) = 3x^2 + 6y^2 + 8xy$ 即为一般的损失函数，或者是说凸优化空间的一个形式. 或者说正定矩阵和凸优化的条件是紧密相连的，对于一个凸优化的函数，minimum $\frac{df}{dx} = 0$, 二阶矩为 $\frac{d^2 f}{dx^2} > 0$，也就是**正定矩阵**。可以印证函数只有一个 minima，随机梯度下降肯定能达到！

Q1: 为什么一定要是正定矩阵?

A1: 因为如果不是正定的话，存在负值，优化找不到对应的最小值，即在 0 点也正好达到 loss 的最小值. $(x, y) = (0, 0)$ 即 $E = 0$.

Definition

Lemma

1.5 lecture 6: Singular Value Decomposition

到了大晚上才突然想起今天是周一，但是时间已经 23 点半了，我要睡啦 先随意放个手写版，其实主要看懂 SVD 的证明那块（其实这个要我打出来我也是会懒了的 hhh，顺着视频看下去很容易懂的），然后从几何转化上理解一下（rotation+ 伸缩 +rotation）

MAO: 后面隐含着的 insight: 一个标准球体从一个侧面角度来看一定是一个椭圆，原始的圆形在奇异值的方向上进行了一定的伸展。和 PCA 最大的不同在于，在 SVD 的过程中，会有一部分的特征空间的损失，保留上方空间就是最好的，或者说剩下的都对应的是零空间

MAO: **PCA 的具体应用是什么？**

6: Singular Value Decomposition (SVD)

$S = Q\Lambda Q^T$

$\Downarrow$ rectangular $Ax \neq \lambda x$

$A = U\Sigma V^T$ not orthogonal

left right

singular ~~奇异矩阵~~ ← 没有 inverse

← determinant 为0 (行列式)

eigenvector in left

eigenvector in right

$\Sigma =$ ~~sig~~ singular values $\begin{bmatrix} \sigma_1 & & \\ & \sigma_2 & \\ & & \sigma_3 \end{bmatrix}$ diagonal, but $\sigma_i \geq 0$

- any matrix can be factor

$A^TA = \underbrace{A^T_{n\times m} A_{m\times n}}_{n\times n} = V \Lambda V^T$ ($\Lambda \geq 0$) ← orthogonal

positive definite $\checkmark$

$(A^TA)^T = A^TA$. Symmetric $\checkmark$

$AA^T = \underbrace{U\Lambda U^T}_{m\times m}$ (但和上面有相同eigenvalue，之前证明过)

$\underline{Av_r = \sigma_r u_r}$ (但某些 $Av_{r+1} = 0$) $\Rightarrow AV = U\Sigma \Rightarrow A = U\Sigma V^T$

$A^TA = V\Sigma^T U^T U \Sigma V^T = V(\Sigma^T\Sigma)V^T$

V: eigenvector of A^TA; $\Sigma^T\Sigma \to \sigma^2 =$ eigenvalue of A^TA

$AA^T = U\Sigma V^T V \Sigma^T U^T = U(\Sigma\Sigma^T)U^T$

U: eigenvector of AA^T; $\Sigma\Sigma^T \to \sigma^2 =$ eigenvalue of AA^T

$\begin{bmatrix} 1 & & \\ & 1 & \\ & & 5 \end{bmatrix}$ double

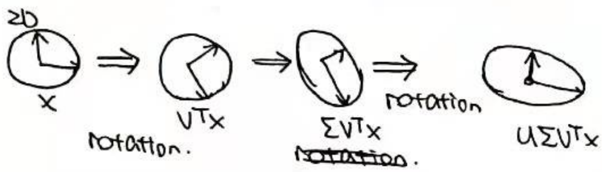
$u_1^T u_2 = 0$ when 找V orthogonal eigenvector of A^TA.

$$\left(\frac{Av_1}{\sigma_1}\right)^T\left(\frac{Av_2}{\sigma_2}\right) = \frac{v_1^T(A^TAv_2)}{\sigma_1\sigma_2} = \frac{v_1^T\sigma_2^2 v_2}{\sigma_1\sigma_2} = \frac{\sigma_2}{\sigma_1} v_1^T v_2 = \text{zero}.$$

这段proof当你找到orth eigen V 一定可得 orth eigen 的 U

$A = U\Sigma V^T$

当乘x时



变换可视化理解.

1.6 lecture 8: Norms of Vectors and Matrices 向量与矩阵的范数

XU: PS：这一节关于范数的基础定义或许非常简单，但太重要了！(广泛应用于各种正则化、度量函数设计、回归模型等领域）

1.6.1 向量范数 (vector norm) 的定义

对于一个向量 $v \in \mathbf{R}^n$,

- p 范数: $l^p = ||v||_p = (\sum_{i=1}^n |v_i|^p)^{\frac{1}{p}}$, 即向量元素绝对值的 p 次方和的 $\frac{1}{p}$ 次幂，表示 x 到零点的 p 阶闵氏距离。
- 0 范数: $l^0 = ||v||_0$，表达的是某向量中非零成分的个数。（当 p 趋于零，可以用这时候的极限来证明）
- 1 范数：$l^1 = ||v||_1 = \sum_{i=1}^n |v_i|$, 即向量元素绝对值之和，x 到零点的曼哈顿距离.
- 2 范数：$l^2 = ||v||_2 = \sqrt{\sum_{i=1}^n |v_i|}$, 也称为 Euclid 范数（欧几里得范数，常用计算向量长度），即向量元素绝对值的平方和再开方，表示 x 到零点的欧式距离。
- (正) 无穷范数，即 ∞-norm: $l^\infty = ||v||_\infty = max|v_i|$, 当 p 趋向于正无穷时，即所有向量元素绝对值中的最大值。
- (负) 无穷范数，即-∞-norm: $l^{-\infty} = ||v||_{-\infty} = min|v_i|$, 当 p 趋向于负无穷时，即所有向量元素绝对值中的最小值。
- S 范数：$||v||_S = \sqrt{v^T S v}$

1.6.2 向量范数的直观可视化

对于二维向量而言：0 范数是坐标轴（除零点），1 范数是菱形，2 范数是圆形，无穷范数是正方形，S 范数在 S ≤ 1 时在二维空间中可直观理解其图像为椭圆 (为什么这里是椭圆可以结合 lecture 6 中矩阵变换进行理解)。

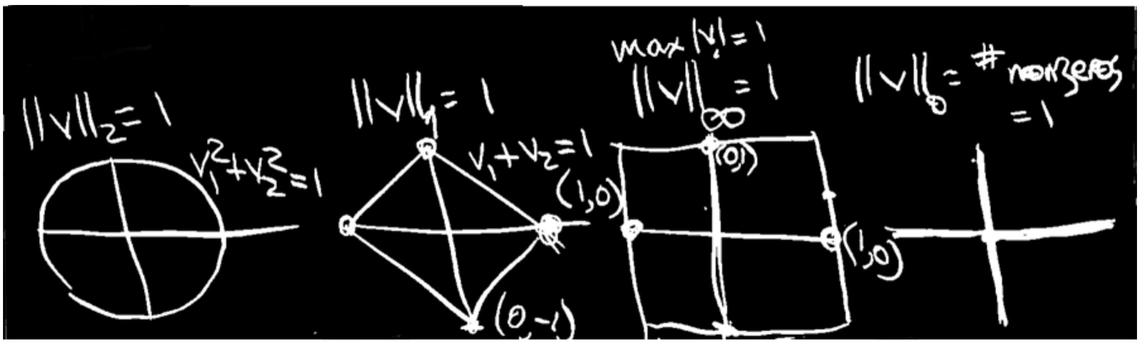


图 1: 向量范数在二维空间的可视化，从左到右为: 2 范数，1 范数，无穷范数和 0 范数。

1.6.3 矩阵范数 (matrix norm) 的定义

对于一个矩阵 $\mathbf{A} \in \mathbf{R}^{m \times n}$,

- 矩阵 **A** 的 2 范数为 $||A||_2 = \sigma_1$，简单的解释为 $||A||_2 = max_x \frac{||Ax||_2}{||x||_2}$。直观地，可以将理解 $||A||_2$ 为对于 $||x||_2$ 在向量空间进行缩放的变换矩阵。
- 矩阵 **A** 的 Frobenius Norm 为：$||A||_F = \sqrt{\sum_{i=1,...,m;j=1,...,n} a_{ij}^2} = \sqrt{\sum_{i=1}^r \sigma_i^2}$，即矩阵元素绝对值的平方和再开平方，另一种计算方式为 $||A||_F = \sqrt{trace(A^TA)}$，这里 $trace()$ 是迹函数。(当矩阵 A 退化为向量时 ($m = 1$ or $n = 1$)，其 Frobenius norm 就是 l^2 范数)
- 矩阵 **A** 的 N-Norm 为 $||A||_N = \sigma_1 + ... + \sigma_r$.

1.7 lecture 9: Four Ways to Solve Least Squares Problems

1.7.1 矩阵的伪逆 A^+

- 如果 A^{-1} 存在, 则 $AA^{-1} = A^{-1}A = I$, $A^+ = A^{-1}$
- 对于不可逆的矩阵: 利用基本空间进行理解, 如2所示, A 乘以列空间中任意一向量 Ax 得到行空间中向量, 这个过程是可逆的, 从而在零空间中应该满足同样的性质（我们在此时引入了伪逆, 即伪逆应该将右零空间投射为左零空间）

  FU: 这种对称的思想在数学概念中的体现, 即通过简单的、可满足的性质对称推广到更宽泛的概念中.

  进一步, 我们对任意矩阵进行 SVD 分解, 通过推广 “矩阵可逆条件下 SVD 分解的逆矩阵” 形式, 我们得到了任意矩阵伪逆的形式:

$$A^+ = V\Sigma^+U^T$$

  此时我们假设 Σ 有且仅有左上角有两个非零元素 σ_1, σ_2, 则 $\Sigma^+\Sigma$ 这两个矩阵相乘后, 有且仅有左上角两个元素为 1.

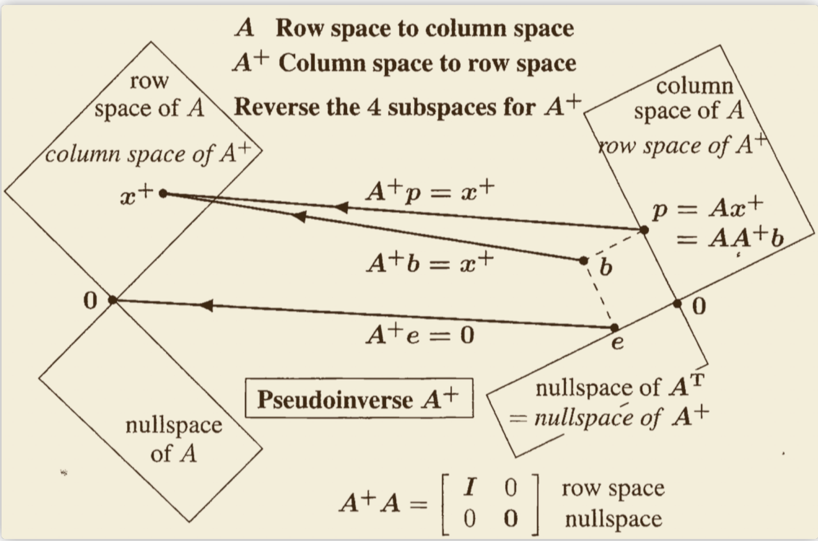


图 2: 伪逆

1.7.2 解决最小二乘问题

Insight: 方程 $Ax = b$ 可能不存在解, 将其转化为误差来获得近似解.

$$\min \|Ax - b\|_2^2 = (Ax - b)^T(Ax - b)$$

求导后, 问题转化为求解:

$$A^T A\hat{x} = A^T b$$

该方法的一种理解: 如果 $Ax = b$ 可能不存在解, 则 b 不在 A 的列空间中, 我们便将 b 投影到 A 的列空间中得到最好的 $\hat{x}$, 如3所示.

$$(Ax)^T(b - A\hat{x}) = x^T A^T(b - A\hat{x}) = 0$$

该式对于任意 x 成立, 所以我们得到

$$A^T(b - A\hat{x}) = 0.$$

即得到了 $A^T A\hat{x} = A^T b$

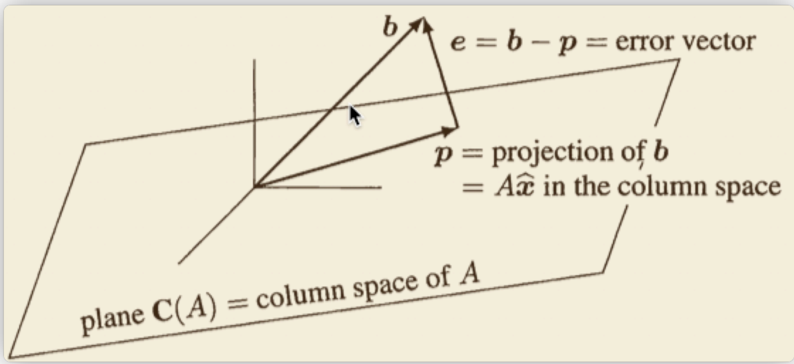


图 3: 将 b 投影 A 的列空间中

方法一: 直接求解上式（但 $A^T A$ 不一定可逆）

方法二: 使用伪逆, $\hat{x} = A^+ b$ (注意有可能 $A^T A$ 不可逆, 但 A 可逆)

方法三: 利用 Gram-Schmidt 正交化, $A = QR$. 代入后得到:$\hat{x} = R^{-1}Q^T b$

方法四: 最小化 $\|b - Ax\|^2 + \sigma^2\|x\|^2$（引入惩罚项）这个在优化理论中被称为 "罚函数方法" , 即引入一个惩罚 $\sigma^2 I$, 将原矩阵构造为可逆矩阵.

FU: 该方法也被称作岭回归, 可见许多概念在不同领域是共通的, 只是换了个叫法而已.

1.8 lecture 10: Survey of Difficulties with $Ax = b$

FU: 这一节课程是对上节课程的补充, 即继续求解 $Ax = b$, 阐述了一些求解中会遇到的问题, 故内容已与上节课程合并.

1.9 lecture 11: Minimizing x Subject to $Ax = b$

1.9.1 最小化 $\|x\|$ 引例

在满足方程 $3x_1 + 4x_2 = 1$ 的条件下，最小化范数 $\|x\|_1, \|x\|_2, \|x\|_\infty$ 从下图可以看出 $\|x\|_p$ 随着 p 从 1 到 2 再到无穷大，交点从 $(1, \frac{1}{4})$ 向右下方移动到 $(\frac{3}{25}, \frac{4}{25})$，后有向右下方移动到 $(\frac{1}{7}, \frac{1}{7})$

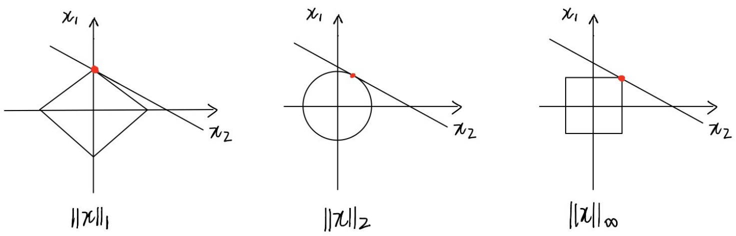


图 4: 约束条件下最小化不同范数的二维状态可视化，从左到右为 1 范数，2 范数和无穷范数

在三维空间中，方程是一个平面。范数的图像变成 3D 菱形，3 球面和 3D 立方体。它们从原点扩张到与平面相交。或者可以有两个约束，两个方程三个未知数，其图像是一条直线。关注在这些不同的情况下或有多少个零点以及 L1 的稀疏度如何。

1.9.2 Gram-Schmidt

Gram-Schmidt 方法回忆: 矩阵 A 的列向量不是正交的，正交化过程设定：

$$A = \begin{bmatrix} | & | & \\ a_1 & a_n & \dots \\ | & | & \end{bmatrix}$$

- q_1 就在 a_1 的方向上，只是做了归一化得到单位向量 $q_1 = \frac{a_1}{\|a_1\|}$。
- q_2 取 a_1 减去 a_2 在 a_1（即 q_1）方向上的投影，得出 $A_2 = a_2 - (a_2^T q_1) q_1$，再取单位向量得到 $q_2 = \frac{A_2}{\|A_2\|}$。
- 同理可依次推出：$A_3 = a_3 - (a_3^T q_1) q_1 - (a_3^T q_2) q_2, q_3 = \frac{A_3}{\|A_3\|}$

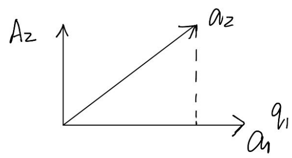


图 5: Gram-Schmidt 中计算 A_2

用 Gram-Schmidt 取 $A = QR$: 矩阵 A 具有 n 个线性无关的列，可以通过 Gram-Schmidt 将矩阵 A 的列向量展开成正交矩阵 Q:

$$A = \begin{bmatrix} | & & | \\ a_1 & \cdots & a_n \\ | & & | \end{bmatrix} \longrightarrow Q = \begin{bmatrix} | & & | \\ q_1 & \cdots & q_n \\ | & & | \end{bmatrix}$$

矩阵 R 的作用是将 Q 的列向量进行线性组合变回矩阵 A 的列向量:

$$R = Q^T A = \begin{bmatrix} - & q_1 & - \\ & \vdots & \\ - & q_1 & - \end{bmatrix} \begin{bmatrix} | & & | \\ a_1 & \dots & a_n \\ | & & | \end{bmatrix}$$

由矩阵乘法的定义，矩阵 R 的第 i 行和第 j 列的元素就是向量 q_i 和 a_j 的内积。

1.9.3 列交换的 Gram-Schmidt

- 前面的处理都按矩阵 A 列向量原始顺序。在进行 LU 分解的时候，如果不做行交换，每次仅找出主元并消掉的其余部分并且重复该过程，则存在风险——主元可能很小，从而给计算带来麻烦。
- 在 Gram-Schmidt 中如果 a2 非常靠近 a1，就会有这种风险。

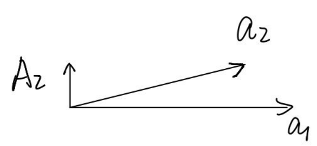


图 6: Gram-Schmidt 中 a2 非常靠近 a1 的情形

- 如果 $A2$ 很小，那么除以其模量产生的舍入误差很大。
- 为了避免这种情况，需要列交换。要做的就是不仅从 a_2 中求正交向量，还要从其余所有 a 中选最大的。

$$A_2 = a_2 - (a_2^T q_1)q_1, a_3 - (a_3^T q_1)q_1, a_n - (a_n^T q_1)q_1$$

- 取其中最大值，然后计算 $q_2 = \frac{A_2}{\|A_2\|}$
- 第三个正交向量是在这两个向量基础上进行操作，也要遍寻剩余所有向量减去两个投影后的最大值。可能表面上看增加了很多运算，但在 Gram-Schmidt 方法中本来就包含这些运算。

1.9.4 Krylov

如果想要求解 $Ax = b$，而 A 是一个较大的矩阵，不能直接用逆矩阵求解。如果矩阵 A 是稀疏矩阵，则 Krylov 是理想的方法。

- 用向量 b 乘上矩阵 A，可以得到一个序列，$b, Ab, A(Ab)......, A^{j-1}b$（注意是用向量乘以矩阵，不是进行矩阵方幂运算）这些向量张成一个空间，即所谓的 Krylov 空间。
- Krylov 空间是一个子空间，求解方程的策略就是在该空间中寻找 $Ax = b$ 的近似解。但是这些向量可能很接近线性相关，因此要取一组更好的基向量，这就需要进行正交化。
- 这就是 Arnoldi 的工作，还有一位匈牙利学者 Lanczos，他们所做的就是如何正交化原基向量。

正交基非常适合映射：

$$x = c_1 q_1 + c_2 q_2 + \cdots + c_n q_n = Qc$$
$$c = Q^{-1} x = Q^T x$$

作 Gram-Schmidt 得到 Krylov 空间的正交基向量：

$$q_1 = \frac{b}{\|b\|} \cdots$$

Arnoldi 所做的适用于任何矩阵，Lanczos 适用于对称矩阵。

1.10 lecture 13: Randomized Matrix Multiplication 矩阵的随机乘法计算加速

XU：这章节思路很直接，就是对于两个矩阵的乘积运算，我们采用一种基于范数（矩阵模长）的大小来选取矩阵中最重要（模长较大）的一些列或者行进行运算的方式来加速

这一讲主要是提出对于两个非常大的矩阵 A 和 B 基于随机采样方式进行乘积加速的方法。具体地，我们对 A 的列向量进行抽样，并对矩阵 B 的相应行向量进行抽样。这里需要取随机抽取概率，然后将抽样结果相加。

$$AB = \begin{bmatrix} | & & | \\ a_1 & \cdots & a_n \\ | & & | \end{bmatrix} \begin{bmatrix} - & b_1^T & - \\ & \vdots & \\ - & b_n^T & - \end{bmatrix} = a_1 b_1^T + \ldots + a_n b_n^T$$

1.10.1 准备知识 (Preliminaries)

- 概率（Probabilities）：我们需要决定抽样概率（相加等于 1）。
- 均值（Mean）：找出随机矩阵乘法 AB 的均值是多少，希望看到随机 AB 的均值就是正确的 AB。
- 方差（Variance）σ^2：方差一定非零，实际上没有任何哪次抽样是完全正确的，可以直接得到矩阵乘积，只有将均值相加时才能得到 AB。我们要选择最佳概率，即了解了系统的工作原理，选择使方差最小化的概率。
- 拉格朗日乘子（Lagrange multiplier）：求函数 f(x1,x2,...) 在 g(x1,x2,...)=0 的约束条件下的极值的方法。其主要思想是引入一个新的参数 λ（即拉格朗日乘子），将约束条件函数与原函数联系到一起，使能配成与变量数量相等的等式方程，从而求出得到原函数极值的各个变量的解。

1.10.2 矩阵随机采样引例

对 1x2 矩阵 $[a, b]$，取两列的概率分别为 p=1$\frac{}{2,\frac{1}{2}s=2} \sum p_j(sample) = \frac{1}{2}(a, 0) + \frac{1}{2}(a, 0) = \frac{1}{2}(a, b)$; 两次随机采样具有相同的均值，故总均值为 (a, b)。

由方差公式 $\sigma^2=\sum p_j(output_j-mean)^2$（或 $\sigma^2=\sum p_j(output_j)^2-mean^2$），我们可以计算得到一次随机抽样的方差 $\sigma^2_{sample}=\frac{1}{2}[(a,0)-(\frac{a}{2},\frac{b}{2})]+\frac{1}{2}[(0,b)-(\frac{a}{2},\frac{b}{2})]=\frac{1}{2}(a^2,b^2)$。

1.10.3 基于范数平方概率的矩阵随机采样方法

思路（Insight）

- 在对于两个非常大的矩阵 A 和 B 基于随机采样方式进行乘积以近似 AB 是，假设 b 比 a 大很多，那么在随机线性代数计算中应该增加选 b 的概率。而前例是取平均随机抽样，即概率是 1/2 和 1/2。我们的努力方向是在保持正确的数学期望的条件下，通过以较高的概率选择重量级元素，来获得更好的方差。
- 对于矩阵运算而言，矩阵中有些列向量的范数较大，可能是更有用的信息，需要得到更高的概率。实际上有两种方法来实现，其一是用列向量的范数平方来加权其概率，下面主要介绍这种方法；另一种方式是用随机数对列向量完成混合，然后对混合的矩阵进行操作。

方法（Method）

- 取 $p_j=||a_j||||b_j^T||/C, C=\sum_1^r||a_j||||b_j^T||$, 其中 $\sum p_j=1$, 即选出矩阵 A 的第 j 列和矩阵 B 的第 j 行的概率与两向量范数乘积成正比。(这里 r 指的是 a 的列数或者 b 的行数)
- 设总抽样次数为 s, 则一次抽样的数学期望值为 $\sum p_j\frac{a_jb_j^T}{sp_j}=\sum\frac{a_jb_j^T}{s}$, 进而 s 次抽样的总和，也就是矩阵乘法 AB 的近似值为 $s\sum p_j\frac{a_jb_j^T}{sp_j}=s\sum\frac{a_jb_j^T}{s}=AB$。
- 我们进一步计算方差，利用方差公式 $\sigma^2=\sum p_j(output_j)^2-mean^2$，可得 $\sigma^2=\frac{1}{s}(C^2-||AB||_F^2)$。

证明（Proof）

- 这里主要利用拉格朗日乘子来证明范数平方概率满足方差的最小化。
- 建立拉格朗日方程 $\mathcal{L}=\sum\frac{||a_j||^2||b_j^T||^2}{sp_j}+\lambda(\sum p_j-1)$, 其中,$\frac{1}{s}||AB||_F^2$ 为常数,$\lambda(\sum p_j-1)=0$。
- 我们对于拉格朗日方程求导，表达式最小值对应于导数为 0，即 $-\frac{||a_j||^2||b_j^T||^2}{sp_j}+\lambda=0$, 可以求得 $p_j=\frac{||a_j||||b_j^T||}{\sqrt{s\lambda}}$, 结合 $(\sum p_j)=1$ 可得 $\sqrt{s\lambda}=C$, 则这组概率就是范数平方概率的表达式, 即得证!

1.11 Low Rank Changes in A and Its Inverse

本节目标：主要研究 A 的变化会怎样影响 A^{-1}

1.11.1 一个公式：

该公式被称作 **matrix inversion lemma** 或者 **Sherman-Morrison-Woodbury formula**

- 在 rank=1 的情况下：(一阶扰动)

$$(I - uv^T)^{-1} = I + \frac{uv^T}{1 - v^T u}$$

是 I 的一种扰动（摄动，perturbation）

- 在 rank=k 的情况下：(k 阶扰动) 求解的关键是改变结合律的顺序

$$(I_n - UV^T)^{-1} = I_n + U(I_k - V^T U)^{-1} V^T$$

- 任意矩阵的扰动:(Sherman-Morrison-Woodbury formula)

$$(A - UV^T)^{-1} = A^{-1} + A^{-1}U(I - V^T A^{-1} U)^{-1} V^T A^{-1}$$

1.11.2 一些本质

为什么这样能简化计算呢，假设矩阵低秩 $m \times 1$，我们本来需要计算的 $uv^T \in m \times m$，在我们求解的过程中使用 $v^T u \in \mathbb{R}$ 进行简化。小矩阵求逆要简单很多。

1.11.3 该公式的应用

- **解决 $(A - uv^T)x = b$ (一阶扰动)**
  MAO：没太理解为什么能求解前面的问题 假设有关 w 的方程 $Aw = b$ 已经求解了，求解 $(A - uv^T)x = b$。我们先求解 $Az = u$，该方程可以和 w 的方程一起求解 (只需要把 w，z 作为同一个矩阵的两列) $A[w; z] = [b; u]$ 然后求解 x:

$$x = w + \frac{wv^T z}{1 - v^T z}$$

- **最小二乘法的新测度**
  新测度： 新的衡量标准，与之前的 $A^T A\hat{x} = A^T b$ 不同，增加了一行/列

$$\begin{bmatrix} A^{\mathrm{T}} & r^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} A \\ r \end{bmatrix} \widehat{x}_{\text{new}} = \begin{bmatrix} A^{\mathrm{T}} & r^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} b \\ b_{m+1} \end{bmatrix}$$

  即：

$$\left[A^{\mathrm{T}} A + r^{\mathrm{T}} r\right] \widehat{x}_{\text{new}} = A^{\mathrm{T}} b + r^{\mathrm{T}} b_{m+1}$$

  含义： 不断更新的最小二乘法（递归最小二乘法），即不断有新数据 (r, b_{m+1}) 产生 (这也是卡尔曼滤波器的思想，但引入了加权最小二乘法：与协方差矩阵、状态方程相关，对原始的递归最小二乘法进行了显著的改善)

1.12 Interlacing Eigenvalues and Low Rank Signals

这节和上节之间存在很大的关联性，所以我们首先要进行一个统一的概括

1.12.1 为什么要研究低秩矩阵？

- 最小的低秩矩阵可以作为构成其他矩阵的基本单元：uv^T 是最小的矩阵
- 世界是通过一个高维空间去观察一个低维流行，SVD 认为其实是由一组低秩矩阵构成的，但不同的低秩矩阵贡献了矩阵的不同部分，小的奇异值对矩阵的贡献很小。
- 即便观测到的是一个秩很高的矩阵，我们通常认为是可能是一个缺失某些值得低秩矩阵，我们会利用这样的假设进行建模（推荐系统）

1.12.2 微观角度观察矩阵变化

之前我们都是观察有效的变化：即 $\delta A = -UV^T$, 现在假设变化范围很小呢？即 dA 对应的是很小的变化 MAO: 这个分子对应的 dt 是一个什么变化？微观和宏观之间存在什么区别

$$B^{-1} - A^{-1} = B^{-1}(A-B)A^{-1}\delta A^{-1} = (A+\delta A)^{-1}(-\delta A)A^{-1}\frac{dA^{-1}}{dt} = A^{-1}\frac{dA}{dt}A^{-1}$$

上述形式其实和 $\frac{1}{x} \rightarrow -\frac{1}{x^2}$ MAO: 微分形式的意义是什么？

另外一个微分！

$$\frac{dA^2}{dt} = A\frac{dA}{dt} + \frac{dA}{dt}A^T \tag{6}$$

1.12.3 从微观角度观察特征值和奇异值的变化

特征值和奇异值的应用差别：特征值多用于和其他矩阵的交互中，奇异值则更关注矩阵本身所具有的性质，比如说模长。

我们把特征值按照微分的形式可以写出来下式

$$A(t)x(t) = \lambda(t)x(t)y^T A(t) = \lambda(t)y^T constraint : y^T x = 1$$

下面的式子对应的是 A 的转置的特征值：A 和 A 的转置特征值相同，但是特征向量应该是不一样，除了对称矩阵然后我们可以得到下式

$$\lambda(t) = y^T(t)A(t)x(t) \tag{7}$$

求导化简可得

$$\frac{d\lambda}{dt} = y^T(t)\frac{d(A)}{d(t)}x(t) \tag{8}$$

MAO: 微观形式的理解可以给我们带来什么帮助呢？

和特征值类似，奇异值的导数也是

$$\frac{d\sigma}{dt} = y^T(t)\frac{d(A)}{d(t)}x(t) \tag{9}$$

1.12.4 从宏观角度的一些理解

交叉理论 S 转化为 $S + u^T u$，特征值分别为: λ 和 γ。有以下的结果

$$\gamma_1 > \lambda_1 > \gamma_2 > \lambda_2 \tag{10}$$

交替进行

一种理解的方式：secular 定理

$$(S + \theta u u^T)v = zv \tag{11}$$

Then

$$\frac{1}{\theta} = \sum_{k=1}^{n} \frac{c_k^2}{z - \lambda_k} \tag{12}$$

永远都是反函数上的点，永远超不过上一个反函数上的点（分区间）

1.13 Questions

- 如何判断一个正定矩阵:1. 求出 A 的所有特征值。若 A 的特征值均为正数，则 A 是正定的;若 A 的特征值均为负数，则 A 为负定的;2. 计算 A 的各阶顺序主子式。若 A 的各阶顺序主子式均大于零，则 A 是正定的；若 A 的各阶主子式中，奇数阶主子式为负，偶数阶为正，则 A 为负定的。
- PCA 的实际应用例：图像高维 feature 的可视化（降维），去噪（de-noising）
- 哈达玛矩阵和小波矩阵在做什么
- 极坐标系和正交变化之间的关系: 一个直观理解：正交变换前后的向量都在极坐标系的同一（高维）球面上，不过向量方向会发生改变。
- 模长和 SVD 的对应关系
- SVD 中左特征向量和右特征向量的关系
- 正交矩阵对任意的 norm 都满足 ||Qx|| = ||X|| 吗
- 卡尔曼滤波是什么?

1.14 Thoughts

- 惩罚项在最优化中的思想
- 近似求解的理解

1.15 Rapidly Decreasing Singular Values

本节循序渐进地讲述了低秩矩阵，低数值秩矩阵，及低秩矩阵的近似。

1.15.1 通过矩阵的奇异值我们能得到什么？

已知: 实矩阵 $X = n \times n$, 该矩阵的唯一奇异值序列 $\sigma_1(x) \geq \sigma_2(x) \geq ... \geq \sigma_n(x) \geq 0$, 我们可以得到以下事实:

若 $\sigma_{k+1}(x) = 0, \sigma_k(x) > 0$:

- $rank(x) = k$, 即矩阵非零奇异值的数量为 k。
- $X = u_1 v_1^T + ... + u_k v_k^T$。通过 SVD 我们知道秩为 k 的矩阵 X 可以分解为 k 个秩 1 矩阵的和, 其中每个秩 1 矩阵都是列向量与列向量转置的乘积。
- $dim(col.space) = dim(row.space) = k$, 即矩阵列空间的维数等于行空间的维数等于 k。

1.15.2 如何定义低秩矩阵？

设想矩阵 X(nxn) 代表一张图片, 矩阵中的每一个元素代表图片的一个像素。如果 X 是一个低秩矩阵, 那么我们有两种方法把图片发送给朋友:

- 发送全部元素， 即发送 n^2 个数据。
- 根据低秩矩阵的性质发送, 发送 k 个秩 1 矩阵, 一个秩 1 矩阵如 $u_1 v_1^T$ 有 2n 个数据, 因此总数据为 2kn 个。

当 $2kn < n^2$, 即 $k < n/2$ 时, 我们称矩阵为低秩矩阵 (low-rank)。在具体实践中, 我们通常要求 $k << n/2$。

1.15.3 低秩矩阵看起来是什么样的？

- 当 rank = 1 时, 矩阵高度对齐
- 三角矩阵不适用于低秩矩阵, 因为三角矩阵的所有奇异值均大于 0
- 矩阵中心有圆形是低秩矩阵 (如日本国旗), 因为我们可以通过分解来限制矩阵的秩。

1.15.4 数值秩矩阵

数值秩和秩很相似, 但是在定义中多了一个摆动空间, 即容忍度 $\epsilon, 0 < \epsilon < 1$。

- 当 $rank_\epsilon(x) = k$ 时,$\sigma_{k+1}(x) \leq \epsilon\sigma_1(x), \sigma_k(x) > \epsilon\sigma_1(x)$
- $rank_0(x) = rank(x)$
- 通过 Eckart-Young 我们知道: 通过奇异值, 我们可以通过低秩矩阵近似得到矩阵 A。
  $\sigma_{k+1}(x) = \parallel X - X_k \parallel_2$
  当我们拥有摆动空间 ϵ 时, 我们得到的是一个近似值而不是准确值。因此我们可以得到如何通过 rank-k 矩阵近似得到矩阵 A, 即 $X_k = best \quad rank - k$。

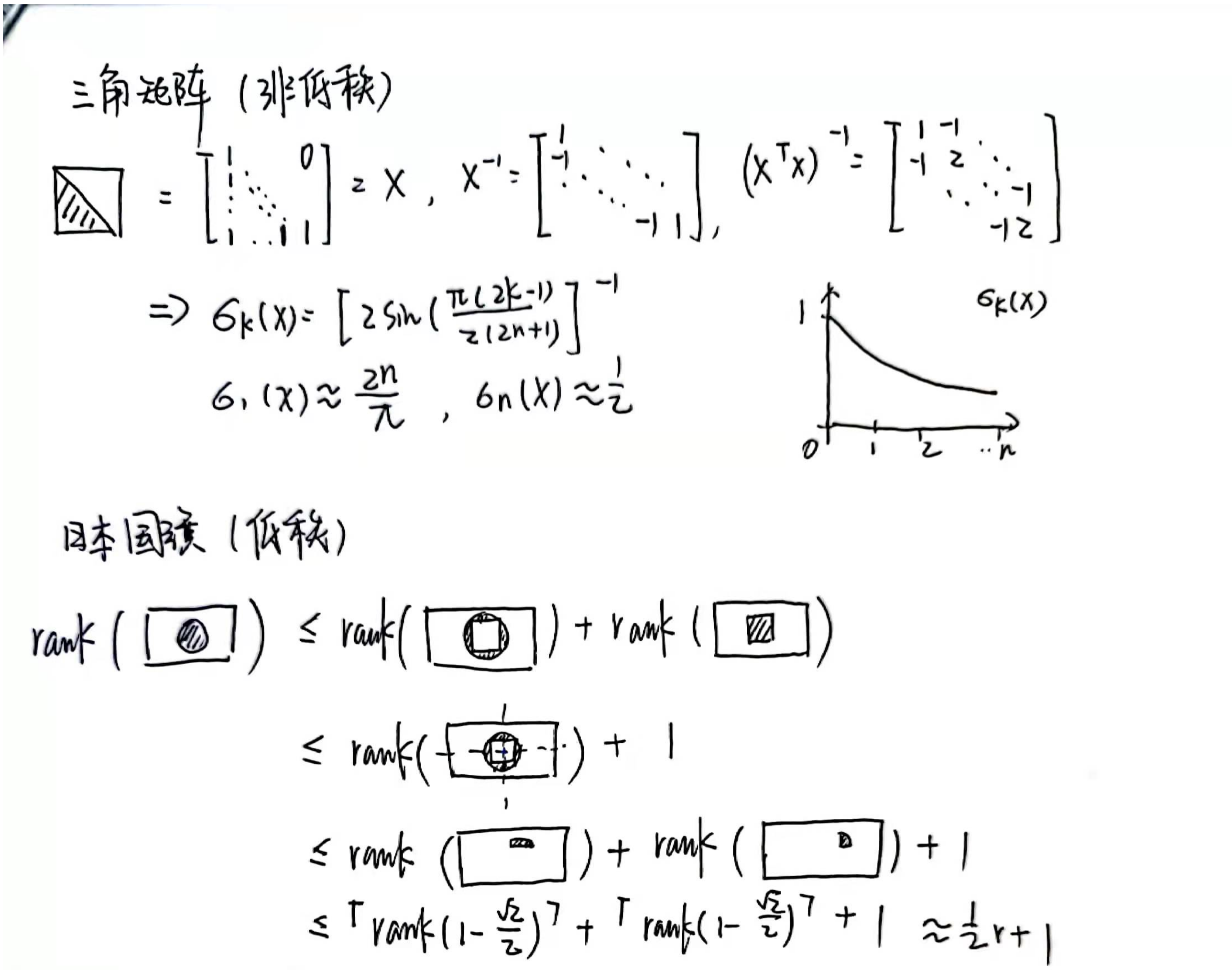


图 7: 三角矩阵与矩阵中央有圆是否低秩的推导

1.15.5 什么样的矩阵是低数值秩的？

- 所有低秩矩阵都是低数值秩矩阵
- 也有许多满秩的, 没有任何奇异值为 0 的矩阵, 它们的奇异值迅速减为 0, 它们是低数值秩的矩阵。

  Hilbert 矩阵:

  $$H_{jk} = \frac{1}{j+k-1}$$

  它是一个满秩但是数值秩很低的矩阵, 这意味着只要给摆动空间 ϵ, 我们就可以通过 rank k 矩阵近似得到 H 矩阵。如果没有摆动空间 $\epsilon(\epsilon = 0)$, 那么此矩阵的所有奇异值均为正数。

  Vandermonde 矩阵 (范德蒙德矩阵):

  $$\begin{bmatrix} 1 & X_1 & \cdots & {X_1}^{n-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_n & \cdots & {X_n}^{n-1} \end{bmatrix}$$

  因为范德蒙德矩阵的数值秩很低, 因此很难求逆。

1.15.6 如何得到低秩矩阵的近似？

关于 x,y 变量的多项式 $p(x,y)=1+x+y$, 矩阵 X 采样于该多项式, 即

$$X_{jk}=p(j,k)=1+j+jk$$

MAO: 如果是三次函数会怎么样呢

分解成矩阵为:

$$\begin{bmatrix}1&1&\cdots&1\\1&1&\cdots&1\\\vdots&\vdots&\ddots&\vdots\\1&1&\cdots&1\end{bmatrix}+\begin{bmatrix}1&1&\cdots&1\\2&2&\cdots&2\\\vdots&\vdots&\ddots&\vdots\\n&n&\cdots&n\end{bmatrix}+\begin{bmatrix}1&2&\cdots&n\\2&4&\cdots&2n\\\vdots&\vdots&\ddots&\vdots\\n&2n&\cdots&n^2\end{bmatrix}$$

易看出第 1,2 个矩阵都只有一个独立列向量, 第三个矩阵可以分解为一个列向量与一个行向量的乘积, 也是秩 1 矩阵, 因此矩阵 X 的秩为 3, 是低秩矩阵。如果 X 来自对一个函数的采样, 那么可以近似该函数, 得到一个低秩的近似, 同时在数值秩上也得到限制。

The world is smooth.

Reade 存在以下论点：如果有两个变量的多项式函数 $p(x,y)=\sum_{s=0}^{m-1}\sum_{t=0}^{m-1}a_{st}x^st^t$，我们可以采样一个矩阵 X，$X_{jk}=p(j,k)$，它会有较低的秩，最多为 m^2 个，$rank(X)\leq m^2$。

以 Hilbert 矩阵为例，Hilbert 矩阵采样于函数 $f(x,y)=\frac{1}{x+y-1}$

$$H_{jk}=\frac{1}{j+k-1}$$

尝试找到 f 函数的多项式近似 p，使得

$$\mid f(x,y)-p(x,y)\mid\leq\frac{\epsilon}{n}\parallel x\parallel_2$$

让 Y 成为 p 的一个采样 $Y_{jk}=p(j,k)$ ，因为 p 是 f 的近似，Y 是 X 的近似，因此 Y 也具有有限的秩，$\parallel x-y\parallel_2\leq\epsilon\parallel x\parallel_2$。如果 X 是从平滑函数中采样得到，我们可以通过多项式近似函数，然后得到多项式秩的近似，矩阵 X 将会有较低的数值秩。

但是 Reade 的论点存在问题，如在 Hilbert 矩阵上效果不好，我们通过多次实践得出

$$rank(H_{1000})=1000, rank\epsilon(H_{1000})=28, \epsilon=10^{-15}$$

而通过 Reade 的理论计算得到的 $rank\epsilon(H_{1000})\leq 719$ 效果不佳。

The world is Sylvester.

Sylverster 方程：

$$AX-XB=C$$

如果想要矩阵 X 的数值秩较低，那么我们需要找到矩阵 A,B,C 满足 Sylverster 方程。

以前文提到的 Hilbert 矩阵为例，我们找到矩阵 A,B,C 满足该方程。

$$\begin{bmatrix} \frac{1}{2} & & & \\ & \frac{3}{2} & & \\ & & \ddots & \\ & & & n-\frac{1}{2} \end{bmatrix} H - H \begin{bmatrix} -\frac{1}{2} & & & \\ & -\frac{3}{2} & & \\ & & \ddots & \\ & & & -n+\frac{1}{2} \end{bmatrix} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}$$

以前文提到的范德蒙德矩阵为例，我们找到矩阵 A,B,C 满足该方程。(* 代表某个数)

$$\begin{bmatrix} x_1 & & & \\ & x_2 & & \\ & & \ddots & \\ & & & x_n \end{bmatrix} V - V \begin{bmatrix} & & & -1 \\ 1 & & & \\ & \ddots & & \\ & & 1 & \end{bmatrix} = \begin{bmatrix} 0 & 0 & \cdots & 0 & * \\ 0 & 0 & \cdots & 0 & * \\ \vdots & \vdots & \ddots & \vdots & * \\ 0 & 0 & \cdots & 0 & * \end{bmatrix}$$

经研究，得到以下结论：

Theorem 如果 A,B 是正规矩阵且矩阵 C 的秩为 r，那么

$$\sigma_{rk+1}(x) \leq Z_k(E,F)\sigma_1(x)$$

其中 Z_K 是 Zolotarev number，E 是 A 的特征值集合，F 是 B 的特征值集合。其中 Zolotarev number 的关键是：E 与 F 的集合是分开的。（如 Hilbert 中 E 取值在 1/2 n-1/2,F 取值在-1/2 -n+1/2 处）而 Hilbert 矩阵的数值秩低正是因为 EF 集合分离使得 Zolotarev number 很小。最终得到的 $rank\epsilon(H_{1000}) = 34$，它比 719 更逼近 28。

1.16 Counting parameters in SVD, LU, QR, Saddle Points

XU：这节课主要是三个比较零散的内容：1.SVD,LU,QR 中的自由参量个数计算；2.saddle point(鞍点) 的计算；3.Rayleigh quotient（瑞利商）的计算。

1.16.1 矩阵不同分解下的自由参数

这里结合各个分解得到的矩阵的相关性质来考虑自由参量个数就比较自然。

- LU:L 是下三角阵，共 $\frac{1}{2}(n-1)n$ 个自由参数；U 是上三角阵，共 $\frac{1}{2}(n+1)n$ 个自由参数。
- QR:Q 的第 k 列共 k 个元素，加上模长为 1 的约束，该列自由参数为 k-1 个, 进而可知对 Q 所有列共有 $\frac{1}{2}(n-1)n$ 个自由参数；R 是上三角阵，共 $\frac{1}{2}(n+1)n$ 个自由参数。
- $X\Lambda X^{-1}$:Λ 共有 n 个自由参数, X 共有 (n-1)n 个。
- $Q\Lambda Q^T$:Λ 共有 n 个自由参数, Q 共有 $\frac{1}{2}(n-1)n$ 个。
- QS:Q 共有 $\frac{1}{2}(n-1)n$ 个自由参数, S 共有 $\frac{1}{2}(n+1)n$ 个。$U_{m\times m}\sum_{m\times n}V_{n\times n}^T$
- (SVD 分解):U 为正交阵, 有 $\frac{1}{2}(m-1)m$ 个,$\sum$ 有 m 个，V 有 $mn-\frac{1}{2}(m+1)m$ 个，故一共 mn 个自由参数。当矩阵的秩为 r 时，一共有 mr+nr-r^2 个自由参量。

1.16.2 约束下的鞍点求解

- 鞍点简介：在直角坐标系中，某个点是一个方向上的极大值，另一个方向上的极小值，则该点是鞍点；在微分方程中，沿着某一方向是稳定的，沿着另一方向不稳定的奇点，就叫鞍点；在泛函分析中，既不是极大值点也不是极小值低点的临界点就叫做鞍点；在矩阵中，一个数在所在行中是最大值，在所在列中是最小值，则被称为鞍点。
- 考虑在 Ax-b=0 约束下，求 $\frac{1}{2}x^TSx$ 的最小值问题。
- 建立 Lagrange 方程：$\mathcal{L}(x,\lambda)=\frac{1}{2}x^TSx+\lambda^T(Ax-b)$
- 对 $\mathcal{L}(x,\lambda)$ 关于 x 和 λ 求导，可以得到矩阵形式鞍点方程如下：

$$\begin{bmatrix} S & A^T \\ A & 0 \end{bmatrix}\begin{bmatrix} x \\ \lambda \end{bmatrix}=\begin{bmatrix} 0 \\ b \end{bmatrix}$$

- 对矩阵左边参量方程的第二行做消元，可得到下面的形式。

$$\begin{bmatrix} S & A^T \\ 0 & -AS^{-1}A^T \end{bmatrix}$$

- 其中，左上为正定矩阵，右下为负定矩阵。由此可判断：前 m 个特征值/主元为正，后 n 个特征值/主元为负。

1.16.3 Rayleigh quotient（瑞利商）的计算

- 对于 $\mathcal{R}(x)$=$\frac{x^TSx}{x^Tx}$
- 当 $x=q_1,\mathcal{R}(x)_{max}=\lambda_1$
- 当 $x=q_n,\mathcal{R}(x)_{min}=\lambda_n$

XU: PS: 瑞利商在 LDA 中有应用

1.17 Saddle Points Continued, Maxmin Principle

主要是关于课程 lab 的解释和提问解答。

1.17.1 Saddle point(from constraint/from $R(x)=\frac{x^TSx}{x^Tx}$)

- from constraint

$$\min \frac{1}{2}x^{\mathrm{T}}Sx, s.t.\ Ax=b$$

Solution:

$$L(x,\lambda)=\frac{1}{2}x^{\mathrm{T}}Sx-\lambda^{\mathrm{T}}(Ax-b)$$

KKT matrix:

$$\begin{bmatrix} S & A^{\mathrm{T}} \\ A & 0 \end{bmatrix}$$

- from $R(x) = \frac{x^{\mathrm{T}}Sx}{x^{\mathrm{T}}x}$(瑞利商)
    - max R
    - min R
    - saddle R

1.17.2 Maxmin principle

鞍点是满足 minmax/maxmin 的点 (结合马鞍的图像)

鞍点的定义: 一阶导数为 0, 二阶 (黑塞矩阵) 为不定阵

1.17.3 Probability

- sample mean: $\mu = \frac{1}{N}(x_1 + \cdots + x_N)$
- expected mean: $m = p_1x_1 + p_2x_2 + \cdots + p_nx_n$
- sample variance: $\frac{1}{N-1}[(x_1 - \mu)^2 + \cdots + (x_N - \mu)^2]$
- Variance: $\mathbb{E}[(x - m)^2]$

1.18 Definition and Inequality

这一节主要讲的是一些统计的基本概念。一些比较基本的期望和方差的概念我们这里就不介绍了。

1.18.1 马尔可夫不等式

assume $X \geq 0$，那么对应的 $P(X(s) \geq a)$ 的概率最多是 $\frac{\bar{X}}{a}$. 比如 $p(X \geq 3) \leq \frac{1}{3}$

检验的方法也非常的简单, 假设后面的都大于期望的话，整体就会超过期望，比如

$$\bar{X} = 0p_0 + 1p_1 + 2p_2 + 3p_3 + 4p_4 + 5p_5 = 1 \tag{13}$$

马尔可夫不等式即为

$$3p_3 + 3p_4 + 3p_5 \leq 1 \tag{14}$$

其实是一个很宽松的条件

1.18.2 切比雪夫不等式

本质上和马尔可夫不等式差不多，只不过是少了一些限制的泛化形式。

对于任意的概率分布 $X(s)$，$P(|X(s) - \bar{X}| \geq a$ 小于等于 $\frac{\sigma^2}{a^2}$ 我们可以轻易的转化为马尔可夫的形式: $Y(s) = (X(s) - \bar{X})^2$

1.18.3 co-variance matrix

这里可以举一个很简单的例子，扔两个硬币就相当于是两个随机变量，我们可以根据不同的共现情况来构造概率矩阵

$$\sigma_{i,j} = \sum_i \sum_j p_{i,j}(x_i - m_1)(y_j - m_2) \tag{15}$$

协方差矩阵的对角线就是方差，并且矩阵式一个正定矩阵，或者说是一个标准的二次型

参考文献