UNIVERSITY OF WISCONSIN-MADISON
COMPUTER SCIENCES DEPARTMENT

Class notes for Math/CS 885: Matrix theory for Numerical Analysis

Spring 1989

Author's address:
Center for the Mathematical Sciences
University of Wisconsin-Madison
610 Walnut St.
Madison WI 53705

notes for Jan.23ff latest change: Mar.4

linear spaces S, T; $A \in L(S,T) :=$ linear map $A : S \to T$;
$S =: \mathrm{dom}(A) =:$ domain of A, $T =: \mathrm{tar}(A) =:$ target of A.
task: given $b \in T$, find $x \in S$ such that $Ax = b$.
at least one solution iff $b \in \mathrm{ran}(A) :=$ range of $A := A(S)$;
at least one solution for every $b \in T$ iff $\mathrm{ran}(A) = T$ (existence iff A onto)
at most one solution iff $0 = \ker(A) :=$ null space of $A := \{x \in S : Ax = 0\}$; (uniqueness iff A 1-1)
exactly one solution for every $b \in T$ iff A is invertible.

---

two ways to connect linear space S linearly to coordinate space $\mathbb{R}^n$:
$V : \mathbb{R}^n \to S$ and $\Lambda : S \to \mathbb{R}^n$

$[v_1, \dots, v_n] := V : \mathbb{R}^n \to S : c \mapsto \sum_{j=0}^{n} v_j c(j)$ with (v_i) an n-sequence in S is most general lm from $\mathbb{R}^n$ to S. If $V \in L(\mathbb{R}^n, S)$ is given, then $v_j = Ve_j$, all j, with e_j the jth **unit vector**.

In this course, any ls S will be a linear subspace of some coordinate space, hence the v_i will be (column) vectors and the map V is just the matrix $[v_1, \dots, v_n]$ with columns $v_1, \dots, v_n$.

special case: $L(\mathbb{R}^n, \mathbb{R}^m) = \mathbb{R}^{m \times n}$

special fact: $m < n$ implies that every $X \in \mathbb{R}^{m \times n}$ has nontrivial nullspace.

special matrix: $I_n := [e_1, \dots, e_n]$ is the nth order identity or unit matrix.

Note: if $[v_1, \dots, v_n] = V \in L(\mathbb{R}^n, S)$ and $A \in L(S,T)$, then $AV = A[v_1, \dots, v_n] = [Av_1, \dots, Av_n] \in L(\mathbb{R}^n, T)$.

elements of $\mathrm{ran}(V)$ are called 'linear combinations of the v_i' (note that I will always write the weights to the **right** of the v_j (WHY?)), and $\mathrm{ran}(V)$ is called the span of the v_i, written $\mathrm{span}(v_i)_1^n := \mathrm{ran}(V)$

won't distinguish between the sequence $(v_1, \dots, v_n)$ and the map $[v_1, \dots, v_n]$

The sequence V is **spanning** (for S) iff the map V is onto

(in that case, call S **finite-dimensional**. We only deal with finite-dimensional spaces)

The sequence V is **linearly independent** iff the map V is 1-1

The sequence V is **basis** (for S) iff the map V is invertible

Fact 1. V not onto implies existence of $w \in S \setminus \mathrm{ran}(V)$ and, for any such w, $[V, w]$ is 1-1 in case V is 1-1 . (proof: $[V, w](c; d) = 0$ implies $-wd \in \mathrm{ran}(V)$, hence, as $w \notin \mathrm{ran}(V)$, $d = 0$, so $Vc = 0$, so $c = 0$ in case V is 1-1.)

Fact 2. V not 1-1 implies that $V = WX$ for some $X \in L(\mathbb{R}^n, \mathbb{R}^{n-1})$, hence if V is onto, then $W : \mathbb{R}^{n-1} \to S$ is also onto. (proof: $c \in \ker(V) \setminus 0$ implies that, for some j, $v_j = Wc$ with $W := V \setminus v_j := [v_1, \dots, v_{j-1}, v_{j+1}, \dots, v_n]$, hence $V = W[e_1, \dots, e_{j-1}, c, e_j, \dots, e_{n-1}] =: WX$.

Theorem: $V \in L(\mathbb{R}^n, S)$ 1-1 and $W \in L(\mathbb{R}^m, S)$ onto implies $n \le m$ with equality iff V is also onto and W is also 1-1.

(proof of implication: W onto implies that, for each v_j, can find $x_j \in \mathbb{R}^m$ so that $v_j = Wx_j$; hence $V = WX$ for some $X \in L(\mathbb{R}^m, \mathbb{R}^n)$. With this, $n > m$ would imply that $\ker(X)$ not trivial, hence $\ker(V)$ not trivial, a contradiction.

proof of equality characterization: This implies that if also V is onto and W 1-1, then also $m \le n$, i.e., equality. It also implies with Fact 1 that, if V is not onto, then $n + 1 \le m$. It also implies with Fact 2 that, if W is not 1-1, then $n \le m - 1$.)

Cor. Any two bases for S have the same cardinality, called the **dimension** of S and written

$$\dim S := \#V := \text{number of terms in the sequence } V$$

Cor. If T and U are l.subspaces of S and $\dim T + \dim U > \dim S$, then $T \cap U \neq 0$.

(proof: Let V, W be bases for T, U resp. and assume that $T \cap U = 0$. Then $[V, W](c; d) = 0$ implies $Vc = -Wd$, hence $Vc, Wd \in \mathrm{ran}(V) \cap \mathrm{ran}(W) = 0$, hence $Vc = 0 = Wd$, hence $c = 0$, $s = 0$. This proves that $[V, W]$ is 1-1, therefore $\#V + \#W \le \dim S$.)

Cor (with Fact 1). Any lin.ind. sequence can be extended to a basis.

Cor. Any linear subspace of a finite-dimensional ls has a basis.

Cor (with Fact 2). Any spanning sequence can be reduced to a basis.

Theorem. $\dim \operatorname{ran}(A) + \dim \ker(A) = \dim \operatorname{dom}(A)$.

(proof: extend basis K for $\ker(A)$ to a basis $V = [K, L]$ for $\operatorname{dom}(A)$. Sufficient to prove: $\dim \operatorname{ran}(A) = \#L$. For this, consider $AL : \mathbb{R}^{\#L} \to \operatorname{ran}(A)$. $\operatorname{ran}(A) = \operatorname{ran}(AV) = \operatorname{ran}(AL)$ since $AK = 0$, i.e., AL is onto. If $ALc = 0$, then $Lc \in \ker(A)$, hence $Lc = Kd$ for some d, therefore $V[-d; c] = -Kd + Lc = 0$, hence, in particular, $c = 0$; in short, AL is also 1-1, therefore a basis for $\operatorname{ran}(A)$.)

Cor. $\dim \operatorname{ran}(A) \le \min\{\dim \operatorname{tar}(A), \dim \operatorname{dom}(A)\}$.

Cor. If $A \in \mathbb{R}^{n\times n}$, then A is 1-1 iff A is onto (since now $\dim \operatorname{tar}(A) = \dim \operatorname{dom}(A)$).

V invertible; then V^{-1} carries $x \in S$ to its **coordinate vector** $V^{-1}x$ wrto the basis V.

$[\lambda_1; \ldots; \lambda_m] := \Lambda : S \to \mathbb{R}^m : s \mapsto (\lambda_i s)_{i=0}^m$ with (λ_i) a sequence of linear functionals on S is the most general element of $L(S, \mathbb{R}^m)$. For given Λ, have $\lambda_i : S \to \mathbb{R} : x \mapsto (Lx)(i)$, all i.

Won't distinguish between the sequence (λ_i) and the linear map Λ.

If $S = \mathbb{R}^n$, then each $\lambda_i \in L(S, \mathbb{R}) = \mathbb{R}^{1\times n}$ is a row matrix (or, row vector), i.e.,

$$\lambda_i x = w_i^T x := \sum_j w_i(j)x(j), \qquad \text{all } x,$$

for some $w_i \in \mathbb{R}^n$. Correspondingly, $\Lambda \in L(\mathbb{R}^n, \mathbb{R}^m) = \mathbb{R}^{m\times n}$ is a matrix, and w_i^T is the ith row of that matrix. Hence, this point of view stresses the rows of matrices, while the earlier one stresses the columns.

In this course, S will always be a linear subspace of some $\mathbb{R}^n$, hence its **dual** $S' := L(S, \mathbb{R})$ can be described by restricting $(\mathbb{R}^n)'$ to S. This means that we can think of each $\lambda \in S'$ as given by some w^T, as is argued in the following.

Fact. If $S \subseteq \mathbb{R}^n$, then, for every $\lambda \in S'$, can find $w \in \mathbb{R}^n$ such that $\lambda x = w^T x$ for all $x \in S$.

(proof. Take V a basis for S, extend it to a basis $B := [V, L]$ for $\mathbb{R}^m$. Then $P := \operatorname{diag}(V, 0)B^{-1}$ projects $\mathbb{R}^m$ onto S in the sense that $\operatorname{ran}(P) = S$ and $P_{|S} = 1$. Consequently, $\lambda P \in L(\mathbb{R}^m, \mathbb{R})$, i.e., $\lambda P = w^T$ for some $w \in \mathbb{R}^m$, and $w_{|S}^T = \lambda P_{|S} = \lambda$.)

Note that there is nothing unique about the w; different choices of L result in different w; but all these w's agree on S.

Thus, the most general linear map from S to $\mathbb{R}^m$ is of the form $\Lambda = (W^T)_{|S}$ for some $W : \mathbb{R}^n \to \mathbb{R}^m$, but if S is a proper subspace of $\mathbb{R}^m$, then different W may give rise to the same map Λ.

Example: $\Lambda = V^{-1}$, with V a basis for S. The corresponding λ_i are the **coordinate functionals** for the basis V.

In order to compute with a linear map $A \in L(S, T)$, we factor it through some coordinate space, i.e., write it as

$$A = VW^T$$

with some $W^T \in L(S, \mathbb{R}^n)$ and $V \in L(\mathbb{R}^n, T)$.

example: take for $V : \mathbb{R}^r \to \operatorname{ran}(A)$ a basis for $\operatorname{ran}(A)$, then $W^T := V^{-1}A : S \to \mathbb{R}^r$ does the job.

More explicitly,

$$A = VW^T = \sum_{j=1}^n v_j w_j^T$$

hence

$$Ax = \sum_j v_j(w_j^T x).$$

How large can n be? Arbitrarily large. From computational point of view, would like n as small as possible. How small can n be? There's a lower limit. Smallest possible n is called the **rank** of A, written $\operatorname{rank}(A)$.

By example, $\operatorname{rank}(A) \le r := \dim \operatorname{ran}(A)$, while necessarily $\operatorname{ran}(V) \supset \operatorname{ran}(A)$, hence $\operatorname{rank}(A) \ge \dim \operatorname{ran}(A)$. Therefore,

$$\operatorname{rank}(A) = \dim \operatorname{ran}(A).$$

Call the factorization/representation VW^T for A **minimal** if $n := \#V = \text{rank}(A)$. If $n = \text{rank}(A)$, then $V : \mathbb{R}^n \to \text{ran}(A)$ necessarily invertible, i.e., V is a basis for $\text{ran}(A)$. Also, $\ker(W^T) = \ker(A)$, since V is 1-1, hence $Ax = 0$ iff $W^T x = 0$.

Since $A = VW^T$ implies $A^T = WV^T$, also $\text{rank}(A) = \text{rank}(A^T) = \dim \text{ran}(A^T)$. Therefore, if $A = VW^T = \sum_{j=1}^{n} v_j w_j^T$ is minimal (i.e., $n = \text{rank}(A)$), then V is a basis for $\text{ran}(A)$ and W is a basis for $\text{ran}(A^T)$, hence $\ker(A) = \text{ran}(A^T)\perp$.

Even if we insist on a minimal factorization, there are still many choices suitable for specific jobs. The best known ones are:

LU or (PL)U (from Gauss elimination, useful for solving $Ax = b$)

QR (from Gram-Schmidt, useful for solving $A^T Ax = A^T b$ (least squares) and for eigenvalue computations)

SVD (from eigenvalue calculations, useful for understanding **robustness** or **condition** of the problem $Ax = b$ and others)

Their numerical construction uses induction (recursion) and **elementary matrices**. Their justification and analysis requires norms, particularly the 2-norm.

A more general factorization for $A \in L(S,T)$ takes the form $V\hat{A}W^T$, which allows for some conditions to be put on W and V.

For example, the factorization $A = V\hat{A}W^{-1}$ with both V and W bases provides a **matrix representation** for A. Note that W^{-1} carries $x \in S$ to its coordinates wrto W, $\hat{A}$ operates on this coordinate vector, and V maps the resulting vector to T. It follows that the matrix $\hat{A}$ is given by $\hat{A} = V^{-1}AW$ showing that the jth column of $\hat{A}$ contains the coordinates of Aw_j wrto V. One says that A and $\hat{A}$ are **equivalent**. If one further insists on having $\hat{A}$ very simple, then one can make it simple indeed, as we saw earlier, viz., of the form $\mathrm{diag}(I_{\mathrm{rank}(A)}, O)$, with the matrix O the zero matrix of the appropriate size.

When $S = T$ and one insists on having $V = W$, it is much harder to keep $\hat{A}$ simple. Now $\hat{A}$ is called **similar** to A. If $\hat{A}$ happens to be diagonal, then all its diagonal entries are **eigenvalues** of A and the columns of V are the corresponding **eigenvectors** in that then $Av_j = \hat{A}(j,j)v_j$, all j (with each v_j not trivial since it is part of a basis). If we can find such V and $\hat{A}$, then A is called **diagonable**. While the diagonable matrices are dense, not all matrices are diagonable. In general, the best in the way of simplicity one can achieve is an upper triangular $\hat{A}$ (the **Schur** form) if one actually wants to construct it stably. Fortunately, this is sufficient for all practical purposes. If one is just proving theorems, then the **Jordan** canonical form is simpler yet. We return to these factorizations soon, when we discuss the computation of eigenvalues and -vectors.

Finally, the factorization $A = V\hat{A}V^T$ is important in the study of the **quadratic form** $x \mapsto x^TAx$ associated with A (or, more precisely, with the symmetric matrix $(A + A^T)/2$). Such $\hat{A}$ is said to be **congruent** to A. For a symmetric A , $\hat{A}$ can be chosen to be diagonal.

---

quick review of norms (sect.2.1 and 2.2) I assume that you have dealt with normed linear spaces and the associated map norm

$$\|A\| := \|A\|_{S,T} := \sup_{x \in S} \|Ax\|_T / \|x\|_S$$

for $A \in L(S,T)$, and are familiar with the fact that $\|A^T\|_{T',S'} = \|A\|_{S,T}$, where, e.g., S' is the dual for S equipped with the dual norm, i.e., $\|\lambda\|_{S'} := \|\lambda\|_{S,\mathbb{R}}$.

In particular, for $x \in S$, $\|x\| = \sup_{\lambda \in S'} (\lambda x)/\|\lambda\|_{S'}$. For example, for the p-norm

$$\|x\|_p := \big(\sum |x(j)|^p\big)^{1/p}$$

including the max-norm

$$\|x\|_\infty := \max_j |x(j)| = \lim_{p \to \infty} \|x\|_p$$

for $x \in \mathbb{R}^n$, the dual norm is the p'-norm with $1/p + 1/p' := 1$.

Cor. *The rank-one map $[y]\lambda : S \to T : x \mapsto y(\lambda x)$ (with y a fixed element of T and λ a fixed element of S') has map norm $\|[y]\lambda\| = \|y\|_T \|\lambda\|_{S'}$.*

Only the three norms, $\|\cdot\|_1, \|\cdot\|_2, \|\cdot\|_\infty$ are of practical interest. The first and last, because the corresponding map norms

$$\|A\|_p = \sup \|Ax\|_p / \|x\|_p = \begin{cases} \max_j \|A(\cdot, j)\|_1, & p = 1; \\ \max_i \|A(i,\cdot)\|_1, & p = \infty; \end{cases}$$

can actually be computed in finitely many arithmetic operations, the middle one because it is the **Euclidean norm**, the norm associated with the scalar product:

$$\|x\|_2^2 = x^T x.$$

For numerical analysts, this norm is so important because there are many linear **isometries** for it, i.e., matrices which preserve this norm, viz. the **orthogonal** matrices, i.e ., any matrix U with $U^TU = I$. For such a matrix and any x,

$$\|Ux\|_2^2 = (Ux)^T Ux = x^T U^T Ux = x^T x = \|x\|_2^2.$$

This implies that any orthogonal matrix is 1-1 (hence a basis for its range), hence invertible, with inverse $U^{-1} = U^T$, in case U is square. It also implies that

$$\|VAW^T\|_2 = \|A\|_2$$

in case both V and W are orthogonal matrices. One example of the plenitude of orthogonal matrices is the

Fact. *Every orthogonal matrix V can be extended to an orthogonal basis for any linear subspace of $\mathbb{R}^n$ containing* $\mathrm{ran}(V)$. *In particular, every linear subspace S of $\mathbb{R}^n$ has an orthogonal basis.*

For the proof, observe that $P := VV^T$ is a projector since $PP = VV^TVV^T = VV^T = P$. It is the **orthogonal projector** for $\mathrm{ran}(V)$, socalled since, for any projector, $P(1-P) = P - P^2 = P - P = 0$, hence, for this projector and for any x, $0 = P(1-P)x = VV^T(x - Px)$, therefore $V^T(x - Px) = 0$ (since V is 1-1), hence $\mathrm{ran}(V) \perp x - Px$. Hence, if $x \in S \backslash \mathrm{ran}(V)$, then $\mathrm{ran}(V) \perp x - Px \neq 0$, so $y := (x - Px)/\|x - Px\|$ is well-defined and $y \perp \mathrm{ran}(V)$, therefore $[V, y]$ is again orthogonal, and, in particular, 1-1. Thus repetition of this process must end exactly when a basis for S is reached.

The numerical process based on this construction is called **Gram-Schmidt**. We will meet numerically preferable algorithms later.

We use the following lemma in the proof of the existence of SVD later.

(1)Lemma. *If $\alpha := A(i,j) = \|A\|_2$, then $Ae_j = \alpha e_j$ and $e_i^T A = \alpha e_i^T$.*

Proof: With $x := A(i,\cdot)^T$, compute

$$\alpha\|x\|_2 = \|A\|_2\|x\|_2 \geq \|Ax\|_2 \geq x^Tx = \|x\|_2^2$$

to find that $\alpha \geq \|x\| = \sqrt{\alpha^2 + \sum_{r\neq i} |A(i,r)|^2}$, hence $A(i,r) = 0$ for $r \neq i$. Use $\|A^T\|_2 = \|A\|_2$ to get that also $A(s,j) = 0$ for $s \neq j$. □

Numerical analysts also use norms on matrices which are, offhand, not map norms. The best known are again the p-norms applied to $A \in \mathbb{R}^{m\times n}$ as a function on $[1,\ldots,m] \times [1,\ldots,n]$. Of these, only the case $p = 2$ is of interest here, the socalled **Frobenius** norm:

$$\|A\|_F := \sqrt{\sum_{i,j} |A(i,j)|^2}.$$

This norm is **consistent with** the 2-norm for vectors in the sense that still

$$\|Ax\|_2 \leq \|A\|_F\|x\|_2$$

(for the simple reason that $\|A\|_2 \leq \|A\|_F$; hence it provides a *computable* upper bound for $\|A\|_2$). Also $\|UAW^T\|_F = \|A\|_F$ for all orthogonal U and W.

---

existence of the SVD

Theorem. *For every $A \in \mathbb{R}^{m\times n}$, can find* **square** *orthogonal V and W so that $A = V\Sigma W^T$ with $\Sigma = \mathrm{diag}(\sigma_1,\ldots,\sigma_p) \in \mathbb{R}^{m\times n}$ and $\sigma_1 \geq \ldots \geq \sigma_p \geq 0$, and $p := \min\{m,n\}$.*

Remark. Note that Σ, though a diagonal matrix by definition, is diagonal in the general sense that its only non-zero entries lie on its 'diagonal'. In particular, the use of the construct diag ... here and in the book does not coincide with its use in MATLAB.

Proof: Since the closed unit ball (in whatever norm) of $\mathbb{R}^n$ is compact, can find unit vector x at which A **takes on its norm**, i.e., for which $Ax = \sigma y$ with $\sigma := \|A\|_2$ and $\|y\|_2 = 1$. Extend x to an orthogonal basis W for $\mathbb{R}^n$ and y to an orthogonal basis V for $\mathbb{R}^m$ and let $\hat{A} := V^TAW$. Then, in particular, $\|\hat{A}\|_2 = \|A\|_2 = \hat{A}(1,1)$ (since $Aw_1 = Ax = \|A\|_2 y = \|A\|_2 v_1$), hence, by (1)Lemma,

$$\hat{A} = \mathrm{diag}(\sigma, B),$$

and induction finishes the proof (since, necessarily, $\|B\|_2 \leq \|A\|_2$; in fact, $\|B\|_2 = \sup_{x(1)=0} \|\hat{A}x\|_2/\|x\|_2 \leq \sup_x \|\hat{A}x\|_2/\|x\|_2 = \|A\|_2$.) □

Cor (i). *Let* $\sigma_r > 0 = \sigma_{r+1}$. *Then* $\operatorname{rank}(A) = r$.

Proof: Then

$$A = A_r := \sum_{j \le r} v_j \sigma_j w_j^T$$

with both $[v_1\sigma_1, \ldots, v_r\sigma_r]$ and $[w_1, \ldots, w_r]$ 1-1, hence A_r provides a minimal factorization for A.

Cor (ii). *If* $A = V\Sigma W^T$ *is a SVD decomposition for* A *and* $\operatorname{rank}(A) \ge k$, *then* $\min\{\|A - B\|_2 : \operatorname{rank}(B) \le k\} = \sigma_{k+1} = \|A - A_k\|_2$.

Proof: By assumption, $\operatorname{rank}(A_k) = k$, and

$$\|A - A_k\|_2 = \|V^T(A - A_k)W\|_2 = \|\operatorname{diag}(0, \ldots, 0, \sigma_{k+1}, \ldots, \sigma_p)\|_2 = \sigma_{k+1}.$$

Also, with $\operatorname{rank}(B) \le k$, $\dim\ker(B) \ge n - k$, hence $\ker(B) \cap \operatorname{ran}([w_1, \ldots, w_{k+1}])$ contains a unit vector z. Then $Bz = 0$ while $Az = \sum_{j \le k+1} v_j \sigma_j (w_j^T z)$, so

$$\|A - B\|_2^2 \ge \|Az - Bz\|_2^2 = \|Az\|_2^2 = \sum_{j \le k+1} \sigma_j^2 (w_j^T z)^2 \ge \sigma_{k+1}.$$

□

This implies that the **singular values** $\sigma_j = \sigma_j(A)$ are independent of the particular choice of V and W.

Since

$$A^T A = (V\Sigma W^T)^T V\Sigma W^T = W\Sigma^T\Sigma W^T,$$

we see that $\Sigma^T\Sigma$ is similar to A^TA, hence $(\sigma_j)^2$ are the eigenvalues of A^TA, another way of seeing that the σ_j do not depend on the particular V and W. On the other hand, this shows that W must consist of **eigenvectors** for A^TA, specifically that w_j must be an eigenvector for A^TA belonging to the eigenvalue $\sigma_j(A)^2$. The analogous argument shows that v_j must be an eigenvector for AA^T belonging to the eigenvalue $\sigma_j(A)^2$.

Thus, while the SVD is strictly speaking, not unique, it is unique (up to signs) in case all the singular values are distinct.

With $r := \operatorname{rank}(A)$, $A = A_r = \sum_{j \le r} v_j \sigma_j w_j^T$. Hence $[(v_j)_{j \le r}]$ is an orthogonal basis for $\operatorname{ran}(A)$, $[(w_j)_{j \le r}]$ is an orthogonal basis for $\operatorname{ran}(A^T)$, $[(v_j)_{j > r}]$ is an orthogonal basis for $\ker(A^T)$, and $[(w_j)_{j > r}]$ is an orthogonal basis for $\ker(A)$. Also,

$$\|A\|_2^2 = \sigma_1(A)^2 = \rho(A^TA) = \rho(AA^T) = \|A^TA\|_2 = \|AA^T\|_2.$$

Approximate inverse

After computing a(n approximate) solution $\tilde{x}$ for the equation $Ax = b$ (with $b \in T$ and $A \in L(S,T)$ given), one judges its quality by looking at the **residual** $r := b - A\tilde{x}$, for that is, off-hand, the only thing one can compute. Since the **error** $e := x - \tilde{x}$ is related to the residual by the equation $Ae = r$, we need a **lower bound** for A in order to obtain a bound on e from r, i.e., some lower bound for the number $\inf_x \|Ax\|/\|x\|$.

Note that, for *invertible* A,

$$\inf_x \|Ax\|/\|x\| = \inf_y \|AA^{-1}y\|/\|A^{-1}y\| = 1/(\sup_y \|A^{-1}y\|/\|y\|) = 1/\|A^{-1}\|,$$

i.e., $\|Ax\| \geq \|x\|/\|A^{-1}\|$ is the sharpest bound available. Unfortunately, it seems to require knowledge of A^{-1}.

There is only one practical way to obtain such a lower bound, viz. via an **approximate inverse** for A.

Call $C \in L(T,S)$ an **approximate inverse** for $A \in L(S,T)$ in case $\|1 - CA\| < 1$.

With

$$E := 1 - CA,$$

an approximate inverse supplies a **lower** bound for A: Since $CAx = x - Ex$, one gets

$$\|C\|\|Ax\| \geq \|CAx\| \geq \|x\| - \|Ex\| \geq (1 - \|E\|)\|x\|,$$

or

$$\|Ax\| \geq \frac{1 - \|E\|}{\|C\|}\|x\|.$$

In particular, $\|x - \tilde{x}\| \leq \frac{\|C\|}{1-\|E\|}\|r\|$.

Lemma. *If the matrix A has C as an approximate inverse, then A is* **bounded below**, *i.e.,* $\inf_x \|Ax\|/\|x\| > 0$, *therefore also A is 1-1. Hence, if A is square, then A is invertible, and $\|A^{-1}\| \leq \|C\|/(1 - \|1 - CA\|)$.*

Proposition.

$$\operatorname{dist}(A, \{C \in L(S) : C \text{ not invertible}\}) = 1/\|A^{-1}\|.$$

Proof: *Consider the rank-one modification $C := A - [x]\lambda$ of A (with $\lambda \in S'$). If $\lambda A^{-1}x = 1$, then $C(A^{-1}x) = (A - [x]\lambda)A^{-1}x = x - x = 0$, i.e., C is not invertible. Further, $\|A - C\| = \|[x]\lambda\| = \|x\|\|\lambda\|$. Thus*

$$\begin{aligned}\operatorname{dist}(A, \{C \in L(S) : C \text{ not invertible}\}) &\leq \inf\{\|x\|\|\lambda\| : \lambda A^{-1}x = 1\} \\ &= 1/\sup_{\lambda,x} \frac{\lambda A^{-1}x}{\|\lambda\|\|x\|} = 1/\sup_x \|A^{-1}x\|/\|x\| \;=\; 1/\|A^{-1}\|.\end{aligned}$$

For the converse inequality, observe that, for a noninvertible C, can find $x \in \ker(C)\backslash 0$, hence get $\|A - C\| \geq \|Ax - Cx\|/\|x\| = \|Ax\|/\|x\| \geq 1/\|A^{-1}\|$. □

Cor. *If A^{-1} takes on its norm at the unit vector x and λ is a linear functional of norm 1 which takes on its norm at $A^{-1}x$, then $A - [x]\lambda/\|A^{-1}\|$ is a noninvertible matrix closest to A.*

If Cr is a good estimate for the error $x - \tilde{x}$ in $\tilde{x}$, then adding it to $\tilde{x}$ should give a better approximation for x. This is at the basis of successful fixed point iteration: The equation $Ax = b$ is equivalent to the equation

$$x = x + C(b - Ax) = Ex + Cb,$$

and the fact that E has norm < 1 ensures convergence of the iteration

$$x_{n+1} := Ex_n + z, \qquad n = 0, 1, 2, \ldots, \tag{2}$$

(regardles of the choice of $z \in S$) since it ensures that the iteration function $g : x \mapsto Ex + z$ is Lipschitz continuous with Lipschitz constant $\kappa = \|E\| < 1$. In fact, $\|x_{n+1} - x_n\| = \|E(x_n - x_{n-1})\| \leq \|E\|\|x_n - x_{n-1}\|$, therefore $\|x_{n+m} - x_n\| \leq \sum_{j<m} \|x_{n+j+1} - x_{n+j}\| \leq \sum_{j<m} \|E\|^{j+n}\|x_1 - x_0\| = \|E\|^n \frac{\|E\|^m - 1}{\|E\| - 1}\|x_1 - x_0\|$ goes to zero as $n \to \infty$. Hence the sequence (x_n) is Cauchy, therefore converges, to some y. On the other hand, any such limit y is necessarily a fixed point of the iteration, i.e., satisfies $y = Ey + z$, i.e., $CAy = z$. In fact, it is the *unique* solution, since if also $v = Ev + z$, then $y - v = E(y - v)$, hence $\|y - v\| \leq \|E\|\|y - v\|$ which, in view of $\|E\| < 1$ can only hold if $\|y - v\| = 0$.

Conclusion. *If C is an approximate inverse for A, then CA is 1-1 and onto, hence invertible. In particular, A is 1-1 and C is onto. Therefore, if also either A is onto or C is 1-1, then both are invertible. In particular, in that case, the limit of the iteration $x_{n+1} = Ex_n + Cb$ is the unique solution of the linear system $A? = b$.*

One reason why Numerical Analysts are pleased to have so many different norms available is precisely because the definition of approximate inverse involves a specific norm, while conclusions like the invertibility of A or the convergence of fixed point iteration are **norm-independent** in the finite-dimensional setting of this course. This means that we can make use of suitably chosen norms to establish these results.

We pick up on this when discussing iterative methods.

Elementary matrices are the main tool in the construction of suitable factorizations. We call $E \in \mathbb{R}^{n\times n}$ an **elementary matrix** if it is a rank-one perturbation of the identity, i.e., of the form

$$E = I_n - vw^T$$

for some n-vectors v and w. Such matrices are easy to apply:

$$Ex = x - v(w^T x),$$

i.e., it involves no more than a scalar product and the subtraction of a scalar multiple of one vector from another. Both of these operations are part of the BLAS. One would never form the matrix E explicitly, but merely store the two vectors v and w. In specific instances, only one of these has to be stored since they are closely related.

The action of such an elementary matrix is easy to visualize: E keeps pointwise fixed the hyperplane $w\perp$, i.e.,

$$Ex = x \quad \forall x \perp w.$$

Thus, if E is invertible, then E^{-1} must keep the same hyperplane pointwise fixed, i.e., necessarily $E^{-1} = I_n - uw^T$ for some suitable u. One computes

$$(I - uw^T)E = (I - uw^T)(I - vw^T) = I - (u + v - u(w^T v))w^T$$

and concludes that

Prop. *$E = I - uw^T$ is invertible iff $w = 0$ or else $u + v = u(w^T v)$, i.e.*

$$u = v/(w^T v - 1),$$

in which case

$$(I - vw^T)^{-1} = I - vw^T/(w^T v - 1).$$

Thus the inverse of an elementary matrix is an elementary matrix of the same kind, hence easily applicable.

Aside: More generally, have the

Sherman-Morrison-Woodbury formula.

$$(A - VW^T)^{-1} = A^{-1} - A^{-1}V(W^T A^{-1}V - 1)^{-1}W^T A^{-1}$$

which holds if both A and $W^T A^{-1}V - 1$ are invertible, hence $A - VW^T$ is also invertible. The special case in which $\#V = 1$ is known as the **Sherman-Morrison formula**.

Cor. *$E = I - cvw^T$ is orthogonal iff $v = w$ and $c = 2/v^T v$.*

(proof: E orthogonal iff $(I - w(cv)^T) = E^T = E^{-1} = (I - (c/(1 - cw^T v))vw^T)$ iff $v = w$ and $c = c/(1 - cv^T v)$, i.e., $2 = cv^T v$.)

Cor. *An orthogonal elementary matrix is also symmetric and selfinverse ($:= E = E^{-1}$).*

(proof:$E = (I - cvv^T)$ implies $E^T = E$. Hence, if also $E^{-1} = E^T$, then E is selfinverse. In fact, for a general square matrix A, any two of the three properties *orthogonal* (i.e., $A^T A = I$, hence $A^{-1} = A^T$), *symmetric* (i.e., $A^T = A$), *selfinverse* (i.e., $A^{-1} = A$), imply the third.)

Because Householder has popularized use of orthogonal elementary matrices (and because they are selfinverse), they are called **Householder reflections** or **Householder transformations**.

EXAMPLES: In applications, one uses elementary matrices to map a specific vector x to some vector y while leaving a whole hyperplane untouched.

For example, the first (and typical) step in **Gauss elimination** consists in subtracting, from the jth row of the matrix A in question, the multiple $m(j)$ of the first row, for all $j > 1$. For the typical column, this means that we are subtracting $m(j)$ times its first entry from its jth, all j. In other words, we change the typical column a to $a - ma(1)$, with $m := [0, m(2), m(3), \ldots]^T$. This is exactly the same as applying the elementary matrix $E_1 := I - me_1^T$. We can check now the proper choice of m, proper in the sense that it maps a_1, ie. the first column of A, to $a_1(1)e_1$. Since $E_1 a_1 = a_1 - ma_1(1)$, we need $m = (a_1 - a_1(1)e_1)/a_1(1)$. Note that $E_1^{-1} = I - me_1^T/(0-1) = I + me_1^T$, i.e., exactly like E_1 itself except for that change in sign.

Thus

$$A = (I + me_1^T) \begin{pmatrix} \alpha & \beta^T \\ 0 & B \end{pmatrix},$$

and induction gives that

$$A = LU,$$

with L lower triangular and U upper triangular **provided** the upper left corner element at each step is nonzero.

For a second example, we may wish to map an arbitrary vector x to a multiple αe_1 of the first unit vector. If we are to do that by a Householder reflection $H = I - cvv^T$ for a certain vector v, then necessarily $|\alpha| = \|x\|_2$ and v must be parallel to $x - \alpha e_1$, hence we choose $v = x - \alpha e_1$, whence $c = 2/v^T v = 2/\|x - \alpha e_1\|_2^2$. One would choose the sign of α so as to make $v^T v$ as large as possible (to avoid cancellation in the calculation of v), i.e., $v = [\text{signum}(x(1))(|x(1)| + \|x\|_2), x(2), x(3), \ldots]^T$. Note that $v^T v = x^T x - 2\alpha x(1) + \alpha^2$, hence $1/c = \|x\|(\|x\| + |x(1)|)$.

If H is constructed in this way from $x := A(:, 1)$ for some given matrix A, then

$$A = H \begin{pmatrix} \alpha & \beta^T \\ 0 & B \end{pmatrix} \tag{1}$$

(remember that such H is self-inverse), and induction gives that

$$A = QR,$$

with Q orthogonal and R upper (or, right) triangular. Note that (1) even holds when $A(:, 1) = 0$, regardless how we might then define the formally undefined v in that case. Note further that A need not be square here. Note also that necessarily $\text{ran}([a_1, \ldots, a_j]) \subseteq \text{ran}([q_1, \ldots, q_j])$, all j, which implies that the contruction of this **QR factorization** also does the job for which mathematicians traditionally rely on Gram-Schmidt. It turns out that Householder is better at it than Gram- Schmidt.

We return to this shortly, in the discussion of Least-squares, aka solution of overdetermined systems.

Givens rotations do not quite fit this pattern of an elementary matrix since they are obtained as a **rank-two** modification of the identity. The typical Givens rotation $G = G_{i,j,\theta}$ is an arbitrary rotation in a particular twodimensional coordinate plane. If this is the plane spanned by the unit vectors e_i and e_j, then $Ge_k = e_k$ for all $k \neq i, j$, while $Ge_i = ce_i + se_j$, $Ge_j = -se_i + ce_j$, with $c = \cos(\theta), s = \sin(\theta)$.

Note that such G is orthogonal and that $(G_{i,j,\theta})^{-1} = G_{i,j,-\theta} = G_{j,i,\theta}$. Givens rotations are of use as similarity transformations, i.e., to change A into $\hat{A} := GAG^{-1}$ in such a way that $\hat{A}(i, j) = 0$. We'll come back to this in the discussion of eigenvalue calculations.

rounding error analysis for Gauss elimination Gauss elimination provides the factorization $PA = LU$ (or $A = PLU$, if that is more convenient), with L unit lower triangular and U upper triangular. The solution x of the system $A? = b$ is obtained by getting first y as the solution of $L? = Pb$, and then x as the solution of $U? = y$. Since pivoting (as recorded in the permutation matrix P) does not introduce any rounding errors, we may assume in the following error analysis that $P = I$, i.e., that we are applying Gauss elimination without pivoting to the (properly rearranged) matrix PA, which we'll call A again.

Calculation of L and U proceeds exactly (except possibly for the temporal order of the intermediate steps) as if we were computing their entries from the requirement that

$$A = LU, \quad \text{with } L \text{ unit lower }, U \text{ upper triangular.}$$

I.e., $A(i,j) = \sum_k L(i,k)U(k,j) = \sum_{k \le \min\{i,j\}} L(i,k)U(k,j)$, therefore

$$A(i,j) = \sum_{k<\min\{i,j\}} L(i,k)U(k,j) + \begin{cases} L(i,i)?, & i \le j; \\ ?U(j,j), & i > j, \end{cases}$$

thus providing the formulae

$$U(i,j) = A(i,j) - \sum_{k<i} L(i,k)U(k,j), \quad i \le j$$

$$L(i,j) = (A(i,j) - \sum_{k<j} L(i,k)U(k,j))/U(j,j), \quad i > j$$

This is not only convenient for handling storage (since it indicates that the interesting entries of L and U can be stored over the corresponding entries of A (even if they are computed by working out those sums gradually, as one would in classical Gauss Elimination)), but it also allows one to account for the rounding errors incurred during the factorization by perturbing the entries of A. Thus, with L and U now the computed factors, we have

$$U(i,j) = A(i,j)\varepsilon^{i-1} - \sum_{k<i} L(i,k)U(k,j)\varepsilon^{i-k+1}, \quad i \le j$$

$$L(i,j)U(j,j)\varepsilon = A(i,j)\varepsilon^{j-1} - \sum_{k<j} L(i,k)U(k,j)\varepsilon^{j+1-k}, \quad i > j$$

with $\varepsilon = (1+\delta)$ and $|\delta| \le \mathbf{u}$. Recall that, under the assumption $n\mathbf{u} \le 0.01$ which we now make, $\varepsilon^r = 1 + r\delta$ for any $r \le n$ and with $|\delta| \le 1.01\mathbf{u}$. Therefore,

$$\sum_k L(i,k)U(k,j) - A(i,j) = \begin{cases} A(i,j)(i-1)\delta - \sum_{k<i} L(i,k)U(k,j)(i-k)\delta, & i \le j \\ A(i,j)(j-1)\delta - \sum_{k<j} L(i,k)U(k,j)(i+1-k)\delta, & i > j \end{cases}$$

or

$$|LU - A| \le (|A| + |L||U|)\mathbf{u}_n$$

with

$$\mathbf{u}_n := n \cdot 1.01 \cdot \mathbf{u}$$

and the inequalities **pointwise** (i.e., $B \le C := \forall\{i,j\}\ B(i,j) \le C(i,j)$), and, for any matrix B, $|B|(i.j) := |B(i,j)|$.

The computational steps in backsolving the triangular systems $L? = b$ and $U? = y$ are exactly of the same kind as used in the construction of the factorization. The same analysis therefore gives that the computed solution y of $L? = b$ satisfies the equation

$$(L+G)y = b, \text{with } |G| \le \mathbf{u}_n|L|$$

and subsequent solving of $U? = y$ provides the computed solution

$$(U + H)x = y, \text{with } |H| \leq \mathbf{u}_n|U|.$$

Altogether, the computed solution x satisfies

$$(L + G)(U + H)x = b.$$

Also,

$$(L + G)(U + H) = LU + LH + GU + GH \ = \ A \ + \ (LU - A) + LH + GU + GH,$$

hence

$$(A + E)x = b, \tag{2}$$

with

$$|E| \leq (|A| + |L||U|)\mathbf{u}_n + |L||U|(1 + 1 + \mathbf{u}_n)\mathbf{u}_n \sim (|A| + 3|L||U|)\mathbf{u}_n.$$

Conclusion. *To the extent that this estimate is accurate, one could monitor the factorization process for trouble simply by tracking (or, at least, estimating) $|L||U|$.*

Conclusion. *To the extent that this estimate is accurate, it should be the goal of any pivoting strategy to keep $|L||U|$ small. (Note that $|L||U| \geq |A|$, hence, at best, $|L||U| = O(|A|)$.)*

In particular, it is not so much near-zero pivots by themselves that produce trouble with Gauss elimination, but the large entries of $|L||U|$ that are usually their consequence.

We showed that elimination solves exactly the 'near-by' problem (2). How much we can conclude from this about the *error* in the computed solution x depends on the **condition** of A.

condition Let x be the solution of $A? = b$, $\hat{x}$ the computed solution, and set $r := b - A\hat{x} =: Ae$. Then $\|e\|/\|A^{-1}\| \leq \|r\| \leq \|A\| \, \|e\|$ and $\|b\|/\|A\| \leq \|x\| \leq \|A^{-1}\| \, \|b\|$ hence

$$\frac{\|r\|}{\|b\|}/\kappa(A) \leq \frac{\|e\|}{\|x\|} \leq \frac{\|r\|}{\|b\|}\kappa(A) \tag{3}$$

with

$$\kappa(A) := \|A\| \, \|A^{-1}\|$$

the **condition** (number) of A. Both inequalities in (3) are *sharp*, i.e., can be made equality by appropriate choice of x and $\hat{x}$.

Conclusion. *The relative residual $\|r\|/\|b\|$ allows conclusions about the size of the relative error $\|e\|/\|x\|$ to the extent that $\kappa(A)$ is close to 1. If $\kappa(A)$ is far from 1 (with respect to the precision used), then a small relative residual gives no information about the relative error; the relative error could be even much smaller or it could be much larger.*

For this reason, good packages like MATLAB provide (automatically) a **condition number estimator** which warns the user when $\kappa(A)\mathbf{u}_n \sim 1$ or worse. For, assuming that pivoting has succeeded in keeping $|L||U| = O(|A|)$, we conclude that the computed solution $\hat{x}$ has a relative residual of the order of $\mathbf{u}_n$, i.e., as good as can be expected, therefore the relative error is no worse than $\kappa(A)\mathbf{u}_n$. Having this O(1) or worse means that the computed solution may have no correct digits.

One *estimates* $\kappa(A)$ with the aid of the computed factorization by trying to generate some right side d of norm 1 for which $A^{-1}d$ is as small as possible. The heuristics are that, with proper pivoting, L^{-1} is not large, hence whatever trouble A^{-1} might have, this already comes out in U^{-1}. Computing U^{-1} and taking its norm would be the obvious thing to do, but it is too expensive; it costs $O(n^3)$ flops. Instead (working in the max-norm), one chooses $d(k) \in \{-1, 1\}$ so as to maximize $|y(k)|$, with $y(k) := (d(k) - \sum_{j>k} U(k, j)y(j))/U(k, k)$ the kth entry of $y := U^{-1}d$ as computed during backsubstitution.

In the scheme actually used in MATLAB (algorithm 4.5-1), the maximization is somewhat more global since it takes into account the effect of choosing $d(k) = 1$ vs $d(k) = -1$ not only on $y(k)$ itself, but also on all the sums $\sum_{j \geq k} U(i,j)y(j)$ for $i < k$ (which enter the subsequent calculation of $y(i)$).

Some efforts have been made to find methods for reducing $\kappa(A)$. They all come down to looking for a norm in which $\kappa(A)$ is small. If (as we mostly do) we only deal with map norms, then $\kappa(A) \geq 1$. The standard way for trying to reduce $\kappa(A)$ is to seek *diagonal* matrices D_l and D_r for which, in the given norm(s), $\kappa(D_l A D_r)$ is as small as possible. (One would restrict the entries of D_l and D_r to have mantissa .1 in whatever floating point arithmetic is being used, in order to avoid roundoff because of such **scaling**.) This leads to **equilibration** of matrices, something I'll skip because I don't fully appreciate its goal in the present context (and this early in the morning).

Factorization as approximate inverse We can think of (2) as telling us that the solution process provides us with the means of computing Cz for given z, with $C := (A+E)^{-1}$ and $|E|$ boundable as shown. Is C even defined? Assuming A to be invertible, we know $A + E$ to be invertible in case $\|E\| < 1/\|A^{-1}\|$, i.e., in case the relative perturbation $\|E\|/\|A\|$ is less than $1/\kappa(A)$. In that case, $\|C\| = \|(A+E)^{-1}\| \leq \|A^{-1}\|/(1 - \|A^{-1}\|\|E\|)$. Then $I - CA = I - C(C^{-1} - E) = CE \leq \|A^{-1}\|\,\|E\|/(1 - \|A^{-1}\|\,\|E\|)$, and this is less than 1 in case

$$\|A^{-1}\|\|E\| < 1/2. \tag{4}$$

In such a case, one could use the backsubstitution process to provide *fixed point iteration*, i.e, the iteration

$$x_{n+1} = x_n + Cr_n, \quad n = 0, 1, 2, \ldots$$

with $r_n := b - Ax_n$ and $x_0 := \hat{x}$, and so hope to obtain an improved solution. This is called **iterative improvement**.

For this, note: Strictly speaking, the error matrix E depends on the particular b to which the calculation was applied, hence we are dealing here with a more complicated object than mere fixed point iteration. But if we can be certain that (4) holds uniformly for all right sides involved, then convergence, to the solution, can be shown, if the calculation is carried out exactly. Since it usually is not, the additional error in the **calculated** residual r_n has to be taken into account. If the residual is computed with the same precision as the (approximate) solutions, then we cannot expect any improvement. For this reason, it is standard to compute $r_n = b - Ax_n$ with *double precision accumulation* of the scalar products $b(i) - A(i,:)x_n$ involved (as can be done easily in many implementations of floating point arithmetic).

overdetermined system:= $A? = b$ with $A \in \mathbb{R}^{m\times n}$ and $m > n$.

$\dim \operatorname{ran}(A) \le \dim \operatorname{dom}(A) = n < m = \dim \operatorname{tar}(A)$ implies that there may be no solution. Actually, this may even happen when $m \le n$, viz. when A fails to be onto because of rank deficiency. In either case, the next best thing is to solve instead

$$\|b - A?\| = \min .$$

This is a much more tricky computational problem than solving a square linear system, unless the norm is chosen so as to make the problem linear again. There is only one such choice of norm, viz. a weighted 2-norm. For this reason, one usually works with that norm, even though another norm may be indicated by the problem background.

Consider now the **LS problem**

$$\|b - A?\|_2 = \min .$$

We are either looking for the element $y = Ax \in \operatorname{ran}(A)$ closest to b, or we are looking explicitly for all x for which Ax is closest to b. The book takes the latter point of view, I prefer the former.

$P_A :=$ orthogonal projector onto $\operatorname{ran}(A)$. Then $P_A b$ is the best approximation (in the 2-norm) to b from $\operatorname{ran}(A)$. Hence, in the terms used in the book,

$$X := \{x : \|b - Ax\|_2 = \min\} = \{x : Ax = P_A b\}.$$

Also, X is a singleton iff A is 1-1. In that case, the book writes $x_{\rm LS} := P_A b$. More generally, the book denotes by $x_{\rm LS}$ the unique point of **minimal** 2-norm in X. As near as I can tell, this so-called **best least squares** solution is the product of mathematical neatness and good for nothing else. The typical situation calls for $P_A b$, and the elements of X become important only if, for some reason, it is important to represent $P_A b$ in the form Ax. In that case, there are usually problem-dependent considerations that give preference to one element of X over the others. Such preferences amount to the addition of homogeneous equations so that the coefficient matrix $[A; B]$ of the enlarged system

$$\begin{bmatrix} A \\ B \end{bmatrix} ? = \begin{bmatrix} b \\ 0 \end{bmatrix} \tag{5}$$

is 1-1.

For **example**, $X = P_A b + \ker(A)$, hence $x_{\rm LS}$ is the error in the best approximation from $\ker(A)$ to $P_A b$. In particular, $x_{\rm LS}$ is the unique element in X perpendicular to $\ker(A)$. Thus, the augmented system (5) with $B = W^T$ and W any spanning sequence for $\ker(A)$ has as its **unique** LS solution the vector $x_{\rm LS}$ of the original system.

pseudo-inverse Recall that, with $A = U\Sigma V^T = \sum_{k\le r} u_k \sigma_k v_k^T$ the SVD for A, $U_r := [u_1, \dots, u_r]$ provides an orthogonal basis for $\operatorname{ran}(A)$, hence $P_A = U_r U_r^T$, therefore $\|b - P_A b\|_2^2 = \sum_{k>r} (u_k^T b)^2$. Further,

$$\begin{aligned} \|b - Ax\|_2^2 &= \|U^T b - \Sigma V^T x\|_2^2 \\ &= \sum_{k\le r} (u_k^T b - \sigma_k v_k^T x)^2 + \sum_{k>r} (u_k^T b)^2. \end{aligned}$$

Hence minimization requires $v_k^T x = u_k^T b/\sigma_k$ for $k \le r$, thus pinning down $v_k^T x$ for $k \le r$. Since V is orthogonal, the x of smallest norm satisfying these conditions will have $v_k^T x = 0$ for $k > r$ (since $\|x\|_2^2 = \|V^T x\|_2^2 = \sum_k (v_k^T x)^2$), as already observed earlier. Conclusion:

$$x_{\rm LS} = \sum_{k\le r} (u_k^T b/\sigma_k) v_k \ =: A^+ b. \tag{6}$$

There is a vast literature on the resulting map

$$A^+ := V\Sigma^+ U^T, \quad \text{with } \Sigma^+ := \operatorname{diag}(\sigma_1^{-1}, \dots, \sigma_r^{-1}, 0, \dots, 0),$$

called the **pseudo-inverse** for A. Note that (for $m \geq n = \text{rank}(A)$), $A^TA = V\Sigma_n^2 V^T$, hence

$$(A^TA)^{-1}A^T = V\Sigma_n^{-1}U^T. \tag{7}$$

If A is invertible, then $A^+ = A^{-1}$. Hence, the map $A \mapsto A^+$ is an extension of the map $A \mapsto A^{-1}$. But note that $\|A^+\|_2 = \sigma_r^{-1}$, hence $\|A^+\| \to \infty$ as A approaches a matrix of lower rank. Thus the ability to define a pseudo-inverse for all A is bought at the price of having the map $A \mapsto A^+$ discontinuous. (Prove that that is necessarily so!). This discontinuity is implicit in the following bound, which need not go to zero as $A \to B$ since $\|A^+\|_2$ might blow up.

(Theorem 6.1-2).

$$\|B^+ - A^+\|_F \leq 2\|B - A\|_F \max\{\|A^+\|_2^2, \|B^+\|_2^2\}.$$

Compare with

$$\|B^{-1} - A^{-1}\| \leq \|B - A\| \, \|A^{-1}\| \, \|B^{-1}\|,$$

valid for invertible maps and arbitrary norms.

sensitivity analysis Standard sensitivity analysis for the LS problem assumes that both A and its perturbed version, $A + \delta A$, have full (column) rank. (Here, $\delta A, \delta b$ are names for the perturbations in A and b.)

Theorem 6.1-3. *$A \in \mathbb{R}^{m\times n}$, A of full rank, all norms are 2-norms,*
$\|b - Ax\| = \min$, $\|(b + \delta b) - (A + \delta A)\hat{x}\| = \min$, $r := b - Ax$, $\hat{r} := (b + \delta b) - (A + \delta A)\hat{x}$,

$$\varepsilon := \max\{\|\delta A\|/\|A\|, \ \|\delta b\|/\|b\|\} < 1/\kappa(A)$$
$$\sin(\theta) := \|r\|/\|b\| < 1.$$

Then, $\cos(\theta) = \|Ax\|/\|b\|$, $\kappa(A) = \sigma_1(A)/\sigma_n(A)$, *and*

$$\|\hat{x} - x\|/\|x\| \leq \varepsilon(2\kappa(A) + \sin(\theta)\kappa(A)^2)/\cos(\theta) + O(\varepsilon^2)$$
$$\|\hat{r} - r\|/\|b\| \leq \varepsilon(1 + 2\kappa(A)) \min\{1, m - n\} + O(\varepsilon^2)$$

Proof: Set $A_t := A + t\delta A$, $b_t := b + t\delta b$. Then $\|b_t - A_t x_t\| = \min$ defines the smooth function $[0, 1] \to \mathbb{R}^n : t \mapsto x_t = (A_t^TA_t)^{-1}A_t b_t$ since, by assumption, $\|t\delta A\| \leq \|\delta A\| < 1/\sigma_n(A)$, hence $\text{rank}(A_t) = n$. Also

$$\hat{x} = x_1 = x_0 + \dot{x}_0 + O(\|\ddot{x}\|),$$

with $\|\ddot{x}\| := \max_{\tau\in[0,1]} \|\ddot{x}_\tau\|$, therefore

$$\|\hat{x} - x\|/\|x\| \leq \|\dot{x}_0\|/\|x\| + O(\|\ddot{x}\|).$$

Differentiate the identity $(A_t^TA_t)x_t = A_t^Tb_t$ to get

$$((\delta A)^TA_t + A_t^T\delta A)x_t + (A_t^TA_t)\dot{x}_t = (\delta A)^Tb_t + A_t^T\delta b. \tag{8}$$

At $t = 0$, this gives

$$((\delta A)^TA + A^T\delta A)x + (A^TA)\dot{x}_0 = (\delta A)^Tb + A^T\delta b,$$

hence

$$\begin{aligned}
\dot{x}_0 &= (A^TA)^{-1}\big(A^T(\delta b - \delta A x) + (\delta A)^T r\big) \\
&\leq \|(A^TA)^{-1}A^T\|\big(\varepsilon\|b\| + \varepsilon\|A\| \, \|x\|\big) + \|(A^TA)^{-1}\|\varepsilon\|A\| \, \|r\| \\
&= \varepsilon\|x\|\Big(\|(A^TA)^{-1}A^T\| \, \|A\|\big(\frac{\|b\|}{\|A\|\|x\|} + 1\big) + \|(A^TA)^{-1}\| \, \|A\|\frac{\|r\|}{\|x\|}\Big)
\end{aligned}$$

Now, from (7), $\|(A^TA)^{-1}A^T\| = 1/\sigma_n(A)$ while $\|(A^TA)^{-1}\| = 1/\sigma_n(A)^2$. Also, $\|b\|/(\|A\| \, \|x\|) \leq \|b\|/\|Ax\| = 1/\cos(\theta)$ and, therefore, $\|r\|/\|x\| = (\|r\|/\|b\|)(\|b\|/\|x\|) \leq \sin(\theta)\|A\|/\cos(\theta)$. This implies that

$$\|\dot{x}_0\|/\|x\| \leq \varepsilon\Big(\kappa(A)\big(1/\cos(\theta) + 1\big) + \kappa(A)^2\sin(\theta)/\cos(\theta)\big)\Big).$$

A differentiation of (8) gives

$$2(\delta A)^T\delta A x_t + 2((\delta A)^TA_t + A_t^T\delta A)\dot{x}_t + (A_t^TA_t)\ddot{x}_t = (\delta A)^T\delta b + (\delta A)^T\delta b.$$

Here all terms, except for the term $(A_t^TA_t)\ddot{x}_t$ are $O(\varepsilon^2)$, hence so is $\ddot{x}_t$.

The proof for the bound on $\|\hat{r} - r\|/\|b\|$ proceeds along similar lines. □

Important point: If $\sin(\theta) \neq 0$, i.e., if the residual r is not zero, then the bound on the relative error in terms of the relative noise in A and b involves the **square** of the condition number. The corresponding bound on the relative change in the residual only involves the condition number itself.

For relatively small relative residual (i.e., for $\sin(\theta) \sim 0$), the noise effect is only proportional to the condition number. In such a situation, it is important to use a method whose sensitivity to noise is only proportional to $\kappa(A)$. The standard method of solving the **normal equations**

$$A^T Ax = A^T b$$

e.g., by Gauss elimination, is not such a method since the condition of its coefficient matrix $A^T A$ is $\kappa(A)^2$.

QR factorization via Householder and (modified) Gram-Schmidt basic idea: factor $A = QR$, with Q orthogonal and R right (or upper) triangular (in the sense that $\mathrm{ran}(i,j) = 0$ for $i > j$).

Notational difficulty: If A is not square, i.e., $A \in \mathbb{R}^{m\times n}$ with $m > n$, then, straighforwardly, also $Q \in \mathbb{R}^{m\times n}$, but R is square. The book prefers for Q to be square (as that is what one gets when using Householder for the construction of the factorization), hence has $R \in \mathbb{R}^{m\times n}$ with the part below row n all zero and useless. This does have the advantage that Q is invertible.

Purpose: If we are to solve the linear system $A? = b$, then we can now look at the equivalent system $R? = Q^T b$ whose solution by backsubstitution is easy. This even works when $m > n$, in which case we obtain thereby the orthogonal projection of b onto $\mathrm{ran}\, A$, i.e., the LS solution to the problem.

Let $A = QR$ be a QR factorization for A. Then

$$\mathrm{ran}\, A(:, 1{:}j) \subset \mathrm{ran}\, Q(:, 1{:}j), \quad \forall j \tag{9}$$

(since then $A(:,j) = Ae_j = QRe_j = Q(\sum_{k\le j} R(k,j)e_k) = \sum_{k\le j} R(k,j)Qe_k$). Hence, if A is 1-1, then necessarily

$$\mathrm{ran}\, A(:, 1{:}j) = \mathrm{ran}\, Q(:, 1{:}j), \quad \forall j,$$

thus $Q(:, 1{:}j)$ provides an orthogonal basis for $\mathrm{ran}\, A(:, 1{:}j)$. Conversely, if Q is orthogonal and satisfies (9), then necessarily $A = QR$ with R right triangular since $R(:,j) = Re_j$ contains the coordinates of $A(:,j)$ with respect to Q and, by (9), $A(:,j) \in \mathrm{ran}\, Q(:, 1{:}j)$.

We can therefore think of obtaining such a QR factorization for A by orthogonalizing the columns of A a la Gram-Schmidt (see Jan.27 notes). This would only provide us with an orthonormal basis for $\mathrm{ran}\, A$. But, for solving $A? = b$, that is all we need. We can also think of it, straightforwardly, as the task of eliminating (see notes Feb.6) the strictly lower triangular part of A but using orthogonal elementary matrices (whose product then provides Q or Q^{-1}). The second point of view has the advantage that it provides Q as a square matrix, thus giving not only an orthonormal basis for $\mathrm{ran}\, A$ but also for its orthogonal complement.

In **Gram-Schmidt**, the typical or $j+1$st step starts with $Q_j := Q(:, 1{:}j)$ already in hand. This means that we can compute the orthogonal projection $P_j x$ of any vector x onto $\mathrm{ran}\, A(:, 1{:}j)$ as $Q_j Q_j^T x$. Its error, $x - P_j x$, is necessarily orthogonal to $\mathrm{ran}\, A_j = \mathrm{ran}\, Q_j$, hence $Q(:, j+1) := (x - P_j x)/\|x - P_j x\|$ provides a suitable next column for Q, provided we take $x = A(:, j+1)$ and *provided* the resulting vector $x - P_j x$ is not zero, i.e., provided $A(:, 1{:}(j+1))$ is 1-1.

The *modification* which gives the name to **modified Gram-Schmidt** concerns the actual calculation of $x - P_j x = x - Q_j Q_j^T x$. In the standard process, one computes the vector $c := Q_j^T x$ and then subtracts $Q_j c$ from x. Now notice that $q_j^T x = q_j^T (x - P_{j-1} x)$ since q_j is perpendicular to $\mathrm{ran}\, Q_{j-1} = \mathrm{ran}\, P_{j-1}$. On the other hand, $\|x - P_{j-1}x\| \le \|x\|$, and usually, the inequality is quite strict. This means that the scalar product $q_j^T (x - P_{j-1}x)$ is apt to involve absolutely smaller terms than the scalar product $q_j^T x$. Since they both give the same result, it must be the case that the calculation $q_j^T x$ involves more cancellation than the calculation $q_j^T (x - P_{j-1}x)$, hence the latter is preferable. This means that, at the beginning of the $j+1$st step, all columns of A will already have been modified j times by subtracting out their orthogonal projection onto $\mathrm{ran}\, A_{j-1} = \mathrm{ran}\, Q_{j-1}$. If we use Q as our working array (which we would initialize to A), then we have an orthonormal basis for A_j in $Q(:, 1{:}j)$, while $Q(:, (j+1){:}n) = (1 - P_{j-1})A(:, (j+1){:}n)$. The task of this step is to take out from columns $j+1, \ldots, n$ also the projection onto the span of q_j by replacing each such column x by $x - q_j q_j^T x$. This leaves the $j+1$st column containing $A(:, j+1) - P_j A(:, j+1)$, and its normalization provides the next column for Q, i.e., the vector q_{j+1}. In practice, one would not carry out this normalization, as it requires taking a square-root. Rather, one would work with the unnormalized vectors, call them $p_1, p_2, \ldots$, and merely record their norm-squares, i.e., the numbers $r_k := \|p_k\|^2$, all k. One obtains $q_k q_k^T x$ from this in the form $p_k p_k^T x / r_k$.

When using **elimination with Householder reflections**, the typical or $j+1$st step starts with the matrix R (which was initialized to A) already upper triangular in its first j columns, i.e.,

$$R = \begin{bmatrix} R_1 & X \\ 0 & S \end{bmatrix}$$

with R_1 square upper triangular. Also, $A = QR$, with Q initialized to I and presently containing the product of the previous j Householder reflections. A Householder reflection $H_j = 1 - c_j v_j v_j^T$ is then constructed so that $H_j R(:, j+1) = [R(1{:}j, j+1); \alpha; 0, \ldots, 0]$, with α necessarily equal to $\pm\|R((j+1){:}m, j+1)\|$. Q and R are then updated:

$$Q \leftarrow QH_j = Q - c_j(Qv_j)v_j^T, \quad R \leftarrow H_j R = R - c_j v_j(v_j^T R).$$

(In fact, Q is usually not constructed explicitly. Rather, the numbers c_j and vectors v_j are stored individually (mostly in R itself) and the sequence $H_0, H_1, \ldots$ is applied whenever application of Q is called for.) Since the construction of H_j requires division by α, there is a difficulty here when $R((j+1){:}m, j+1) = 0$. In that case, $Ae_{j+1} = QRe_{j+1} = \sum_{k \le j} q_k R(k, j+1)$, i.e., $Ae_{j+1} \in \operatorname{ran} Q_j = \operatorname{ran} A_j$ (assuming that R_1 is invertible), i.e., $A(:, 1{:}j+1)$ fails to be 1-1.

There is a straightforward way of dealing with this difficulty of a **rank-deficient** A. From the elimination point of view, there is no need for this elimination step since the strictly lower triangular part of the current column is already zero. Hence the choice $H_j = I$ will do. The corresponding action in Gram-Schmidt is simply to ignore the $j+1$st column and go on to the next. Either way, the final factorization obtained this way is deficient since now R fails to have an invertible upper triangular part. One deals with this by formally swapping the useless $j+1$st column of A for a more useful, i.e., independent later column of A.

This solution of the problem of rank-deficiency takes it for granted that we can tell a zero when we see one. In finite-precision arithmetic, that is a hard problem. The best we can do here is the following: if $x - P_j x = O(\mathbf{u}x)$, then, **within the precision used**, $x \in \operatorname{ran} P_j = \operatorname{ran} Q_j = \operatorname{ran} A(:, 1{:}j)$.

In the elimination approach, one similarly checks whether $\|R((j+1){:}m, j+1)\| = O(\mathbf{u}\|A(:, j+1)\|)$ and deems $A(:, j+1)$ to be in $\operatorname{ran} A(:, 1{:}j)$ in that case.

In either method, one uses in practice **column pivoting** by choosing, at the $j+1$st step, from among the columns not yet used the one with the largest norm. This requires some normalization at the beginning of the process. It is not necessary to *recompute* the norms of the remaining columns since we can keep a running total: In elimination, each column is modified by a unitary matrix, hence does not change its norm. On the other hand, if we have in hand $r_k := \|R(j{:}m, k)\|^2$ from the previous step, then we can easily update it for the current step by $r_k \leftarrow r_k - R(j,k)^2$. In modified Gram-Schmidt, one similarly keeps track of the current norm (square) of each column in the (working) matrix Q, $r_k := \|Q(:,k)\|^2$, say . During the $j+1$st step, one subtracts from each such column $x := Q(:,k)$ (for $k > j$) the vector $q_j(q_j^T x)$, thus can update $r_k \leftarrow r_k - (q_j^T x)^2$.

In either method, one produces, in the end, a factorization $A\Pi = QR$ for some permutation matrix Π and with R of the form

$$R = \begin{bmatrix} R_{11} & R_{12} \\ 0 & 0 \end{bmatrix}$$

where R_{11} is an invertible upper triangular matrix of order $r := \operatorname{rank}(A)$. Then $Q^T b = Q^T A? = R\Pi^T ?$ has the so-called **basic** solution

$$\Pi^T x_B = \operatorname{diag}(R_{11}^{-1}, 0) Q^T b.$$

If you are after x_{LS}, though, you would have to work still harder here. You now know that

$$X = x_B + \Pi \begin{bmatrix} R_{11}^{-1}(-R_{12}) \\ I \end{bmatrix} \mathbb{R}^{n-r},$$

hence must compute a best approximation to x_B from that space at the right. On the other hand, remember the earlier comments about a problem-motivated approach to removing rank-deficiency.

Note that Householder holds a slight edge here over Gram-Schmidt since, in the former method, the vectors treated in step $j+1$ have only $m-j$ entries, while they always have m entries in the latter method. It can be shown that Householder elimination and *modified* Gram-Schmidt have similar stability properties (which are better than those of Gauss elimination in case the latter is applicable).

If A is very sparse in a regular way, then it might pay to use **Givens rotations** rather than Householder reflections during the elimination, since the former can be used effectively to eliminate the occasional nonzero strictly lower triangular entry. But the implementation can be hairy.

Householder bi-diagonalization is of use in the LS problem. It is also handy, ultimately, in the construction of the SVD (to be discussed after the QR method for computing eigenvalues).

The goal: factor A as UBV^T, with U and V orthogonal, and B **(upper) bidiagonal**, i.e., $B(i,j) \neq 0 \Longrightarrow j = i, i+1$. This is as close as one can get to a SVD without iteration.

The purpose: to provide better information about rank(A) (though nothing beats the SVD for that) for use in the LS problem. More importantly, to provide a good starting point for the iterative construction of the SVD.

The means: use Householder reflections on both sides to alternately zero out a column below the diagonal and the corresponding row to the right of the first super-diagonal.

Thus in the first step (after initializing B to A, and U and V to I), construct H_0 to carry $B(:,1)$ to $\alpha_1 e_1$ and update $B \leftarrow H_0 B$, $U \leftarrow UH_0$. Then construct K_0 to carry the current $B(1,:)^T$ to $B(1,1)e_1 + \beta_1 e_2$. In particular, K_0 leaves e_1 fixed, hence the resulting update $B \leftarrow BK_0^T (= BK_0)$ will not destroy the zeros already obtained in $B(:,1)$. Also update $V \leftarrow VK_0$. Now B is of the form

$$B = \begin{bmatrix} \alpha_1 & \beta_1 e_1^T \\ 0 & B_{22} \end{bmatrix},$$

and the process is applied to the submatrix B_{22}. - You get the picture.

If $m >> n$, it is faster to carry out all the left steps first since this avoids having to apply the K_j to the many rows $B(i,:)$ with $i > n$. It does require a final redoing of the factorization including further steps on the left, but this will then only involve the matrix $B(1{:}n, 1{:}n)$.

The detection of rank-deficiency may be helped by going to this bidiagonal factorization. Still, it is easy to come up with bidiagonal matrices whose elements are all of size 1, yet whose condition number is large. For **example**, the nth order bidiagonal matrix C_n with typical row $[\ldots, 0, 1, -2, 0, \ldots]$ is like a first-order difference equation. The nullspace of the bi-infinite version of C_n is spanned by the vector $\left(2^{-j}\right)_{j=-\infty}^{\infty}$. Hence C_n maps $x = [2^{n-1}, 2^{n-2}, \ldots, 2^0]$ to e_n, therefore $\|C_n^{-1}\|_\infty \geq 2^{n-1}$. (In fact, $C_n^{-1} e_j = [2^{j-1}, 2^{j-2}, \ldots, 2^0, 0, \ldots, 0]$, hence $\|C_n^{-1}\|_\infty = 2^{n-1}$.)

The singular values of such C_n behave similarly to those of B_n computed earlier. They lie between 1 and 3, clustering against 1, except for $\sigma_n(C_n) \sim 2^{-n}$. Only the actual calculation of the SVD (or of C_n^{-1}) would reveal this near rank-deficiency.

In effect, the columns of C_n would be strongly linearly independent if it weren't for the first column. In particular, the singular values of $C_n(:, 2{:}n)$ are perfectly fine.

Iterative improvement is possible (in the full-rank case) with the observation that the problem $\|b - Ax\| = \min$ is equivalent to the linear system

$$\begin{bmatrix} I & A \\ A^T & 0 \end{bmatrix} ? = \begin{bmatrix} b \\ 0 \end{bmatrix}$$

whose unique solution is $[r; x]$ (with $r := b - Ax$, as usual).

In fact, this equivalent formulation is useful in the construction of **iterative** methods for the solution of the LS problem in case A is suitably sparse. It is also to be preferred to the **normal equations** formulation

$$A^T A ? = A^T b$$

since the latter's coefficient matrix has condition $\kappa_2(A)^2$ while the former's has only condition $\kappa_2(A)$.

Generalized Least Squares concerns the problem

$$v^T v = \min \quad \text{subject to } b = Ax + Bv$$

which, for invertible B, is equivalent to

$$\|B^{-1}(b - Ax)\| = \min .$$

(Indeed, set $v := B^{-1}(b - Ax)$.) Paige solves this minimization problem this way: He factors $A = [Q_1, Q_2] * [R; 0]$ with R square right triangular, and then obtains an orthogonal matrix P so that $Q_2^T BP = [0, S]$, with S square right triangular and, correspondingly, $P = [P_1, P_2]$. (If $A \in \mathbb{R}^{m\times n}$ and B is square, then B has order m, R has order n, $Q_2^T B \in \mathbb{R}^{m-n\times m}$, hence S has order $m - n$.) Then $b = Ax + Bv$ is equivalent to

$$\begin{bmatrix} Q_1^T b \\ Q_2^T b \end{bmatrix} = Q^T b = \begin{bmatrix} R \\ 0 \end{bmatrix} x + \begin{bmatrix} Q_1^T BP_1 & Q_1^T BP_2 \\ 0 & S \end{bmatrix} \begin{bmatrix} P_1^T v \\ P_2^T v \end{bmatrix}.$$

Consequently, from the bottom half, the triangular system $S? = Q_2^T b$ has the solution $P_2^T v$ from which v is determined. With v in hand, the top half provides the system $R? = Q_1^T b+$ known stuff, whose solution is x.

Similarity A and B are **similar** := for some invertible V, $A = VBV^{-1}$. If $A : S \to S$ and $V : \mathbb{R}^n \to S$ (i.e, V is a basis for S), then $B := V^{-1}AV$ is a matrix, called **the matrix representation of A wrto the basis V**. In a picture:

$$\begin{array}{ccc} & A & \\ S & \longrightarrow & S \\ V\uparrow & & \uparrow V \\ \mathbb{R}^n & \underset{B}{\longrightarrow} & \mathbb{R}^n \end{array}$$

In the best of circumstances, A is **diagonable**, i.e, can find a *diagonal* matrix $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots)$ similar to A. For then we can study the powers of A and, more generally, $f(A)$ for functions f such as polynomials and limits of polynomials in terms of the much simpler matrices $f(\Lambda) = \mathrm{diag}(f(\lambda_1), f(\lambda_2), \ldots)$, using the fact that then

$$f(A) = Vf(\Lambda)V^{-1},$$

(at least for polynomials f; for other functions f it would be a definition made plausible by a limit process).

It so happens that not all matrices are diagonable (although the generic square matrix is, i.e., the diagonable matrices are dense). But every (square) matrix is similar to an upper triangular matrix, and this often helps already in the study of $f(A)$. In fact, such a similarity is available with V unitary, i.e., $V^HV = I$ (which is the complex version of an orthogonal matrix). The resulting upper triangular matrix **unitarily similar** to A is called the **Schur form** for A; see below.

Since such an upper triangular matrix T is invertible iff all its diagonal entries are nonzero, it follows that $\lambda - A$ is not invertible iff $\lambda - T$ is not invertible iff $\lambda = T(j,j)$ for some j. Thus construction of an upper triangular matrix similar to A requires (implicitly or explicitly) the construction of the **spectrum** of A, i.e., the set

$$\lambda(A) := \{\lambda \in \mathbb{F} : \lambda - A \text{ is not invertible } \}.$$

The points in the spectrum of A are called **eigenvalues** of A, and any vector in $\ker(\lambda - A)$ is called an **eigenvector** of A belonging to λ. In particular, v_1 must be an eigenvector of A belonging to the eigenvalue $T(1,1)$.

This anticipates the result proved later that every A is similar to an upper triangular matrix, and this proof requires us to know that every A has eigenvalues. This latter fact is most quickly seen by considering the **minimal polynomial** for A, i.e., the polynomial p of smallest degree for which $p(A) = 0$.

(10)Lemma. *For every A, there exists a unique monic polynomial p of minimal degree for which $p(A) = 0$. Any root of the minimal polynomial is an eigenvalue of A and conversely.*

Indeed, with A of order n, the n^2+1 powers $A^0, A^1, \ldots, A^{n^2}$ must be linearly dependent (since they lie in the n^2-dimensional space $\mathbb{R}^{n\times n}$). This means that, for some nontrivial coefficient vector $[a(0), a(1), \ldots, a(n^2)]$, $p(A) := \sum_j A^j a(j) = 0$. This shows that there are such nontrivial **annihilating** polynomials p, hence there must be one of smallest degree and with leading coefficient 1.

If now λ is any root of this p, then $p = (\cdot - \lambda)q$, hence $0 = p(A) = (A-\lambda)q(A)$. If now $A-\lambda$ were invertible, then it would follow that already $q(A) = 0$ even though q is of smaller degree than p, a contradiction to the assumed minimality of p. Conversely, for any λ, $p - p(\lambda) = (\cdot - \lambda)q$, hence $-p(\lambda)I = p(A) - p(\lambda) = (A-\lambda)q(A)$, and the left side is invertible when $p(\lambda) \neq 0$, therefore so must the right side be. Hence $p(\lambda) \neq 0$ implies $\lambda \notin \lambda(A)$. □

In principle, it is possible to construct the minimal polynomial in finitely many flops (though there is that practical difficulty of determining the smallest j for which $A^j \in \mathrm{span}(A^k)_{k<j}$ since this requires the decision of whether something is zero). With the minimal polynomial p in hand, the problem of calculating the spectrum $\lambda(A)$ of A is 'reduced' to the problem of polynomial root-finding. In fact, the two problems

are *equivalent*. E.g., MATLAB finds the roots of an arbitrary monic polynomial $p =: ()^n - \sum_{j=1}^n ()^{j-1} a(j)$ by finding the eigenvalues of its **companion matrix**

$$A_p := \left[\begin{array}{ccc|c} & 0 & & \\ -- & -- & -- & \\ & & & a \\ & I & & \\ & & & \end{array}\right], \qquad \text{i.e. } A_p e_j = \begin{cases} e_{j+1}, & j < n; \\ \sum_k e_k a(k), & j = n. \end{cases}$$

I claim that p is the minimal polynomial for A_p: First, $A_p^{j-1} e_1 = e_j$ for $j \le n$, hence $A_p^0, A_p^1, \ldots, A_p^{n-1}$ are linearly independent (since the vectors $A_p^0 e_1, A_p^1 e_1, \ldots, A_p^{n-1} e_1$ are), therefore A_p's minimal polynomial must have degree at least n. Further, $A_p^n e_1 = A_p e_n = \sum_k e_k a(k) = \sum_k A_p^{k-1} e_1 a(k)$, hence $p(A_p) e_1 = 0$. But then, for any j, $p(A_p) e_j = p(A_p) A_p^{j-1} e_1 = A_p^{j-1} p(A_p) e_1 = 0$, therefore $p(A_p) = 0$.

We conclude from (10)Lemma that every matrix has eigenvalues, because the fundamental theorem of algebra assures us that every nontrivial polynomial has roots, provided that we admit complex numbers as roots (and eigenvalues). For this reason, we now switch from the reals to the complex numbers. This means that we use the **conjugate transpose** or **Hermitian** A^H instead of A^T and, correspondingly, use **unitary** matrices U, i.e., square matrices with $U^H U = I$, instead of orthogonal ones. We are also more interested in a **hermitian** matrix , i.e., A with $A^H = A$, than in a symmetric one. We will also be concerned with **normal** matrices, i.e., matrices with $A^H A = A A^H$.

One calls the linear subspace $\operatorname{ran} X$ **invariant** for A (or, an invariant subspace of A) if $AX \subset \operatorname{ran} X$, i.e., if $AX = XB$ for some B. Thus, $T := V^{-1} A V$ is upper triangular iff for each j, $V_{\le j} := [v_1, \ldots, v_j]$ spans an invariant subspace for A.

(11)Lemma. *$AX = XB$ for some (not necessarily square) $X \neq 0$ implies that $\lambda(A) \cap \lambda(B) \neq \emptyset$.*

Note that, if here X is invertible, then A and B are similar, hence $\lambda(A) = \lambda(B)$ in that case.

For the **proof** of the Lemma, write $X = U \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} V^H$ with U and V unitary, and Σ diagonal of order $r := \operatorname{rank}(X)$, hence invertible. Then, with $C := U^H A U, D := V^H B V$,

$$\begin{bmatrix} C_{11}\Sigma & 0 \\ C_{21}\Sigma & 0 \end{bmatrix} = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix} \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} = U^H A X V = U^H X B V = \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix} = \begin{bmatrix} \Sigma D_{11} & \Sigma D_{12} \\ 0 & 0 \end{bmatrix}.$$

Therefore, $C_{11}\Sigma = \Sigma D_{11}$, i.e., C_{11} and D_{11} are similar, hence $\lambda(C_{11}) = \lambda(D_{11}) \neq \emptyset$. Also, $C_{21} = 0, D_{21} = 0$, consequently

$$\lambda(A) = \lambda(C) \supset \lambda(C_{11}) = \lambda(D_{11}) \subset \lambda(D) = \lambda(B).$$

□

One calls the linear subspace $\operatorname{ran} X$ **reducing** for A if it is invariant for A and some complementary subspace (i.e., some subspace $\operatorname{ran} Y$ with $\operatorname{ran} X \cap \operatorname{ran} Y = \{0\}$ and $\operatorname{ran} X + \operatorname{ran} Y = \operatorname{dom}(A)$) is also invariant for A. Assuming that X is 1-1, this means that there is an extension of X to a basis $V = [X, Y]$ for $\operatorname{dom}(A)$ so that both $\operatorname{ran} X$ and $\operatorname{ran} Y$ are A-invariant. In that case,

$$V^{-1} A V = T := \begin{bmatrix} T_{11} & 0 \\ 0 & T_{22} \end{bmatrix}$$

with $A_{|\operatorname{ran} X} = X T_{11} X^{-1}$ and $A_{|\operatorname{ran} Y} = Y T_{22} Y^{-1}$.

Having obtained such a pair of reducing spaces for A, we can look at each of these and try to break them up into smaller reducing spaces, and if we succeed, we can try to break the smaller spaces into yet smaller reducing spaces. Since we are starting off with a finite-dimensional space, we will, after finitely many such attempts, reach a basis $V = [V_1, \ldots, V_s]$ for which each $\operatorname{ran} V_j$ is a minimal A-invariant space. We call such a basis V **maximally reducing** for A. The corresponding matrix $J = V^{-1} A V$ is block-diagonal, and, with a natural choice of the basis V_j for the 'summand' $\operatorname{ran} V_j$, all j, it is the **Jordan canonical form** for

A. From a computational point of view, this canonical form is of no interest since it does not depend stably on A.

Instead, Numerical Analysts work with the Schur form, which is only upper triangular, but does depend stably on A. In fact, the basis for it is even unitary. **Unitary similarity** is particularly important since it preserves properties concerning the scalar product such as **normality**, i.e., the property that $A^HA = AA^H$, or the conjugate symmetry $A^H = A$ of a **hermitian** A. Thus, if $A = VBV^{-1}$ for a unitary V, then $V^{-1} = V^H$, hence $A^H = VB^HV^{-1}$, i.e., the hermitian of A is represented by the hermitian of the representer of A. In particular, A is normal [hermitian] iff B is normal [hermitian].

The essential point for the understanding of the Schur form is the following

(12)Lemma. *If X is 1-1 and $\operatorname{ran} X$ is A-invariant, i.e., $AX = XB$ for some B, then, with U an orthonormal basis for $\operatorname{ran} X$, and $V = [U, W]$ an orthonormal basis for $\operatorname{dom}(A)$, $T := V^HAV$ is block upper triangular, i.e.,*

$$V^HAV = \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix}.$$

Moreover, if A is normal, then V^HAV is block-diagonal, i.e., $T_{12} = 0$.

For the **proof**, note that $T_{21} = W^HAU = 0$ since $A \operatorname{ran} U \subset \operatorname{ran} U \perp W$. As to the rest, if A is normal, then also T must be normal. In particular,

$$T_{11}T_{11}^H + T_{12}T_{12}^H = T_{11}^HT_{11}. \tag{13}$$

But, for any matrix C,

$$\operatorname{trace}(CC^H) = \sum_j (CC^H)(j,j) = \sum_j \sum_k |C(j,k)|^2 = \|C\|_F^2 = \operatorname{trace}(C^HC).$$

Thus (13) reads

$$\|T_{11}\|_F^2 + \|T_{12}\|_F^2 = \|T_{11}\|_F^2,$$

which implies $T_{12} = 0$. □

Corollary (Schur Form). *For any square matrix A, can find a unitary matrix V so that V^HAV is upper triangular, i.e.,*

$$V^HAV = \Lambda + N,$$

with Λ diagonal and N strictly upper triangular.

The inductive **proof** relies on the fact that every square matrix has an eigenvalue, λ, say. If v_1 is a corresponding eigenvector of unit length, then any extension to an o.n. basis V will give

$$V^HAV = \begin{bmatrix} \lambda & b^H \\ 0 & B \end{bmatrix}.$$

Induction provides a unitary U so that $T := U^HBU$ is upper triangular, hence

$$W^HAW = \begin{bmatrix} \lambda & b^H \\ 0 & T \end{bmatrix},$$

with $W := V \operatorname{diag}(1, U)$ again unitary, is upper triangular. □

Note that we had the choice here of any particular $\lambda \in \lambda(A)$, hence we can prescribe the order in which we would like to have the eigenvalues of A appear in the Schur form.

Corollary. *A is normal iff A is unitarily similar to a diagonal matrix. A is hermitian iff A is unitarily similar to a real diagonal matrix.*

For the **proof**, recall that, with V unitary, VBV^H is normal [hermitian] iff B is normal [hermitian]. Hence, need only prove that a normal matrix is unitarily similar to a diagonal matrix. But, with $A = VTV^H$ and T upper triangular, (12)Lemma implies that T must be diagonal. □

Note that, for $T := V^{-1}AV$ upper triangular,

$$\det(t-A) = \det(V)\det(t-T)\det(V^{-1}) = \det(t-T) = \prod_j (t - T(j,j)).$$

This shows that $\lambda \in \lambda(A)$ iff λ is a root of the **characteristic polynomial of** A, i.e., the polynomial $t \mapsto \det(t-A)$. Moreover, each eigenvalue of A appears as a diagonal entry of such an upper triangular matrix similar to A exactly as many times as it appears as a root of the characteristic polynomial. This is the **algebraic multiplicity** of the eigenvalue.

Note that, with $A = V(\Lambda + N)V^H$ a Schur form for A, $\|A\|_F^2 = \|\Lambda\|_F^2 + \|N\|_F^2$. Since $\|\Lambda\|_F^2 = \sum_j |\lambda_j|^2$ (in which each eigenvalue of A appears according to its algebraic multiplicity) depends only on A, as does $\|A\|_F$, this shows that $\|N\|_F$ is independent of how we choose the unitary matrix V. The number $\|N\|_F^2$ is called A's **departure from normality**. This shows that, for a nonnormal matrix, we have to go to nonunitary bases V if we want to make $V^{-1}AV$ simpler than just upper triangular.

For this, assume that T is block upper triangular,

$$T = \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix} \tag{14}$$

with T_{11}, T_{22} square (but not necessarily of the same order). We look for a similarity transformation $Y^{-1}TY$ which leaves the two diagonal blocks as well as $0 = T_{21}$ unchanged, but turns that block T_{12} into zero. In the general case, this forces Y to be of the block form

$$Y = \begin{bmatrix} I & Z \\ 0 & I \end{bmatrix},$$

and so gives

$$Y^{-1}TY = \begin{bmatrix} I & -Z \\ 0 & I \end{bmatrix} \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix} \begin{bmatrix} I & Z \\ 0 & I \end{bmatrix} = \begin{bmatrix} T_{11} & T_{12} + T_{11}Z - ZT_{22} \\ 0 & T_{22} \end{bmatrix}.$$

Hence this similarity is successful provided we can choose Z so that

$$\varphi(Z) := T_{11}Z - ZT_{22} = -T_{12}.$$

Since the map φ so defined is linear and $D(\varphi) = T(\varphi)$, we will succeed, in general, iff φ is invertible, i.e., iff φ is 1-1. For this reason, the following lemma is important.

Lemma. *The linear map $\varphi : X \mapsto AX - XB$ is invertible iff $\lambda(A) \cap \lambda(B) = \emptyset$.*

For the **proof**, note that if $\lambda \in \lambda(A) \cap \lambda(B)$, then $Av = \lambda v$ and $w^H B = \lambda w^H$ for some nontrivial v and w, consequently, $vw^H \neq 0$, yet $\varphi(vw^H) = Avw^H - vw^H B = \lambda vw^H - \lambda vw^H = 0$. The converse was proved in (11)Lemma. □

Corollary. *Every A is similar to a block diagonal matrix $J = \operatorname{diag}(J_1, J_2, \ldots, J_t)$ with $\lambda(J_s) = \{\lambda_s\}$, all s, and $\lambda_i \neq \lambda_j$ for $i \neq j$, hence $\lambda(A) = \{\lambda_j : j = 1, \ldots, t\}$.*

Indeed, if we take J to be the block-diagonal matrix similar to A obtained from a maximal reducing basis X for A, then necessarily each block J_s has only one eigenvalue. For, if such J_s had more than one, then we could get its Schur form T in the form (14) with both T_{11} and T_{22} of positive order and with disjoint spectra. But this would imply that $\operatorname{ran}(T_{11})$ and $\operatorname{ran}(T_{22})$ are nontrivial reducing spaces for J_s, hence for A, thus contradicting the maximality of the reducing basis we started out with. □

From this Corollary, it is only a small step, along the same line of argument, to the Jordan canonical form. Here is the step, described for the typical minimal reducing subspace $Z := \operatorname{ran} V_s$ which provides the diagonal block J_s in a maximally refined block-diagonal representation $J = \operatorname{diag}(J_1, \ldots, J_t) = V^{-1}AV$ for A, with $V = [V_1, \ldots, V_t]$ the corresponding maximal reducing basis.

On $Z = \operatorname{ran} V_s$, $A_{|Z}$ takes the simple form $\lambda_s + N$, with $\lambda(N) = \{0\}$. By (10)Lemma, this implies that N is **nilpotent**, i.e., $N^q = 0$ for some q. The smallest q for which this holds is called the **degree of nilpotency** of N. We claim that necessarily this smallest q equals $\dim Z$. Indeed, with q as defined, there must be $x \in Z$ for which $N^{q-1}x \neq 0$, and, that being so, there must be $y^H \in Z'$ (e.g., $y = N^{q-1}x$) for which $y^H N^{q-1}x \neq 0$. Define $x_j := N^{q-j}x, y_i^H := y^H N^{i-1}$, all $i, j = 1, \ldots, q$, and consider their **Gramian**, i.e., the matrix $Y^H X := [y_1, \ldots, y_q]^H [x_1, \ldots, x_q]$. We compute

$$\big(Y^H X\big)(i,j) = y_i^H x_j = y^H N^{i-1+q-j} x = \begin{cases} 0, & \text{for } i > j; \\ y^H N^{q-1} x \neq 0, & \text{for } i = j. \end{cases}$$

This means that $Y^H X$ is triangular with nonzero diagonals, hence invertible. This implies that both X and Y are 1-1, and that $\operatorname{ran} X \cap \ker Y^H = \{0\}$, therefore $\operatorname{ran} X$ and $\ker Y^H$ are complementary subspaces for Z. Moreover, $\operatorname{ran} X$ is N-invariant, since $Nx_j = x_{j-1}$ or 0, all j, while also $\ker Y^H$ is N-invariant, since $Y^H x = 0$ is equivalent to $y^H N^i x = 0$ for all i, and this implies that also $Y^H N^i N x = 0$ for all i. We conclude that $\operatorname{ran} X$ and $\ker Y^H$ are complementary subspaces which are both N-invariant, hence they are reducing (for N, hence also for $A_{|Z}$). Since Z cannot be further reduced, one of these subspaces must be trivial. Since $\operatorname{ran} X$ is not, it must be $\ker Y^H$. Conclusion: $q = \dim \operatorname{ran} X = \dim Z$, i.e., X is a basis for Z. Further, $Ax_j = (\lambda_s + N)x_j = \lambda_s x_j + x_{j-1}$, all j (with $x_0 := N^q x = 0$). Therefore,

$$A_{|Z} = X^{-1} \begin{bmatrix} \lambda_s & 1 & & \\ & \cdot & \cdot & \\ & & \cdot & \cdot \\ & & & \cdot & 1 \\ & & & & \lambda_s \end{bmatrix} X,$$

and this shows the promised typical **Jordan block** in the **Jordan canonical form**. Note that $\ker N = \operatorname{span} x_1$, hence, up to scalar multiples, x_1 is the unique eigenvector for A in $Z = \operatorname{ran} X = \operatorname{ran} V_s$.

We conclude that, for each $\lambda \in \lambda(A)$, there are exactly $\dim \ker(\lambda - A)$ Jordan blocks associated with λ. This is the **geometric** multiplicity of λ. In contrast, the **algebraic** multiplicity of λ is the sum of the orders of the Jordan blocks associated with λ, i.e., the number of times λ appears as a diagonal element of J. This implies that the algebraic multiplicity is at least as large as the geometric multiplicity. One calls λ a **defective** eigenvalue if the two multiplicities do not agree. Thus A is diagonable iff none of its eigenvalues is defective, i.e., iff all its Jordan blocks are 1×1, i.e., iff there are enough eigenvectors to staff a basis.

The order $\alpha = \alpha(\lambda)$ of the largest Jordan block for λ is called the **ascent** of λ. It is the smallest r for which $\ker(A - \lambda)^r = \ker(A - \lambda)^{r+1}$. It is also the order to which λ is a root of the minimal polynomial for A. The space $\ker(A - \lambda)^\alpha = \cup_r \ker(A - \lambda)^r$ is the **eigenspace** associated with λ. It is a reducing space, the complementary A-invariant space is $\operatorname{ran}(A - \lambda)^\alpha$.

perturbations Since eigenvalue calculations in finite-precision arithmetic only produce approximations, it is important to understand the continuity of the map

$$A \mapsto \lambda(A).$$

Right-off, there is some question as to whether to take the spectrum $\lambda(A)$ as a **subset** or a **sequence** or a **suite**. In the first case, we would only pay attention to the eigenvalues and not their multiplicities, in the second case, we would write them down including (algebraic) multiplicities, e.g., we would think of the spectrum as the sequence of diagonal entries in a Schur form for A. This is a bit unsatisfactory, since there is arbitrariness in the order in which the eigenvalues appear in such a sequence. For this reason, a **suite** or **multiset** (or **bag**, like a bag of marbles) is more satisfactory, for here we would merely record the set of eigenvalues together with their (algebraic) multiplicities, but not enforce any particular order.

Theorem. *The map $A \mapsto \lambda(A)$ from the matrix A to the suite $\lambda(A)$ is continuous.*

The **proof** is all yours (see Problem P.7.1-6).

All standard perturbation arguments rely on the following observation. If $\mu \in \lambda(B+E)\backslash\lambda(B)$, then $B-\mu$ is invertible, while $B+E-\mu = (B-\mu)(I+(B-\mu)^{-1}E) = (I+E(B-\mu)^{-1})(B-\mu)$ is not, hence

$$\|(B-\mu)^{-1}\|\,\|E\| \geq \left\{ \begin{array}{c} \|(B-\mu)^{-1}E\| \\ \|E(B-\mu)^{-1}\| \end{array} \right\} \geq 1$$

in *any* norm. Here are some examples.

Gershgorin. *$\lambda(A) \subset \bigcup_i D_i$, with*

$$D_i := \{z \in \mathbb{C} : |z - A(i,i)| \leq \sum_{j \neq i} |A(i,j)|\}$$

the typical **Gershgorin disk**. *Moreover, any isolated disk will contain exactly one eigenvalue of A.*

For the **proof**, let $D := \mathrm{diag}(A(1,1), A(2,2), \ldots)$ and take $\lambda \in \lambda(A)\backslash\lambda(D)$. Then, with $N := A - D$,

$$A - \lambda = (D-\lambda) + N = (D-\lambda)(I + (D-\lambda)^{-1}N),$$

hence $\|(D-\lambda)^{-1}N\| \geq 1$. By taking the max-norm, we get

$$1 \leq \max_i \sum_j |A(i,i)-\lambda|^{-1}|N(i,j)| = (\max_i \sum_{j \neq i} |A(i,j)|)/|A(i,i)-\lambda|.$$

The continuity of the function $t \mapsto \lambda(D+tN)$ provides the conclusion about isolated disks. □

Note that, by going to A^H, (or by using the 1-norm instead, but with the factor $(D-\lambda)^{-1}$ on the other side), we also know that $\lambda(A) \subset \bigcup_i D_i$, with

$$D_i := \{z \in \mathbb{C} : |z - A(i,i)| \leq \sum_{j \neq i} |A(j,i)|\}.$$

Bauer-Fike. *$\mathrm{dist}\,(\lambda(A+E), \lambda(A)) \leq \kappa(X)\|E\|$, with X any basis which diagonalizes A, and in any norm for which $\|\,\mathrm{diag}(d_1, d_2, \ldots)\| = \|d\|_\infty$.*

For the **proof**, take $\mu \in \lambda(A+E)\backslash\lambda(A)$ and set $\Lambda := X^{-1}AX$. Then $A+E-\mu = X(\Lambda - \mu + X^{-1}EX)X^{-1}$ is not invertible, therefore

$$\|(\Lambda-\mu)^{-1}X^{-1}EX\| \geq 1.$$

But, $\|(\Lambda-\mu)^{-1}\| = \max_i |\lambda_i - \mu|^{-1} = 1/\,\mathrm{dist}\,(\mu, \lambda(A))$. □

If A is not diagonable, one can at times get good results from the Schur form.

Schur. $\operatorname{dist}(\lambda(A+E),\lambda(A)) \leq \max\{\theta,\theta^{1/q}\}$, *with* $\theta := \|E\|_2 \sum_{k<q} \|N\|_2^k$ *and* $\Lambda + N$ *a Schur form for* A, *and with* q *the degree of nilpotency of the strictly upper triangular part* N *of the Schur form.*

For the **proof**, take $\mu \in \lambda(A+E)\backslash\lambda(A)$. Then $I-(\mu-A)^{-1}E$ is defined and not invertible. Therefore,

$$1 \leq \|(\mu-A)^{-1}E\|_2 \leq \|(\mu-A)^{-1}\|_2\|E\|_2.$$

But $\|(\mu-A)^{-1}\|_2 = \|(\mu-\Lambda)-N^{-1}\|_2$ while N is nilpotent of degree q, hence

$$(\mu-A)^{-1} = \sum_{k<q} \big((\mu-\Lambda)^{-1}N\big)^k (\mu-\Lambda)^{-1},$$

and so, with $\delta := \operatorname{dist}(\mu,\lambda(A)) = 1/\|\mu-\Lambda\|_2$, get

$$1 \leq \sum_{k<q} \big(\|N\|_2/\delta\big)^k \|E\|_2/\delta \leq \|E\|_2 \sum_{k<q} \|N\|_2^k \ \max\{\delta^{-0},\delta^{-1},\ldots,\delta^{-q+1}\}/\delta.$$

Consequently, $\delta \max\{1,\delta^{q-1}\} \leq \theta$, etc. □

Whether by Bauer-Fike or by Schur, we see potential sensitivity of eigenvalues in case A is strongly non-normal, for then either $\kappa(X)$ or $\|N\|_2^q$ is large. Put positively, for a normal A,

$$\operatorname{dist}(\lambda(A+E),\lambda(A)) = O(\|E\|_2).$$

Such strong continuous dependence is available for individual eigenvalues to the extent that they are 'normal'.

Specifically, assume that $\lambda \in \lambda(A)$ is a **simple** eigenvalue, i.e., there is just one Jordan block for λ and this block is of order 1. Then this situation will persist in a sufficiently small neighborhood of A. Specifically, for an arbitrary F, with $\|F\|_2 = 1$ say, there exists an $\varepsilon = \varepsilon(A,\lambda,F) > 0$ so that, for all $t \in [-\varepsilon,\varepsilon]$, there will be $\lambda(t) \in \lambda(A+tF)$ with corresponding normalized eigenvector $x(t)$ so that the map $t \mapsto [\lambda(t),x(t)]$ is smooth.

Let y^H be a normalized left eigenvector of A belonging to λ. Claim: $y^Hx \neq 0$. Indeed, if V is a maximal reducing basis for A and, without loss of generality, $J := V^{-1}AV$ has λ as its first diagonal element, then v_1 is necessarily an eigenvector for A belonging to λ, hence, as $\ker(\lambda-A)$ is one-dimensional by assumption, equal to $x = x(0)$ except for a scaling factor. Assume without loss that, in fact, $V = [x,\ldots]$, and consider $[w,\ldots] := (V^{-1})^H$. Then

$$w^HA = w^H(VJV^{-1}) = e_1^HJV^{-1} = \lambda e_1^HV^{-1} = \lambda w^H,$$

hence w^H is a nontrivial left eigenvector for A belonging to λ, and $w^Hx = 1$. Since also $\ker(\lambda-A)^H$ is one-dimensional, it follows that any left eigenvector y^H belonging to λ is a scalar multiple of w^H, hence $y^Hx \neq 0$.

With this, we can differentiate the equation

$$(A+tF)x(t) = \lambda(t)x(t)$$

and evaluate at $t=0$ to get

$$A\dot{x}(0) + Fx = \dot{\lambda}(0)x + \lambda\dot{x}(0),$$

apply y^H to both sides to get $y^HFx = \dot{\lambda}(0)y^Hx$ (since $y^HA\dot{x}(0) = \lambda y^H\dot{x}(0)$), hence

$$|\dot{\lambda}(0)| = |y^HFx|/|y^Hx| \leq 1/|y^Hx|.$$

This shows that a simple eigenvalue is sensitive to perturbations to the extent that its normalized left and right eigenvector fail to be parallel.

notes for mar.15 latest change: 19feb09

The following analysis is valid for any compact linear map on a normed linear space, hence useful in the study of the eigenstructure of differential and integral operators and their approximation. But I find it helpful even in the simple case of a matrix discussed here.

resolvent The map

$$R : \mathbb{C}\backslash\sigma(A) \to \mathbb{R}^{n\times n} : z \mapsto (A-z)^{-1}$$

is called the **resolvent** of A. It is continuous on its domain (since $\mathbb{C} \to \mathbb{R}^{n\times n} : z \mapsto A - z$ is). This implies that R is bounded on any compact subset of $\backslash\sigma(A)$.

R is also differentiable, hence analytic, since

$$R(z) - R(z') = R(z)((A-z') - (A-z))R(z') = (z-z')R(z)R(z'),$$

therefore $dR(z)/dz = (R(z))^2$. This makes complex variable theory available as a convenient and powerful tool for the analysis of the resolvent and, ultimately, the spectral properties of A.

Remark If you are uncomfortable with operator-valued functions, consider instead the complex-valued function $z \mapsto \lambda R(z)x$ for arbitrary $\lambda \in X^*$ and arbitrary $x \in X$.

The spectrum $\sigma(A)$ comprises the singularities of R.

(15) Theorem. *Suppose that*

$$\sigma(A) = \sigma_1 \dot\cup \sigma_2$$

and that it is possible to find a simple closed curve Γ in $\mathbb{C}\backslash\sigma(A)$ which encloses σ_1 and excludes σ_2. Then (i)

$$P := \frac{-1}{2\pi\mathrm{i}} \int_\Gamma R(z)\,\mathrm{d}z \tag{16}$$

defines a bounded lprojector which commutes with A, hence (ii) both $M_1 := \operatorname{ran} P$ and $M_2 := \operatorname{ran}(1-P) = \ker P$ are A-invariant closed lss's, and $X = M_1 \oplus M_2$, and (iii) $A_j := A_{|M_j} \in bL(M_j)$, and (iv) $\sigma(A_j) = \sigma_j$.

Proof: (i) P is bounded (e.g., by $(|\Gamma|/2\pi)\max\|R(\Gamma)\|$). To see that P is a projector, note that P is unchanged if we deform Γ as long as we don't cross $\sigma(A)$ in the process. Therefore

$$\begin{aligned} P^2 = \frac{-1}{2\pi\mathrm{i}}\int_\Gamma \frac{-1}{2\pi\mathrm{i}}\int_{\Gamma'} R(z)R(z')\,\mathrm{d}z'\,\mathrm{d}z &= \frac{-1}{2\pi\mathrm{i}}\int_\Gamma \frac{-1}{2\pi\mathrm{i}}\int_{\Gamma'} \frac{R(z)-R(z')}{z-z'}\,\mathrm{d}z'\,\mathrm{d}z \\ &= \frac{-1}{2\pi\mathrm{i}}\int_\Gamma R(z)\,\mathrm{d}z = P, \end{aligned}$$

the crucial equality by Cauchy's formula if we (as we may) choose Γ' close to Γ but enclosing it, hence

$$\int_{\Gamma'} \frac{R(z)}{z-z'}\,\mathrm{d}z' = R(z)\int_{\Gamma'} \frac{\mathrm{d}z'}{z-z'} = -2\pi\mathrm{i}R(z),$$

while

$$\int_\Gamma \frac{R(z')}{z-z'}\,\mathrm{d}z = R(z')\int_\Gamma \frac{\mathrm{d}z}{z-z'} = 0.$$

(ii) Since $R(z) = (A-z)^{-1}$ commutes with A, so does P, hence M_j is invariant under A (i.e., $AM_j \subset M_j$).

(iii) Therefore $A_j := A_{|M_j} \in L(M_j)$ and

$$R_j(z) := R(z)_{|M_j} = (A_j - z)^{-1}.$$

(iv) In particular, $\sigma(A_j) \subset \sigma(A)$, while if both $(A_1 - z)$ and $(A_2 - z)$ are boundedly invertible, then so is $(A - z)$. Therefore

$$\sigma(A_1) \cup \sigma(A_2) = \sigma(A).$$

On the other hand, for any z' outside Γ,

$$R(z')P = \frac{-1}{2\pi\mathrm{i}}\int_\Gamma R(z')R(z)\,\mathrm{d}z = \frac{-1}{2\pi\mathrm{i}}\int_\Gamma \frac{R(z')-R(z)}{z'-z}dz = \frac{1}{2\pi\mathrm{i}}\int_\Gamma \frac{R(z)}{z'-z}\,\mathrm{d}z$$

which shows that $R_1 = RP_{|M_1}$ has no singularities outside Γ. Correspondingly, R_2 has no singularities inside Γ. Hence, $\sigma(A_j) = \sigma_j$. □

M_1 **contains** $\ker(A-z)^r$ **for any** r **and any** $z \in \sigma_1$ **since** $(A-z)^r x = 0$ **implies** $(A-z)^r(1-P)x = 0$**, yet** $(A-z)$ **is invertible on** $M_2 = \operatorname{ran}(1-P)$**, hence** $(1-P)x = 0$**.**

We call such P a **spectral projector** for A and note that it commutes with any spectral projector Q of A. Further, P and Q are **disjoint** (i.e., $PQ = QP = 0$) in case the curves used in their definition exclude one another.

Note that $P = 1$ in case $\sigma_2 = \emptyset$. This is a special case of the assertion that

$$f(A) = \frac{1}{2\pi \mathrm{i}} \int_\Gamma f(z)(z-A)^{-1}\,\mathrm{d}z \tag{17}$$

in case f is analytic on an open set containing $\sigma(A)$, and Γ is chosen in that open set and surrounding $\sigma(A)$. Note the exact analogue to Cauchy's formula.

Strictly speaking, (17) is an assertion only in case f is a polynomial or, more generally, a rational function, for then we have an alternative notion of what $f(A)$ might be. For more general f, (17) serves as a natural definition which is sensible since we can approximate such f uniformly by polynomials, hence obtain this definition as the limit.

The spectral projectors provide a reasonable means of singling out just which invariant subspaces of A and $A+E$ to compare when considering the sensitivity of an invariant subspace of A. They are precisely those reducing subspaces associated with a certain subset of the eigenvalues of A, i.e., the subspaces $S := \operatorname{ran} P$, with P the spectral projector given by (16) and Γ a curve enclosing some piece of $\lambda(A)$ (but intersecting no part of $\lambda(A)$).

For small enough E, the spectrum of $A + tE$ will not cross Γ as t goes from 0 to 1. In fact, if $\|E\| < \min_{z\in\Gamma} \|(A-z)^{-1}\|^{-1}$, then $(A + tE - z)$ will remain invertible for all $z \in \Gamma$ and all $t \in [0\,..\,1]$ and

$$P_t := \frac{-1}{2\pi \mathrm{i}} \int_\Gamma (A + tE - z)^{-1}\,\mathrm{d}z \tag{18}$$

is well-defined and depends continuously on t. In particular, $\lim_{t\to 0} P_t = P$ and $\operatorname{rank} P_t$ is independent of t. This implies that the corresponding reducing subspace $S_t := \operatorname{ran} P_t$ has the same dimension as S and converges to $S = S_0$ as $t \to 0$. But you'll notice that much depends here on just how close the rest of the spectrum of A is to the part enclosed by Γ, as this influences the minimum of $\|(A-z)^{-1}\|^{-1}$ for $z \in \Gamma$ and thereby the size of the perturbation E allowable.

Now choose a basis V for S. Then we can write $P = VW^T$ for some dual basis W. In fact, W must then be a basis for $\operatorname{ran} P^T$. As

$$P^T = \frac{-1}{2\pi \mathrm{i}} \int_\Gamma (A^T - z)^{-1}\,\mathrm{d}z, \tag{19}$$

we conclude that W is the unique basis dual to V for the reducing subspace of A^T associated with the same piece of $\lambda(A^T) = \lambda(A)$.

Note that $W^T V = 1$, hence $J := W^T AV$ provides the matrix representation for $A_{|S}$ with respect to the basis V.

Consider, in particular, the situation when Γ encircles the eigenvalue λ and no other point from $\lambda(A)$. Then $S = \operatorname{ran} P$ is the eigenspace associated with λ, i.e., $S = \cup_j \ker(A-\lambda)^j = \ker(A-\lambda)^\alpha$, with $\alpha = \alpha(\lambda)$ the **ascent** of λ, i.e., the size of the largest Jordan block associated with λ. Further, $\dim S = a = a(\lambda)$ is the algebraic multiplicity of λ while $\dim\ker(A-\lambda)$ is its geometric multiplicity. Further, $\operatorname{ran} P^T = \ker(A^T-\lambda)^\alpha$.

Now consider perturbations E small enough so that P is 1-1 on $S_1 := \operatorname{ran} P_1$. Then there is a basis V_1 for S_1 for which $PV_1 = V$. We can compare the spectrum of $(A+E)_{|S_1}$ with that of $A_{|S}$ by comparing the matrix representation $J = V^{-1}AV$ for $A_{|S}$ with the matrix representation $J_1 := V_1^{-1}(A+E)V_1$ for $A_{|S_1}$. Specifically, on S, $V^{-1} = W^T$ (since $P = VW^T = 1$ on S) while, on S_1, $V_1^{-1} = W^T$ since $W^T V_1 = W^T P V_1 = W^T V = 1$. Therefore

$$J = W^T AV = W^T APV_1 = W^T PAV_1 = W^T AV_1,$$

while

$$J_1 = W^T(A+E)V_1.$$

Consequently,

$$J_1 - J = W^T E V_1. \tag{20}$$

From this identity, we can read off all kinds of qualitative information, particularly when A is a differential or integral operator and $A + E$ its discrete approximation.

In the present context, we only make use of the immediate consequence

$$\|J_1 - J\| \leq \|W^T\| \, \|E\| \, \|V_1\|. \tag{21}$$

If we start off with an o.n. basis V for S, hence $\|V\|_2 = 1$, then the dual basis W from $\operatorname{ran} P^T$ is uniquely determined. The size of W is a measure of the inclination between $S = \operatorname{ran} P$ and $\operatorname{ran} P^T$. In fact, $\|P\|_2 = \sup_x \|VW^T x\|_2/\|x\|_2 = \sup_x \|W^T x\|_2/\|x\|_2 = \|W\|_2$. For a normal A, P is an orthogonal projector, i.e., W is also o.n. and that is the best situation. But, to the extent that λ fails to be normal, the effect of the perturbation E on λ is magnified. The other factor in (21) is V_1. To first order, its norm will differ from $\|V\|$ by something of order $\|E\|$, hence provide only second-order terms for the difference $J - J_1$.

Generically, the single eigenvalue λ of J breaks up into $a = a(\lambda)$ simple eigenvalues of J_1 as the result of the perturbation. How close these will be to λ will depend on the precise eigenstructure of J.

Claim:

$$|\lambda - \lambda_1| = O(\|J - J_1\|^\alpha) \quad \forall \lambda_1 \in \lambda(J_1). \tag{22}$$

Indeed, since $(\lambda - J)^\alpha = 0$,

$$|\lambda - \lambda_1|^\alpha \leq \|(\lambda - J_1)^\alpha\| = \|(\lambda - J_1)^\alpha - (\lambda - J)^\alpha\| \leq \operatorname{const}_{\|J\|,\|J_1\|} \|J - J_1\|,$$

using the fact that the map $B \mapsto B^\alpha$ is locally Lipschitz continuous. (Note that $B^\alpha - C^\alpha = \sum_{j=1}^{\alpha-1} B^{\alpha-j-1}(B - C)C^j$, hence

$$\|B^\alpha - C^\alpha\| \leq \big(\sum_{j=0}^{\alpha-1} \|B\|^{\alpha-j-1} \|C\|^j \big) \|B - C\|.)$$

On the other hand, the average of all a eigenvalues of J_1 (counting multiplicities) is within $\|J - J_1\|$ of λ since this average equals $\operatorname{trace} J_1/a = \lambda - \operatorname{trace}(J - J_1)/a$.

In a practical situation, one would not be able to identify the reducing subspace S belonging to λ. But, in observing the spectrum of $A + E$ as E becomes smaller, one would be able to recognize that some $\lambda \in \lambda(A)$ is defective by seeing certain of the approximate eigenvalues relatively slowly cluster. If their average then turns out to approach a limit at a much faster rate, the hypothesis of a defective eigenvalue would be confirmed.

power-boundedness The following important theorem is a simple consequence of knowing the Schur form. It relates the **spectral radius**

$$\rho(A) := \max\{|\lambda| : \lambda \in \lambda(A)\}$$

of A to the possible **power-boundedness** of A, i.e., to the condition that

$$\sup_k \|A^k\| < \infty.$$

Since all norms in a finite-dimensional linear space are equivalent, a subset is bounded in one norm iff it is bounded in any other norm, hence power-boundedness holds in all norms if it holds in one.

Particular norms for A can be found in the form $\|VAV^{-1}\|$, with V any basis (this would correspond to taking the map-norm for A corresponding to the vector norm $x \mapsto \|Vx\|$). In particular, *a matrix A is power-bounded iff $\sup_{i,j,k} |B^k(i,j)| < \infty$ for every (some) matrix representation $B = VAV^{-1}$ for A.* But it may be easier to prove it in one norm than in others. For **example**, the matrix

$$A := \begin{bmatrix} .9 & 10^{10} \\ 0 & .9 \end{bmatrix}$$

is power-bounded; in fact, A is **convergent**, i.e., $\lim_k A^k = 0$. But if you look at the first few powers of A in the max-norm or the 1-norm, you'd never guess that A is power-bounded, let alone convergent.

We start with the following

(23)Observation. *Any Jordan matrix J of order > 1 belonging to an eigenvalue λ of absolute value 1 fails to be power-bounded.*

For the **proof**, observe that

$$\begin{bmatrix} \lambda & a \\ 0 & \lambda \end{bmatrix} \begin{bmatrix} \mu & b \\ 0 & \mu \end{bmatrix} = \begin{bmatrix} \lambda\mu & \lambda b + a\mu \\ 0 & \lambda\mu \end{bmatrix},$$

hence, by induction, $\begin{bmatrix} \lambda & 1 \\ 0 & \lambda \end{bmatrix}^k = \begin{bmatrix} \lambda^k & k\lambda^{k-1} \\ 0 & \lambda^k \end{bmatrix}$ and, for $|\lambda| = 1$, the $(1,2)$-entry goes to ∞ with k. □

Remark You have here a very special case of the following remarkable result concerning Jordan-like matrices: For any reasonable function f and Jordan-like matrix

$$J := \begin{bmatrix} \lambda_1 & 1 & & & \\ & \cdot & \cdot & & \\ & & \cdot & \cdot & \\ & & & \cdot & 1 \\ & & & & \lambda_n \end{bmatrix},$$

the matrix $f(J)$ has the divided difference $[\lambda_i, \ldots, \lambda_j]f$ as its (i,j)-entry (with $f(J)(i,j) = 0$ for $i > j$).

(24)Lemma. *For every (square) matrix A and every norm on* $\mathrm{dom}(A)$,

$$\rho(A) \le \|A\|,$$

and this inequality is sharp. Equality is possible in some norm iff all absolutely largest eigenvalues of A are non-defective.

This implies that, for every $\tau > \rho(A)$, there is some map-norm in which $\|A\| < \tau$, hence for which $\|(A/\tau)^k\| < 1$, hence for which $\lim_{k\to\infty} \|(A/\tau)^k\| = 0$. Consequently,

(25)Corollary. *For every (square) matrix A and every norm on* $\mathrm{dom}(A)$, $\|A^k\| = o(\tau^k)$ *for all* $\tau > \rho(A)$.

For the **proof of the lemma**, observe first that, for all $\lambda \in \lambda(A)$, there is some nontrivial eigenvector x, hence $\|A\| \geq \|Ax\|/\|x\| = |\lambda|$, i.e., $\lambda(A)$ lies inside the disk of radius $\|A\|$.

The proof that, by proper choice of the vector norm on $\mathrm{dom}(A)$ we can make the corresponding map norm $\|A\|$ as close as we like to $\rho(A)$, is a bit harder. For this, we begin with an o.n. basis V for which $V^{-1}AV =: \Lambda + N$ with Λ diagonal and N strictly upper triangular. Then we scale this o.n. basis to the basis $W := VD$, with $D := \mathrm{diag}(d^1, d^2, d^3, \ldots)$ for some positive d. Then we find that

$$W^{-1}AW = D^{-1}(\Lambda + N)D = \Lambda + D^{-1}ND,$$

with $D^{-1}ND(i,j) = d^{-i}N(i,j)d^j = d^{j-i}N(i,j)$. Since N is strictly upper triangular, we have $N(i,j) = 0$ for $i \geq j$, hence, for any $d < 1$, $|D^{-1}ND| < d|N|$, hence

$$\|W^{-1}AW\|_\infty \leq \|\Lambda\|_\infty + \|D^{-1}ND\|_\infty \leq \rho(A) + d\|N\|_\infty,$$

and this can be made arbitarily close to $\rho(A)$ by choosing d small enough. By keeping d positive, we insure that W is a basis for $\mathrm{dom}(A)$, hence $x \mapsto \|W^{-1}x\|_\infty$ is a norm on $\mathrm{dom}(A)$ whose corresponding map-norm for A is, indeed, the number $\|W^{-1}AW\|$.

Finally, we can achieve equality in case all absolutely largest eigenvalues of A are non-defective. For, in that case, the space S spanned by a maximally linearly independent set of corresponding eigenvectors (i.e., the sum of the corresponding nullspaces $\ker(A-\lambda)$) is reducing, and, with T a corresponding complementary invariant subspace, we know that $B := A_{|T}$ has spectral radius $< \rho(A) = \|A_{|S}\|$. Therefore, by what we already know, we can find some norm on T for which $\|A_{|T}\|$ is close enough to $\rho(A_{|T})$ to guarantee that it is less than $\rho(A) = \|A_{|S}\|$. But then, in the resulting norm on $\mathrm{dom}(A) = S \oplus T$, $\|A\| = \|A_{|S}\| = \rho(A)$. The converse is proved as part of the following Theorem. □

(26)Theorem. *The (square) matrix A is power-bounded iff either $\rho(A) < 1$ or else $\rho(A) = 1$ with every absolutely largest eigenvalue non-defective.*

By (24)Lemma, A is even convergent in case $\rho(A) < 1$, while A cannot possibly be power-bounded in case $\rho(A) > 1$ (since then, for some λ and some unit vector x, $\|A\| \geq \|A^k x\| = |\lambda|^k \to \infty$). Hence, for the **proof**, it is sufficient to consider A with $\rho(A) = 1$. If every absolutely largest eigenvalue of A is non-defective, then, by (24)Lemma, $\|A\| = 1$ for some map-norm, hence A is power-bounded. If on the other hand, some absolutely largest eigenvalue λ of A is defective, then there is a reducing subspace on which A looks like a Jordan matrix of order > 1 with an eigenvalue of absolute value 1, hence, by (23)Observation, A fails to be power-bounded already on this reducing subspace. □

In a *practical*, i.e., finite-precision arithmetic, setting, this theorem has to be augmented to take account of the fact that we usually cannot know the matrix A better than within a relative error of $\varepsilon := \mathbf{u}\|A\|$, and the same holds for the computed powers. Thus it is not so much A's spectrum $\lambda(A)$ or its spectral radius $\rho(A)$ that predicts the behavior of the computed sequence (A^k) of powers, as it is the **approximate spectrum** $\lambda_\varepsilon(A)$ and the **approximate spectral radius** $\rho_\varepsilon(A) := \sup|\lambda_\varepsilon(A)|$, with

$$\lambda_\varepsilon(A) := \bigcup_{\|E\| \leq \varepsilon} \lambda(A+E) = \{\lambda : \|(A-\lambda)^{-1}\| \geq 1/\varepsilon\}.$$

This definition depends on the particular vector norm used, but, whatever that norm, its dependence on ε is the more interesting feature. The earlier perturbation theory makes clear that, for a normal A (and in the 2-norm), $\rho_\varepsilon(A) = \rho(A) + O(\varepsilon)$, hence some attention has to be paid in case $\rho(A)$ is close to 1. Much more attention has to be paid for a general A since, according to that earlier perturbation theory, $\rho_\varepsilon(A) = \rho(A) + O(\varepsilon^a)$, with a the maximum ascent of all maximal eigenvalues of A. Thus, for non-normal, and particular for defective, matrices, even the noise associated with finite-precision arithmetic (never mind the errors possibly incurred when computing A) could drastically raise the spectral radius of A as used in computations, hence lead to drastically different conclusions concerning the boundedness of its powers. This

is a very important issue in the study of iterative methods, be it for the solution of linear systems per se or for the solutions of discretized ODEs or PDEs.

In the **power method**, one constructs the sequence

$$z_k := A^k z_0, \qquad k = 1, 2, \ldots$$

in hopes that, for large k, the ratios $z_{k+1}(i)/z_k(i)$ provide good estimates for the absolutely largest eigenvalue of A. This hope is justified in case the absolutely largest eigenvalue of A is **dominant**, i.e., is the only absolutely largest eigenvalue, even counting algebraic multiplicity.

For, with λ the dominant eigenvalue for A, $X := \ker(A - \lambda)$ is reducing for A. Further, with Y the corresponding complementary A-invariant subspace, $B := A_{|Y}$ has spectral radius $< \rho(A) = |\lambda|$. Hence, for any $\varepsilon > 0$, we can choose a norm on $\mathrm{dom}(A)$ so that $\|B\| < \rho(B) + \varepsilon$, hence can make $\|B/\lambda\|$ as close to $r := \rho(B)/\rho(A)$ as we like. Further, with $z_0 = x_z + y_z$ for $x_z \in X$ and $y_z \in Y$,

$$z_k = \lambda^k x_z + B^k y_z = \lambda^k (x_z + (B/\lambda)^k y_z).$$

This shows that $z_k/\lambda^k = x_z + O(r^k)$. Consequently, $z_{k+1}(i)/z_k(i) = \lambda + O(r^k)$, provided $x_z(i) \neq 0$.

Thus, to the extent that the ratio $r = \rho(B)/\rho(A)$ of the next-largest eigenvalue to the dominant eigenvalue is less than 1, the power method provides an effective means for obtaining good estimates for the dominant eigenvalue *and for a corresponding eigenvector.*

The method of **orthogonal iterations** is the following generalization of the power method: With Q_0 an orthogonal matrix having p columns, compute

$$Q_k R_k := Z_k := A Q_{k-1} \quad , k = 1, 2, \ldots.$$

Here, $Q_k R_k$ is the QR factorization of Z_k, and the hope is that $\mathrm{ran}\, Q_k$ might converge to an A-invariant subspace. This hope is justified in case A has a **dominant** p-dimensional A-invariant subspace, i.e, a subspace S of dimension p so that

$$\max\{|\lambda| : \lambda \in \lambda(A) \backslash \lambda(A_{|S})\} < \min\{|\lambda| : \lambda \in \lambda(A_{|S})\}.$$

This means that, with $|\lambda_1| \geq |\lambda_2| \geq \ldots$ the spectrum of A ordered by magnitude and counted including multiplicities, $|\lambda_p| > |\lambda_{p+1}|$, and S is the unique invariant subspace associated with the eigenvalues $\lambda_1, \ldots, \lambda_p$. The book uses the notation

$$D_p(A)$$

for this invariant subspace. In this case, we can provide

Theorem 7.3-1. *If $D_p(A)$ is well-defined, then, starting with almost any orthogonal $Q_0 \in \mathbb{C}^{n \times p}$, we have*

$$\mathrm{dist}\,(D_p(A), \mathrm{ran}\, Q_k) \leq c|\lambda_{p+1}/\lambda_p|^k,$$

with c a constant that depends on A's departure from normality and on the **separation** $\mathrm{sep}(T_{11}, T_{22})$ *between the complementary diagonal blocks in an block-upper triangular form for A in which T_{11} represents $A_{|D_p(A)}$.*

Proof: A convenient way to measure the distance between two subspaces S_j is provided by the number

$$\mathrm{dist}\,(S_1, S_2) := \|P_1 - P_2\|_2,$$

with P_j the orthogonal projector onto S_j. Let $[Q_j, U_j]$ be o.n. bases for $\mathbb{C}^n$, with $S_j = \mathrm{ran}\, Q_j$, hence $P_j = Q_j Q_J^H$. Then

$$\begin{aligned}
\|P_1 - P_2\|_2 &= \|Q_1 Q_1^H - Q_2 Q_2^H\|_2 = \|[Q_1, U_1]^H (Q_1 Q_1^H - Q_2 Q_2^H)[Q_2, U_2]\|_2 \\
&= \| \begin{bmatrix} Q_1^H (Q_1 Q_1^H - Q_2 Q_2^H) Q_2 & Q_1^H Q_1 Q_1^H U_2 \\ -U_1^H Q_2 Q_2^H Q_2 & 0 \end{bmatrix} \|_2 = \| \begin{bmatrix} 0 & Q_1^H U_2 \\ -U_1^H Q_2 & 0 \end{bmatrix} \|_2 \\
&= \max\{\|Q_1^H U_2\|_2, \|U_1^H Q_2\|_2\} = \|Q_1^H U_2\|_2 = \|U_1^H Q_2\|_2.
\end{aligned}$$

(Last equality is equivalent to $\|U_{21}\|_2 = \|U_{12}\|_2$ in case $[U_{11}, U_{12}; U_{21}, U_{22}]$ unitary with U_{11}, U_{22} square.) This means that $\operatorname{dist}(D_p(A), \operatorname{ran} Q_k) = \|Q_k^H Q_\beta\|_2$, if

$$Q =: [Q_\alpha, Q_\beta]$$

is an o.n. basis for which $\operatorname{ran} Q_\alpha = D_p(A)$. For such a basis, $T := Q^H A Q$ is necessarily block-upper triangular, i.e.

$$T = Q^H A Q = \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix}.$$

Now consider the iteration. Observe (by induction) that $A^k Q_0 = Q_k(R_k \cdots R_1) =: Q_k R$, hence

$$T^k Q^H Q_0 = Q^H Q_k R.$$

Since $\lambda(T_{11}) \cap \lambda(T_{22}) = \emptyset$, we know that

$$\begin{bmatrix} I & -X \\ 0 & I \end{bmatrix} \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix} \begin{bmatrix} I & X \\ 0 & I \end{bmatrix} = \begin{bmatrix} T_{11} & 0 \\ 0 & T_{22} \end{bmatrix}$$

with $\varphi(X) := T_{11}X - XT_{22} = -T_{12}$. This gives the equation

$$\begin{bmatrix} T_{11}^k & 0 \\ 0 & T_{22}^k \end{bmatrix} \begin{bmatrix} V_0 - XW_0 \\ W_0 \end{bmatrix} = \begin{bmatrix} V_k - XW_k \\ W_k \end{bmatrix} R, \tag{27}$$

with $[V_j; W_j] := Q^H Q_j$, all j, hence

$$\|W_k\|_2 = \operatorname{dist}(D_p(A), \operatorname{ran} Q_k),$$

the quantity we want to show goes to zero as $k \to \infty$.

Now, for almost all orthogonal Q_0, $V_0 - XW_0 = (Q_\alpha^H - XQ_\beta^H)Q_0$ is invertible, since $(Q_\alpha^H - XQ_\beta^H) = [I, -X]Q^H$ is onto (regardless of X since $[I, -X]$ is onto). This implies, from the first part of (27), that

$$R^{-1} = (T_{11}^k(V_0 - XW_0))^{-1}(V_k - XW_k),$$

therefore, from the second part of (27),

$$W_k = T_{22}^k W_0 R^{-1} = T_{22}^k W_0 (T_{11}^k (V_0 - XW_0))^{-1}(V_k - XW_k).$$

This gives the estimate

$$\operatorname{dist}(D_p(A), \operatorname{ran} Q_k) \le \|T_{22}^k\|_2 \, \|T_{11}^{-k}\|_2 \, \|(V_0 - XW_0)^{-1}\|_2 \, \|W_0\|_2 \, \|V_k - XW_k\|_2.$$

Here we estimate $\|V_k - XW_k\|_2 \le (1 + \|X\|_F)$, a constant that depends on the separation of the spectrum of T_{11} and T_{22} but, in any case, does not depend on k, and neither does $\|(V_0 - XW_0)^{-1}\|_2 \, \|W_0\|_2$. But, by the result on power-boundedness, and since $\rho(T_{22}) = |\lambda_{p+1}|$, $\rho(T_{11}^{-1}) = 1/|\lambda_p|$, for any positive ε, $\|T_{22}^k\| = o((|\lambda_{p+1}| + \varepsilon)^k)$, and $\|T_{11}^{-k}\| = o((|\lambda_p| - \varepsilon)^{-k})$, therefore, finally,

$$\operatorname{dist}(D_p(A), \operatorname{ran} Q_k) = o\Big(\Big(\frac{|\lambda_{p+1}| + \varepsilon}{|\lambda_p| - \varepsilon}\Big)\Big).$$

□

The **QR method** can be understood as complete orthogonal iteration. Starting with a unitary Q_0 (i.e., with $p = n$), we can, for each p, interpret $Q_k(:, 1{:}p)$ as the result of having orthogonal iteration starting with $Q_0(:, 1{:}p)$. Consequently, if $|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n|$, then $D_p(A)$ is defined for every p and the theorem tells us that $Q_k(:, 1{:}p)$ converges to a basis for it. This implies that

$$T_k := Q_k^H A Q_k$$

converges to an upper triangular matrix, thus providing a Schur form for A in the limit.

The actual QR method organizes the calculation as follows: We have

$$T_{k-1} = Q_{k-1}^H A Q_{k-1} = Q_{k-1}^H (A Q_{k-1}) = Q_{k-1}^H (Q_k R_k),$$

while

$$T_k = Q_k^H A Q_k = Q_k^H A Q_{k-1} Q_{k-1}^H Q_k = R_k (Q_{k-1}^H Q_k).$$

Thus, T_k is obtained from T_{k-1} by computing the QR factorization

$$T_{k-1} = (Q_{k-1}^H Q_k) R_k$$

and then multiplying its factors in the reverse order. This gives the iteration:

For $k = 1, 2, \ldots$:

step 1: $A =: QR$, with Q unitary and R right-triangular.

step 2: $A := RQ$.

In this raw form, the iteration is not very effective and also quite expensive. Its convergence proof above relies on special assumptions and, as the proof indicated, the convergence rate can be quite slow in case eigenvalues cluster. Also, each step takes a complete QR factorization, i.e. takes $O(n^3)$ flops.

The latter complaint is dealt with by the assumption that A is already in upper Hessenberg form (which can always be achieved by a unitary similarity transformation). Now the computation of $A =: QR$ takes only $O(n^2)$ flops, and, fortunately, RQ is again upper Hessenberg.

The slow convergence is dealt with by (implicit) shifts and deflation. Thus streamlined, the QR method generates the spectrum of a matrix in remarkably fast time. MATLAB provides the command eigmovie(A) which illustrates the iteration for a symmetric A.

Hessenberg matrices An **(upper) Hessenberg** matrix H has zeros below the first subdiagonal, i.e., $H(i,j) = 0$ for $i > j+1$. Such matrices are of importance for eigenvalue calculations because (i) it is possible to bring any matrix of order n into Hessenberg form by $n-2$ Householder reflections, and (ii) one step of the QR iteration applied to a Hessenberg matrix costs only $O(n^2)$ (rather than $O(n^3)$) flops and produces again a Hessenberg matrix.

Hessenberg Q-R step Assuming H is in upper Hessenberg form, its QR factorization is most efficiently computed by Givens rotations. Walking down the diagonal, one rotates rows k and $k+1$ to bring a zero into position $(k+1,k)$, $k = 1, \ldots, n-1$. To compute from this $\bar{H} := RQ = Q^H HQ$, it is only necessary to rotate columns k and $k+1$ in the same way, and in the same order, i.e., for $k = 1, \ldots, n-1$. Since the workarray H was upper triangular after that set of row rotations, columns k and $k+1$ have no entries below the diagonal when it comes time to rotate them, hence the same will be true after their rotation except that there now might be a nonzero $(k+1,k)$ entry. Thus $\bar{H}$ is again upper Hessenberg.

In the procedure, we obtain Q as the product

$$Q = \prod_{k<n} \mathrm{diag}(I, \begin{bmatrix} c_k & s_k \\ -s_k & c_k \end{bmatrix}^H, I)$$

of Givens rotations, with $\begin{bmatrix} c_k & s_k \\ -s_k & c_k \end{bmatrix}$ the rotation applied to rows k and $k+1$ at the kth step, thus producing $R := Q^H H$. For this reason, it was just the right thing to follow this up by rotating columns k and $k+1$ with the very same rotation $\begin{bmatrix} c_k & s_k \\ -s_k & c_k \end{bmatrix}$, and in the order $k = 1, 2., \ldots$, since that constructs $RQ = Q^H HQ$, the new upper Hessenberg matrix similar to the old.

Application of this fast QR iteration requires an initial similarity transformation to bring the original matrix into upper Hessenberg form. This can be done in various ways. One could use Givens rotations as above to zero out all entries below the Hessenberg matrix, following each zeroing out by the corresponding rotation of columns, to keep the similarity. Assuming that one starts at the upper left and works column by column, the zero entries gained during row rotation will not be destroyed by the corresponding column rotation, as the latter will always involve columns to the right of the one currently being worked on.

For the same reason, use of Householder reflection works. Here one would reflect the first column to a vector $[\mathrm{x}; \mathrm{x}; 0; \ldots; 0]$ with just the first two entries nonzero. The corresponding Householder reflection will keep e_1 fixed, hence its subsequent application from the right (to keep the similarity) will not change the first column, i.e., will retain the zeros just gained.

This settles the efficiency of a QR-method step. Next we take on the other objection to the QR-method, its slow or nonexistent convergence. For this, we need some facts about Hessenberg matrices.

unreduced Hessenberg A Hessenberg matrix H is called **unreduced** by the book if its first subdiagonal contains no zero, i.e., if $\prod_i H(i+1,i) \neq 0$. A general Hessenberg matrix breaks apart into its unreduced blocks which may be treated separately if we are after the spectrum of the matrix. Thus the QR iteration and its variants may be viewed as converging toward a Hessenberg matrix with small enough unreduced blocks to make the calculation of the spectrum a triviality.

An unreduced Hessenberg matrix H of order n has its first $n-1$ columns linearly independent. This implies that its kernel has dimension no bigger than 1. In fact, if it has a nontrivial kernel, then necessarily its last column is linearly dependent on the preceding columns. Since the jth column of the right-triangular matrix R in a QR-factorization for H contains the coordinates of the jth column He_j with respect to the basis $Q(:, 1{:}j)$ for a space containing $\mathrm{ran}\, H(:, 1{:}j)$ (and equal to it in case $H(:, 1{:}j)$ is of full rank), the following is true.

(28)Lemma. *If $H = QR$ is a QR-factorization of the unreduced and non-invertible Hessenberg matrix H, then $R(n,n) = 0$.*

(29)Corollary. *If H is unreduced Hessenberg, then, for any $\mu \in \lambda(H)$, the QR-factorization $QR = (H - \mu I)$ must have $R(n,n) = 0$, therefore $\bar{H} := RQ + \mu I$ must be reduced, with $\bar{H}(n, n-1) = 0$.*

This observation is at the heart of the QR-iteration with shifts. Note that the matrix $\bar{H} = RQ + \mu I = Q^H(H - \mu I)Q + \mu I = Q^H HQ$ obtained from H in such a shifted QR-step is again unitarily similar to H; only the way in which the unitary matrix Q was obtained has become a little bit more complicated. It will be important later to know that the Q can be constructed 'implicitly', i.e., without actually carrying out the shifts. This is the content of the next result, which concerns the **Uniqueness of the Q in a unitary similarity to an unreduced Hessenberg** matrix.

(30)(Implicit Q Theorem). *If Q, V are both (real) orthogonal, and $H := Q^T AQ, G := V^T AV$ are both upper Hessenberg, then $q_1 = v_1$ and $H(2,1) \cdots H(k, k-1) \neq 0$ implies $q_j = \pm v_j$ and $|H(j, j-1)| = |G(j, j-1)|$ for $j \leq k+1$.*

Proof: Have $WH = GW$ with $W := V^T Q$ and, by assumption, $w_1 = e_1$. Therefore $Wh_i = Gw_i$, i.e.,

$$H(i+1,i)w_{i+1} = Gw_i - \sum_{j \leq i} H(j,i)w_j.$$

Assuming that we already know that $w_j = \pm e_j$ for $j \leq i$ (as we do when $i = 1$), this implies that $H(i+1,i)w_{i+1} \in \text{ran}[e_1, \ldots, e_{i+1}]$ (since then $Gw_i = Ge_i$ is in that space), hence, if also $H(i+1,i) \neq 0$, then $\text{ran}[e_1, \ldots, e_i] \perp w_{i+1} \in \text{ran}[e_1, \ldots, e_{i+1}]$ which implies that also $w_{i+1} = \pm e_{i+1}$. In particular,

$$H(i+1,i)(\pm e_{i+1}) = G(\pm e_i) - \sum_{j \leq i} H(j,i)(\pm e_j),$$

which implies that $H(i+1,i) = \pm G(i+1,i)$. But this last conclusion also holds if $H(i+1,i) = 0$ since then

$$0 = \pm \sum_{j \leq i+1} G(j,i)e_j - \sum_{j \leq i} H(j,i)(\pm e_j) \in \pm G(i+1,i)e_{i+1} + \text{ran}[e_1, \ldots, e_i],$$

therefore also $G(i+1,i) = 0$. □

Remark. In a complex context, we would merely have to replace here $\pm e_j$ by $\exp(i\theta)e_j$, i.e., by the jth unit vector multiplied by an arbitrary (complex) sign.

Schur is to Jordan as Hessenberg is to Krylov

Breakup into unreduced blocks A general Hessenberg matrix breaks apart into its unreduced blocks, and these may be treated separately if we are after the spectrum of the matrix. Thus the QR iteration and its variants may be viewed as converging toward a Hessenberg matrix with small enough unreduced blocks to make the calculation of the spectrum a triviality.

The calculation is indeed trivial in case such an unreduced block is of order 1 or 2. In the latter case, with $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ the unreduced block, one simply solves the corresponding characteristic equation

$$(a - \lambda)(d - \lambda) = cb$$

for its two roots. This possibility is particularly important since one would like, if possible, to avoid complex arithmetic in case the original matrix A is real. In that case, A may well have complex eigenvalues, but these would come in complex conjugate pairs, hence could be found as the two eigenvalues of an unreduced block of order 2.

This fact can be used, in the proof of the existence of the Schur form, to show (by induction) that, for an arbitrary *real* matrix A, there exists an *orthogonal* matrix V so that $V^T AV$ is block upper triangular, with diagonal blocks of order 1 or 2, and with each order-two diagonal block having two complex conjugate eigenvalues. To make the proof work, one would observe that, for a nonreal eigenvalue $\lambda = \gamma + i\mu \in \lambda(A)$ there must be real vectors y and $z \neq 0$ so that $A(y + iz) = (\gamma + i\mu)(y + iz) = (\gamma y - \mu z) + i(\gamma z + \mu y)$, i.e.

(since A is real) $A[y,z] = [y,z]\begin{bmatrix} \gamma & \mu \\ -\mu & \gamma \end{bmatrix}$. This shows that $\operatorname{ran}[y,z]$ is a two-dimensional invariant subspace for A, hence an o.n. basis $[v_1, v_2]$ for it extended to an o.n. basis V for $\mathbb{R}^n$ provides a block-upper triangular matrix representation for A with an upper diagonal block of order 2. Its lower diagonal block is taken care of by induction hypothesis.

This shows that a real QR-iteration should generically converge to a block upper triangular matrix whose diagonal blocks are of order 1 or 2.

shifts The second objection to the raw QR method concerns its slow rate of convergence. This is remedied by careful use of shifts. Recall that the convergence of the upper left principal submatrix $H(1{:}j, 1{:}j)$ to an upper triangular matrix depended on having a dominant eigenspace of order j, i.e., on having $|\lambda_j| > |\lambda_{j+1}|$ (when we write out the spectrum of A in order, counting multiplicities). The greater the gap, the faster the convergence. Also, once $H(j+1, j)$ is small enough to be zero within the precision used, we declare it to be zero and are then free to concentrate on the smaller matrices $H(1{:}j, 1{:}j)$ and $H((j+1){:}n, j+1{:}n)$.

In principle, we can force a big gap by shifting the spectrum, i.e., by considering $H - \mu I$ rather than H, since now the important sequence is the newly ordered sequence $|\lambda_1 - \mu| \geq |\lambda_2 - \mu| \geq \cdots$ which may show a large gap $|\lambda_{n-1} - \mu| > |\lambda_n - \mu|$ in case μ is close to $\lambda(A)$.

In fact, recall from (29)Corollary that the shifted QR-step

$$H - \mu I =: QR, \qquad \bar{H} := RQ + \mu I$$

with any $\mu \in \lambda(H)$ produces from an unreduced H a similar Hessenberg matrix $\bar{H}$ with $H(n, n-1) = 0$ (and $H(n,n) = \mu$). This gives the hope (often not disappointed) that, with μ a point close to $\lambda(H)$, the very same shifted QR-step will produce a more nearly reduced $\bar{H}$.

In particular, if

$$|H(n, n-1)| \ll \min\{|H(n-1, n-1)|, |H(n,n)|\}, \tag{31}$$

then the shifted QR-step with $\mu := H(n,n)$ is recommended, as it can be shown to result in a **quadratic** convergence to a reduced Hessenberg matrix with the $(n, n-1)-$entry equal to zero, hence its $(n,n)-$entry an eigenvalue.

This observation is based on the following calculation. If

$$H((n-1){:}n, (n-1){:}n) =: \begin{bmatrix} \mathrm{x} & \mathrm{x} \\ \varepsilon & \mu \end{bmatrix}$$

is the lower right principal submatrix of H of order 2, then, before the last row rotation, it will look like

$$\begin{bmatrix} a & b \\ \varepsilon & 0 \end{bmatrix},$$

causing the last rotation to be $\begin{bmatrix} c & s \\ -s & c \end{bmatrix}$ with $s^2 + c^2 = 1$ and $-sa + c\varepsilon = 0$ and giving the new lower right principal submatrix

$$\begin{bmatrix} ca + s\varepsilon & cb \\ 0 & -sb \end{bmatrix},$$

where $\begin{bmatrix} -a & \varepsilon \\ s & c \end{bmatrix}[s; c] = [0; 1]$, hence $s = \varepsilon/(ca + s\varepsilon)$, therefore, since $\begin{bmatrix} c & s \\ -s & c \end{bmatrix}[a; \varepsilon] = [ca + s\varepsilon; 0]$, $s^2 = \varepsilon^2/\|[ca + s\varepsilon; 0]\|_2^2 = (\varepsilon/\|[a;\varepsilon]\|_2)^2$. The subsequent column rotations will affect the $(n, n-1)-$entry only at the last rotation, and this gives

$$H((n-1){:}n, (n-1){:}n) = \begin{bmatrix} ca + s\varepsilon & cb \\ 0 & -sb \end{bmatrix}\begin{bmatrix} c & -s \\ s & c \end{bmatrix} = \begin{bmatrix} \mathrm{x} & \mathrm{x} \\ -s^2 b & \mathrm{x} \end{bmatrix} \bar{H}((n-1){:}n, (n-1){:}n),$$

therefore $\bar{H}(n,n-1) = \varepsilon^2 b/(\varepsilon^2+a^2)$. Consequently, for $\varepsilon \ll |a|$, we get $\bar{H}(n,n-1) = O(|H(n,n-1)|^2)$.

double shift The final detail to be taken care of is the fact that, in case of complex eigenvalues, we cannot expect condition (31) always to hold, i.e., we cannot expect $H(n,n)$ always to be a good approximation to an eigenvalue of H for the simple reason that it is real.

Francis managed to come up with a variant in which one carries out two shifts in succession, by μ_1 and μ_2, with μ_1, μ_2 the eigenvalues of

$$G := H((n-1){:}n, (n-1){:}n) =: \begin{bmatrix} a & b \\ c & d \end{bmatrix}.$$

For, this can be done in real arithmetic and in $O(n^2)$ flops, as follows.

Set

$$\begin{aligned} H - \mu_1 I &=: U_1 R_1, \\ H_1 &:= R_1 U_1 + \mu_1 I, \\ H_1 - \mu_2 I &=: U_2 R_2, \\ H_2 &:= R_2 U_2 + \mu_2 I. \end{aligned}$$

Then $H_j = R_j U_j + \mu_j I = U_j^H (H_{j-1} - \mu_j I) U_j + \mu_j I = U_j^H H_{j-1} U_j$ (with $H_0 := H$), hence $H_2 = (U_1 U_2)^H H (U_1 U_2)$.

But

$$\begin{aligned} (H-\mu_1 I)(H-\mu_2 I) &= U_1 R_1 (U_1 R_1 - (\mu_2 - \mu_1) I) \\ &= U_1 (H_1 - \mu_1 I) R_1 - U_1 R_1 (\mu_2 - \mu_1) I \\ &= U_1 (H_1 - \mu_2 I) R_1 = U_1 U_2 R_2 R_1. \end{aligned}$$

This implies that the needed matrix $Z := U_1 U_2$ is obtainable as the orthogonal matrix in the QR-factorization ZR for the **real** matrix

$$M := (H-\mu_1 I)(H-\mu_2 I) = H^2 - (\mu_1+\mu_2)H + \mu_1\mu_2 = H^2 - \operatorname{trace}(G)H + \det(G).$$

Here, we are using the fact that μ_1, μ_2 are the eigenvalues for that order-two lower right principal submatrix G of the real matrix H, hence $\mu_1 + \mu_2 = a + d = \operatorname{trace}(G)$ and $\mu_1\mu_2 = ad - bc = \det(G)$, and both these numbers are real. Of course, we make the assumption here that the QR-factorization for M is (essentially) unique and that is false. But we do know by (30)Theorem the following:

Observation. *If both $Z^T H Z$ and $Z_1^T H Z_1$ are Hessenberg and one is unreduced, with Z, Z_1 orthogonal and $Z(:,1) = Z_1(:,1)$, then $Z = Z_1 D$ for some diagonal matrix $D := \operatorname{diag}(1, \pm 1, \pm 1, \ldots)$, hence $Z^T H Z$ and $Z_1^T H Z_1$ are essentially the same.*

This means that, even when μ_1, μ_2 are complex, we can obtain H_2 entirely by real arithmetic and in $O(n^2)$ flops by the following steps:

Step 1 Compute the first column of $M = H^2 - \operatorname{trace}(G)H + \det(G)$.

Step 2 Determine the Householder reflection P_0 which carries Me_1 to αe_1.

Step 3 Compute, in order, the Householder reflections P_j so that $P_j(P_{j-1}\ldots P_0 H)$ is upper Hessenberg in its first j columns, $j = 1, \ldots, n-2$, while each such P_j leaves e_1 fixed.

Step 4 Finish up by forming $\bar{H} = H_2 = (Z^T H)Z$, with $Z := P_0 P_1 \ldots P_{n-2}$.

It follows that Z must be essentially $U_1 U_2$, as desired, provided the new H is still unreduced. But, if it isn't, so much the better since we now have an H (similar to the original matrix A) which breaks into smaller blocks.

It is important to note that $Me_1 = [\mathrm{x}; \mathrm{x}; \mathrm{x}; 0; 0; \ldots]$, hence the similarity transformation $H \mapsto P_0 H P_0$ only affects the first three rows and also the first three columns. This implies, by induction, that the subsequent transformations $X \mapsto P_j X P_j$ each only affect rows $j+1, j+2, j+3$ and also columns $j+1, j+2, j+3$, i.e, each such transformation takes $O(n)$ flops. The entire transformation $H \mapsto \bar{H} = H_2 = Z^T H Z$ therefore only takes $O(n^2)$ flops, as promised.

drift We have not discussed the possibility of **drift** since the book doesn't mention it. This possibility exists since the calculations do not explicitly refer back to the original A nor even to the first Hessenberg matrix constructed. The only way to combat this drift is to accumulate the various orthogonal matrices into one matrix Q, and to restart the entire process (including initial reduction to Hessenberg form) from the matrix $Q^H AQ$. The book does mention the importance of an initial **scaling** $A \mapsto D^{-1}AD$, with the diagonal matrix D so chosen that all rows and columns of $D^{-1}AD$ have approximatedly the same size.

eigenvector calculations use **inverse iteration**, i.e., the power method applied to the matrix $(A-\mu I)^{-1}$, with μ a very good approximation to a particular eigenvalue λ of A, hence $1/(\lambda-\mu)$ likely to be the dominant eigenvalue of $(A-\mu I)^{-1}$, consequently the iteration fast (i.e., in one or two steps) converging to an eigenvector of $inv(A-\mu)$ belonging to $1/(\lambda-\mu)$, i.e., an eigenvector of A belonging to λ.

This is strange at first sight since, after all, $A-\mu$ is (nearly) singular. Hence it should be difficult to compute $(A-\mu)^{-1}x$, and it is. But the point of the calculation is not so much to solve the equation

$$(A-\mu)? = x \tag{32}$$

as it is to produce an eigenvector of A belonging to that nearby eigenvalue λ, and that is exactly what the numerical solution of (32) does well, the more nearly singular, the better.

If this is done at the end of QR iteration, then we can expect the approximation μ to be within $\mathbf{u}\|A\|$ of an eigenvalue λ of A. Also, we have in hand, from the beginning step, the upper Hessenberg matrix $H := T_0 := U_0^T A U_0$ unitarily similar to A. Thus the solution of the linear system

$$(H-\mu)z_1 = z_0$$

can be carried out (e.g., by Gauss elimination with partial pivoting) in $O(n^2)$ flops. Let $H-\mu = U\Sigma V^T$ be an SVD. Then $\sigma_n \sim \mathbf{u}\|A\|$ is about as small as we can make it. Therefore, $\|(H-\mu)v_n\| = \sigma_n \sim \mathbf{u}\|H\|$, i.e., v_n is about as good an (approximate) eigenvector corresponding to $\lambda \sim \mu$ as we can hope to get. Further,

$$z_1 = \sum_j v_j (u_j^T z_0)/\sigma_j \sim v_n (u_n^T z_0)/\sigma_n,$$

if λ is a simple eigenvalue, hence $\sigma_{n-1} > \sigma_n$. We cannot hope to improve on this by subsequent iteration.

For a **multiple** eigenvalue λ, we cannot expect this process to work. This is due to the fact that it is now not possible to single out a particular eigenvector. (This a nice example of the paradox that 'the more there are, the harder they are to find'.) Orthogonal iteration was invented for this since, in such a case, there is, in the limit, an o.n. basis for the invariant subspace associated with the dominant eigenvalue.

More generally, one takes the (block) upper triangular Schur form $T = Q^T A Q$ for A obtained in the limit, makes certain that all repeated eigenvalues occur consecutively, and then uses further similarity transformations (of the kind already discussed) to transform T to block-diagonal form, with different diagonal blocks having disjoint spectra, and each diagonal block having just one eigenvalue, or else a complex conjugate pair of eigenvalues.

This requires the ability to prescribe the order in which A's eigenvalues appear along the diagonal. Since any permutation can be achieved as the product of transpositions, i.e., of interchanges of neighboring terms, it is sufficient to know how to reorder two neighboring diagonal entries in an upper triangular matrix using similarity. For this, it is sufficient to know how to transform the matrix $G = \begin{bmatrix} a & \mathrm{x} \\ 0 & d \end{bmatrix}$ with $a \neq b$ by similarity to the matrix $\bar{G} = \begin{bmatrix} d & \mathrm{x} \\ 0 & a \end{bmatrix}$. This means that we are looking for a basis V for which the matrix representation $\bar{G} = V^{-1} G V$ for G has the first column $[d;0]$, i.e., for which v_1 is an eigenvector of G belonging to the eigenvalue d. If we also insist that V be orthogonal, this determines V up to sign. In particular, then necessarily $\bar{G}(2,2) = a$ since $\bar{G}$ is upper triangular, hence $\bar{G}(2,2) \in \lambda(\bar{G}) = \lambda(G) = \{a, d\}$ and d already appears as $\bar{G}(1,1)$.

To get an eigenvector v of G belonging to d, use inverse iteration, i.e., solve

$$\begin{bmatrix} a-d & \mathrm{x} \\ 0 & 0 \end{bmatrix} v = 0$$

say, to get $(a-d)v(1) + \mathrm{x}v(2) = 0$, e.g., $v = [\mathrm{x}; d-a]$. Thus the Givens rotation Q which carries e_1 to a multiple of v, i.e., for which $Q^T v$ is a multiple of e_1, would give $Q^T G Q = \bar{G}$.

Once the Schur form T for A has been reordered to have all repeated eigenvalues appear consecutively (there has to be version that will bring 2-by-2 diagonal blocks with the same complex conjugate pair next to each other), one would partition T accordingly (see above), and then make use of the following material from mar.3 recalled here through the miracle of modern information processing:

Assume that T is block upper triangular,

$$T = \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix} \tag{33}$$

with T_{11}, T_{22} square (but not necessarily of the same order). We look for a similarity transformation $Y^{-1}TY$ which leaves the two diagonal blocks as well as $0 = T_{21}$ unchanged, but turns that block T_{12} into zero. In the general case, this forces Y to be of the block form

$$Y = \begin{bmatrix} I & Z \\ 0 & I \end{bmatrix},$$

and so gives

$$Y^{-1}TY = \begin{bmatrix} I & -Z \\ 0 & I \end{bmatrix} \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix} \begin{bmatrix} I & Z \\ 0 & I \end{bmatrix} = \begin{bmatrix} T_{11} & T_{12} + T_{11}Z - ZT_{22} \\ 0 & T_{22} \end{bmatrix}.$$

Hence this similarity is successful provided we can choose Z so that

$$\varphi(Z) := T_{11}Z - ZT_{22} = -T_{12}.$$

Since the map φ so defined is linear and $D(\varphi) = T(\varphi)$, we will succeed, in general, iff φ is invertible, i.e., iff φ is 1-1 , i.e., iff $\lambda(A) \cap \lambda(B) = \emptyset$, as we proved earlier.

Implementation requires the solution of the **Sylvester equation**

$$FZ - ZG = C \tag{34}$$

for Z where, in our case, both $F := T_{11}$ and $G := T_{22}$ are block upper triangular with diagonal blocks of order 1 or 2. The book calls such matrices **quasi-triangular**. This quasi-triangularity makes for an easy inductive solution (the Bartels-Stewart algorithm): The kth column of the equation (34) reads

$$Fz_k - \sum_{i \le k+1} z_i G(i,k) = c_k.$$

If $G(k+1,k) = 0$, then, in fact,

$$Fz_k - z_k G(k,k) = c_k + \sum_{i \le k} z_i G(i,k)$$

which can be solved quickly (since F is quasi-triangular) for z_k if we already know z_i for $i < k$.

If $G(k+1,k) \neq 0$, then we combine this with the next equation and get (since then, by quasi-triangularity, $G(k+2,k+1) = 0$) the system

$$\begin{bmatrix} F - G(k,k) & -G(k+1,k) \\ -G(k,k+1) & F - G(k+1,k+1) \end{bmatrix} \begin{bmatrix} z_k \\ z_{k+1} \end{bmatrix} = \begin{bmatrix} c_k \\ c_{k+1} \end{bmatrix} + \sum_{i<k} \begin{bmatrix} z_i G(i,k) \\ z_i G(i,k+1) \end{bmatrix}.$$

By interlacing the equations from the first block with those of the second, a banded system is obtained which can be solved in $O(n^2)$ flops.

Since this process breaks down when F and G share eigenvalues, it is not surprising that it experiences numerical difficulties in case $\lambda(F)$ and $\lambda(G)$ are 'close'. Explicitly, one defines the **separation** between (the spectra of) two square matrices B and C, not necessarily of the same order, as the largest lower bound for the linear map $\varphi : X \mapsto BX - XC$ with respect to the Frobenius norm, i.e.,

$$\operatorname{sep}(B,C) := \min_X \|BX - XC\|_F / \|X\|_F.$$

This number quantifies just how invertible φ is. Thus $\operatorname{sep}(F,G)$ measures how well we can solve the Sylvester equation (34) in the presence of noise.

The book considers only **real symmetric** matrices. But the theory for their eigenstructure is identical with that of **hermitian** matrices, hence this is what will be discussed. Throughout, let A be a hermitian matrix, i.e., $A^H = A$.

By Schur, $A = U^H T U$ for some unitary U and some upper triangular T, for any matrix A. But, A being hermitian implies that T is hermitian, which implies that T is diagonal with real diagonal entries. Conclusion: A is unitarily similar to a real diagonal matrix. In particular, if A is real, then U can be chosen orthogonal.

For various reasons of convenience, we order the real eigenvalues of A from left to right (as we write) rather than from right to left (as the book does). We write:

$$\lambda_1(A) \leq \lambda_2(A) \leq \ldots \leq \lambda_n(A)$$

and denote the corresponding sequence of o.n. eigenvectors by $[u_1, \ldots, u_n]$.

The **perturbation theory** for the spectrum of a *hermitian* matrix A is remarkably simple and elegant. It centers on the **Rayleigh quotient**

$$R_A(x) := x^H A x / x^H x,$$

whose maximum/minimum equals the largest/smallest eigenvalue of A, as follows easily from the observation that

$$R_A(x) = R_A(Uy) = \frac{y^H U^H A U y}{y^H U^H U y} = \frac{\sum_j |y(j)|^2 \lambda_j}{\sum_j |y(j)|^2} \ \in \ [\min_j \lambda_j, \max_j \lambda_j] = [\lambda_1(A), \lambda_n(A)].$$

This is the **Rayleigh principle**. It is a very special case of the following **MMM Theorem** (or, Principle, if you go in for such things), which is the convenient amalgam of the 'N.N. maximin' theorem with the 'Courant-Fischer minimax' theorem:

MaxiMiniMax Theorem. *For a hermitian matrix A with eigenvalues $\lambda_1(A) \leq \lambda_2(A) \leq \ldots \leq \lambda_n(A)$ (and corresp. o.n. eigenvectors $u_1, \ldots, u_n$), and any j,*

$$\max_{\dim G < j} \min_{x \perp G} R_A(x) \ = \ \lambda_j(A) \ = \ \min_{j \leq \dim H} \max_{x \in H} R_A(x),$$

with G and H arbitrary linear subspaces.

Proof: If $\dim G < j \leq \dim H$, then one can find $y \in (G \perp \cap H) \backslash 0$, therefore

$$\min_{x \perp G} R_A(x) \ \leq R_A(y) \leq \ \max_{x \in H} R_A(x).$$

Hence,

$$\max_{\dim G < j} \min_{x \perp G} R_A(x) \ \leq \ \min_{j \leq \dim H} \max_{x \in H} R_A(x).$$

On the other hand, for $G = \mathrm{ran}[u_1, \ldots, u_{j-1}]$ and $H = \mathrm{ran}[u_1, \ldots, u_j]$,

$$\min_{x \perp G} R_A(x) \ = \ \lambda_j(A) \ = \ \max_{x \in H} R_A(x).$$

□

The MMM theorem has various useful (and immediate) corollaries.

Interlacing Theorem. *If the matrix B is obtained from the hermitian matrix A by crossing out the last row and column (i.e., $B = A(1:n-1, 1:n-1)$), then*

$$\lambda_j(A) \le \lambda_j(B) \le \lambda_{j+1}(A), \quad j < n.$$

Proof: With $J : \mathbb{F}^{n-1} \to \mathbb{F}^n : x \mapsto [x; 0]$, we have $R_B(x) = R_A(Jx)$ and $\operatorname{ran} J = e_n \perp$. Hence

$$\begin{aligned}
\lambda_j(B) &= \max_{\dim G < j} \min_{x \perp G} R_A(Jx) \\
&= \max_{\dim G < j} \min_{y \perp JG + \operatorname{span}(e_n)} R_A(y) \\
&\le \max_{\dim G \le j+1} \min_{y \perp G} R_A(y) = \lambda_{j+1}(A).
\end{aligned}$$

Also,

$$\begin{aligned}
\lambda_j(B) &= \min_{j \le \dim H} \max_{x \in H} R_A(Jx) \\
&= \min_{j \le \dim H} \max_{y \in JH} R_A(y) \\
&\ge \min_{j \le \dim H} \max_{y \in H} R_A(y) = \lambda_j(A).
\end{aligned}$$

□

A different, simpler, application of the MMM theorem is based on the following observation: If

$$f(t) \le g(t) \quad \forall t,$$

then this inequality persists if we take on both sides the maximum or minimum over the same set T, i.e., then

$$\max_{t \in T} f(t) \le \max_{t \in T} g(t), \qquad \min_{t \in T} f(t) \le \min_{t \in T} g(t).$$

It even persists if we further take the miminum or maximum over the same family $\mathbf{T}$ of subsets T, e.g., then also

$$\max_{T \in \mathbf{T}} \min_{t \in T} f(t) \le \max_{T \in \mathbf{T}} \min_{t \in T} g(t).$$

Consequently,

Corollary. *If $R_A(x) \le R_B(x) + c$ for some constant c and all x, then*

$$\lambda_j(A) \le \lambda_j(B) + c, \quad \forall j.$$

This gives

Weyl's inequalities. *If $A = B + C$, with A, B, C hermitian, then*

$$\lambda_j(B) + \lambda_1(C) \le \lambda_j(A) \le \lambda_j(B) + \lambda_n(C), \quad \forall j.$$

Proof: Since $\lambda_1(C) \le R_C(x) \le \lambda_n(C)$ (by Rayleigh's principle), while $R_B(x) + R_C(x) = R_A(x)$, the preceding corollary provides the proof. □

These inequalities are particularly useful when C is 'small' in some sense. E.g., the book considers (in Cor.8.1-5) the special case $C = \tau cc^H$ with c of unit length. Then C is a rank-one matrix, with 0 as $n-1$-fold eigenvalue, while also $Cc = \tau c$, hence also τ an eigenvalue. Consequently, $\lambda_1(C) = \min\{0, \tau\}$ and $\lambda_n(C) = \max\{0, \tau\}$, and Weyl's inequalities do the rest.

The book also brings (without proof) the

Wielandt-Hoffman Theorem. *If A and E are both hermitian, then*

$$\sum_j \big(\lambda_j(A+E) - \lambda_j(A)\big)^2 \leq \sum_j \lambda_j(E)^2.$$

But I find the following max-norm version more directly useful.

max-norm Wielandt-Hoffman. *If A and E are both hermitian, then*

$$\max_j |\lambda_j(A+E) - \lambda_j(A)| \leq \max_j |\lambda_j(E)|.$$

Proof: By Weyl's inequalities,

$$|\lambda_j(A) - \lambda_j(B)| \leq \max\{|\lambda_1(A-B)|, |\lambda_n(A-B)|\} = \rho(A-B).$$

□

Finally, a totally different application of the MMM Theorem is

Sylvester's Law of Inertia. *Any two congruent Hermitian matrices have the same number of positive, zero, and negative eigenvalues.*

Proof: It is sufficient to prove that if $B = X^H AX$ for some hermitian A and some invertible X, then $\lambda_j(A) > 0$ implies $\lambda_j(B) > 0$. For this, we observe that, by the MMM Theorem, $\lambda_j(A) > 0$ implies that R_A is positive somewhere on any j-dimensional subspace, while (also by the MMM Theorem), for some j-dimensional subspace H,

$$\lambda_j(B) \;=\; \max_{x \in H} R_B(x) = \max_{x \in H} R_A(Xx) R_{X^H X}(x),$$

and this is necessarily positive, since $\dim XH = j$ and $R_{X^H X}(x) = (\|Xx\|_2/\|x\|_2)^2$ is positive for any $x \neq 0$. □

Sturm sequences provide a very elegant way to separate the eigenvalues of a **Jacobi matrix**, i.e., a real symmetric tridiagonal matrix

$$T_n := \begin{bmatrix} a_1 & b_2 & & & & \\ b_2 & a_2 & b_3 & & & \\ & \cdot & \cdot & \cdot & & \\ & & \cdot & \cdot & \cdot & \\ & & & \cdot & \cdot & b_n \\ & & & & b_n & a_n \end{bmatrix}.$$

The key for this is the **three-term recurrence relation**

$$p_r(x) = (a_r - x)p_{r-1}(x) - b_r^2 p_{r-2}(x), \quad r = 1, 2, \ldots,$$

satisfied by the characteristic polynomials

$$p_r(x) := \det(T_r - x), \quad r = 1, 2, \ldots,$$

a definition which, in light of the recurrence relation is suitably augmented by the agreements

$$p_{-1} := 0, \quad p_0 := 1.$$

(Remark. Assuming that T_n is unreduced, the sequence (p_r) is orthogonal with respect to some discrete measure $\sum\{w(\xi)[\xi] : p_n(\xi) = 0\}$, a fact that summarizes the various statements about the zeros of the p_r about to be detailed. More on this later.)

We now need a way to count the sign changes in a sequence or vector:

$$S(a) := \#\{j : \operatorname{sign}(a(j)) \neq \operatorname{sign}(a(j+1))\},$$

where $\operatorname{ran}\operatorname{sign} = \{-1, 1\}$. This definition is a bit hazy on just what sign(0) might be and, offhand, that's really up to you. To make this precise, one defines

$$S^+(a) := \max S(a), \quad S^-(a) := \min S(a),$$

with the extremum taken over all possible assignments of 1 or -1 to sign$(a(j))$ in case $a(j) = 0$.

We are applying this notion to the particular sequence

$$P(x) := [p_0(x), \ldots, p_n(x)],$$

i.e., we define

$$s(x) := S(P(x)).$$

This function is piecewise constant, with jumps only at points x at which $S^+(P(x)) \neq S^-(P(x))$. This can happen only at the finitely many points x at which $p_r(x) = 0$ for some r. If $r < n$, then, from the recurrence,

$$p_{r+1}(x) = -b_{r+1}^2 p_{r-1}(x)$$

at such a point. If now also $p_{r-1}(x) = 0$, then (assuming that T_n is unreduced) it would follow that also $0 = p_{r-2}(x) = p_{r-3}(x) = \ldots p_0(x) = 1$, an impossibility. Hence, assuming T_n to be unreduced, it follows that $s(x-) = s(x+)$ at such a point since, necessarily,

$$p_{r+1}(x)p_{r-1}(x) < 0 \quad \text{whenever} \quad p_r(x) = 0. \tag{35}$$

Thus, (since p_0 never vanishes,) s can have a jump at x only when $p_n(x) = 0$, in which case it is a jump of ± 1 only, since then necessarily $p_{n-1}(x) \neq 0$. On the other hand, since $p_r(x) = (-x)^r + lot$, it follows that

$$s(x) = \begin{cases} 0, & x \to -\infty; \\ n, & x \to \infty. \end{cases}$$

This implies that s must jump at least n times, and, since it can only jump at the zeros of p_n and p_n can have at most n zeros, it follows that p_n has exactly n real zeros, hence these must all be simple. in other words,

$$s(x+) - s(x-) =: \operatorname{jump}_x s = 1 \quad \text{whenever} \quad p_n(x) = 0,$$

and

$$s(x-) = \#\{\xi < x : p_n(\xi) = 0\}.$$

Therefore also $p_n'(\xi)p_{n-1}(\xi) < 0$ whenever $p_n(\xi) = 0$.

This implies the same for the zeros of p_{n-1}. Also, since the earlier perturbation analysis gave us that the eigenvalues of T_{n-1} separate those of T_n, we now conclude from (35) that they *strictly* separate, i.e., that

$$\lambda_1(T_n) < \lambda_1(T_{n-1}) < \lambda_2(T_n) < \lambda_2(T_{n-1}) < \cdots < \lambda_{n-1}(T_{n-1}) < \lambda_n(T_n).$$

This proves:

Sturm's Theorem. *If T_n is unreduced, then the zeros of $p_n := \det(T_n - \cdot)$ are all real and simple and are strictly separated by the zeros of p_{n-1}. Moreover*

$$S([p_0(x-), \ldots, p_n(x-)]) = \#\{\xi < x : p_n(\xi) = 0\}.$$

Since the recurrence provides the entire vector $P(x) = [p_0(x), \ldots, p_n(x)]$, it is easy to compute $s(x) = S(P(x))$ as one evaluates $p_n(x)$, hence easy to localize the zeros of p_n and obtain them quickly by bisection followed by modified regula falsi and the like.

The corresponding eigenvectors of T_n are readily found by inverse iteration, particularly fast because T_n is tridiagonal. Note that, in such inverse iteration, we would be computing, in effect, the L-D-L^T factorization $LDL^T := T_n - \mu$ (with L unit lower bidiagonal and D diagonal, and μ our approximation to a particular eigenvalue of T_n). Since the resulting $D =: \operatorname{diag}(d_1, \ldots, d_n)$ is *congruent* to $T_n - \mu$, it follows from Sylvester's Inertia Law that

$$\#\{j : d_j < 0\} = \#\{\xi < \mu : p_n(\xi) = 0\},$$

suggesting another method for localizing and approximately determining the zeros of p_n.

Rayleigh quotient iteration is based on the following facts: With $\mu := R_A(x) = x^T Ax/x^T x$ the Rayleigh quotient for A at x, and z computed a la inverse iteration as the solution of the linear system $(A - \mu)? = x$, we compute

$$(A - xz^T/z^T z)z = Az - x = \mu z,$$

showing that μ is an eigenvalue of the perturbed matrix $A - xz^T/z^T z$. Hence, by Bauer-Fike, $\operatorname{dist}(\mu, \lambda(A)) \leq \|xz^T/z^T z\|_2 = 1/\|z\|_2$ if, as we may, we take x to be normalized.

For the resulting iteration $x_{k+1} := z_{k+1}/\|z_{k+1}\|_2$ with $(A - \mu_k)z_{k+1} = x_k$ and $\mu_k := R_A(x_k)$, the pair (x_k, μ_k) converges cubically to an eigenpair, as the book's 2-by-2 example demonstrates. In order to shoot for a particular eigenvalue of A, one would start the iteration with a μ close to that eigenvalue, and use inverse iteration to obtain from it a first x.

Lanczos Method can be viewed as a refinement of the power method: If (as we assume) A is real symmetric, with $R_A(x) := x^TAx/x^Tx$ its Rayleigh quotient, then, for almost any y, we would expect $R_A(A^ky)$ to converge to its absolutely largest eigenvalue, as $k \to \infty$. On the other hand, we know that this absolutely largest eigenvalue is either $\min_x R_A(x)$ or $\max_x R_A(x)$, hence would expect the appropriate extremum taken over $x \in \text{span}\{y, Ay, A^2y, \ldots, A^ky\}$ to be a much better approximation to that absolutely largest eigenvalue than that simple ratio $R_A(A^ky)$.

Here is a more explicit analysis. The typical element of the **Krylov space**

$$\mathbb{K}(A, y, j) := \text{span}(A^k y)_{k<j}$$

is of the form $p(A)y$ for some $p \in \pi_{<j} :=$ polynomials of degree $< j$. Thus, with $y =: \sum_k z_k c(k)$ the expansion of y wrto an orthonormal basis consisting of eigenvectors of A, with corresponding eigenvalues

$$\lambda_1 \le \ldots \le \lambda_n,$$

we compute $(p(A)y)^T A\, p(A)y = \sum_k c(k)^2 p(\lambda_k)^2 \lambda_k$, therefore, with $\|y\|_2^2 = \sum_k c(k)^2 = 1$ wlog,

$$\begin{aligned}
\lambda_1 \le R_A(p(A)y) &= \lambda_1 + \frac{\sum_k c(k)^2 p(\lambda_k)^2 (\lambda_k - \lambda_1)}{\sum_k c(k)^2 p(\lambda_k)^2} \\
&\le \lambda_1 + (\lambda_n - \lambda_1) \frac{\sum_{k>1}}{c(1)^2 p(\lambda_1)^2 + \sum_{k>1}} \\
&\le \lambda_1 + (\lambda_n - \lambda_1) \frac{1 - c(1)^2}{c(1)^2} \; \frac{\max_{k>1} p(\lambda_k)^2}{p(\lambda_1)^2}.
\end{aligned}$$

Assume without loss that λ_1 is, indeed, the absolutely largest eigenvalue of A. Our use of the Rayleigh quotient $R_A(A^{j-1}y)$ to estimate λ_1 corresponds to the choice $p := ()^{j-1}$, hence provides the estimate

$$\lambda_1 \le R_A(A^{j-1}y) \le \lambda_1 + (\lambda_n - \lambda_1) t^{-2(j-1)}, \quad t := |\lambda_1/\mu|,$$

with μ the absolutely second largest eigenvalue of A.

By choosing different $p \in \pi_{<j}$, we may hope to get closer to λ_1. Specifically, we may seek to minimize

$$b(p) := \frac{\max_{k>1} |p(\lambda_k)|}{|p(\lambda_1)|}$$

over all $p \in \pi_{<j}$. It is well-known that the minimum of the related function

$$B(p) := \frac{\max_{\lambda \in [\lambda_2, \lambda_n]} |p(\lambda)|}{|p(\lambda_1)|} \;\ge\; b(p)$$

(which dominates b) is taken on by the **Chebyshev polynomial** of order j for the interval $[\lambda_2, \lambda_n]$, i.e., the polynomial $C_{j-1}(2\frac{\cdot - \lambda_2}{\lambda_n - \lambda_2} - 1)$, with C_k the Chebyshev polynomial of degree k. Recall that $|C_k|$ is bounded by 1 on $[-1, 1]$ but grows the fastest outside that interval when compared with any other polynomial of degree $\le k$ which is absolutely bounded by 1 on $[-1, 1]$. Also, $C_k(t) \sim (t + \sqrt{t^2 - 1})^k/2$ for t outside $[-1, 1]$. Thus, replacing the interval $[\lambda_2, \lambda_n]$ by the larger (hence less advantageous) interval $[-|\mu|, |\mu|]$, we get here

$$\lambda_1 \le R_A(p(A)y) \le \lambda_1 + O\big((t + \sqrt{t^2 - 1})^{-2j}\big)$$

instead of

$$\lambda_1 \le R_A(A^{j-1}y) \le \lambda_1 + O(t^{-2j})$$

for the power method with Rayleigh quotient, or

$$\lambda_1 = (A^j y)(k)/(A^{j-1}y)(k) + O(t^{-j})$$

for the straight power method. The book provides more explicit numerical comparisons.

Note also that the power method only goes after the absolutely largest eigenvalue, while the above analysis could equally well have been applied to the largest or smallest eigenvalue. In other words, having taken the trouble to compute the Krylov sequence $y, Ay, A^2y, \ldots, A^{j-1}y$, we could have maximimized as well as minimized the Rayleigh quotient over that subspace. This leads to Lanczos' Method.

There are various ways of viewing that method. Most directly, Lanczos method consists in deriving from $y, Ay, A^2y, \ldots, A^{j-1}y$ an equivalent o.n. sequence $q_1, q_2, \ldots, q_j$ (AHA! that's the reason the book keeps talking about $A^{j-1}y$ all the time) as one would with Gram-Schmid, i.e., so that $\text{ran}[q_1, \ldots, q_k] = \text{ran}[y, Ay, \ldots, A^{k-1}y] = \mathbb{K}(A, y, k)$. If Q is an extension of the o.n. sequence to an o.n. basis for $\mathbb{R}^n$, then the matrix representation $T := Q^T AQ$ for A wrto this basis has a tridiagonal upper left principal submatrix of order j. For, by construction, $Aq_k \in A(\text{ran}[q_1, \ldots, q_k]) = A\mathbb{K}(A, y, k) \subset \mathbb{K}(A, y, k+1) = \text{ran}[q_1, \ldots, q_{k+1}]$, hence $T(:, k) = Q^T AQe_k = Q^T Aq_k$ has nonzero entries only in positions $1, \ldots, k+1$. On the other hand, T is real symmetric since A is, hence also $T(k, :)$ has nonzero entries only in its first $k+1$ positions. This implies that

$$AQ_j = Q_j T_j + r_j e_j^T, \tag{36}$$

with

$$T_j := \begin{bmatrix} a_1 & b_1 & & & & \\ b_1 & a_2 & b_2 & & & \\ & \cdot & \cdot & \cdot & & \\ & & \cdot & \cdot & \cdot & \\ & & & \cdot & \cdot & b_{j-1} \\ & & & & b_{j-1} & a_j \end{bmatrix}$$

and

$$r_j := (A - a_j)q_j - b_{j-1}q_{j-1}.$$

(Note the different way of subscripting the off-diagonal entries here.) In particular, since T_j is a principal submatrix of a matrix representation for A with respect to an on. basis,

$$\lambda_1(A) \le \lambda_1(T_j) = \min R_A(\mathbb{K}(A, y, j)) \le \max R_A(\mathbb{K}(A, y, j)) = \lambda_j(T_j) \le \lambda_n(A).$$

This shows that we can obtain best possible estimates from $\mathbb{K}(A, y, j)$ for the extreme eigenvalues of A by computing the extreme eigenvalues of T_j.

It was Lanczos' genius to turn this construction around: We are usually not really interested in the **Lanczos vectors** q_k per se. Rather, we are only interested in the tridiagonal matrix T_j since, through the symmetric QR (or other) methods, we can efficiently extract the Rayleigh quotient extrema sought as the extremal eigenvalues of T_j. But the entries of T_j we can hope to compute directly from the requirement that equation (36) should hold, as follows. Assume that we already know a_i, b_i for $i < k$ and also know q_i for $i \le k$. We read off that

$$Aq_k = b_{k-1}q_{k-1} + a_k q_k + b_k q_{k+1},$$

hence we can compute a_k as

$$q_k^T Aq_k / q_k^T q_k = a_k,$$

and therefore know the vector

$$r_k := (A - a_k)q_k - b_{k-1}q_{k-1} = b_k q_{k+1}.$$

We can therefore compute

$$b_k := \|r_k\|_2,$$

and, assuming that $r_k \neq 0$, also get

$$q_{k+1} := r_k / b_k.$$

Note that we only needed q_{k-1} and q_k here, hence need to retain now only q_k and q_{k+1} for the next step.

This very convenience is also the down-fall of the method. Since q_{k+1} is derived from q_k and q_{k-1}, drift sets in eventually, i.e., with increasing k, the computed q_k increasingly fail to be orthogonal to earlier q_i. This not only makes the computed eigenvalues of the computed T_j drift away from those of A, but also introduces *ghost* eigenvalues, i.e., repetitions of eigenvalues already approximated well earlier on. Reorthogonalization seems to be the only remedy. This is unfortunate since Lanczos method is, offhand, ideal for computing the extremal eigenvalues of a large sparse A and the sparsity advantage is lost if reorthogonalization is needed.

Error analysis for Lanczos method Paige shows that the equation

$$AQ_j = Q_j T_j + r_j e_j^T$$

is satisfied by the computed Q_j, T_j and r_j to within $O(\mathbf{u}A)$, i.e., as well as expected. The trouble with Lanczos method has its source in 'small' r_j which leads to amplification of noise when $q_{j+1} = r_j/\|r_j\|_2$ is computed, thus spoiling the expected orthogonality (though not the normality) of the computed Q_j. This may have serious consequences for the quality of the eigenvalues of the computed T_j as approximations to the eigenvalues of A, as the following analysis shows.

From p.272 of the book, we recall (without sqrt) the

Proposition (Theorem 8.1-8). *If A, B are symmetric and $E := AQ_1 - Q_1 B$ with $Q_1 \in \mathbb{R}^{n\times k}$ orthogonal, then*

$$\max_{\lambda\in\lambda(B)} \min_{\mu\in\lambda(A)} |\lambda - \mu| =: \operatorname{dist}(\lambda(B), \lambda(A)) \le \|AQ_1 - Q_1 B\|_2.$$

Proof: Extend Q_1 in any way to an o.n. basis $Q := [Q_1, Q_2]$. Then

$$Q^T A Q = \begin{bmatrix} B & 0 \\ 0 & Q_2^T A Q_2 \end{bmatrix} + \begin{bmatrix} Q_1^T E & E^T Q_2 \\ Q_2^T E & 0 \end{bmatrix} =: C + F.$$

Hence $|\lambda_j(C) - \lambda_j(A)| \le \|F\|_2$. But $\lambda(B) \subset \lambda(C)$, while $F[x;y] = [Q_1^T E x + E^T Q_2 y; Q_2^T E x]$, hence $\|F[x;y]\|_2^2 = \|Q_1^T E x + E^T Q_2 y\|_2^2 + \|Q_2^T E x\|_2^2 \le \|Q_1^T E x\|_2^2 + \|Q_2^T E x\|_2^2 + \|E^T Q_2 y\|_2^2 = \|Ex\|_2^2 + \|E^T Q_2 y\|_2^2 \le \|E\|_2^2 \, \|[x;y]\|_2^2$, and so $\|F\|_2 \le \|E\|_2$. □

We conclude that, actually, 'small' r_j is good **if** Q_j is o.n. If it isn't, then the best we can say is the following.

Corollary. *If A, B are symmetric and $F := AX - XB$ with $X \in \mathbb{R}^{n\times k}$ satisfying $\sigma_k(X) > 0$, then*

$$\operatorname{dist}(\lambda(B), \lambda(A)) \le \|AX - XB\|_2/\sigma_k(X).$$

Proof: Let $X =: QR$ with Q o.n. and R square upper triangular, therefore $\|R^{-1}\|_2 = 1/\sigma_k(X)$. Then $AQ - Q(RBR^{-1}) = E := (AX - XB)R^{-1}$, hence the proposition does it. □

We conclude from this that

$$\operatorname{dist}(\lambda(T_j), \lambda(A)) \le (\|r_j\|_2 + \|E_j\|_2)/\sigma_j(Q_j),$$

with

$$E_j := AQ_j \; - \; (Q_j T_j + r_j e_j^T).$$

Again, we know that $\|E_j\|_2 = O(\mathbf{u}\|A\|_2)$. What is troubling is the failure of the computed Q_j to stay orthogonal, as this leads to 'small' $\sigma_j(Q_j)$.

The straightforward remedy is reorthogonalization, in which each computed q_j is further modified by orthogonalizing it wrto all the earlier q_r, a very expensive procedure. An interesting modification merely orthogonalizes q_j wrto a set of 'converged' Ritz vectors, thus making certain that the corresponding extremal eigenvalues are not introduced anew (as 'ghost' eigenvalues). In this view, the extremal eigenvalues of T_j are good approximations to corresponding eigenvalues of A to the extent that the corresponding Ritz vectors have converged. Here are the details (from Theorem 8.1-11).

If A is hermitian and $Q \in \mathbb{F}^{n\times j}$ is o.n., and

$$D := \operatorname{diag}(\theta_1, \ldots, \theta_j) := Z^H (Q^H A Q) Z$$

is a Schur decomposition for $Q^H A Q$, then

$$AQZ - QZD = (I - QQ^H) AQZ.$$

The diagonal entries of D are called **Ritz values** and the columns of

$$[y_1, \ldots, y_j] := QZ$$

are the corresponding **Ritz vectors**. It follows that

$$Ay_k - \theta_k y_k = (AQZ - QZD)e_k,$$

hence

$$\|Ay_k - \theta_k y_k\|_2 \leq \|(I - QQ^H)AQ\|_2.$$

This is relevant here since $A = Q_j T_j Q_j^H + r_j e_j^H Q_j^H$, while $Z_j^H T_j Z_j =: \mathrm{diag}(\theta_1, \ldots, \theta_j)$ for some suitable o.n. matrix Z_j. In particular, $y_k := Q_j Z_j e_k$ is the approximate eigenvector or *Ritz* vector corresponding to the approximate eigen- (or *Ritz*) value θ_k. If $\|Ay_k - \theta_k y_k\|_2 = O(\mathbf{u}\|A\|_2)$, then the pair (y_k, θ_k) is as close to an eigenpair for A as we are entitled to get. We want subsequent q_i to stay away from this direction, and, in exact arithmetic, they would, since they would be orthogonal to $\mathrm{ran}\, Q_j$ which contains these y_k. In finite precision arithmetic, this orthogonality fails and this leads to a reintroduction of these eigendirections. in **selective reorthogonalization**, one keeps a record of such converged Ritz pairs and only insists that subsequent q_i be orthogonal to them. There are tricks concerning just how one determines which Ritz pairs have converged without keeping track of all the q_i ...

Jacobi matrices, three-term recurrence relations, and orthogonal polynomials Throughout, T_n is an unreduced **Jacobi matrix**, i.e., a real symmetric tridiagonal matrix

$$T_n := \begin{bmatrix} a_1 & b_1 & & & & \\ b_1 & a_2 & b_2 & & & \\ & \cdot & \cdot & \cdot & & \\ & & \cdot & \cdot & \cdot & \\ & & & \cdot & \cdot & b_{n-1} \\ & & & & b_{n-1} & a_n \end{bmatrix}$$

with $b_i > 0$ for all j. Associated with T_n are the monic polynomials

$$p_r := \begin{cases} 0, & r < 0; \\ 1, & r = 0; \\ \det(\cdot - T_r), & r > 0. \end{cases}$$

These polynomials satisfy the **three-term recurrence relation**

$$p_r(x) = (x - a_r)p_{r-1}(x) - b_{r-1}^2 p_{r-2}(x), \quad r = 1, 2, \ldots.$$

Conversely, if we know the sequence $(p_r)_{r \le n}$, then we can construct the corresponding Jacobi matrix by running the recurrence backwards. In fact, we can do this if we know just p_n and p_{n-1}. For, if we already know p_r and p_{r-1}, then, by the recurrrence, necessarily $p_r = ()^r - a_r()^{r-1} + lot$, which provides us with a_r. After that, we know $p_r - (\cdot - a_r)p_{r-1} = -b_{r-1}^2 p_{r-2}$, hence can determine b_{r-1} and p_{r-2} from the requirement that p_{r-2} be monic.

If we try this for an arbitrary pair p_n, p_{n-1} of monic polynomials, we should not succeed since we proved earlier results about interlacing of zeros of the polynomials associated with a Jacobi matrix. We will run into trouble when we have to take squareroots of negative numbers. What is the most general sequence of polynomials obtainable this way? From the Sturm sequence results, we know that, at a minimum,

$$p_n = \prod_{i=1}^{n} (\cdot - \xi_i), \quad p_{n-1} = \prod_{j=1}^{n-1} (\cdot - \mu_j), \quad \text{with } \xi_i < \mu_i < \xi_{i+1}, \text{all } i. \tag{37}$$

As we will see, these necessary conditions are also sufficient.

Associated with any inner product $\langle , \rangle$, there is its sequence of orthogonal polynomials, i.e., the unique sequence (p_r) of monic polynomials, $p_r = ()^r + lot$, all r, for which $\langle p_r, p_s \rangle = 0$ whenever $r \neq s$. There is, in fact, at most one such sequence, since, by its definition, each p_r is in $()^r + \pi_{<r}$, yet perpendicular to $\pi_{<r} = \text{span}(p_s)_{s<r}$, hence must be the error in the b.a. to $()^r$ from $\pi_{<r}$ wrto the norm $\| \cdot \| := \langle \cdot, \cdot \rangle^{1/2}$. This argument also proves existence of the sequence.

The actual (numerical) construction is best done by making use of the fact that such a sequence of orthogonal polynomials must satisfy a three-term recurrence relation (AHA!), provided multiplication by $()^1$ is symmetric with respect to the inner product, i.e., provided

$$\langle ()^1 f, g \rangle = \langle f, ()^1 g \rangle.$$

For, in that case,

$$\langle ()^1 p_{r-1}, p_s \rangle = \langle p_{r-1}, ()^1 p_s \rangle = 0 \quad \forall s < r - 2$$

since $()^1 p_s \in \pi_{s+1}$. This means that p_r is obtainable from $()^1 p_{r-1}$ by merely subtracting the component of p_{r-1} and of p_{r-2} to get something perpendicular to p_{r-1} and p_{r-2}. Such a modification does not destroy the orthogonality to π_{r-3} since p_{r-1} and p_{r-2} are both orthogonal to it. This gives

$$p_r = ()^1 p_{r-1} - p_{r-1} \langle ()^1 p_{r-1}, p_{r-1} \rangle / \|p_{r-1}\|^2 - p_{r-2} \langle ()^1 p_{r-1}, p_{r-2} \rangle / \|p_{r-2}\|^2,$$

or

$$p_r = (\cdot - a_r)p_{r-1} - b_{r-1}^2 p_{r-2},$$

with

$$\begin{aligned} a_r &:= \langle ()^1 p_{r-1}, p_{r-1}\rangle / c_{r-1} \\ b_{r-1}^2 &:= \langle ()^1 p_{r-1}, p_{r-2}\rangle / c_{r-2} = \langle p_{r-1}, ()^1 p_{r-2}\rangle / c_{r-2} = c_{r-1}/c_{r-2} \\ c_r &:= \|p_r\|^2 \end{aligned}$$

Note that the p_r so computed does have the correct normalization since its leading coefficient is the same as that of p_{r-1}, i.e., 1 by induction. Also note that the recurrence can start with $r = 1$, taking $b_0 := 0$.

All this goes through even if $\langle , \rangle$ is only semi-definite, except that we lose uniqueness. One could regain it by insisting on the 'best' b.a., i.e., the one one would obtain by using Gram-Schmid while ignoring all p_r whose 'norm' is zero.

For **example**, consider the discrete inner product

$$\langle f, g\rangle := \sum_{\xi\in\Xi} w(\xi) f(\xi) g(\xi)$$

with Ξ a finite point set, $\#\Xi = n$ say, and the **weight** function w positive on Ξ. This is definite on $\pi_{<n}$, but gives the nontrivial polynomial $\prod_{\xi\in\Xi}(\cdot - \xi)$ zero 'norm'. In fact, this is necessarily p_n for this inner product, since it lies in $()^n + \pi_{<n}$ and has smallest possible norm while, in $\pi_{<n}$, this discrete inner product is definite, hence there is a unique b.a. from that space to any function. If we continue now, ignoring p_n in the calculation of p_{n+1} since p_n is zero as far as the inner product is concerned, then also p_{n+1} is unique. It is the unique element of $()^{n+1} + \pi_{<n}$ of zero 'norm'. Thus, by induction, we obtain p_j for any $j \geq n$ as the error in the polynomial interpolant to $()^j$ from $\pi_{<n}$ at Ξ.

We noticed earlier that we can compute $p_1, \ldots, p_n$ if we know p_n and p_{n-1}. One way would be to run the recurrence backwards, a very unstable process. The other is to find a suitable inner product and generate the sequence and the corresponding Jacobi matrix by forward recurrence. We try this for the discrete inner product, choosing $\Xi = Z(p_n) := \{\xi : p_n(\xi) = 0\} = \{\xi_1, \ldots, \xi_n\}$, i.e., with p_n and p_{n-1} satisfying (37).

With this choice p_n is indeed orthogonal to $\pi_{<n}$, regardless of the choice of the weights w. If we succeed in choosing the weights positive and so as to make p_{n-1} perpendicular to $\pi_{<n-1}$, we will be done, by the uniqueness of the orthogonal polynomials.

This requires that the linear functional

$$L : \pi_{n-1} \to \mathbb{R} : f \mapsto \langle f, p_{n-1}\rangle = \sum_{i=1^n} (w p_{n-1})(\xi_i) f(\xi_i)$$

have π_{n-2} in its kernel. Since π_{n-2} is the kernel of the divided difference $[\Xi] := \sum_{i=1}^n \prod_{j\neq i}(\xi_i - \xi_j)^{-1}\ [\xi_i]$ as a linear functional on π_{n-1}, this implies that, for some γ, $L = \gamma[\Xi]$. Consequently,

$$w = \gamma/(p_{n-1}p_n') \quad \text{on } \Xi$$

(using the fact that $\prod_{j\neq i}(\xi_i - \xi_j) = p_n'(\xi_i)$). In particular,

$$\|p_{n-1}\|^2 = L(p_{n-1}) = \gamma \sum_i p_{n-1}(\xi_i)/p_n'(\xi_i).$$

On the other hand, since by (37) p_{n-1} has exactly one zero (counting multiplicity) in each interval (ξ_i, ξ_{i+1}) and each ξ_i is a simple zero of p_n, it follows that, for all i, $\mathrm{signum}(p_{n-1}p_n')(\xi_i) = \mathrm{signum}(p_{n-1}p_n')(\xi_n) = 1$, hence, choosing the arbitrary scalar $\gamma = 1$, the weights we get are indeed positive. □

We will meet three-term recurrence relations again in the conjugate gradient method, and for the same reason that we saw them in the consideration of eigenvalues of symmetric tridiagonal matrices and in the Lanczos method, viz., because there are orthogonal polynomials in the background. Conversely, facts about orthogonal polynomials (e.g., nodes and weights for Gaussian quadrature formulae) are often most easily derived by looking at eigenvalues and -vectors of the associated Jacobi matrix.

For example, if we define $P_r = p_r/(b_1 \cdots b_r)$, all r, then $b_r P_r = (\cdot - a_r)P_{r-1} - b_{r-1}P_{r-2}$, all r. This shows that, for each $\xi \in Z(p_n) = Z(P_n)$, the vector $[P_0(\xi); \ldots; P_{n-1}(\xi)]$ is a (nontrivial) eigenvector of the Jacobi matrix T_n belonging to the eigenvalue ξ, hence can be obtained from the normalized eigenvectors for T_n provided by the symmetric QR method.

The Conjugate Gradient Method (often called the 'conjugant' gradient method, an illustration of the power of alliteration in language formation) was derived by Hestenes and Stiefel as a direct method for solving a linear system. It is possible, as the book shows, to spend quite some time deriving it by looking at steepest descent, then conjugate descent directions and the like. It seems more direct and less surprising to think of it instead as a particular instance of least squares approximation by polynomials.

Specifically, the cg method consists in computing

$$\min_{p\in\pi_k} \|x - p(A)b\|_A$$

for given b and given positive definite hermitian A and for $k = 0, 1, 2, \ldots$, thus obtaining a best approximation (=:b.a.) x_k from

$$\Pi_k := \mathrm{ran}[b, Ab, \ldots, A^k b]$$

(note, once again, the Krylov sequence) to the solution x of the equation $A? = b$. The norm here is the one associated with the inner product

$$\langle y, z\rangle_A := y^H A z,$$

i.e.,

$$\|y\|_A^2 := y^H A y.$$

This makes it possible to compute $\langle x, y\rangle_A$ for any y as

$$\langle x, y\rangle_A = b^H y,$$

and that is all we need to be able to do in order to construct the b.a. to x from any linear subspace in the standard way, i.e., by constructing an orthogonal basis $p_0, \ldots, p_k$ for that subspace and then finding the b.a. as

$$\sum_j p_j \langle x, p_j\rangle_A / \|p_j\|_A^2.$$

Our space Π_k is derived from the function space π_k of all polynomials of degree $\le k$, and there are standard methods available to compute the b.a. to some function f from π_k wrto the norm derived from an inner product $\langle , \rangle$: we would generate the corresponding sequence $p_0, p_1, \ldots, p_k$ of (monic) orthogonal polynomials with the aid of the three-term recurrence and calculate the b.a. in the form

$$f_k := \sum_{j\le k} p_j \langle f, p_j\rangle / \langle p_j, p_j\rangle.$$

We would do this inductively, generating first p_k from p_{k-1} and p_{k-2} via the recurrence:

$$p_k = (()^1 - a_k)p_{k-1} - b_{k-1}^2 p_{k-2},$$

with

$$a_k := \langle ()^1 p_{k-1}, p_{k-1}\rangle / c_{k-1}, \quad b_{k-1}^2 := c_{k-1}/c_{k-2}, \quad c_k := \langle p_k, p_k\rangle,$$

and then f_k from f_{k-1} and p_k via $f_k := f_{k-1} + p_k\langle f, p_k\rangle / c_k$. In fact, we would recall the earlier discussion of Modified Gram-Schmidt, hence would compute $\langle f, p_k\rangle$ more accurately as $\langle \varepsilon_{k-1}, p_k\rangle$, with $\varepsilon_{k-1} := f - f_{k-1}$, i.e., would update

$$d_k := p_k \langle \varepsilon_{k-1}, p_k\rangle / c_k, \quad f_k := f_{k-1} + d_k, \quad \varepsilon_k := \varepsilon_{k-1} - d_k.$$

We would initialize the whole process with

$$p_{-1} := 0 =: f_{-1},\ p_0 := ()^0, \quad , \varepsilon_{-1} := f.$$

By defining the inner product by

$$\langle f, g\rangle := (f(A)b)^H A(g(A)b),$$

we have

$$\|x - p(A)b\|_A = \|f - p\|$$

provided we take the function $f := ()^{-1}$, as then $f(A)b = A^{-1}b = x$. This makes the standard machinery for LS approximation by polynomials recalled in the preceding paragraph available for the efficient calculation of the b.a. to x from Π_k. The essence of the cg method lies in the hope that, already for 'small' k, this approximate solution of $A? = b$ is good enough. In any case, one is assured that, for some **finite** k, the approximate solution is exact, at least if the calculation was carried out in exact arithmetic.

You may have a moment's hesitation here, since we are using here functions of matrices. There is no difficulty as long as f and g are polynomials, but we are proposing here to apply it also to the function $()^{-1}$. Of course, the notation is suggestive: $f(A) = A^{-1}$ in this case. But that seems merely slick and by definition. Just to avoid any questions here, we have (at least) two choices.

(i) Since A is hermitian positive definite, it is unitarily similar to a diagonal matrix with positive entries, $A = U^H \Lambda U$, in which case $f(A) = U^H f(\Lambda) U$, hence

$$\langle f, g\rangle = (U^H f(\Lambda) U b)^H U^H \Lambda U U^H g(\Lambda) U b = (Ub)^H f(\Lambda)^H \Lambda g(\Lambda)(Ub) = \sum_i \lambda_i |Ub(i)|^2 \overline{f(\lambda_i)} g(\lambda_i). \tag{38}$$

This looks, once again, just like some ordinary discrete inner product.

(ii) Observe that we can think of A^{-1} also as a polynomial in A, as follows: Let p be the minimal polynomial for A, and write it as $p = ()^1 q - a$ for some polynomial q and with $a := -p(0)$. Since p is A's minimal polynomial, all of its roots must be eigenvalues of A, hence A's invertibility implies that $a \neq 0$. But this implies that $Aq(A) = aI$, hence $A^{-1} = q(A)/a$. Hence we can take $f := q/a$ in the above. This second point of view (as did the first implicitly) gives us a strong hint as to the success of conjugate gradient: The degree of the minimal polynomial for A may well be much smaller than the order n of the matrix A. In that case, we will obtain the solution to the linear system $A? = b$ much earlier than after n steps.

If we now carry out the calculation, we do not want to deal explicitly with the functions p_k, f_k and ε_k. Rather, we are only interested in the **vectors**

$$p_k := p_k(A)b, \quad x_k := f_k(A)b, \quad \varepsilon_k := \varepsilon_k(A)b. \tag{39}$$

(I trust that the use of p_k for both the polynomial p_k and the vector $p_k(A)b$ stresses the close connection and doesn't confuse.) We observe that, for any functions f and g,

$$\langle f, g\rangle = \langle f(A)b, g(A)b\rangle_A := (f(A)b)^H A(g(A)b).$$

Hence, in terms of the vectors (39), the earlier calculation reads

$$p_k = (A - a_k)p_{k-1} - b_{k-1}^2 p_{k-2},$$

with

$$a_k := \langle Ap_{k-1}, p_{k-1}\rangle_A / c_{k-1}, \quad b_{k-1}^2 := c_{k-1}/c_{k-2}, \quad c_k := \langle p_k, p_k\rangle_A,$$
$$d_k := p_k \langle \varepsilon_{k-1}, p_k\rangle_A / c_k, \quad x_k := x_{k-1} + d_k, \quad \varepsilon_k := \varepsilon_{k-1} - d_k.$$

We would initialize the whole process with

$$p_{-1} := 0 =: f_{-1}, \quad p_0 := b, \quad \varepsilon_{-1} := x := A^{-1}b.$$

In the calculations, we would make use of the fact that $\langle \varepsilon_j, y\rangle_A = (A\varepsilon_j)^H y = r_j^H y$, with

$$r_j := A\varepsilon_j = b - Ax_j$$

the **residual** of the approximate solution x_j. Once we agree to keep track of the residual, then there are certain simplifications possible because of the fact that two vectors y and z are A-orthogonal (or **conjugate** (AHA!)) iff Ay and z are orthogonal, i.e., $(Ay)^H z = 0$. Here are the details.

The vector p_j has the form $p_j(A)b$ for some polynomial p_j of degree j. Hence $[p_0, \dots, p_k]$ provides an A-orthogonal basis for Π_k. Also, $x_{k-1} = f_{k-1}(A)b \in \Pi_{<k}$, hence $r_{k-1} = b - Ax_{k-1} \in b + A\Pi_{<k} \subset \Pi_k$. On the other hand, $r_{k-1} = A\varepsilon_{k-1}(A)b$, with the function ε_{k-1} perpendicular to the function space π_{k-1}, hence

$$r_{k-1}^H p_j = 0 \quad \forall j < k. \tag{40}$$

This implies that

$$r_{k-1}^H r_j = 0 \quad \forall j < k-1. \tag{41}$$

It also implies that r_{k-1} has some nontrivial component in the p_k-direction, i.e., a nontrivial scalar multiple of p_k can be obtained by modifying r_{k-1} to be A-orthogonal to $\Pi_{<k}$. Yet, since $A\Pi_{k-2} \subset \Pi_{<k} \perp r_{k-1}$ by (40), r_{k-1} is already A-orthogonal to Π_{k-2}, hence we can obtain (a nontrivial scalar multiple of) p_k (which we call again just p_k, even though it might now be differently normalized) as

$$p_k = r_{k-1} + \beta_k p_{k-1}, \tag{42}$$

with

$$\beta_k := -\langle r_{k-1}, p_{k-1}\rangle_A / \langle p_{k-1}, p_{k-1}\rangle_A.$$

With p_k available, we can proceed to get $x_k = f_k(A)b$ by adding to x_{k-1} the p_k-component of x, i.e., the p_k-component of r_{k-1}, i.e., by computing

$$x_k = x_{k-1} + \alpha_k p_k \quad \text{with } \alpha_k := r_{k-1}^H p_k / p_k^H A p_k.$$

This also gives

$$r_k := r_{k-1} - \alpha_k A p_k.$$

From this and (41),

$$\|r_k\|_2^2 = -\alpha_k \langle r_k, p_k\rangle_A, \tag{43}$$

while from (42) and (40)

$$r_{k-1}^H p_k = \|r_{k-1}\|_2^2.$$

Hence

$$\alpha_k = \|r_{k-1}\|_2^2 / \|p_k\|_A^2.$$

From this and (43) (with $k-1$ rather than k),

$$\beta_k = -(\|r_{k-1}\|_2^2/(-\alpha_{k-1}))/\|p_{k-1}\|_A^2 = \frac{\|r_{k-1}\|_2^2 \|p_{k-1}\|_A^2}{\|r_{k-2}\|_2^2 \|p_{k-1}\|_A^2} = \|r_{k-1}\|_2^2/\|r_{k-2}\|_2^2.$$

Altogether, this gives the **conjugate gradient iteration**:

$$p_k := r_{k-1} + \beta_k p_{k-1}, \quad x_k := x_{k-1} + \alpha_k p_k, \quad r_k := r_{k-1} - \alpha_k A p_k,$$

with

$$\beta_k := \|r_{k-1}\|_2^2/\|r_{k-2}\|_2^2, \quad \alpha_k := \|r_{k-1}\|_2^2/\|p_k\|_A^2.$$

We would start with $k = 1$, having initialized the whole process with

$$p_{-1} := 0 =: x_{-1},\ p_0 := b =: r_{-1}.$$

The process becomes undefined when $\|r_{k-2}\|_2 = 0$ and/or $\|p_k\|_A = 0$. In the first case, we would have stopped at $k-2$ since we would have found the exact solution to $A? = b$. In the second case, we conclude (A being positive definite) that the vector $p_k = 0$, hence, with (42) and (40), already $r_{k-1} = 0$, and this implies that we would have stopped at $k-1$.

In exact arithmetic, the cg iteration converges in finitely many steps, viz. with step $k-1$ in case

$$k = \#\{\lambda \in \lambda(A) : b \not\perp \ker(A - \lambda)\}.$$

In noisy arithmetic, it is more appropriate to treat it as an iteration, hence to apply standard analysis of iterative methods to it.

Convergence of cg can be seen as follows. Since $x_k = f_k(A)b$ is the b.a. from π_k to $()^{-1}$ wrto the norm

$$\|f\|^2 = \|f(A)b\|_A^2 = \sum_i \lambda_i |Ub(i)|^2 |f(\lambda_i)|^2,$$

we readily estimate

$$\|x - x_k\|_A^2 \leq \|b\|_A^2 (\operatorname{dist}((\,)^{-1}, \pi_k)_{\infty,[a,b]})^2,$$

with $[a, b]$ any interval containing $\lambda(A)$. The smallest such interval is the interval

$$[m, M] := [\lambda_1(A), \lambda_n(A)].$$

This indicates that cg might not be very effective when $\lambda_1(A)$ is 'close' to 0. It also indicates that there may be real troubles (and there are) when A is real symmetric and invertible but not positive definite, as then we must approximate $()^{-1}$ on both sides of 0, and that is much harder.

In our situation of a positive definite A, we need to compute, or at least bound, the number

$$E_k := \operatorname{dist}((\,)^{-1}, \pi_k)_{\infty,[m,M]}.$$

A linear change of variables which carries the interval $[m, M]$ to the interval $[-1, 1]$ carries 0 to the point

$$r := \frac{M+m}{M-m} = \frac{\kappa+1}{\kappa-1}$$

with

$$\kappa := \kappa_2(A) = \lambda_{\max}(A)/\lambda_{\min}(A) = M/m.$$

This implies that we can compute E_k as

$$E_k = \frac{2}{M-m} \operatorname{dist}((r - \cdot)^{-1}, \pi_k)_{\infty,[-1,1]}.$$

It turns out that

$$g := (r - \cdot)^{-1}$$

is one of the very few nontrivial functions for which it is possible to write down a formula for its best uniform approximation from π_k on $[-1, 1]$, - Bernstein did that some time ago. This means that Bernstein also has a formula for E_k (source: Timan, page 76):

$$E_k = \frac{1}{(r^2-1)(r+\sqrt{r^2-1})^k}.$$

But, up to small terms, we can obtain such a result immediately as follows. For any function p,

$$\|g - p\|_\infty \in [1/(r+1), 1/(r-1)] \|1 - (r - \cdot)p\|_\infty.$$

On the other hand,

$$\min_{p \in \pi_k} \|1 - (r - \cdot)p\|_\infty = \min\{\|q\|_\infty : q \in \pi_{k+1}, q(r) = 1\} = 1/\max_{q \in \pi_{k+1}} q(r)/\|q\|_\infty = 1/C_{k+1}(r),$$

with C_j the Chebyshev polynomial (for the interval $[-1, 1]$) of degree j. From the recurrence relation for Chebyshev polynomials (cf. apr.26 discussion of Lanczos method),

$$C_j(t) = [(t + \sqrt{t^2-1})^j + (t - \sqrt{t^2-1})^j]/2,$$

hence

$$(E_k)^{1/k} \to 1/(r + \sqrt{r^2 - 1})$$

monotonely from below as $k \to \infty$. In terms of the condition $\kappa = \kappa_2(A) = M/m$ of A, we have $r^2 - 1 = \frac{(\kappa+1)^2-(\kappa-1)^2}{(\kappa-1)^2} = 4\kappa/(\kappa-1)^2$, hence

$$\frac{1}{r + \sqrt{r^2 - 1}} = \frac{\kappa - 1}{1 + \kappa + 2\sqrt{\kappa}} = \frac{\sqrt{\kappa}^2 - 1}{(\sqrt{\kappa} + 1)^2} = \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}.$$

This shows that, for every k,

$$\|x - x_k\|_A^2 \leq \|b\|_A^2 \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^{2k}.$$

This error analysis has replaced dist $()_{\infty,\Lambda}$ by dist $()_{\infty,[m,M]}$, and that is often a very pessimistic change. In practice, convergence is much faster. Heuristically, the method reduces the contributions of the eigenvectors belonging to extreme eigenvalues fastest. This means that, in some sense, the interval on which we have to approximate $()^{-1}$ from π_k shrinks as k grows, giving us a better condition number to work with.

preconditioned cg is a natural response to the error analysis for cg just given. Since the convergence rate is 'slow' to the extent that $\kappa_2(A)$ is 'large', one makes an 'easily invertible' change of coordinates $x \to \tilde{x} := Cx$. In terms of these coordinates, $\tilde{x}$ satisfies the linear system $AC^{-1}\tilde{?} = b$, or $\tilde{A}\tilde{?} = \tilde{b}$, with $\tilde{A} := C^{-1}AC^{-1}$. The hope is that $\kappa_2(\tilde{A})$ is smaller than $\kappa_2(A)$. For example, writing $A = C^TC$ (as we may by, e.g., using the LDL^T factorization of A), we get $\tilde{A} = I$, hence the perfect condition 1. Of course, if we are willing to compute such a C, we might as well solve $A? = b$ directly. Thus the point is to pick up an 'easily invertible' C for which nevertheless $\kappa_2(\tilde{A})$ is 'small'. Popular candidates for C include the (squareroot of the) diagonal of A or the C obtained by 'incomplete factorization' in which $A \sim C^TC$ with C as sparse as A.

Since we are interested ultimately in solving $A? = b$, one modifies the earlier description of the cg iteration to maintain r_k and x_k rather than $\tilde{r}_k$ and $\tilde{x}_k$. Note that, with $\tilde{x}_k$ the approximate solution obtained and $\tilde{r}_k := \tilde{b} - \tilde{A}\tilde{x}_k$ its residual, $C^{-1}\tilde{r}_k$ is **not** the residual in the approximate solution $x_k := C^{-1}\tilde{x}_k$. Rather,

$$r_k := b - Ax_k = b - AC^{-1}\tilde{x}_k = b - C\tilde{A}\tilde{x} = C\tilde{r}_k.$$

This introduces a notational difficulty which the book resolves by defining

$$z_k := C^{-1}\tilde{r}_k = M^{-1}r_k, \quad M := C^2, \quad r_k : b - Ax_k.$$

These z_k come in handy when we need to calculate $(\tilde{r}_k)^H(\tilde{r}_k) = (C^{-1}r_k)^HC^{-1}r_k = z_k^Hr_k$. We do not bother to construct the $\tilde{p}_k$ explicitly, but merely carry the $p_k = C^{-1}\tilde{p}_k = C^{-1}(\tilde{r}_{k-1} + \beta_k\tilde{p}_{k-1}) = z_{k-1} + \beta_kp_{k-1}$. This gives the following **pre-conditioned cg iteration** (obtained here by copying from the may.3 note and modifying):

$$p_k := z_{k-1} + \beta_kp_{k-1}, \quad x_k := x_{k-1} + \alpha_kp_k, \quad r_k := r_{k-1} - \alpha_kAp_k,$$

with

$$Mz_{k-1} := r_{k-1}$$

and

$$\beta_k := z_{k-1}^Hr_{k-1}/z_{k-2}^Hr_{k-2}, \quad \alpha_k := z_{k-1}^Hr_{k-1}/\|p_k\|_A^2.$$

We would start with $k = 1$, having initialized the whole process with

$$p_{-1} := 0 =: x_{-1}, \; p_0 := z_{-1}, r_{-1} := b.$$

The **Perron-Frobenius** theory asserts that any irreducible nonnegative matrix has an eigenvalue equal to its spectral radius and that this eigenvalue is algebraically (hence also geometrically) simple, with the corresponding eigenvector positive in all its entries. The standard arguments (e.g., in Varga, or Franklin) are somewhat long. But the gist of the argument is easy to give.

Call a matrix A **nonnegative** and write this $A \geq 0$ if it is nonnegative as a function, i.e., if $A(i,j) \geq 0$ for all i, j. Such a matrix maps the nonnegative orthant

$$\mathbb{R}^n_+ := \{x \in \mathbb{R}^n : x \geq 0\}$$

into itself. Call A **positive** and write $A > 0$ in case all its entries are positive. A positive matrix maps $\mathbb{R}^n_+$ into its interior, i.e., the map

$$S_+ \to S_+ : x \mapsto Ax/\|Ax\|$$

on

$$S_+ := \{x \in \mathbb{R}^n_+ : \|x\| = 1\}$$

is continuous and a strict contraction, hence has a unique fixed point. (You might prefer here to use the 1-norm, for then S_+ is a cut-out from the hyperplane $\sum_j x(j) = 1$.) This means that we can find x in the

relative interior of S_+ for which $Ax = \|Ax\|x$. In particular, x is an eigenvector for A with corresponding positive eigenvalue $\lambda := \|Ax\|$.

Note that we found this eigenpair by the power method and, by the argument, we would have reached it starting from any nontrivial nonnegative vector. Since, for any nontrivial vector y (in particular, for any eigenvector of A) we can find some positive vector z with $z^T y \neq 0$, it follows that λ must be the largest eigenvalue of A and that its ascent must be one.

Assume now that there is also an eigenvector y for λ independent of x. Then we may assume that y has some negative component, hence there is a point z in the segment $[y, x]$ on the boundary of $\mathbb{R}^n_+$ and this z cannot be zero (since x and y are linearly independent) and is also an eigenvector for λ. But that is nonsense since A carries any nontrivial vector in $\mathbb{R}^n_+$ into the interior.

This is the original theory, due to Oskar Perron. Frobenius improved it by weakening the conditions on A. Specifically, Frobenius showed the same conclusion under the weaker assumption that A is nonnegative but **irreducible**, meaning that, for every pair of distinct indices i, j, there is a sequence $i = i_1, i_2, \ldots, i_r = j$ so that $\prod_k A(i_k, i_{k+1}) \neq 0$. The nonnegativity implies that $x \mapsto Ax/\|x\|$ maps S_+ into itself, and the irreducibility guarantees that the map has at most one fixed point and that this fixed point necessarily lies in the (relative) interior. With this, the geometric simplicity is demonstrated as before.

determinants are often brought into courses such as this quite unnecessarily. But when they are useful, they are remarkably so. The use of determinants is a bit bewildering to the beginner. I have found that, of the many, many determinant identities available, in the end I have always managed with just one, viz. **Sylvester's determinant identity**, and this is nothing but Gauss elimination.

For $A \in \mathbb{R}^{m\times n}$, and $\mathbf{i}, \mathbf{j}$ sequence with ranges $1, \dots, m$ and $1, \dots, n$ resp., $A(\mathbf{i};\mathbf{j})$ is the matrix fashioned by the rows and columns of A so indicated. In MATLAB, we would have $A(\mathbf{i};\mathbf{j}) = A(\mathbf{i},\mathbf{j})$. Precisely,

$$A(\mathbf{i};\mathbf{j})(i,j) := A(\mathbf{i}(i),\mathbf{j}(j)), \quad \forall i,j.$$

For $[a_1,\dots,a_n] = A \in \mathbb{F}^{n\times n}$,

$$\det[a_1,\dots,a_n] = \det(A)$$

is, by definition, a multilinear alternating form in its n arguments. At a minimum, this means that

$$\mathbb{F}^n :\to \mathbb{F} : a \mapsto \det[a, a_2, \dots, a_n]$$

is linear, and that

$$\det[..., a, ..., b, ...] = -\det[..., b, ..., a, ...]$$

(with the various ellipses indicating the other arguments, held fixed). Note that this **alternation** property implies that the determinant is zero if two of its arguments coincide, which, together with the linearity, implies that

$$\det[..., a + \alpha b, ..., b, ...] = \det[..., a, ..., b, ...].$$

It follows that

$$\det[a_1,\dots,a_n] = \det[\dots, \sum_i e_i a_j(i), \dots] = \sum_{\mathbf{i}\in\{1,\dots,n\}^n} \det[e_{\mathbf{i}(1)},\dots,e_{\mathbf{i}(n)}] \prod_j a_j(\mathbf{i}(j)),$$

just using linearity. By the alternation property, most of these summands are zero. Only those determinants $\det[e_{\mathbf{i}(1)},\dots,e_{\mathbf{i}(n)}]$ for which $\mathbf{i} \in \$_{\mathbf{n}}$:= symmetric group of order n, i.e., for which $\mathbf{i}$ is a permutation of the sequence $\mathbf{n} := \{1,\dots,n\}$, are not automatically zero. Further, for such $\mathbf{i}$,

$$\det[e_{\mathbf{i}(1)},\dots,e_{\mathbf{i}(n)}] = (-)^{\mathbf{i}} \det(I)$$

by the alternation property, with $(-)^{\mathbf{i}} = \pm 1$ depending on whether it takes an even or an odd number of interchanges to change $\mathbf{i}$ into a strictly increasing sequence. (This requires the proof that it is always possible to convert any such $\mathbf{i}$ into an increasing sequence by interchanges (easy) and that, while this might be done in many ways, the parity of such number of interchanges is independent of how we got from $\mathbf{i}$ to $\mathbf{n}$.) Thus det is entirely pinned down once we know the number $\det(I)$. It is customary to choose

$$\det(I) := 1.$$

On the other hand, starting with the resulting formula

$$\det[a_1,\dots,a_n] = \sum_{\mathbf{i}\in\$_{\mathbf{n}}} (-)^{\mathbf{i}} \prod_j a_j(\mathbf{i}(j)) \tag{44}$$

as a definition, one readily verifies that det so defined satisfies the three properties claimed for it.

The same argument shows that

$$\begin{aligned}\det(AB) &= \det[\dots, \sum_i a_j b_1(i), \dots] \\ &= \sum_{\mathbf{i}\in\$_{\mathbf{n}}} \det[a_{\mathbf{i}(1)},\dots,a_{\mathbf{i}(n)}] \prod_j b_j(\mathbf{i}(j)) \\ &= \sum_{\mathbf{i}\in\$_{\mathbf{n}}} \det(A)(-)^{\mathbf{i}} \prod_j b_j(\mathbf{i}(j)) = \det(A)\det(B).\end{aligned}$$

In particular, with

$$B := [e_1, \ldots, e_{j-1}, x, e_{j+1}, \ldots, e_n],$$

we have $\det(B) = x(j)$, hence

$$\det(A)x(j) = \det(AB) = \det[a_1, \ldots, a_{j-1}, Ax, a_{j+1}, \ldots, a_n], \tag{45}$$

therefore (with $b := Ax$)

$$x(j) = \det[a_1, \ldots, a_{j-1}, b, a_{j+1}, \ldots, a_n]/\det(A), \quad \forall j,$$

in case A is **nonsingular**, i.e., $\det(A) \neq 0$. This is **Cramer's rule** for the entries of the solution of $A? = b$. It shows that a nonsingular A is invertible since, for a noninvertible A, we can find x and j with $x(j) \neq 0$ for which $Ax = 0$, hence the right side of (45) is zero, and so must $\det(A)$ be since $x(j) \neq 0$. (Equivalently, it shows that the columns of A are linearly independent.)

On the other hand, if A is invertible, then $1 = \det(I) = \det(AA^{-1}) = \det(A)\det(A^{-1})$, hence A is nonsingular and $\det(A^{-1}) = \det(A)^{-1}$. Thus invertibility and nonsingularity coincide, accounting for the great popularity of determinants as a means of telling whether or not a given matrix is invertible. But, if such telling is done numerically, then it is usually faster just to solve $A? = b$ directly than to compute $\det(A)$.

One verifies directly from the formula (44) that $\det A^T = \det A$. Also from that formula, one obtains the very useful **Laplace's expansion by the last column**:

$$\begin{aligned}
\det A &= \sum_{\mathbf{i} \in \$_{\mathbf{n}}} a_n(\mathbf{i}(n))(-)^{\mathbf{i}} \prod_{j<n} a_j(\mathbf{i}(j)) \\
&= \sum_{i=1}^{n} a_n(i) \sum_{\mathbf{i} \in \$_{\mathbf{n}\setminus i}} a_n(\mathbf{i}(n))(-)^{\mathbf{i},i} \prod_{j<n} a_j(\mathbf{i}(j)) \\
&= \sum_{i=1}^{n} a_n(i)(-)^{n-i} \det A(\mathbf{n}\setminus i; 1, \ldots, n-1)
\end{aligned}$$

using the fact that $(-)^{\mathbf{i},i} = (-)^{\mathbf{i}}(-)^{n-i}$. Finally, also from (44),

$$\det T = \prod_j T(j,j)$$

for any **triangular** matrix T, since for any $\mathbf{i}$ other than $\mathbf{n}$, there is a j with $\mathbf{i}(j) < j$ and a j with $\mathbf{i}(j) > j$, hence triangularity implies that $\prod_j T(j, \mathbf{i}(j)) = 0$ for all $\mathbf{i} \neq \mathbf{n}$. Thus, for general A, the fastest way to compute $\det A$ is via an LU (or PLU) factorization.

Now comes the promised Sylvester's Identity which is a direct consequence of the following neat description of Gauss elimination.

With $\mathbf{k} := \{1, \ldots, k\}$, consider the matrix

$$B(i,j) := \det A(\mathbf{k}, i; \mathbf{k}, j).$$

Then, on expanding by entries of the last row,

$$B(i,j) = A(i,j)\det A(\mathbf{k};\mathbf{k}) - \sum_{r \leq k} A(i,r)(-)^{k-r} \det A(\mathbf{k}; (\mathbf{k}\setminus r), j).$$

This shows that

$$B(:,j) \in A(:,j)\det A(\mathbf{k};\mathbf{k}) + \operatorname{span} A(:,\mathbf{k}),$$

while, directly, $B(i,j) = 0$ for $i \in \mathbf{k}$ since then $\det A(\mathbf{k}, i; \mathbf{k}, j)$ has two rows the same.

In the same way,

$$B(i,:) \in A(i,:) \det A(\mathbf{k};\mathbf{k}) + \operatorname{span} A(\mathbf{k},:),$$

while,directly, $B(i,j) = 0$ for $j \in \mathbf{k}$. Thus, if $\det A(\mathbf{k};\mathbf{k}) \neq 0$, then, for $i > k$,

$$B(i,:)/\det A(\mathbf{k};\mathbf{k})$$

provides the ith row of the matrix obtained from A after k steps of Gauss elimination. In other words, the matrix $S := B/\det A(\mathbf{k};\mathbf{k})$ provides the **Schur complement** $S(k+1,\dots,n;k+1,\dots,n)$ in A of the **pivot block** $A(\mathbf{k};\mathbf{k})$.

Since such row elimination is done by elementary matrices with determinant equal to 1, it follows that

$$\det A = \det A(\mathbf{k};\mathbf{k}) \det S(k+1,\dots,n;k+1,\dots,n).$$

Since, for any $\#\mathbf{i} = \#\mathbf{j}$, $B(\mathbf{i},\mathbf{j})$ depends only on the square matrix $\det A(\mathbf{k},\mathbf{i};\mathbf{k},\mathbf{j})$, this implies

Sylvester's determinant identity. *If*

$$S(i,j) := \det A(\mathbf{k},i;\mathbf{k},j)/\det A(\mathbf{k};\mathbf{k}), \quad \forall i,j,$$

then

$$\det S(\mathbf{i};\mathbf{j}) = \det A(\mathbf{k},\mathbf{i};\mathbf{k},\mathbf{j})/\det A(\mathbf{k};\mathbf{k}).$$