APPLIED LINEAR ALGEBRA

ABSTRACT. This course contains a review of basic linear algebra and many graduate level materials about several topics.

CONTENTS

1. 9/28: Introduction

We might go through some of the applications of linear algebra here.

1.1. Differential equations(finite difference method).

Example 1.1. *An example of one differential equation one might solve is the following:* $\alpha, \beta \in \mathbb{R}, f \in C([0,1]), c \in C([0,1] \to [0,\infty))$ *(c is continuous on [0,1] and is positive),*

$$\begin{cases} -u''(x) + c(x)u(x) = f(x) \quad \textit{where } x \in [0,1] \\ u(0) = \alpha, u(1) = \beta \end{cases}$$

Now we know that if c is a constant, then there is an explicit solution. However, with x in higher dimension or with c varying, there is no analytic solution, or at least pretty hard to find one. Therefore we need a numerical approximation.

So we use the finite difference method to solve the question. We divide $[0,1]$ into n intervals of equal size and let $x_i = \frac{i}{n}, c_i = c(x_i), f_i = f(x_i)$, and we want to approximate $u(x_i)$ by u_i.

So the idea is simply to replace the u'' with u via the approximation by Taylor:

$$u''(x_i) = \frac{u_{i+1} - 2u_i + u_{i-1}}{(1/n)^2}$$

which yields

$$\begin{cases} \frac{2u_i - u_{i-1} - u_{i+1}}{(1/n)^2} + c_i u_i = f_i \quad i = 1, 2 \ldots n-1 \\ u_0 = \alpha, u_n = \beta \end{cases}$$

writing it in matrix form we have $Au = b$ where

$$A = \begin{pmatrix} 2n^2 + c_1 & -1 & 0 & \ldots & & 0 \\ -1 & 2n^2 + c_2 & -1 & 0 & \ldots & 0 \\ 0 & -1 & 2n^2 + c_3 & -1 & \ldots & 0 \\ \vdots & & & \ddots & & \vdots \\ 0 & \ldots & & & -1 & 2n^2 + c_{n-1} \end{pmatrix} \text{ and } b = \begin{pmatrix} f_1 + \alpha n^2 \\ f_2 \\ \vdots \\ f_{n-2} \\ f_{n-1} + \beta n^2 \end{pmatrix}$$

Moreover, suppose u is a solution of a BVP that is C^4, then for u_i determined by the above computation has the property of

$$\max_{i=0,1,\dots,n} \left| u_i^{(n)} - u(x_i) \right| \leq \frac{c}{n^2} \sup_{0 \leq x \leq 1} \left| u^{(4)}(x) \right|$$

1.2. **Image expression via SVD.** We want to model image data as a matrix $A_{m \times n}$.

With SVD, we can write $A = V \tilde{\Sigma} U^*$ where $\tilde{\Sigma} \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix}$ and $\Sigma = diag(\mu_1, \mu_2, \dots, \mu_r)$ with $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_r$, the eigenvalues of A^*A, and U, V are both unitary matrices, i.e. $U^*U = UU^* = I$. Observing closely this decomposition we get

$$A = \sum_{i=1}^{r} \mu_i v_i \cdot \bar{u_i}^T$$

Since μ_i is in descending order, we might say that the first k elements represent the most information that can be restored with a k-rank matrix. We might in some sense say that $\sum_{i=1}^{k} \mu_i v_i \cdot \bar{u_i}^T$ is the "best approximation of A among rank k matrices." This means that for $K \ll r$, we would need a much smaller storage space!

1.3. **Other focuses.** Other focuses include Least square problems and some Modeling, etc.

2. 10/3: Review of some basic Linear Algebra results

2.1. **Gram-Schmidt orthogonalization.** Let our base field of the matrix space be $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$, so our vector space is $\mathbb{K}^d$.

Def 2.1. *The inner product of x and y, written as $< x, y >$, for $x, y \in \mathbb{K}^d$ is*

$$< x, y >= \begin{cases} \sum x_i y_i & \mathbb{K} = \mathbb{R} \\ \sum x_i \bar{y}_i & \mathbb{K} = \mathbb{C} \end{cases}$$

Now, recall that

- $x \perp y$ iff $< x, y >= 0$.
- x is a unit vector iff $< x, x >= 1$.
- $span < x_1, \dots, x_p >= \{\alpha_1 x_1 + \cdots + \alpha_p x_p | \alpha_1 \dots \alpha_p \in \mathbb{R}\}$
- $\{x_1, \dots, x_n\}$ is a collection of linearly independent vectors if $\forall \alpha_1, \dots, \alpha_n \in \mathbb{K}$, $\sum \alpha_i x_i = 0 \Rightarrow \alpha_i = 0, \forall i$.
- $\{x_1, \dots, x_n\}$ is an orthonormal family if $< x_i, x_j >= \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases}$, where we sometimes also write $|x| = ||x|| = (< x, x >)^{1/2}$, the first equality holds when x is finite dimensional.

Theorem 2.1. *(Gram-Schmidt decomposition) Let $\{x_1, \dots, x_n\}$ be a linearly independent set of vectors in $\mathbb{K}^d$. Then $\exists$ orthonormal family $\{y_1, \dots, y_n\} \subseteq \mathbb{K}^d$ s.t.*

$$span\{y_1, \dots, y_p\} = span\{x_1, \dots, x_p\}$$

for all $1 \leq p \leq n$.

Remark 2.1.

- *When $\mathbb{K} = \mathbb{R}$, the family $\{y_1, \dots, y_n\}$ is unique up to changes in signs.*
- *When $\mathbb{K} = \mathbb{C}$, the family $\{y_1, \dots, y_n\}$ is unique up to a scalar $|\alpha| = 1$.*

Proof. (induction on n)

For $n = 1$, let $y_1 = \frac{x_1}{|x_1|}$ then we are done.

Suppose claim holds for all $n \leq N$ for some $N \geq 1$ (IH). Let $\{x_1, \dots, x_N + 1\}$ be a linearly independent set of vectors, we can choose $\{y_1, \dots, y_N\} \subseteq \mathbb{K}^d$ such that $span\{y_1, \dots, y_p\} = span\{x_1, \dots, x_p\}$ for all $1 \leq p \leq N$ (*).

The only thing we need to do now is to search for the suitable y_{N+1} such that

$$span\{y_1, \dots, y_N + 1\} = span\{x_1, \dots, x_N + 1\}$$

which means $y_{N+1} = \sum_{i=1}^{N+1} \alpha_i x_i$ for some $\alpha_1, \dots, \alpha_{N+1}$. Moreover, since (*), we can find $\beta_1, \dots, \beta_N$ such that $\sum_{i=1}^{N} \alpha_i x_i = \sum_{i=1}^{N} \beta_i y_i$.

Thus, any candidate vector y_{N+1} must satisfy

$$y_{N+1} = \alpha_{N+1} x_{N+1} + \sum_{i=1}^{N} \beta_i y_i.$$

Since we want $< y_{N+1}, y_i >= 0$, we have by taking inner product with y_i the following

$$\alpha_{N+1} < x_{N+1}, y_i > + \beta_i = 0 \quad \Rightarrow \quad \beta_i = -\alpha_{N+1} < x_{N+1}, y_i >$$

which implies

$$\begin{aligned} y_{N+1} &= \alpha_{N+1} x_{N+1} + \sum_{i=1}^{N} \beta_i y_i = \alpha_{N+1} x_{N+1} + \sum_{i=1}^{N} -\alpha_{N+1} < x_{N+1}, y_i > y_i \\ &= \alpha_{N+1} \left(x_{N+1} - \sum_{i=1}^{N} < x_{N+1}, y_i > y_i \right). \end{aligned}$$

Afterwards we can simply normalize y_{N+1} and check that it works.

□

Following are some definitions and results from basic linear algebra.

2.2. **Properties of matrices.**

Def 2.2. *For $A \in \mathcal{M}_{n,p}(\mathbb{K})$, the kernel, image/range, and rank of A is defined as*

- $\ker(A) = \{x \in \mathbb{K}^p | Ax = 0\}$;
- $\mathrm{Im}(A) = \{Ax | x \in \mathbb{K}^p\} \subset \mathbb{K}^n$;
- $\mathrm{rank}(A) = \dim(\mathrm{Im}(A))$.

Remark 2.2.

(1) $\ker(A), \mathrm{Im}(A)$ *are subspaces of* $\mathbb{K}^p, \mathbb{K}^n$*, respectively.*
(Recall that $A \subset S$ is a subspace $\iff \forall x, y \in A, \alpha \in \mathbb{K}, x + \alpha y \in A$.)

(2) The dimension of a subspace of $\mathbb{K}^d$ is the number of elements in a spanning linearly independent set of vectors (a basis).
To show this, we claim: let $A \subset \mathbb{K}^d$ be a subspace. If $\{v_1, \dots, v_k\}$ is a basis and $\{w_1, \dots, w_l\}$ is another basis, then $k = l$.
The main idea of the proof is that if $l < k$ then we could exhibit a dependence relation. But it's fairly easy so we skip it.

Lemma 2.2. *For $A \in \mathcal{M}_n(\mathbb{K})$, the following are equivalent:*

(1) A is invertible, i.e. $\exists B \in \mathcal{M}_n(\mathbb{K})$ such that $AB = BA = I$;
(2) $\ker(A) = \{0\}$;
(3) $\mathrm{Im}(A) = \mathbb{K}^n$;
(4) $\exists B \in \mathcal{M}_n(\mathbb{K})$ such that $AB = id_n$;
(5) $\exists B \in \mathcal{M}_n(\mathbb{K})$ such that $BA = id_n$.

Note that in (4) and (5) above, $B = A^{-1}$. Also, if $A, B \in \mathcal{M}_n(\mathbb{K})$ are invertible, then AB is invertible and $(AB)^{-1} = B^{-1}A^{-1}$.

Def 2.3. *For $A = (a_{ij})_{1\le i,j\le n}$, the trace of A is such that* $\mathrm{tr}(A) = \sum_{i=1}^{n} a_{ii}$.

Lemma 2.3. *For $A, B \in \mathcal{M}_n(\mathbb{K})$,* $\mathrm{tr}(AB) = \mathrm{tr}(BA)$.

Proof.

$$\begin{aligned}
\mathrm{tr}(AB) &= \mathrm{tr}\left(\begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} \end{bmatrix}\begin{bmatrix} b_{11} & \dots & b_{1n} \\ \vdots & \ddots & \vdots \\ b_{n1} & \dots & b_{nn} \end{bmatrix}\right) \\
&= \mathrm{tr}\left(\begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} + \cdots + a_{1n}b_{n1} & & \\ & a_{21}b_{12} + a_{22}b_{22} + \cdots + a_{2n}b_{n2} & \\ & & \ddots \end{bmatrix}\right) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}b_{ji} = \sum_{i=1}^{n}\sum_{j=1}^{n} b_{ij}a_{ji} = \mathrm{tr}(BA)
\end{aligned}$$

by rearrangement of finite sums. □

There are a few different perspectives of what determinant is, like the volumn of parallelepiped, decomposition to blocks, permutations, etc.. And we will go with the following.

Def 2.4.

- *$\sigma : \{1, \dots, n\} \to \{1, \dots, n\}$ is a permutation of $\{1, \dots, n\}$ if it's 1-1 and onto;*
- *Let S_n be the set of all permutations of $\{1, \dots, n\}$, call it the permutation group of n.*
- *For $\sigma \in S_n$, the signature of σ is $\varepsilon(\sigma) = (-1)^{p(\sigma)}$ where*

$$p(\sigma) = \sum_{1\le i\le j\le n} Inv_\alpha(i,j) \quad with \quad Inv_\alpha(i,j) = \begin{cases} 0 & \sigma(i) \le \sigma(j) \\ 1 & Otherwise. \end{cases}$$

Def 2.5. *If $A = (a_{ij}) \in \mathcal{M}_n(\mathbb{K})$, then the determinant of A is*

$$\det(A) = \sum_{\sigma\in S_n}\left[\varepsilon(\sigma)\prod_{i=1}^{n} a_{i,\sigma(i)}\right].$$

Def 2.6. *If $A = (a_{ij}) \in \mathcal{M}_n(\mathbb{K})$, then the permanent of A is*

$$\det(A) = \sum_{\sigma \in S_n} \left[\prod_{i=1}^{n} a_{i,\sigma(i)} \right].$$

In general, permanent is much harder to compute.

Remark 2.3. *For $A, B \in \mathcal{M}_n(\mathbb{K})$,*

- *Our definition gives the same value as the definition through recursive computation.*
- $\det(AB) = \det(A)\det(B) = \det(BA)$.
- $\det(A) = \det(A^T)$
- *A is invertible* $\iff \det(A) \neq 0$.

Let's now look at some special matrices.

Def 2.7.

- $A \in \mathcal{M}^n(\mathbb{K})$ *is diagonal if* $a_{ij} = 0$ *for* $i \neq j$.
- $T \in \mathcal{M}^n(\mathbb{K})$ *is upper-triangular if* $t_{i,j} = 0$ *when* $i > j$.
- $T \in \mathcal{M}^n(\mathbb{K})$ *is lower-triangular if* $t_{i,j} = 0$ *when* $i < j$.

Lemma 2.4. *If T is lower triangular in $\mathcal{M}^n(\mathbb{K})$, then if T^{-1} exists it's also lower triangular with diagonal entries given as the reciprocal of the diagonal entries of T. In addition, if $T' \in \mathcal{M}^n(\mathbb{K})$ is also lower triangular, then so is TT', with its diagonal entries $tt'_{ii} = t_{ii}t'_{ii}$.*

Def 2.8. *For $A \in \mathcal{M}^n(\mathbb{C}), A = (a_{ij})_{ij}$, the adjoint/Hermitian transpose matrix A^* of A is given by $A^* = \bar{A}^T = (\bar{a_{ji}})_{ij}$.*

Remark 2.4. *If $x, y \in \mathbb{C}^d$, then $< x, y >= x^T \bar{y}$ and thus $< Ax, y >=< x, A^* y >$.*

Def 2.9.

(1) For $A \in \mathcal{M}^n(\mathbb{C})$,
- *A is self-adjoint/Hermitian if $A = A^*$.*
- *A is unitary if $A^{-1} = A^*$, i.e. $AA^* = A^*A = I$.*
- *A is normal if $AA^* = A^*A$.*

(2) For $A \in \mathcal{M}^n(\mathbb{R})$,
- *A is symmetric if $A = A^T$.*
- *A is orthogonal if $A^{-1} = A^T$, i.e. $AA^T = A^TA = I$.*
- *A is normal if $AA^T = A^TA$.*

3. 10/5: Block matrices, Spectral theory of matrices

For a start, let's just quickly mention that any operation of Gaussian elimination can be view as multiplication by a square matrix, where the square matrix is the result of the operations acted on I. Row operation matrices are multiplied on left, where as column operation matrices are multiplied on right.

3.1. **Block matrices.**

If $A \in \mathcal{M}^n(\mathbb{C})$ and $(n_I)_{1\leq I\leq p}$ is a finite set of positive integers with $\sum n_I = n$. We define things this way so that, as we'll see later, n_I is the "side-length" of the i-th diagonal block.

Def 3.1. *Let $\{e_1, \dots, e_n\} \in \mathbb{C}^n$ be the canonical basis where $e_i = (0, \dots, 1, \dots, 0)$ where the 1 appears on the ith term.*

Now, let $v_i \in \mathbb{C}^n$ be the subspace spanned by sets of n_I base vectors arranged in order, i.e. $v_1 = span\{e_1, \dots, e_{n_1}\}$ and $v_i = span\{e_{1+\sum_{k=1}^{i-1} n_k}, \dots, e_{\sum_{k=1}^{i} n_k}\}$.

Now, let $A_{I,J}$ be block submatrix of size $n_I \times n_J$, which responds to the restriction A to domain v_I and codomain v_J. Then

$$A = \begin{pmatrix} A_{1,1} & \dots & A_{1,p} \\ \vdots & \ddots & \vdots \\ A_{p,1} & \dots & A_{p,p} \end{pmatrix}.$$

Remark 3.1.

(1) Diagonal blocks are square matrices, whereas others are not necessarily so.
(2) If $A = (A_{I,J}), B = (B_{I,J})$ for the same partition, then $C = AB$ where $C_{IJ} \sum_{K=1}^{p} A_{I,K} B_{K,J}$ and C has the same block structure.
(3) Not all operations have block analogs, for instance, there's no block determinant rule. However, some nice identities holds, e.g,

$$\det \begin{pmatrix} A & C \\ 0 & B \end{pmatrix} = \det A \det B.$$

3.2. **Spectral Theory of Matrices over the complex plane.**

Def 3.2. *For $A \in \mathcal{M}^n(\mathbb{C})$,*

(1) The characteristic polynomial $P_A : \mathbb{C} \to \mathbb{C}$ is defined as $P_A = \det(A - \lambda I)$.
(2) The polynomial P_A is a polynomial of degree n and therefore has n roots (possibly repeating) in $\mathbb{C}$ by fundamental theorem of algebra. These roots are the eigenvalues of A; the set of eigenvalues is the spectrum of A, usually denoted $\sigma(A)$.
(3) For $\lambda \in \sigma(A)$, it's algebraic multiplicity is the multiplicity of λ as a root of P_A.
(4) If $\lambda \in \sigma(A)$ has algebraic multiplicity 1 then we call it a simple eigenvalue, otherwise we call it a multiple eigenvalue.

(5) An eigenvector is a non-zero vector $x \in \mathbb{C}^n$ such that $Ax = \lambda x$ for some $\lambda \in \sigma(A)$.
(6) The spectral radius of A is $\rho(A) := \max_{\lambda \in \sigma(A)} |\lambda|$.

Remark 3.2. *Let $A \in \mathcal{M}^n(\mathbb{C})$ be given,*

(1) $\lambda \in \sigma(A)$ implies that there exists an eigenvector associated to λ.
Reason:

$$P_A(\lambda) = 0 \;\Rightarrow\; \det(A - \lambda I) = 0 \;\Rightarrow\; \ker(A - \lambda I) \neq \{0\}$$

so there is some eigenvector in it.
Recall, however, that the eigenvector is not unique since any linear combination of eigenvectors associated with λ is also one. So what really matters is the span of these eigenvectors.
(2) If $\exists x \neq 0$ with $Ax = \lambda x$, then λ is an eigenvalue of A.
(3) Even if $A \in \mathcal{M}_n(\mathbb{R})$, A may have complex eigenvalues, since $\mathbb{C}$ is the algebraic closure of $\mathbb{R}$.
(4) Both the characteristic polynomial and the eigenvalues are invariant under change of basis. In other words, for any invertible matrix $Q \in \mathcal{M}_n(\mathbb{C})$, $P_{QAQ^{-1}} = P_A$ and $\sigma(QAQ^{-1}) = \sigma(A)$.
Reason:

$$\begin{aligned}\det(QAQ^{-1} - \lambda I) &= \det(Q(A - \lambda I)Q^{-1}) \\ &= \det(Q)\det(Q^{-1})\det(A - \lambda I) \\ &= \det(A - \lambda I)\end{aligned}$$

(5) If A is Hermitian, then all its eigenvalues are real.
Reason: if $\lambda \in \sigma(A)$ with eigenvector $u \neq 0$, then

$$\lambda||u||^2 = \lambda < u, u > = < Au, u > = < u, A^*u > = < u, Au > = \bar{\lambda}||u||^2.$$

Now we drive towards a "spectral decomposition" result.

Def 3.3. *For $A \in \mathcal{M}_n(\mathbb{C})$, $\lambda \in \sigma(A)$, the eigenspace associated to eigenvalue λ is*

$$E_\lambda := \ker(A - \lambda I) = \{x \in \mathbb{C}^n : Ax = \lambda x\}.$$

Note that the multiplicity of an eigenvalue is not equal to the dimension of the eigenspace associated to it.

Def 3.4. *The subspace $F_\lambda = \bigcup_{k \geq 1} \ker(A - \lambda I)^k$ is the generalized eigenspace associated to $\lambda \in \sigma(A)$.*

Note that the union in definition 3.4 is really nothing but a finite union. To see this we note that there are only finitely many distinct $\ker(A - \lambda I)^k$. Indeed,

$$x \in \ker(A - \lambda I)^k \;\Rightarrow\; (A - \lambda I)^k x = 0 \;\Rightarrow\; x \in \ker(A - \lambda I)^{k+1}$$

so $\dim(\ker(A-\lambda I)^k)$ increases with k. Yet since there's only finite dimension at its maximum, the union is only a finite union.

Def 3.5. *Let* $P \in \mathbb{C}[x] = \{$*polynomials with base field* $\mathbb{C}\}$ *and* $A \in \mathcal{M}_n(\mathbb{C})$. *Then a matrix polynomial is a function* $P : \mathcal{M}_n(\mathbb{C}) \to \mathcal{M}_n(\mathbb{C})$ *where*

$$P(A) = a_0 I + a_1 A + a_2 A^2 + \cdots + a_d A^d.$$

4. 10/10: Spectral Decomposition and triangular reduction.

Continue with our discussion of polynomials of matrices, we have

Lemma 4.1. *If $X \in C^d$ satisfies $Ax = \lambda x$ for some $\lambda \in \mathbb{C}$, then $P(A)x = P(\lambda)x$ for any polynomial P. In particular, this gives that $P(\lambda) \in \sigma(P(A))$ if $\lambda \in \sigma(A)$.*

Theorem 4.2. *(Cayley-Hamilton Theorem) Given $A \in \mathcal{M}_n(\mathbb{C})$, let $P_A \in \mathbb{C}[x]$ be the characteristic polynomial of A. Then*

$$P_A(A) = 0.$$

The idea of this result is easy: since $P_A(\lambda) = \det(A - \lambda I)$, thus by formally plugging in we have $P_A(A) \approx \det(A - A) = 0$. However, this is not in any sense proper proof since the plug in step is not defined. But a dedicated plug in will yield the result.

In particular, consider the smallest degree polynomial that vanishes at A. Normalize it so that the leading coefficient is 1, which we call the minimal polynomial. The Cayley-Hamilton Theorem says that there is an upper bound to the degree of the minimal polynomial of any A.

Now look at some notations: If $F_1, \dots, F_p \subset \mathbb{C}^n$ are subspaces, we write

$$\mathbb{C}^n = \oplus_{i=1}^p F_i$$

if $\forall x \in \mathbb{C}^n$ can be written uniquely as $x = \sum_{i=1}^{p} x_i$ where $x_i \in F_i$, $1 \le i \le p$.

This is somehow similar to how direct sums are originally defined: $C = A \oplus B$ if $C = A + B = \{a + b | a \in A, b \in B\}$ and $A \cap B = \{0\}$ for subspaces A, B.

More generally, $C = A_1 \oplus \cdots \oplus A_k$ if $C = \{a_1 + \cdots + a_k | a_i \in A_i\}$ and

$A_i \cap (A_1 + \dots A_{i-1} + A_{i+1} + \cdots + A_k) = \{0\}$.

4.1. **Spectral Decomposition.**

Theorem 4.3. *Suppose $A \subset M_n(\mathbb{C})$ has p distinct eigenvalues $\lambda_1, \dots, \lambda_p$ with each λ_i having algebraic multiplicity n_i. Then, the generalized eigenspaces F_{λ_i} satisfy that* $\dim F_{\lambda_i} = n_i$ *for $1 \le i \le p$ and $\mathbb{C}^n = \oplus_{i=1}^p F_{\lambda_i}$ with $F_{\lambda_i} = \ker\left[(A - \lambda_i I)^{n_i}\right]$.*

Remark 4.1.

(1) Consider the generalized F_{λ_i} and choose a basis B_i, $1 \le i \le p$, then the theorem states that $B := \cup_{i=1}^p B_i$ is a basis for $\mathbb{C}_n$.
(2) This theorem encodes everything we need for a Jordan-decomposition. Let P denote the change of basis matrix from the standard basis $\{e_1, \dots, e_n\} \subset \mathbb{C}^n$ to B, i.e. $P^{-1}e_i = b_i$.

Then since $A(F_{\lambda_i}) \subset F_{\lambda_i}$ for all i, we know that $P_{-1}AP = [A]_{E\to B}$ is a block diagonal matrix, i.e $P_{-1}AP = \begin{pmatrix} A_1 & & 0 \\ & \ddots & \\ 0 & & A_p \end{pmatrix}$ with each $A_i \in \mathcal{M}_n(\mathbb{C})$.

Next, by choosing b_i appropriately, we can make each A_i upper triangular with diagonal entries λ_i. With a further refinement, we can get the the Jordan canonical form (not really a definition, but merely mentioning).

4.2. **Matrices and their Triangular forms.**

Def 4.1. *We say $A \in \mathcal{M}_n(\mathbb{C})$ can be reduced to upper/lower triangular form if $\exists P \in \mathcal{M}_n(\mathbb{C})$ that is non-singular (or invertible) and a upper/lower triangular matrix T such that $A = PTP^{-1}$.*

Similarly, A is said to be diagonalizable if $A = PDP^{-1}$ for suitable P and diagonal matrix D.

Remark 4.2.

(1) If A can be reduced to a triangular matrix T, then we say that A and T are similar matrices.
(2) The action of the matrix T corresponds to the same linear transform $x \mapsto Ax$ expressed with respect to a different basis (P).
(3) When A is diagonalizable, the corresponding P has eigenvectors as its columns.
(4) If A can be reduced to triangular (or diagonal) form, then the eigenvalues of A are the diagonal entrices of T (or D), repeated accordingly to their algebraic multiplicity.

Before we go into the proof of a triangular decomposition, let's settle down the notation of a change of basis.

The regular expression of x as a vector is in fact the vector's representation in the standard basis. More generally, we say that x has representation $[b_1, \dots, b_n]^T$ in V if $x = \sum b_i v_i$.

Given two basis sets $\mathcal{A} = \{u_1, \dots, u_n\}$, $\mathcal{B} = \{v_1, \dots, v_n\}$ for the vector space V, for linear map $T \in L(V, V)$, we have

$$[T]_{\mathcal{B}} = S_{\mathcal{A}\to\mathcal{B}}[T]_{\mathcal{A}}S_{\mathcal{B}\to\mathcal{A}} = S^{-1}_{\mathcal{B}\to\mathcal{A}}[T]_{\mathcal{A}}S^{-1}_{\mathcal{A}\to\mathcal{B}}$$

where $S_{\mathcal{B}\to\mathcal{A}} = \left([v_1]_{\mathcal{A}}, \dots, [v_n]_{\mathcal{A}}\right)$.

Proposition 4.4. *$\forall A \in \mathcal{M}_n(\mathbb{C})$ can be reduced to (upper) triangular form.*

Proof.

We proof by induction on the dimension n.

When $n = 1$ it is obviously true.

Suppose the result holds for $n = N - 1$, then for $n = N$.

For $\lambda_1 \in \sigma(A)$, we choose a corresponding eigenvector v_1. Then we choose $N-1$ vectors v_2, v_N such that $\{v_1, \dots, v_n\}$ is a basis for $\mathbb{C}^n$.

For $j \in \{2, \dots, N\}$ we may therefore choose $\alpha_j \in \mathbb{C}$, $b_{ij} \in \mathbb{C}$ for $i = 2, \dots, N$ such that

$$Av_j = \alpha_j v_1 + \sum_{i=2}^{N} b_{ij} v_i \tag{4.1}$$

and let $B = (b_{ij})_{i-1,j-1} \in M_{n-1}(\mathbb{C})$. Then

$$[A]_{v_1,\dots,v_n} = P_1^{-1} A P_1 = \begin{pmatrix} \lambda_1 & \alpha_2 & \dots & \alpha_n \\ 0 & & & \\ \vdots & & B & \\ 0 & & & \end{pmatrix}$$

By IH, $\exists P_2 \in M_{N-1}(\mathbb{C})$ non-singular such that $P_2^{-1} B P_2 = T_2$ where T_2 is upper triangular matrix. We define $P_3 = \begin{pmatrix} 1 & 0 \\ 0 & P_2 \end{pmatrix}$ and we are done. This is because $P = P_1 P_3$, and if we (for simplicity) define $\tilde{\beta} = \tilde{\alpha} P_2$, where $\tilde{\beta} = (\beta_2, \dots, \beta_n)$ and $\tilde{\alpha} = (\alpha_2, \dots, \alpha_n)$. Then

$$P^{-1} A P = \begin{pmatrix} \lambda_1 & \beta_2 & \dots & \beta_n \\ 0 & & & \\ \vdots & & P_2^{-1} B P_2 & \\ 0 & & & \end{pmatrix} = \begin{pmatrix} \lambda_1 & \beta_2 & \dots & \beta_n \\ 0 & & & \\ \vdots & & T_2 & \\ 0 & & & \end{pmatrix}$$

is upper triangular. □

5. 10/12: Schur factorization, Diagonalization of Matrices, and min-max characterization.

5.1. **Schur factorization.**

As a corollary of triangular decomposition, we have the Schur factorization.

Corollary 5.1. *For all $A \in \mathcal{M}_n(\mathbb{C})$, $\exists U \in \mathcal{M}_n(\mathbb{C})$ unitary such that $U^{-1}AU$ is triangular.*

Proof.

Let E be the standard basis for $\mathbb{C}^n$ and use proposition 4.4 to choose a basis $\{f_1, \dots, f_n\}$ which renders A as an upper triangular, i.e. choose f_1 to f_n such that with $P = \begin{pmatrix} f_1, \dots, f_n \end{pmatrix}$, $P^{-1}AP$ is upper triangular.

Choose the same P as defined above, then apply the Gram-Schmidt decomposition to choose $\{g_1, \dots, g_n\} \subset \mathbb{C}^n$ as orthonormal set of vector such that for $1 \leq i \leq n$,

$$\text{span}\{f_1, \dots, f_n\} = \text{span}\{g_1, \dots, g_n\}.$$

Since $AP = PT$, by looking at the columns of AP we get for every i

$$Af_i = (AP)_{ith-col} = (PT)_{ith-col} \in \text{span}\{f_1, \dots, f_n\}.$$

Now, since we can express g_i as a linear combination of f_1 through f_i,

$$\text{span}\{g_1, \dots, g_n\} \subset \text{span}\{f_1, \dots, f_n\}$$

and thus there exists upper triangular R such that $AU = UR$ where $U = \begin{pmatrix} g_1, \dots, g_n \end{pmatrix}$ is unitary. □

5.2. **Diagonalization if Matrices.**

Proposition 5.2. *Suppose $A \in \mathcal{M}_n(\mathbb{C})$ has p distinct eigenvalues $\lambda_1, \dots, \lambda_p$, then*

$$A \text{ is diagonalizable} \iff \mathbb{C}^n = \bigoplus_{i=1}^{p} E_{\lambda_i}$$

where the second part is equivalent to $(E_{\lambda_i} = F_{\lambda_i})$, for all i.

One might be surprised to see a rather short proof on this point. But really this is because it has a much stronger condition.

Proof. ($\Leftarrow$) : If $\mathbb{C}^n = \bigoplus_{i=1}^{p} E_{\lambda_i}$, then A is diagonalizable with respect to the basis obtained by taking a union of basis for each E_{λ_i}.

($\Rightarrow$) : if $A = PDP^{-1}$, then the columns of P is a linearly independent eigenvectors for A_i, so $\mathbb{C}^n = \bigoplus_{i=1}^{p} E_{\lambda_i}$. □

Remark 5.1.

- *Not every matrix is diagonalizable, for instance,* $A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$.

- *There's no easy characterization to a matrix A being diagonalizable.*

Theorem 5.3. *For $A \in \mathcal{M}_n(\mathbb{C})$, A is normal iff A is unitarily diagonalizable (P is unitary).*

Some remarks before the proof:

Remark 5.2.

- *If $A \in \mathcal{M}_n(\mathbb{R})$ is symmetric, then A is diagonalizable.*
- *If one allows the basis "encoded" in P to be non-orthonormal, then, there are diagonalizable matrices $A = PDP^{-1}$ which are not normal.*
- $A = UDU^* \iff A = \sum_{i=1}^{n} \lambda_i u_i u_i^*$.

Proof. (Theorem 5.3)

($\Leftarrow$) : This direction is by direct computation.

($\Rightarrow$) : Suppose A is normal, then we can find $U \in \mathcal{M}_n(\mathbb{C})$ unitary such that $U^{-1}AU = T$ is upper triangular. Now,

$$TT^* = U^*AUU^*A^*U = U^*AA^*U = T^*T$$

which means T is normal. We claim that this implies T is a diagonal matrix.

To see this, we note that

$$\sum_{i=1}^{n} t_{1i}^2 = (TT^*)_{1,1} = (T^*T)_{1,1} = t_{11}^2$$

where $A_{1,1}$ means the first row and first column entry.

Since each summand on the left is non-negative, we get $t_{1i} = 0$ for all $i \neq 1$. Doing this for all diagonal entries of TT^* we see easily that the matrix is diagonal. □

Theorem 5.4. *For $A \in \mathcal{M}_n(\mathbb{C})$, A is Hermitian iff A is diagonalizable with respect to an orthonormal basis and have real eigenvalues.*

Proof. ($\Rightarrow$) : A is Hermitian $\Rightarrow$ A is normal $\Rightarrow$ A is diagonalizable with respect to an orthonormal basis.

Moreover, A is Hermitian itself implies that all it's eigenvalues are real (remark 3.2 (5)).

($\Leftarrow$) : This is simply by computation. □

5.3. **Another view on eigenvalues: variational/min-max characterization of eigenvalues.**

Def 5.1. *For $A \in \mathcal{M}_n(\mathbb{C})$ and Hermitian, the Rayley quotient is the function*

$$R_A : \mathbb{C}^n \backslash \{0\} \to \mathbb{R}$$

determined by

$$R_A(x) = \frac{< Ax, x >}{< x, x >}.$$

Remark 5.3. *For A Hermitian,* $x \in \mathbb{C}^n$, $< Ax, x >=< x, A^*x >=< x, Ax >= \overline{< Ax, x >}$, *so* $< Ax, x > \in \mathbb{R}$ *and* $R_A \in \mathbb{R}$.

Theorem 5.5. *For* $A \in \mathcal{M}_n(\mathbb{C})$ *and Hermitian, the smallest eigenvalue*

$$\lambda_1 = \min_{x \neq 0} \frac{< Ax, x >}{< x, x >} = \min_{|x|=1} < Ax, x >$$

and the largest eigenvalue

$$\lambda_n = \max_{x \neq 0} \frac{< Ax, x >}{< x, x >} = \max_{|x|=1} < Ax, x > .$$

Proof. We will prove the result for λ_1 and that the usage of min is justified. The other part is the same.

A Hermitian A has the spectrum $\sigma(A) \subset \mathbb{R}$, so we can order them in the manner $\lambda_1 \leq \cdots \leq \lambda_n$ (listing with multiplicity), and there is a orthonormal basis of corresponding eigenvectors $\{x_1, \ldots, x_n\}$.

For $y \in \mathbb{C}^n$, choose $\tilde{y}_1$ to $\tilde{y}_n \in \mathbb{C}$ such that $y = \sum_{i=1}^{n} \tilde{y}_i x_i$. Then, for all $y \in \mathbb{C}^n$, we have

$$Ay = \sum_{i=1}^{n} \tilde{y}_i A x_i = \tilde{y}_i \lambda_i x_i$$

and thus

$$\begin{aligned} < Ay, y > &= \left\langle \sum_{i=1}^{n} \tilde{y}_i \lambda_i x_i, y \right\rangle = \sum_{i=1}^{n} \tilde{y}_i \lambda_i < x_i, y >= \sum_{i=1}^{n} \sum_{j=1}^{n} \tilde{y}_i \lambda_i < x_i, \tilde{y}_j x_j > \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} \tilde{y}_i \lambda_i \tilde{y}_j < x_i, x_j >= \sum_{i=1}^{n} |\tilde{y}_i|^2 \lambda_i \\ &\geq \lambda_1 \sum_{i=1}^{n} |\tilde{y}_i|^2 = \lambda_1 < y, y > \end{aligned}$$

Thus, $R_A \geq \lambda_1$ for all $y \in \mathbb{C}^n, y \neq 0$.

Moreover, for $\forall y$ such that $||y|| = 1$ we have $< Ay, y > \geq \lambda_1$.

Since there exist an eigenvector corresponding to λ_1, apply that and we see the min is attained.

□

6. 10/17: Courant-Fisher min-max method, SVD

Proposition 6.1. *For $A \in \mathcal{M}_n(\mathbb{C})$ Hermitian with eigenvalues $\lambda_1, \dots, \lambda_n$ in increasing order, then for $i = 2, \dots, n$,*

$$\lambda_i = \min_{x \perp span\{x_1, \dots, x_{i-1}\}} R_A(x)$$

where x_j is the eigenvector associated to λ_j.

Remark 6.1.

(1) The proof of the proposition is nothing different from when $i = 2$, just when taking the inner product with y_i, ther first $i - 1$ terms are 0.
(2) A similar characterization in terms of maximizing $R_A(x)$ holds.
(3) Notation wise, $x \perp span\{x_1, \dots, x_{i-1}\}$ denotes the set $\{z \in \mathbb{C}^n; \langle z, y\rangle = 0\}$.

6.1. Courant-Fisher Min-max method.

Theorem 6.2. *(Courant Fisher): For $A \in \mathcal{M}_n(\mathbb{C})$ Hermitian, $\lambda \leq \cdots \leq \lambda_n$. Then for all $i = 1, \dots, n$,*

$$\lambda_i = \max_{\text{all possible } \{a_1, \dots, a_{i-1}\}} \min_{x \perp span\{a_1, \dots, a_{i-1}\}} R_A(x).$$

Remark 6.2.

(1) When $i = 1$, we are just minimizing over everything.
(2) A similar characterization holds with min-max.

The idea of the proof is that, really, we have a max and a min, so using them well should give us both a upper bound and lower bound.

Proof. ($\lambda_i \leq$): We can simply choose $(a_1, \dots, a_{i-1}) = (x_1, \dots, x_{i-1})$ and we are done.

($\lambda_i \geq$): Note that for any $\{(a_1, \dots, a_{i-1})\}$ in $\mathbb{C}^n$, we have that

$$\dim(x \perp \text{span}\{a_1, \dots, a_{i-1}\}) \geq n - i + 1$$

and

$$\dim(\text{span}\{x_1, \dots, x_i\}) = i.$$

Basic knowledge of dimension then implies

$$S = \left(\text{span}\{a_1, \dots, a_{i-1}\}\right)^{\perp} \cap \{\text{span}\{x_1, \dots, x_i\}\} \neq \{0\}.$$

Thus we have moved the min from a characterization of arbitrary a_n to the eigenvectors by comparing the minimization of the intersection:

$$\min_{x \perp \text{span}\{a_1, \dots, a_{i-1}\}} R_A(x) \leq \min_{x \in S} R_A(x) \leq \max_{x \in \text{span}\{x_1, \dots, x_i\}} R_A(x) = \lambda_i$$

thus we are done. □

6.2. **SVD.**

Lemma 6.3. *$\forall A \in \mathcal{M}_n(\mathbb{C})$, the nxn matrix A^*A is Hermitian and has real non-negative eigenvalues.*

Proof. We know already Hermitian matrices have real eigenvalues.

By computation:

$$(A^*A)^* = A^*(A^*)^* = A^*A$$

so it is Hermitian. Further,

$$\langle A^*Ax, x\rangle = \langle \lambda x, x\rangle = \lambda\langle x, x\rangle$$

which means

$$\lambda = \frac{\langle A^*Ax, x\rangle}{\langle x, x\rangle} = \frac{\langle Ax, Ax\rangle}{\langle x, x\rangle} = \frac{||Ax||^2}{||x||^2} \geq 0.$$

□

Def 6.1. *The singular values for A are the square roots of the eigenvalues of AA^*.*

Remark 6.3.

(1) The same of course applies to AA^*. Indeed,

Lemma 6.4. *If $A, B \in \mathcal{M}_n(\mathbb{C})$ then non-zero eigenvalues of AB and BA are the same.*

Proof. The idea of the proof is simply to suppose (λ, x) is an eigenvalue, eigenvector pair for AB, then $ABx = \lambda x$, and if $\lambda \neq 0$, $x \notin \ker(B)$.

Thus, we left multiply both sides by B and get $BABx = \lambda Bx \quad \Rightarrow \quad \lambda \in \sigma(BA)$.

□

(2) When A is a square matrix, it is non-singular iff all singular values are non-zero.

(3) If $A \in \mathcal{M}_n(\mathbb{C})$ is normal, then the singular values of A are $\{|\lambda_i| : \lambda_i \in \sigma(A)\}$.

To see this, note that by spectral theorem for normal matrices, $A = U^*DU$, which since U unitary, we can see the result.

(4) In particular, above remark implies that the spectral radius of a normal matrix is the same as A^*.

Theorem 6.5. *(SVD factorization.) Let $A \in \mathcal{M}_{m,n}(\mathbb{C})$ be a matrix with r positive singular values $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_r > 0$. Set $\Sigma = diag(\mu_1, \mu_2, \ldots, \mu_r)$ and $\tilde{\Sigma} = \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix} \in \mathcal{M}_{m,n}(\mathbb{R})$.*

Then there exists unitary matrices $U \in M_n(\mathbb{C})$, $V \in M_m(\mathbb{C})$ wuch that

$$A = V\tilde{\Sigma}U^*.$$

Proof. WLOG, assume $m \geq n$ since otherwise look at A^*.

Recall that eigenvalues of A^*A are given by μ_i^2. Since A^*A is Hermitian, so its diagonalizable via orthonormal basis U. We claim that this is the U we want and our goal is to find V below.

Write $U = [u_1, \dots, u_n]$, then $A^*Au_i = \mu_i^2 u_i$ so that

$$A^*AU = [\mu_1^2 u_1, \dots, \mu_n^2 u_n] = U \cdot \operatorname{diag}(\mu_1^2, \dots, \mu_n^2).$$

Left multiplying by U^* we get

$$U^*A^*AU = \operatorname{diag}(\mu_1^2, \dots, \mu_n^2) = \tilde{\Sigma}^T\tilde{\Sigma}.$$

On the other hand we also know that

$$\langle Au_i, Au_j \rangle = \langle A^*Au_i, u_j \rangle = \mu_i^2 \langle u_i, u_j \rangle = \mu_i^2 \delta_{ij}$$

since $\{Au_i, Au_j\}$ are pairwise orthogonal for $1 \leq i, j \leq r$.

Also, $Au_i = 0$ for $r < i \leq n$.

Now we normalize these vectors by setting $v_i = \dfrac{Au_i}{\mu_i}$ and for the rest of v are chosen such that $v_1, \dots, v_m$ forms an orthonormal basis.

Now we can compute and check:

$$\begin{aligned} V\tilde{\Sigma}U^* &= [v_1, \dots, v_m]\tilde{\Sigma}U^* = [\mu_1 v_1, \dots, \mu_n v_n]_{mxn}U^* \\ &= [Au_1, \dots, Au_r, 0, \dots, 0]U^* = [Au_1, \dots, Au_n]U^* \\ &= AUU^* = A \end{aligned}$$

□

7. 10/19: Remarks on SVD; few words on natural subspaces associated to a matrix; Norms

7.1. More on SVD.

Remark 7.1.

(0) When $A \in \mathcal{M}_{m,n}(\mathbb{R})$, U, V can be chosen as real valued.

(1) The theorem implies that for real valued non-singular $A \in \mathcal{M}_n(\mathbb{R})$, it takes the unit sphere $\mathbb{S}^{n-1}$ into an ellipsoid.

In other words, suppose $A = V\Sigma U^T$, we have the following question and answer:

Q: What is $V\Sigma U^T\mathbb{S}^{n-1}$?

A: Since U is orthogonal, and any orthogonal matrix maps the space into another orthogonal basis, we have $U^T\mathbb{S}^{n-1} = \mathbb{S}^{n-1}$ (caring only the shape).

Further, Σ is a diagonal matrix, which means that

$$\Sigma\mathbb{S}^{n-1} = \left\{ (x_1', \dots, x_n') \in \mathbb{R}^n \,:\, \sum_{i=1}^{n} \left(\frac{x_i}{\mu_i}\right)^2 = 1 \right\} \sim E^{n-1}$$

for E^{n-1} is an ellipsoid.

And V is yet another rotation/flipping, and we are done.

(2) Applying SVD to square matrices is NOT the same as diagonalization. The reason behind this is that U and V are two different basis.

(3) $\operatorname{rank}(A) = r \leq \min\{m, n\}$ for r defined in the proof.

(4) How much different choice is there in choosing the right SVD?

A: The really free choice are those columns of U and V that we manually added for them to be orthogonal.

Def 7.1. *(Moore-Penrose pseudoinverse) Given a matrix $A \in \mathcal{M}_{m,n}(\mathbb{C})$ with SVD*

$$A = V\tilde{\Sigma}U^*$$

and the pseudoinverse $A^\dagger \in \mathcal{M}_{n,m}(\mathbb{C})$ is the matrix

$$A^\dagger = U\tilde{\Sigma}^\dagger V^*$$

where $\tilde{\Sigma}^\dagger = \begin{pmatrix} \Sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix} \in \mathcal{M}_{n,m}(\mathbb{R})$.

Why is this pseudoinverse useful? Let's see the following remarks:

Remark 7.2. *Suppose that $m \geq n$.*

(0) Some useful equations are:

- $A^\dagger A = U\tilde{\Sigma}^\dagger\tilde{\Sigma}U^* = U \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} U^* = \sum_{i=1}^{r} u_i u_i^*$.

- $AA^\dagger = U\tilde{\Sigma}\tilde{\Sigma}^\dagger U^* = V\begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix}V^* = \sum_{i=1}^{r} v_i v_i^*$.
- $A = \sum_{i=1}^{r} \mu_i u_i v_i^*$ where as $A^\dagger = \sum_{i=1}^{r} \frac{1}{\mu_i} u_i v_i^*$.

(1) If $\text{rank}(A) = n \leq m$, then $A^\dagger = (A^*A)^{-1}A^*$, which means that if A happened to be a square matrix then $A^* = A^\dagger$.
(2) It turns out that $A^\dagger$ is the unique matrix X such that all of below holds:
   (i) $AXA = A$,
   (ii) $XAX = X$,
   (iii) $XA = (XA)^*$,
   (iv) $AX = (AX)^*$.
(3) The "least square" (we will define later) solution to $Ax = b$, i.e. $\min_x ||Ax - b||$ is obtained by $x^\dagger = A^\dagger b$.

Maybe let's see one more application of SVD.

Proposition 7.1. *(Polar Decomposition) For all $A \in \mathcal{M}_n(\mathbb{R})$, there exists $Q, S \in \mathcal{M}_n(\mathbb{R})$ where Q is orthonormal and S is symmetric and semi-positive definite such that $A = QS$.*

Remark 7.3.

(1) This is a matrix generalization of $z = re^{i\theta}$. By saying this I mean that A is decomposed into two matrices, one controlling stretching solely, and one rotation solely.
(2) Note that if A is invertible, then S is positive definite.

Proof. The idea is just to write $A = V\tilde{\Sigma}U^T = VU^TU\tilde{\Sigma}U^T$. By simply letting $S = U\tilde{\Sigma}U^T$ and $Q = VU^T$ we are done. □

As one can imagine, we can use it to separate stretching from rotation in deformations of solids.

7.2. **A few word on natrual subspaces associated to a matrix.**

For $A \in \mathcal{M}_{m,n}(\mathbb{R})$, let

$$A = \begin{pmatrix} | & & | \\ a_1 & \dots & a_n \\ | & & | \end{pmatrix} = \begin{pmatrix} — & \tilde{a_1} & — \\ & \vdots & \\ — & \tilde{a_n} & — \end{pmatrix}.$$

Then we have

- $\text{Col}(A) = \text{span}\{a_1, \dots, a_n\} = \text{Im}(A) = \{Ax | x \in \mathbb{R}^n\}$;
- $\text{rank}(A) = \dim(\text{Im}(A))$;
- $\ker(A) = \text{Null}(A) = \{x \in \mathbb{R}^n | Ax = 0\}$;
- $\text{Row}(A) = \text{span}\{\tilde{a_1}, \dots, \tilde{a_n}\} = \text{Col}(A^T)$;

- "(Left null space:) " $\ker(A^T) = \{y \in \mathbb{R}^m \,:\, A^T y = 0\}$.

and it's worth mentioning that the following holds automatically:

- $\dim(\ker(A)) = n - \operatorname{rank}(A)$;
- $\dim(\operatorname{Row}(A)) = \operatorname{rank}(A)$;
- $\dim(\ker(A^T)) = m - \operatorname{rank}(A)$;
- $\ker(A) = \operatorname{Row}(A)^{\perp}$;
- $\ker(A^T) = \operatorname{Col}(A)^{\perp}$.

Let's focus now on $A \in \mathcal{M}_n(\mathbb{R})$. By SVD $A = V\tilde{\Sigma}U^T$ with $r = \operatorname{rank}(A)$, we have

$$U = \underbrace{\begin{pmatrix} | & & | \\ u_1 & \dots & u_r \\ | & & | \end{pmatrix}}_{\substack{\text{Orthonormal} \\ \text{basis of} \\ \text{Row}(A)}} \underbrace{\begin{pmatrix} | & & | \\ u_{r+1} & \dots & u_n \\ | & & | \end{pmatrix}}_{\substack{\text{Orthonormal} \\ \text{basis of} \\ \ker(A)}}, \quad V = \underbrace{\begin{pmatrix} | & & | \\ v_1 & \dots & v_r \\ | & & | \end{pmatrix}}_{\substack{\text{Orthonormal} \\ \text{basis of} \\ \text{Col}(A)}} \underbrace{\begin{pmatrix} | & & | \\ v_{r+1} & \dots & v_m \\ | & & | \end{pmatrix}}_{\substack{\text{Orthonormal} \\ \text{basis of} \\ \ker(A^T)}}.$$

7.3. **Norms.**

We now move towards some numerical considerations such as how to measure errors and convergence, etc.

Def 7.2. *For $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$, a norm $||\cdot|| : \mathbb{K}^d \to [\infty)$ is a function that satisfies*

(i) Positive definitenes: $||x|| \geq 0$ with $||x|| = 0$ iff $x = 0$.
(ii) Homogeneous: $||\lambda x|| = |\lambda| \cdot ||x||$ for all $x \in \mathbb{K}^d$, $\lambda \in \mathbb{K}$.
(iii) Triangular inequality: $||x + y|| \leq ||x|| + ||y||$ for all $x, y \in \mathbb{K}^d$.

Def 7.3. *For vector space V over $\mathbb{K}$, $\langle \cdot, \cdot \rangle$ is a inner product if it satisfies*

(i) $\langle v, v \rangle \geq 0$ for all v.
(ii) $\langle v, v \rangle = 0$ iff $v = 0 \in V$.
(iii) $\langle \alpha_1 w_1 + \alpha_2 w_2, v \rangle = \alpha_1 \langle w_1, v \rangle + \alpha_2 \langle w_2, v \rangle$.
(iv) $\langle v, w \rangle = \overline{\langle w, v \rangle}$ for all $w, v \in V$.

Let's see some examples on this:

Example 7.1.

(1) A scalar product (since the output is a scalar) $||x|| = \sqrt{\langle x, x \rangle}$ defines a norm on $\mathbb{K}^d$. However, not every norm can be defined by an inner product.

(2) *The Euclidean norm on* $\mathbb{K}^d$ *is*

$$||x||_2 = \left(\sum_{i=1}^{d} |x_i|^2 \right)^{\frac{1}{2}}.$$

(3) *Similarly, the p-norm is defined as*

$$||x||_p = \left(\sum_{i=1}^{d} |x_i|^p \right)^{\frac{1}{p}}$$

for $1 \leq p < \infty$ *since we want it to be a norm (satisfy the triangular inequality).*

In particular, when $p = 1$ *we call that norm the "Manhattan norm" since in big cities the streets are all in grids. So does our norm look like so, be cause we always go in grids to find the norm when in practice.*

(4) *A weighted p-norm is defined as*

$$||x||_{p,\omega} = \left(\sum_{i=1}^{d} \omega_i |x_i|^p \right)^{\frac{1}{p}}$$

where the parameter $\omega = (\omega_1, \ldots, \omega_d)$ *and each* $\omega_i \geq 0$.

(5) *For a real positive definite symmetric matrix A, the following defines a norm:*

$$||x||_A = \left(x^T A x\right)^{\frac{1}{2}} = \left(\sum_{i,j=1}^{n} a_{ij} x_i x_j \right)^{\frac{1}{2}}.$$

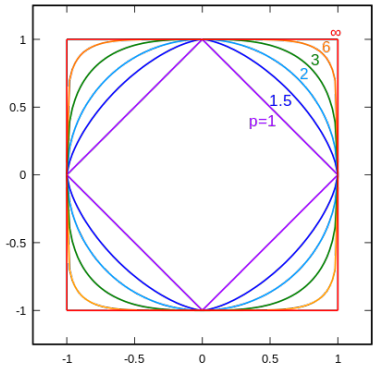
Note that in particular, when $W = diag(w)$, *we have* $||x||_W = ||x||_{2,w}$.

(6) *The* ∞*-norm is defined as*

$$||x||_\infty = \max_{1 \leq i \leq d} |x_i| = \lim_{p \to \infty} ||x||_p$$

where the last limit will be shown later.

Remark 7.4. *The following is the unit ball associated to different p-norms:*



We'll end today's discussion by the fact about compariring norms.

- For $p \geq 1$, $x \in \mathbb{K}^d$, we have that
$$|x_i| \leq \left(\sum_{i=1}^{d} |x_i|^p \right)^{\frac{1}{p}} \rightsquigarrow ||x||_\infty \leq ||x||_p.$$
- For $p \geq 1$, $x \in \mathbb{K}^d$,
$$||x||_p = \left(\sum_{i=1}^{d} |x_i|^p \right)^{\frac{1}{p}} \leq \left(\sum_{i=1}^{d} ||x_i||_\infty^p \right)^{\frac{1}{p}} = ||x||_\infty d^{\frac{1}{p}}.$$
- $x \in \mathbb{K}^d$,
$$||x||_2 = \left(\sum_{i=1}^{d} |x_i|^2 \right)^{\frac{1}{2}} \leq \sum_{i=1}^{d} \left(|x_i|^2 \right)^{\frac{1}{2}} = \sum_{i=1}^{d} |x_i|.$$
Note that we can get the inequality $(a+b)^{\frac{1}{2}} \leq a^{\frac{1}{2}} + b^{\frac{1}{2}}$ simply by taking the derivative of the function
$$t \mapsto (a+t)^{\frac{1}{2}} - a^{\frac{1}{2}} - t^{\frac{1}{2}}$$
and see that it is decreasing for $a \geq 0, t \geq 0$, and when both is 0 its value is 0.

8. 10/24: Norms on matrices

Let's start with a few further remarks on norms:

Remark 8.1.

(1) Norms are continuous functions. To see why, we use the triangle inequality and get

$$||x|| = ||x - y + y|| \leq ||x - y|| + ||y|| \Rightarrow \Big|||x|| - ||y||\Big| \leq ||x - y||$$

where the last inequality is exactly the continuous condition.

In particular, $x \mapsto ||x||$ is uniformly continuous (and even Lipschitz).

(2) On $\mathbb{R}^d$, Cauchy-Schwartz implies $x \cdot y \leq ||x||_2||y||_2$. Now

$$||x||_1 = \sum_{i=1}^{d} |x_i| = (\pm 1, \dots, \pm 1) \cdot x \leq ||(\pm 1, \dots, \pm 1)||_2||x||_2 = \sqrt{d}||x||_2$$

which, combining with

$$||x||_2 \leq ||x||_1$$

says

$$||x||_2 \leq ||x||_1 \leq \sqrt{d}||x||_2$$

that they are of the same order.

The following fact is a generation of the above remark 2.

Proposition 8.1. *If E is a finitely dimensional vector space, then all norms of E are equivalent in the sense that for all norms $||\cdot||, ||\cdot||'$, $\exists c, C > 0$ such that*

$$c||x|| \leq ||x||' \leq C||x||$$

for all $x \in E$.

For instance we have

$$||x||_\infty \leq ||x||_2 \leq \sqrt{n}||x||_\infty$$

and

$$\frac{1}{n}||x||_1 \leq ||x||_\infty \leq ||x||_1.$$

Note that this is true even for "crazy" choice of norms such as

$$||x||_\alpha = \left(\sum_{i=1}^{d} |x_i|^n\right)^{\frac{1}{n}} \quad \text{and} \quad ||x||_\beta = x^T \begin{pmatrix} 3 & -1 & & \\ -1 & 3 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 3 \end{pmatrix} x.$$

Remark 8.2. *Why do we care about this?*

Answer: for instance, when we have a sequence $x_n \to x$, which means $||x_n - x|| \to 0$ for some norm. Then this proposition 8.1 tells us that whichever norm we use here would be fine. In addition, if we can show an algorithm converges in a specific norm, then it is convergent in any other norm as well.

8.1. **Norm on matrices.**

We first justify some initial explanation of norms of matrices. Since we can view $\mathcal{M}_n(\mathbb{K})$ as a vector space of $d = n^2$ over $\mathbb{K}$, we can consider norms on the space. This "can view" part can be justified by a statement that says the two spaces are isomorphic.

We now define some norms on matrices.

Def 8.1.

(1) *Forbinus Norm (Euclidean norm, Schur norm):*

$$||A||_{\mathcal{F}} = \left(\sum_{1 \le i,j \le n} |a_{ij}|^2 \right)^{\frac{1}{2}}.$$

(2) *Hölder's q-norm: for $q \ge 1$,*

$$||A||_{l^q} := \left(\sum_{1 \le i,j \le n} |a_{ij}|^q \right)^{\frac{1}{q}} =: ||A||_{H,q}.$$

(3) *co-norm:*

$$||A||_{l^\infty} = \max_{1 \le i,j \le n} |a_{i,j}| = ||A||_{H,\infty}.$$

We will see that, even though above norms all seemed to come from the p-vector norms, some of them are nicer that others in terms of how they captures matrix multiplications.

Def 8.2. *A norm $|| \cdot ||$ on $\mathcal{M}_n(\mathbb{K})$ is a matrix norm if $\forall A, B \in \mathcal{M}_n(\mathbb{K})$ we have $||AB|| \le ||A|| \cdot ||B||$.*

Let's test the norms in definition 8.1 and see why some is better.

- $|| \cdot ||_{\mathcal{F}}$ is a matrix norm (by Cauchy):

$$\begin{aligned} ||AB||_{\mathcal{F}} &= \sum_{1 \le i,j \le n} |(AB)_{ij}|^2 = \sum_{1 \le i,j \le n} \left| \sum_{k=1}^{n} a_{ik} b_{kj} \right|^2 \\ &\le \sum_{1 \le i,j \le n} \left(\sum_{k=1}^{n} |a_{ik}|^2 \right) \left(\sum_{k=1}^{n} |b_{kj}|^2 \right) = ||A||_{\mathcal{F}} ||B||_{\mathcal{F}}. \end{aligned}$$

- $|| \cdot ||_{l^\infty}$ is not a matrix norm as we can consider A to be the matrix with every entry equal to 1. Then $||A^2||_{l^\infty} = n \ne 1 = ||A||_{l^\infty}^2$.

For many purposes we will work with matrix norms associated to vector norms on the space $\mathbb{K}^n$.

Def 8.3. *Let $||\cdot||_*$ be a vector norm on $\mathbb{K}^n$, then the norm*

$$||A||_* := \sum_{\substack{x\neq 0\\ x\in\mathbb{K}^n}} \frac{||Ax||_*}{||x||_*}$$

is a matrix norm on $\mathcal{M}_n(\mathbb{K})$ that is subordinate to a vector norm.

Geometrically, this norm encodes the value of how far the map $x \rightsquigarrow Ax$ sends the vector from the unit circle.

Def 8.4. *For $A \in L(\mathbb{C}^m, \mathbb{C}^n)$, the operator norm is defined as*

$$||A||_{a,b} = \sup_{x\neq 0} \frac{||Ax||_b}{||x||_a}$$

for suitably defined a,b norms. Note that operator norms aren't necessarily matrix norms since two matrix may not even be able to multiply.

Remark 8.3.

(1) The $||A||_*$ defined this way is indeed a matrix norm:

$$||AB||_* = \sup_{x\neq 0} \frac{||ABx||_*}{||x||_*} \leq \sup_{x\neq 0} \frac{||A||_*||Bx||_*}{||x||_*} = ||A||_*||B||_*$$

where the inequality is definition of subordinate norms.

For general operator norms though, if we can get over the well-define issue, then we really have the matrix norm property: for $A \in \mathbb{C}^{m\times n}$, $B \in \mathbb{C}^{n\times p}$, we have

$$||AB||_{a,c} \leq ||A||_{c,b} \cdot ||B||_{a,b}.$$

(2) We also have (by homogeneity) for $A \in \mathcal{M}_n(\mathbb{K})$,

$$||A||_* = \sup_{\substack{x\in\mathbb{K}^n\\ ||x||_*=1}} ||Ax||_* = \sup_{\substack{x\in\mathbb{K}^n\\ ||x||_*\leq 1}} ||Ax||_*.$$

(3) $||I_n||_* = 1$ for all subordinate norms.

Proposition 8.2. *Let $||\cdot||$ be a subordinate matrix norm on $\mathcal{M}_n(\mathbb{K})$, then for $A \in \mathcal{M}_n(\mathbb{K})$, $\exists x_A \in \mathbb{K}^n\backslash\{0\}$ such that*

$$||A|| = \frac{||Ax_A||}{||x_A||}.$$

Proof. The idea of the proof is easy. We know that $f : x \mapsto ||Ax||$ is continuous on the closed bounded subset $\{x \in \mathbb{K}^n : ||x|| = 1\}$ of $\mathbb{K}$, then since $\mathbb{K}^n$ is finite dimensional, the set is compact thus f attains its maximum. □

Remark 8.4.

(1) Let $\tilde{x}_A = \dfrac{x_A}{||x_A||}$, *then* $||A\tilde{x}_A|| = ||A||$.

(2) Not all matrix norms are subordinate to a vector norm. For instance, the Forbinus norm has that $||I||_F = \sqrt{n}$, *so it's not a subordinate norm for* $n \geq 2$.

(3) Many properties of subordinate norms are inherited from its generating vector norm. For instance, for $A \in \mathcal{M}_n(\mathbb{K})$,

$$\frac{1}{n^{\frac{1}{p}}}||A||_\infty \leq ||A||_p \leq n^{\frac{1}{p}}||A||_\infty$$

which is easily get by plugging in the definitions.

(4) The computation of $||A||_p$ *for* $p \geq 1$ *are in general difficult. But for* $p = 1, 2, \infty$ *we are in better position, as we will show in the next propsition.*

Proposition 8.3. *(Special matrix norms) For* $A \in \mathcal{M}_n(\mathbb{K})$,

(a) $||A||_1 = \max\limits_{1\leq j\leq n} \sum\limits_{i=1}^{n} |a_{ij}|$. *(Max of the sum of columns.)*

(b) $||A||_\infty = \max\limits_{1\leq i\leq n} \sum\limits_{j=1}^{n} |a_{ij}|$. *(Max of the sum of rows.)*

(c) $||A||_2 = ||A^*||_2 = \mu_1$ *where* μ_1 *denotes the largest singular value of* A.

Proof.

(a):

($\leq$): Since

$$\begin{aligned} ||Ax||_1 = \sum_{i=1}^{n} \left| \sum_{j=1}^{n} a_{ij}x_j \right| &\leq \sum_{j=1}^{n} \left(|x_j| \sum_{j=1}^{n} \left| a_{ij} \right| \right) \\ &\leq \sum_{j=1}^{n} |x_j| \left(\max_j \sum_{j=1}^{n} \left| a_{ij} \right| \right) = \left(\max_j \sum_{j=1}^{n} \left| a_{ij} \right| \right) ||x||_1 \end{aligned}$$

we have

$$||A||_1 = \sum_{x\neq 0} \frac{||Ax||_1}{||x||_1} \leq \max_j \sum_{j=1}^{n} \left| a_{ij} \right|.$$

($\geq$): We simply choose the regular base vector e_j corresponding the column that has the largest absolute sum.

(b):

($\leq$):

$$||Ax||_\infty = \max_{1\leq i\leq n} \left| \sum_{j=1}^{n} a_{ij}x_j \right| \leq \max_{1\leq i\leq n} ||x||_\infty \sum_{j=1}^{n} |a_{ij}| = ||x||_\infty \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}|.$$

($\geq$):

For the row that has the largest absolute sum, we normalize it and take the conjugate, call it x_*.

Then if $A = 0$ clearly we are done. If $A \neq 0$ then since $||x_*|| = 1$, by plugging in we know

$$||A||_\infty \geq ||Ax_*||_\infty = \sum_{j=1}^{n} |a_{i^*,j}|$$

where i^* is the index of that biggest row.

(c):

A trick to begin is to note the following:

$$\frac{||Ax||_2^2}{||x||_2^2} = \frac{\langle Ax, Ax\rangle}{\langle x, x\rangle} = \frac{\langle A^*Ax, x\rangle}{\langle x, x\rangle}.$$

Since A^*A is Hermitian and positive semi-definite, it is diagonalizable with respect to an orthonormal basis of eigenvectors. So we obtaine

$$\sup_{x\neq 0} \frac{\langle A^*Ax, x\rangle}{\langle x, x\rangle} = \max_{i=1,2,\ldots,n} \lambda_i(A^*A) = \lambda_1 = \mu_1^2$$

where eigenvalues are ordered and notations are consistent with preceding lectures.

So $||A||_2 = \sqrt{\mu_1^2} = \mu_1$ since singular values are positive real numbers.

To see $||A^*||_2 = \mu_1$, we can of course do the above proof with substitution, but there's one computation that is really nice and efficient:

$$\frac{||Ax||_2^2}{||x||_2^2} = \frac{\langle A^*Ax, x\rangle}{\langle x, x\rangle} \leq \frac{||A^*Ax||_2||x||_2}{||x||_2^2} \leq ||A^*A||_2 \leq ||A^*||_2||A||_2$$

where for the inequalities we used Cauchy, definition of subordinate norms, and matrix norm property. So it really connects everything.

Then by simply taking the sup on the left side we obtain

$$||A||_2^2 \leq ||A^*||_2||A||_2 \Rightarrow ||A||_2 \leq ||A^*||_2$$

and by a symmetry argument we get $||A^*||_2 \leq ||A||_2$, thus we are done.

□

9. 10/26: Matrix norms and Spectral theory, Matrix approximation

9.1. **Matrix norms and spectral theory.**

Remark 9.1.

(1) $A \in \mathcal{M}_n(\mathbb{R})$ *can also be viewed as an element of* $\mathcal{M}_n(\mathbb{C})$*. Similarly, given a norm* $||\cdot||_{\mathbb{C}}$ *on* $\mathbb{C}^n$*, we can define* $||\cdot||_{\mathbb{R}}$ *on* $\mathbb{R}$ *and let* $||x||_{\mathbb{R}} = ||x||_{\mathbb{C}}$ *for* $x \in \mathbb{R}^n$*.*

Therefore, we obtain 2 subordinate matrix norms given

$$||A||_{\mathbb{C}} = \sup_{\substack{x\in\mathbb{C}^n \\ x\neq 0}} \frac{||Ax||_{\mathbb{C}}}{||x||_{\mathbb{C}}}, ||A||_{\mathbb{R}} = \sup_{\substack{x\in\mathbb{R}^n \\ x\neq 0}} \frac{||Ax||_{\mathbb{R}}}{||x||_{\mathbb{R}}}$$

these two expressions coincide for vector norms $||\cdot||_1, ||\cdot||_2, ||\cdot||_\infty$*, but not for general norms.*

So we follow the convention that, from now on, all subordinate norms (even for real matrices) are defined over $\mathbb{C}^n$*.*

(2) $A \mapsto \rho(A) = \max\{\lambda_i\}$*, the spectral radius of* A *is not a norm on* $\mathbb{C}^{n\times n}$*.*

To show this, we note that $\rho\left(\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}\right) = 0$ *for a non-zero matrix.*

A useful lemma that we will repeatedly use today is

Lemma 9.1. *If* $U \in M_n(\mathbb{C})$ *is unitary, then* $\forall A \in \mathcal{M}_n(\mathbb{C})$

$$||UA||_2^2 = ||AU||_2^2 = ||A||_2^2.$$

Proof. Note that U unitary implies

$$||UA||_2^2 = \sup_{\substack{x\in\mathbb{C}^n \\ x\neq 0}} \frac{\langle U^*UAx, Ax\rangle}{\langle x, x\rangle} = \sup_{\substack{x\in\mathbb{C}^n \\ x\neq 0}} \frac{\langle Ax, Ax\rangle}{\langle x, x\rangle} = ||A||_2^2.$$

For right multiplication we either use transpose or we can do the following: let $y = Ux$ then we have

$$||y||_2 = ||Ux||_2 = \langle Ux, Ux\rangle^{\frac{1}{2}} = \langle x, x\rangle^{\frac{1}{2}} = ||x||_2$$

and thus

$$||AU||_2^2 = \sup_{\substack{x\in\mathbb{C}^n \\ x\neq 0}} \frac{||AUx||_2^2}{||x||_2^2} = \sup_{\substack{y\in\mathbb{C}^n \\ y\neq 0}} \frac{||Ay||_2^2}{||U^{-1}y||_2^2} = \sup_{\substack{y\in\mathbb{C}^n \\ y\neq 0}} \frac{||Ay||_2^2}{||y||_2^2} = ||A||_2^2.$$

□

Corollary 9.2. *If* $A \in \mathcal{M}_n(\mathbb{C})$ *is normal, then*

$$||A||_2 = \rho(A).$$

Proof. Recall that for A normal, there exists unitary U such that $A = UDU^*$ where $D = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$.

Now, the above lemma gives us

$$||A||_2 = ||UDU^*||_2 = ||\operatorname{diag}(\lambda_1, \dots, \lambda_n)||_2 = \rho(A).$$

□

Does the spectral radius $\rho(A)$ relates to other more general matrix norms? The answer is encoded in the following proposition.

Proposition 9.3. *Let $A \mapsto ||A||$ be a matrix norm defined on $\mathcal{M}_n(\mathbb{C})$. Then $\rho(A) \leq ||A||$ for all $A \in \mathcal{M}_n(\mathbb{C})$.*

Proof. Given $A \in \mathcal{M}_n(\mathbb{C})$, let $\lambda \in \mathbb{C}$ be an eigenvalue with $|\lambda| = \rho(A)$ and associated eigenvector x. Then we choose $y \in \mathbb{C}^n$ such that $xy^* \neq 0$. Then we have

$$(Ax)y^* = \lambda(xy^*) \quad \Rightarrow \quad ||\lambda xy^*|| = |\lambda| \cdot ||xy^*|| = ||Axy^*|| \leq ||A|| \cdot ||xy^*||$$

which yields $|\lambda| \leq ||A||$ due to our choice of y. □

Corollary 9.4. *For $A \in \mathcal{M}_n(\mathbb{C})$,*

$$\rho(A) \leq \min\{||A||_1, ||A||_\infty\} = \min\left\{ \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}|, \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}| \right\}$$

where the equality is attained using the corresponding base vector.

The above result is kind of surprising. The natural thing to ask now is what about bounding the spectral radius from below.

Proposition 9.5. *Given $A \in \mathcal{M}_n(\mathbb{C})$ and $\varepsilon > 0$, there exists a subordinate matrix norm $B \mapsto ||B||_{A,\varepsilon}$ such that*

$$||A||_{A,\varepsilon} \leq \rho(A) + \varepsilon.$$

Proof. By Schur factorization, we can find unitary U such that $T = U^*AU$ where T is upper triangular with diagonal entries the eigenvalues of A.

For $\delta > 0$, set $D_\delta = \operatorname{diag}(1, \delta, \dots, \delta^{n-1})$ and define

$$T_\delta := D_\delta^{-1} T D_\delta = D_\delta^{-1} U^* A U D_\delta = (U D_\delta)^{-1} A (U D_\delta).$$

Then, with the convenient fact that right multiplying a diagonal matrix is to multiply each column by the associated diagonal term, and left multiplying is to multiply each row, we get the following:

$$T_\delta = D_\delta^{-1} T D_\delta = D_\delta^{-1} \begin{pmatrix} t_{11} & \delta t_{12} & \delta^2 t_{13} & \dots & \delta^{n-1} t_{1n} \\ & \delta t_{22} & \delta^2 t_{23} & \dots & \delta^{n-1} t_{2n} \\ & & \ddots & \ddots & \vdots \\ & & & & \delta_{n-1} t_{nn} \end{pmatrix} = \begin{pmatrix} t_{11} & \delta t_{12} & \delta^2 t_{13} & \dots & \delta^{n-1} t_{1n} \\ & t_{22} & \delta t_{23} & \dots & \delta^{n-2} t_{2n} \\ & & \ddots & \ddots & \vdots \\ & & & & t_{nn} \end{pmatrix}.$$

Now, given $\varepsilon > 0$, we can choose small δ such that

$$\sum_{j=i+1}^{n} \delta^{j-i} |t_{ij}| < \varepsilon, 1 \le i \le n$$

i.e. the sum of non-diagonal terms in each row is less than ε. This means that

$$||T_\delta||_\infty \le \rho(A) + \varepsilon.$$

Now we can define the matrix norm $|| \cdot ||$ by

$$||B|| := ||(U D_\delta)^{-1} B (U D_\delta)||_\infty$$

where U, δ are chosen as above, depending on A, ε. So it makes sense to say that the norm really depends on A and ε.

It is indeed a subordinate norm since it's generating vector norm is

$$x \mapsto ||(U D_\delta)^{-1} x||_\infty$$

which can be shown easily.

And as we've shown above,

$$||A|| = ||(U D_\delta)^{-1} (U D_\delta) T_\delta (U D_\delta)^{-1} (U D_\delta)||_\infty = ||T_\delta||_\infty \le \rho(A) + \varepsilon$$

so it satisfy the condition in proposition. □

Remark 9.2.

This proposition may seem quite useless since the norm has to be dependent on A. But since we know all norms are equivalent, we really can use this to bound $\rho(A)$ in other norms, which is extremely powerful.

Similar definitions for matrix norms can be made on $A \in \mathcal{M}_{m,n}(\mathbb{K})$:

Def 9.1.

- *The generalized Forbinius norm and Holder norms:*

$$||A||_F = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2 \right)^{\frac{1}{2}}$$

and of course we can change this 2 to q to get the Holder's q norm, even for $q = \infty$.

- *(As we've talked last time) the operator norm*

$$||A||_{m,n} = \sup_{\substack{y\in\mathbb{C}^n \\ y\neq 0}} \frac{||Ax||_m}{||x||_n}.$$

9.2. **An approximation.**

We now put some of the discussion into practice. Since norms can be used to discuss how close or far away a matrix is from another matrix, we can find best approximation in this sense.

The approximation we are doing today is the best approximation of matrices in $\mathcal{M}_{m,n}(\mathbb{C})$ among fixed rank competitors.

Proposition 9.6. *Suppose $A = V\tilde{\Sigma}U^*$ is the SVD for $A \in \mathcal{M}_{m,n}(\mathbb{C})$ with r non-zero singular values of A arranges from large to small.*

Then for each $1 \leq k \leq r$, the matrix

$$A_k = \sum_{i=1}^{k} \mu_i v_i u_i^*$$

satisfies

$$||A - A_k||_2 \leq ||A - X||_2$$

for all $X \in \mathcal{M}_{m,n}(\mathbb{C})$ with $\operatorname{rank}(X) = k$. Thus it is a best approximation in this sense. Moreover, we can explicitly compute

$$||A - A_k||_2 = \mu_{k+1}.$$

Proof. As is proven in homework 3, $A = \sum_{i=1}^{r} \mu_i v_i u_i^*$. Thus

$$A - A_k = \sum_{i=k+1}^{r} \mu_i v_i u_i^* = \begin{pmatrix} | & & | \\ v_{k+1} & \dots & v_r \\ | & & | \end{pmatrix} \begin{pmatrix} \mu_{k+1} & & \\ & \ddots & \\ & & \mu_r \end{pmatrix} \begin{pmatrix} — & u_{k+1}^* & — \\ & \vdots & \\ — & u_r^* & — \end{pmatrix}$$

and thus by setting $D = \operatorname{diag}(0, \dots, 0, \mu_{k+1}, \dots, \mu_r, 0, \dots, 0)$ we have

$$A - A_k = VDU^*.$$

Now by lemma 9.1 we have

$$||A - A_k||_2 = ||VDU^*||_2 = ||D||_2 = \mu_{k+1}.$$

WLOG we do the following with $m = n$ (the general case is similar).

To show that the best approximation result we first note that

$$||Ax||_2 \geq \mu_{k+1}||x||_2 \quad \text{for all} \quad x \in \operatorname{span}\{u_1, \dots, u_{k+1}\}. \tag{9.1}$$

The reason is that for x in the above span, we can write $x = \sum_{i=1}^{k+1} \alpha_i u_i$, which implies

$$U^* x = U^* \left(\sum_{i=1}^{k+1} \alpha_i u_i \right) = \sum_{i=1}^{k+1} \alpha_i U^* u_i = \sum_{i=1}^{k+1} \alpha_i e_i = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_k + 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

and thus

$$\tilde{\Sigma} U^* x = \begin{pmatrix} \mu_1 \alpha_1 \\ \vdots \\ \mu_{k+1} \alpha_k + 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

so

$$||Ax|| = ||V \tilde{\Sigma} U^* x||_2 = ||\tilde{\Sigma} U^* x||_2 \geq \mu_{k+1} ||x||_2.$$

For the sake of contradiction we now suppose $X \in \mathcal{M}_n(\mathbb{C})$ is a matrix of rank k with $||A - X||_2 < ||A - A_k||_2$. Then, $\forall x \in \ker(X) \backslash \{0\}$ we have

$$\frac{||Ax||_2}{||x||_2} = \frac{||(A - X)x||_2}{||x||_2} \leq \sup_{y \neq 0} \frac{||(A - X)y||_2}{||y||_2} = ||A - X||_2$$

since $Xx = 0$ and the last equality is by definition, then we have

$$||Ax||_2 \leq ||A - X||_2 ||x||_2 < ||A - A_k||_2 ||x||_2 = \mu_{k+1} ||x||_2 \text{ for } x \in \ker(X) \backslash \{0\} \quad (9.2)$$

So (9.1) and (9.2) form a contradiction if $\text{span}\{u_1, \ldots, u_{k+1}\} \cap \ker(X) \neq \{0\}$.

The above statement follows by a simple dimension argument: If the above two subspaces have only a trivial intersection, then their sum would be a subspace of dimension $n + 1$ of C^n, which is impossible, so we are done.

To sum up, choosing $x \in \text{span}\{u_1, \ldots, u_{k+1}\} \cap \ker(X)$ and $x \neq 0$ we get

$$\mu_{k+1} ||x||_2 \leq ||Ax||_2 < \mu_{k+1} ||x||_2$$

a contradiction! So we are done.

□

10. 10/31: Sequence and series of matrices; Algorithms for matrix computation

We work in $\mathcal{M}_n(\mathbb{C})$.

10.1. **Sequence and series of matrices.**

Def 10.1. *A sequence of matrices A_i for $i \geq 1$ is convergent to a limiting matrix $A \in \mathcal{M}_n(\mathbb{C})$ if, for a matrix norm $||\cdot||$,*

$$\lim_{i\to\infty} ||A_i - A|| = 0.$$

Note that since all norms are equivalent, this really works for any norm if it works for one. Therefore, we are justified to write

$$\lim_{i\to\infty} A_i = A.$$

Def 10.2. *A matrix series $\sum_{j=0}^{\infty} A_j$ is said to be convergent if the sequence $S_i = \sum_{j=0}^{i} A_j$ converges to some matrix S_*.*

Def 10.3. *Given a sequence $(a_i)_{i\geq 1}$ in $\mathbb{C}$, the associated matrix power series is*

$$\sum_{i=0}^{\infty} a_i A^i.$$

Remark 10.1. *If $\sum_{i=0}^{\infty} a_i A^i$ converges for some A, then $a_i A^i \to 0$ as $i \to \infty$, whereas the converse is not true.*

Proposition 10.1. *For $A \in \mathcal{M}_n(\mathbb{C})$, the following are equivalent:*

(i) $A^i \to 0$ *as* $i \to \infty$.
(ii) $A^i x \to 0$ *as* $i \to \infty$ *for all* $x \in \mathbb{C}^n$.
(iii) $\rho(A) < 1$.
(iv) There is a subordinate matrix norm $||\cdot||$ *with* $||A|| < 1$.

Proof. (idea)

$(i) \Rightarrow (ii)$: This is just because $||A^i x|| \leq ||A^i|| \cdot ||x||$ for the induced matrix norm.

$(ii) \Rightarrow (iii)$: If $\rho(A) \geq 1$, then we can find the corresponding eigenvector such that $A^i x = \lambda^i x$ does not go to 0. Contradiction.

$(iii) \Rightarrow (iv)$: By proposition 9.5, this is direct.

$(iv) \Rightarrow (i)$: We choose the norm as in (iv), then, since it's a matrix norm we have $||A^i|| \leq ||A||^i \to 0$. □

Theorem 10.2. *Suppose* $(a_i) \in \mathbb{C}$ *defines a power series* $\sum_{i=0}^{\infty} az^i$ *in* $\mathbb{C}$ *with radius of convervgence* $R > 0$. *Then for any* $A \in \mathcal{M}_n(\mathbb{C})$ *with* $\rho(A) < R$, *the series* $\sum_{i=0}^{\infty} aA^i$ *converges in* $\mathcal{M}_n(\mathbb{C})$.

Proof. Note that $\rho(A) < R$ implies that we can choose a subordinate matrix norm $||\cdot||$ with $||A|| < R$ (again, by proposition 9.5). We use this norm in the following.

Now, to show the convergence of the series, it suffices to show that the sequence of partial sums

$$S_i = \sum k = 1^i a_k A^k$$

converges. And since we are in Euclidean space, we only need to show that it's Cauchy.

But this follows from the simple bound

$$|| \sum_{k=j+1}^{i} a_k A^k || \leq \sum_{k=j+1}^{i} |a_k| ||A||^k$$

and the fact that power series in $\mathbb{C}$ are absolutely convergent in the interior of the disk of convergence. (note that $||A||$ is just a complex number under this specific norm.) □

Def 10.4. *If* $f : \mathbb{C} \to \mathbb{C}$ *is analytic on* $\{z \in \mathbb{C} || z| < R, R > 0\}$, *i.e. it can be written as a power series*

$$f = \sum_{i=0}^{\infty} a_i z^i \ \ for \ \ |z| < R.$$

Then, for $A \in \mathcal{M}_n(\mathbb{C})$ *with* $\rho(A) < R$, *we define* $f(A) = \sum_{i=0}^{\infty} a_i A^i$.

Example 10.1.

(1) We can thus define the exponential of a matrix (since $x \mapsto e^x$ *is entire (analytic on the whole* $\mathbb{C}$*)). Then we can define*

$$e^A = \sum_{i=0}^{\infty} \frac{A^i}{i!}.$$

(2) As another natural example, the analytic function

$$z \mapsto \frac{1}{1-z} = 1 + z + z^2 + \ldots$$

is analytic in the unit disc $\{z; |z| < 1\}$.

Proposition 10.3. *Suppose* $A \in \mathcal{M}_n(\mathbb{C})$ *is such that* $\rho(A) < 1$, *then* $(I - A)$ *is a non-singular matrix in* $\mathcal{M}_n(\mathbb{C})$ *with*

$$(I - A)^{-1} = \sum_{i=0}^{\infty} A^i.$$

Proof. We've showed earlier that $A^i \to 0$ as $i \to \infty$ since $\rho(A) < 1$. Moreover, for any $p \geq 1$, we have that the partial sum to p

$$(I - A)\sum_{i=0}^{p} A^i = I - A + A - A^2 + \cdots + A^p - A^{p+1} = I - A^{p+1}$$

which means that if we take the limit as $p \to \infty$,

$$\left(\sum_{i=0}^{\infty} A^i\right)(I - A) = I - 0.$$

□

Remark 10.2. *Proposition 10.3 implies that* $\{A \in \mathcal{M}_n(\mathbb{C}) : \det(A) \neq 0\}$ *is an open set in* $\mathcal{M}_n(\mathbb{C})$.

Indeed, for any subordinate norm $||\cdot||$ *and non-singular* A*, if* B *satisfies*

$$||B - A|| < \frac{1}{||A^{-1}||}$$

then

$$\rho(A^{-1}(A - B)) \leq ||A^{-1}(A - B)|| \leq ||A^{-1}|| \cdot ||(A - B)|| < 1$$

such that $I - A^{-1}(A - B)$ *is invertible, and so is*

$$A(I - A^{-1}(A - B)) = A - (A - B) = B.$$

(the idea is just to see that any B *close to* A*,* B *is invertible.)*

10.2. **Algorithms for Matrix Multiplication.**

Def 10.5. *We call the complexity of a matrix algorithm the number of multiplication or division required to execute.*

Remark 10.3.

(1) We do not keep track of additions or subtractions, or square roots unless we note otherwise.

A notation we will use is that, for a problem of size n (either a matrix or a vector),

$$\mathrm{Nop}(n) = \#\ \textit{of multiplications and divisions required.}$$

(2) We're most interested in the asymptotic behavior of Nop(n) for large n. In particular, we are most interested in its leading order term.

(3) These counts are tied to a specific algorithm/computational procedure involving computations with real and complex numbers.

We look at a classical, basic (and still ongoing!) algorithm: Matrix Multiplication.

A regular computation process has n multiplications in practicing each term in the output matrix, and there's n^2 terms, so the complexity is $O(n^3)$.

A major break through is made by Strassen, 1969, that has only $\text{Nop}(n) = O\left(n^{\log_2 7}\right)$. We prove this by showing this for only 2 by 2 matrices.

Lemma 10.4. *We can compute*

$$AB_{2x2} = \begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22} + b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{pmatrix}$$

with 7 multiplication and 18 addition, instead of 8 multiplications.

Proof. (idea) We can write out the needed matrix with a new set of multiplications. We only list a few of them:

$$\begin{cases} m_1 = (a_{12} - a_{22})(b_2 1 + b_{22}) \\ m_2 = (a_{11} + a_{22})(b_{11} + b_{22}) \\ m_4 = (a_{11} + a_{12})b_{22} \\ m_5 = a_{11}(b_{12} - b_{22}) \end{cases}$$

and with this construction we can write

$$AB_{11} = m_1 + m_2 - m_4 + m_5$$

and in the similar manner we use seven multiplications and some extra additions to do the computation. □

This results also holds for $2x2$ block matrices. Let's see how it may apply: for $A \in \mathcal{M}_{2^k}(\mathbb{C})$ and split this matrix into 4 blocks of size $2^{k-1}x2^{k-1}$ and apply lemma 10.4 including the additions, we get

$$\text{Nop}(2^k) = 7\,\text{Nop}(2^{k-1})$$

with $18(2^k)$ additions. Then, an induction argument implies

$$\text{Nop}(2^k) = 7^k(\text{Nop}(1)) + 18\sum_{i=0}^{k-1} 7^i 4^{k-1-i} \leq 7^k(\text{Nop}(1) + 6)$$

and thus

$$\text{Nop}(n) \leq c \cdot n^{\log_2 7}.$$

11. 11/7: More on Straussen; Linear systems of equations; matrix conditioning.

11.1. **More on straussen.**

The Straussen's method (and its improvements) gives improvements beyond $C = AB$.

Def 11.1. *Let Nop(n) denote the complexity of the best algorithm performing a matrix operation(possibly unknown). We call the bound Nop(n)$\leq cn^\alpha$ for $n \geq 0$ with c, α independent of n. The Asymptotic complexity of the operation and write this as Nop(n)$= O(n^\alpha)$.*

Theorem 11.1. *Each of the following as the same asymptotic complexity in the sense that if any has an algorithm computing it with complexity $O(n^\alpha), \alpha \geq 2$, then so do the other three:*

(i) Given A, B find $C = AB$.
(ii) Find the inverse of A.
(iii) Find the determinant of A.
(iv) Given A, b, find the solution to $Ax = b$.

Remark 11.1.
- *Gaussian elimination can get us all of the above, and Gaussian elimination can be done in at most $O(n^3)$.*
- *We won't prove all of these equivalences. We will, however, sketch the equivalence between (i) and (ii). The hard part of this prove is that we don't have a pecific algorithm for either.*

Proof. (ii)⇒(i):

Note that

$$\begin{pmatrix} I & A & 0 \\ 0 & I & B \\ 0 & 0 & I \end{pmatrix}^{-1} = \begin{pmatrix} I & -A & AB \\ 0 & I & -B \\ 0 & 0 & I \end{pmatrix}$$

we can find AB by finding an inverse. So, we can compute AB within $O((3n)^\alpha) = O(n^\alpha)$ complexity.

(i)⇒(ii):

Note that

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} + A^{-1}B\Delta^{-1}CA^{-1} & -A^{-1}B \\ -\Delta^{-1}CA^{-1} & \Delta^{-1} \end{pmatrix}$$

where $\Delta = D - CA^{-1}B$. This operation requires inverting 2 smaller $n/2$ by $n/2$ matrices and six multiplications. We obtain

$$I(2n) \leq 2I(n) + 6M(n)$$

for which we've neglected additions. If we do this iteratively we will get

$$I(2^k) \leq 2^k I(1) + 6\sum_{i=1}^{k-1} 2^{k-i-1} M(2^i) \leq c2^k + c' \sum_{i=1}^{k-1} 2^{k-i-1+\alpha_i}$$

since $\alpha \geq 2$

$$\Rightarrow I(2^k) \leq c2^{\alpha k}$$

$$\Rightarrow I(n) = O(n^\alpha).$$

□

11.2. **Overview of linear systems of equations.**

We are interested in the following problem: for $A \in \mathcal{M}_{n,p}(\mathbb{K})$, $b \in \mathbb{K}^n$, find $x \in \mathbb{K}^p$ such that $Ax = b$.

Theorem 11.2. *For the square case ($n = p$), A is non-singular $\iff$ $\exists$ a unique solution $x = A^{-1}b$ to $Ax = b$;*

Likewise, if A is singular, then either

- *$b \in \text{range}(A)$ so that $\exists x_0$ such that the solution set to the system is*

$$\{x + v | v \in \ker(A)\}$$

- *or that $b \notin \text{range}(A)$ and $Ax = b$ has no solution.*

Proposition 11.3. *(Cramer's formula) Suppose $A \in \mathcal{M}_n(\mathbb{R})$ is non-singular with columns given by*

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

and consider $Ax = b$. Then the solution x in R^n is given by

$$x_i = \frac{\det \begin{pmatrix} | & & | & | & | & | \\ a_1 & \dots & a_{i-1} & b & a_{i+1} & \dots & a_n \\ | & & | & | & | & | \end{pmatrix}}{\det(A)}$$

Proof. The main idea is to recall that one characterization of determinant is an alternating n-linear functions of the column vectors.

In practice, writing

$$Ax = A(xe_1 + \cdots + xe_n) = x_1 Ae_1 + \cdots + x_n Ae_n = x_1 a_1 + \cdots + x_n a_n$$

and the condition $Ax = b$ implies that

$$\det \begin{pmatrix} | & & | & | & | & | \\ a_1 & \dots & a_{i-1} & b & a_{i+1} & \dots & a_n \\ | & & | & | & | & | \end{pmatrix} = \sum_{j=1}^{n} x_j \det \begin{pmatrix} | & & | & | & | & | \\ a_1 & \dots & a_{i-1} & a_j & a_{i+1} & \dots & a_n \\ | & & | & | & | & | \end{pmatrix}$$

but if we look at the expression we will find that only when $i = j$, the summand is not 0. Hense we have that

$$\det\begin{pmatrix} | & & | & | & | & | \\ a_1 & \dots & a_{i-1} & b & a_{i+1} & \dots & a_n \\ | & & | & | & | & | \end{pmatrix} = x_i \det(A)$$

which gives us what we want. □

Remark 11.2. *(1) This is not efficient computationally. So it requires $n!$ operations of using recursive definition of the determinant.*

(2) In fact, we'll see, that we can compute the solution to linear systems with the same complexity as computing the determinant. So Cramer's rule is never a practical computational rule.

Now let's look at some simple matrices:

(1) If A is diagonal, then solving $Ax = b$ requires n multiplication/division.
(2) If A is unitary, then solving $Ax = b$ requires n^2 multiplication/division since we only need to compute $x = A^{-1}b = A^*b$.
(3) If A is lower triangular, then solving $Ax = b$ requires $O(n^2)$ multiplication/division. The algorithm is the forward substitution:

$$\begin{aligned} x_1 &= \frac{b_1}{a_{11}} \\ x_2 &= \frac{b_2 - a_{21}x_1}{a_{22}} \\ &\vdots \\ x_n &= \frac{b_n - a_{n1}x_1 - a_{n2}x_2 - \cdots - a_{nn}x_n}{a_{nn}} \end{aligned}$$

which requires $1 + 2 + \cdots + m = O(n^2)$ complexity.
(4) If A is upper triangular, we use the backward substitution, which is the same as the forward substitution above except that we start backwards up.

Remark 11.3. *(1) To solve $Ax = b$, we do not have to compute A^{-1} and in amny cases that is too computationally expensive to be worth doing.*

(2) If we can solve $Ax = b$, we can compute A^{-1} as follows: For $i = 1, \dots, n$, solve $Ax = e_i \rightsquigarrow x_i \in \mathbb{K}^n$ and

$$A^{-1} = \begin{pmatrix} | & & | \\ x_1 & \dots & x_n \\ | & & | \end{pmatrix}$$

this is true because

$$A^{-1} = A \begin{pmatrix} | & & | \\ x_1 & \dots & x_n \\ | & & | \end{pmatrix} = \begin{pmatrix} | & & | \\ Ax_1 & \dots & Ax_n \\ | & & | \end{pmatrix} = \begin{pmatrix} | & & | \\ e_1 & \dots & e_n \\ | & & | \end{pmatrix} = I.$$

(3) The spectral information about A that we've found when putting A into triangular or diagonal form won't play a role in numerical solution for $Ax = b$. The computation of eigenvalues is in general a harder task computationally.

(4) Some readings to do are: Sec 5.2-Sec 5.31.

11.3. **Matrix conditioning.**

In characterizing the numerically behavior of various algorithms, it'll be useful to understand how perturbations in A, b affect the solution x to $Ax = b$.

Our setting is this: For $\varepsilon > 0, B \in M_n(\mathbb{K}), \gamma \in \mathbb{K}^n$, Set $A_\varepsilon = A + \varepsilon B, b_\varepsilon = b + \varepsilon\gamma$. Recall that the set of non-singular *nxn* matrices is open, thus A_ε is non-singular for small ε and we can let x_ε be the solution to

$$A\varepsilon x_\varepsilon = b_\varepsilon.$$

Now we can notice that

$$A_\varepsilon^{-1} = (I + \varepsilon A^{-1}B)^{-1}A^{-1}$$

this is because

$$(I + \varepsilon A^{-1}B)^{-1}A^{-1}(A + \varepsilon B) = (I + \varepsilon A^{-1}B)^{-1}(I + \varepsilon A^{-1}B).$$

Further, we see that

$$(I + \varepsilon A^{-1}B)^{-1} = I - \varepsilon A^{-1}B + O(\varepsilon^2)$$

by computation up to first order.

Hence we have

$$\begin{aligned} x_\varepsilon &= (I + \varepsilon A^{-1}B)^{-1}A^{-1}(b + \varepsilon\gamma) \\ &= I - \varepsilon A^{-1}B + O(\varepsilon^2)(x + \varepsilon A^{-1}\gamma) \\ &= x + \varepsilon A^{-1}(\gamma - Bx) + O(\varepsilon^2). \end{aligned}$$

This gives that for a vector norm $||\cdot||$ and its corresponding subordinate matrix norm the following estimate:

$$||x_\varepsilon - x|| \leq \varepsilon||A^{-1}(\gamma - Bx)|| + O(\varepsilon^2) \leq \varepsilon||A^{-1}||(||\gamma|| + ||B|| \cdot ||x||)|| + O(\varepsilon^2)$$

where since $||b|| = ||Ax|| \leq ||A||||x||$ plugging it in we have

$$||x_\varepsilon - x|| \leq \varepsilon||x|| \cdot ||A|| \cdot ||A^{-1}|| \left(\frac{||\gamma||}{||b||} + \frac{||B||}{||A||} \right) || + O(\varepsilon^2). \tag{11.1}$$

Def 11.2. *The condition number of a matrix $A \in \mathcal{M}_n(\mathbb{K})$ relative to a subordinate matrix norm $||\cdot||$ is given by* $\text{cond}(A) = ||A|| \cdot ||A^{-1}||$.

Remark 11.4. *(1)* $\mathrm{cond}(A) \geq 1$ *by properties of matrix norm.*

(2) The above perturbation bound is

$$\frac{||x_\varepsilon - x||}{||x||} \leq \mathrm{cond}(A)\left(\frac{||A_\varepsilon - A||}{||A||} + \frac{||b_\varepsilon - b||}{||b||}\right) + O(\varepsilon^2)$$

since $||A_\varepsilon - A|| = \varepsilon||B||$, $||b_\varepsilon - b|| = \varepsilon||\gamma||$ *and the estimation (11.1).*

i.e. the relative error in x (up to the first order of ε*) is controlled by the product of* $\mathrm{cond}(A)$ *and the relative errors in A and b.*

This tells us that the condition number $\mathrm{cond}(A)$ *measures how much errors in* A, b *are amplified for* $Ax = b$*.*

Proposition 11.4. *For* $A \in \mathcal{M}_n(\mathbb{R})$ *non-singular,* $b \in \mathbb{R}^n$*,* $b \neq 0$*, and* $\delta b \in \mathbb{R}^n$*. Note that here* $\delta b \neq \delta \cdot b$*, in fact there's no* δ *as a number defined.*

If x *solves* $Ax = b$ *and* $x + \delta x$ *solves* $A(x + \delta x) = b + \delta b$*, then we have*

$$\frac{||\delta x||}{||x||} \leq \mathrm{cond}(A)\frac{||\delta b||}{||b||}.$$

Proof. The proof is actually quite short.

Note that $AA(x + \delta x) = b + \delta b$ means that $A\delta x = \delta b$. Then by matrix norm we have

$$||\delta x|| \leq ||A|| \cdot ||\delta b||.$$

Recall that $||b|| = ||Ax|| \leq ||A|| \cdot ||x||$ we then have

$$\frac{||\delta x||}{||x||} \leq \frac{||A^{-1}|| \cdot ||\delta b|| \cdot ||A||}{||b||} = \mathrm{cond}(A)\frac{||\delta b||}{||b||}.$$

□

12. 11/9: More on condition numbers

12.1. On properties of condition numbers.

Remark 12.1. *(on Proposition 11.4) The evaluation in proposition 11.4 is optimal, i.e. given invertible $A \in \mathcal{M}_n(\mathbb{R})$, $\exists x_0 \neq 0$ such that*

$$||Ax_0|| = ||A|| \cdot ||x_0||$$

and $x_1 \neq 0$ such that

$$||A^{-1}x_1|| = ||A^{-1}|| \cdot ||x_1||.$$

Now, we only need to take $b = Ax_0, \delta_b = x_1$, which gives us $x = x_0$ and

$$\delta_x = A^{-1}\delta b = A^{-1}x_1.$$

So

$$\frac{||\delta x||}{||x||} = \frac{||A^{-1}x_1||}{||x_0||} = \frac{||A^{-1}|| \cdot ||x_1|| \cdot ||A||}{||A|| \cdot ||x_0||} = \text{cond}(A)\frac{||\delta b||}{||b||}.$$

Similarly, if it is not b, but A that is fluctuating, $\text{cond}(A)$ is still a handful tool.

Proposition 12.1. *For $A \in \mathcal{M}_n(\mathbb{R})$ non-singular, let $b \in \mathbb{R}^n$ that is non-zero, and $Ax = b$, further, let $\delta_A \in \mathcal{M}_n(\mathbb{R})$. If δ_x is such that*

$$(A + \delta_A)(x + \delta_x) = b$$

then

$$\frac{||\delta_x||}{||x + \delta_x||} \leq \text{cond}(A)\frac{||\delta_A||}{||A||}$$

with optimality similar to above.

Proof. Note that $A\delta_x + \delta_A(x + \delta_x) = b - Ax = 0$. This means

$$||\delta_x|| = ||A^{-1}\delta_A(-x - \delta_x)|| \leq ||A^{-1}|| \cdot ||\delta_A|| \cdot ||x + \delta_x|| \leq \text{cond}(A)\frac{1}{||A||}||\delta_A|| \cdot ||x + \delta_x||$$

which thus means

$$\frac{||\delta_x||}{||x + \delta_x||} \leq \text{cond}(A)\frac{||\delta_A||}{||A||}.$$

□

Note that we use $\frac{||\delta_x||}{||x + \delta_x||}$ rather than $\frac{||\delta_x||}{||x||}$. This is just to be easier and it doesn't affect much.

Remark 12.2.

(1) It's most common to use $\text{cond}(A) = \text{cond}_p(A) = ||A||_P||A^{-1}||_p$, where $p = 1, 2, \infty$, and the norm is the subordinate matrix norm.

(2) $A \in \mathcal{M}_n(\mathbb{R})$ means

$$\frac{1}{n}\operatorname{cond}_2(A) \leq \operatorname{cond}_1(A) \leq n \operatorname{cond}_2(A)$$

and

$$\frac{1}{n}\operatorname{cond}_\infty(A) \leq \operatorname{cond}_2(A) \leq n \operatorname{cond}_\infty(A)$$

which means

$$\frac{1}{n^2}\operatorname{cond}_1(A) \leq \operatorname{cond}_\infty(A) \leq n^2 \operatorname{cond}_1(A).$$

(3) If $A \in \mathcal{M}_n(\mathbb{C})$ is non-singular, then $\operatorname{cond}(A) = \operatorname{cond}(A^{-1})$.

(4) For $A \in \mathcal{M}_n(\mathbb{C})$, $\alpha \in \mathbb{C}$, $\alpha \neq 0$, then $\operatorname{cond}(\alpha A) = \operatorname{cond}(A)$.

(5) For $A \in \mathcal{M}_n(\mathbb{C})$, we have that

$$\operatorname{cond}_2(A) = \frac{\mu_1}{\mu_n}$$

where μ_1 is the largest singular value of A and μ_n is the smallest.

(6) If $A \in \mathcal{M}_n(\mathbb{C})$ is normal, then

$$\operatorname{cond}_2(A) = \rho(A)\rho(A^{-1}) = \frac{|\lambda_{max}|}{|\lambda_{min}|}.$$

(7) For $U \in \mathcal{M}_n(\mathbb{C})$ unitary, then $\operatorname{cond}_2(U) = 1$ and for $A \in \mathcal{M}_n(\mathbb{C})$

$$\operatorname{cond}_2(AU) = \operatorname{cond}_2(UA) = \operatorname{cond}_2(A).$$

(8) Since $\rho(A) \leq ||A||$ for any matrix norm, thus

$$\rho(A)\rho(A^{-1}) \leq ||A|| \cdot ||A^{-1}|| = \operatorname{cond}(A).$$

So if A is normal, then $\operatorname{cond}_2(A) \leq \operatorname{cond}(A)$ for any norm.

(9) We say that $A \in \mathcal{M}_n(\mathbb{C})$ is well-conditioned if $\operatorname{cond}(A)$ is close to 1.
It is ill-conditioned if not.

(10) Unitary matrices are well conditioned.

(11) There are 2 interpretations of $\operatorname{cond}_2(A)$:
   (i) Geometric: Recall that $A(\mathbb{S}^{n-1})$ is an ellipsoid in $\mathbb{R}^n$ and the length of its semi-axis are given by the singular values. Then $\operatorname{cond}_2(A)$ represents how flat the ellipsoid is.
   (ii) Analytic configuration: $\operatorname{cond}_2(A)$ represents the relative distance from A to the class of singular matrices

$$S_n(\mathbb{C}) := \{B \in \mathcal{M}_n(\mathbb{C}) | B \text{ non-singular}\}.$$

   This is encoded in Lemma 12.2 below.

(12) One can also define $\operatorname{cond}(A)$ for a non-square or singular matrix. The idea is to use psudo-inverse $A^\dagger$:

$$\operatorname{cond}(A) = ||A|| \cdot ||A^\dagger||.$$

Now we prove the lemma mentioned in (11) above.

Lemma 12.2. *For $A \in \mathcal{M}_n(\mathbb{C})$ non-singular, we have*

$$\frac{1}{\text{cond}_2(A)} = \inf\left\{ \frac{||A-B||_2}{||A||_2} \,:\, B \in S_n(\mathbb{C}) \right\}.$$

Note that this means how close A is with the set of singular matrices, with proper scaling.

Proof. It suffices to show that

$$\frac{1}{||A^{-1}||_2} = \inf_{B \in S_n(\mathbb{C})} ||A-B||_2$$

and for an equality with an inf in it, it's common trick to prove both sides.

($\leq$): For the purpose of contradiction, if $B \in S_n(\mathbb{C})$ satisfies

$$||A-B||_2 < \frac{1}{||A^{-1}||_2}$$

then we would have

$$||A^{-1}(A-B)||_2 \leq ||A^{-1}||_2 ||A-B||_2 < 1$$

such that due to proposition 10.3

$$I - A^{-1}(A-B) = A^{-1}B$$

would be a non-singular matrix. But then B^{-1} would be also non-singular, which is a contradiction!

($\geq$): Choose $u \in \mathbb{C}^n$ with $||u||_2 = 1$ such that $||A^{-1}||_2 = ||A^{-1}u||_2$.

Now, set

$$B_0 = A - \frac{u(A^{-1}u)^*}{||A^{-1}||_2^2}$$

then

$$B_0(A^{-1}u) = u - \frac{u(A^{-1}u)^*(A^{-1}u)}{||A^{-1}||_2^2} = u - \frac{\langle A^{-1}u, A^{-1}u\rangle - u}{||A^{-1}u||_2^2} = u - u = 0$$

hence, $A^{-1}u$ is in the kernel of B_0, so $B_0 \in S_n(\mathbb{C})$ since $||A^{-1}u|| \neq 0$.

So we want to choose $B = B_0$, and let's check that it satisfies the required inequality.

$$||A_{B_0}||_2 = \frac{||u(A^{-1}u)^*||_2}{||A^{-1}||_2^2} = \left(\max_{x=0} \frac{||u(A^{-1}u)^* x||_2}{||x||_2} \right) \frac{1}{||A^{-1}||_2^2}$$

where the second equality is by noting that the numerator is actually the norm of a matrix. Continue and we get

$$||AB_0||_2 = \max_{x \neq 0} \frac{\langle A^{-1}u, x\rangle ||u||_2}{||x||_2} \frac{1}{||A^{-1}||_2^2} \leq \max_{x \neq 0} \frac{||x||_2 ||A^{-1}u||_2}{||x||_2} \frac{1}{||A^{-1}||_2^2} = \frac{1}{||A^{-1}||_2}$$

where we used $||u||_2 = 1$, $||A^{-1}u||_2 = ||A^{-1}||_2$, and Cauchy's inequality.

So we are done. □

12.2. **Estimations of condition number.**

In general, the condition number of A is hard to compute. But we also don't care that much about computing it precisely. But we do care about approximations of it.

We focus on $p = 2$ here.

Suppose $A \in \mathcal{M}_n(\mathbb{R})$ has SVD factorization $A = V\Sigma U^*$ where the singular values μ_i are arranged in decreasing order, and u_i, v_i denote the columns of U, V, as usual.

Since we can represent arbitrary $x \in \mathbb{R}^n$ by $x = \sum_{i=1}^{n} x_i v_i$. Since $AU = V\Sigma$ we have $Au_i = \mu_i v_i$ and

$$||A||_2 = \max_{x \neq 0} \frac{||Ax||_2}{||x||_2} = \mu_1 = ||Au_1||_2$$

plus for a similar reason we have

$$||A^{-1}||_2 = \max_{x \neq 0} \frac{||A^{-1}x||_2}{||x||_2} = \frac{1}{\mu_n} = ||A^{-1}v_n||_2.$$

We try to (heuristically) approximate these values by restricting the maximums to

$$Q := \{x \in \mathbb{R}^n : x_i \in \{-1, 1\}\}$$

with the idea that if $\alpha \leq ||A||_2$ and $\beta \leq ||A^{-1}||_2$ we have $\alpha\beta \leq \mathrm{cond}_2(A)$.

Since we have the LU decomposition $A = LU$ where L is lower triangular and U upper triangular, we only consider the case for A is triangular due to the following reason: for $A \in \mathcal{M}_n(\mathbb{C})$,

$$\mathrm{cond}(A) = ||LU|| \cdot ||U^{-1}L^{-1}|| \leq ||L|| \cdot ||L^{-1}|| \cdot ||U|| \cdot ||U^{-1}|| = \mathrm{cond}(L)\,\mathrm{cond}(U).$$

Now we suppose A is lower triangular and we first approximate $||A||_2$.

Note that for $Ax = y$ we have $y_i = (Ax)_i = a_{ii}x_i + \sum_{j=1^{i-1}} a_{ij}x_j$ due to the triangular structure of A.

We'll soon find out how this helps to maximize $||Ax||_2$ over Q: we just happily choose $x_1 = 1$ and "greedily" choose $x_i \in \{-1, 1\}$ such that it maximizes $|y_i|$ at each step. Note that this really doesn't guarantee we've maximized $|y|$, but at least we get an approximate

$$||A||_2 \geq \frac{||y||_2}{\sqrt{n}}$$

where $||x||_2 = \sqrt{n}$. Using the forward substitution we get

$$||A^{-1}||_2 \geq \frac{||y||_2}{\sqrt{n}}$$

with a similar process.

Since this y can be concretely computed, it at least gives us a lower bound of $\text{cond}(A)$.

12.3. **Condition numbers and computation.**

When A is ill-conditioned, instead of solving $Ax = b$ it is often useful to solve $C^{-1}Ax = C^{-1}b$ for a suitable preconditioner non-singular matrix C.

The main goal here is to find the best choice of $C \sim A$ such that $C^{-1}A \sim I$. In this sense it's best to have $C = A^{-1}$ since it just solves the system, but to compute A^{-1} is just as hard, so we need to find a balance between.

Next time we will look for some structure to invert A.

13. 11/14: Direct Methods; LU decomposition; Cholesky decomposition

A trick to compute the inverse of a matrix is to truncate the series that a matrix satisfies. For instance, if $||I - A|| < 1$, then

$$A^{-1} = (I - (I - A))^{-1} = I + \sum_{k=1}^{\infty} (I - A)^k$$

and hence we can truncate the sum.

13.1. **Direct Methods.**

We want to use Gaussian Elimination.

Theorem 13.1. *Let $A \in \mathcal{M}_n(\mathbb{C})$ be given, then $\exists M \in \mathcal{M}_n(\mathbb{C})$ non-singular such that $T = MA$ is upper triangular.*

Remark 13.1.

(1) Along the way we will see that we can compute Mb without actually forming b.
(2) To find x, we can now apply back substitution to solve $Tx = MAx = Mb$.

Proof. We construct a sequence of matrices $A^{(k)}$ (it's not a power just bad notation) for $1 \leq k \leq n$, and let $A^{(1)} = A$. Then we write $A^{(k)} = (a_{ij}^k)$.

Now we can first choose a permutation matrix such that $\tilde{A}^{(1)} = P_1 A^{(1)}$ has entry $\tilde{a}_{11}^1 \neq 0$. Note that if no such entry exists, then A already has all zeros below the diagonal in the first column, then we can just let $A^{(2)} = A^{(1)}$.

Now, we left multiply $\tilde{A}^{(1)}$ by

$$E^{(1)} = \begin{pmatrix} 1 & & & \\ -\frac{\tilde{a}_{21}^1}{\tilde{a}_{11}^1} & 1 & & \\ \vdots & & \ddots & \\ -\frac{\tilde{a}_{n1}^1}{\tilde{a}_{11}^1} & & & 1 \end{pmatrix}$$

to form

$$A^{(2)} = E^{(1)} \tilde{A}^{(1)} \begin{pmatrix} \tilde{a}_{11}^1 & \tilde{a}_{12}^1 & \dots & \tilde{a}_{1n}^1 \\ 0 & & & \\ \vdots & & (a_{ij}^2) & \\ 0 & & & \end{pmatrix}$$

where

$$a_{ij}^2 = \tilde{a}_{ij}^1 - \frac{\tilde{a}_{i1}^1}{\tilde{a}_{ii}^1} \tilde{a}_{1j}^1$$

for $2 \leq i, j \leq n$.

Now, we can repeat to get $A^{(n)} = (E^{(n-1)} P_{n-1} \dots E^{(1)} P_1) A = MA$. □

To check that M is non-singular, we can simply compute

$$\det(M) = \prod_{i=1}^{n} \det(P_i) \prod_{i=1}^{n} \det(E^{(i)}) = \pm 1$$

since the det P_i is ± 1 and $\det(E^{(i)}) = 1$.

Remark 13.2.

(1) At each step, we only modify the entries in row k through n and columns k through n. This is because we keep the first $k-1$ rows and columns the same.

(2) Pivoting is essential to numerical stability considerations. Some heuristics and modifications are

- *We would usually be interested in the largest pivot from the point of view of rounding error. Also we might do partial pivoting.*
- *We can even perform row swaps to do optimization even when the entry is already non-zero.*

(3) If no permutations are involved, then Gaussian elimination corresponds to writing $A = LU$ with L lower triangular and U upper triangular. Also, we can directly write down L without multiplying.

Given $A \in \mathcal{M}_n(\mathbb{R})$, we can write look at the first k top left submatrices, call them Δ^k.

13.2. **LU decomposition.**

Proposition 13.2. *Suppose $A \in \mathcal{M}_n(\mathbb{R})$ is such that all diagonal submatrices Δ^k are non-singular, then there exists a unique pair L, U in $\mathcal{M}_n(\mathbb{R})$ such that L is lower triangular and U is upper triangular with $l_{ii} = 1$ and $A = LU$. This means that one can do Gaussian Elimination without rotation, or pure LU decomposition.*

Remark 13.3.

(1) Why do we care? Well, if we compute $A = LU$, then we can solve $Ax = b$ by solving $Ly = b$ with $\mathrm{Nop}\, O\left(\frac{n^2}{2}\right)$ multiplication or division, then solve $Ux = y$ with the same order of complexity with backwards substitution.

(2) The normalization in L means that $\det(A) = \det(U)$.

(3) Given $A = LU$, we can compute A^{-1} by solving $Ax_i = e_i$.

(4) How to understand this diagonal submatrices condition? One thing to notice is that it is almost always OK in practice if $A \in \mathcal{M}_n(\mathbb{R})$ is symmetric positive definite.

Why is that? Either you've heard it from undergraduate class or you can just argue this way: if Δ^k is singular for some k, then we could find $(x_1, \dots, x_k)^T \in \ker(\Delta^k)$ then simply extend this vector to enough dimension, i.e. $x = (x_1, \dots, x_k, 0, \dots, 0)^T$ and we have

$$\langle x, Ax \rangle = \sum_{i=1}^{k} x_i A x_i = \sum_{i=1}^{k} x_i (\Delta^k x)_i = 0$$

which means that A is not positive definite. And by contrapositive we are done.

Proof. (of LU decomposition): To do this we perform Gaussian Elimination.

Step 1: We would like to show that all pivots are non-zero without any row swaps. This follows by an induction argument:

- $a_{11} = \Delta^1$ is non-zero since non-singular;
- Suppose the first k pivots are non-zero, then we can apply $k-1$ steps of Gaussian Elimination to write
$$\begin{pmatrix} \Delta^k & A_{12} \\ A_{21} & A_{22} \end{pmatrix} = \begin{pmatrix} L_{11}^k & 0 \\ L_{21}^k & I \end{pmatrix} = \begin{pmatrix} U_{11}^k & A_{12}^k \\ A_{21}^k & A_{22}^k \end{pmatrix}$$
now we have $L_{11}^k U_{11}^k = \Delta^k$ with U_{11}^k upper triangular and L_{11}^k lower triangular with diagonal all 1. Thus
$$\det\left(U_{11}^k\right) = \det\left(\left(L_{11}^k\right)^{-1}\Delta^k\right) \neq 0$$
which yields what we want in this step.

Step 2: Form L.

Note that $A^k = E^{k-1}\ldots E^1 A^1$ with
$$E^k = \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & \ddots & & \\ & & -l_{k+1,k} & \ddots & \\ & & \vdots & & \\ & & -l_{n,k} & & 1 \end{pmatrix} \quad \text{where} \quad l_{ik} = \frac{a_{ik}^k}{a_{kk}^k}$$
but since $U = E^n \ldots E^1 A$ we have $A = (E^1)^{-1}\ldots(E^n)^{-1}U$ and we can just compute $L = (E^1)^{-1}\ldots(E^n)^{-1}$, which is simple because the inverses of E are easy to compute: just flipping the signs on the non-diagonal non-zero terms. So we can compute:
$$E^k = \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & \ddots & & \\ & & l_{k+1,k} & \ddots & \\ & & \vdots & & \\ & & l_{n,k} & & 1 \end{pmatrix} \quad \text{and hence} \quad L = \begin{pmatrix} 1 & & & & \\ l_{21} & \ddots & & & \\ l_{31} & l_{32} & \ddots & & \\ \vdots & & \ddots & \ddots & \\ l_{n1} & & \ldots & l_{n,n-1} & 1 \end{pmatrix}$$

Step 3: Uniqueness.

Suppose $A = L_1U_1 = L_2U_2$, then $L_2^{-1}L_1 = U_2U_1^{-1} = C$. Since C is both upper triangular and lower triangular, C is diagonal. And since L_1 and L_2 both have 1s on the diagonal, $C = I$ and the decomposition is unique. □

Remark 13.4.

(1) $A = LU \Rightarrow \text{cond}(A) \leq \text{cond}(L)\,\text{cond}(U)$.

(2) One can use an "incomplete" LU factorization (to an approximation of A) to compute columns of an approximation for A^{-1}*, the preconditioner.*

(3) Once can compute efficiently column by column using

$$a_{ij} = \sum_{k=1}^{\min\{i,j\}} l_{i,k}u_{k,j}$$

To be exact, let's compute the first column and a general column:

Column 1:

- $a_{11} = l_{11}u_{11} \rightsquigarrow u_{11} = a_{11}$.
- $a_{21} = l_{21}u_{11} \rightsquigarrow l_{21} = \dfrac{a_{21}}{a_{11}}$.
- $\vdots$
- $a_{n1} = l_{n1}u_{11} \rightsquigarrow l_{n1} = \dfrac{a_{n1}}{a_{11}}$.

and for column j we have

- $a_{1j} = l_{11}u_{1j}$.
- $\vdots$
- $a_{jj} = \sum_{k=1}^{j} l_{jk}u_{kj} \rightsquigarrow u_{jj} = a_{jj} - \sum_{k=1}^{j-1} l_{ik}u_{kj}$.
- *And from then on we can compute* $l_{j+1,j}, \ldots$ *due to the same calculation.*

With this procedure (storing values as needed) the LU factorization requires $O\left(\frac{n^3}{3}\right)$ *complexity.*

13.3. **Cholesky factorization.**

For $A \in \mathcal{M}_n(\mathbb{R})$ symmetric positive definite, the idea is to write $A = BB^T$ with B-lower triangular.

Theorem 13.3. *For* $A \in \mathcal{M}_n(\mathbb{R})$ *symmetric positive definite, there exists a unique real lower triangular matrix* B *such that* $A = BB^T$ *and the diagonal entries of* B *are strictly positive.*

Proof. By our result on LU factorization, there exists unique L,U with $A = LU$. Set $D = \text{diag}(\sqrt{u_{11}}, \ldots, \sqrt{u_{nn}})$. Recall that A is positive definite really implies that we can do it.

Now, let $B = LD$ and $C = D^{-1}U$ such that A = BC. We want to show that $C = B^T$.

To do this, we note that $A = A^T \Rightarrow BC = C^TB^T$ such that

$$C^T = BC(B^T)^{-1} \Rightarrow C(B^T)^{-1} = B^{-1}C^T.$$

But note that $C(B^T)^{-1}$ is upper triangular and $BC(B^T)^{-1}$ is lower triangular, so they are both diagonal. Moreover, direct computation shows that the diagonal entries of B and C are all equal. Hence $B^{-1}C^T = 1$.

Uniqueness: If $A = B_1 B_1^T = B_2 B_2^T$, then $B_2^{-1} B_1 = B_2^T (B_1^T)^{-1}$, which means that this is diagonal, so we can let

$$D := \text{diag}(d_1, \dots, d_n) = B_2^{-1} B_1 = B_2^T (B_1^T)^{-1}.$$

So we have $B_1 = B_2 D$ and

$$A = B_1 B_1^T = B_2 D D^T B_2^T = B_2 D^2 B_2^T$$

for which we can move B_2 and B_2^T to the left side and get $I = D^2$. Thus we know $d_i = \pm 1$. But since B_1, B_2 are all positive definite, we know that $d_i > 0$, which yields that $D = I$ and $B_1 = B_2$.

□

Remark 13.5.

(1) As for the LU factorization,

$$A = BB^T \Rightarrow a_{ij} = \sum_{k=1}^{n} b_{ik} b_{jk} = \sum_{k=1}^{\min\{i,j\}} b_{ik} b_{jk}$$

and we can derive a column by column computational scheme. See AK P113-P114.

(2) $A = BB^T$ *means* $\det(A) = \det(B)^2$.

(3) $A = BB^T$ *means* $\text{cond}_2(A) = \text{cond}_2(B)^2$. *This is becuase*

$$||XX^*||_2 = ||X^*X||_2 = ||X||_2^2.$$

(4) Computing the Cholesky factorization uses $O(n^3/6)$ *multiplication/divisions.*

14. 11/16: QR FACTORIZATION; LEAST SQUARE PROBLEM

14.1. **QR factorization.**

For $A \in \mathcal{M}_n(\mathbb{R})$ non-singular, we want to write $A = QR$ where Q is orthogonal and R is upper triangular. We'll see later that this is an easy application of the Gram-Schmidt process.

Using this, we can solve $Ax = b$ by the following process:

(1) Factor $A = QR$;
(2) Compute $Q^T b = Rx$;
(3) Use back substitution to solve $Rx = Q^T b$ for x.

Theorem 14.1. *Let $A \in \mathcal{M}_n(\mathbb{R})$ be non-singular, then there exists unique pair QR such that Q is n by n orthogonal, and R is an upper triangular matrix with diagonal entries positive (we need this to say unique) with $A = QR$.*

A quick remark is that we can generalize to singular matrices and non-square matrices.

Proof. Write $A = \begin{pmatrix} | & & | \\ a_1 & \dots & a_n \\ | & & | \end{pmatrix}$ and since A is non-singular, the collection $\{a_1, \dots, a_n\}$ is a basis for $\mathbb{R}^n$. Now we can apply Gram-Schmidt to obtain orthonormal basis $q_1, \dots, q_n$ given by

$$q_i' = a_i - \sum_{k-1}^{i-1} q_k \langle q_k, a_i \rangle \quad \text{and} \quad q_i = \frac{q_i'}{||q_i'||}.$$

The justification of the above construction is in the second lecture. Hence $A = QR$.

Uniqueness: If $A = Q_1 R_1 = Q_2 R_2$, then $T = Q_2^T Q_1 = R_2 R_1^{-1}$ that is upper triangular due to the second expression. Moreover, we have

$$TT^T = (Q_2^T Q_1)(Q_1^T Q_2) = I$$

which means that $I = TT^T$, which, since T is positive definite, is a Cholesky factorization of I.

But $I = II^T$ and $T = I$ due to uniqueness of Cholesky factorization. □

Curiously, even though QR decomposition is in itself easy, the proof of uniqueness is much simpler due to our result above.

Remark 14.1.

(1) The above uniqueness result uses that the diagonal entries of R are positive, which really is essential.
(2) Using the Gram-Schmidt procedure of finding $A = QR$ uses $O(n^3)$ operations. Moreover, it's numerically unstable.

An alternative to this is the Householder algorithm.

(3) $A = QR$ *means that* $\text{cond}_2(A) = \text{cond}_2(R)$.

14.2. **Application: Least squares.**

The question we are interested in is, for $A \in \mathcal{M}_{n,p}(\mathbb{R})$, we want to find $x \in \mathbb{R}^p$ such that it minimizes $x \rightsquigarrow ||b - Ax||_2$.

Lemma 14.2. *$x \in \mathbb{R}^p$ solves the above problem iff $AA^*x = A^*b$.*

Proof. Note that $A^*A \in \mathcal{M}_n(\mathbb{R})$.

$\Rightarrow$:

Suppose that $||b - Ax||_2^2 \leq ||b - Ay||_2^2$ for all $y \in \mathbb{R}^p$. If we fix $z \in \mathbb{R}^p$ and $t \in \mathbb{R}$ we will have for $y = x + tz$

$$||b - Ax||^2 \leq ||b - A(x + tz)||^2 = ||b - Ax||^2 + 2t\langle Ax - b, Az\rangle + t^2||Az||^2$$

and by dividing both sides by $|t|$

$$0 \leq 2\frac{t}{|t|}\langle Ax - b, Az\rangle + |t|||Az||^2$$

now we let $t \to 0^+$ and $t \to 0^-$ we get

$$2\langle Ax - b, Az\rangle \geq 0 \geq 2\langle Ax - b, Az\rangle$$

which implies

$$\langle Ax - b, Az\rangle = 0 \quad \Rightarrow \quad \langle A^*Ax - A^*b, z\rangle = 0$$

and since z is arbitrary $A^*Ax = A^*b$.

$\Leftarrow$:

For $z \in \mathbb{R}^p$, expand to get

$$\begin{aligned} ||b - Az||^2 &= ||b - A(x - z + x)||^2 = ||b - Ax||^2 + 2\langle Ax - b, A(z - x)\rangle + ||A(z - x)||^2 \\ &= ||b - Ax||^2 + 2\langle A^*Ax - A^*b, z - x\rangle + ||A(z - x)||^2 \\ &= ||b - Ax||^2 + 0 + ||A(z - x)||^2 \end{aligned}$$

and since $||A(z - x)||^2 \geq 0$ we have $||b - Az||^2 \geq ||b - Ax||^2$. □

Theorem 14.3. *If $A \in \mathcal{M}_{n,p}(\mathbb{R})$, then there always exists $x \in \mathbb{R}^p$ such that $A^*Ax = A^*b$. This is sometimes called the normal equation.*

Proof. If A^*A is non-singular, we are done. Suppose A^*A is singular, we show that $\text{Im}(A^*) \subset \text{Im}(A^*A)$ and thus $A^*b \in \text{Im}(A^*A)$ for all b.

Recall that $\text{Im}(A^*) = \ker(A)^\perp$ so that

$$\mathbb{R}^p = \ker(A) \oplus \text{Im}(A^*) = \ker(A^*A) \oplus \text{Im}(A^*A).$$

So we only need to show that $\ker(A^*A) \subset \mathrm{Im}(A^*A)$. But this is because if $x \in \ker(A^*A)$, then

$$\langle A^*Ax, x\rangle = 0 \Rightarrow \langle Ax, Ax\rangle = 0 \Rightarrow x \in \ker(A).$$

□

15. 11/28: Closing up on direct methods; Iterative methods

15.1. More on direct methods.

Proposition 15.1. *$A^*Ax = A^*b$ has exactly 1 solution $\iff$ $\ker(A) = \{0\}$.*

A partial proof to this is that if x_1, x_2 are both solutions, then $x_1 - x_2 \in \ker(A^*A) \subset \ker(A)$, which is because of the last thing we shown last time.

Approach 1: Using Cholesky factorization.

The method of solving via normal equations are to solve the equation $A^*Ax = A^*b$ when $\ker(A) = \{0\}$ and use the fact that $A^*A \in \mathcal{M}_n(\mathbb{R})$ that is real symmetric positive definite. Also, we do this using Cholesky factorization.

The cost here are

- Forming $\dfrac{np(p+1)}{2}$ operations;
- The total count is
$$\frac{p^3}{6} + pn + p^2$$
where the three terms corresponding to solving via Cholesky, computing the right hand side, and using forward and backward substitution.

Remark 15.1.

(1) It is usually the case that $n \gg p$, which means that the main cose will be $\dfrac{np(p+1)}{2}$.
(2) But this method amplifies the rounding errors as $\operatorname{cond}(A^*A) > \operatorname{cond}(A)$ in most case.

Approach 2: Using QR factorization.

Given $A \in \mathcal{M}_n(\mathbb{R})$, suppose we've found the corresponding upper-trig R and orthogonal Q, then since $||Qx||_2 = ||X||_2$ (pass the adjoint in inner product), the least square problem $\min ||b - Ax||_2$ is equivalent to $\min_x ||Rx - Q^*b||_2$. And we can solve this via back substitution.

Suppose $n > p$ and $ker(A) = \{0\}$, then to apply Gram-Schmidt to columns of A is the process of QR factorization of non-square matrices, and we have

$$||b - Ax||^2 = ||Q^*b - Rx||^2 = ||(Q^*b)_p - R_{11}x||^2 + ||(Q^*b)_{n-p}||.$$

And simply back-substitution solves the whole problem. But the same issue is that Gram-Schmidt process is not stable too.

Approach 3: Householder matrix (skip).

15.2. Iterative Methods.

The strategy is that, in order to solve $Ax = b$, we generate a sequence of approximate solutions x_k such that $x_k \to x$, the true solution.

Def 15.1. *For $A \in \mathcal{M}_n(\mathbb{R})$ non-singular, A has a splitting (M, N) with M, N both $\in \mathcal{M}_n(\mathbb{R})$ if M is non-singular and $A = M - N$.*

Def 15.2. *Iterative method based on the splitting (M, N) is defined by fixing $x_0 \in \mathbb{R}^n$ and letting x_{k+1} solve the system $Mx_{k+1} = Nx_k + b$.*

The essential reason for doing this is that we can choose M as a matrix that is easier to invert than A. To justify this method, we simply has to note that if $x_k \to x$, then $Mx = Nx+b$ implies our original system $Ax = b$.

Def 15.3. *An iterative method is said to converge if $x_k \to x$ fo any choice of initial data x_0.*

Some notations that we'll use in the following pages are

$$\begin{cases} r_k : b - Ax_k & \text{the residue} \\ \varepsilon_k := x_k - x & \text{the error.} \end{cases}$$

Remark 15.2. *(1) Iterative scheme converge $\iff \varepsilon_k \to 0 \iff r_k = A\varepsilon_k = 0$.*

(2) The core step of the iterative method is $x_{k+1} = M^{-1}Nx_k + M^{-1}b$. Since the matrix $M^{-1}N$ is the matrix that we'll really use to do computations, we call it the iteration/amplification matrix.

Theorem 15.2. *The iterative method associated with the splitting (M, N) converges $\iff$ $\rho(M^{-1}N) < 1$.*

Proof. For $k \geq 1$, we have

$$\varepsilon_k = x_k - x = (M^{-1}Nx_{k-1} + M^{-1}b) - (M^{-1}Nx + M^{-1}b) = M^{-1}N(x_{k-1} - x) = M^{-1}N\varepsilon_k$$

and thus $\varepsilon_k = (M^{-1}N)\varepsilon_0$. By our result on convergence of powers of matrices (proposition 10.1), we have that $\forall x_0$

$$\varepsilon_k \to 0 \iff \rho(M^{-1}N) < 1.$$

□

15.3. **Richardson's/gradient method.**

We will explain in next class why this method is also called the gradient method. For now, we just want to show that it's an iterative method with

$$M = \alpha^{-1}I, N = \alpha^{-1}I - A, M^{-1}N = B_\alpha = I - \alpha A$$

which implies that $x_{k+1} = x_k + \alpha(b - Ax_k)$.

Let's consider this. If $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A, then the eigenvalues of B are $1 - \alpha\lambda_i$. And by our theorem above, the iterative method converges $\iff \forall i, |1 - \alpha\lambda_i| \leq 1$.

In particular, if all the eigenvalues of A are positive, then

$$0 < \alpha < \frac{2}{\rho(A)}.$$

15.4. **Jacobi's method.**

Given A, let $D = \text{diag}(a_{11}, \dots, a_{nn})$ be the diagonal terms of A, then we have Jacobi's method is besed on the splitting $(D, D - A)$, and whose iterative matrix is $J = M^{-1}N = I - D^{-1}A$

Remark 15.3.

(1) The method is well defined if $\det(D) \neq 0$.
(2) To study convergence, several results are available. For instance, the "strict diagonal dominance of A", i.e. $|a_{ii}| > \sum_{i \neq j} |a_{ij}|$ *suffices to show convergence.*

Theorem 15.3. *If A is Hermitian positive definite, and if (M, N) is a splitting of A, then $M^* + N$ is also Hermitian and that if $M^* + N$ is positive definite, then $\rho(M^{-1}N) < 1$.*

Proof. To see that $M^* + N$ is Hermitian, we can write

$$M^* + N = M^* - N^* + N^* - N =^* +(N^* - N)$$

which is a sum of 2 Hermitian matrices.

Now, suppose $M^* + N$ is positive definite and consider the vector norm on $\mathbb{R}^n$ that is defined by

$$|x|_A = \sqrt{\langle Ax, x \rangle}.$$

Now since $\rho(M^{-1}N) \leq ||M^{-1}N||_A$ (since $\rho(A) \leq$ any matrix norm), we only need to show that $||M^{-1}N||_A < 1$. For this, we may choose $v \in \mathbb{R}^n$ with $|v|_A = 1$ and

$$||M^{-1}N||_A = |M^{-1}Nv||_A.$$

Now, setting $w = M^{-1}Av$ we have

$$\begin{aligned}
|M^{-1}Nv|_A^2 &= \langle AM^{-1}Nv, M^{-1}Nv \rangle = \langle AM^{-1}(M - A)v, M^{-1}(M - A)v \rangle \\
&= \langle Av - Aw, v - w \rangle = \langle Av, v \rangle - \langle Aw, v \rangle - \langle Av, w \rangle + \langle Aw, w \rangle \\
&= 1 - \langle w, Av \rangle - \langle Av, w \rangle + \langle Aw, w \rangle \\
&= 1 - \langle w, Mw \rangle - \langle Mw, w \rangle + \langle Aw, w \rangle \\
&\text{(linearity)} = 1 - \langle (M^* + M - A)w, w \rangle = 1 - \langle (M^* + N)w, w \rangle < 1
\end{aligned}$$

since $M^* + N$ is non-singular and that M, A non-singular imples $w \neq 0$. □

Applying the above theorem to Jacobi's method we see that when A is real Hermitian, the following condition together imples that the iteration converges:

(i) A is positive definite.
(ii) $M^* + N = D^* + D - A = 2D - A$ is positive definite.

16. Gauss-Seidel Method; Improving the convergence

16.1. **Gauss-Seidel Method.**

Given $A = (a_{ij}) \in \mathcal{M}_n(\mathbb{R})$, write $A = D - E - F$ with

- D: the diagonal part of A;
- $-E$: the lower triangular part of A excluding the diagonal;
- $-F$: the upper triangular part of A excluding the diagonal.

Then, the Gauss-Seidel Method is the iterative method associated to the scheme $M = D - E$, $N = F$ with the iteratiion matrix $G = M^{-1}N = (D - E)^{-1}F$.

Remark 16.1.

(1) $D - E$ is lower triangular and the method is well defined provided that D is non-singular.

(2) $(D - E)^{-1}$ can be computed efficiently since it's triangular.

(3) If A is Hermitian postive definite, then $M^* + N = D$ is Hermitian positive definite.
Thus, the last theorem from last time implies that the Gauss-Seidel method is convergent in this case.

(4) This method can be improved: Method of successive overrelaxation (SOR): A variant of Gauss=Seidel with improved convergence rate is based on the relaxation method:

$$N = \frac{1}{w}D - E; N = \frac{1 - w}{w}D - F$$

for $w \in (0, \infty)$.

16.2. **Improving convergence.**

We do Richardson's method + steepest descend, and eventually we'll get to conjugate gradient.

How does descent enter the game here? Consider minimizing

$$f(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle$$

for $A \in \mathcal{M}_n(\mathbb{R})$ that is symmetric. We'll see later that by taking the derivative, the coefficient $\frac{1}{2}$ goes away and we're left with our wanted system.

Proposition 16.1.

$$(\nabla f)(x) = Ax - b$$

and moreover, if A is positive definite, then f admits a unique minimum x_0 that solves $Ax = b$.

Proof. We compute

$$\partial_{x_k} f = a_{kk}x_k + \frac{1}{2}\sum_{i \neq k} a_{ik}x_i + \frac{1}{2}\sum_{i \neq k} a_{ki}x_i - b_k = \sum_i a_{ik}x_i - b_k = (Ax - b)_k$$

where the middle step used the fact that A is symmetric.

Now we prove the second part. Since A is real symmetric, we can diagonalize it. Let $\lambda_1, \dots, \lambda_n$ be the eigenvalues corresponding to the eigenvectors x_i, then

$$x = \sum_{i=1}^{n} \alpha_i x_i; b = \sum_{i=1}^{n} \beta_i x_i.$$

Then

$$f(x) = \frac{1}{2} \sum_{i=1}^{n} \lambda_i \alpha_i^2 - \sum_{i=1}^{n} \beta_i \alpha_i = \frac{1}{2} \sum_{i=1}^{n} \left[\lambda_i \left(\alpha_i - \frac{\beta_i}{\lambda_i} \right) \right].$$

Note that we only have the freedom to choose α_i, which by the positive definiteness of A we only need to minimize each term, i.e. choosing $\alpha_i = \dfrac{\beta_i}{\lambda_i}$ such that

$$x = \sum_{i=1}^{n} \frac{\beta_i}{\lambda_i} x_i$$

is the unique solution of $Ax = b$. □

If you think about it, this makes very much sense since it's nothing but finding the minimum of a convex function.

Proposition 16.2. *Suppose A is real symmetric positive definite and f is as defined above. For $F \subset \mathbb{R}^n$ is a subspace of $\mathbb{R}^n$, $\exists x_0 \in F$ such that*

$$f(x_0) \leq f(x), \forall x \in F.$$

Moreover, x_0 is the unique vector in F such that $\langle Ax_0 - b, y \rangle = 0$ for $\forall y \in F$.

Proof. (Idea) Let P be the orthogonal projection from $\mathbb{R}^n \to F$, then

$$f(Py) = \frac{1}{2} \langle APy, Py \rangle + \langle b, Py \rangle = \frac{1}{2} \langle P^* APy, y \rangle + \langle P^* b, y \rangle$$

and

$$\min_y f(Py) = \min_{x \in F} f(x).$$

But the above theorem can be applied to show the result with $\tilde{A} = P^* AP, \tilde{b} = P^* b$. □

Theorem 16.3. *Suppose that A is real symmetric positive definite as above, then $x \in \mathbb{R}^n$ minimizes $f \iff (\nabla f)(x) = 0$. Moreover, if $x \in \mathbb{R}^n$ is such that $(\nabla f)(x) \neq 0$, then $\forall \alpha \in \left(0, \frac{2}{\rho(A)}\right)$*

$$f(x - \alpha \cdot \nabla f(x)) < f(x).$$

This visually means that we've found a direction that is pointing downward.

Remark 16.2. *This motivates the iterative method*

$$x_{k+1} = x_k - \alpha \nabla f(x_k) = x_k + \alpha(b - Ax)$$

which is the Richardson's method we've seen last time. Now we can see why it's called the gradient method or steepest method.

Proof. We already know that f attains minimum at the unique solution to $Ax = b$. Now, let's suppose that $x \in \mathbb{R}^n$ satisfies $(\nabla f)(x) \neq 0$. Set $\delta = -\alpha(Ax - b)$, then

$$\begin{aligned} f(x+\delta) &= \frac{1}{2}\langle A(x+\delta), x+\delta\rangle = f(x) + \frac{1}{2}\langle A\delta, \delta\rangle + \langle Ax - b, \delta\rangle \\ \text{(by Cauchy's)} \ &\leq f(x) + \frac{1}{2}||A||_2||\delta||^2 + \langle Ax - b, \delta\rangle \\ &= f(x) + \frac{1}{2}\rho(A)||\delta||^2 - \alpha\langle Ax - b, Ax - b\rangle \\ &= f(x)\left(\frac{1}{2}\rho(A)\alpha^2 - \alpha\right)||Ax - b||^2 < f(x) + 0 \end{aligned}$$

where since $\alpha \leq \dfrac{2}{\rho(A)}$ we get the last inequality. Thus, since $\nabla f(x) = Ax - b \neq 0$ we have

$$f(x+\delta) < f(x).$$

□

It's all good but it might happen that some times the steepest direction is just a local estimation and the method with result in the overstepping of some deeper points.

So a further modification is made such that we can set variable step size α_k that minimizes $\alpha \mapsto f(x_k) - \alpha \nabla f(x_k))$.

Remark 16.3. *If A is positive definite and symmetric, the expression for the above idea is*

$$\alpha_k = \frac{||Ax_k - b||^2}{\langle A(Ax_k - b), Ax_k - b\rangle}$$

which is harder to analysis. Also, even though theoretically it's better, it really does not have much better convergence.

16.3. **More involved refinements.**

Def 16.1. *For $r \in \mathbb{R}^n, k \geq 0$, the Krylov space associated to r and A is the space*

$$K_k = \text{span}\{r, Ar, \ldots, A^k r\}.$$

Remark 16.4.

(1) We write the order k Krylov space as K_k, and we have $K_k \subset K_{k+1}$, for $\forall k \geq 0$.
(2) For all $r_o \neq 0$ in $\mathbb{R}^n$, there exists k_0, called the "Krylov critical dimension" such that $\dim K_k = k_+ 1$ *for* $0 \leq k \leq k_0$ *and* $\dim K_k = K_0 + 1$ *for* $k \geq k_0$.

(3) The gradient iteration with constant step size has its residue $r_k = b - Ax_k$ satisfies

(i) $r_k \in K_k(r_0, A)$ with reason:

$$x_{k+1} = x_k + \alpha_k r_k \Rightarrow b - Ax_{k-1} = b - A(x_k - \alpha A r_k)$$

which means

$$r_{k+1} = r_k - \alpha_k A r_k$$

that means each time the iteration goes into K_{k+1} via an induction argument.

(ii) $x_{k+1} \in [x_0 + K_k] := \{x : x - x_0 \in K_k\}$.

So one way to improve on the gradient method is to think about abandoning the explicit iteration, but keep only (ii) above. We call this "conjugate gradient method." How do we specify this?

There are 2 equivalent ideas:

(a) $x_{k+1} \in [x_0 + K_k]$ chosen such that $r_{k+1} \perp K_k$.
(b) $x_{k+1} \in [x_0 + K_k]$ chosen such that it minimizes $f(x)$ in $[x_0 + K_k]$.

Proposition 16.4. *If A is symmetric positive definite and x_k is a sequence obtained by conjugate gradient. Let $r_k = b - Ax_k, d_k = x_{k+1} - x_k$, then we have the below results:*

(i)

$$K_k(r_0, A) = \text{span}\{r_0, \dots, r_k\} = \text{span}\{d_0, \dots, d_k\}.$$

(ii) $\{r_k : 0 \le k \le n-1\}$ *is pairwise orthogonal.*

(iii) $\{d_k\}$ *is conjugate with respect to A, i.e.* $\forall 0 \le l < k \le n-1$

$$\langle Ad_k, d_l \rangle = 0.$$

Now we just list the algorithm of the conjugate gradient method and leave it there.

- $p_0 = r_0 = b - Ax_0$;
- $x_{k+1} = x_k + \alpha_k p_k$, $\alpha_k = \dfrac{||r_k||^2}{\langle Ap_k, p_k \rangle}$;
- $r_{k+1} = r_k - \alpha_k A p_k$;
- $p_{k+1} = r_{k+1} + \beta_k p_k$, $\beta_k = \dfrac{||r_{k+1}||^2}{||k||^2}$.

which converges in n steps.

17. 12/5: Exercise questions

Exercise 17.1. *Let $A \in \mathcal{M}_n(\mathbb{R})$ be given and let σ_i denote its singular values in decreasing order, show that*

$$\sigma_1 = \sup_{\substack{x,y\in\mathbb{R}^n \\ x\neq 0, y\neq 0}} \frac{\langle Ax, y\rangle}{||x|| \cdot ||y||}.$$

Proof. Recall that $\sigma_1 = ||A||_2$.

($\leq$): we have

$$\sigma_1 = ||A||_2 = \sup_{x\neq 0} \frac{||Ax||}{||x||} \leq \sup_{\substack{x,y\in\mathbb{R}^n \\ x\neq 0, y\neq 0}} \frac{\langle Ax, y\rangle}{||x|| \cdot ||y||}.$$

($\geq$): now we try to bound an inner product from above, and the trick is to use Cauchy-Schwartz inequality.

For $\forall x, y \neq 0$, we have

$$\langle Ax, y\rangle \leq |\langle Ax, y\rangle| \leq ||Ax|| \cdot ||y||$$

and thus

$$\frac{\langle Ax, y\rangle}{||x|| \cdot ||y||} \leq \frac{||Ax||}{||x||} \leq \sigma_1$$

where by taking sup to both sides we get the result. □

Exercise 17.2. *Let $||A||_{\mathcal{F}}$ be the Frobenius norm of A, show that*

$$||A||_2 \leq ||A||_{\mathcal{F}}.$$

Proof. We can write any $x \in \mathbb{R}^n$ as $x = \sum_{i=1}^{n} x_i e_i$.

Now, note that

$$||Ax|| = ||A \sum_{i=1}^{n} x_i e_i|| = || \sum_{i=1}^{n} x_i Ae_i|| \leq \sum_{i=1}^{n} |x_i| \cdot ||Ae_i||.$$

Now we note that this is nothing but an inner product for which we use Cauchy-Schwartz:

$$\begin{pmatrix} |x_1| \\ \vdots \\ |x_n| \end{pmatrix} \cdot \begin{pmatrix} ||Ae_1|| \\ \vdots \\ ||Ae_n|| \end{pmatrix} \leq \left(\sum_{i=1}^{n} |x_i|^2 \right)^{\frac{1}{2}} \left(\sum_{i=1}^{n} ||Ae_i||^2 \right)^{\frac{1}{2}} = ||x|| \cdot ||A||_{\mathcal{F}}.$$

□