Chapter 2

Here we learned about norms on spaces of matrices. Two cases are here important:

i) Induced matrix norms

Definition 0.0.1. *Let* $\mathbf{A} \in \mathbb{K}^{m\times n}$ *and we equip* $\mathbb{K}^n$ *with* $\|\cdot\|_{(n)}$ *and* $\mathbb{K}^m$ *with* $\|\cdot\|_{(m)}$. *The induced matrix norm is then*

$$\|\mathbf{A}\|_{(m,n)} = \sup_{\substack{\mathbf{x}\in\mathbb{K}^n \\ \mathbf{x}\neq 0}} \frac{\|\mathbf{A}\mathbf{x}\|_{(m)}}{\|\mathbf{x}\|_{(n)}} = \sup_{\substack{\mathbf{x}\in\mathbb{K}^n \\ \|\mathbf{x}\|_{(n)}=1}} \|\mathbf{A}\mathbf{x}\|_{(m)} \tag{1}$$

ii) Matrix norm on the vector space of matrices

Definition 0.0.2. *A function* $\|\cdot\| : \mathbb{K}^m \to \mathbb{R}$ *is called a norm if*

1. $\|\mathbf{x}\| \geqslant 0$ *for all* $\mathbf{x} \in \mathbb{K}^m$ *and* $\|\mathbf{x}\| = 0 \Leftrightarrow \mathbf{x} = \mathbf{0}$
2. $\|\mathbf{x}+\mathbf{y}\| \leqslant \|\mathbf{x}\| + \|\mathbf{y}\|$ *for all* $\mathbf{x},\mathbf{y} \in \mathbb{K}^m$
3. $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|$ *for all* $\mathbf{x} \in \mathbb{K}^m$ *and for all* $\alpha \in \mathbb{K}$.

We have also seen central statements like

- Young's product inequality:

Lemma 0.0.1 (Young's product inequality)**.** *Let* $a, b \in \mathbb{R}_{\geqslant 0}$. *Then*

$$ab \leqslant \frac{1}{p}a^p + \frac{1}{q}b^q \tag{2}$$

for $1 \leqslant p, q \leqslant \infty$ *and* $\frac{1}{p} + \frac{1}{q} = 1$.

Proof. Let $a, b \in \mathbb{R}_{\geqslant 0}$, and $t = \frac{1}{p}$ and $1 - t = \frac{1}{q}$. Then

$$\ln(ta^p + (1-t)b^q) \underset{(*)}{\geqslant} t\ln(a^p) + (1-t)\ln(b^q) = \ln(a) + \ln(b) = \ln(ab) \tag{3}$$

where we used that ln in concave in $(*)$. □

- Hölder inequality:

Theorem 0.1 (Hölder inequality)**.** *Let* $\mathbf{x},\mathbf{y} \in \mathbb{K}^m$. *Then*

$$|\langle\mathbf{x},\mathbf{y}\rangle| \leqslant \|\mathbf{x}\|_p\|\mathbf{y}\|_q, \tag{4}$$

where $1 \leqslant p, q \leqslant \infty$ *and* $\frac{1}{p} + \frac{1}{q} = 1$.

- Cauchy–Schwarz inequality:

Corollary 0.1.1 (Cauchy–Schwarz inequality)**.** *For* $p, q = 2$ *the Hölder inequality yields*

$$|\langle\mathbf{x},\mathbf{y}\rangle| \leqslant \|\mathbf{x}\|_2\|\mathbf{y}\|_2. \tag{5}$$

- Minkowski inequality

  Theorem 0.2 (Minkowski inequality)**.** *Let* $\mathbf{x}, \mathbf{y} \in \mathbb{K}^m$ *and* $p \geqslant 1$*. Then*

$$\|\mathbf{x} + \mathbf{y}\|_p \leqslant \|\mathbf{x}\|_p + \|\mathbf{y}\|_p \tag{6}$$

- The standard inner product on matrix spaces:

  Definition 0.2.1. *Let* $\mathbf{A}, \mathbf{B} \in \mathbb{K}^{m \times n}$*. The standard inner product of matrices is defined as*

$$\langle \mathbf{A}, \mathbf{B} \rangle = \operatorname{Tr}(\mathbf{A}^* \mathbf{B}), \tag{7}$$

- The Frobenius norm is induced by the standard inner product:

  Proposition 0.2.1. *The Frobenius norm is induced by the standard inner product of matrices, i.e., for* $\mathbf{A} \in \mathbb{K}^{m \times n}$

$$\|\mathbf{A}\|_{\mathrm{F}} = \sqrt{\langle \mathbf{A}, \mathbf{B} \rangle}. \tag{8}$$

- Unitarily invariant norms:

  Theorem 0.3. *For any* $\mathbf{A} \in \mathbb{K}^{m \times n}$ *and unitary* $\mathbf{U} \in \mathbb{K}^{m \times m}$*, we have*

$$\|\mathbf{UA}\|_2 = \|\mathbf{A}\|_2 \quad \text{and} \quad \|\mathbf{UA}\|_{\mathrm{F}} = \|\mathbf{A}\|_{\mathrm{F}} \tag{9}$$

  Proof. Let $\mathbf{A} \in \mathbb{K}^{m \times n}$ and $\mathbf{U} \in \mathbb{K}^{m \times m}$ be unitary. Then

$$\|\mathbf{UAx}\|_2 = \sqrt{\mathbf{x}^* \mathbf{A}^* \mathbf{U}^* \mathbf{UAx}} = \|\mathbf{Ax}\|_2 \Rightarrow \|\mathbf{UA}\|_2 = \|\mathbf{A}\|_2 \tag{10}$$

  and

$$\|\mathbf{UA}\|_{\mathrm{F}} = \sqrt{\operatorname{Tr}((\mathbf{UA})^* \mathbf{UA})} = \sqrt{\operatorname{Tr}(\mathbf{A}^* \mathbf{A})} = \|\mathbf{A}\|_{\mathrm{F}} \tag{11}$$

  □

In the homework assignments you have seen central statements like:

- Hermitian matrices have real-valued eigenvalues.
- Skew hermitian matrices have purely imaginary eigenvalues
- And matrix inequalities: $\|x\|_2 \leqslant \sqrt{m}\|x\|_\infty$

  Proof. Starting from the definition of the 2-norm, we find

$$\|x\|_2 = \left(\sum_{i=1}^{m} |x_i|^2 \right)^{1/2} \leqslant \left(\sum_{i=1}^{m} \max_{i \in [m]} |x_i|^2 \right)^{1/2} = \sqrt{m} \left(\max_{i \in [m]} |x_i|^2 \right)^{1/2} = \sqrt{m} \left(\max_{i \in [m]} |x_i| \right),$$

  hence $\sqrt{m}\|x\|_\infty$. The inequality is sharp for $x = (1, 1, \ldots, 1)^\top$, i.e., the vector with all entries equal to one, since $\|x\|_2 = \sqrt{m}$ and $\|x\|_\infty = 1$. □

Chapter 3

The most central object of this course and large parts of numerical linear algebra; the singular value decomposition:

Definition 0.3.1. *Let* $\mathbf{A} \in \mathbb{K}^{m\times n}$*. We call the factorization*

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*, \tag{12}$$

where $\mathbf{U} \in \mathbb{K}^{m\times m}$ *and* $\mathbf{V} \in \mathbb{K}^{n\times n}$ *are unitary, and* $\mathbf{\Sigma} \in \mathbb{K}^{m\times n}$ *is diagonal, singular value decomposition of* $\mathbf{A}$*.*

Theorem 0.4. *Every matrix* $\mathbf{A} \in \mathbb{K}^{m\times n}$ *has a singular value decomposition and the singular values* $\{\sigma_i\}$ *are uniquely determined. Moreover, if* $\mathbf{A}$ *is square and* σ_i *distinct, the left and right singular vectors* $\{\mathbf{u}_j\}$ *and* $\{\mathbf{v}_j\}$ *are uniquely determined up to complex signs, i.e., complex scaling factors of length one.*

The proof is long but parts can be asked:

- Then A^*A is positive semi-definite, indeed,

$$\mathbf{x}^*\mathbf{A}^*\mathbf{A}\mathbf{x} = (\mathbf{A}\mathbf{x})^*(\mathbf{A}\mathbf{x}) = \|\mathbf{A}\mathbf{x}\|_2 \geqslant 0. \tag{13}$$

Proposition 0.4.1. *Let* $\mathbf{A} \in \mathbb{K}^{m\times n}$*. Then* $\|\mathbf{A}\|_2 = \sigma_{\max}(\mathbf{A})$*, i.e., the largest singular value.*

Proof. Let $\mathbf{A} \in \mathbb{K}^{m\times n}$, with singular value decomposition $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ and σ_1 being the largest singular value. Then for $\|\mathbf{x}\|_2 = 1$

$$\|\mathbf{A}\mathbf{x}\|_2^2 = \langle \mathbf{x}, \mathbf{A}^*\mathbf{A}\mathbf{x}\rangle = \sum_{i=1}^{n} \sigma_i^2 \langle \mathbf{x}, \mathbf{v}_i\mathbf{v}_i^*\mathbf{x}\rangle \leqslant \sigma_1^2 \sum_{i=1}^{n} |\mathbf{v}_i^*\mathbf{x}|^2 = \sigma_1^2\|\mathbf{V}^*\mathbf{x}\|^2 \leqslant \sigma_1^2\|\mathbf{V}^*\|^2 = \sigma_1^2 \tag{14}$$

which is tight for $\mathbf{x} = \mathbf{v}_1$. □

We learned two key applications:

- Low-rank approximation
- Moore-Penrose inverse:

  Definition 0.4.1. *Let* $\mathbf{A} \in \mathbb{K}^{m\times n}$*. The matrix* $A^+ \in \mathbb{K}^{n\times m}$ *is called the pseudo inverse (Moose-Penrose) inverse of* $\mathbf{A}$ *if*

$$\begin{array}{llll} \text{i)} & \mathbf{A}\mathbf{A}^+\mathbf{A} = \mathbf{A} & \text{iii)} & (\mathbf{A}\mathbf{A}^+)^* = \mathbf{A}\mathbf{A}^+ \\ \text{ii)} & \mathbf{A}^+\mathbf{A}\mathbf{A}^+ = \mathbf{A}^+ & \text{iv)} & (\mathbf{A}^+\mathbf{A})^* = \mathbf{A}^+\mathbf{A} \end{array}$$

That have different properties:

Low-rank

Theorem 0.5 (Eckast-Young-Mirsky – spectral norm). *Let* $\mathbf{A} \in \mathbb{K}^{m\times m}$ *with* $\text{rank}(\mathbf{A}) = r$. *For any* k *with* $1 \leqslant k < r$*, define*

$$A_k = \sum_{j=1}^{k} \sigma_j u_j v_j^*. \tag{15}$$

Then

$$\|\mathbf{A} - \mathbf{A}_k\|_2 = \inf_{\substack{\mathbf{B}\in\mathbb{K}^{m\times m}\\ \text{rank}(\mathbf{B})\leqslant k}} \|\mathbf{A} - \mathbf{B}\|_2 = \sigma_{k+1} \tag{16}$$

Proof. First note that

$$\|A - A_k\|_2 = \sigma_{k+1}. \tag{17}$$

It remains to show that $\mathbf{A}_k$ is the infimum. To that end, assume the exist $\mathbf{B}_k = \mathbf{X}\mathbf{Y}^*$ where $\mathbf{X}, \mathbf{Y}$ have k-columns and that

$$\|\mathbf{A} - \mathbf{B}_k\|_2 < \|\mathbf{A} - \mathbf{A}_k\|_2 = \sigma_{k+1}. \tag{18}$$

However, since

$$\text{rank}(\mathbf{Y}) = k < k + 1 = \text{rank}([\mathbf{v}_1|...|\mathbf{v}_{k+1}]) \tag{19}$$

there exists a linear combination of right singular vectors of

$$\mathbf{w} = c_1\mathbf{v}_1 + ... + c_{k+1}\mathbf{v}_{k+1} \tag{20}$$

with

$$\mathbf{Y}^*\mathbf{w} = \mathbf{0}. \tag{21}$$

W.l.o.g. we assume $\mathbf{w}$ is normalized, otherwise we normalize $\mathbf{w}$. Then,

$$\|\mathbf{A} - \mathbf{B}_k\|_2^2 \geqslant \|(\mathbf{A} - \mathbf{B}_k)\,\mathbf{w}\|_2^2 = \|A\mathbf{w}\|_2^2 = c_1^2\sigma_1^2 + ... + c_{k+1}^2\sigma_{k+1}^2 \geqslant \sigma_{k+1}^2 \tag{22}$$

□

Theorem 0.6 (Courant-Fisher min-max – singular values). *For* $\mathbf{A} \in \mathbb{K}^{m\times n}$*, we have*

$$\sigma_k = \max_{\substack{V\subset\mathbb{K}^n\\ \dim(V)=k}} \min_{\substack{\|\mathbf{v}\|=1\\ \mathbf{v}\in V}} \|\mathbf{A}\mathbf{v}\|_2 \tag{23}$$

and

$$\sigma_{k+1} = \min_{\substack{V\subset\mathbb{K}^n\\ \dim(V)=n-k}} \max_{\substack{\|\mathbf{v}\|=1\\ \mathbf{v}\in V}} \|\mathbf{A}\mathbf{v}\|_2. \tag{24}$$

Theorem 0.7 (Weyl's inequality). *Let* $\mathbf{A}, \mathbf{B} \in \mathbb{K}^{m\times n}$ *and denote its singular values by* $\sigma_i(\mathbf{A})$ *and* $\sigma_i(\mathbf{B})$*, respectively. We then have*

$$\sigma_{i+j-1}(\mathbf{A} + \mathbf{B}) \leqslant \sigma_i(\mathbf{A}) + \sigma_j(\mathbf{B}). \tag{25}$$

Theorem 0.8 (Eckert-Young-Mirsky for Frobenins norm). *Let* $\mathbf{A} \in \mathbb{K}^{m\times m}$ *with* $\text{rank}(\mathbf{A}) = r$. *For any* k *with* $1 \leqslant k < r$*, define*

$$\mathbf{A}_k = \sum_{j=1}^{k} \sigma_j \mathbf{u}_j \mathbf{v}_j^*. \tag{26}$$

Then

$$\|\mathbf{A} - \mathbf{A}_k\|_F = \inf_{\substack{\mathbf{B}\in\mathbb{K}^{m\times m}\\ \text{rank}(\mathbf{B})\leqslant k}} \|\mathbf{A} - \mathbf{B}\|_F = \sqrt{\sigma_{k+1}^2 + ... + \sigma_r^2}. \tag{27}$$

Moore Penrose inverse

Proposition 0.8.1. *Let* $\mathbf{A} \in \mathbb{K}^{m\times n}$ *with* $m \leqslant n$ *and* $\mathbf{A}^+$ *its Moore-Penrose inverse. Then*

$$\text{range}(\mathbf{A}^+) \perp \ker(\mathbf{A}). \tag{28}$$

Proof. Let $\mathbf{A} \in \mathbb{K}^{m\times n}$ with $\mathbf{A}^+$ its Moore-Penrose inverses. Recall that

$$\mathbf{A}\mathbf{A}^+\mathbf{A} = \mathbf{A} \quad \text{and} \quad (\mathbf{A}^+\mathbf{A})^* = \mathbf{A}^+\mathbf{A}. \tag{29}$$

Moreover let $\mathbf{y} \in \text{range}(\mathbf{A}^+)$, i.e., $\mathbf{y} = \mathbf{A}^+\mathbf{b}$ for some $\mathbf{b} \in \mathbb{K}^m$, and $\mathbf{x} \in \ker(\mathbf{A})$. Then

$$\langle \mathbf{y}, \mathbf{x}\rangle = \langle \mathbf{A}^+\mathbf{b}, \mathbf{x}\rangle = \langle \mathbf{A}^+\mathbf{A}\mathbf{A}^+\mathbf{b}, \mathbf{x}\rangle = \langle \mathbf{A}^+\mathbf{b}, (\mathbf{A}^+\mathbf{A})^*\mathbf{x}\rangle = \langle \mathbf{A}^+\mathbf{b}, \mathbf{A}^+\mathbf{A}\mathbf{x}\rangle = 0 \tag{30}$$

□

Theorem 0.9. *Let* $\mathbf{A} \in \mathbb{K}^{m\times n}$*, the Moore-penrose inverse* $\mathbf{A}^+$ *is unique.*

Proposition 0.9.1. *Let* $\mathbf{A} \in \mathbb{K}^{m\times n}$ *with* $m > n$ *and* $\text{rank}(\mathbf{A}) = n$*. Then*

$$\mathbf{A}^+ = (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*. \tag{31}$$

Theorem 0.10. *If* $\mathbf{A} \in \mathbb{K}^{m\times m}$ *attains an inverse, then* $\mathbf{A}^{-1} = \mathbf{A}^+$*.*

Proof. Note that

$$\mathbf{I} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{A}\mathbf{A}^+\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}\mathbf{A}^+ \tag{32}$$

hence

$$\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{I} = \mathbf{A}^{-1}\mathbf{A}\mathbf{A}^+ = \mathbf{A}^+ \tag{33}$$

□

Theorem 0.11. *Let* $\mathbf{A} \in \mathbb{K}^{m\times n}$ *with* $\mathbf{A}^+ \in \mathbb{K}^{n\times m}$ *its pseudo inverse, then*

$$\left(\mathbf{A}^+\right)^+ = \mathbf{A}. \tag{34}$$

Application of MP inverse:

The MP inverse solves the over-determined least squares problem, i.e., minimize

$$\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2. \tag{35}$$

where $\mathbf{A} \in \mathbb{K}^{m\times n}$ and $m \geqslant n$ – we say "$\mathbf{A}$ is tall and skinny". We have more equations than variables and consequently zero solutions to the system. We therefore seek $\mathbf{x} \in \mathbb{K}^n$ that minimizes the above residual, i.e.,

$$\min_{\mathbf{x}\in\mathbb{K}^n} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2. \tag{36}$$

To that end, we compute the gradient of with respect to $\mathbf{x}$:

$$\frac{\partial}{\partial \mathbf{x}}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 = 2\mathbf{A}^*(\mathbf{A}\mathbf{x} - \mathbf{b}). \tag{37}$$

Enforcing first-order optimality yields the normal equation

$$\mathbf{A}^*\mathbf{A}\mathbf{x} = \mathbf{A}^*\mathbf{b}. \tag{38}$$

Assuming $\mathbf{A}^*\mathbf{A}$ is invertible, which holds if $\mathbf{A}$ has full rank, we can solve the normal equation, i.e.,

$$\mathbf{x} = (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\mathbf{b} = \mathbf{A}^+\mathbf{b}. \tag{39}$$

Chapter 4

This Chapter was all about QR factorization. We learned

Definition 0.11.1. *Let* $\mathbf{P} \in \mathbb{K}^{m \times m}$*. We call* $\mathbf{P}$ *a projector if and only if*

$$\mathbf{P}^2 = \mathbf{P}, \tag{40}$$

i.e., $\mathbf{P}$ *is idempotent.*

Remark 0.11.1. *This definition includes both, orthogonal and non-orthogonal projectors. To avoid confusion, we call non-orthogonal projectors oblique projectors.*

Proposition 0.11.1. *If* $\mathbf{P} \in \mathbb{K}^{m \times m}$ *is a projector, then* $\mathbf{I} - \mathbf{P}$ *is also a projector.*

Proof. Note that

$$(\mathbf{I} - \mathbf{P})^2 = (\mathbf{I} - \mathbf{P})(\mathbf{I} - \mathbf{P}) = \mathbf{I} - 2\mathbf{P} + \mathbf{P}^2 = \mathbf{I} - \mathbf{P} \tag{41}$$

which shows the claim. □

Definition 0.11.2. *Let* $\mathbf{P} \in \mathbb{K}^{m \times m}$ *be a projector. We call* $\mathbf{P}$ *an orthogonal projector if and only if*

$$\langle \mathbf{Px}, \mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{Py} \rangle \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{K}^m, \tag{42}$$

i.e., $\mathbf{P} \in \mathbb{H}_m(\mathbb{K})$.

Definition 0.11.3. *Let* $\mathbf{A} \in \mathbb{K}^{m \times n}$*. We call the factorization*

$$\mathbf{A} = \mathbf{QR} \tag{43}$$

where $\mathbf{Q} \in \mathbb{K}^{m \times m}$ *unitary, and* $\mathbf{R} \in \mathbb{K}^{m \times n}$ *is an upper triangular matrix, a QR-factorization of* $\mathbf{A}$.

Remark 0.11.2. *We shall now take a closer look at the QR-factorization:*
Consider a reduced QR-factorization of $\mathbf{A} \in \mathbb{K}^{m \times n}$ *with* $n \leqslant m$*, i.e.,*

$$[\mathbf{a}_1|...|\mathbf{a}_n] = [\mathbf{q}_1|...|\mathbf{q}_n] \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1_n} \\ 0 & r_{22} & \cdots & r_{2_n} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & r_{nn} \end{bmatrix} \tag{44}$$

hence

$$\begin{aligned}
\mathbf{a}_1 &= r_{11}\mathbf{q}_1 & \Leftrightarrow \quad \mathbf{q}_1 &= \frac{\mathbf{a}_1}{r_{11}} = \frac{\mathbf{a}_1}{\|\mathbf{a}_1\|} \\
\mathbf{a}_2 &= r_{12}\mathbf{q}_1 + r_{22}\mathbf{q}_2 & \Leftrightarrow \quad \mathbf{q}_2 &= \frac{\mathbf{a}_2 - r_{12}\mathbf{q}_1}{r_{22}} = \frac{\mathbf{a}_2 - \langle \mathbf{q}_1, \mathbf{a}_2 \rangle \mathbf{q}_1}{r_{22}} = \frac{(\mathbf{I} - \mathbf{q}_1\mathbf{q}_1^*)\mathbf{a}_2}{\|(\mathbf{I} - \mathbf{q}_1\mathbf{q}_1^*)\mathbf{a}_2\|} \\
\mathbf{a}_3 &= r_{13}\mathbf{q}_1 + r_{23}\mathbf{q}_2 + r_{33}\mathbf{q}_3 & \Leftrightarrow \quad \mathbf{q}_3 &= \frac{\mathbf{a}_3 - r_{13}\mathbf{q}_1 - r_{23}\mathbf{q}_2}{r_{33}} = \frac{(\mathbf{I} - \mathbf{q}_1\mathbf{q}_1^* - \mathbf{q}_2\mathbf{q}_2^*)\mathbf{a}_3}{\|(\mathbf{I} - \mathbf{q}_1\mathbf{q}_1^* - \mathbf{q}_2\mathbf{q}_2^*)\mathbf{a}_3\|} \\
& & \vdots & \\
\mathbf{a}_i &= \sum_{j=1}^{i} r_{ji}\mathbf{q}_j & \Leftrightarrow \quad \mathbf{q}_i &= \frac{\mathbf{a}_i - \sum_{j=1}^{i-1} r_{ji}\mathbf{q}_j}{r_{ii}} = \frac{(\mathbf{I} - \sum_{j=1}^{i-1} \mathbf{q}_j\mathbf{q}_j^*)\mathbf{a}_i}{\|(\mathbf{I} - \sum_{j=1}^{i-1} \mathbf{q}_j\mathbf{q}_j^*)\mathbf{a}_i\|}
\end{aligned} \tag{45}$$

This gave rise to three algorithms

- Classical Gram-Schmidt
- Modified Gram-Schmidt
- Iterative Gram-Schmidt

Together with their operational count.

Definition 0.11.4. *Let $\mathbf{v} \in \mathbb{K}^n$ be a normal vector defining a hyperplane. The transformation*

$$f_{\mathrm{H}} : \mathbb{K}^n \to \mathbb{K}^n \ : \ x \mapsto x - 2\langle x, v \rangle v$$

is the Householder transformation about the hyperplane defined by the normal vector $\mathbf{v} \in \mathbb{K}^n$.

Proposition 0.11.2. *Let $\mathbf{v} \in \mathbb{K}^n$ be a normal vector defining a hyperplane and f_{H} be the Householder transformation about the hyperplane defined by the normal vector $\mathbf{v} \in \mathbb{K}^n$. Then f_{H} is a linear map and its matrix representation is*

$$\mathbf{P}_{\mathbf{v}} = \mathbf{I} - 2\mathbf{v}\mathbf{v}^*$$

Proposition 0.11.3. *Let $\mathbf{v} \in \mathbb{K}^n$ be a normal vector defining a hyperplane and f_{H} be the Householder transformation about the hyperplane defined by the normal vector $\mathbf{v} \in \mathbb{K}^n$. The householder matrix $\mathbf{P}_{\mathbf{v}}$ fulfills:*

i)	Hermitian	$(\mathbf{P}_{\mathbf{v}} = \mathbf{P}_{\mathbf{v}}^*)$	iv)	$\mathbf{P}_{\mathbf{v}}$ has eigenvalues ± 1
ii)	Unitary	$(\mathbf{P}_{\mathbf{v}}^{-1} = \mathbf{P}_{\mathbf{v}}^*)$	v)	$\det(\mathbf{P}_{\mathbf{v}}) = -1$
iii)	Involutory	$(\mathbf{P}_{\mathbf{v}}^{-1} = \mathbf{P}_{\mathbf{v}})$		

Proof. First note that

$$\mathbf{P}_{\mathbf{v}}^* = (\mathbf{I} - 2\mathbf{v}\mathbf{v}^*)^* = \mathbf{I} - 2\mathbf{v}\mathbf{v}^* = \mathbf{P}_{\mathbf{v}} \tag{46}$$

which shows i). Next, we consider

$$\mathbf{P}_{\mathbf{v}}^2 = (\mathbf{I} - 2\mathbf{v}\mathbf{v}^*)(\mathbf{I} - 2\mathbf{v}\mathbf{v}^*) = \mathbf{I} - 4\mathbf{v}\mathbf{v}^* + 4\mathbf{v}\mathbf{v}^* = \mathbf{I} \tag{47}$$

showing that $\mathbf{P}_{\mathbf{v}}$ is involutory. This in turn yields that $\mathbf{P}_{\mathbf{v}}$ is unitary, since

$$\mathbf{P}_{\mathbf{v}}^{-1} = \mathbf{P}_{\mathbf{v}} = \mathbf{P}_{\mathbf{v}}^*. \tag{48}$$

Note that for $\mathbf{u} \perp \mathbf{v}$ we have $\mathbf{P}_{\mathbf{v}}\mathbf{u} = \mathbf{u}$. Since there are $n-1$ linearly independent vectors $\mathbf{u} \in \mathbb{K}^n$ fulfilling $\mathbf{u} \perp \mathbf{v}$, the eigenspace of $\mathbf{P}_{\mathbf{v}}$ corresponding to the eigenvalue $\lambda = 1$ is $n-1$ dimensional. Moreover $\mathbf{P}_{\mathbf{v}}\mathbf{v} = -\mathbf{v}$, showing iv). By iv), we know that $\mathbf{P}_{\mathbf{v}}$ is diagonalizable with $n-1$ eigenvalues $\lambda_1 = 1$ and one eigenvalue $\lambda_1 = -1$. Aplying the determinant multiplication Theorem we have

$$\det(\mathbf{P}_{\mathbf{v}}) = \det \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & 1 & 0 \\ 0 & \dots & 0 & -1 \end{bmatrix} = (-1)\, 1^{n-1} = -1 \tag{49}$$

□

The most important application:

Householder QR

We also did its operational count and argued how to keep it low:

Work with Householder vectors

Definition 0.11.5. *Let* $i, j \in [\![m]\!]$ *and* $\theta \in [0, 2\pi)$*. A matrix* $\mathbf{G}(i, j, \theta) \in \mathbb{K}^{m \times m}$ *defined through*

$$[\mathbf{G}(i,j,\theta)]_{l,m} = \begin{cases} 1 & \text{, if } l = m, \text{ and } l \neq i, j \\ \cos(\theta) & \text{, if } l = m = i, j \\ \sin(\theta) & \text{, if } l = i, \text{ and } m = j \\ -\sin(\theta) & \text{, if } l = j, \text{ and } m = i \\ 0 & \text{, else.} \end{cases} \tag{50}$$

is called Givens rotation around θ *in the* i*-*j*-plane.*

Proposition 0.11.4. *Givens rotations are orthogonal matrices, i.e.,* $\mathbf{G}^\top = \mathbf{G}^{-1}$.

Remark 0.11.3. *Givens rotations indeed rotate in the* i*-*j*-plane. Consider*

$$\mathbf{G}(i,j,\theta)\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_{i-1} \\ cx_i - sx_j \\ x_{i+1} \\ \vdots \\ cx_j + sx_i \\ x_{j+1} \\ \vdots \\ x_m \end{bmatrix} \Bigg] = \begin{bmatrix} c & -s \\ s & c \end{bmatrix} \begin{bmatrix} x_i \\ x_j \end{bmatrix} \tag{51}$$

substituting c *and* s *with* $\cos(\theta)$ *and* $\sin(\theta)$*, respectively, we see that this corresponds to a (counter-clockwise) rotation through an angle* θ *in the* i*-*j*-plane.*

We designed an algorithm that uses Givens rotations to compute a QR factorization and discussed the operational count, and how to keep it low:

Track only the Givens angles.

Chapter 5

The topic of Chapter 5 was accuracy. We distinguish three "error-contributing" parts

i) Conditioning of a problem

ii) Floating point errors

iii) Algorithmic stability

0.1 Conditioning of a problem

Definition 0.11.6. *Consider the problem $f : X \to Y$, where $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ are normed vector spaces. Let δx be a perturbation on x and define $\delta f = f(x + \delta x) - f(x)$. Then the absolute condition number is defined as*

$$\hat{\kappa}_f(x) = \lim_{\delta \to 0} \sup_{\|\delta x\| \leqslant \delta} \frac{\|\delta f\|_Y}{\|\delta x\|_X} \tag{52}$$

Proposition 0.11.5. *Consider the problem $f : X \to X$, where $(X, \|\cdot\|)$ is a normed vector space. Let f be differentiable, then*

$$\hat{\kappa}_f(x) = \|Df(x)\| \tag{53}$$

Definition 0.11.7. *Consider the problem $f : X \to Y$, where $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ are normed vector spaces. Let δx be a perturbation on x and define $\delta f = f(x + \delta x) - f(x)$. Then the relative condition number in x is defined as*

$$\kappa_f(x) = \lim_{\delta \to 0} \sup_{\|\delta x\| \leqslant \delta} \left(\frac{\|\delta f\|_Y}{\|f(x)\|_Y} \frac{\|x\|_X}{\|\delta x\|_X} \right) = \hat{\kappa}_f(x) \frac{\|x\|_X}{\|f(x)\|_Y} \tag{54}$$

Proposition 0.11.6. *Let $\mathbf{A} \in \mathbb{K}^{m \times n}$ and consider the problem*

$$f : \mathbb{K}^n \to \mathbb{K}^m \; ; \; \mathbf{x} \mapsto \mathbf{A}\mathbf{x}. \tag{55}$$

Then

$$\kappa_f(\mathbf{x}) = \|\mathbf{A}\| \frac{\|\mathbf{x}\|}{\|\mathbf{A}\mathbf{x}\|} \tag{56}$$

Corollary 0.11.1. *Let $\mathbf{A} \in \mathbb{K}^{m \times m}$ be non-singular. Then*

$$\kappa_f(\mathbf{x}) \leqslant \|\mathbf{A}\| \|\mathbf{A}^{-1}\| \tag{57}$$

Remark 0.11.4. *Let $\mathbf{A} \in \mathbb{K}^{m \times m}$ and considering the problem*

$$f : \mathbb{K}^n \to \mathbb{K}^m \; ; \; \mathbf{x} \mapsto \mathbf{A}\mathbf{x}, \tag{58}$$

we note that

$$\kappa_f(\mathbf{x}) \leqslant \sup_{\substack{\mathbf{x} \in \mathbb{K}^m \\ \|\mathbf{x}\| \neq 0}} \kappa_f(\mathbf{x}) = \|\mathbf{A}\| \|\mathbf{A}^{-1}\| \tag{59}$$

constitutes a worst-case scenario. We therefore denote the condition number of a matrix

$$\kappa(\mathbf{A}) := \|\mathbf{A}\| \|\mathbf{A}^{-1}\|. \tag{60}$$

Note that if we impose $\|\cdot\|_2$ *on* $\mathbb{K}^m$ *we have*

$$\|A^{-1}\|_2 = \frac{1}{\sigma_m} \tag{61}$$

and therewith

$$\kappa(\mathbf{A}) = \frac{\sigma_1}{\sigma_m} \tag{62}$$

This argument can furthermore be extended to linear problems defined by general matrices $\mathbf{A} \in \mathbb{K}^{m\times n}$*, with the adjustment that*

$$\kappa(\mathbf{A}) := \|\mathbf{A}\|\|\mathbf{A}^+\|. \tag{63}$$

Again imposing the spectral norm, we have

$$\|A^+\|_2 = \frac{1}{\sigma_n} \tag{64}$$

and therewith

$$\kappa(\mathbf{A}) = \frac{\sigma_1}{\sigma_n} \tag{65}$$

where $\mathbf{A}$ *was assumed to have full rank and* $n < m$*.*

Proposition 0.11.7. *Let* $\mathbf{b} \in \mathbb{K}^m$ *and consider the problem*

$$f : \mathrm{GL}(m) \to \mathbb{K}^m \ ; \ \mathbf{A} \mapsto \mathbf{A}^{-1}\mathbf{b}. \tag{66}$$

Then

$$\kappa_f(\mathbf{A}) \leqslant \kappa(\mathbf{A}) \tag{67}$$

Proof. For the considered problem we need to quantify

$$\delta\mathbf{x} = (\mathbf{A} + \delta\mathbf{A})^{-1}\mathbf{b} - \mathbf{A}^{-1}\mathbf{b} \tag{68}$$

To that end we consider the inverse problem

$$\begin{aligned} & \mathbf{b} = (\mathbf{A} + \delta\mathbf{A})(\mathbf{x} + \delta\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{A}\delta\mathbf{x} + \delta\mathbf{A}\mathbf{x} + \delta\mathbf{A}\delta\mathbf{x} = \mathbf{b} + \mathbf{A}\delta\mathbf{x} + \delta\mathbf{A}\mathbf{x} \\ \Leftrightarrow \quad & \mathbf{0} = \mathbf{A}\delta\mathbf{x} + \delta\mathbf{A}\mathbf{x} \\ \Leftrightarrow \quad & \delta\mathbf{x} = -\mathbf{A}^{-1}(\delta\mathbf{A})\mathbf{x} \end{aligned} \tag{69}$$

therefore

$$\|\delta\mathbf{x}\| \leqslant \|\mathbf{A}^{-1}\|\|\delta\mathbf{A}\|\|\mathbf{x}\|. \tag{70}$$

This yields that

$$\kappa_f(\mathbf{A}) = \lim_{\delta\to 0} \sup_{\|\delta\mathbf{A}\|\leqslant\delta} \left(\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \frac{\|\mathbf{A}\|}{\|\delta\mathbf{A}\|} \right) \leqslant \frac{\|\mathbf{A}^{-1}\|\|\delta\mathbf{A}\|\|\mathbf{x}\|}{\|\mathbf{x}\|} \frac{\|\mathbf{A}\|}{\|\delta\mathbf{A}\|} = \|\mathbf{A}^{-1}\|\|\mathbf{A}\| = \kappa(\mathbf{A}) \tag{71}$$

□

Floating point arithmetics

Definition 0.11.8. *Consider $x \in \mathbb{R}$, and let*

i) $b \in \mathbb{N}_+$ *be the basis*

ii) $\delta \in \{\pm 1\}$ *be the sign*

iii) $e \in \mathbb{Z}$ *the exponent*

We call

$$x = \delta \left(\sum_{n=1}^{\infty} a_k b^{-k} \right) b^e \tag{72}$$

the b-adic representation x, where $(a_k)_{k \in \mathbb{N}} \subseteq \mathbb{N}$ with $0 \leqslant a_k < b$ for all k.

Definition 0.11.9. *Let $b \in \mathbb{N}_+$ be the basis and $x \in \mathbb{R}$ with b-adic representation*

$$x = \delta \left(\sum_{n=1}^{\infty} a_k b^{-k} \right) b^e. \tag{73}$$

We call

$$\hat{x} = \delta \left(\sum_{n=1}^{m} a_k b^{-k} \right) b^e \tag{74}$$

the m-floating point representation of x. We call m the mantissa length.

Remark 0.11.5. *We are here mostly concerned with a binary and finite representation of real numbers, i.e., $b = 2$ and $m < \infty$. We here may moreover define the normalized representation i.e.*

$$\hat{x} = \delta \left(1 + \sum_{n=1}^{m} a_k b^{-k} \right) b^e = \mathrm{fl}_{b,m,e}(x). \tag{75}$$

Note that this (potentially) results in a shift in the exponent, yet it allows us a broader range of numbers to represent as we have an implicit leading one.

Examples:

- IEEE 754 64-bit standard
- IEEE 754 32-bit standard

Definition 0.11.10. *We define the machine epsilon as*

$$\varepsilon ps = \inf\{\varepsilon \in \mathbb{R}_{>0} \mid \mathrm{fl}_{b,m,e}(1 + \varepsilon) > 0\}. \tag{76}$$

Remark 0.11.6. *The fundamental axiom of floating point arithmetic states that for all $x, y \in \mathcal{F}$, there exists a ε with $|\varepsilon| \leqslant \varepsilon ps$, s.t.*

$$x \circledast y = x \star y(1 + \varepsilon). \tag{77}$$

Put differently, every floating point operation is exact up to a relative error of size at most εps.

Numerical stability

Definition 0.11.11. *Given a problem $f : X \to Y$, where $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ are normed vectors spaces, and $\hat{f}$ is an algorithm that approximates f. We call*

$$\|f(x) - \hat{f}(x)\|_Y \tag{78}$$

the absolute forward error of $\hat{f}$ in x, and

$$\frac{\|f(x) - \hat{f}(x)\|_Y}{\|f(x)\|_Y} \tag{79}$$

the relative forward error. We call the algorithm $\hat{f}$ (forward) stable if

$$\frac{\|f(x) - \hat{f}(x)\|_Y}{\|f(x)\|_Y} \in \mathcal{O}(\varepsilon ps). \tag{80}$$

Definition 0.11.12. *Given a problem $f : X \to Y$, where $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ are normed vectors spaces, and $\hat{f}$ is an algorithm that approximates f. We define the backward error of $\hat{f}(x)$ as*

$$\min \left\{ \frac{\|\delta \mathbf{x}\|}{\|\mathbf{x}\|} \;\middle|\; \hat{f}(\mathbf{x}) = f(\mathbf{x} + \delta \mathbf{x}) \right\} \tag{81}$$

We say that $\hat{f}$ is backward stable if and only if for all $x \in X$ there exists a $\hat{x} \in X$ with $\|x - \hat{x}\|/\|x\| \in \mathcal{O}(\varepsilon ps)$ such that

$$\hat{f}(x) = f(\hat{x}) \tag{82}$$

Proposition 0.11.8. *$f : X \to Y$, where $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ are normed vectors spaces, and let f be well-conditioned. Then, an algorithm that is backward stable is also forward stable stable.*

Chapter 6

The subject of this chapter was matrix factorizations, in particular, LU and Cholesky.

Definition 0.11.13. *Let $\mathbf{A} \in \mathbb{K}^{m \times m}$. We call the factorization*

$$\mathbf{A} = \mathbf{LU} \tag{83}$$

where $\mathbf{L} \in \mathbb{K}^{m \times m}$ is lower triangular and $\mathbf{U} \in \mathbb{K}^{m \times m}$ is upper triangular an LU-factorization of $\mathbf{A}$.

Proposition 0.11.9. *The matrix $\mathbf{L}$ is given by*

$$[\mathbf{L}]_{j,k} = l_{j,k} = \frac{\mathbf{A}_{jk}}{\mathbf{A}_{kk}} \tag{84}$$

for $j \geqslant k$.

Algorithm 1 LU factorization (without pivoting)

Require: $\mathbf{A} \in \mathbb{K}^{m \times m}$
Ensure: $\mathbf{L} \in \mathbb{K}^{m \times m}$, $\mathbf{U} \in \mathbb{K}^{m \times m}$
  $\mathbf{U} \leftarrow \mathbf{A}$
  $\mathbf{L} \leftarrow \mathbf{I}$
  for k=1 to m-1 **do**
    for j=k+1 to m **do**
      $l_{jk} \leftarrow u[j,k]/u[k,k]$
      $u[j,k:m] \leftarrow u[j,k:m] - l_{jk}\, u[k,k:m]$
    end for
  end for

Algorithm 2 LU factorization with partial pivoting

Require: $\mathbf{A} \in \mathbb{K}^{m \times m}$
Ensure: $\mathbf{L} \in \mathbb{K}^{m \times m}$, $\mathbf{U} \in \mathbb{K}^{m \times m}$ and $\mathbf{P} \in \mathbb{K}^{m \times m}$
  $\mathbf{U} \leftarrow \mathbf{A}$
  $\mathbf{L} \leftarrow \mathbf{I}$
  $\mathbf{P} \leftarrow \mathbf{I}$
  for k=1 to m-1 **do**
    Select $i \geqslant k$ s.t. $|U[i,k]| \geqslant |U[j,k]|$ for all $j \geqslant k$
    $\mathbf{U}[k,k:m] \leftrightarrow \mathbf{U}[i,k:m]$ (swap rows)
    $\mathbf{L}[k,1:k-1] \leftrightarrow \mathbf{L}[i,1:k-1]$ (swap rows)
    $\mathbf{P}[k,1:m] \leftrightarrow \mathbf{P}[i,1:m]$ (swap rows)
    for j=k+1 to m **do**
      $\mathbf{L}[j,k] \leftarrow \mathbf{U}[j:k]/\mathbf{U}[k,k]$
      $\mathbf{U}[j,k:m] \leftarrow \mathbf{U}[j,k:m] - \mathbf{L}[j,k]\,\mathbf{U}[k,k:m]$
    end for
  end for

Definition 0.11.14. *Let* $\mathbf{A} \in \mathbb{H}_m(\mathbb{K})$ *with* $0 \prec \mathbf{A}$ *. We call the factorization*

$$\mathbf{A} = \mathbf{L}\mathbf{L}^* \tag{85}$$

where $\mathbf{L} \in \mathbb{K}^{m \times m}$ *is lower triangular a Cholesky factorization of* $\mathbf{A}$*.*

Proposition 0.11.10. *Let* $\mathbf{A} \in \mathbb{H}_m(\mathbb{K})$ *with* $0 \prec \mathbf{A}$*, and* $\mathbf{X} \in \mathbb{K}^{m \times n}$ *with* $m \geqslant n$ *be full rank. Then*

$$0 \prec \mathbf{X}^*\mathbf{A}\mathbf{X} \tag{86}$$

is hermitian.

Proof. We first note that

$$(\mathbf{X}^*\mathbf{A}\mathbf{X})^* = \mathbf{X}^*\mathbf{A}^*\mathbf{X} = \mathbf{X}^*\mathbf{A}\mathbf{X}. \tag{87}$$

Moreover, since $\mathbf{X}$ is full rank, we know that $\mathbf{X}\mathbf{x} \neq \mathbf{0}$ for all $\mathbf{x} \neq \mathbf{0}$. Hence,

$$\mathbf{x}^*(\mathbf{X}^*\mathbf{A}\mathbf{X})\mathbf{x} = (\mathbf{X}x)^*\mathbf{A}(\mathbf{X}\mathbf{x}) > 0 \tag{88}$$

since $0 \prec \mathbf{A}$. □

Corollary 0.11.2. *Let* $\mathbf{A} \in \mathbb{H}_m(\mathbb{K})$ *with* $0 \prec \mathbf{A}$*, then any principal submatrix is hermitian and positive definite.*

Proposition 0.11.11. *Let* $\mathbf{A} \in \mathbb{H}_m(\mathbb{K})$*. Then* $0 \prec \mathbf{A}$ *if and only if all eigenvalues are positive.*

Lemma 0.11.1. *Let* $\mathbf{A} \in \mathbb{H}_m(\mathbb{K})$ *with* $0 \prec \mathbf{A}$*, i.e.,*

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & \mathbf{w}^* \\ \mathbf{w} & \mathbf{K} \end{bmatrix} \tag{89}$$

with $a_{1,1} > 0$*. Then, the Schur complement*

$$\mathbf{S} = \mathbf{K} - \frac{1}{a_{1,1}}\mathbf{w}\mathbf{w}^* \tag{90}$$

is positive definite.

Proof. Since $a_{1,1} > 0$ the Schur complement is well-define, and

$$\mathbf{S}^* = \left(\mathbf{K} - \frac{1}{a_{1,1}}\mathbf{w}\mathbf{w}^*\right)^* = \mathbf{K}^* - \frac{1}{a_{1,1}}\mathbf{w}\mathbf{w}^* = \mathbf{K} - \frac{1}{a_{1,1}}\mathbf{w}\mathbf{w}^* \tag{91}$$

Consider $\mathbf{y} \in \mathbb{K}^{m-1}$ with $\mathbf{y} \neq 0$ and define $x = -\frac{1}{a_{1,1}}\mathbf{w}^*\mathbf{y} \in \mathbb{K}$. Then

$$0 < \begin{bmatrix} x \\ \mathbf{y} \end{bmatrix}^* \mathbf{A} \begin{bmatrix} x \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} x \\ \mathbf{y} \end{bmatrix}^* \begin{bmatrix} a_{1,1}x + \mathbf{w}^*\mathbf{y} \\ x\mathbf{w} + \mathbf{K}\mathbf{y} \end{bmatrix} = \begin{bmatrix} x \\ \mathbf{y} \end{bmatrix}^* \begin{bmatrix} 0 \\ \mathbf{S}\mathbf{y} \end{bmatrix} = \mathbf{y}^*\mathbf{S}\mathbf{y} \tag{92}$$

Hence, $0 \prec \mathbf{S}$. □

Theorem 0.12. *Every hermitian and positive definite matrix has a unique Cholesky factorization.*

Algorithm 3 Cholesky factorization (without pivoting, naïve)

Require: $\mathbf{A} \in \mathbb{K}^{m \times m}$
Ensure: $\mathbf{R} \in \mathbb{K}^{m \times m}$ upper triangular s.t. $\mathbf{A} = \mathbf{R}^*\mathbf{R}$
  for k=1 to m-1 **do**
    $A[k+1:m, k+1:m] \leftarrow \mathbf{A}[k+1:m, k+1:m] - \frac{1}{\mathbf{A}[k,k]}\mathbf{A}[k+1:m, k]\mathbf{A}[k+1:m, k]^*$
    $A[k, k:m] \leftarrow A[k, k:m]/\sqrt{A[k,k]}$
  end for
  $A[m,m] \leftarrow A[m,m]/\sqrt{A[m,m]}$

Algorithm 4 pivoted Cholesky factorization

Require: $\mathbf{A} \in \mathbb{H}_m(\mathbb{R})$ and $0 \preccurlyeq \mathbf{A}$; $\varepsilon > 0$
Ensure: Low-rank approximation $\mathbf{A}_k = \sum_{i=1}^k \ell_i \ell_i^\top$ s.t. $\|\mathbf{A} - \mathbf{A}_k\|_1 \leqslant \varepsilon$
  $k \leftarrow 1$
  $\mathbf{d} \leftarrow \mathrm{diag}(\mathbf{A})$
  $\delta \leftarrow \|\mathbf{d}\|_1$
  $\pi = (1, 2, ..., m)$
  while $\delta > \varepsilon$ **do**
    $i \leftarrow \mathrm{argmax}\{\mathbf{d}[\pi_j] \mid j = k, k+1, ..., m\}$
    $\pi_k \leftrightarrow \pi_i$ (swap entries in the vector)
    $\ell_{k,\pi_k} \leftarrow \sqrt{\mathbf{d}[\pi_k]}$
    for $j = k+1$ to m **do**
      $\ell_{k,\pi_j} \leftarrow \mathbf{A}[\pi_k, \pi_j] - \sum_{p=1}^{k-1} \ell_{p,\pi_k}\ell_{p,\pi_j}/\ell_{k,\pi_k}$
      $\mathbf{d}[\pi_j] \leftarrow \mathbf{d}[\pi_j] - \ell_{k,\pi_j}^2$
    end for
    $\delta \leftarrow \sum_{j=k+1}^m \mathbf{d}[\pi_j]$
    $k \leftarrow k+1$
  end while

Definition 0.12.1. *Let* $\mathbf{M} \in \mathbb{K}^{m \times n}$ *and*

$$\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_k) \subseteq [\![m]\!] \quad \text{and} \quad \boldsymbol{\alpha}^c = [\![m]\!] \backslash \boldsymbol{\alpha} \tag{93}$$

and

$$\boldsymbol{\beta} = (\beta_1, ..., \beta_\ell) \subseteq [\![n]\!] \quad \text{and} \quad \boldsymbol{\beta}^c = [\![n]\!] \backslash \boldsymbol{\beta}. \tag{94}$$

We denote

$$\mathbf{M}[\boldsymbol{\gamma}, \boldsymbol{\delta}] \tag{95}$$

the $(\boldsymbol{\gamma}, \boldsymbol{\delta})$*-block in* $\mathbf{M}$.
The Schur complement of $\mathbf{M}[\boldsymbol{\alpha}, \boldsymbol{\beta}]$ *in* $\mathbf{M}$ *is*

$$\mathbf{M}/\mathbf{M}[\boldsymbol{\alpha}, \boldsymbol{\beta}] = \mathbf{M}[\boldsymbol{\alpha}^c, \boldsymbol{\beta}^c] - \mathbf{M}[\boldsymbol{\alpha}^c, \boldsymbol{\beta}]\left(\mathbf{M}[\boldsymbol{\alpha}, \boldsymbol{\beta}]\right)^\dagger \mathbf{M}[\boldsymbol{\alpha}, \boldsymbol{\beta}^c]. \tag{96}$$

Proposition 0.12.1. *Let* M *be a square matrix partitioned as*

$$\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}. \tag{97}$$

Let $\mathbf{A}$ *be nonsingular, then*

$$\det(\mathbf{M}/\mathbf{A}) = \det(\mathbf{M})/\det(\mathbf{A}). \tag{98}$$

Chapter 7

This chapter covers eigenvalue problems and basic algorithms to numerically approximate their solutions.

Definition 0.12.2. *Let $\mathbf{A} \in \mathbb{K}^{m\times m}$. We call the pair $(\lambda, \mathbf{v}) \in \mathbb{K} \times \mathbb{K}^m$ with $\mathbf{v} \neq \mathbf{0}$ an eigenpair if and only if*

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v} \tag{99}$$

We call λ the eigenvalue and $\mathbf{v}$ a to λ corresponding eigenvector. We call the set of all eigenvalues of $\mathbf{A}$, the spectrum of $\mathbf{A}$ denoted by $\Lambda(\mathbf{A})$.

Definition 0.12.3. *Let $\mathbf{A} \in \mathbb{K}^{m\times m}$. We call the decomposition*

$$\mathbf{A} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^{-1} \tag{100}$$

with $\mathbf{\Lambda} \in \mathbb{K}^{m\times m}$ diagonal and $\mathbf{X} \in \mathbb{K}^{m\times m}$ non-singular an eigenceomposition of $\mathbf{A}$.

Definition 0.12.4. *Let $\mathbf{A} \in \mathbb{K}^{m\times m}$, and let $\lambda \in \mathbb{K}$ be an eigenvalue of $\mathbf{A}$. We define*

$$E_\lambda = \{\mathbf{v} \in \mathbb{K}^m \mid \mathbf{A}\mathbf{v} = \lambda\mathbf{v}\} \tag{101}$$

as the eigenspace corresponding to λ. We call the dimension of E_λ the geometric multiplicity of λ.

Definition 0.12.5. *Let $\mathbf{A} \in \mathbb{K}^{m\times m}$. We call*

$$p_{\mathbf{A}}(z) = \det(z\mathbf{I} - \mathbf{A}) \tag{102}$$

the characteristic polynomial of $\mathbf{A}$.

Theorem 0.13. *The scalar $\lambda \in \mathbb{K}$ is an eigenvalue of $\mathbf{A}$ if and only if*

$$p_{\mathbf{A}}(\lambda) = 0. \tag{103}$$

Definition 0.13.1. *Let $\mathbf{X} \in \mathbb{K}^{m\times m}$ be non-singular, then we call $\mathbf{X}^{-1}\mathbf{A}\mathbf{X}$ the similarity transformed of $\mathbf{A}$ under $\mathbf{X}$. We call two matrices $\mathbf{A}, \mathbf{B} \in \mathbb{K}^{m\times m}$ similar if and only if there exists a non-singular matrix $\mathbf{X} \in \mathbb{K}^{m\times m}$ such that*

$$\mathbf{A} = \mathbf{X}^{-1}\mathbf{B}\mathbf{X} \tag{104}$$

Theorem 0.14. *Let $\mathbf{A} \in \mathbb{K}^{m\times m}$ and $\mathbf{X} \in \mathbb{K}^{m\times m}$ be non-singular. Then $\mathbf{A}$ and $\mathbf{X}^{-1}\mathbf{A}\mathbf{X}$ have the same characteristic polynomial, eigenvalues with the same algebraic and geometric multiplicity.*

Proof. We first note that

$$p_{\mathbf{X}^{-1}\mathbf{A}\mathbf{X}}(z) = \det(z\mathbf{I}-\mathbf{X}^{-1}\mathbf{A}\mathbf{X}) = \det(\mathbf{X}^{-1})\det(z\mathbf{I}-\mathbf{A})\det(\mathbf{X}) = \det(z\mathbf{I}-\mathbf{A}) = p_{\mathbf{A}}(z). \tag{105}$$

Hence, $\mathbf{A}$ and $\mathbf{X}^{-1}\mathbf{A}\mathbf{X}$ have the same characteristic polynomial, therewith the same eigenvalues at the same algebraic multiplicity. Next, we note that if E_λ is an eigenspace of $\mathbf{A}$, then $\mathbf{X}^{-1}E_\lambda$ is the corresponding eigenspace of $\mathbf{X}^{-1}\mathbf{A}\mathbf{X}$. Since $\mathbf{X}$ is non-singular

$$\dim(E_\lambda) = \dim(\mathbf{X}^{-1}E_\lambda) \tag{106}$$

□

Theorem 0.15. *Let* $\mathbf{A} \in \mathbb{K}^{m \times m}$*. The algebraic multiplicity of an eigenvalue* $\lambda \in \mathbb{K}$ *is at least as great as its geometric multiplicity.*

Proof. Let $\dim(E_\lambda) = n$. We then form a matrix $\hat{\mathbf{V}} = [\mathbf{v}_1|...|\mathbf{v}_n]$ whose columns are an orthonormal basis of E_λ and orthonormally extend it to $\mathbf{V} \in \mathbb{K}^{m \times m}$. This yields

$$\mathbf{B} = \mathbf{V}^*\mathbf{A}\mathbf{V} = \begin{bmatrix} \lambda\mathbf{I}_n & \mathbf{C} \\ \mathbf{0} & \mathbf{D} \end{bmatrix} \tag{107}$$

and therewith

$$\det(z\mathbf{I}_m - \mathbf{B}) = \det(z\mathbf{I}_n - \lambda\mathbf{I}_n)\det(z\mathbf{I}_{m-n} - \lambda\mathbf{D}) = (z-\lambda)^n\det(z\mathbf{I}_{m-n} - \lambda\mathbf{D}) \tag{108}$$

□

Definition 0.15.1. *We call an eigenvalue whose algebraic multiplicity supersedes its geometric multiplicity defective. A matrix that has one or more defective eigenvalues is called a defective matrix.*

Theorem 0.16. *A matrix* $\mathbf{A} \in \mathbb{K}^{m \times m}$ *is non-defective if and only if it has an eigendecomposition.*

Proof. First, assume $\mathbf{A} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^{-1}$. Since $\mathbf{\Lambda}$ is diagonal it is non-defective. Therefore $\mathbf{A}$ is non-defective by Theorem 0.14.
Second, we assume that $\mathbf{A}$ is non-defective. This in turn means that $\mathbf{A}$ has m linearly independent eigenvectors $\mathbf{v}_1, ..., \mathbf{v}_m$– note that eigenvectors to different eigenvalues are linearly independent. Defining $\mathbf{X} = [\mathbf{v}_1|...|\mathbf{v}_m]$ yields

$$\mathbf{A}\mathbf{X} = \mathbf{X}\mathbf{\Lambda} \Leftrightarrow \mathbf{A} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^{-1} \tag{109}$$

□

Definition 0.16.1. *Let* $\mathbf{A} \in \mathbb{K}^{m \times m}$*. We call* $\mathbf{A}$ *unitarily diagonalizable if and only if*

$$\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^* \tag{110}$$

where $\mathbf{Q} \in \mathbb{K}^{m \times m}$ *is unitary and* $\mathbf{\Lambda} \in \mathbb{K}^{m \times m}$ *is diagonal.*

Definition 0.16.2. *Let* $\mathbf{A} \in \mathbb{K}^{m \times m}$*. We say that* $\mathbf{A}$ *is normal if and only if*

$$\mathbf{A}^*\mathbf{A} = \mathbf{A}\mathbf{A}^*. \tag{111}$$

Definition 0.16.3. *Let* $\mathbf{A} \in \mathbb{K}^{m \times m}$*. We call the factorization*

$$\mathbf{A} = \mathbf{Q}\mathbf{T}\mathbf{Q}^* \tag{112}$$

where $\mathbf{Q} \in \mathbb{K}^{m \times m}$ *is unitary and* $\mathbf{T}$ *is upper triangular, a Schur factorization of* $\mathbf{A}$*.*

Theorem 0.17. *Every matrix* $\mathbf{A} \in \mathbb{C}^{m \times m}$ *has a Schur factorization*

Proof. We prove this by induction.

<u>$m = 1$</u>**:** The claim follows directly since

$$a = 1 \cdot a \cdot 1. \tag{113}$$

Induction hypothesis: Every matrix $\mathbf{A} \in \mathbb{C}^{m \times m}$ has a Schur factorization.

<u>$m \to m+1$</u>**:** Let $\mathbf{A} \in \mathbb{C}^{(m+1)\times(m+1)}$ and $(\lambda, \mathbf{v})$ be an eigenpair and let $\|\mathbf{v}\| = 1$. We then extend $\mathbf{v}$ unitarily to a basis which yields $\mathbf{U} = [\mathbf{v}|\mathbf{u}_2|...|\mathbf{u}_m] \in \mathbb{C}^{m \times m}$ unitary. This yields

$$\mathbf{U}^*\mathbf{A}\mathbf{U} = \begin{bmatrix} \lambda & \mathbf{B} \\ \mathbf{0} & \mathbf{C} \end{bmatrix} \tag{114}$$

By induction hypothesis, there exists a Schur factorization of $\mathbf{C}\mathbb{C}^{m \times m}$, i.e.,

$$\mathbf{C} = \mathbf{V}^*\mathbf{T}\mathbf{V}^*. \tag{115}$$

defining

$$\mathbf{Q} = \mathbf{U} \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{V} \end{bmatrix} \tag{116}$$

yields

$$\mathbf{Q}^*\mathbf{A}\mathbf{Q} = \begin{bmatrix} \lambda & \mathbf{B}\mathbf{V} \\ \mathbf{0} & \mathbf{T} \end{bmatrix} \tag{117}$$

□

Remark 0.17.1. *Above, we have seen three eigenvalue-revealing factorizations:*

1. $\mathbf{A} = \mathbf{X}\boldsymbol{\lambda}\mathbf{X}^{-1}$ *holds for non-defective matrices.*
2. $\mathbf{A} = \mathbf{Q}\boldsymbol{\lambda}\mathbf{Q}^*$ *holds for normal matrices.*
3. $\mathbf{A} = \mathbf{Q}\mathbf{T}\mathbf{Q}^*$ *holds for any matrix.*

Numerical approaches

Generally, build upon a two-phase procedure:

i) Bring the matrix close to an eigenvalue revealing factorization,i.e., upper Hessenberg form

ii) Apply various methods – depending on the problem – to compute the eigenvalue revealing factorization.

Definition 0.17.1. *Let* $\mathbf{A} \in \mathbb{H}_m(\mathbb{R})$*,* $\mathbf{x} \in \mathbb{R}^m$ *we call*

$$r(\mathbf{x}) = \frac{\mathbf{x}^\top\mathbf{A}\mathbf{x}}{\mathbf{x}^\top\mathbf{x}} \tag{118}$$

Theorem 0.18. *The pair* $(r(\mathbf{x}), \mathbf{x})$ *is an eigenpair of* $\mathbf{A} \in \mathbb{H}_m(\mathbb{R})$ *if and only if* $\mathbf{x}$ *is a stationary point of* $r(\cdot)$*.*

Proof. We compute the gradient

$$\frac{\partial}{\partial x_j} r(\mathbf{x}) = \frac{2(\mathbf{A}\mathbf{x})_j}{\mathbf{x}^\top\mathbf{x}} - \frac{(\mathbf{x}^\top\mathbf{A}\mathbf{x})2x_j}{(\mathbf{x}^\top\mathbf{x})^2} = \frac{2}{\mathbf{x}^\top\mathbf{x}}\left(\mathbf{A}\mathbf{x} - r(\mathbf{x})\mathbf{x}\right)_j. \tag{119}$$

Hence, if $(r(\mathbf{x}), \mathbf{x})$ is an eigenpair then $\nabla r(\mathbf{x}) = \mathbf{0}$ and conversely, $\nabla r(\mathbf{x}) = \mathbf{0}$ implies that

$$\mathbf{A}\mathbf{x} - r(\mathbf{x})\mathbf{x} = \mathbf{0} \Leftrightarrow \mathbf{A}\mathbf{x} = r(\mathbf{x})\mathbf{x} \tag{120}$$

□

Algorithm 5 Power method

Require: $\mathbf{A} \in \mathbb{H}_m(\mathbb{R})$
Ensure: $(\lambda, \mathbf{v})$ largest eigenpair
  $\mathbf{v}^{(0)} \leftarrow \mathbf{v}$ some vector with $\|\mathbf{v}\| = 1$
  for $k = 1, 2, \ldots$ **do**
    $\mathbf{w} \leftarrow \mathbf{A}\mathbf{v}^{(k-1)}$
    $\mathbf{v}^{(k)} \leftarrow \frac{\mathbf{w}}{\|\mathbf{w}\|}$
    $\lambda^{(k)} \leftarrow (\mathbf{v}^{(k)})^\top \mathbf{A}\mathbf{v}^{(k)}$
  end for

Theorem 0.19. *Suppose* $|\lambda_1| > |\lambda_2| \geqslant |\lambda_3| \geqslant \ldots \geqslant |\lambda_m| > 0$ *and* $q_1^\top v^{(0)} \neq 0$. *Then*

$$\|v^{(k)} - (\pm)q_1\| \in \mathcal{O}\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right) \quad \text{and} \quad |\lambda^{(k)} - \lambda_1| \in \mathcal{O}\left(\left|\frac{\lambda_2}{\lambda_1}\right|^{2k}\right) \tag{121}$$

Algorithm 6 Inverse Power method

Require: $\mathbf{A} \in \mathbb{H}_m(\mathbb{R})$, μ
Ensure: eigenpair $(\lambda_k, \mathbf{v})$ where $(\lambda_k - \mu)^{-1} > (\lambda_i - \mu)^{-1}$ for all $i \neq k$
  $\mathbf{v}^{(0)} \leftarrow \mathbf{v}$ some vector with $\|\mathbf{v}\| = 1$
  for $k = 1, 2, \ldots$ **do**
    Solve $(\mathbf{A} - \mu\mathbf{I})\mathbf{w} = \mathbf{v}^{(k-1)}$ for $\mathbf{w}$
    $\mathbf{v}^{(k)} \leftarrow \frac{\mathbf{w}}{\|\mathbf{w}\|}$
    $\lambda^{(k)} \leftarrow (\mathbf{v}^{(k)})^\top \mathbf{A}\mathbf{v}^{(k)}$
  end for

Theorem 0.20. *Let* $\mathbf{A} \in \mathbb{H}_m(\mathbb{R})$, *and* $k \in [\![m]\!]$ *and* $j \in [\![m]\!]$ *be such that*

$$(\lambda_k - \mu)^{-1} > (\lambda_j - \mu)^{-1} > (\lambda_i - \mu)^{-1} \tag{122}$$

for all $i \neq k, j$, *and let* $\mathbf{q}_k^\top \mathbf{v}^{(0)} \neq 0$. *Then the iterates of the inverse power method satifsy*

$$\|v^{(\ell)} - (\pm)\mathbf{q}_k\| \in \mathcal{O}\left(\left|\frac{\mu - \lambda_k}{\mu - \lambda_j}\right|^\ell\right) \quad \text{and} \quad |\lambda^{(\ell)} - \lambda_k| \in \mathcal{O}\left(\left|\frac{\mu - \lambda_k}{\mu - \lambda_j}\right|^{2\ell}\right) \tag{123}$$

Algorithm 7 Rayleigh quotient iteration

Require: $\mathbf{A} \in \mathbb{H}_m(\mathbb{R})$, μ
Ensure: eigenpair $(\lambda_k, \mathbf{v})$ where $(\lambda_k - \mu)^{-1} > (\lambda_i - \mu)^{-1}$ for all $i \neq k$
  $\mathbf{v}^{(0)} \leftarrow \mathbf{v}$ some vector with $\|\mathbf{v}\| = 1$
  $\lambda^{(0)} \leftarrow \mu$
  for $k = 1, 2, \ldots$ **do**
    Solve $(\mathbf{A} - \lambda^{(k-1)}\mathbf{I})\mathbf{w} = \mathbf{v}^{(k-1)}$ for $\mathbf{w}$
    $\mathbf{v}^{(k)} \leftarrow \frac{\mathbf{w}}{\|\mathbf{w}\|}$
    $\lambda^{(k)} \leftarrow (\mathbf{v}^{(k)})^\top \mathbf{A}\mathbf{v}^{(k)}$
  end for

Theorem 0.21. *The Rayleigh quotient iteration converges to an eigenpair for almost all starting vectors – meaning for all except a measure zero set. When it converges, the convergence is locally cubic, i.e.,*

$$\|\mathbf{v}^{(\ell+1)} - (\pm)\mathbf{q}_k\| \in \mathcal{O}\left(\|\mathbf{v}^{(\ell)} - (\pm)\mathbf{q}_k\|^3\right) \quad \text{and} \quad |\lambda^{(\ell)} - \lambda_k| \in \mathcal{O}\left(|\lambda^{(\ell)} - \lambda_k|^3\right) \tag{124}$$

Algorithm 8 Practical QR algorithm

Require: $\mathbf{A} \in \mathbb{H}_m(\mathbb{R})$
Ensure:
  $(\mathbf{Q}^{(0)})^\top \mathbf{A}^{(0)} \mathbf{Q}^{(0)} \leftarrow \mathbf{A}$ where $\mathbf{A}^{(0)}$ is upper Hessenberg matrix
  for $k = 1, 2, ...$ **do**
    $\mu^{(k)} \leftarrow \mathbf{A}^{k-1}_{m,m}$ Rayleigh shift
    $\mathbf{Q}^{(k)}\mathbf{R}^{(k)} = \mathbf{A}^{(k-1)} - \mu^{(k)}\mathbf{I}$
    $\mathbf{A}^{(k)} = \mathbf{R}^{(k)}\mathbf{Q}^{(k)} + \mu^{(k)}\mathbf{I}$

    If $|\mathbf{A}^{(k)}_{j,j+1}| < \varepsilon$ deflate the matrix

$$\mathbf{A}^{(k)} = \begin{bmatrix} \mathbf{A}_1 & 0 \\ 0 & \mathbf{A}_2 \end{bmatrix}$$

    and apply QR algorithm to $\mathbf{A}_1$,$\mathbf{A}_2$.
  end for

We have seen two shifts:

i) Rayleigh shift

ii) Wilkinson shift

0.2 Other algorithms

- Jacobi method
- Bisection method

Proposition 0.21.1. *Let* $\mathbf{A}$ *be tridiagonal with non-zero off-diagonal elements. Then the eigenvalues of each principle submatrix* $\mathbf{A}^k$ *of size* $k \times k$ *are distinct*

$$\lambda_1^{(k)} < \lambda_2^{(k)} < ... < \lambda_k^{(k)} \tag{125}$$

and the eigenvalues are strictly interlaced, i.e.,

$$\lambda_j^{(k+1)} < \lambda_j^{(k)} < \lambda_{j+1}^{(k+1)} \tag{126}$$

Sturm sequence:

$$1 \to \det(\mathbf{A}^{(1)}) \to \det(\mathbf{A}^{(2)}) \to ... \to \det(\mathbf{A}^{(m)}). \tag{127}$$

Chapter 8

This chapter was about the implementation of SVD computing algorithms.

1. Bidiagonalize $\mathbf{A} = \mathbf{UBV}^*$
   - Golub–Kahan (GK)
   - Lawson-Hanson-Chan (LHC)
2. Compute tridiagonalization of $\mathbf{B}$
   - diagonalize with bisection

Consider the matrix

$$\underset{\mathbf{A}}{\begin{bmatrix} * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \end{bmatrix}} \xrightarrow{\mathbf{U}_1^* \cdot} \underset{\mathbf{U}_1^*\mathbf{A}}{\begin{bmatrix} * & * & * & * \\ 0 & * & * & * \\ 0 & * & * & * \\ 0 & * & * & * \\ 0 & * & * & * \\ 0 & * & * & * \end{bmatrix}} \xrightarrow{\cdot \mathbf{V}_1} \underset{\mathbf{U}_1^*\mathbf{A}\mathbf{V}_1}{\begin{bmatrix} * & * & 0 & 0 \\ 0 & * & * & * \\ 0 & * & * & * \\ 0 & * & * & * \\ 0 & * & * & * \\ 0 & * & * & * \end{bmatrix}} \to \ldots \to \underset{\mathbf{U}_4^*\mathbf{U}_3^*\mathbf{U}_2^*\mathbf{U}_1^*\mathbf{A}\mathbf{V}_1\mathbf{V}_2}{\begin{bmatrix} * & * & 0 & 0 \\ 0 & * & * & 0 \\ 0 & 0 & * & * \\ 0 & 0 & 0 & * \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}} \tag{128}$$

Golub-Kahan:

$$\mathcal{O}(4mn^2 - \frac{4}{3}n^3) \tag{129}$$

Lawson-Hanson-Chan:

$$\mathcal{O}(2mn^2 + 2n^3) \tag{130}$$

Since the matrix is in bidiagonal form, i.e.,

$$\mathbf{B} = \begin{bmatrix} a_1 & b_1 & & \mathbf{0} \\ & a_2 & \ddots & \\ & & \ddots & b_{n-1} \\ \mathbf{0} & & & a_n \end{bmatrix} \tag{131}$$

a first and naïve approach would be to compute

$$\mathbf{B}^\top\mathbf{B} = \begin{bmatrix} a_1^2 & b_1 a_1 & & \\ b_1 a_1 & b_1^2 + a_2^2 & \ddots & \\ & \ddots & \ddots & b_{n-1}a_{n-1} \\ & & b_{n-1}a_{n-1} & b_{n-1}^2 + a_n^2 \end{bmatrix} \tag{132}$$

However, at its core, this is performing the product $\mathbf{A}^\top\mathbf{A}$ which squares the condition number. An alternative approach is to similarity transform the surrogate matrix

$$\begin{bmatrix} \mathbf{0} & \mathbf{B} \\ \mathbf{B}^\top & \mathbf{0} \end{bmatrix} \tag{133}$$

by the permutation

$$\mathbf{P} : \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ \vdots \\ 2n-1 \\ 2n \end{bmatrix} \mapsto \begin{bmatrix} n+1 \\ 1 \\ n+2 \\ 2 \\ \vdots \\ n+n \\ n \end{bmatrix} \tag{134}$$

which yields

$$\mathbf{S} = \mathbf{P} \begin{bmatrix} \mathbf{0} & \mathbf{B} \\ \mathbf{B}^\top & \mathbf{0} \end{bmatrix} \mathbf{P}^\top = \begin{bmatrix} 0 & a_1 & & & & & \\ a_1 & 0 & b_1 & & & & \\ & b_1 & 0 & a_2 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & a_{n-2} & 0 & b_{n-1} & \\ & & & & b_{n-1} & 0 & a_n \\ & & & & & a_n & 0 \end{bmatrix} \tag{135}$$

We may then compute the eigenvalues of $\mathbf{S}$ which correspond to the singular values of $\pm\sigma(\mathbf{B})$.

Chapter 9

We here focused on Kyrlov subspace methods, in particular the CG method.
Depending on the problem different names arise

	$Ax = b$	$Ax = \lambda x$
$A = A^*$	CG	Lanczos
$A \neq A^*$	GMRES CGN BCG et al.	Arnoldi

We derive and investigate the CG algorithm. To that end, consider the *quadratic test function*:
Let $0 \prec \mathbf{A} \in \mathbb{H}_n(\mathbb{R})$ and $\mathbf{b}, \mathbf{x} \in \mathbb{R}^n$

$$\phi(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{x}^\top \mathbf{b}$$

The gradient of ϕ is given by

$$\nabla\phi(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}$$

Hence, at the critical point $\mathbf{x}_*$ we have

$$\nabla\phi(\mathbf{x}_*) = 0 \quad \Leftrightarrow \quad \mathbf{A}\mathbf{x}_* = \mathbf{b}$$

This critical point is unique!

i) Note that

$$\nabla^2\phi(\mathbf{x}) = \mathbf{A} \succ \mathbf{0}$$

$\Rightarrow \mathbf{x}_*$ is a minimum

ii) $\nabla^2\phi(\mathbf{x})$ is constant $\Rightarrow \phi$ is convex.

Numerically, we can find the minimum using a linesearch method, i.e., an iterative optimization method. We start with an initial guess $\mathbf{x}_0$
Update as

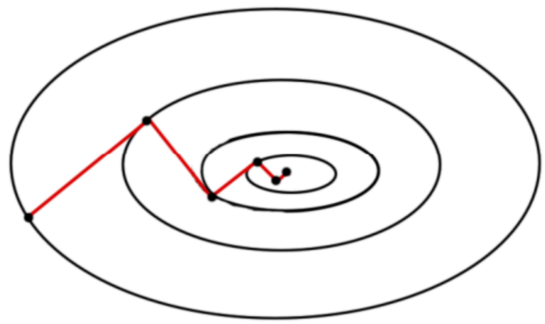
$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k$$

where $\mathbf{p}_k$ is the *search direction* and α_k is the *step length* Remember

$$\nabla\phi(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}$$

points towards largest increase of ϕ in $\mathbf{x}$.
$\Rightarrow$ Search direction should be $\mathbf{p}_k = -\nabla\phi(\mathbf{x}_k) = \mathbf{r}(\mathbf{x}_k)$

What about the step length?



Idea: Walk until we no longer descend!

$$0 \overset{!}{=} \partial_{\alpha_k} \phi(\mathbf{x}_{k+1}) \quad \Rightarrow \quad \alpha_k = \frac{\mathbf{r}_k^\top \mathbf{r}_k}{\mathbf{r}_k^\top \mathbf{A} \mathbf{r}_k}$$

We say a set of vectors $\{\mathbf{p}_1, ..., \mathbf{p}_k\}$ are conjugate w.r.t. the SPD matrix $\mathbf{A}$ iff

$$\mathbf{p}_i^\top \mathbf{A} \mathbf{p}_j = 0 \quad \forall i \neq j$$

Claim: n $\mathbf{A}$-conjugate vectors form a basis of $\mathbb{R}^n$. Then

$$\mathbf{x}_* = \sum_{i=1}^{n} c_i \mathbf{p}_i \quad \Rightarrow \quad \mathbf{A}\mathbf{x}_* = \sum_{i=1}^{n} c_i \mathbf{A} \mathbf{p}_i$$

hence

$$\mathbf{p}_k^\top \mathbf{b} = \sum_{i=0}^{n-1} c_i \mathbf{p}_k^\top \mathbf{A} \mathbf{p}_i = c_k \mathbf{p}_k^\top \mathbf{A} \mathbf{p}_k \quad \Rightarrow \quad c_k = \frac{\mathbf{p}_k^\top \mathbf{b}}{\mathbf{p}_k^\top \mathbf{A} \mathbf{p}_k}$$

If we have sequence of $\mathbf{A}$-conjugate vecorts we can solve for $\mathbf{x}_*$

Zeroth Iteration:

- We start with $\mathbf{x}_0 = \mathbf{0} \in \mathbb{R}^n$
- Compute the residual

$$\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$$

- Compute the search direction

$$\mathbf{p}_0 = -\nabla \phi(\mathbf{x}_0) = \mathbf{r}_0$$

- Compute the step length

$$\alpha_0 = \frac{\mathbf{p}_0^\top \mathbf{r}_0}{\mathbf{p}_0^\top \mathbf{A} \mathbf{p}_0}$$

- Update the iterate

$$\mathbf{x}_1 = \mathbf{x}_0 + \alpha_0 \mathbf{p}_0$$

kth iteration:

- Compute the residual

$$\mathbf{r}_k = \mathbf{b} - \mathbf{A}\mathbf{x}_k = -\nabla \phi(\mathbf{x}_k)$$

- Make the gradient conjugate to the previous $\{\mathbf{p}_0, ..., \mathbf{p}_{k-1}\}$

$$\mathbf{p}_k = \mathbf{r}_k - \sum_{i=0}^{k-1} \frac{\mathbf{p}_i^\top \mathbf{A}\mathbf{r}_k}{\mathbf{p}_i^\top \mathbf{A}\mathbf{p}_i}\mathbf{p}_i$$

- Compute the step length

$$\alpha_k = \frac{\mathbf{p}_k^\top \mathbf{r}_k}{\mathbf{p}_k^\top \mathbf{A}\mathbf{p}_k}$$

- Update the iterate

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k$$

Why is this a Krylov subspace method?

- Claim: $\mathbf{x}_k \in \mathcal{K}_k = \text{span}(\mathbf{b}, \mathbf{A}\mathbf{b}, ..., \mathbf{A}^{k-1}\mathbf{b})$
- Claim: $\mathbf{r}_k \perp \mathcal{K}_k$
- Theorem:
  $\mathbf{x}_k \in \mathcal{K}_k$ is the unique point that minimizes $\|\mathbf{e}_k\|_A$ with $\mathbf{e}_k = \mathbf{x}_* - \mathbf{x}_k$ and $\|\mathbf{e}_k\|_{\mathbf{A}} \leqslant \|\mathbf{e}_{k-1}\|_{\mathbf{A}}$ and $\mathbf{e}_\ell = 0$ for some $\ell \geqslant n$.