I - Linear Algebra Basics and Functional Analysis Snippets

Math 728 D - Machine Learning & Data Science - Spring 2019

Contents

Examples

$\mathbb{R}, \mathbb{C}$ fields of real resp. complex numbers, $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$

$\mathbb{K}^n := \{\mathbf{x} = (x_1, \ldots . x_n)^\top : x_i \in \mathbb{K},\ i = 1, \ldots, n\}$

Example 1

Find the intersection of two lines in the plane

$$L_1 = \{\mathbf{x} \in \mathbb{R}^2 : a_{1,1}x_1^* + a_{1,2}x_2^* = b_1\}, \quad L_2 = \{\mathbf{x} \in \mathbb{R}^2 : a_{2,1}x_1^* + a_{2,2}x_2^* = b_1\}$$

$$\mathbf{x}^* \in L_1 \cap L_2 \quad \Leftrightarrow \quad \begin{aligned} a_{1,1}x_1^* + a_{1,2}x_2^* &= b_1 \\ a_{2,1}x_1^* + a_{2,2}x_2^* &= b_1 \end{aligned} \quad \Leftrightarrow \quad \begin{pmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{pmatrix} \begin{pmatrix} x_1^* \\ x_2^* \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \quad \Leftrightarrow \quad \boldsymbol{A}\mathbf{x}^* = \mathbf{b}$$

Matrix-vector multiplication:

$$\boldsymbol{A} = \begin{pmatrix} a_{1,1}, & \ldots, & a_{1,n} \\ \vdots & & \vdots \\ a_{m,1}, & \ldots, & a_{m,n} \end{pmatrix} = \Big(\underbrace{\mathbf{a}^1, \ldots, \mathbf{a}^n}_{\text{columns of } \boldsymbol{A}}\Big) \in \mathbb{K}^{m \times n} \rightsquigarrow \boldsymbol{A}\mathbf{x} := x_1\mathbf{a}^1 + \cdots + x_n\mathbf{a}^n \in \mathbb{K}^m$$

The Core Questions

Given $\boldsymbol{A} = (a_{i,j})_{i,j=1}^{m,n} \in \mathbb{K}^{m\times n}$, $\mathbf{b} \in \mathbb{K}^m$, find $\mathbf{x} \in \mathbb{K}^n$ s.t.

$$\begin{array}{lcccl} x_1 a_{1,1} & + \cdots + & x_n a_{1,n} & = & b_1 \\ \vdots & & \vdots & \vdots & \vdots \\ x_1 a_{m,1} & + \cdots + & x_n a_{m,n} & = & b_m \end{array} \quad \Leftrightarrow \quad \boldsymbol{A}\mathbf{x} = \mathbf{b} \quad \Leftrightarrow \quad \sum_{j=1}^{n} x_j \mathbf{a}^j = \mathbf{b}$$

The driving issue in Linear Algebra is the "solvability" of such systems, more precisely:

under which conditions on $\boldsymbol{A}, \mathbf{b}$ does there

- exist exactly one solution $\mathbf{x} \in \mathbb{K}^n$
- exist at least one solution $\mathbf{x} \in \mathbb{K}^n$
- no solution?
- what is the structure of the set of solutions?

These questions can be best answered by viewing the matrix $\boldsymbol{A}$ as a *mapping*

$$\boldsymbol{A} : \mathbb{K}^n \to \mathbb{K}^m, \quad \boldsymbol{A} : \mathbf{x} \mapsto \boldsymbol{A}\mathbf{x}$$

What has been used?

$$x \in \mathbb{K},\ \mathbf{a} \in \mathbb{K}^m \ \rightsquigarrow\ x\mathbf{a} := \begin{pmatrix} xa_1 \\ \vdots \\ x\mathbf{a}_m \end{pmatrix} \in \mathbb{K}^m, \qquad \mathbf{a},\ \mathbf{b} \in \mathbb{K}^m,\ \rightsquigarrow\ \mathbf{a}+\mathbf{b} := \begin{pmatrix} a_1 + b_1 \\ \vdots \\ a_m + b_m \end{pmatrix} \in \mathbb{K}^m$$

Definition 2

A set $\mathbb{V}$ is a *Vector Space* (linear space) over the field $\mathbb{K}$ ($\mathbb{K}$-vector space) if the following holds:

- $\exists$ an "addition" $+ : \mathbb{V} \times \mathbb{V} \to \mathbb{V}$ which is commutative and associative
- $\exists$ a "multiplication" $x \in \mathbb{K}, \mathbf{a} \in \mathbb{V} \mapsto x\mathbf{a} \in \mathbb{V}$ which is distributive

Example 3

1. $\mathbb{V} = \mathbb{K}^m$ *is a vector space over* $\mathbb{K}$
2. $C^k([a,b]) := \{f : [a,b] \to \mathbb{R} : f^{(k)}$ *is continuous in every* $x \in [a,b]\}$ *is a vector space over* $\mathbb{R}$ *where*
$$(af)(x) := af(x), \quad (f+g)(x) := f(x) + g(x), \quad x \in [a,b]$$
3. $$\mathbb{P}_n := \left\{ p(x) : p(x) = \sum_{j=0}^{n} a_j x^j,\ a_j \in \mathbb{R},\ j = 0, \ldots, n \right\}$$

is a vector space over $\mathbb{R}$ *with the same operations as in (2)*

Forms, Scalar (Inner) Products

Let $\mathbb{X}, \mathbb{Y}$ be an $\mathbb{K}$-vector spaces

$\underline{\mathbb{K} = \mathbb{R}:}$ $b(\cdot,\cdot) : \mathbb{X} \times \mathbb{Y} \to \mathbb{R}$ is called bilinear if $\forall\, x, z \in \mathbb{X},\ y, v \in \mathbb{Y},\ \alpha, \beta \in \mathbb{R}$

$$b(\alpha x + \beta z, y) = \alpha b(x,y) + \beta b(z,y), \quad b(x, \alpha y + \beta v) = \alpha b(x,y) + \beta b(x,v)$$

In other words $b(\cdot, y) : \mathbb{X} \to \mathbb{R}$, $b(x,\cdot) : \mathbb{Y} \to \mathbb{R}$ are for fixed $x \in \mathbb{X}, y \in \mathbb{Y}$ linear functionals

$b(\cdot,\cdot) : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ is called symmetric if $b(x,y) = b(y,x)$, $x, y \in \mathbb{X}$

$\underline{\mathbb{K} = \mathbb{C}:}$ $b(\cdot,\cdot) : \mathbb{X} \times \mathbb{Y} \to \mathbb{C}$ is called sesquilinear if for fixed $y \in \mathbb{Y}$, $x \in \mathbb{X}$, $b(\cdot, y) : \mathbb{X} \to \mathbb{K}$ is linear and $b(x,\cdot) : \mathbb{Y} \to \mathbb{K}$ is semi-linear, i.e., $b(x, \alpha y) = \overline{\alpha} b(x,y)$. Hence
$\forall\, x, z \in \mathbb{X},\ y, v \in \mathbb{Y},\ \alpha, \beta \in \mathbb{R}$

$$b(\alpha x + \beta z, y) = \alpha b(x,y) + \beta b(z,y), \quad b(x, \alpha y + \beta v) = \overline{\alpha} b(x,y) + \overline{\beta} b(x,v)$$

A sesquilinear form $b(\cdot,\cdot) : \mathbb{X} \times \mathbb{X} \to \mathbb{C}$ is called hermitian if $b(x,y) = \overline{b(y,x)}$

Definition 4

Let $b(\cdot,\cdot) : \mathbb{X} \times \mathbb{X} \to \mathbb{K}$ be a symmetric, resp. hermitian form when $\mathbb{K} = \mathbb{R}$ resp. $\mathbb{K} = \mathbb{C}$. Then $b(\cdot,\cdot) : \mathbb{X} \times \mathbb{X} \to \mathbb{K}$ is called a scalar product (inner product, dot product) if it is positive definite, i.e.,

$$b(x,x) > 0 \quad \forall\, x \in \mathbb{X} \setminus \{0\}$$

Examples

- $\mathbb{K} = \mathbb{R}, \mathbb{X} = \mathbb{R}^n, \mathbb{Y} = \mathbb{R}^m, \boldsymbol{A} \in \mathbb{R}^{n\times m}, \quad b(\mathbf{x}, \mathbf{y}) := \mathbf{y}^\top \boldsymbol{A}\mathbf{x}$
- $\mathbb{X} = \mathbb{Y} = \mathbb{R}^n$

$$b(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle_n := \mathbf{y}^\top \mathbf{x} = \sum_{k=1}^{n} x_k y_k =: \mathbf{x} \cdot \mathbf{y}$$

  show that this is a scalar product on $\mathbb{R}^n$
- $\mathbb{K} = \mathbb{C}, \mathbb{X} = \mathbb{C}^n$

$$b(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle_n := \mathbf{y}^* \mathbf{x} := \sum_{k=1}^{n} x_k \overline{y}_k, \qquad \mathbf{y}^* := \overline{\mathbf{y}}^\top$$

  show that this is a scalar product on $\mathbb{C}^n$
- $\mathbb{K} = \mathbb{R}, \mathbb{X} = C([a, b])$

$$\langle f, g \rangle_{[a,b]} := \int_a^b f(x) g(x) dx$$

  is a scalar product on $C([a, b])$

Dimension

Recall:

- a vector space $\mathbb{V}$ is generated by a set $\{v_1 \ldots, v_k\} \subset \mathbb{V}$ ($\mathbb{V} = \text{span}\{v_1, \ldots, v_k\}$) if every $v \in \mathbb{V}$ can be written as linear combination of the v_j, $j = 1, \ldots, k$, i.e., for some $c_1, \ldots, c_k \in \mathbb{K}$

$$v = \sum_{j=1}^{k} c_j v_j$$

- there exists a minimal $n \in \mathbb{N}$ such that $\mathbb{V}$ can be generated by n elements of $\mathbb{V}$. n is called the *dimension* of $\mathbb{V}$, $n = \dim \mathbb{V}$
- a minimal generating set is called a *basis* of $\mathbb{V}$.

Exercise 5

- *Determine* $\dim \mathbb{K}^n, \dim \mathbb{P}_n$*, exhibit bases*
- *What is the dimension of* $C^k([a, b])$*?*
- *The elements of a basis* $\{v_1, \ldots, v_n\}$ *are* linearly independent*, i.e.,*

$$\sum_{j=1}^{n} c_j v_j = 0 \Rightarrow c_j = 0, \quad j = 1, \ldots, n$$

Linear Mappings

Definition 6

$\mathbb{X}, \mathbb{Y}$ vector spaces (over $\mathbb{K}$), $L : \mathbb{X} \to \mathbb{Y}$ is called a *linear* mapping (operator) if for any $a, b \in \mathbb{K}$, $v, w \in \mathbb{X}$

$$L(av + bw) = aLv + bLw \in \mathbb{Y}$$

- L is injective (one-to-one) if $L(v) = 0 \Rightarrow v = 0$
- L is surjective (onto) if for $y \in \mathbb{Y}$ there exists $x \in \mathbb{X}$ such that $L(x) = y$
- L is bijective (one-to-one and onto, invertible) if L is injective and surjective

The set of all linear mappings from $\mathbb{X}$ to $\mathbb{Y}$ is denoted by $\mathcal{L}(\mathbb{X}, \mathbb{Y})$

- Any matrix $\boldsymbol{A} \in \mathbb{K}^{m\times n}$ induces a *linear* mapping (operator) from $\mathbb{K}^n$ to $\mathbb{K}^m$ by

$$\mathbf{x} \in \mathbb{K}^n \mapsto \boldsymbol{A}\mathbf{x} \in \mathbb{K}^m \quad (\text{recall } \boldsymbol{A}\mathbf{x} = x_1\mathbf{a}^1 + \cdots + x_n\mathbf{a}^n) \rightsquigarrow \boldsymbol{A}(a\mathbf{x}^1 + b\mathbf{x}^2) = a\boldsymbol{A}\mathbf{x}^1 + b\boldsymbol{A}\mathbf{x}^2$$

- $\mathbb{V}$ a vector space over $\mathbb{K}$, a (linear) mapping $\ell : \mathbb{V} \to \mathbb{K}$ is called a (linear) *functional*

Examples:

- $\mathbb{V} = C^0([a, b])$, $\ell(f) := \int_a^b f(x)\omega(x)dx$
- $\mathbb{V}$ as before, $\delta_{x_0} f := f(x_0)$
- $\mathbb{V} = \mathbb{R}^n$, $\mathbf{a} \in \mathbb{R}^n$ fixed, $\ell(\mathbf{x}) := \sum_{j=1}^n a_j x_j =: \mathbf{a} \cdot \mathbf{x} = \mathbf{a}^\top \mathbf{x}$

Linear Mappings some further notions and facts

1. If $\mathbb{X}, \mathbb{Y}$ are $\mathbb{K}$-vector spaces then $\mathcal{L}(\mathbb{X}, \mathbb{Y})$ is a $\mathbb{K}$-vector space; identify the "+" in $\mathcal{L}(\mathbb{X}, \mathbb{Y})$
2. $\ker(L) := \{x \in \mathbb{X} : L(x) = 0\}$ is a linear subspace of $\mathbb{X}$
3. $\mathrm{ran}(L) := \{y \in \mathbb{Y} : \exists\, x \in \mathbb{X}, \text{ s.t. } y = L(x)\}$ is a linear subspace of $\mathbb{Y}$
4. let $\dim \mathbb{X} = n < \infty$, $L \in \mathcal{L}(\mathbb{X}, \mathbb{Y})$ bijective implies $\dim \mathbb{Y} = n$
5. compositions: $L \in \mathcal{L}(\mathbb{X}, \mathbb{Y})$, $B \in \mathcal{L}(\mathbb{V}, \mathbb{X})$, then LB, defined by $LB(v) := L(B(v))$, $v \in \mathbb{V}$, belongs to $L(\mathbb{V}, \mathbb{Y})$
6. let $L \in \mathcal{L}(\mathbb{X}, \mathbb{Y})$ be bijective, then $L^{-1} : \mathbb{Y} \to \mathbb{X}$, defined by $L(x) = y \Leftrightarrow L^{-1}(y) = x$, belongs to $\mathcal{L}(\mathbb{Y}, \mathbb{X})$; L^{-1} is called *inverse* of L and

$$LL^{-1} = I_{\mathbb{Y}}, \quad L^{-1}L = I_{\mathbb{X}}, \quad (I_{\mathbb{V}} v = v,\ v \in \mathbb{V},\ \text{identity on } \mathbb{V})$$

Thus, the set of bijective mappings in $\mathcal{L}(\mathbb{X}, \mathbb{X})$ form a *group* with respect to composition as multiplication.

Proposition 7

- *every $\mathbb{K}$-vector space $\mathbb{X}$ of dimension $\dim \mathbb{X} = n$ is* isomorphic *to $\mathbb{K}^n$, in the sense that there exists a bijection $B \in \mathcal{L}(\mathbb{X}, \mathbb{K}^n)$ such that $\mathbb{K}^n = \{B(x) : x \in \mathbb{X}\}$. Hence any two n-dimensional $\mathbb{K}$-vector spaces are isomorphic*
- *If $\dim(\mathbb{X}) = n, \dim(\mathbb{Y}) = m$, then $\mathcal{L}(\mathbb{X}, \mathbb{Y})$, $\mathcal{L}(\mathbb{K}^n, \mathbb{K}^m)$, $\mathbb{K}^{mn}$, $\mathbb{K}^{m \times n}$ are all isomorphic to each other*

Proof of Proposition 7

Let $\mathbb{X}$ be an n-dimensional $\mathbb{K}$-vector space and $L \in \mathcal{L}(\mathbb{X}, \mathbb{X})$. Then there exists a basis $\Phi = \{\phi_1, \ldots, \phi_n\} \subset \mathbb{X}$ of $\mathbb{X}$. Hence, every $x \in \mathbb{X}$ has a unique representation

$$x = \sum_{k=1}^{n} x_k \phi_k \in \mathbb{X}. \tag{3.1}$$

Define $B_\Phi : \mathbb{X} \to \mathbb{K}^n$ by $B_\Phi(x) := \mathbf{x} = (x_1, \ldots, x_n)^\top$ (B_Φ depends on the choice of basis). Clearly B_Φ is linear and injective (by definition of a basis), i.e., $B_\Phi \in \mathcal{L}(\mathbb{X}, \mathbb{K}^n)$. Also $M : \mathbb{K}^n \to \mathbb{X}$, defined by $M(\mathbf{x}) := \sum_{k=1}^{n} x_k \phi_k$ is linear, injective and obviously $M(B_\Phi(x)) = x$. Thus $M = B_\Phi^{-1}$.

Now suppose $\mathbb{X}, \mathbb{Y}$ are K-vector spaces with bases $\Phi = \{\phi_1, \ldots, \phi_n\} \subset \mathbb{X}$, $\Psi = \{\psi_1, \ldots, \psi_m\} \subset \mathbb{Y}$, respectively. In particular, for each $k \in \{1, \ldots, n\}$ one has $B_\Psi(L(\phi_k)) = (c_{1,k} \ldots, c_{m,k})^\top =: \mathbf{c}^k \in \mathbb{K}^m$. Thus, for $\mathbf{C} = (\mathbf{c}^1, \ldots, \mathbf{c}^n) \in \mathbb{K}^{m \times n}$

$$y := L(x) = L\Big(\sum_{k=1}^{n} x_k \phi_k\Big) = \sum_{k=1}^{n} x_k L(\phi_k) = \sum_{r=1}^{m} \Big(\sum_{k=1}^{n} c_{r,k} x_k\Big) \psi_r = B_\Psi^{-1}(\mathbf{Cx}) \in \mathbb{Y}.$$

In other words

$$L(x) = y \quad \Leftrightarrow \quad \mathbf{Cx} = \mathbf{y} \quad \text{for} \quad \mathbf{y} = B_\Psi(y),\ \mathbf{x} = B_\Phi(x) \tag{3.2}$$

In that sense $L \in \mathcal{L}(\mathbb{X}, \mathbb{Y})$ is represented by $\mathbf{C} \in \mathcal{L}(\mathbb{K}^n, \mathbb{K}^m) \equiv \mathbb{K}^{m \times n}$

Linear Mappings, Matrices — some further notions and facts

- by (5) above $\boldsymbol{A} \in \mathbb{K}^{m\times n}$, $\mathbf{B} \in \mathbb{K}^{n\times r} \Rightarrow \boldsymbol{AB} = (\boldsymbol{A}\mathbf{b}^1, \ldots, \boldsymbol{A}\mathbf{b}^r) \in \mathbb{K}^{r\times m}$
- change of bases: suppose $\mathbf{a}^1, \ldots, \mathbf{a}^n$, and $\mathbf{b}^1, \ldots, \mathbf{b}^n$ are both bases for $\mathbb{K}^n$, then there exists a unique $\mathbf{M} \in \mathbb{K}^{n\times n}$ such that $\mathbf{b}^j = \mathbf{M}\mathbf{a}^j$, $j = 1, \ldots, n$

  In fact: $\boldsymbol{A} := (\mathbf{a}^1, \ldots, \mathbf{a}^n), \mathbf{B} := (\mathbf{b}^1, \ldots, \mathbf{b}^n) \in \mathbb{K}^{n\times n}$ then $\mathbf{B} = (\mathbf{B}\boldsymbol{A}^{-1})\boldsymbol{A}$, i.e., $\mathbf{M} = \mathbf{B}\boldsymbol{A}^{-1}$
- $\boldsymbol{A} = (\mathbf{a}^1, \ldots, \mathbf{a}^n) \rightsquigarrow \mathrm{ran}(\boldsymbol{A}) = \mathrm{span}\,\{\mathbf{a}^1, \ldots, \mathbf{a}^n\}$
- $\mathrm{rank}(\boldsymbol{A}) := \dim(\mathrm{ran}(\boldsymbol{A})) = \#(\text{linearly independent columns of } \boldsymbol{A})$
- $\boldsymbol{A} \in \mathbb{K}^{m\times n}$, $\ker(\boldsymbol{A}) = \{\mathbf{x} \in \mathbb{K}^n : \boldsymbol{A}\mathbf{x} = 0\} = \mathrm{ran}(\boldsymbol{A}^\top)^\perp$, where

$$\boldsymbol{A} = (a_{i,j})_{i,j=1}^{m,n} \in \mathbb{K}^{m\times n} \quad \rightsquigarrow \quad \boldsymbol{A}^\top := (a_{j,i})_{j,i=1}^{n,m} \in \mathbb{K}^{n\times m} \text{ (transpose of } \boldsymbol{A})$$

and

$$\mathbb{V} \subset \mathbb{K}^n \;\rightsquigarrow\; \mathbb{V}^\perp := \{\mathbf{y} \in \mathbb{K}^n : \mathbf{y}\cdot\mathbf{v} = 0,\ \mathbf{v} \in \mathbb{V}\}$$

Exercise 8

- *show that* $\#(\textit{linearly independent columns of } \boldsymbol{A}) = \#(\textit{linearly independent rows of } \boldsymbol{A})$
- *show that for* $\boldsymbol{A} \in \mathbb{K}^{m\times n}$ *one has* $m = \dim(\mathrm{ran}(\boldsymbol{A})) + \dim(\ker(\boldsymbol{A}^\top))$

Solvability of Linear Systems

Given $\boldsymbol{A} \in \mathbb{K}^{m\times n}$, $\mathbf{b} \in \mathbb{K}^m$, solve $\boldsymbol{A}\mathbf{x} = \mathbf{b}$:

Exercise 9

- *There is a unique solution if and only if* $\mathbf{b} \in \operatorname{ran}(\boldsymbol{A})$, $m \geq n$, $\operatorname{rank}(\boldsymbol{A}) = n$.
- *There exists a unique solution for* **every** $\mathbf{b} \in \mathbb{R}^m$ *if and only if* $m = n$ *and* $\operatorname{rank}(\boldsymbol{A}) = n$. *In this case* $\mathbf{x} = \boldsymbol{A}^{-1}\mathbf{b}$.
- *For* $n = m$ *there exists a unique solution for* **every** $\mathbf{b} \in \mathbb{R}^m$ *if and only if* $\ker(\boldsymbol{A}) = \{0\}$.
- *Recall:* $\det : \mathbb{K}^{n\times n} \to \mathbb{K}$ *is a multilinear anti-symmetric mapping and* $\boldsymbol{A}^{-1}$ *exists if and only if* $\det(\boldsymbol{A}) \neq 0$ *in which case we call* $\boldsymbol{A}$ non-singular.
- *More than one solution exists if and only if infinitely many solutions exist. The solution set is the* affine space

$$S(\boldsymbol{A}) = \mathbf{x}^0 + \ker(\boldsymbol{A}) = \{\mathbf{x} = \mathbf{x}^0 + \mathbf{y} : \mathbf{y} \in \ker(\boldsymbol{A})\}$$

Notice: searching for a "solution" of $\boldsymbol{A}\mathbf{x} = \mathbf{b}$ makes sense only if $\boldsymbol{A} \in \mathbb{K}^{n\times n}$ is non-singular, that is solvability depends **only** on $\boldsymbol{A}$, why? In all other cases the notion of "solution" has to be generalized.

Hermitian and Positive Definite Matrices

For $\boldsymbol{A} = (a_{k,j})_{k,j=1}^{m,n} \in \mathbb{K}^{m\times n}$ the transpose $\boldsymbol{A}^\top$ is defined by $\boldsymbol{A}^\top = (a_{j,k})_{k,j=1}^{n,m} \in \mathbb{K}^{n\times m}$ (reflect across the diagonal)

Recall: i imaginary unit, $i^2 = -1$, $z = x + iy \in \mathbb{C}$, $\overline{z} := x - iy \rightsquigarrow z\overline{z} = |z|^2 = x^2 + y^2$

$\mathbb{K} = \mathbb{C}$: hermitian conjugate:

$$\boldsymbol{A}^* := \overline{\boldsymbol{A}}^\top = (\overline{a_{j,k}})_{j=1,k}^{n,m} \in \mathbb{C}^{n\times m} \quad (\boldsymbol{A}^* = \boldsymbol{A}^\top \text{ when } \mathbb{K} = \mathbb{R})$$

Definition 10

$\boldsymbol{A} \in \mathbb{K}^{n\times n}$ is called symmetric ($\mathbb{K} = \mathbb{R}$), resp. hermitian ($\mathbb{K} = \mathbb{C}$) if $\boldsymbol{A} = \boldsymbol{A}^*$

A hermitian matrix $\boldsymbol{A} \in \mathbb{K}^{n\times n}$ is called positive (semi-)definite if

$$\mathbf{x}^* \boldsymbol{A}\mathbf{x} > (\geq) 0 \quad \forall\, \mathbf{x} \in \mathbb{K}^n \setminus \{0\}$$

Remark 11

For $\boldsymbol{A} \in \mathbb{K}^{n\times n}$, one has $\langle \boldsymbol{A}\mathbf{x}, \mathbf{y}\rangle_n = \langle \mathbf{x}, \boldsymbol{A}^*\mathbf{y}\rangle_n$, i.e., $\boldsymbol{A}^*$ is called the adjoint of $\boldsymbol{A}$ for reasons to be explained later.

Examples of Symmetric/Hermitian Positive Definite Matrices

- **I** is in each "club"
- For any $\mathbf{B} \in \mathbb{K}^{m\times n}$ the matrices $\boldsymbol{A} := \mathbf{B}^*\mathbf{B} \in \mathbb{K}^{n\times n}$, $\boldsymbol{A} := \mathbf{B}\mathbf{B}^* \in \mathbb{K}^{m\times m}$ are hermitian positive semi-definite. Under which conditions on **B** is which version positive definite?
- Suppose that $\{\phi_1, \ldots, \phi_n\} \subset \mathbb{X}$ where $\mathbb{X}$ is a $\mathbb{K}$-vector space and $\langle\cdot,\cdot\rangle_{\mathbb{X}}$ is a scalar product on $\mathbb{X}$. Then the Gramian matrix

$$\mathbf{G} := \left(\langle\phi_k, \phi_j\rangle_{\mathbb{X}}\right)_{k,j=1}^{n,n} \in \mathbb{K}^{n\times n} \tag{4.1}$$

  is hermitian positive semi-definite. Under which condition on the set $\{\phi_1, \ldots, \phi_n\} \subset \mathbb{X}$ is **G** positive definite?

Remark 12

Any hermitian positive definite matrix is nonsingular and its inverse is also hermitian positive definite.

Proof: Suppose $\det(\boldsymbol{A}) = 0, \Rightarrow \exists\, \mathbf{x} \in \mathbb{K}^n \setminus \{0\}$ s.t. $\boldsymbol{A}\mathbf{x} = 0, \Rightarrow \mathbf{x}^*\boldsymbol{A}\mathbf{x} = 0$ which is a contradiction for $\mathbf{x} \neq 0$. Now show that $\boldsymbol{A}^{-1} = (A^*)^{-1} \overset{!}{=} (\boldsymbol{A}^{-1})^* \Leftrightarrow \mathbf{I} = \boldsymbol{A}^*(\boldsymbol{A}^{-1})^*$. Indeed, (since $\mathbf{B}^*\mathbf{C}^* = (\mathbf{C}\mathbf{B})^*$) one has $\boldsymbol{A}^*(\boldsymbol{A}^{-1})^* = (\boldsymbol{A}^{-1}\boldsymbol{A})^* = \mathbf{I}^* = \mathbf{I}$. □

Further Properties of Hermitian Positive Definite Matrices

Proposition 13

Let $\boldsymbol{A} \in \mathbb{K}^{n\times n}$ be hermitian positive semi-definite, then:

1. *All principal submatrices of $\boldsymbol{A}$ are hermitian positive semi-definite. In particular, all diagonal entries are real and non-negative $a_{k,k} \geq 0$, $k = 1, \ldots, n$.*
2. *All eigenvalues of $\boldsymbol{A}$ are real and nonnegative.*
3. *The maximum modulus entry of $\boldsymbol{A}$ occurs on the diagonal of $\boldsymbol{A}$.*

Proof: For any $J \subset \{1, \ldots, n\}$ let $\boldsymbol{A}_J := \left(a_{j,k}\right)_{(j,k)\in J\times J} \in \mathbb{K}^{\#(J)\times\#(J)}$. For $J \neq \emptyset$ and any $\mathbf{x} \in \mathbb{K}^n \setminus \{0\}$ supported in J let $\tilde{\mathbf{x}} := \mathbf{x}|_J \in \mathbb{K}^{\#(J)}$. Since $\tilde{\mathbf{x}} \neq \mathbf{0}$ we have $0 < \mathbf{x}^*\boldsymbol{A}\mathbf{x} = \tilde{\mathbf{x}}^*\boldsymbol{A}_J\tilde{\mathbf{x}}$. Since $\tilde{\mathbf{x}} \neq \mathbf{0}$ is arbitrary (1) follows. As for (2), suppose $\boldsymbol{A}\mathbf{x} = \lambda\mathbf{x}$, $\lambda \neq 0$. Then $0 < \mathbf{x}^*\boldsymbol{A}\mathbf{x} = \lambda\mathbf{x}^*\mathbf{x} = \lambda\sum_{k=1}^n |x_k|^2$ which implies $\lambda > 0$. Concerning (3), we know from Theorem 30 (later below) that then all principal minors (determinants of principal submatrices) are nonnegative. Now suppose, that $|a_{j,k}| = \max_{1\leq r,s\leq n} |a_{r,s}|$ and $j \neq k$. Then (w.l.o.g. $j < k$)

$$0 \leq \det\begin{pmatrix} a_{j,j} & a_{k,j} \\ a_{j,k} & a_{k,k} \end{pmatrix} = a_{j,j}a_{k,k} - a_{k,j}a_{j,k} = a_{j,j}a_{k,k} - |a_{j,k}|^2$$

which means that $|a_{j,k}|^2 \leq a_{j,j}a_{k,k}$ and hence $\max\{a_{j,j}, a_{k,k}\} \geq |a_{j,k}|$. □

Unitary Matrices

$\boldsymbol{A} \in \mathbb{K}^{n\times n}$ is called unitary (orthogonal if $\mathbb{K} = \mathbb{R}$): $\boldsymbol{Q} \in \mathcal{O}_n = \mathcal{O}_n(\mathbb{K})$ if $\boldsymbol{Q}^*\boldsymbol{Q} = \mathbf{I}$

Properties:

- $\boldsymbol{Q}^*\boldsymbol{Q} = \mathbf{I}$ means that the columns $\mathbf{q}^j$, $j = 1, \ldots, n$, are an orthonormal basis of $\mathbb{K}^n$, i.e.,

$$(\mathbf{q}^j)^*\mathbf{q}^k = \sum_{r=1}^{n} \overline{q_{r,j}} q_{r,k} = \delta_{j,k}, \quad j,k = 1,\ldots,n$$

- $\boldsymbol{Q} \in \mathcal{O}_n \Leftrightarrow \boldsymbol{Q}^* \in \mathcal{O}_n$, i.e., the rows of $\boldsymbol{Q}$ also form an orthonormal basis
- $\boldsymbol{Q}, \tilde{\boldsymbol{Q}} \in \mathcal{O}_n \Rightarrow \boldsymbol{Q}\tilde{\boldsymbol{Q}} \in \mathcal{O}_n$, i.e., $\mathcal{O}_n$ is a multiplicative group
- $|\det(\boldsymbol{Q})| = 1$ $\quad (\text{since } 1 = \det(\boldsymbol{Q}^*\boldsymbol{Q}) = \det(\boldsymbol{Q}^*)\det(\boldsymbol{Q}) = \overline{\det(\boldsymbol{Q})}\det(\boldsymbol{Q}) = |\det(\boldsymbol{Q})|^2)$

Permutation matrices belong to $\mathcal{O}_n$: let $\pi : \{1,\ldots,n\} \to \{1,\ldots,n\}$ be a permutation, $\mathbf{e}^j := (0,\ldots,0,1,0\ldots,0)^\top = (\delta_{k,j})_{k=1}^j$ the jth coordinate vector. Then

$$\boldsymbol{P}_\pi := (\mathbf{e}^{\pi(1)}, \ldots \mathbf{e}^{\pi(n)}) \in \mathcal{O}_n \quad \text{and} \quad \boldsymbol{A}\boldsymbol{P}_\pi = (\mathbf{a}^{\pi(1)}, \ldots, \boldsymbol{A}^{\pi(n)})$$

Hence $\boldsymbol{P}_\pi^\top \boldsymbol{A}$ permutes the rows of $\boldsymbol{A}$ according to π

Normed Linear Spaces

How to "measure" vectors, functions?

Let $\mathbb{V}$ be a $\mathbb{K}$-vector space (finite or infinite dimensional). Any mapping $\|\cdot\| : \mathbb{V} \to \mathbb{R}_+$ is called a norm on $\mathbb{V}$ if

(N1) $\|v\| = 0 \quad \Rightarrow \quad v = 0$

(N2) $\|\alpha v\| = |\alpha| \, \|v\|$, $\alpha \in \mathbb{K}$, $v \in \mathbb{V}$

(N3) $\|v + z\| \leq \|v\| + \|z\|$, $v, z \in \mathbb{V}$ (triangle inequality)

The pair $(\mathbb{V}, \|\cdot\|)$ is called a a normed linear space. When the choice of the norm is clear we simply say that $\mathbb{V}$ is a normed linear space.

Remark 14

The absolute value is a norm on $\mathbb{V} = \mathbb{K}$. Up to a scaling factor this is the only norm on $\mathbb{K} = \mathbb{R}$. For higher dimensional spaces many different norms exist. Their choice depends on the purpose or application. Every norm is a Lipschitz continuous mapping with Lipschitz constant equal to one, i.e.,

$$\big|\|v\| - \|z\|\big| \leq \|v - z\|, \quad v, z \in \mathbb{V}. \tag{5.1}$$

To see (5.1), note that by (N3) $\|v\| = \|v - z + z\| \leq \|v - z\| + \|z\| \Rightarrow \|v\| - \|z\| \leq \|v - z\|$.

The same argument shows that $\|z\| - \|v\| \leq \|v - z\|$.

Examples Vector and Sequence Norms

p-norms on $\mathbb{K}^n$, $1 \le p \le \infty$:

$$\|\mathbf{x}\|_p := \Big(\sum_{j=1}^{n} |x_j|^p\Big)^{1/p}, \quad 1 \le p < \infty, \qquad \|\mathbf{x}\|_\infty := \max_{j=1,\dots,n} |x_j|$$

For any fixed positive "weights" $\omega_j, j = 1, \dots, n$, the "weighted" counterparts

$$\|\mathbf{x}\|_{p,\omega} := \Big(\sum_{j=1}^{n} \omega_j |x_j|^p\Big)^{1/p} \tag{5.2}$$

define norms as well.

p-norms on $\mathbb{K}^\infty$ (sequences), $1 \le p \le \infty$:

$$\|\mathbf{x}\|_p := \Big(\sum_{j=1}^{\infty} |x_j|^p\Big)^{1/p}, \quad 1 \le p < \infty, \qquad \|\mathbf{x}\|_\infty := \sup_{j=1,\dots,n} |x_j|$$

In this case one often uses the notation $\|\cdot\|_{\ell_p}$ and ℓ_p denotes the space of all sequences for which $\|\mathbf{x}\|_{\ell_p}$ is finite. When using specific index sets $\mathcal{I}$ different from $\mathbb{N}$, one writes $\ell_p(\mathcal{I})$ viewing a sequence as a function mapping a discrete domain $\mathcal{I}$ to $\mathbb{K}$. To stress the dependence on a finite dimension *n* we also write briefly $\ell_p^n = \ell_p(\{1, \dots, n\})$

Examples Vector and Sequence Norms

$\|\cdot\|_p, \|\cdot\|_{\ell_p}$ are indeed norms:

- For $p = 1, \infty$ this follows directly corresponding properties of the absolute value.
- For $1 < p < \infty$ properties (N1), (N2) for $\|\cdot\|_p, \|\cdot\|_{\ell_p}$ follow directly as well. The triangle inequality (N3) is in this case less obvious. The main tool for verifying (N3) as well is the following important inequality:

Hölder's Inequality: for $\frac{1}{p} + \frac{1}{p^*} = 1$ one has

$$\left| \sum_{j \in \mathcal{I}} x_j \bar{y}_j \right| \leq \|\mathbf{x}\|_{\ell_p(\mathcal{I})} \|\mathbf{y}\|_{\ell_{p^*}(\mathcal{I})}, \quad \mathbf{x} \in \ell_p(\mathcal{I}),\ \mathbf{y} \in \ell_{p^*}(\mathcal{I}). \tag{5.3}$$

We'll show later for the special case $p = 2$ how this implies the triangle inequality (N3).

Defining $\mathbf{x}\bar{\mathbf{y}} := (x_j \bar{y}_j)_{j \in \mathcal{I}}$, (5.3) implies

$$\|\mathbf{x}\bar{\mathbf{y}}\|_{\ell_1} \leq \|\mathbf{x}\|_{\ell_p(\mathcal{I})} \|\mathbf{y}\|_{\ell_{p^*}(\mathcal{I})}.$$

Operator Norms

One can "measure" the mapping properties of a linear operator by monitoring how it distorts the unit ball:

$$\|L\|_{\mathcal{L}(\mathbb{X},\mathbb{Y})} := \sup_{x\in\mathbb{X};\|x\|_{\mathbb{X}}=1} \|L(x)\|_{\mathbb{Y}} = \sup_{x\in\mathbb{X}\setminus\{0\}} \frac{\|L(x)\|_{\mathbb{Y}}}{\|x\|_{\mathbb{X}}}$$

Verify: $\|L\|_{\mathcal{L}(\mathbb{X},\mathbb{Y})}$ is a norm on $\mathcal{L}(\mathbb{X},\mathbb{Y})$

The operator norm $\|L\|_{\mathcal{L}(\mathbb{X},\mathbb{Y})}$ is induced by and thus depends on the norms $\|\cdot\|_{\mathbb{X}}, \|\cdot\|_{\mathbb{Y}}$.

- L is called bounded if $\|L\|_{\mathcal{L}(\mathbb{X},\mathbb{Y})} < \infty$
- $\rightsquigarrow \|L(x)\|_{\mathbb{Y}} = \frac{\|L(x)\|_{\mathbb{Y}}}{\|x\|_{\mathbb{X}}}\|x\|_{\mathbb{X}} \le \sup_{y\in\mathbb{X}\setminus\{0\}} \frac{\|L(y)\|_{\mathbb{Y}}}{\|y\|_{\mathbb{X}}}\|x\|_{\mathbb{X}} \rightsquigarrow$

$$\|L(x)\|_{\mathbb{Y}} \le \|L\|_{\mathcal{L}(\mathbb{X},\mathbb{Y})}\|x\|_{\mathbb{X}}, \quad x \in \mathbb{X}. \tag{5.4}$$

  In other words, if $L \in \mathcal{L}(\mathbb{X},\mathbb{Y})$ is bounded there exists a constant $C < \infty$ such that $\|L(x)\|_{\mathbb{Y}} \le C\|x\|_{\mathbb{X}}$, $x \in \mathbb{X}$, and $C = \|L\|_{\mathcal{L}(\mathbb{X},\mathbb{Y})}$ is the smallest such constant.
- A bounded linear operator is Lipschitz continuous with Lipschitz constant $\|L\|_{\mathcal{L}(\mathbb{X},\mathbb{Y})}$:

$$\|L(x) - L(z)\|_{\mathbb{Y}} = \|L(x-z)\|_{\mathbb{Y}} \le \|L\|_{\mathcal{L}(\mathbb{X},\mathbb{Y})}\|x - z\|_{\mathbb{X}};$$

- $L \in \mathcal{L}(\mathbb{X},\mathbb{Y}), R \in \mathcal{L}(\mathbb{Y},\mathbb{W}) \rightsquigarrow \|R \circ L\|_{\mathcal{L}(\mathbb{X},\mathbb{W})} \le \|R\|_{\mathcal{L}(\mathbb{Y},\mathbb{W})}\|L\|_{\mathcal{L}(\mathbb{X},\mathbb{Y})}$.

Matrix Norms

When viewing $\boldsymbol{A} \in \mathbb{K}^{m\times n}$ as a "vector" in $\mathbb{K}^{mn}$ one could, in principle, use p-norms for matrices as well. However, these norms would not reflect the properties of $\boldsymbol{A}$ when viewed as a linear operator from $\mathbb{K}^n$ to $\mathbb{K}^m$. Here are some operator norms for matrices: let $\boldsymbol{A} \in \mathbb{K}^{m\times n} \equiv \mathcal{L}(\mathbb{K}^n, \mathbb{K}^m)$

$$\|\boldsymbol{A}\|_\infty = \|\boldsymbol{A}\|_{\mathcal{L}(\ell_\infty,\ell_\infty)} = \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|\boldsymbol{A}\mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty} = \max_{j=1,\ldots,m} \sum_{k=1}^{n} |a_{j,k}| \quad \text{(maximal row sum)}$$

$$\|\boldsymbol{A}\|_1 = \|\boldsymbol{A}\|_{\mathcal{L}(\ell_1,\ell_1)} = \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|\boldsymbol{A}\mathbf{x}\|_1}{\|\mathbf{x}\|_1} = \max_{k=1,\ldots,n} \sum_{j=1}^{m} |a_{j,k}| \quad \text{(maximal column sum)}$$

The so called spectral norm, induced by the Euclidean norm $\|\cdot\|_2$ requires the notion of eigenvalue of a matrix. The set of eigenvalues of $\boldsymbol{A}$ is often called "spectrum" of $\boldsymbol{A}$. This will be discussed in more detail a little later. Here it sufficies to know that $\lambda \in \mathbb{K}, \mathbf{x} \in \mathbb{K}^n$ are called eigenvalue and eigenvector of $\boldsymbol{A}$ if $\boldsymbol{A}\mathbf{x} = \lambda\mathbf{x}$, i.e., $\boldsymbol{A}$ just strechtes or shortens an eigenvector while preserving its direction. Moreover (see Proposition 13 later below) the matrix $\boldsymbol{A}^*\boldsymbol{A}$ has only nonnegative eigenvalues. We denote by $\lambda_{\max}(\boldsymbol{A}^*\boldsymbol{A})$ the largest eigenvalue of $\boldsymbol{A}^*\boldsymbol{A}$. Then one can show that for $\boldsymbol{A} \in \mathbb{K}^{m\times n}$

$$\|\boldsymbol{A}\|_2 = \|\boldsymbol{A}\|_{\mathcal{L}(\ell_2^n,\ell_2^m)} = \sqrt{\lambda_{\max}(\boldsymbol{A}^*\boldsymbol{A})} \tag{5.5}$$

Unitary Matrices "Love" the Euclidean Norm

Proposition 15

Let $\boldsymbol{Q} \in \mathcal{O}_n$

1. *Unitary matrices map Euclidean spheres into themselves:* $\|\boldsymbol{Q}\mathbf{x}\|_2 = \|\mathbf{x}\|_2$, $\mathbf{x} \in \mathbb{K}^n$;
2. *they have minimal condition numbers:* $\kappa_2(\boldsymbol{Q}) = 1$;
3. *multiplying a matrix with a unitary matrix does not change the spectral norm of the matrix:* $\|\boldsymbol{A}\|_2 = \|\boldsymbol{QA}\|_2 = \|\boldsymbol{AQ}\|_2$;
4. *multiplying a matrix with a unitary matrix does not change the spectral norm of the matrix:* $\kappa_2(\boldsymbol{A}) = \kappa_2(\boldsymbol{QA}) = \kappa_2(\boldsymbol{AQ})$

Proof: (1): $\|\boldsymbol{Q}\mathbf{x}\|_2^2 = (\boldsymbol{Q}\mathbf{x})^* \boldsymbol{Q}\mathbf{x} = \mathbf{x}^* \boldsymbol{Q}^* \boldsymbol{Q}\mathbf{x} = \mathbf{x}^*\mathbf{x} = \|\mathbf{x}\|_2^2$;

(2): (1) $\Rightarrow \|\boldsymbol{Q}\|_2 = 1 \Rightarrow \|\boldsymbol{Q}^*\|_2 = \|\boldsymbol{Q}^{-1}\|_2 = 1 \Rightarrow \kappa_2(\boldsymbol{Q}) = \|\boldsymbol{Q}\|_2 \|\boldsymbol{Q}^{-1}\|_2 = 1$;

Well-Posedness

A problem is called well posed if

1. There **exists** a solution;
2. the solution is **unique**;
3. the solution depends **continuously** on the data

The problem is called ill-posed if at least one of these requirements fails to hold.

- To deal with ill-posed data one has to regularize the problem, i.e., one seeks a nearby well-posed problem and solves this one
- In finite dimensions “continuous dependence” is independent of a specific norm. In the infinite dimensional case the choice of the norm may be essential.

Condition of a Problem, Condition Numbers

The notion of Condition of a problem attempts to quantify how strongly the output depends on perturbations of the input data.

The **relative condition** of a linear operator $L \in \mathcal{L}(\mathbb{X}, \mathbb{Y})$ can be estimated by the (relative) condition number $\kappa_{\mathbb{X},\mathbb{Y}}(L)$ of L which is the smallest constant κ satisfying

$$\frac{\|L(x) - L(\tilde{x})\|_{\mathbb{Y}}}{\|L(x)\|_{\mathbb{Y}}} \leq \kappa \frac{\|x - \tilde{x}\|_{\mathbb{X}}}{\|x\|_{\mathbb{X}}}, \quad x, \tilde{x} \in \mathbb{X}, \tag{5.6}$$

hence quantifying how the relative output accuracy is controlled by the relative input error.

Proposition 16

One has

$$\kappa_{\mathbb{X},\mathbb{Y}}(L) = \frac{\sup_{x \in \mathbb{X}; \|x\|_{\mathbb{X}}=1} \|L(x)\|_{\mathbb{Y}}}{\inf_{x; \|x\|_{\mathbb{X}}=1} \|L(x)\|_{\mathbb{Y}}} = \frac{\|L\|_{\mathcal{L}(\mathbb{X},\mathbb{Y})}}{\inf_{x; \|x\|_{\mathbb{X}}=1} \|L(x)\|_{\mathbb{Y}}} \tag{5.7}$$

If L is invertible one has

$$\kappa_{\mathbb{X},\mathbb{Y}}(L) = \|L\|_{\mathcal{L}(\mathbb{X},\mathbb{Y})} \|L^{-1}\|_{\mathcal{L}(\mathbb{Y},\mathbb{X})}. \tag{5.8}$$

Thus, L and L^{-1} have the same condition number.

$\kappa_{\mathbb{X},\mathbb{Y}}(L)$ is the ratio of maximal expansion and maximal compression that can be caused by L.

Proof of Proposition 16

$$\|L(x) - L(\tilde{x})\|_{\mathbb{Y}} = \|L(x - \tilde{x})\|_{\mathbb{Y}} \overset{(5.4)}{\leq} \|L\|_{\mathcal{L}(\mathbb{X},\mathbb{Y})} \|x - \tilde{x}\|_{\mathbb{X}}, \tag{5.9}$$

and

$$\|L(x)\|_{\mathbb{X}} = \frac{\|L(x)\|_{\mathbb{Y}}}{\|x\|_{\mathbb{X}}} \|x\|_{\mathbb{X}} \geq \inf_{x' \neq 0} \frac{\|L(x')\|_{\mathbb{Y}}}{\|x'\|_{\mathbb{X}}} \|x\|_{\mathbb{X}} = \inf_{\|x'\|_{\mathbb{X}}=1} \|L(x')\|_{\mathbb{Y}} \|\mathbf{x}\|_{\mathbb{X}} \tag{5.10}$$

provide

$$\frac{\|L(x) - L(\tilde{x})\|_{\mathbb{Y}}}{\|L(x)\|_{\mathbb{Y}}} \leq \frac{\|L\|_{\mathcal{L}(\mathbb{X},\mathbb{Y})}}{\inf_{\|x'\|_{\mathbb{X}}=1} \|L(x')\|_{\mathbb{Y}}} \frac{\|x - \tilde{x}\|_{\mathbb{X}}}{\|x\|_{\mathbb{X}}}$$

which shows that $\kappa_{\mathbb{X},\mathbb{Y}}(L) \leq \frac{\|L\|_{\mathcal{L}(\mathbb{X},\mathbb{Y})}}{\inf_{\|x'\|_{\mathbb{X}}=1} \|L(x')\|_{\mathbb{Y}}}$. Show that one cannot do better.

Regarding (5.8), notice

$$\frac{1}{\inf_{\|x'\|_{\mathbb{X}}=1} \|L(x')\|_{\mathbb{Y}}} = \sup_{x' \neq 0} \frac{\|x'\|_{\mathbb{X}}}{\|L(x')\|_{Y}} \overset{(y=L(x'))}{=} \sup_{y \in \mathbb{Y}} \frac{\|L^{-1}(y)\|_{\mathbb{X}}}{\|y\|_{\mathbb{Y}}} = \|L^{-1}\|_{\mathcal{L}(\mathbb{Y},\mathbb{X})}$$

which completes the proof □

Condition Numbers and Residuals

Let $\boldsymbol{A} \in \mathbb{K}^{n\times n} \equiv \mathcal{L}(\mathbb{K}^n, \mathbb{K}^n)$ be nonsingular, let $\|\cdot\|$ be any norm on $\mathbb{K}^n$, and $\kappa(\boldsymbol{A}) = \|\boldsymbol{A}\|\|\boldsymbol{A}^{-1}\|$ the corresponding condition number.

Question: suppose that $\tilde{\mathbf{x}}$ is an approximate solution of $\boldsymbol{A}\mathbf{x} = \mathbf{b}$. The residual $\mathbf{r} := \mathbf{b} - \boldsymbol{A}\tilde{\mathbf{x}}$ is a computable and hence known quantity. What does the residual tell us about the unknown error $\tilde{\mathbf{x}} - \mathbf{x}$?

$$\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \leq \kappa(\boldsymbol{A}) \frac{\|\mathbf{r}\|}{\|\mathbf{b}\|} \tag{5.11}$$

Hence, the smaller $\kappa(\boldsymbol{A})$ the more accurate is the information provided by the (known) residual about the (unknown) error. To see (5.11), note

$$\|\mathbf{x} - \tilde{\mathbf{x}}\| = \|\boldsymbol{A}^{-1}\boldsymbol{A}(\mathbf{x} - \tilde{\mathbf{x}})\| = \|\boldsymbol{A}^{-1}(\mathbf{b} - \boldsymbol{A}\tilde{\mathbf{x}})\| \leq \|\boldsymbol{A}^{-1}\|\|\mathbf{b} - \boldsymbol{A}\tilde{\mathbf{x}}\| = \|\boldsymbol{A}^{-1}\|\|\mathbf{r}\|, \quad \|\mathbf{b}\| \leq \|\boldsymbol{A}\|\|\mathbf{x}\|$$

which gives (5.11).

Why does this imply immediately also

$$\frac{\|\mathbf{b} - \boldsymbol{A}\tilde{\mathbf{x}}\|}{\|\mathbf{b}\|} = \frac{\|\mathbf{r}\|}{\|\mathbf{b}\|} \leq \kappa(\boldsymbol{A}) \frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\|\mathbf{x}\|}? \tag{5.12}$$

Further Comments

- Applying an operator $L \in \mathcal{L}(\mathbb{X}, \mathbb{Y})$ to some input $x \in \mathbb{X}$, i.e., $L : x \mapsto L(x)$ (e.g. a matrix vector multiplication $\boldsymbol{A} : \mathbf{x} \mapsto \boldsymbol{A}\mathbf{x}$) and solving an "operator equation" $L(x) = y$,.e., $y \mapsto L^{-1}(y)$ ($\mathbf{x} \mapsto \boldsymbol{A}^{-1}\mathbf{b}$) are mathematical operations with the same condition numbers.

- A linear operator on a finite dimensional space is always bounded (hence continuous). In fact, on account of Proposition 7, this has to be verified only for matrices and some fixed norm. For instance, the maximal row-sum norm is trivially bounded. Whenever $\boldsymbol{A}$ is non-singular $\boldsymbol{A}^{-1}$ has for the same reason a finite norm as well. Therefore,

$$\boldsymbol{A}\mathbf{x} = \mathbf{b} \quad \text{is a well-posed problem} \quad \Leftrightarrow \quad \boldsymbol{A}^{-1} \text{ exits} \quad \Leftrightarrow \quad \kappa(\boldsymbol{A}) < \infty \text{ (any norm)}$$

Stability of a Basis

Suppose that $\mathbb{X}$ is a finite-dimensional normed linear space and $\Phi = \{\phi_1, \ldots, \phi_n\} \subset \mathbb{X}$ is a basis for $\mathbb{X}$. We wish to quantify the quality of the basis in the following sense: suppose we know only perturbed coefficients $\tilde{x}_j$, $j = 1, \ldots, n$ of the representation $x = \sum_{j=1}^n x_j \phi_j$, how much does $\tilde{x} = \sum_{j=1}^n \tilde{x}_j \phi_j$ differ from x? If we decide to measure the perturbation of the coefficient vector in the norm $\|\cdot\|_p$ on $\mathbb{K}^n$, say, we ask to estimate the relative error $\frac{\|\tilde{x}-x\|_{\mathbb{X}}}{\|x\|_{\mathbb{X}}}$ in terms of the relative error

$$\frac{\|\tilde{\mathbf{x}} - \mathbf{x}\|_p}{\|\mathbf{x}\|_p} = \frac{\|B_\Phi(\tilde{x}) - B_\Phi(x)\|_p}{\|B_\Phi(x)\|_p} \quad \overset{???}{\leftrightarrow} \quad \frac{\|\tilde{x} - x\|_{\mathbb{X}}}{\|x\|_{\mathbb{X}}} \tag{5.13}$$

Thus, the relation between these relative errors is precisely described by the condition number $\kappa_{\mathbb{X},\ell_p^n}(B_\Phi)$ of the mapping $B_\Phi : \mathbb{X} \to \mathbb{K}^n$, induced by (3.1). By the previous comments, B_Φ is a linear and bijective mapping between finite dimensional spaces and thus has a bounded condition number. Hence, the basis Φ is "the more stable" the smaller $\kappa_{\mathbb{X},\ell_p^n}(B_\Phi)$ is. Stability of a basis Φ (in a quantitative sense) is therefore synonimous to a small condition number of the coordinate mapping B_Φ.

Finite vs. Infinite Dimensional Spaces

$\ell_p(\mathcal{I}) := \{\mathbf{x} \in \mathbf{K}^{\mathcal{I}} : \|\mathbf{x}\|_{\ell_p(\mathcal{I})} < \infty\}$, $B_r(\ell_p(\mathcal{I})) := \{\mathbf{x} \in \mathbf{K}^{\mathcal{I}} : \|\mathbf{x}\|_{\ell_p(\mathcal{I})} \leq r\}$ ball of radius r.

There are essential differences between $\#(\mathcal{I}) = n < \infty$ and $\#(\mathcal{I}) = \infty$: $(\mathbb{K}^{\mathcal{I}} = (\mathbb{K} \setminus \{\infty\})^{\mathcal{I}})$

- $\|\mathbf{x}\|_p < \infty$ for all $\mathbf{x} \in \mathbb{K}^n$. For $\#(\mathcal{I}) = \infty$ there exist $\mathbf{x} \in \mathbb{K}^{\mathcal{I}}$ for which $\|\mathbf{x}\|_{\ell_p(\mathcal{I})} = \infty$, i.e., $\ell_p(\mathcal{I})$ is a strict subset of $\mathbb{K}^{\mathcal{I}}$ and $\ell_p(\mathcal{I})$ differs from $\ell_q(\mathcal{I})$ when $p \neq q$.
- For $\#(\mathcal{I}) = \infty$ closed balls $B_r(\ell_p(\mathcal{I}))$ are not compact (they are for $\#(\mathcal{I}) < \infty$).
- Visualize $B_1(\ell_p^2)$ for different p, including $p = 1, 2, \infty$
- Determine the volume of $B_1(\ell_p^n)$ for $p = 1, 2, \infty$.
- If you draw N points in $[-1, 1]^n = B_1(\ell_\infty^n)$ randomly according to the uniform distribution. How many of those points do you expect to find on average in the Euclidean ball $B_1(\ell_2^n)$ and in the ℓ_1-ball $B_1(\ell_1^n)$?
- Linear operators on infinite dimensional spaces are no longer automatically bounded.

Finite vs. Infinitely Dimensional Spaces

How different can different norms be?

Proposition 17

Let $\mathbb{V}$ be $\mathbb{K}$-vector space. If $\dim(\mathbb{V}) = n < \infty$ then all norms on $\mathbb{V}$ are equivalent, i.e., for any two norms $\|\cdot\|_*, \|\cdot\|_{**}$ on $\mathbb{V}$ there exists constants $0 < c, C < \infty$ such that

$$c\|v\|_* \leq \|v\|_{**} \leq C\|v\|_*, \quad \forall\, v \in \mathbb{V}. \tag{5.14}$$

This is in essence a consequence of the fact that any closed bounded subsets of a finite-dimensional space is compact and that continuous functions on compact sets attain their extrema in those sets.

In particular, recall:

Theorem (Heine-Borel):

All bounded closed sets in $\mathbb{K}^n$ are compact.

Sketch of Proof of Prop. 17: Note that (5.14) is equivalent to

$$c \leq \|v\|_{**} \leq C \quad \forall\, v \in B_1((\mathbb{V}, \|\cdot\|_*)). \tag{5.15}$$

Now, fix a basis $\{\phi, \ldots, \phi_n\}$ for $\mathbb{V}$. As shown earlier every $v \in \mathbb{V}$ has a unique coefficient vector $\mathbf{v} = \Phi(v) \in \mathbb{K}^n$. Define $\|v\|_n := \|\mathbf{v}\|_\infty$. Then it suffices to show that (5.15) holds for any norm $\|\cdot\|_{**}$ on $\mathbb{V}$ and $\|\cdot\|_* = \|\cdot\|_n$. To that end, recall
Moreover, we know that every continuous function attains its minimum and maximum on a compact set. The function $F(\mathbf{v}) := \left\| \sum_{j=1}^n v_j \phi_j \right\|_{**} = \|B_\Phi^{-1}(\mathbf{v})\|_{**}$ is clearly continuous because $B_\Phi^{-1} : \mathbf{v} \to \sum_{j=1}^n v_j \phi_j$ is easily seen to be continuous and $v \to \|v\|_{**}$ is by (5.1) even Lipschitz continuous. Since by Heine-Borel, $B_1(\ell_\infty^n)$ is compact, $F(\mathbf{v})$ attains its minimum and maximum on $B_1(\ell_\infty^n)$. Since by linear independence, $B_\Phi^{-1}(\mathbf{v}) \neq 0$ iff $\mathbf{v} \neq \mathbf{0}$ (N1) implies that $c := \min_{\mathbf{v} \in B_1(\ell_\infty^n)} \|B_\Phi^{-1}(\mathbf{v})\|_{**} > 0$. The upper bound is handeled analogously. □

Finite vs. Infinitely Dimensional Spaces

Remark 18

The statement in Proposition 17 needs to be read with caution. The equivalence constants c, C in (5.14) *depend on the dimension n. For instance,*

$$\|\mathbf{x}\|_\infty = \max_{j=1,\dots,n} |x_j| \le \Big(\sum_{j=1}^{n} |x_j|^2\Big)^{1/2} = \|\mathbf{x}\|_2 \le \sqrt{n}\|\mathbf{x}\|_\infty.$$

Hence, the larger the dimension the more the choice of a norm matters.

In infinite dimensions there could be issues with convergence. In that regard, the ℓ_p-spaces are not so bad in the following sense

Definition 19

A normed linear space $(\mathbb{V}, \|\cdot\|)$ is called complete if every Cauchy sequence in $\mathbb{V}$ converges to an element in $\mathbb{V}$, i.e.,

$$\|v_k - v_n\| \to 0, \quad k, n \to \infty, \quad \Rightarrow \quad \exists\, v \in \mathbb{V}, \text{ s.t. } \|v_k - v\| \to 0, \quad k \to \infty.$$

Complete normed linear spaces are called Banach spaces.

Norms on Function Spaces

Remark 20

Obviously, completeness is useful to be sure that, if one constructs a sequence, the Cauchy property (which one may be able to verify) already ensures that this sequence has a limit in the space one is working in.

- *All finite dimensional spaces, endowed with any norm, are always complete.*
- *$\ell_p(\mathcal{I})$ are Banach spaces.*

Things get more subtle when dealing with function spaces. One can formally define again for $1 \leq p \leq \infty$ and a given function $f : \Omega \subset \mathbb{R}^n \to \mathbb{K}$, say:

$$\|f\|_{L_p(\Omega)} := \Big(\int_\Omega |f(x)|^p dx \Big)^{1/p}, \quad 1 \leq p < \infty, \quad \|f\|_{L_\infty(\Omega)} := \sup_{x \in \Omega} |f(x)|. \tag{5.16}$$

This is an obvious analogy to sequence norms replacing discrete arguments $j \in \mathcal{I}$ and summation by continuous arguments x and integration.

Again, the verification of properties (N1), (N2) (and (N3) when $p = \infty$) is straight forward. For $1 < p < \infty$ the triangle inequality (N3) - also referred to as Minkowski's inequality, rests again the continuous version of Hölder's lequality

$$\Big| \int_\Omega f(x)\overline{g(x)} dx \Big| \leq \|f\|_{L_p(\Omega)} \|g\|_{L_{p^*}(\Omega)}, \quad \frac{1}{p} + \frac{1}{p^*}. \tag{5.17}$$

L_p-Spaces Continued

Can be skipped, just for interested readers

When Ω is a bounded domain and f is continuous, the expressions (5.16) are well defined and $C(\Omega;\mathbb{K}) := \{f : \Omega \to \mathbb{K} : f \text{ continuous in } \Omega\}$, endowed with any of the norms $\|\cdot\|_{L_p(\Omega)}$ becomes a normed linear space. However, completeness is an issue.

Remark 21

Consider $\Omega = (-1,1) \subset \mathbb{R}$, $f_n(x) := 1$, $x \in (-1,0]$, $f_n(x) := \max\{0, 1 - xn\}$. One easily checks that this is a Cauchy sequence for every $1 \le p < \infty$ but not for $p = \infty$. All f_n are continuous, but the f_n don't have a limit in $C((-1,1))$. In fact, the pointwise limit is $f(x) = 1$, $x \in (-1,0]$, $f(x) = 0$, $x \in (0,1)$ which is also the limit in $L_p((-1,1))$ for $1 \le p < \infty$. This hints at the facts:

- *$(C(\Omega), L_p(\Omega))$ is not a complete normed linear space for $p < \infty$;*
- *$(C(\Omega), \|\cdot\|_{L_\infty(\Omega)})$ is complete and hence a Banach space.*

Regarding a suitable notion of normed linear spaces for $1 \le p < \infty$, completeness is an issue. First, one should employ the right notion of integration is used. The conceptually simple Riemann integration will not lead to complete spaces. Instead, integration is always understood in the Lebesgue sense based on Measure Theory. Starting form measurable sets and measurable functions, the space $L_p(\Omega)$ is defined as the closure of step functions with respect to the above norms. So these spaces are complete by definition. Then the following has to be kept in mind:

- The elements of the spaces $(L_p(\Omega), \|\cdot\|_{L_p(\Omega)})$ are, stricty speaking only equivalence classes of functions, where elements of one class differ only on sets of measure zero. For instance, points have measure zero in $\mathbb{R}$, points, lines, curves have mesure zero in $\mathbb{R}^2$, etc. Therefore, it does not make sense to ask for point values of elements in $L_p(\Omega)$. Keeping this in mind, we simply speak of "functions" in $L_p(\Omega)$.
- In the definition of $L_\infty(\Omega)$ which is a strictly larger Banach space than $C(\Omega)$ with the same norm.

Pre-Hilbert Spaces

Norms induced by scalar products play an important role:

Remark 22

Suppose $\mathbb{V}$ is $\mathbb{K}$-vector space with a scalar product $\langle \cdot, \cdot \rangle_{\mathbb{V}}$. Then

$$\|v\|_{\mathbb{V}} := \langle v, v \rangle_{\mathbb{V}}^{1/2} \tag{5.18}$$

is a norm on $\mathbb{V}$. A linear space with a scalar product and associated norm is called a Pre-Hilbert space. If $\mathbb{V}$ is complete under this norm it is called a Hilbert space. Hilbert spaces are in some sense closest to finite dimensional Euclidean spaces.

Properties (N1), (N2) of a norm follow directly from the properties of a scalar product. The triangle inequality (N3) follows from the Cauchy-Schwarz Inequality

$$|\langle v, w \rangle_{\mathbb{V}}| \leq \|v\|_{\mathbb{V}} \|w\|_{\mathbb{V}}, \quad v, w \in \mathbb{V}. \tag{5.19}$$

In fact,

$$\begin{aligned} \|v+w\|_{\mathbb{V}}^2 &= \langle v+w, v+w \rangle_{\mathbb{V}} = \langle v, v \rangle_{\mathbb{V}} + \langle w, w \rangle_{\mathbb{V}} + \langle v, w \rangle_{\mathbb{V}} + \langle w, v \rangle_{\mathbb{V}} \\ &= \|v\|_{\mathbb{V}}^2 + \|w\|^2 + \langle v, w \rangle_{\mathbb{V}} + \langle w, v \rangle_{\mathbb{V}} \leq \|v\|_{\mathbb{V}}^2 + \|w\|^2 + 2\|v\|_{\mathbb{V}} \|w\|_{\mathbb{V}} \\ &= \left(\|v\|_{\mathbb{V}} + \|w\|_{\mathbb{V}}\right)^2 \end{aligned}$$

Proof of the Cauchy Schwarz Inequality

Assume that $v, w \neq 0$ (otherwise the inequality is trivial). Then for any $\lambda \in \mathbb{K}$

$$0 \quad \leq \quad \langle v - \lambda w, v - \lambda w \rangle_{\mathbb{V}} = \|v\|_{\mathbb{V}}^2 + |\lambda|^2 \|w\|_{\mathbb{V}}^2 - \lambda \langle w, v \rangle_{\mathbb{V}} - \bar{\lambda} \langle v, w \rangle_{\mathbb{V}}.$$

Choose $\lambda = \frac{\langle v, w \rangle_{\mathbb{V}}}{\|w\|_{\mathbb{V}}^2}$ to obtain

$$0 \leq \|v\|_{\mathbb{V}}^2 + \frac{|\langle w, v \rangle_{\mathbb{V}}|^2}{\|w\|_{\mathbb{V}}^4} \|w\|_{\mathbb{V}}^2 - 2 \frac{|\langle w, v \rangle_{\mathbb{V}}|^2}{\|w\|_{\mathbb{V}}^2} = \|v\|_{\mathbb{V}}^2 - \frac{|\langle w, v \rangle_{\mathbb{V}}|^2}{\|w\|_{\mathbb{V}}^2},$$

which implies (5.19). □

Examples

- $(\mathbb{V}, \|\cdot\|) = (\mathbb{K}^n, \|\cdot\|_2)$,

$$\langle \mathbf{x}, \mathbf{y} \rangle_n = \mathbf{y}^* \mathbf{x}, \qquad \|\mathbf{x}\|_2^2 = \mathbf{x}^* \mathbf{x} = \sum_{j=1}^{n} |x_j|^2$$

- $(\ell_2(\mathbb{N}), \|\cdot\|_{\ell_2(\mathbb{N})})$

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbb{N}} = \sum_{j \in \mathbb{N}} x_j \overline{y_j} = \mathbf{y}^* \mathbf{x}$$

- $(L_2(\Omega), \|\cdot\|_{L_2(\Omega)})$,

$$\langle f, g \rangle_{L_2(\Omega)} = \int_\Omega f(x) \overline{g(x)} dx, \qquad \|f\|_{L_2(\Omega)}^2 = \int_\Omega |f(x)|^2 dx$$

- $(L_{2,\pi}, \|\cdot\|_{L_2(-\pi,\pi)})$, 2π-periodic square integrable functions

$$\langle f, g \rangle_{(-\pi,\pi)} := \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) \overline{g(x)} dx,$$

$$e_k(x) := e^{ikx}, \quad \hat{f}(k) := \langle f, e_k \rangle_{(-\pi,\pi)} = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{-ikx} dx \quad \text{(Fourier coefficients)}$$

Dual Spaces

Let $(\mathbb{X}, \|\cdot\|_{\mathbb{X}})$ be a Banach space. The collection of all bounded linear functionals $\mathbb{X}^* := \mathcal{L}(\mathbb{X}, \mathbb{K})$ is also a Banach space under the norm

$$\|g\|_{\mathbb{X}^*} := \sup_{x \in \mathbb{X}\setminus\{0\}} \frac{g(x)}{\|x\|_{\mathbb{X}}}. \tag{5.20}$$

Examples:

- $C^k(\Omega) :=$ collection of all continuously differentiable functions on Ω, $\|f\|_{C^k(\Omega)} := \max_{0\le j\le k} \|f^{(j)}\|_{L_\infty(\Omega)}$; then $\delta_{x_0}^{(k)} : f \mapsto f^{(k)}(x_0)$ belongs to $(C^k(\Omega))^*$ because

$$\|\delta_{x_0}^{(k)}\|_{(C^k(\Omega))^*} = \sup_{f\in C^k(\Omega)} \frac{\delta_{x_0}^{(k)}(f)}{\|f\|_{C^k(\Omega)}} = \sup_{f\in C^k(\Omega)} \frac{f^{(k)}(x_0)}{\|f\|_{C^k(\Omega)}} \le 1$$

- $\boldsymbol{A} \in \mathbb{K}^{m\times n}$, $\mathbf{y} \in \mathbb{K}^n$ fixed, then $g(\mathbf{x}) := (\boldsymbol{A}\mathbf{y})^*\mathbf{x}$ belongs to $(\ell_2^m)^* = \mathcal{L}((\mathbb{K}^m, \|\cdot\|_2), \mathbb{K})$ and $\|g\|_{(\ell_2^m)^*} = \|\boldsymbol{A}\mathbf{y}\|_2$

Theorem 23

Riesz Representation Theorem: let $(\mathbb{V}, \|\cdot\|_{\mathbb{V}})$ be a Hilbert space. There exists a linear mapping $R = R_{\mathbb{V}^*\to\mathbb{V}} : \mathbb{V}^* \to \mathbb{V}$ such that for any $g \in \mathbb{V}^*$, $v \in \mathbb{V}$, one has $g(v) = \langle v, Rg\rangle_{\mathbb{V}}$ and $\|Rg\|_{\mathbb{V}} = \|g\|_{\mathbb{V}^*}$, i.e., R is an isometry $\|R\|_{\mathcal{L}(\mathbb{V}^*,\mathbb{V})} = 1$.

Orthogonal Projections and Best Approximations

Given a (finite dimensional) subspace $\mathbb{U}$ of a normed linear space $\mathbb{V}$ and given any $v \in \mathbb{V}$, the problem of best approximation to v from $\mathbb{U}$ is to find an element $u(v) \in \mathbb{U}$ that is closest to v with respect to a given norm. In general, this is a difficult problem but when $\mathbb{V}$ is a Hilbert space, it amounts to a linear projection.

Theorem 24

Let $(\mathbb{V}, \|\cdot\|_{\mathbb{V}})$ be a Hilbert space and let $\mathbb{U} \subset \mathbb{V}$ be a finite dimansional subspace (this holds in greater generality). Given any $v \in \mathbb{V}$, then some $u(v) \in \mathbb{U}$ satisfies

$$\|v - u(v)\|_{\mathbb{V}} = \min_{u \in \mathbb{U}} \|v - u\|_{\mathbb{V}} \tag{5.21}$$

if and only if

$$\langle v - u(v), u\rangle_{\mathbb{V}} = 0 \quad \forall\, u \in \mathbb{U} \quad \textit{i.e.,} \quad v - u(v) \perp \mathbb{U} \tag{5.22}$$

Remark 25

1. *If $u \in \mathbb{U}$ and $\langle u, w\rangle_{\mathbb{V}} = 0$, $\forall\, w \in \mathbb{U}$ then $u = 0$, i.e., the only element in a linear space that is orthogonal to all elements in the same space is the zero element. In fact, for $w = u$ one has $0 = \langle u, u\rangle_{\mathbb{V}} = \|u\|^2_{\mathbb{V}}$ which, by (N1) implies $u = 0$.*
2. *Given $v \in \mathbb{V}$ there exists at most one $\bar{u} \in \mathbb{U}$ such that $\langle v - \bar{u}, u\rangle_{\mathbb{V}} = 0$, $\forall\, u \in \mathbb{U}$. In fact, suppose $u_1, u_2 \in \mathbb{U}$ have that property. Then $0 = \langle v - u_1, w\rangle_{\mathbb{V}} - \langle v - u_2, w\rangle_{\mathbb{V}} = \langle u_2 - u_1, w\rangle_{\mathbb{V}}$, $\forall\, w \in \mathbb{U}$, so, by (1), $u_1 = u_2$.*

Proof of Theorem 24

The proof uses a standard variational argument. Consider a candidate $u \in \mathbb{U}$ for (5.21) and any $t \in \mathbb{R}$, $w \in \mathbb{U}$. We wish to see under which circumstances a perturbation $u + tw \in \mathbb{U}$ could do better.

$$\begin{aligned} \|v - (u + tw)\|_{\mathbb{V}}^2 &= \langle (v-u) - tw, (v-u) - tw \rangle_{\mathbb{V}} = \|v-u\|_{\mathbb{V}}^2 + t^2 \|w\|_{\mathbb{V}}^2 \\ &\quad -2t\mathrm{Re}(\langle v-u, w \rangle_{\mathbb{V}}) \end{aligned} \tag{5.23}$$

(5.21) $\Rightarrow$ (5.22): Suppose $u = u(v)$ is the minimizer but there exists a $w \in \mathbb{U}$, $w \neq 0$ such that $\langle v-u, w \rangle_{\mathbb{V}} = \alpha \neq 0$, and hence $\langle v-u, iw \rangle_{\mathbb{V}} \neq 0$. Without loss of generality we can assume $\beta := \mathrm{Re}(\alpha) > 0$, $\|w\|_{\mathbb{V}} = 1$. Then, by optimality of $u = u(v)$ we must have

$$0 < \|v - (u + tw)\|_{\mathbb{V}}^2 - \|v-u\|_{\mathbb{V}}^2 = t^2 - 2t\beta \quad \forall\, t \in \mathbb{R}.$$

But choosing $t = \beta$, yields $\beta^2 - 2\beta^2 < 0$, which is a contradiction.

(5.22) $\Rightarrow$ (5.21): By (5.23) we have in this case $\|v - (u+w)\|_{\mathbb{V}}^2 - \|v-u\|_{\mathbb{V}}^2 = \|w\|_{\mathbb{V}}^2$ for all $w \in \mathbb{U}$, which implies (5.21). $\square$

Orthogonal Projections and Best Approximations

Remark 26

Theorem 24 doesn't explicitly state existence of a minimizer only the equivalence of (5.21) *and* (5.22). *Existence of a minimizer in* (5.21) *can be argued as follows: the search can be restricted to a ball in* $\mathbb{U}$. *Since* $\mathbb{U}$ *is finite dimensional this ball is compact. Since* $F(u) = \|v - u\|_{\mathbb{V}}$ *is a continuous function it attains its minimum. The uniqueness of the minimizer, by the equivalence of* (5.21) *and* (5.22), *follows from Remark 25, (1).*

Theorem 27

Let $(\mathbb{V}, \|\cdot\|_{\mathbb{V}})$ *be a Hilbert space and* $\mathbb{U} \subset \mathbb{V}$ *a finite dimensional subspace. For every* $v \in \mathbb{V}$ *there exists a unique* $u = u(v) \in \mathbb{U}$ *such that*

$$\langle v - u(v), w\rangle_{\mathbb{V}} = 0, \quad \forall\, w \in \mathbb{U}, \tag{5.24}$$

i.e., the difference $v - u(v)$ *is perpendicular to the subspace* $\mathbb{U}$. *Hence,* $P_{\mathbb{U}} : \mathbb{V} \to \mathbb{U}$ *given by* $P_{\mathbb{U}} v = u(v)$ *is well-defined and has the following properties:*

1. $P_{\mathbb{U}}$ *is linear and idempotent* $P_{\mathbb{U}} \circ P_{\mathbb{U}} = P_{\mathbb{U}}$, *i.e.,* $P_{\mathbb{U}}$ *is a projector.*
2. $P_{\mathbb{U}}$ *is self-adjoint, i.e.,* $\langle P_{\mathbb{U}} v, z\rangle_{\mathbb{V}} = \langle v, P_{\mathbb{U}} z\rangle_{\mathbb{V}}$, $v, z \in \mathbb{V}$.
3. $\|P_{\mathbb{U}}\|_{\mathcal{L}(\mathbb{V},\mathbb{V})} = 1$.

Proof of Theorem 27

$P_{\mathbb{U}}$ is well-defined: This follows from Remark 25, (2) and Remark 26.

(1): It is also linear because for all $w \in \mathbb{U}$ one has for all $w \in \mathbb{U}$

$$\begin{aligned}\langle P_{\mathbb{U}}(v_1+v_2)-(P_{\mathbb{U}}v_1+P_{\mathbb{U}}v_2), w\rangle_{\mathbb{V}} &= \langle P_{\mathbb{U}}(v_1+v_2)-(v_1+v_1), w\rangle_{\mathbb{V}} + \langle (v_1+v_1)\\ &\quad -(P_{\mathbb{U}}v_1+P_{\mathbb{U}}v_2), w\rangle_{\mathbb{V}}\\ &= 0+\langle v_1-P_{\mathbb{U}}v_1, w\rangle_{\mathbb{V}} + \langle v_2-P_{\mathbb{U}}v_2, w\rangle_{\mathbb{V}} = 0\end{aligned}$$

Since both $P_{\mathbb{U}}(v_1+v_2)$ and $P_{\mathbb{U}}v_1+P_{\mathbb{U}}v_2$ belong to $\mathbb{U}$, Remark 25, (1), implies $P_{\mathbb{U}}(v_1+v_2)=P_{\mathbb{U}}v_1+P_{\mathbb{U}}v_2$ In the same way one verifies $P_{\mathbb{U}}(\alpha v)=\alpha P_{\mathbb{U}}v$ which shows the first part of (1). Moreover, $0=\langle P_{\mathbb{U}}v-P_{\mathbb{U}}(P_{\mathbb{U}}v), w\rangle_{\mathbb{V}}$ which, by Remark 25, (1), again shows that $P_{\mathbb{U}}v=P_{\mathbb{U}}(P_{\mathbb{U}}v)$ which is (1).

(2): $0=\langle P_{\mathbb{U}}v, z-P_{\mathbb{U}}z\rangle_{\mathbb{V}} \rightsquigarrow \langle P_{\mathbb{U}}v, z\rangle_{\mathbb{V}} = \langle P_{\mathbb{U}}v, P_{\mathbb{U}}z\rangle_{\mathbb{V}}$. Likewise, $0=\langle v-P_{\mathbb{U}}v, P_{\mathbb{U}}z\rangle_{\mathbb{V}} \rightsquigarrow \langle v, P_{\mathbb{U}}z\rangle_{\mathbb{V}} = \langle P_{\mathbb{U}}v, P_{\mathbb{U}}z\rangle_{\mathbb{V}}$. Thus $\langle P_{\mathbb{U}}v, z\rangle_{\mathbb{V}} = \langle v, P_{\mathbb{U}}z\rangle_{\mathbb{V}}$.

(3): $\|P_{\mathbb{U}}v\|_{\mathbb{V}}^2 = \langle P_{\mathbb{U}}v, P_{\mathbb{U}}v\rangle_{\mathbb{V}} \overset{(1),(2)}{=} \langle v, P_{\mathbb{U}}v\rangle_{\mathbb{V}} \overset{CS}{\leq} \|P_{\mathbb{U}}v\|_{\mathbb{V}}\|v\|_{\mathbb{V}}$. $\Rightarrow \|P_{\mathbb{U}}\|_{\mathcal{L}(\mathbb{V},\mathbb{V})} \leq 1$,
$\|P_{\mathbb{U}}u\|_{\mathbb{U}} \overset{(1)}{=} \|u\|_{\mathbb{U}} \Rightarrow$ (3). □

Pythagoras' Theorem:

$$\begin{aligned}\|v-P_{\mathbb{U}}v\|_{\mathbb{V}}^2 &= \langle v-P_{\mathbb{U}}v, v-P_{\mathbb{U}}v\rangle_{\mathbb{V}} = \langle v, v\rangle_{\mathbb{V}} + \langle P_{\mathbb{U}}v, P_{\mathbb{U}}v\rangle_{\mathbb{V}} - \langle P_{\mathbb{U}}v, v\rangle_{\mathbb{V}} - \langle v, P_{\mathbb{U}}v\rangle_{\mathbb{V}}\\ &= \|v\|_{\mathbb{V}}^2 + \|P_{\mathbb{U}}v\|_{\mathbb{V}}^2 - 2\|P_{\mathbb{U}}v\|_{\mathbb{V}}^2 = \|v\|_{\mathbb{V}}^2 - \|P_{\mathbb{U}}v\|_{\mathbb{V}}^2\end{aligned}$$

Computing Orthogonal Projections/Best Approximations

Suppose $\Phi = \{\phi_1, \ldots, \phi_n\}$ is a basis for $\mathbb{U}$. Since

$$\langle v - P_{\mathbb{U}} v, u\rangle_{\mathbb{V}} = 0, \ \forall\, u \in \mathbb{U} \quad \Leftrightarrow \quad \langle v - P_{\mathbb{U}} v, \phi_k\rangle_{\mathbb{V}} = 0, \quad \text{for } k = 1, \ldots, n,$$

substituting $P_{\mathbb{U}} v = \sum_{j=1}^{n} u_j \phi_j$, yields (see (4.1))

$$\sum_{j=1}^{n} u_j \langle \phi_j, \phi_k\rangle_{\mathbb{V}} = \langle v, \phi_k\rangle_{\mathbb{V}} =: b_k, \quad k = 1, \ldots, n, \quad \Leftrightarrow \quad \mathbf{G}_\Phi \mathbf{u} = \mathbf{b} \tag{5.25}$$

where $\mathbf{G}_\Phi = \big(\langle \phi_k, \phi_j\rangle_{\mathbb{V}}\big)_{j,k=1}^{n}$ is the Gramian matrix associated with Φ. We already know that $\mathbf{G}_\Phi$ hermitian positive definite and hence non-singular, so that the computing $P_{\mathbb{U}} v$ amounts to solving a linear system of equations for the unknown expansion coefficient vector $\mathbf{u} = (u_1, \ldots, u_n)^\top$.

Orthonormal Bases: The orthogonal projection is very easy to compute if the basis Φ is orthonormal, i.e., $\langle \phi_j, \phi_k\rangle_{\mathbb{V}} = \delta_{j,k}$, $j, k = 1, \ldots, n$. In fact, then $\mathbf{G}_\Phi = \mathbf{I}$ and

$$u_j = \langle v, \phi_j\rangle_{\mathbb{V}}, \quad j = 1, \ldots, n. \tag{5.26}$$

Example: $\mathbb{V} = L_{2,\pi}$, show that the $e_k(x) = e^{ikx}$, $k \in \mathbb{Z}$, form an orthonormal system respect to $\langle f, g\rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x)\overline{g(x)}dx$, so that orthogonal projection to the space spanned by the $2n+1$ first harmonics e_k, $-n \le k \le n$, is the Fourier partial sum

$$S_n(f; x) = \sum_{|k|\le n} \hat{f}(k) e^{ikx}, \quad \hat{f}(k) = \langle f, e_k\rangle, \quad |k| \le n.$$

Gram-Schmidt Orthogonalization

Let $\mathbb{U}$ be an n-dimensional linear space with scalar product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \| = \langle \cdot, \cdot \rangle^{1/2}$. Given a basis $\{\phi_1, \ldots, \phi_n\}$ of $\mathbb{U}$ one can always generate an orthonormal basis $\{\psi_1, \ldots, \psi_n\}$ as follows:

- normalize $\psi_1 := \frac{\phi_1}{\|\phi_1\|}$;
- while $k < n$, given an orthonormal basis $\{\psi_1, \ldots, \psi_k\}$ for $\operatorname{span}\{\phi_1, \ldots, \phi_k\}$, let

$$\psi'_{k+1} := \phi_{k+1} - \sum_{j=1}^{k} \langle \phi_{k+1}, \psi_j \rangle \psi_j$$

and normalize

$$\psi_{k+1} := \frac{\psi'_{k+1}}{\|\psi'_{k+1}\|}.$$

Show that $\{\psi_1, \ldots \psi_n\}$, generated in this fashion, is indeed an orthonormal basis of $\mathbb{U}$.

Orthonormal Bases Continued

Remark 28

An orthonormal basis is "optimally stable" in the sense of (5.13)*, i.e.,*

$$\kappa_{\mathbb{U},\ell_2^n}(B_\Psi) = 1. \tag{5.27}$$

In fact, for any $u = \sum_{j=1}^n u_j\psi_j \in \mathbb{U}$, $\Psi = \{\psi_1, \dots, \psi_n\}$ *orthonormal, one has*

$$\|u\|^2 = \langle u, u\rangle = \left\langle \sum_{j=1}^n u_j\psi_j, \sum_{k=1}^n u_k\psi_k \right\rangle = \mathbf{u}^*\mathbf{G}_\Psi\mathbf{u} = \mathbf{u}^*\mathbf{u} = \|\mathbf{u}\|_2^2$$

which implies $\|B_\Psi^{-1}(\mathbf{u})\| = \|B_\Psi(u)\|_2 = 1$.

Consider the monomial basis $\phi_j(x) := x^j$, $j = 0, \dots, n$ of $\mathbb{U} := \mathbb{P}_n$ equipped with the scalar product $\langle f, g\rangle := \int_0^1 f(x)g(x)dx$ ($\mathbb{K} = \mathbb{R}$). What can you say about the condition number of B_Φ in this case. Intuitively, it is very large for large n. What does this mean about the Gram-Schmidt process turning Φ into an orthonormal basis Ψ for $\mathbb{P}_n$? Since this is a change of bases it can be represented by a matrix whose condition describes the stability of the change of bases.

Eigenvectors, Eigenvalues

Let $\mathbb{X}$ be a $\mathbb{K}$-vector space. Then $\lambda \in \mathbb{K}$, $x \in \mathbb{X}$ are called eigenvalue, resp. (right-)eigenvector of $L \in \mathcal{L}(\mathbb{X}, \mathbb{X})$ if

$$Lx = \lambda x \qquad \text{We also say } (\lambda, x) \text{ is an eigenpair of } L. \text{ Also } (\lambda^{-1}, \mathbf{x}) \text{ is an eigenpair of } L^{-1} \text{ if it exists} \tag{6.1}$$

Proposition 29

When $\mathbb{X}$ is finite dimensional there always exists an eigenpair $(\lambda, x) \in \mathbb{C} \times \mathbb{C}^n$. x is determined only up to normalization.

Proof: Fix bases Φ, Ψ for $\mathbb{X}, \mathbb{Y}$ as before. Then, by Proposition 7 and (3.2) we have

$$L(x) = \lambda x \;\Leftrightarrow\; \mathbf{C}\mathbf{x} = \lambda \mathbf{x} \;\Leftrightarrow\; (\mathbf{C} - \lambda \mathbf{I})\mathbf{x} = \mathbf{0}, \;\Leftrightarrow\; \det(\mathbf{C} - \lambda \mathbf{I}) = 0. \tag{6.2}$$

By Laplace's expansion rule of determinants, it follows that is a polynomial in $\mathbb{P}_n(\mathbb{K})$ of degree at most n. This is called the characteristic polynomial. By the Fundamental Theorem of Algebra, any polynomial in $\mathbb{P}_n(\mathbb{K})$ has exactly n (possibly complex) roots (counting multiplicities). □

It therefore suffices (in the finite dimensional case) to understand eigenproblems for matrices $\boldsymbol{A} \in \mathbb{K}^{n \times n}$, $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$:

$$\boldsymbol{A}\mathbf{x} = \lambda \mathbf{x} \tag{6.3}$$

Theorem 30

For $\boldsymbol{A} \in \mathbb{C}^{n \times n}$ one has $\det(\boldsymbol{A}) = \prod_{k=1}^{n} \lambda_k$, where $\lambda_1, \ldots, \lambda_n \subset \mathbb{C}$ are the n eigenvalues of $\boldsymbol{A}$.

Spectral Theorem for Hermitian Matrices

Theorem 31

Let $\boldsymbol{A} \in \mathbb{K}^{n\times n}$ be hermitian, then:

1. *all eigenvalues of $\boldsymbol{A}$ are real;*
2. *eigenvectors corresponding to distinct eigenvalues are pairwise orthogonal;*
3. *there exists an orthonormal basis of $\mathbb{K}^n$, consisting of eigenvectors.*

Hence, $\boldsymbol{A}$ is diagonalizable, i.e., there exists a unitary matrix $\mathbf{U} \in \mathcal{O}_n(\mathbb{K})$ such that

$$\mathbf{U}^* \boldsymbol{A} \mathbf{U} = \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n), \quad \lambda_j \in \mathbb{R},\ j = 1, \ldots, n.$$

Proof: (1): suppose $\boldsymbol{A}\mathbf{x} = \lambda\mathbf{x} \Rightarrow \mathbf{x}^*\boldsymbol{A}\mathbf{x} = \lambda\mathbf{x}^*\mathbf{x}$. Since $\mathbf{x}^*\mathbf{x} > 0$ we also have $\overline{\lambda}\mathbf{x}^*\mathbf{x} = \overline{\mathbf{x}^*\boldsymbol{A}\mathbf{x}} = (\mathbf{x}^\top\overline{\boldsymbol{A}\mathbf{x}})^\top = \overline{\mathbf{x}}^\top\overline{\boldsymbol{A}}^\top\mathbf{x} = \mathbf{x}^*\boldsymbol{A}^*\mathbf{x} = \mathbf{x}^*\boldsymbol{A}\mathbf{x} = \lambda\mathbf{x}^*\mathbf{x} \Rightarrow \overline{\lambda} = \lambda$.

(2): Assume that $\boldsymbol{A}\mathbf{x} = \lambda\mathbf{x}$, $\boldsymbol{A}\mathbf{y} = \mu\mathbf{y}$ but $\mathbf{y}^*\mathbf{x} \neq 0$. Since $(\boldsymbol{A}^*)^{-1} = \boldsymbol{A}^{-1}$ one has $\mathbf{y}^*\mathbf{x} = \mathbf{y}^*\boldsymbol{A}^*\boldsymbol{A}^{-1}\mathbf{x} = \overline{\mu}\lambda^{-1}\mathbf{y}^*\mathbf{x} = \mu\lambda^{-1}\mathbf{y}^*\mathbf{x}$, where we have used (1). Since $\mathbf{y}^*\mathbf{x} \neq 0$, this can only holde when $\lambda = \mu$ which is a contradiction.

(3): By Proposition 29, there exists an eigenpair $(\lambda_1, \mathbf{u}^1)$ $(\|\mathbf{u}^1\|_2 = 1)$ for $\boldsymbol{A}$. Let

$$\mathbf{V}_1 := \{\mathbf{z} \in \mathbb{K}^n : \mathbf{z}^*\mathbf{u}^1 = \langle \mathbf{u}^1, \mathbf{z}\rangle_n = 0\}$$

be the orthogonal complement of span $\{\mathbf{u}^1\}$. Observe next that $\boldsymbol{A}$ maps $\mathbb{V}_1$ into itself.

Proof of (3) continued:

In fact, when $\mathbf{z} \in \mathbb{V}_1$

$$(\boldsymbol{A}\mathbf{z})^* \mathbf{u}^1 = \mathbf{z}^* \boldsymbol{A}^* \mathbf{u}^1 = \mathbf{z}^* \boldsymbol{A} \mathbf{u}^1 = \mathbf{z}^* \lambda \mathbf{u}^1 = \lambda \mathbf{z}^* \mathbf{u}^1 = 0,$$

i.e., $\mathbf{z} \in \mathbb{V}_1$ implies $\boldsymbol{A}\mathbf{z} \in \mathbb{V}_1$. Therefore, the operator $L(\mathbf{x}) := \boldsymbol{A}\mathbf{x}$ belongs to $\mathcal{L}(\mathbb{V}_1, \mathbb{V}_1)$. By Proposition 29 there exists an eigenpair $(\lambda_2, \mathbf{u}^2) \in \mathbb{C} \times \mathbb{V}_1$ and $\langle \mathbf{u}^1, \mathbf{u}^2 \rangle_n = 0$. Clearly $\dim \mathbb{V}_1 = n - 1$. Now, by the same reasoning, we consider the orthogonal complement $\mathbb{V}_2$ to $\mathbf{u}^2$ in $\mathbb{V}_1$ (which is therefore also orthogonal to $\mathbf{u}^1$) and find the next eigenpair $(\lambda_3, \mathbf{u}^3) \in \mathbb{V}_2$. Since each time the dimension of the subsequent orthogonal complement decreases by one, this process terminates after $n - 1$ steps. □

Remark 32

- *Not every matrix $\boldsymbol{A} \in \mathbb{K}^{n \times n}$ can be diagonalized (is similar to a diagonal matrix, meaning there exists a matrix $\mathbf{C}$ such that $\mathbf{C}^{-1}\boldsymbol{A}\mathbf{C} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$). It is diagonalizable iff there exists a basis of eigenvactors.*
- *For $\mathbb{K} = \mathbb{R}$ the eigenvectors are in $\mathbb{R}^n$, i.e., $\mathbf{U} \in \mathcal{O}_n(\mathbb{R})$ is an orthogonal matrix.*

Spectral Theorem for Unitary Matrices

Theorem 33

$\boldsymbol{Q} \in \mathcal{O}_n(\mathbb{K})$ unitary, then:

1. *all eigenvalues of $\boldsymbol{Q}$ have absolute value equal to one;*
2. *eigenvectors corresponding to distinct eigenvalues are pairwise orthogonal;*
3. *there exists an orthonormal basis of $\mathbb{K}^n$, consisting of eigenvectors.*

Hence, $\boldsymbol{Q}$ is diagonalizable, i.e., there exists a unitary matrix $\mathbf{U} \in \mathcal{O}_n(\mathbb{K})$ such that

$$\mathbf{U}^* \boldsymbol{Q} \mathbf{U} = \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n), \quad |\lambda_j| = 1,\ j = 1, \ldots, n.$$

Proof: (1): asume that $\boldsymbol{Q}\mathbf{x} = \lambda \mathbf{x}$, $\mathbf{x} \neq \mathbf{0}$, $\rightsquigarrow\ \overline{\lambda}\lambda \mathbf{x}^*\mathbf{x} = (\boldsymbol{Q}\mathbf{x})^* \boldsymbol{Q}\mathbf{x} = \mathbf{x}^* \boldsymbol{Q}^* \boldsymbol{Q}\mathbf{x} = \mathbf{x}^*\mathbf{x} \Rightarrow\ |\lambda|^2 = 1$.

(2): Assume that $\boldsymbol{Q}\mathbf{x} = \lambda \mathbf{x}$, $\boldsymbol{Q}\mathbf{y} = \mu \mathbf{y}$ but $\mathbf{y}^*\mathbf{x} \neq 0 \rightsquigarrow \mathbf{y}^*\mathbf{x} = \mathbf{y}^* \boldsymbol{Q}^* \boldsymbol{Q}\mathbf{x} = \overline{\mu}\lambda \mathbf{y}^*\mathbf{x} \Rightarrow \overline{\mu}\lambda = 1 \Rightarrow \lambda/\mu = 1$ which is a contradiction to $\mu \neq \lambda$.

(3): Let $\boldsymbol{A}\mathbf{x} = \lambda \mathbf{x}$ and $\mathbb{V}_1 := \{\mathbf{z} \in \mathbb{K}^n : \mathbf{z}^*\mathbf{x} = 0\}$. Now for any $\mathbf{z} \in \mathbb{V}_1$ one has $(\boldsymbol{Q}\mathbf{z})^*\mathbf{x} = \mathbf{z}^* \boldsymbol{Q}^*\mathbf{x} = \mathbf{z}^* \boldsymbol{Q}^{-1}\mathbf{x} = \lambda^{-1}\mathbf{z}^*\mathbf{x} = 0$. Hence, the orthogonal complement of any eigenvector is an invariant subspace of $\boldsymbol{Q}$, i.e., $\boldsymbol{Q}\mathbb{V}_1 \subseteq \mathbb{V}_1$. As in the proof of Theorem 31, one can therefore successively peel off invariant subspaces of decreasing dimension which are orthogonal to previously found eigenvectors. □

Application: Direct Solvers for Systems of Linear Equations

Common principle: given $\boldsymbol{A} \in \mathbb{K}^{n\times n}$ find a **factorization** $\boldsymbol{A} = \mathbf{CB}$, where both $\mathbf{B}, \mathbf{C} \in \mathbb{K}^{n\times n}$ are "easy to invert"

then, since $\boldsymbol{A}\mathbf{x} = \mathbf{C}\underbrace{(\mathbf{B}\mathbf{x})}_{=:\mathbf{y}} = \mathbf{b}$

$$\text{first solve} \quad \mathbf{C}\mathbf{y} = \mathbf{b} \;\rightarrow\; \mathbf{y}, \quad \text{then solve} \quad \mathbf{B}\mathbf{x} = \mathbf{y} \;\rightarrow\; \mathbf{x}$$

- One "difficult" solve is traded against two "easy solves"
- Such factorizations are typically generated in a step-wise fashion using, for instance Gauß elimination, in which case **C** is a product of "simple" lower triangular matrices, or rotations, in which case **C** is a unitary matrix successively built from products of "simple" unitary matrices.
- The factors belong to special matrix classes to be discussed below

Easy to invert matrices ...

- $\mathbf{I} = (\delta_{i,j})_{i,j=1}^{n,n}$ trivial $\mathbf{I}^{-1} = \mathbf{I}$
- Upper/lower triangular matrices:

$$\boldsymbol{R} = (r_{i,j})_{i,j=1}^{n,n}, \quad j < i \Rightarrow r_{i,j} = 0, \qquad \mathbf{L} = (\ell_{i,j})_{i,j=1}^{n,n}, \quad j > i \Rightarrow \ell_{i,j} = 0,$$

  $\boldsymbol{R}$ non-singular if and only if $\det(\boldsymbol{R}) = \prod_{j=1}^{n} r_{j,j} \neq 0$.

  Backsubstitution: $\boldsymbol{R}\mathbf{x} = \mathbf{b} \Rightarrow x_n = b_n / r_{n,n}$, knowing $x_n, x_{n-1}, \ldots, x_{n-j} \Rightarrow$

$$x_{n-j-1} = \left(b_{n-j-1} - \sum_{k=0}^{j} r_{n-j-1,n-k} x_{n-k} \right) / r_{n-j-1,n-j-1}$$

  Complexity: $\sim n^2$ flops
- same for $\mathbf{L}\mathbf{x} = \mathbf{b}$
- Orthogonal ($\mathbb{K} = \mathbb{R}$)/unitary ($\mathbb{K} = \mathbb{C}$) matrices: $\boldsymbol{Q} \in \mathcal{O}_n$ iff $\boldsymbol{Q}^* \boldsymbol{Q} = \mathbf{I}$. i.e

$$\boldsymbol{Q}\mathbf{x} = \mathbf{b} \quad \Leftrightarrow \quad \mathbf{x} = \boldsymbol{Q}^* \mathbf{b}$$

Properties of Triangular Matrices

- An upper triangular matrix $\boldsymbol{R} \in \mathbb{K}^{n\times n}$ is non-singular iff $r_{j,j} \neq 0$, $j = 1, \ldots, n$, and

$$\det(\boldsymbol{R}) = \prod_{j=1}^{n} r_{j,j}.$$

- The inverse of an upper (lower) triangular non-singular matrix is upper (lower) triangular.
- Products of upper (lower) triangular matrices (of the same dimension) are again upper (lower) triangular. Hence, such matrices form a multiplicative group with $\mathbf{I}$ as the neutral element.

Exercise 34

Prove the above statements

L*R* Factorization

Theorem 35

For every nonsingular $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ *there exists a permutation matrix* $\boldsymbol{P}$ *a normalized lower triangular matrix* $\mathbf{L}$*, i.e.,* $\ell_{j,j} = 1$*,* $j = 1, \ldots, n$*, and an upper triangular matrix* $\boldsymbol{R}$*, such that*

$$\boldsymbol{PA} = \mathbf{L}\boldsymbol{R}$$

- The construction of this factorization is a by-product of Gauß-elimination with pivoting, i.e., $\boldsymbol{A}$ is reduced to an upper triangular matrix $\boldsymbol{R}$ by columnwise elimination of all non-zero entries below the diagonal, starting with the first column. More precisely, at the jth step rows are exchanged so as to move the largest entry in column j to th diagonal position (multiplication by a permutation matrix $\boldsymbol{P}_j$, followed by an elimination step, realized by multiplication with a specific lower triangular normalized matrix $\mathbf{L}_j$ (Frobenius matrix). Thus

$$\boldsymbol{R} = \mathbf{L}_{n-1}\boldsymbol{P}_{n-1}\mathbf{L}_{n-2}\cdots\mathbf{L}_1\boldsymbol{P}_1\boldsymbol{A} = \underbrace{\tilde{\mathbf{L}}_{n-1}\cdots\tilde{\mathbf{L}}_1}_{=:\mathbf{L}^{-1}}\underbrace{\boldsymbol{P}_{n-1}\cdots\boldsymbol{P}_1}_{=:\boldsymbol{P}}\boldsymbol{A}$$

- The computational cost is $\sim n^3/3$ multiplications

Exercise 36

What is the complexity of computing $\boldsymbol{A}^{-1}$*?*

What is the complexity of computing $\det(\boldsymbol{A})$*?*

QR Factorization

Theorem 37

For every $\boldsymbol{A} \in \mathbf{K}^{m\times n}$ there exist $\boldsymbol{Q} \in \mathcal{O}_m$ and an upper triangular $\boldsymbol{R} \in \mathbb{K}^{m\times n}$ (with the natural interpretation when $m \neq n$) such that

$$\boldsymbol{A} = \boldsymbol{QR}$$

Remark 38

There is neither any restriction on the dimension $m \times n$ nor on ranks.

The factorization is computed by repeated multiplication by special "elementary" unitary matrices so as to successively eliminate sub-triangular entries:

$$\boldsymbol{Q}_{n-1}\boldsymbol{Q}_{n-2}\cdots\boldsymbol{Q}_1\boldsymbol{A} = \boldsymbol{R} \quad \Leftrightarrow \quad \underbrace{\boldsymbol{Q}_1^*\boldsymbol{Q}_2^*\cdots\boldsymbol{Q}_{n-2}^*\boldsymbol{Q}_{n-1}^*}_{=:\boldsymbol{Q}}\boldsymbol{R} = \boldsymbol{A}$$

An ingredient: dyades - rank-one matrices: $\mathbf{x} \in \mathbb{K}^m, \mathbf{y} \in \mathbb{K}^n \rightsquigarrow$

$$\mathbf{x}\mathbf{y}^* = \left(\bar{y}_1\mathbf{x}, \ldots, \bar{y}_n\mathbf{x}\right) \in \mathbb{K}^{m\times n}$$

Hence, for $\mathbf{z} \in \mathbb{K}^n$ one has $(\mathbf{x}\mathbf{y}^*)\mathbf{z} = \mathbf{x}(\mathbf{y}^*\mathbf{z}) = \langle \mathbf{z}, \mathbf{y}\rangle_n\mathbf{x}$, i.e., $\mathrm{rank}(\mathbf{x}\mathbf{y}^*) = 1$

Householder-Reflections

Given $\mathbf{a} \in \mathbb{K}^n$ let

$$\boldsymbol{Q}_{\mathbf{v}} := \mathbf{I} - \frac{2}{\mathbf{v}^*\mathbf{v}}\mathbf{v}\mathbf{v}^*$$

The following facts can be verified by corresponding calculations:

1. $\boldsymbol{Q}_{\mathbf{v}}^* = \boldsymbol{Q}_{\mathbf{v}}$
2. $\boldsymbol{Q}_{\mathbf{v}}^2 = \mathbf{I}$, i.e., $\boldsymbol{Q}_{\mathbf{v}} \in \mathcal{O}_n$
3. $\boldsymbol{Q}_{\mathbf{v}} = \boldsymbol{Q}_{\alpha\mathbf{v}}$ for any $\alpha \in \mathbb{K} \setminus \{0\}$
4. $\boldsymbol{Q}_{\mathbf{v}}\mathbf{v} = -\mathbf{v}$; interprete this geometrically in terms of the hyperplane

$$H_{\mathbf{v}} := \{\mathbf{x} \in \mathbb{K}^n : \mathbf{v}^*\mathbf{x} = 0\}$$

Claim:

Given $\mathbf{a} \in \mathbb{K}^n$, take $\mathbf{v}(\mathbf{a}) = \mathbf{v} := \mathbf{a} + \operatorname{sgn}(a_1)\|\mathbf{a}\|_2\mathbf{e}^1$, $\mathbf{e}^j := (0, \ldots, 0, 1, 0, \ldots, 0)^\top$. Then

$$\boldsymbol{Q}_{\mathbf{v}}\mathbf{a} = -\operatorname{sgn}(a_1)(\|\mathbf{a}\|_2\mathbf{e}^1,$$

i.e., $Q_{\mathbf{v}}$ rotates $\mathbf{a}$ into a multiple of the first coordinate vector. The particular sign is chosen so as to avoid numerical cancellation in the first component of $\mathbf{a} + \operatorname{sgn}(a_1)\|\mathbf{a}\|_2\mathbf{e}^1$

QR-Factorization

Recall:

$$\mathbf{v}(\mathbf{a}) = \mathbf{v} := \mathbf{a} + \operatorname{sgn}(a_1)\|\mathbf{a}\|_2 \mathbf{e}^1 \tag{6.4}$$

- Given $\boldsymbol{A} = (\mathbf{a}^1, \mathbf{a}^2, \ldots, \mathbf{a}^n)$, $\mathbf{a}^j \in \mathbb{K}^m$, choose $\mathbf{v}^1 := \mathbf{v}(\mathbf{a}^1) \rightsquigarrow$

$$\boldsymbol{Q}_1 := \boldsymbol{Q}_{\mathbf{v}^1}\boldsymbol{A} = \begin{pmatrix} \tilde{a}_{1,1} & 0 \\ 0 & \tilde{\boldsymbol{A}}_1 \end{pmatrix} =: \boldsymbol{A}^1, \quad \tilde{a}_{1,1} = a_{1,1} + \operatorname{sgn}(a_{1,1})\|\mathbf{a}^1\|_2, \quad \tilde{\boldsymbol{A}}^1 \in \mathbb{K}^{(m-1)\times(n-1)}$$

- Given $\boldsymbol{A}_j$, let $\tilde{\boldsymbol{Q}}_{j+1} = \boldsymbol{Q}_{\mathbf{v}((\tilde{\boldsymbol{A}}_j)^1)} \in \mathcal{O}_{n-j}$ where $(\tilde{\boldsymbol{A}}_j)^1$ is the first column in $\tilde{\boldsymbol{A}}_j \in \mathbb{K}^{(m-j)\times(n-j)}$ and set

$$\boldsymbol{Q}_{j+1} := \begin{pmatrix} \mathbf{I}_j & \mathbf{0} \\ \mathbf{0} & \tilde{\boldsymbol{Q}}_{j+1} \end{pmatrix} \in \mathcal{O}_m \quad \rightsquigarrow \boldsymbol{Q}_{j+1}\boldsymbol{A}_j = \boldsymbol{A}_{j+1} \qquad (\mathbf{I}_j \text{ is the } j \times j \text{ identity matrix})$$

Then $\boldsymbol{A}_n =: \boldsymbol{R}$ is upper triangular and

$$\boldsymbol{Q}_{n-1} \cdots \boldsymbol{Q}_1 \boldsymbol{A} = \boldsymbol{R}, \quad \rightsquigarrow \quad \boldsymbol{A} = \boldsymbol{Q}_1^* \cdots \boldsymbol{Q}_{n-1}^* \boldsymbol{R} = \boldsymbol{Q}\boldsymbol{R}$$

Orthonormalization in $\mathbb{K}^n$

Suppose that $\{\mathbf{b}^1, \ldots, \mathbf{b}^n\}$ is a basis for $\mathbb{K}^n$. One could generate an orthonormal basis for $\mathbb{K}^n$ by applying Gram-Schmidt orthogonalization. Alternatively, let $\mathbf{B}$ be the matrix with columns $\mathbf{b}^j$ and compute a QR factorization

$$\mathbf{B} = \boldsymbol{QR} \quad \Leftrightarrow \quad \boldsymbol{Q} = \mathbf{B}\boldsymbol{R}^{-1},$$

i.e., the columns $\mathbf{q}^j$ of $\boldsymbol{Q}$ form an orthonormal basis of $\mathbb{K}^n$ as well. Recall that $\kappa_2(\mathbf{B}) = \kappa_2(\boldsymbol{R})$ (since $\boldsymbol{Q}$ is unitary). But $\boldsymbol{Q}$ can be computed, without inverting $\boldsymbol{R}$, by a successive application of Householder reflections, each being unitary and numerically stable. This process being stable the interrelation between the basis $\{\mathbf{q}^1, \ldots, \mathbf{q}^n\}$ and the basis $\{\mathbf{b}^1, \ldots, \mathbf{b}^n\}$ is given by the transformation $\boldsymbol{R}^{-1}$ whose condition number equals the one of $\mathbf{B}$. Thus if $\{\mathbf{b}^1, \ldots, \mathbf{b}^n\}$ has poor stability properties $\boldsymbol{R}^{-1}$ is poorly conditioned. This is, however, not used in the computation of $\boldsymbol{Q}$!

Compare this with the Gram-Schmitd process!

Definition and Existence

The following factorization has a multitude of applications in data science and machine learning. Unlike the spectral decompositions or *LR*-factorization it holds for matrices of any dimension (as the *QR*-factorization)

Theorem 39

For every matrix $\boldsymbol{A} \in \mathbb{K}^{m\times n}$ *there exist unitary matrices* $\mathbf{U} \in \mathcal{O}_m$, $\mathbf{V} \in \mathcal{O}_n$ *and a diagonal matrix* $\mathbf{S} \in \mathbb{R}^{\mathbf{m}\times\mathbf{n}}$

$$\mathbf{S} = \operatorname{diag}(\sigma_{\mathbf{1}}, \ldots, \sigma_{\mathbf{p}}, \mathbf{0}, \ldots, \mathbf{0}), \quad \mathbf{p} := \min\{\mathbf{m}, \mathbf{n}\},$$

with

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0, \tag{6.5}$$

so that

$$\mathbf{U}^* \boldsymbol{A} \mathbf{V} = \mathbf{S} \quad \Leftrightarrow \quad \boldsymbol{A} = \mathbf{U}\mathbf{S}\mathbf{V}^* \tag{6.6}$$

The second relation in (6.6) can be written as

$$\boldsymbol{A} = \sum_{k=1}^{p} \sigma_k \mathbf{u}^k (\mathbf{v}^k)^* \tag{6.7}$$

where $\mathbf{u}^k, \mathbf{v}^k$ are the columns of $\mathbf{U}, \mathbf{V}$, respectively.

Proof of Theorem 39: Assume $\boldsymbol{A} \neq \mathbf{0}$, let

$$\sigma_1 := \|\boldsymbol{A}\|_2 = \max_{\|\mathbf{x}\|_2=1} \|\boldsymbol{A}\mathbf{x}\|_2 > 0.$$

Let $\mathbf{v} \in \mathbb{K}^n$, with $\|\mathbf{v}\|_2 = 1$ a vector satisfying $\|\boldsymbol{A}\|_2 = \|\boldsymbol{A}\mathbf{v}\|_2$ and $\mathbf{u} := \frac{1}{\sigma_1}\boldsymbol{A}\mathbf{v} \in \mathbb{K}^m$. Then one has for $\mathbf{u}$ that $\|\mathbf{u}\|_2 = \|\boldsymbol{A}\mathbf{v}\|_2/\sigma_1 = 1$. We can extend the vectors $\mathbf{v}$ und $\mathbf{u}$ to orthogonormal bases $\{\mathbf{v}, \tilde{\mathbf{v}}_2, \ldots, \tilde{\mathbf{v}}_n\}$ resp. $\{\mathbf{u}, \tilde{\mathbf{u}}_2, \ldots, \tilde{\mathbf{u}}_m\}$ of $\mathbb{K}^n$ resp. $\mathbb{K}^m$. We view the elements of these bases as columns of corresponding unitary matrices $\mathbf{V}_1 \in \mathcal{O}_n(\mathbb{K})$, $\mathbf{U}_1 \in \mathcal{O}_m(\mathbb{K})$:

$$\begin{aligned} \mathbf{V}_1 &= \begin{pmatrix} \mathbf{v} & \tilde{\mathbf{V}}_1 \end{pmatrix} \in \mathbb{K}^{n\times n}, \qquad \text{unitary,} \\ \mathbf{U}_1 &= \begin{pmatrix} \mathbf{u} & \tilde{\mathbf{U}}_1 \end{pmatrix} \in \mathbb{K}^{m\times m}, \qquad \text{unitary .} \end{aligned}$$

Since $\tilde{\mathbf{u}}_i^* \boldsymbol{A}\mathbf{v} = \sigma_1 \tilde{\mathbf{u}}_i^* \mathbf{u} = 0$, $i = 2, \ldots, m$ the matrix $\mathbf{U}_1^* \boldsymbol{A}\mathbf{V}_1$ has the form

$$\boldsymbol{A}_1 := \mathbf{U}_1^* \boldsymbol{A}\mathbf{V}_1 = \begin{pmatrix} \sigma_1 & \mathbf{w}^* \\ \mathbf{0} & \mathbf{B} \end{pmatrix} \in \mathbb{K}^{m\times n},$$

with $\mathbf{w} \in \mathbb{K}^{n-1}$. From

$$\left\| A_1 \begin{pmatrix} \sigma_1 \\ \mathbf{w} \end{pmatrix} \right\|_2 = \left\| \begin{pmatrix} \sigma_1^2 + \mathbf{w}^* w \\ \mathbf{B}\mathbf{w} \end{pmatrix} \right\|_2 \geq \sigma_1^2 + \mathbf{w}^*\mathbf{w} = \left\| \begin{pmatrix} \sigma_1 \\ \mathbf{w} \end{pmatrix} \right\|_2^2$$

and $\|A\|_2 = \|A_1\|_2$ it follows that

$$\sigma_1 = \|A_1\|_2 \geq \frac{\|A_1 \begin{pmatrix} \sigma_1 \\ \mathbf{w} \end{pmatrix}\|_2}{\|\begin{pmatrix} \sigma_1 \\ \mathbf{w} \end{pmatrix}\|_2} \geq \sqrt{\sigma_1^2 + \mathbf{w}^*\mathbf{w}} \;\Rightarrow\; \mathbf{w} = 0 \;\Rightarrow\; \mathbf{U}_1^* \boldsymbol{A}\mathbf{V}_1 = \begin{pmatrix} \sigma_1 & \mathbf{0} \\ \mathbf{0}^* & \mathbf{B} \end{pmatrix} \in \mathbb{K}^{m\times n}.$$

This proves the assertion for $m = 1$ or $n = 1$.

For $m, n > 1$ one can use induction. Suppose that $\mathbf{U}_2^* \mathbf{B} \mathbf{V}_2 = \mathbf{S_2}$ with $\mathbf{U}_2 \in \mathcal{O}_{m-1}(\mathbb{K})$, $\mathbf{V}_2 \in \mathcal{O}_{n-1}(\mathbb{K})$ and $\mathbf{S_2} \in \mathbb{R}_+^{(\mathbf{m-1})\times(\mathbf{n-1})}$ is diagonal. For the largest diagonal entry σ_2 of $\mathbf{S_2}$ we have again $\sigma_2 := \|B\|_2 \leq \|\mathbf{U}_1^* \boldsymbol{A} \mathbf{V}_1\|_2 = \|A\|_2 = \sigma_1$. Furthermore, with the unitary matrices

$$\mathbf{U} = \mathbf{U}_1 \begin{pmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & \mathbf{U}_2 \end{pmatrix}, \; \mathbf{V} = \mathbf{V}_1 \begin{pmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & \mathbf{V}_2 \end{pmatrix}$$

we obtain the decomposition

$$\mathbf{U}^* \boldsymbol{A} \mathbf{V} = \begin{pmatrix} \sigma_1 & \mathbf{0}^* \\ \mathbf{0} & \mathbf{S_2} \end{pmatrix}$$

from which the assertion follows by induction. □

Basic Properties

Proposition 40

For $\boldsymbol{A} \in \mathbb{K}^{m\times n}$, $\mathbf{U}^*\boldsymbol{A}\mathbf{v} = \mathbf{S}$ as before with

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_p = 0, \quad p = \min\{m, n\},$$

one has:

1. $\boldsymbol{A}\mathbf{v}^j = \sigma_j \mathbf{u}^j, \quad \boldsymbol{A}^*\mathbf{u}^j = \sigma_j \mathbf{v}^j, j = 1, \ldots, p.$
2. $\operatorname{rank}(\boldsymbol{A}) = r,$
3. $\operatorname{range}(\boldsymbol{A}) = \operatorname{span}\{\mathbf{u}^1, \ldots, \mathbf{u}^r\}$, $\ker(\boldsymbol{A}) = \operatorname{span}\{\mathbf{v}^{r+1}, \ldots, \mathbf{v}^n\}.$
4. $\|\boldsymbol{A}\|_2 = \sigma_1.$
5. *The strictly positive singular values $\sigma_k, k \leq r$ are the square roots of the (strictly positive) eigenvalues of $\boldsymbol{A}^*\boldsymbol{A}$:*
$$\sigma_j = \sqrt{\lambda_j(\boldsymbol{A}^*\boldsymbol{A})}, \quad j = 1, \ldots, r.$$

(1) - (4) follow immediately from $\boldsymbol{A} = \mathbf{USV}^*$ (and its hermitian conjugate) and $\|\boldsymbol{A}\|_2 = \|\mathbf{S}\|_2 = \sigma_1$. Since $\boldsymbol{A}^*\boldsymbol{A} = \mathbf{VS}^*\mathbf{U}^*\mathbf{USV}^* = \mathbf{V}\operatorname{diag}(\sigma_1^2, \ldots, \sigma_r^2, \mathbf{0}, \ldots, \mathbf{0})\mathbf{V}^*$, we have determined the spectral decomposition of $\boldsymbol{A}^*\boldsymbol{A}$ which confirms (5).

More Properties and the Pseudo-Inverse of $\boldsymbol{A}$

- Likewise one could have argued $\boldsymbol{A}\boldsymbol{A}^* = \mathbf{U}\mathbf{S}\mathbf{S}^*\mathbf{U}^*$ which reveals the roles of $\mathbf{U}, \mathbf{V}$ regarding the eigenspaces of $\boldsymbol{A}\boldsymbol{A}^*, \boldsymbol{A}^*\boldsymbol{A}$, respectively.
- Denoting by $\mathbf{U}_r, \mathbf{V}_r$ ($r = \mathrm{rank}(\boldsymbol{A})$ as above) the matrices formed by the first r columns of $\mathbf{U}, \mathbf{V}$, respectively, which refines (6.7)

$$\boldsymbol{A} = \sum_{k=1}^{r} \sigma_k \mathbf{u}^j (\mathbf{v}^j)^*. \tag{6.8}$$

  If $r \ll p = \min\{m, n\}$, computing first $c_j := (\mathbf{v}^j)^*\mathbf{x}$, $j = 1, \ldots, r$ and then summing $\sum_{k=1}^{r} \sigma_k c_k \mathbf{u}^k$, requires roughly $r(n+m)$ operations as opposed to the order of mn operations when applying $\boldsymbol{A}$ to $\mathbf{x}$ directly.
- For $\mathbf{S}$ as above define $\mathbf{S}^\dagger := \mathrm{diag}(\sigma_1^{-1}, \ldots, \sigma_r^{-1}, \mathbf{0}, \ldots, \mathbf{0}) \in \mathbb{R}^{\mathbf{n}\times\mathbf{m}}$ (the dimension of the transpose with inverses of the positive singular values on the diagonal). One then defines the pseudo-inverse of $\boldsymbol{A}$ as:

$$\boldsymbol{A}^\dagger := \mathbf{V}\mathbf{S}^\dagger\mathbf{U}^* \in \mathbb{K}^{\mathbf{n}\times\mathbf{m}}. \tag{6.9}$$

  One checks that

$$\boldsymbol{A}^\dagger = \begin{cases} \boldsymbol{A}^{-1} & \text{if} \quad m = n \text{ and } \boldsymbol{A} \text{ is non-singular,} \\ (\boldsymbol{A}^*\boldsymbol{A})^{-1}\boldsymbol{A}^* & \text{if} \quad m > n \text{ and } \mathrm{rank}(\boldsymbol{A}) = n. \end{cases}$$

An Application of the Pseudo-Inverse Least Squares

- $\boldsymbol{A} \in \mathbb{K}^{n\times n}$ non-singular $\rightsquigarrow$ $\boldsymbol{A}\mathbf{x} = \mathbf{b}$ has a unique solution $\mathbf{x} = \boldsymbol{A}^{-1}\mathbf{b}$
  $\mathbf{b} \to \mathbf{x}$ is a linear process;
- $\boldsymbol{A} \in \mathbb{K}^{m\times n}$, $m > n$, it makes no longer sense to pose $\boldsymbol{A}\mathbf{x} = \mathbf{b}$. Instead: find

$$\hat{\mathbf{x}} \in \underset{\mathbf{x}\in\mathbb{K}^n}{\operatorname{argmin}} \|\boldsymbol{A}\mathbf{x} - \mathbf{b}\|_2 \tag{6.10}$$

Comment: one could take as well any other norm but, as shown later, the Euclidean norm has significant advantages and comes with a favorable statistical interpretation

- $\boldsymbol{A}$ has full rank n $\rightsquigarrow$ (6.10) has the unique solution $\hat{\mathbf{x}} = (\boldsymbol{A}^*\boldsymbol{A})^{-1}\boldsymbol{A}^*\mathbf{b}$
  again $\mathbf{b} \to \hat{\mathbf{x}}$ is a linear process;
- $\operatorname{rank}(\boldsymbol{A}) < \min\{n, m\} \rightsquigarrow$ (6.10) has infinitely many solutions, but $\rightsquigarrow$ $\hat{\mathbf{x}} = \boldsymbol{A}^\dagger\mathbf{b}$ is the unique minimizer of minimal Euclidean norm

See Theorem 45 later below for more details.

Remark 41

Finding $\hat{\mathbf{x}}$ in (6.10) is called a Least Squares problem (minimizing the squares of residual components). It plays an eminent role for the design of estimators in machine learning, especially regression. It can be seen as generalizing the solution of linear systems which is well-posed only if one has as many equations (conditions) as unknowns. Least squares methods will therefore be discussed in more detail later.

Best Low-Rank Approximation

Many applications of the SVD, used later, are based on the following fact:

Theorem 42

Let $\boldsymbol{A} = \mathbf{USV}^*$ *and* $\operatorname{rank}(\boldsymbol{A}) = r \leq p := \min\{m, n\}$. *Defining the truncated matrices* $\mathbf{U}_k, \mathbf{V}_k$ *as before, one has for* $k \leq p$

$$\|\boldsymbol{A} - \mathbf{U}_k\mathbf{S}_\mathbf{k}\mathbf{V}_\mathbf{k}^*\|_\mathbf{2} = \min\{\|\boldsymbol{A} - \mathbf{B}\|_\mathbf{2} : \mathbf{B} \in \mathbb{K}^{\mathbf{m}\times\mathbf{n}}, \operatorname{rank}(\mathbf{B}) \leq \mathbf{k}\} = \sigma_{\mathbf{k+1}}. \tag{6.11}$$

Moreover, defining the Frobenius norm $\|\boldsymbol{A}\|_F := \left(\sum_{j,k=1}^{m,n} |a_{j,k}|^2\right)^{1/2} = (\operatorname{trace}(\boldsymbol{A}^*\boldsymbol{A}))^{1/2}$, *one has*

$$\|\boldsymbol{A} - \mathbf{U}_k\mathbf{S}_\mathbf{k}\mathbf{V}_\mathbf{k}^*\|_\mathbf{F}^\mathbf{2} = \sum_{\mathbf{j}>\mathbf{k}} \sigma_\mathbf{j}^\mathbf{2}. \tag{6.12}$$

Proof: Since the spectral norm of a diagonal matrix is the maximal diagonal entry in absolute value, one has $\|\boldsymbol{A} - \mathbf{U}_k\mathbf{S}_\mathbf{k}\mathbf{V}_\mathbf{k}^*\|_\mathbf{2} = \|\mathbf{U}(\mathbf{S} - \mathbf{S}_\mathbf{k})\mathbf{V}^*\|_\mathbf{2} = \|\mathbf{S} - \mathbf{S}_\mathbf{k}\|_\mathbf{2} = \sigma_{\mathbf{k+1}}$. It remains to show that for every matrix $\mathbf{B}$ of rank k one has $\|\mathbf{B} - \boldsymbol{A}\|_2 \geq \sigma_{k+1}$. Now suppose $\mathbf{B} \in \mathbb{K}^{m\times n}$ has rank $\leq k$. Then $\dim(\ker(\mathbf{B})) \geq n - k$. Hence $\mathbb{U} := \ker(\mathbf{B}) \cap \operatorname{span}\{\mathbf{v}^1, \ldots, \mathbf{v}^{k+1}\} \neq \{\mathbf{0}\}$. Let $\mathbf{z} = \sum_{j=1}^{k+1} \alpha_j \mathbf{v}^j \in \mathbb{U}$, $\sum_{j=1}^{k+1} |\alpha|^2 = 1$. Then

$$\|\boldsymbol{A} - \mathbf{B}\|_2^2 \geq \|(\boldsymbol{A} - \mathbf{B})\mathbf{z}\|_2^2 = \|\boldsymbol{A}\mathbf{z}\|_2^2 \overset{Prop.40,(1)}{=} \sum_{j=1}^{k+1} \sigma_j^2|\alpha_j|^2 \geq \sigma_{k+1}\sum_{j=1}^{k+1} |\alpha_j|^2 = \sigma_{k+1}. \qquad \square$$

A Greedy Characterization — Principal Component Analysis (PCA)

Given a point cloud $\{\mathbf{a}^j : j = 1, \ldots, n\} \subset \mathbb{R}^m$, find its best simultaneous approximation by a line $L_1 := \{\mathbf{x} = t\mathbf{u} : \mathbf{u} \in \mathbb{R}^m, \|\mathbf{u}\|_2 = 1, t \in \mathbb{R}\}$, i.e.,

$$\mathbf{u}^1 = \operatorname*{argmin}_{\substack{\mathbf{u}\in\mathbb{R}^m \\ \|\mathbf{u}\|_2=1}} \sum_{i=1}^{n} \|\mathbf{a}^i - (\mathbf{u}^*\mathbf{a}^i)\mathbf{u}\|_2^2. \tag{6.13}$$

Lemma 43

The unit vector $\mathbf{u}^1$*, defined by* (6.14) *is characterized as follows*

$$\mathbf{u}^1 = \operatorname*{argmax}_{\substack{\mathbf{u}\in\mathbb{R}^m \\ \|\mathbf{u}\|_2=1}} \mathbf{u}^* \boldsymbol{A}\boldsymbol{A}^* \mathbf{u}, \quad = \max_{\substack{\mathbf{u}\in\mathbb{R}^m \\ \|\mathbf{u}\|_2=1}} \mathbf{u}^* \boldsymbol{A}\boldsymbol{A}^* \mathbf{u} \lambda_{\max}(\boldsymbol{A}\boldsymbol{A}^*) = \sigma_1^2. \tag{6.14}$$

Remark 44

Thus, $\mathbf{u}^1$ *is the normalized eigenvector for the largest eigenvalue of* $\boldsymbol{A}\boldsymbol{A}^*$*. This matrix will later be seen as a "covariance" matrix in a statistical context, and the direction* $\mathbf{u}^1$ *maximizes the "variance".*

Proof of Lemma 43: Recall that

$$\operatorname{trace}(\boldsymbol{A}^*\boldsymbol{A}) = \sum_{j=1}^{n}(\boldsymbol{A}^*\boldsymbol{A})_{j,j} = \sum_{k=1}^{n}\sum_{j=1}^{m} a_{k,j}^2 = \sum_{j=1}^{n} \|\mathbf{a}^j\|_2^2.$$

Therefore

$$\begin{aligned}\sum_{j=1}^{n} \|\mathbf{a}^i - (\mathbf{u}^*\mathbf{a}^j)\mathbf{u}\|_2^2 &= \sum_{j=1}^{n} \|\mathbf{a}^j\|_2^2 - 2(\mathbf{a}^j)^*(\mathbf{u}^*\mathbf{a}^j)\mathbf{u} + (\mathbf{u}^*\mathbf{a}^j)^2 = \sum_{j=1}^{n} \|\mathbf{a}^j\|_2^2 - (\mathbf{u}^*\mathbf{a}^j)^2 \\ &= \operatorname{trace}(\boldsymbol{A}^*\boldsymbol{A}) - \|\boldsymbol{A}^*\mathbf{u}\|_2^2 = \operatorname{trace}(\boldsymbol{A}^*\boldsymbol{A}) - \mathbf{u}^*\boldsymbol{A}\boldsymbol{A}^*u. \qquad (6.15)\end{aligned}$$

Maximizing $\mathbf{u}^*\boldsymbol{A}\boldsymbol{A}^*u$ over $\|\mathbf{u}\|_2 = 1$, minimizes the left hand side of (6.15) as claimed. As hinted at in Remark 44, this yields the first eigenvector and maximal eigenvalue of $\boldsymbol{A}\boldsymbol{A}^*$ and hence, by Proposition 40 (4), the first column of the matrix $\mathbf{U}$ in the SVD and the square of the first singular value σ_1. □

A Greedy Characterization — Principal Component Analysis (PCA)

- $\mathbf{u}^1$ given by (6.14);
- given $\mathbf{u}^1, \ldots \mathbf{u}^k$, determine $\mathbf{u}^{k+1}$ by

$$\mathbf{u}^{k+1} = \underset{\substack{\mathbf{u}\in\mathbb{R}^m, \|\mathbf{u}\|_2=1 \\ \mathbf{u}\perp\mathbf{u}, \ldots, bu^k}}{\operatorname{argmax}} \mathbf{u}^* \boldsymbol{A}\boldsymbol{A}^* \mathbf{u}, \tag{6.16}$$

$\rightsquigarrow$ this is the successive construction of the columns $\mathbf{u}^j$ of the $\mathbf{U}$ in SVD $\boldsymbol{A} = \mathbf{U}\boldsymbol{S}\mathbf{V}^*$.

Statistical Interpretation: If $\mathbf{a}^j$ are random samples of an m-variate random variable the $\boldsymbol{A}\boldsymbol{A}^*$ is the corresponding covariance matrix which is diagonalized by the above process PCA (i.e., by the SVD). $\mathbf{u}^1$ is the direction that maximizes the variance - largest contribution to the total variance of the underlying distribution - providing most information.

If the distribution is normal with mean zero - $\mathcal{N}(0, \sigma)$ - then the $\mathbf{u}^j$ are uncorrelated and statistically independent $\rightsquigarrow$ cluster analysis, pattern recognition, feature discrimination,...more on this later ...

Example: consumer behavior; the vector $\mathbf{a}^j$ contains characteristic consumer traits such as age, gender, income, debts, location, etc. $\rightsquigarrow$ PCA helps identifying those possibly few characteristics with largest impact on buying patterns

Least Squares Method

Theorem 45

Let $\boldsymbol{A} \in \mathbb{K}^{m\times n}$, $\mathbf{b} \in \mathbb{K}^m$ *and consider*

$$\|\boldsymbol{A}\mathbf{x} - \mathbf{b}\|_2 \to_{\mathbf{x}\in\mathbb{K}^n} \min \tag{6.17}$$

Then

1. $\mathbf{x} \in \mathbb{K}^n$ *is a solution of* (6.17) *if and only if*

$$\boldsymbol{A}^*\boldsymbol{A}\mathbf{x} = \boldsymbol{A}^*\mathbf{b} \qquad \textit{called "normal equations"}. \tag{6.18}$$

2. (6.17) *has a unique solution* $\mathbf{x}$ *if and only if* $n \leq m$ *and* $\operatorname{rank}(\boldsymbol{A}) = n$.
3. *For any* $n, m \in \mathbb{N}$ *and* $\operatorname{rank}(\boldsymbol{A}) \leq \min\{m, n\}$ *arbitrary, there exists a unique* $\tilde{\mathbf{x}} \in \mathbb{K}^n$ *satisfying*

$$\|\boldsymbol{A}\tilde{\mathbf{x}} - \mathbf{b}\|_2 = \min_{\mathbf{x}\in\mathbb{K}^n} \|\boldsymbol{A}\mathbf{x} - \mathbf{b}\|_2 \text{ and } \|\tilde{u}\|_2 \leq \|\mathbf{x}'\|_2 \text{ for which } \|\boldsymbol{A}\mathbf{x}' - \mathbf{b}\|_2 = \min. \tag{6.19}$$

Moreover, $\tilde{\mathbf{x}}$ *is given by* $\tilde{\mathbf{x}} = \boldsymbol{A}^\dagger\mathbf{b}$. *(minimal norm solution)*

Proof of Theorem 45: Let $\mathbb{V} = \mathbb{K}^m, \mathbb{U} := \mathrm{ran}(\boldsymbol{A}) = \{\mathbf{u} = \boldsymbol{A}\mathbf{x} : \mathbf{x} \in \mathbb{K}^n\} \subset \mathbb{V}$. **ad (1):** Then, by Theorem 24,

$$\begin{aligned}
\|\boldsymbol{A}\mathbf{x} - \mathbf{b}\|_2 = \min_{\mathbf{x}' \in \mathbb{K}^n} \|\boldsymbol{A}\mathbf{x}' - \mathbf{b}\|_2 \quad &\Leftrightarrow \quad \mathbf{u}^*(\boldsymbol{A}\mathbf{x} - \mathbf{b}) = 0, \quad \forall\, \mathbf{u} \in \mathbb{U} = \mathrm{ran}(\boldsymbol{A}) \\
&\Leftrightarrow \quad (\boldsymbol{A}\mathbf{z})^*(\boldsymbol{A}\mathbf{x} - \mathbf{b}) = 0, \quad \forall\, \mathbf{z} \in \mathbb{K}^n \\
&\Leftrightarrow \quad \mathbf{z}^*\left(\boldsymbol{A}^*\boldsymbol{A}\mathbf{x} - \boldsymbol{A}^*\mathbf{b}\right) = 0, \quad \forall\, \mathbf{z} \in \mathbb{K}^n \\
&\overset{\textit{Remark } 25,(1)}{\Leftrightarrow} \quad \boldsymbol{A}^*\boldsymbol{A}\mathbf{x} - \boldsymbol{A}^*\mathbf{b} = 0, \qquad \rightsquigarrow \text{ (6.18).}
\end{aligned}$$

ad (2): If $m \geq n = \mathrm{rank}(\boldsymbol{A})$ then $\boldsymbol{A}^*\boldsymbol{A}$ is hermitian positive definite and hence non-singular (see Remark 12), which by (1) confirms (2).

By (2), it suffices to consider the case $\mathrm{rank}(\boldsymbol{A}) < \min\{m, n\}$. In this case the solution set

$$S(\boldsymbol{A}, \mathbf{b}) := \{\mathbf{x} \in \mathbb{K}^n : \|\boldsymbol{A}\mathbf{x} - \mathbf{b}\|_2 = \min\}$$

is easily seen to be given by

$$S(\boldsymbol{A}, \mathbf{b}) = \mathbf{x}^0 + \ker(\boldsymbol{A}) = \{\mathbf{x}^0 + \mathbf{z} : \mathbf{z} \in \mathbb{K}^n,\ \boldsymbol{A}\mathbf{z} = 0\}, \tag{6.20}$$

where $\mathbf{x}^0$ is any fixed minimizer of $\|\boldsymbol{A}\mathbf{x} - \mathbf{b}\|_2$. Taking $\mathbf{U} := \ker(\boldsymbol{A}) \subset \mathbb{V} := \mathbb{K}^n$, we know again from Theorem 24 that there exists a unique $\bar{\mathbf{z}} \in \ker(\boldsymbol{A}) = \mathbb{U}$ such that $\|\mathbf{x}^0 - \bar{\mathbf{z}}\|_2 = \min_{\mathbf{z} \in \ker(\boldsymbol{A})} \|\mathbf{x}^0 - \mathbf{z}\|_2$ and $\bar{z}$ is characterized by $\mathbf{z}^*(\mathbf{x}^0 - \bar{\mathbf{z}}) = 0$, $\mathbf{z} \in \ker(\boldsymbol{A})$. Hence $\tilde{\mathbf{x}} := \mathbf{x}^0 - \bar{\mathbf{z}} \in S(\boldsymbol{A}, \mathbf{b})$ is the unique minimal norm minimizer.

Proof of Theorem 45 continued: It remains to show that $\tilde{\mathbf{x}}$ is given by $\boldsymbol{A}^\dagger \mathbf{b}$.

To that end, let $\mathbf{y} := \boldsymbol{A}^\dagger \mathbf{b}$ and recall that $\boldsymbol{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^*$, $\boldsymbol{A}^\dagger = \mathbf{V}\boldsymbol{\Sigma}^\dagger\mathbf{U}^*$.

Show that $\mathbf{y} \in S(\boldsymbol{A}, \mathbf{b})$ by verifying that $\boldsymbol{A}^*\boldsymbol{A}\mathbf{y} = \boldsymbol{A}^*\mathbf{b}$ (see (1)):

$$\begin{aligned} \boldsymbol{A}^*\boldsymbol{A}\mathbf{y} &= (\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^*)^*(\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^*)\boldsymbol{A}^\dagger\mathbf{b} = \mathbf{V}\boldsymbol{\Sigma}^*\mathbf{U}^*\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^*\mathbf{V}\boldsymbol{\Sigma}^\dagger\mathbf{U}^*\mathbf{b} \\ &= \mathbf{V}\boldsymbol{\Sigma}^*\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\dagger\mathbf{U}^*\mathbf{b} = \mathbf{V}\boldsymbol{\Sigma}^*\mathbf{U}^*\mathbf{b} = \boldsymbol{A}^*\mathbf{b}, \end{aligned}$$

i.e., $\mathbf{y}$ is a solution of the normal equations and hence a minimizer.

Show that $\mathbf{y}$ has minimal Euclidean norm:

As shown before, this is equivalent to showing that $\mathbf{y} \perp \ker(\boldsymbol{A})$. By Proposition 40, it suffices to show that $(\mathbf{v}^k)^*\mathbf{y} = 0$, $k = r+1, \ldots, n$, where $r = \operatorname{rank}(\boldsymbol{A})$. To that end,

$$(\mathbf{v}^k)^*\mathbf{y} = (\mathbf{v}^k)^*\boldsymbol{A}^\dagger\mathbf{b} = (\mathbf{v}^k)^*\mathbf{V}\boldsymbol{\Sigma}^\dagger\mathbf{U}^* = (\mathbf{v}^\mathbf{k})^*\mathbf{V}_\mathbf{r}\boldsymbol{\Sigma}_\mathbf{r}^\dagger\mathbf{U}_\mathbf{r}^* = \mathbf{0}$$

since $k > r$ and the columns of $\mathbf{V}$ are pairwise orthogonal. Hence $\mathbf{y}$ solves Problem (6.19) and, by uniqueness, agrees with $\tilde{\mathbf{x}}$. □

References I

[1] G. H. Golub, C. F. van Loan.
Matrix Computations.
3. Aufl., Oxford University Press, 1996.

[2] Y. Saad.
Iterative Methods for Sparse Linear Systems.
2. Aufl. SIAM Publications, 2003.

[3] R. DeVore and G.G. Lorentz,
Constructive Approximation, vol. 303,
Springer Grundlehren, Springer, Berlin-Heidelberg, 1993.