March 26, 2018

Lecture 3 – Factorisations

A Decompositions

Exercise 3.1. **The stability of decompositions.** Let $\mathbf{A}$, $\mathbf{L}$ and $\mathbf{U}$ be $n \times n$ non-singular matrices such that $\mathbf{A} = \mathbf{LU}$. We are interested in the quantities

$$\gamma_1 \equiv \frac{\| \, |\mathbf{L}| \, |\mathbf{U}| \, \|}{\|\mathbf{A}\|} \quad \text{and} \quad \gamma_2 \equiv \frac{\|\mathbf{L}\| \, \|\mathbf{U}\|}{\|\mathbf{A}\|}$$

for some induced norm.

(a) Show that $\gamma_2 \leq \min(\mathcal{C}(\mathbf{L}), \mathcal{C}(\mathbf{U}))$, where $\mathcal{C}(\mathbf{A})$ is the condition number of a matrix $\mathbf{A}$ with respect to the used norm.

(b) Show that for p-norms, $\gamma_1 \leq \gamma_2$.

(c) Now, suppose $\mathbf{L}$ is unitary. Show that $\gamma_2 = 1$ and $\gamma_1 \leq n$ w.r.t. the 2-norm and $\gamma_2 = 1$, $\gamma_1 \leq \sqrt{n}$ w.r.t. the F-norm .

B Orthonormal matrices

Orthonormal matrices (see Lecture 0.C) play an important role in Numerical Linear Algebra. The columns represent an orthonormal basis of a subspace. Stable computations require the use of well-conditioned bases. The best conditioned basis is an orthonormal one.

In many methods high dimensional problems are approximately solved by projecting them onto low dimensional spaces, thus turning the 'hard' high dimensional problem into an 'easy' low dimensional one. To have optimal stability orthogonal projection are often used.

A map $\mathbf{P}$ from $\mathbb{C}^n$ onto a subspaces $\mathcal{V}$ of $\mathbb{C}^n$ is an **orthogonal projection** onto $\mathcal{V}$ if

$$\begin{aligned} &\mathbf{Px} \in \mathcal{V} \ \text{ for any } \mathbf{x} \in \mathbb{C}^n, \\ &\mathbf{P}\,\mathbf{x} = \mathbf{x} \ \text{ for any } \mathbf{x} \text{ in } \mathcal{V}, \\ &\mathbf{x} - \mathbf{Px} \perp \mathbf{Px} \ \text{ for any } \mathbf{x} \in \mathbb{C}^n. \end{aligned}$$

The first two properties define a **projection** onto $\mathcal{V}$, the third property makes $\mathbf{P}$ orthogonal. The orthogonal projection onto $\mathcal{V}$ is unique as we will learn in the next exercise. Non-orthogonal projections are also called **skew projections**.

Orthogonal projections are conveniently defined by means of an orthonormal basis in the image space $\mathcal{V}$ of the projection, as is explained in the next exercise. In Exercise 3.6, we will see that any basis of $\mathcal{V}$ can be used to define the orthogonal projection. In part (d) of this exercise, we will also see that how skew projections can be used to project orthogonal to $\mathcal{V}$.

Exercise 3.2. **Orthogonal projections.** In this exercise we use the standard inner product and associated 2-norm:

$$(\mathbf{x}, \mathbf{y}) = \mathbf{y}^* \mathbf{x} = \mathbf{y}^{\mathrm{H}} \mathbf{x} = \sum \bar{y}_j x_j. \tag{3.1}$$

Let $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_k]$ be an $n \times k$ matrix.

(a) Show that $\mathbf{V}$ is orthonormal if and only if $\mathbf{V}^*\mathbf{V} = I$, where I is the $k \times k$ identity matrix.

(b) Is the product $\mathbf{VV}^*$ defined (i.e., do the dimensions match)? Is this product also equal to an identity matrix if $\mathbf{V}$ is orthonormal?

(c) Suppose $\mathbf{V}$ is orthonormal.
Show that both $\mathbf{P} \equiv \mathbf{VV}^*$ and $\mathbf{Q} \equiv \mathbf{I} - \mathbf{VV}^*$ are projections (i.e., $\mathbf{PP} = \mathbf{P}$) and Hermitian ($\mathbf{P}^* = \mathbf{P}$).
Show that $\mathbf{P}$ projects orthogonally on the $\mathcal{V} \equiv \text{span}(\mathbf{V}) \equiv \text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$.
Show that any map (matrix) $\widetilde{\mathbf{P}}$ that *projects orthogonally* onto $\mathcal{V}$ is equal to $\mathbf{P}$: $\mathbf{P}$ is the

orthogonal projection onto $\mathcal{V}$.
Show that any *Hermitian* map (matrix) $\widetilde{\mathbf{P}}$ that *projects* onto $\mathcal{V}$ is equal to $\mathbf{P}$
Show that $\mathbf{Q}$ is also an orthogonal projection (onto what space?).

(d) How many flop does it take to compute $\mathbf{Px}$? And to compute $\mathbf{Qx}$?
Note that, for any $\mathbf{x} \in \mathbb{C}^n$, with $\mathbf{x}_{\mathcal{V}} \equiv \mathbf{Px}$ and $\mathbf{x}_{\mathcal{V}^\perp} \equiv \mathbf{x} - \mathbf{Px} = \mathbf{Qx}$, we have that

$$\mathbf{x} = \mathbf{x}_{\mathcal{V}} + \mathbf{x}_{\mathcal{V}^\perp}, \quad \mathbf{x}_{\mathcal{V}} \in \mathcal{V}, \quad \mathbf{x}_{\mathcal{V}^\perp} \perp \mathcal{V}:$$

the orthogonal projection $\mathbf{P}$ onto $\mathcal{V}$ provides an efficient way to compute the component the component in the subspace $\mathcal{V}$ as well as its orthogonal complement of any vector.

(e) Show that, if $\mathbf{V}$ is orthonormal,

$$\mathbf{I} - \mathbf{V}\mathbf{V}^* = \mathbf{I} - \sum_{j=1}^{k} \mathbf{v}_j\mathbf{v}_j^* = \prod_{j=1}^{k}(\mathbf{I} - \mathbf{v}_j\mathbf{v}_j^*).$$

(Note: If I is the $k \times k$ identity, then $I = \sum_{j=1}^{k} e_j e_j^*$. Therefore, if $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_k]$ is an $n \times k$ matrix and $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_k]$ is an $m \times k$ matrix, then we have that

$$\mathbf{A}\mathbf{B}^* = \sum_{j=1}^{k} \mathbf{A} e_j e_j^* \mathbf{B}^* = \sum_{j=1}^{k} \mathbf{a}_j \mathbf{b}_j^*.$$

Here, we expressed $\mathbf{AB}^*$ as the sum of rank one matrices.)

Exercise 3.3. **One dimensional range.** Let $\mathbf{A}$ be an $n \times k$ matrix.

(a) Prove that $\mathcal{R}(\mathbf{A}) = \mathcal{N}(\mathbf{A}^*)^\perp$ (see also Exercise 0.4(c)).
Assume $\mathcal{N}(\mathbf{A}^*) = \text{span}(\mathbf{v})$ for some $\mathbf{v}$ with $\|\mathbf{v}\|_2 = 1$ (that is, $k = n$ or $n-1$).

(b) Prove that $\mathcal{R}(\mathbf{A}) = \{(\mathbf{I} - \mathbf{v}\,\mathbf{v}^*)\mathbf{y} \,\big|\, \mathbf{y}\}$.
Let $\mathbf{b}$ be a k vector.

(c) Let $\mathbf{x}$ be the least square solution of $\mathbf{Ax} = \mathbf{b}$, i.e., $\mathbf{x} = \text{minarg}_{\widetilde{\mathbf{x}}} \|\mathbf{A}\widetilde{\mathbf{x}} - \mathbf{b}\|_2$. Prove that

$$\mathbf{b} - \mathbf{Ax} = \mathbf{v}(\mathbf{v}^*\mathbf{b}), \qquad \|\mathbf{b} - \mathbf{Ax}\|_2 = |\mathbf{v}^*\mathbf{b}|.$$

(d) Let $\mathbf{c}$ be a k-vector such that $\mathbf{v}^*\mathbf{c} \neq 0$, $\mathbf{c} \perp \mathbf{b}$ and $\|\mathbf{c}\|_2 = 1$.
Show that $\mathbf{b} + \beta\mathbf{c} \in \mathcal{R}(\mathbf{A})$ for some scalar β. Conclude that $(\mathbf{I} - \mathbf{c}\,\mathbf{c}^*)\mathbf{Ax} = \mathbf{b}$ for some $\mathbf{x}$ and

$$\mathbf{b} - \mathbf{Ax} = \mathbf{c}\,\frac{\mathbf{v}^*\mathbf{b}}{\mathbf{v}^*\mathbf{c}}, \qquad \|\mathbf{b} - \mathbf{Ax}\|_2 = \frac{|\mathbf{v}^*\mathbf{b}|}{|\mathbf{v}^*\mathbf{c}|}$$

Householder reflections of Exercise 3.4 and Givens rotations of Exercise 3.5 are important elementary unitary matrices that allow stable transformations and that are frequently use to bring matrices to a simpler form. For a typical application, see Exercise 3.16.

Exercise 3.4. **Householder reflections.** Let $\mathbf{V} \equiv [\mathbf{v}_1, \ldots, \mathbf{v}_k]$ be an orthonormal $n \times k$ matrix. Let $\mathcal{V} \equiv \text{span}(\mathbf{V})$. Consider the **Householder reflection**

$$\mathbf{H} \equiv \mathbf{I} - 2\mathbf{V}\mathbf{V}^*.$$

(a) Use the results Exercise 3.2 to show that $\mathbf{Hx}$ reflects $\mathbf{x}$ with respect to the 'mirror' $\mathcal{V}^\perp$ (with $n = 3$ and $k = 1$, $\mathcal{V}^\perp$ is a plane). Show that $\mathbf{H}$ is unitary, Hermitian, and the inverse of $\mathbf{H}$ is $\mathbf{H}$ itself.

(b) Discuss the computational costs to perform a matrix-vector product $\mathbf{Hx}$. Discuss the memory requirements to store $\mathbf{H}$.

(c) Let $\mathbf{x}$ and $\mathbf{y}$ be non-trivial vectors in $\mathbb{C}^n$.
For a vector $\mathbf{v}$ and a scalar τ (in $\mathbb{C}$) consider the following two statements

$$\|\mathbf{v}\|_2 = 1, \qquad (\mathbf{I} - 2\mathbf{v}\mathbf{v}^*)\mathbf{x} = \tau\mathbf{y}, \tag{3.2}$$

and

$$\text{i)}\ |\tau| = \frac{\|\mathbf{x}\|_2}{\|\mathbf{y}\|_2}, \quad \text{ii)}\ \tau\mathbf{x}^*\mathbf{y} \in \mathbb{R}, \quad \text{iii)}\ \mathbf{v} = \rho(\mathbf{x} - \tau\mathbf{y}) \ \text{with}\ |\rho| = \frac{1}{\|\mathbf{x} - \tau\mathbf{y}\|_2}. \tag{3.3}$$

Show that (3.2) $\Leftrightarrow$ (3.3).

Note that ρ in (3.3) is not unique: ρ can be replaced by $\zeta\rho$ for any *sign* ζ, i.e., $\zeta \in \mathbb{C}$ with $|\zeta| = 1$. However, this sign is irrelevant for $\mathbf{v}\mathbf{v}^*$, and therefore, for $\mathbf{v}$. Except for a factor -1, the scalar τ is determined by the properties i) and ii) of (3.3), ($-\tau$ has these two properties if that is the case for τ): either $\tau\mathbf{x}^*\mathbf{y} > 0$ or $\tau\mathbf{x}^*\mathbf{y} < 0$ (if $\mathbf{x}^*\mathbf{y} \neq 0$). This essentially leaves two choices for $\mathbf{v}$. What do you expect to be the best choice for minimising the effect of rounding errors in the Householder reflection in (3.2)? Also discus the situation where $\mathbf{x}^*\mathbf{y} = 0$.

Note that $\mathbf{I} - (2/\widetilde{\mathbf{v}}^*\widetilde{\mathbf{v}})\,\widetilde{\mathbf{v}}\,\widetilde{\mathbf{v}}^*$ is a Householder reflection (i.p., unitary) regardless whether $\widetilde{\mathbf{v}}$ is an accurate approximation of the desired $\mathbf{v}$.

(d) Let $\mathbf{w} = (w_j)$ be an n-vector. Put

$$\tau \equiv -\text{sign}(w_1)\,\|\mathbf{w}\|_2, \qquad \widetilde{\mathbf{v}} \equiv \mathbf{w} - \tau\mathbf{e}_1, \qquad \beta \equiv \frac{2}{\widetilde{\mathbf{v}}^*\widetilde{\mathbf{v}}}.$$

Show that, $\mathbf{I} - \beta\,\widetilde{\mathbf{v}}\,\widetilde{\mathbf{v}}^*$ is a Householder reflection,

$$\widetilde{\mathbf{v}}^*\widetilde{\mathbf{v}} = 2\|\mathbf{w}\|_2(\|\mathbf{w}\|_2 + |w_1|), \quad \text{and} \quad (\mathbf{I} - \beta\,\widetilde{\mathbf{v}}\,\widetilde{\mathbf{v}}^*)\mathbf{w} = \tau\,\mathbf{e}_1.$$

With $\mathbf{v} \equiv \widetilde{\mathbf{v}}/\|\widetilde{\mathbf{v}}\|_2$, we have that $\mathbf{I} - \beta\,\widetilde{\mathbf{v}}\,\widetilde{\mathbf{v}}^* = \mathbf{I} - 2\,\mathbf{v}\,\mathbf{v}^*$. Nevertheless, we prefer the first expression. Why?

Exercise 3.5. **Givens rotations.** An $n \times n$ matrix $\mathbf{G}$ is a **Givens rotation** if $\mathbf{G}$ rotates in an (i,j)-plane, i.e., in span$([\mathbf{e}_i, \mathbf{e}_j])$. Using MATLAB notation, $\mathbf{G}$ is the $n \times n$ identity matrix except for the submatrix $\mathbf{G}([i,j],[i,j])$ which is equal to

$$\mathbf{G}([i,j],[i,j]) = \begin{bmatrix} c & s \\ -\bar{s} & c \end{bmatrix}, \quad \text{where} \quad c \equiv \cos(\phi),\ |s|^2 = 1 - c^2,$$

for some $\phi \in [0, 2\pi)$.

(a) Show that $\mathbf{G}$ is unitary.

(b) In practice, a $t \in \mathbb{C}$, a tangent value, is determined and c and s are computed as

$$c = \frac{1}{\sqrt{1 + |t|^2}} \quad \text{and} \quad s = tc \tag{3.4}$$

Let $\mathbf{A}$ be an $n \times n$ matrix.

(c) Show there is a Givens rotation (that rotates in the (i,j)-plane) such that the (i,j)-entry of $\mathbf{G}^*\mathbf{A}\mathbf{G}$ is zero.

(d) What is the effect on the eigenvalues (of $\mathbf{A}$ as compared to $\mathbf{G}^*\mathbf{A}\mathbf{G}$) if c and s are computed using (3.4) and the error in t is large?

(e) The t that shows up in (c) is the root of a degree two polynomial. How do you compute the desired root in view of rounding errors?

(f) Not that multiplication of the bottom row of $\mathbf{G}([i,j],[i,j])$ with -1 turns $\mathbf{G}$ into a unitary Hermitian matrix (actually a Householder reflection) that, as Givens rotations, also can be used to bring matrix entries to 0.

Exercise 3.6. **Orthogonal and skew projections.** Let $\mathbf{V}$ be an $n \times k$ matrix, $k \le n$. Note that $\mathbf{V}$ is not required to be orthogonal. Put $M \equiv \mathbf{V}^*\mathbf{V}$ and $\mathcal{V} \equiv \mathrm{span}(\mathbf{V})$.

Prove that M is non-singular if and only if $\mathbf{V}$ has full rank.

(a) Show that $\mathbf{P} \equiv \mathbf{V}M^{-1}\mathbf{V}^*$ defines the orthogonal projection onto $\mathcal{V}$ if M is non-singular.

(b) Put $E \equiv I - \mathbf{V}^*\mathbf{V}$. Then $M = I - E$ (ideally $E = 0$). Suppose $\|E\| < 1$.
Then M is non-singular (why?). Show that

$$\mathbf{I} - \mathbf{V}(I + E)\mathbf{V}^* = (\mathbf{I} - \mathbf{V}\mathbf{V}^*)^2 \tag{3.5}$$

and, more generally,

$$\mathbf{I} - \mathbf{V}(I + E + E^2 + \ldots + E^m)\mathbf{V}^* = (\mathbf{I} - \mathbf{V}\mathbf{V}^*)^{m+1}.$$

Conclude that

$$\lim_{m\to\infty} (\mathbf{I} - \mathbf{V}\mathbf{V}^*)^m = \mathbf{I} - \mathbf{V}M^{-1}\mathbf{V}^*.$$

Hence, repeating Gram–Schmidt (cf., Lecture 0.C) leads to an orthonormal projection even if the basis is not fully orthonormal (as will happen in rounded arithmetic). Statement (3.5) can be used to prove that twice repeated Gram–Schmidt leads to a nearly orthonormal basis (with $\|E\|_2^2 = \mathcal{O}(\mathbf{u})$). There is a variant that controles the number of repetitions by tracking the tangent of the angle between span($\mathbf{V}$) and the vector to be orthogonalized against span($\mathbf{V}$): see, Alg. 3.1

(c) Consider the case where $1 \ll k \ll n$. Discuss the costs of computing $(\mathbf{I} - \mathbf{V}\mathbf{V}^*)\mathbf{x}$, $(\mathbf{I} - \mathbf{V}M^{-1}\mathbf{V}^*)\mathbf{x}$ and $(\mathbf{I} - \mathbf{V}\mathbf{V}^*)^2\mathbf{x}$ for a vector $\mathbf{x}$ (given $\mathbf{V}$ and considering efficient implementations).

Let $\mathbf{W}$ be a full rank $n \times k$ matrix. Put $T \equiv \mathbf{V}^*\mathbf{W}$. Assume that T is non-singular.

(d) Prove that $\mathbf{W}T^{-1}\mathbf{V}^*$ is a projection onto $\mathcal{W} \equiv \mathrm{span}(\mathbf{W})$, while $\Pi \equiv \mathbf{I} - \mathbf{W}T^{-1}\mathbf{V}^*$ projects onto $\mathcal{V}^\perp$ with null space $\mathcal{W}$: Π is a skew projection that projects orthogonal to $\mathcal{V}$ along $\mathcal{W}$.

Exercise 3.7. **Sherman–Morrison–Woodbury formula.**
Consider low rank matrices $\mathbf{V}$ and $\mathbf{W}$ both of size $n \times k$. Put $M \equiv \mathbf{V}^*\mathbf{V}$ and $T \equiv \mathbf{W}^*\mathbf{V}$.

We are interested in the solution $\mathbf{y}$ of the problem

$$\text{solve} \quad (\mathbf{I} - \mathbf{V}\mathbf{W}^*)\mathbf{y} = \mathbf{x} \quad \text{for } \mathbf{y}. \tag{3.6}$$

(a) Assume $\mathbf{w}$ is an eigenvector of $\mathbf{V}\mathbf{W}^*$ with eigenvalue $\lambda \neq 0$. Show that $\mathbf{W}^*\mathbf{w}$ is an eigenvector of T with eigenvalue λ (in particular $\mathbf{W}^*\mathbf{w} \neq \vec{0}$). Prove that $\lambda \neq 0$ is an eigenvalue of $\mathbf{V}\mathbf{W}^*$ if and only if λ is an eigenvalue of T. Conclude that

$$\mathbf{I} - \mathbf{V}\mathbf{W}^* \text{ is non-singular} \quad \Leftrightarrow \quad I - T \text{ is non-singular.}$$

(b) Prove the **Sherman-Morrison-Woodbury formula** for the solution $\mathbf{y}$ of (3.6):

$$\mathbf{y} = \mathbf{x} + \mathbf{V}(I - T)^{-1}\mathbf{W}^*\mathbf{x} \quad \text{if } I - T \text{ is non-singular.} \tag{3.7}$$

For more insight, assume M is non-singular and decompose the solution $\mathbf{y}$ of (3.6) into its orthogonal projection onto span($\mathbf{V}$) and its orthogonal complement (cf., Exercise 3.6(a)):

$$\mathbf{y} = (\mathbf{I} - \mathbf{V}M^{-1}\mathbf{V}^*)\mathbf{y} + \mathbf{V}M^{-1}\mathbf{V}^*\mathbf{y}.$$

(c) Show that $(\mathbf{I} - \mathbf{V}M^{-1}\mathbf{V}^*)\mathbf{y} = (\mathbf{I} - \mathbf{V}M^{-1}\mathbf{V}^*)\mathbf{x}$.

(d) Show that $(\mathbf{I} - \mathbf{V}\mathbf{W}^*)\mathbf{V}M^{-1}\mathbf{V}^*\mathbf{y} = \mathbf{V}(I - T)M^{-1}\mathbf{V}^*\mathbf{y} = \mathbf{V}\mathbf{W}^*\mathbf{x} + \mathbf{V}(I - T)M^{-1}\mathbf{V}^*\mathbf{x}$. Hence, if $I - T$ is non-singular, we have that $M^{-1}\mathbf{V}^*\mathbf{y} = (I - T)^{-1}\mathbf{W}^*\mathbf{x} + M^{-1}\mathbf{V}^*\mathbf{x}$.

(e) Now, prove (3.7) using the above decomposition of $\mathbf{y}$ and the results of (c) and (d).

Exercise 3.8. **Conditioning.** For $\alpha, \beta \in \mathbb{C}$ such that $|\alpha|^2 + |\beta|^2 = 1$, consider the 2×2 matrix

$$A \equiv \begin{bmatrix} \alpha & 1 \\ \beta & 0 \end{bmatrix} = [r, s]$$

with r the first column of A and s the second column.

(a) Show that the 2-norm condition number $\mathcal{C}_2(A)$ of A is equal to

$$\mathcal{C}_2(A)=\sqrt{\frac{1+|\alpha|}{1-|\alpha|}}=\frac{1}{\tan\frac{1}{2}\angle(r,s)}.$$

(b) Consider the $n\times\ell$ matrix $\mathbf{A}=[\mathbf{r},\mathbf{v}_2,\ldots,\mathbf{v}_\ell]$ with $\mathbf{V}\equiv[\mathbf{v}_2,\ldots,\mathbf{v}_\ell]$ orthonormal and $\mathbf{r}$ normalised. Prove that $\mathcal{C}_2(\mathbf{A})$ is equal to the reciprocal of the tangent of half the angle between $\mathbf{r}$ and the space spanned by $\mathbf{V}$. (Hint let $\mathbf{s}$ be the normalised vector $\mathbf{V}\mathbf{V}^*\mathbf{r}$. Show that the condition number of $\mathbf{A}$ and of the matrix $[\mathbf{r},\mathbf{s}]$ are the same.)

(c) With $\mathbf{A}$ as in (b), put $\mathbf{r}_\ell\equiv\mathbf{r}-\mathbf{V}\mathbf{V}^*\mathbf{r}$. Suppose that $\|\mathbf{r}_\ell\|_2\ll 1$, prove that

$$\mathcal{C}_2(\mathbf{A})\approx\frac{2}{\|\mathbf{r}_\ell\|_2}.$$

(d) Let $\mathbf{A}$ be an $n\times\ell$ matrix. Show that the condition number of $\mathbf{A}$ is larger than the reciprocal of the tangent of half the angle between the ith column of $\mathbf{A}$ and the space spanned by the other columns of $\mathbf{A}$.

C Singular value decomposition

Let $\mathbf{A}$ be an $n\times k$ matrix. The decomposition

$$\mathbf{A}=\mathbf{V}\,\Sigma\,Q^* \tag{3.8}$$

is a **singular value decomposition** (SVD) of $\mathbf{A}$ if

$\mathbf{V}$ is $n\times n$ unitary, $\mathbf{Q}$ is $k\times k$ unitary,

$\Sigma=\mathrm{diag}(\sigma_1,\ldots,\sigma_k)$ is $n\times k$ diagonal, i.e., $\Sigma_{ij}=0$ if $i\neq j$ and $\sigma_j\equiv\Sigma_{jj}$

such that $\sigma_1\geq\sigma_2\geq\ldots\geq\sigma_k\geq 0$.

The σ_k are **singular values** of $\mathbf{A}$, the columns of $\mathbf{V}$ are **left singular vectors**, the columns of Q are **right singular vectors**.

The following theorem states the existence of an SVD.

Theorem 3.1 *Let $\mathbf{A}$ be an $n\times k$ matrix.*

There are an orthonormal $n\times k$ matrix $\mathbf{V}_1$, a unitary $k\times k$ matrix Q and an $k\times k$ diagonal matrix $\Sigma_1=\mathrm{diag}(\sigma_1,\ldots,\sigma_k)$ such that

$$\mathbf{A}=\mathbf{V}_1\,\Sigma_1\,Q^* \quad \textit{and} \quad \sigma_1\geq\sigma_2\geq\ldots\geq\sigma_k\geq 0. \tag{3.9}$$

$\mathbf{V}_1$ can be extended to a unitary $n\times n$ matrix $\mathbf{V}=[\mathbf{V}_1,\mathbf{V}_2]$, Σ_1 can be extended with rows of zeros to an $n\times k$ diagonal matrix $\Sigma=[\Sigma_1^{\mathrm{T}},\mathbf{0}^{\mathrm{T}}]^{\mathrm{T}}$. Then $\mathbf{A}=\mathbf{V}\Sigma Q^*$, the SVD of $\mathbf{A}$. In particular, $\sigma_1,\ldots,\sigma_k$ are the singular values of $\mathbf{A}$. The singular values are unique.

Let ℓ be the largest index for which $\sigma_\ell>0$ (i.e., $\sigma_{\ell+1}=\ldots=\sigma_k=0$). Let $\mathbf{V}_3$ and Q_3 consist of the first ℓ columns of $\mathbf{V}_1$, and Q, respectively, and let Σ_3 be the $\ell\times\ell$ diagonal matrix $\Sigma_3\equiv diag(\sigma_1,\ldots,\sigma_\ell)$. Then, $\mathbf{V}_3$ is $n\times k$ orthonormal, Q_3 is $k\times\ell$ orthonormal and $\mathbf{A}=\mathbf{V}_3\,\Sigma_3\,Q_3^*$.

The last decomposition in the theorem, $\mathbf{A}=\mathbf{V}_3\,\Sigma_3\,Q_3^*$, is the **economical form** of the SVD.

In contrast to the singular values, the singular vectors are not unique: they allow multiplication by a sign, i.e., by a $\zeta\in\mathbb{C}$ with $|\zeta|=1$. If singular values coincide, then any orthonormal basis of the space spanned by left singular vectors with the same singular value can replace these left singular vectors. Similarly for the right singular vectors.

Exercise 3.9. ***Proof of Theorem* 3.1*.***

(a) Prove that there is an $k \times k$ unitary matrix Q such that $Q^*\mathbf{A}^*\mathbf{A}Q = \Lambda$ is $k \times k$ diagonal. The diagonal entries λ_i of Λ or positive. We may assume they are in decreasing order.

(b) Let $\Sigma_1 \equiv \mathrm{diag}(\sigma_1, \ldots, \sigma_k)$ with $\sigma_j \equiv \sqrt{\lambda_j}$. Put $\mathbf{V}_1 \equiv \mathbf{A}Q\Sigma_1^{-1}$. Prove that $\mathbf{V}_1$ is orthonormal. Prove (3.9).

(c) Show that any $n \times k$ orthonormal matrix $\mathbf{V}_1$ can be extended to an $n \times n$ unitary matrix $\mathbf{V} = [\mathbf{V}_1, \mathbf{V}_2]$.

(d) Show that the SVD in economical form exists.

The economical form of the SVD allows us to characrterize the Moore–Penrose pseudo inverse (as introduced in Exercise 0.14).

Theorem 3.2 *Let* $\mathbf{A}$ *be an* $n \times k$ *matrix.*
Let $\mathbf{A} = \mathbf{V}\Sigma Q^*$ *be the SVD in ecomical form, i.e.,* Σ *is an* $\ell \times \ell$ *non-singular diagonal matrix with* $\ell \leq k$*. Then, for the Moore–Penrose pseudo-inverse* $\mathbf{A}^\dagger$ *we have that*

$$\mathbf{A}^\dagger = Q\Sigma^{-1}\mathbf{V}^*$$

If $\mathbf{A}^*\mathbf{A}$ *is non-singular (i.e.,* $n \geq \ell = k$*), then*

$$\mathbf{A}^\dagger = (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*.$$

Exercise 3.10. ***Proof of Theorem* 3.2*.***

(a) Prove the first claim of Theorem 3.2.

(b) Prove the second claim.

Exercise 3.11. **SVD and least square problems.** For an $n \times k$ matrix $\mathbf{A}$, let $\mathbf{A} = \mathbf{V}\Sigma Q^*$ be its singular value decomposition (economical form): $\mathbf{V}$ is $n \times k$ orthonormal, Q is $k \times k$ unitary, Σ is $k \times k$ diagonal. For a $\mathbf{b} \in \mathbb{C}^n$, consider the least square problems

$$\mathbf{A}x = \mathbf{b} \quad \text{and} \quad \Sigma y = \widetilde{\mathbf{b}} \equiv \mathbf{V}^*\mathbf{b}$$

with solution x and y, respectively, where the **least square solution** x of the problem $\mathbf{A}x = \mathbf{b}$ is $x \equiv \mathrm{argmin}\{\|\mathbf{b} - \mathbf{A}x'\|_2 \,\big|\, x' \in \mathbb{C}^k\}$.

(a) Give an expression that relates x and y.

(b) Give an expression for the least square, least norm solution of $\Sigma y = \widetilde{\mathbf{b}}$ in terms of $\widetilde{\mathbf{b}} = (\widetilde{b}_j)$. Is the least square, least norm solution unique?

(c) Interpret the y results for x.

Exercise 3.12. The SVD can be used to discuss the sharpness of the estimates in Theorem 1.9. Let $\mathbf{A}$ be an $n \times n$ matrix and $\mathbf{x}$ and n-vector.

(a) Let $\delta > 0$.
Show there is an $n \times n$ matrix Δ such that $\|\Delta\|_2 = \delta$ and $\|\mathbf{A}\Delta\mathbf{x}\|_2 = \|\mathbf{A}\|_2\,\|\Delta\|_2\,\|\mathbf{x}\|_2$

(b) Characterise the situation where $\|\mathbf{A}\mathbf{x}\|_2 = \|\mathbf{A}\|_2\,\|\mathbf{x}\|_2$.

Exercise 3.13. Let $\mathbf{A}$ be $n \times k$.
Put $\sigma_{\min} \equiv \min\{\|\mathbf{A}\mathbf{y}\|_2 \,\big|\, \|\mathbf{y}\|_2 = 1\}$, $\sigma_{\max} \equiv \max\{\|\mathbf{A}\mathbf{y}\|_2 \,\big|\, \|\mathbf{y}\|_2 = 1\}$.

(a) Prove that $\sigma_{\min} = 1/\|\mathbf{A}^{-1}\|_2$ in case $k = n$ and $\mathbf{A}$ is non-singular.

(b) Prove that $\sigma_{\min}$ is the smallest singular value of $\mathbf{A}$ and $\sigma_{\max}$ is the largest singular value.

(c) Let $\mathbf{A} = \mathbf{QR}$ be the (economical form) QR-decomposition. Show that $\mathbf{R}^*\mathbf{R}$ is the Cholesky decomposition of $\mathbf{A}^*\mathbf{A}$.

Exercise 3.14. **Angles between subspaces.** Let $\mathcal{V}$ and $\mathcal{W}$ be linear subspace of $\mathbb{C}^n$. Let $\mathbf{x} \in \mathbb{C}^n, \mathbf{x} \neq \mathbf{0}$. Then the **angle** $\angle(\mathbf{x}, \mathcal{V})$ between $\mathbf{x}$ and $\mathcal{V}$ is defined by

$$\angle(\mathbf{x}, \mathcal{V}) \equiv \min\{\angle(\mathbf{x}, \mathbf{v}) \Big| \mathbf{v} \in \mathcal{V}, \mathbf{v} \neq \mathbf{0}\}. \tag{3.10}$$

Let $\mathbf{V}$ be an $n \times k$ orthonormal matrix that spans $\mathcal{V}$.

(a) Let $\mathbf{x}_V$ be the orthogonal projection onto $\mathcal{V}$. Show that $\mathbf{x}_V = \mathbf{V}(\mathbf{V}^*\mathbf{x})$.
Prove that $\angle(\mathbf{x}, \mathcal{V}) = \frac{1}{2}\pi$ if $\mathbf{x}_V = \mathbf{0}$ and $\angle(\mathbf{x}, \mathcal{V}) = \angle(\mathbf{x}, \mathbf{x}_V)$ if $\mathbf{x}_V \neq \mathbf{0}$.

(b) Define

$$\phi(\mathcal{W}, \mathcal{V}) \equiv \max\{\angle(\mathbf{w}, \mathcal{V}) \Big| \mathbf{w} \in \mathcal{W}, \mathbf{w} \neq \mathbf{0}\}. \tag{3.11}$$

Is the definition $\angle(\mathcal{W}, \mathcal{V}) \equiv \phi(\mathcal{W}, \mathcal{V})$ acceptable? (Hint: Consider a 2-dimensional space $\mathcal{V}$ and a 1-dimensional space $\mathcal{W}$ and conclude that with this definition, $\angle(\mathcal{W}, \mathcal{V}) \neq \angle(\mathcal{V}, \mathcal{W})$.)

Define

$$\angle(\mathcal{W}, \mathcal{V}) \equiv \min(\phi(\mathcal{W}, \mathcal{V}), \phi(\mathcal{V}, \mathcal{W})). \tag{3.12}$$

Let $\mathbf{W}$ be an $n \times \ell$ orthonormal matrix that spans $\mathcal{W}$.

(c) Assume $\ell \leq k$.
Prove that

$$\cos(\angle(\mathcal{W}, \mathcal{V})) \text{ is the smallest singular value of } \mathbf{V}^*\mathbf{W}. \tag{3.13}$$

Prove that

$$\sin(\angle(\mathcal{W}, \mathcal{V})) = \|(\mathbf{I} - \mathbf{V}\mathbf{V}^*)\mathbf{W}\mathbf{W}^*\|_2. \tag{3.14}$$

(d) Prove that $\angle(\mathcal{W}, \mathcal{V}) \equiv \phi(\mathcal{W}, \mathcal{V})$ if $\ell \leq k$, while $\phi(\mathcal{W}, \mathcal{V}) = \frac{1}{2}\pi$ if $\ell > k$.

D Orthonormalisation

Let $\mathbf{A}$ be an $n \times k$ matrix, $n \geq k$.

$$\mathbf{A} = \mathbf{Q}\, R = \mathbf{Q}'\, R' \quad \text{where} \quad R \equiv \begin{bmatrix} R' \\ \mathbf{0} \end{bmatrix}$$

is a **QR-decomposition** or **QR-factorisation** of $\mathbf{A}$ if $\mathbf{Q}$ is $n \times n$ unitary and R is $n \times k$ upper triangular.

Let R' be the $\ell \times k$ upper block of R with $\ell \leq k$ such that R can be obtained from R' by appending R' with zeros. Let $\mathbf{Q}'$ be the left $n \times \ell$ block of $\mathbf{Q}$. Then $\mathbf{A} = \mathbf{Q}'\, R'$. This decomposition is a so-called '**economical form**' of the QR-decomposition of $\mathbf{A}$.
In Matlab `[Q,R]=qr(A)` and `[Qp,Rp]=qr(A,'0')`, respectively.

Theorem 3.3 *Let* $\mathbf{A}$ *be an* $n \times k$ *matrix.*
Then $\mathbf{A}$ *has a QR-decomposition and* span $(\mathbf{A})$ = span $(\mathbf{Q})$.
If $\mathbf{A} = \mathbf{Q}R$ *is a QR decomposition and* D *is a* $k \times k$ *diagonal such that* $|D| = I$,
then $\mathbf{A} = (\mathbf{Q}D)(D^{-1}R)$ *is also a QR decomposition. If* $\mathbf{A}$ *has full rank, then these are the only QR-decompositions: the QR-decomposition is unique up to signs.*[1]

The Gram–Schmidt process proves the existence of a QR-decomposition. Alg. 3.2 produces the (economical form of the) QR-decompostion: it relies on (a variant of) the Gram–Schmidt process (of Alg. 3.1).

Exercise 3.15. **Gram–Schmidt.** Let $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_k]$ be an $n \times k$ matrix, $n \geq k$.

(a) Prove that the columns of $\mathbf{Q}'$ form an orthonormal basis of $\mathcal{R}(\mathbf{A})$, the space spanned by the columns of $\mathbf{A}$.

[1] For each j, let $\pi(j)$ be such that $r_{\pi(j),j} \neq 0$ and $r_{ij} = 0$ for all $i > \pi(j)$. If we assume the decomposition to be ordered such that π is not decreasing, $\pi(j) \leq \pi(j+1)$, then the decomposition is unique up to signs, also in case $\mathbf{A}$ does not have full rank. If $\mathbf{A}$ has full rank, then $\pi(j) = j$ all j.

```
%% Orthogonalise:
%% Classical Gram-Schmidt
$\vec{h} = \mathbf{Q}^*\mathbf{a}$,  $\mathbf{v} = \mathbf{a} - \mathbf{Q}\,\vec{h}$
$\nu = \|\mathbf{v}\|_2$
%% Normalise:
NORMALISE
```

```
Select a $\tau > 0$
%% Orthogonalise:
%% Repeated Gram-Schmidt
$\vec{h} \equiv \mathbf{Q}^*\mathbf{a}$,  $\mathbf{v} = \mathbf{a} - \mathbf{Q}\,\vec{h}$
$\nu = \|\mathbf{v}\|_2$,  $\mu = \|\vec{h}\|_2$
while ($0 < \nu < \tau\,\mu$)
    $\vec{g} = \mathbf{Q}^*\mathbf{v}$,  $\mathbf{v} \leftarrow \mathbf{v} - \mathbf{Q}\,\vec{g}$
    $\nu = \|\mathbf{v}\|_2$,  $\mu = \|\vec{g}\|_2$,  $\vec{h} \leftarrow \vec{h} + \vec{g}$
end while
%% Normalise:
NORMALISE
```

```
%% Orthogonalise:
%% Modified Gram-Schmidt
$\mathbf{v} = \mathbf{a}$
for $i = 1, \ldots, \ell$ do
    $h_i \equiv \mathbf{q}_i^*\mathbf{v}$,  $\mathbf{v} \leftarrow \mathbf{v} - \mathbf{q}_i\,h_i$
end for
$\nu = \|\mathbf{v}\|_2$
%% Normalise:
NORMALISE
```

```
NORMALISE:
If  $\nu > 0$
    $\vec{h} \leftarrow (\vec{h};\nu)$,  $\mathbf{q} = \mathbf{v}/\nu$
elseif (EXPAND & $\ell < n$)
    $\mathbf{q} = $ ORTH($\mathbf{Q}$, rand$_n$),  $\vec{h} \leftarrow (\vec{h};0)$
else
    $\mathbf{q} = [\,]$
end if
```

ALGORITHM 3.1. $[\mathbf{q},\vec{h}] =$ ORTH$(\mathbf{Q},\mathbf{a})$: an n-vector $\mathbf{a}$ is orthogonalised against an $n \times \ell$ orthonormal matrix $\mathbf{Q} = [\mathbf{q}_1,\ldots,\mathbf{q}_\ell]$. If $\mathbf{a}$ is not in the span of $\mathbf{Q}$, then the output vector $\mathbf{q}$ is orthogonal to $\mathbf{Q}$, normalised, and $\mathbf{a}$ is in span($[\mathbf{Q},\mathbf{q}]$): $\mathbf{a} = [\mathbf{Q},\mathbf{q}]\,\vec{h}$. The output vector $\vec{h} = (h_i)$ is an $(\ell+1)$-vector, $h_{\ell+1}\mathbf{q}$ is the component of $\mathbf{a}$ orthogonal to $\mathbf{Q}$. If $\mathbf{Q} = [\,]$, then $\vec{h} = (h_1)$, $h_1 = \|\mathbf{a}\|_2$, $\mathbf{q} = \mathbf{a}/h_1$.
If $\mathbf{a}$ is in the span of $\mathbf{Q}$, then $\mathbf{q}$ is an empty vector (an $n \times 0$ vector in MATLAB's terminology) and $\vec{h}$ is an ℓ-vector: $\mathbf{a} = \mathbf{Q}\,\vec{h}$. Or, if expansion of $\mathbf{Q}$ is required (i.e., EXPAND is 'true') and is allowed by the dimensions (i.e., $\ell < n$), then $\mathbf{Q}$ is expanded with a normalised random vector that is orthogonal to $\mathbf{Q}$, i.e., a random vector is orthogonalised against $\mathbf{Q}$ using ORTH (recursively). Then, $\vec{h}$ is an $(\ell+1)$-vector with last coordinate equal to 0.
Several variants of the Gram–Schmidt process can be used for the orthogonalising (the left panels and the top right panel). The variants have different stability properties (see Lecture 3.E below). The normalisation step 'NORMALISE' (right bottom panel) is the same for all variants.
In practice, the condition '$\|\mathbf{v}\|_2 > 0$' (i.e., $\nu > 0$) in 'NORMALISE' and in repeated Gram–Schmidt has to be replaced by one that accommodates for the effect of rounding errors as $\|\mathbf{v}\|_2 > \bar{\xi}(\ell+n)\sqrt{\ell}\,\|\mathbf{a}\|_2$ with $\bar{\xi}$ the relative machine precision (see the discussion in the paragraph containing (1.30)).

Consider the variants in ALG. 3.1 of the Gram–Schmidt process as introduced in Section C in Lecture 0.

(b) Show that (in exact arithmetic), these variants (in combination with ALG. 3.2) are equivalent to the one in ALG. 0.1, i.e, they produce the same quantities. Check that modified and classical Gram–Schmidt require the same number of flops (if $\mathbf{A}$ is of full rank, then $2nk^2$, neglecting costs of order nk flop).

(c) Show that the classical variant is also in rounded arithmetic equivalent to the one in ALG. 0.1. Argue why this is not the case for the modified variant.

(d) Argue that $\frac{\nu}{\mu}$ is (an estimate for) the tangent of the angle between $\mathbf{a}_j$ and the span($\mathbf{Q}$). Note that $\|\mathbf{a}_j\|_2$ can be computed from ν and μ, that is, without additional n-dimensional operations.

The **DGKS** (Daniel–Grag–Kaufmann–Stewart) repetition **criterion**, $\nu < \tau\mu$, in repeated Gram–Schmidt requires repeated orthogonalisation if the tangent is smaller than τ (typical value for τ is 0.7). In practise, repetition is required at most once. For a more extensive

```
$\mathbf{Q}=[\,]$, $R=[\,]$, $\ell=0$
for $j=1,\ldots,k$ do
    $[\mathbf{q},\vec{h}\,]=\text{ORTH}(\mathbf{Q},\mathbf{a}_j)$
    if $\mathbf{q}\neq[\,]$
        $\mathbf{Q}\leftarrow[\mathbf{Q},\mathbf{q}]$, $R\leftarrow[R\,;\vec{0}^*_{j-1}]$, $\ell\leftarrow\ell+1$
    end if
    $R\leftarrow[R,\vec{h}\,]$
end for
```

ALGORITHM 3.2. The QR-decomposition of an $n\times k$ matrix $\mathbf{A}=[\mathbf{a}_1,\ldots,\mathbf{a}_k]$ is computed: $\mathbf{A}=\mathbf{Q}R$. $\mathbf{Q}$ is an orthonormal $n\times\ell$ matrix with $\ell\leq k$; $\ell=k$ if $\mathbf{A}$ is of full rank. R is an $\ell\times k$ upper triangular matrix.
The algorithm uses the orthogonalisation procedure ORTH from ALG. 3.1. Here, $[R\,;\,\vec{0}^*_{j-1}]$ indicates that the $\ell\times(j-1)$ matrix R is extended with a row of zeros: $\vec{0}_{j-1}$ is the $(j-1)$-vector of zeros. $[R,\vec{h}\,]$ is the matrix R extended with one column by the vector $\vec{h}$.

discussion on the stability of this Gram–Schmidt variant, see Lecture 3.E below.

(e) Argue that, in contrast to modified and classical Gram–Schmidt, repeated Gram–Schmidt is (much less) insensitive to perturbations of (the orthogonality of) $\mathbf{Q}$. (Hint: see Exercise 3.6)

(f) Show that $r_{jj}\neq 0$ is each step j ($\ell=j$) if $\mathbf{A}$ is of full rank.

(g) Show that the QR-decomposition of a full rank matrix $\mathbf{A}$ is unique up to signs (i.e., prove the last statement of Theorem 3.3).

The following exercise gives a procedure, **Householder QR**, to construct a QR-decomposition, using Householder reflections. This construction is in some sense optimal stable. The reason is that a Householder reflection is unitary no matter how inaccurate the vector $\mathbf{v}$ is.

Exercise 3.16. **Householder QR.** Let $\mathbf{A}$ be an $n\times k$ matrix.

We show that the QR decomposition can be computed by the application of k (or $k-1$ if $n=k$) appropriate Householder reflections:

$$\mathbf{A}^{(0)}\equiv\mathbf{A},\quad \mathbf{A}^{(j)}=\mathbf{H}_{\mathbf{v}_j}\mathbf{A}^{(j-1)}\quad (j=1,\ldots,k)$$

where

$$\mathbf{H}_\mathbf{v}\equiv\mathbf{I}-\frac{2}{\mathbf{v}^*\mathbf{v}}\mathbf{v}\,\mathbf{v}^*,\quad\text{i.e.,}\quad \mathbf{H}_\mathbf{v}\mathbf{x}=\mathbf{x}-\mathbf{v}\beta\quad\text{with}\quad\beta\equiv 2\frac{\mathbf{v}^*\mathbf{x}}{\mathbf{v}^*\mathbf{v}}\tag{3.15}$$

and $\mathbf{v}_j$ as to be discussed below.

(a) Show that there is an n-vector $\mathbf{v}_1$ such that $\mathbf{H}_{\mathbf{v}_1}(\mathbf{A}e_1)=\tau\mathbf{e}_1$ for some scalar τ (e_1 and $\mathbf{e}_1$ are the standard basis vectors of dimension k, n, respectively).

(b) Show there are vectors $\mathbf{v}_1,\ldots,\mathbf{v}_k$ such that the lower triangular part of the left $n\times j$ block of $\mathbf{A}^{(j+1)}$ consists of zeros ($j=1,\ldots,k$). Note that the first $j-1$ coordinates of $\mathbf{v}_j$ consists of zeros.

(c) Show that with

$$\mathbf{Q}\equiv\mathbf{H}_{\mathbf{v}_1}\cdot\ldots\cdot\mathbf{H}_{\mathbf{v}_k},\tag{3.16}$$

we have that $\mathbf{A}=\mathbf{Q}\mathbf{A}^{(k)}$ is a QR-decomposition of $\mathbf{A}$: $\mathbf{R}\equiv\mathbf{A}^{(k)}$ is upper triangular.

Exercise 3.17. **Costs of Householder QR.** Notation as in Exercise 3.16.

Note that the storage of $\mathbf{v}_1,\ldots,\mathbf{v}_k$ and of $\mathbf{R}$ requires the same memory as for storing $\mathbf{A}$ plus the storage for one additional k-vector.

(a) Suppose the n-vector $\mathbf{v}$ and the scalar $\mathbf{v}^*\mathbf{v}/2$ is available. Show that $\mathbf{H}_\mathbf{v}\mathbf{x}$ can be computed with $4n$ flop. Conclude that the computation of $\mathbf{A}^{(1)}$ (given $\mathbf{v}_1$ and $\rho_1 \equiv \mathbf{v}_1^*\mathbf{v}_1$) requires $4nk+2k$ flop.

(b) If we neglect lower order cost terms, then Householder QR requires $2nk^2 - \frac{2}{3}k^3$ flop to compute $\mathbf{v}_1, \ldots, \mathbf{v}_k$ and $\mathbf{R}$. (Actually, the costs to compute the $\mathbf{v}_j$ and ρ_j is in the neglected lower order terms.) Note that the term '$-\frac{2}{3}k^3$' is missing in the costs of Gram–Schmidt (cf. Exercise 3.15(b)). In particular, if $k \ll n$ then the costs of Gram–Schmidt and Householder QR are comparable.

(c) In many applications, we can work with the factorised form of $\mathbf{Q}$ (cf., (3.16)). If an explicit expression for $\mathbf{Q}$ is required, then this can be obtained by $\mathbf{Q} = \mathbf{Q}^{(k)}$ where

$$\mathbf{Q}^{(0)} = \mathbf{I}_k, \qquad \mathbf{Q}^{(j)} = \mathbf{H}_{\mathbf{v}_{k+1-j}}\mathbf{Q}^{(j-1)} \quad (j = 1, 2, \ldots, k),$$

where $\mathbf{I}_k$ is the left $n \times k$ block of the $n \times n$ identity matrix $\mathbf{I}$. This requires $2nk^2$ flop.

For $k \ll n$, this doubles the costs would make Householder twice as expensive as Gram–Schmidt. Moreover, the high stability of Householder QR is somewhat affected.

The following exercise contains an application of QR-decompositions.

Exercise 3.18. **Least square problems.**
Let $\mathbf{A}$ be an $n \times k$ matrix, $k \leq n$, and let $\mathbf{b} \in \mathbb{C}^n$, $\mathbf{b} \neq \mathbf{0}$.

(a) Prove Theorem 0.5

(b) Show there is an $x \in \mathbb{C}^k$ (a so-called **least square solution**) such that

$$\|\mathbf{b} - \mathbf{A}x\|_2 \leq \|\mathbf{b} - \mathbf{A}y\|_2 \quad \text{for all } y \in \mathbb{C}^k.$$

Show that x is characterised by the property

$$\mathbf{b} - \mathbf{A}x \perp \mathbf{A}y \quad (y \in \mathbb{C}^k)$$

which is equivalent to (**normal equations**)

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

Let $\mathbf{A} = \mathbf{Q}R$ be the QR decomposition (in its most economical form, i.e., $\mathbf{Q}$ is $n \times m$ and R is $m \times k$ with $m \leq k$ and R has full row rank).

(c) Show that the least square solution x satisfies

$$Rx = \mathbf{Q}^*\mathbf{b}.$$

Prove that $\mathbf{Q}\mathbf{Q}^*\mathbf{b}$ is the orthogonal projection of $\mathbf{b}$ onto the range of $\mathbf{A}$. Show that

$$\mathbf{A}\mathbf{x} = \mathbf{Q}\mathbf{Q}^*\mathbf{b}, \quad \mathbf{r} \equiv \mathbf{b} - \mathbf{A}\mathbf{x} = \mathbf{b} - \mathbf{Q}\mathbf{Q}^*\mathbf{b} :$$

the **residual r** is orthogonal complement of $\mathbf{b}$ with respect to range of $\mathbf{A}$.

(d) Is the least square solution unique?

(e) Note that $x + y$ is a least square solution if x is a least square solution and $\mathbf{A}y = 0$.
Let x be a least square solution with smallest $\|\cdot\|_2$-norm. Prove that $x \perp \{y \,\big|\, \mathbf{A}y = 0\}$ and $x = \mathbf{A}^*\mathbf{z}$ for some $\mathbf{z} \in \mathbb{C}^n$.

(f) Is the least square, least norm solution unique?

Hessenberg decomposition.
The matrix $\mathbf{R}$ in a QR-decomposition of $\mathbf{A}$, is the matrix of $\mathbf{A}$ with repect to the standard basis in domain space and the orthonormal basis $\mathbf{q}_1, \ldots, \mathbf{q}_k$ in image space. The resulting matrix $\mathbf{R}$ has a simple structure (upper triangular). For many applications where $\mathbf{A}$ is square, for instance in eigenvalue computations, it is desirable to have the same basis in domain space as in image space. Unfortunately, the matrix of $\mathbf{A}$ with respect to such a basis can not be as simple as triangular. But it can be upper Hessenberg: $\mathbf{H} = (H_{ij})$ is upper **Hessenberg** if its lower triangular entries, except for the first lower co-diagonal are zero: $H_{ij} = 0$ if $i - j > 1$.

The following theorem tells us that Hessenberg decompositions exist.

Theorem 3.4 *Let* $\mathbf{A}$ *be an* $n \times n$ *matrix. There is an* **Hessenberg decomposition**, *that is, there are* $n \times n$ *matrices* $\mathbf{V}$ *and* $\mathbf{H}$ *such that*

$$\mathbf{AV} = \mathbf{VH}, \quad \mathbf{V}^*\mathbf{V} = \mathbf{I}, \quad \textit{and} \quad \mathbf{H} \textit{ is Hessenberg.} \tag{3.17}$$

If the first column of $\mathbf{V}$ *is fixed, then the Hessenberg decomposition is unique up to signs, that is, the* $\mathbf{v}_j$ *are unique up to scaling of the* $\mathbf{v}_j$ *with factors of the form* $e^{i\phi}$ *(*± 1 *in the real case).*

The following exercise gives an explicit construction of an Hessenberg decomposition using Householder reflections. For matrices of low dimension, this construction can be very useful. For general matrices, it proves the above theorem.

Exercise 3.19. **Hessenberg decomposition.** Let $\mathbf{A}$ be an $n \times n$ matrix.

(a) Let $\mathbf{a}_1 = (a_{11}, \widetilde{\mathbf{a}}_1^{\mathrm{T}})^{\mathrm{T}}$ be the first column of $\mathbf{A}$. Construct a $\|\cdot\|_2$-normalised vector $\mathbf{v}_1$ of the form $\mathbf{v}_1 = (0, \widetilde{\mathbf{v}}_1^{\mathrm{T}})^{\mathrm{T}}$ such that the Householder reflection $\mathbf{H}_1 \equiv \mathbf{I} - 2\mathbf{v}_1\mathbf{v}_1^*$ maps $\mathbf{a}_1$ to a vector of the form $(a_{11}, *, 0, \ldots, 0)^{\mathrm{T}}$. What is the form of the first column the matrix $\mathbf{A}_2 \equiv \mathbf{H}_1\mathbf{A}\mathbf{H}_1^*$.

(b) Repeat this procedure with a normalised vector $\mathbf{v}_2$ of the form $(0, 0, *, *, \ldots, *)^{\mathrm{T}}$ such that the associated Householder reflection maps the second column of $\mathbf{A}_2$ to a vector of the form $(*, *, *, 0, \ldots, 0)^{\mathrm{T}}$. What is the form of the first two columns of $\mathbf{A}_3 \equiv \mathbf{H}_2\mathbf{A}_2\mathbf{H}_2^*$?

(c) Repeat this procedure and conclude that there is a unitary matrix $\mathbf{Q}$ $(= \mathbf{H}_{n-1} \ldots \mathbf{H}_2\mathbf{H}_1)$ such that $\mathbf{QAQ}^*$ is upper Hessenberg: there is a Hessenberg decomposition (3.17).

(d) Is the Hessenberg decomposition unique (consider 2×2 matrices)?

(e) Relate the Hessenberg decomposition to a QR-decomposition of $[\mathbf{v}_1, \mathbf{AV}]$. Select the first column of $\mathbf{V}$. Show that then the Hessenberg decomposition (5.10) is unique up to signs.

(f) Show that the computation of $\mathbf{v}_1, \ldots, \mathbf{v}_n$ and the upper Hessenberg matrix $\mathbf{QAQ}^*$ requires $\frac{8}{3}n^3$ flop (neglecting $\mathcal{O}(n^2)$ terms). Note that a vector $\mathbf{c}_n \equiv \mathbf{Qb}$ can be recursively computed as $\mathbf{c}_{j+1} = \mathbf{H}_j\mathbf{c}_j \quad (j = 1, 2, \ldots, n-1)$, where $\mathbf{c}_1 \equiv \mathbf{b}$. Show that in this way $\mathbf{Qb}$ can be computed with $2n^2$ flop. Conclude that there is no need to form the matrix $\mathbf{Q}$ explicitly.

Bi-diagonalisation. If $\mathbf{A}$ is $n \times n$ and we accept to work with a basis (say, $\mathbf{Q} = [\mathbf{q}_1, \ldots, \mathbf{q}_n]$) in image space that differs from the one in domain space (as, $\mathbf{e}_1, \ldots, \mathbf{e}_n$), then the QR-decomposition shows that we can obtain a matrix representation of $\mathbf{A}$ that is upper triangular: $\mathbf{A} = \mathbf{QR}$. With an appropriate non-trivial orthonormal basis in domain space we can simplify the matrix structure even further: the matrix can be bidiagonal (see Th. 3.5). A matrix $\mathbf{B} = (B_{ij})$ is upper **bidiagonal** if there are non-zeros only on the diagonal and the first upper co-diagonal: $B_{ij} = 0$ if $i - j < -1$ or $i - j > 0$.

Theorem 3.5 *There are* $n \times n$ *matrices* $\mathbf{V}$*,* $\mathbf{U}$ *and* $\mathbf{B}$ *such that*

$$\mathbf{V}^*\mathbf{AU} = \mathbf{B}, \quad \mathbf{V}^*\mathbf{V} = \mathbf{I}, \quad \mathbf{U}^*\mathbf{U} = \mathbf{I}, \quad \textit{and} \quad \mathbf{B} \textit{ upper bi-diagonal.} \tag{3.18}$$

In the following exercise we use a procedure similar to the one in Exercise 3.19 to perform the **bi-diagonalisation**: both $\mathbf{V}$ and $\mathbf{U}$ can be obtained as products of Householder reflections.

Exercise 3.20. ***Proof of Theorem*** 3.5. Let $\mathbf{A}$ be an $n \times n$ matrix.

(a) Let $\mathbf{a}$ be the first column of $\mathbf{A}$. Construct a $\|\cdot\|_2$-normalised vector $\mathbf{v}_1$ of the form $\mathbf{v}_1$ such that the Householder reflection $\mathbf{H}_1 \equiv \mathbf{I} - 2\mathbf{v}_1\mathbf{v}_1^*$ maps $\mathbf{a}_1$ to a vector of the form $(*, 0, \ldots, 0)^{\mathrm{T}}$.

(b) With $\mathbf{A}^{(1)} \equiv \mathbf{H}_1\mathbf{A}$, Let $\mathbf{a}^* = (a_{11}, \widetilde{\mathbf{a}}_1)$ be the first row of $\mathbf{A}^{(1)}$. Construct a $\|\cdot\|_2$-normalised vector $\mathbf{u}_1$ of the form $\mathbf{u}_1 = (0, \widetilde{\mathbf{u}}_1^{\mathrm{T}})^{\mathrm{T}}$ such that the Householder reflection $\mathbf{H}'_1 \equiv \mathbf{I} - 2\mathbf{u}_1\mathbf{u}_1^*$ when applied to the right maps $\mathbf{a}^*$ to a vector of the form $(a_{11}, *, 0, \ldots, 0)$ $(\mathbf{a}^*\mathbf{H}'_1 = (a_{11}, *, 0, \ldots, 0)$; cf., Exercise 3.19).

(c) Put $\mathbf{A}_1 \equiv \mathbf{A}^{(1)}\mathbf{H}'_1$. Repeat this procedure with a normalised vector $\mathbf{v}_2$ of the form $(0, *, *, \ldots, *)^{\mathrm{T}}$ such that the associated Householder reflection maps the second column of $\mathbf{A}_1$ to a vector of the form $(*, *, 0, \ldots, 0)^{\mathrm{T}}$.

(d) Repeat this procedure and conclude that the statement concerning the existence of the bi-diagonalisation in (3.18) of Theorem 3.5 is correct.

(e) relate the singular values of $\mathbf{A}$ to the singular values of $\mathbf{B}$.

(f) Show that the procedure in (d) can also be applied in case $\mathbf{A}$ is $(n+1)\times n$.

As we will see in Section E below, Householder QR is optimal stable: the obtained matrix $\mathbf{Q}$, in factorised form (3.16), is unitary, rounding errors when being casted in backward error form, can be bounded by machine precision times $|\mathbf{A}|$. To analyse the stability of modified Gram–Schmidt, the results in the following exercise are useful. Here, modified Gram-Schmidt is related to a Householder QR, a relation that even holds in rounded arithmetic.

Exercise 3.21. Householder QR and Modified Gram–Schmidt. Let $\mathbf{V}\equiv[\mathbf{v}_1,\ldots,\mathbf{v}_k]$ be an orthonormal $n\times k$ matrix. Let e_j be the jth standard basis k-vector. Put $\mathbf{w}_j\equiv[-e_j^{\mathrm{T}},\mathbf{v}_j^{\mathrm{T}}]^{\mathrm{T}}$.

(a) Show that $\mathbf{H}_j\equiv\mathbf{I}-\mathbf{w}_j\mathbf{w}_j^*$ is a Householder reflection.

(b) Prove that

$$\mathbf{H}_k\cdot\ldots\cdot\mathbf{H}_1=\mathbf{I}-\sum_{j=1}^{k}\mathbf{w}_j\mathbf{w}_j^*=\left[\begin{array}{cc}\mathbf{0} & \mathbf{V}^*\\ \mathbf{V} & \mathbf{I}-\mathbf{V}\mathbf{V}^*\end{array}\right].$$

Let $\mathbf{A}$ be an $n\times k$ matrix of rank k. Let 0 be the $k\times k$ matrix of zeros. Apply Householder reflection to find the QR-decomposition of $\mathbf{A}^{(1)}\equiv\left[\begin{array}{c}0\\ \mathbf{A}\end{array}\right]$: $\mathbf{A}^{(j+1)}=\mathbf{H}_j\mathbf{A}^{(j)}$ for $j=1,\ldots,k$ with $\mathbf{H}_j$ Householder reflections and $\mathbf{A}^{(k+1)}=\left[\begin{array}{c}R\\ \mathbf{0}\end{array}\right]$, where R is a $k\times k$ upper triangular matrix and $\mathbf{0}$ is the $n\times k$ matrix of zeros (see Exercise 3.16).

(c) Show that R is non-singular.

(d) Show that $\mathbf{A}=\mathbf{V}R$, where $\mathbf{V}$ and the Householder reflections are related as indicated above.

(e) Prove that, also in rounded arithmetic, this way of computing a QR-decomposition of $\mathbf{A}$ is equivalent to computing the QR-decomposition with Modified Gram–Schmidt.

E Rounding errors

We follow the conventions as introduced in Section E of Lecture 1. Below we analyse the effect of rounding errors in the algorithms for computing a QR-decomposition. For more details, we refer to [1, Ch.18].

We first discuss the accuracy in the classical and the repeated Gram–Schmidt variant of Alg. 3.1.

As observed in the paragraph containing (1.30), the rounding errors in the classical Gram–Schmidt step $\mathbf{a}-\mathbf{Q}(\mathbf{Q}^*\mathbf{a})$ are equal to $\mathbf{Q}\,\delta_1+\Delta\vec{h}+\delta_2$, where the errors δ_1, Δ, δ_2 come from computing $\mathbf{Q}^*\mathbf{a}$, $\mathbf{Q}\,\vec{h}$, and $\mathbf{a}-\mathbf{a}'$, respectively. Here $\vec{h}\equiv\mathbf{Q}^*\mathbf{a}$ and $\mathbf{a}'\equiv\mathbf{Q}\vec{h}$. We have the sharp upper bounds $\|\delta_1\|_2\leq n\,\bar{\xi}\,\|\,|\mathbf{Q}^*|\,\|_2\,\|\mathbf{a}\|_2\leq n\,\sqrt{\ell}\,\bar{\xi}\,\|\mathbf{a}\|_2$, $\|\Delta\|_2\leq\ell\,\xi\,\|\,|\mathbf{Q}|\,\|_2\leq\ell\,\sqrt{\ell}\,\bar{\xi}$ and $\|\delta_2\|\leq\bar{\xi}\,\|\mathbf{v}\|_2$. Since $\|\vec{h}\|_2\leq\|\mathbf{a}\|_2$, we obtain the upper bound $\bar{\xi}\,(n+\ell)\,\sqrt{\ell}\,\|\mathbf{a}\|_2$ on the rounding errors. In practice $\ell\ll n$. Therefore, the principal error component is $\mathbf{Q}\delta_1$, which is in the span of $\mathbf{Q}$:

$$\mathbf{v}\equiv\mathbf{a}-\mathbf{Q}(\mathbf{Q}^*\mathbf{a})\quad\Rightarrow\quad\mathbf{v}^\star=\mathbf{v}+\delta\quad\text{with}\quad\delta\approx\mathbf{Q}\,\delta_1,\quad\|\delta_1\|_2\leq n\,\sqrt{\ell}\,\bar{\xi}\,\|\mathbf{a}\|_2.$$

Here, $\mathbf{v}^\star$ is the version of $\mathbf{v}$ computed in floating point arithmetic. In particular, $\mathbf{q}^\star=\mathbf{v}^\star/\|\mathbf{v}^\star\|_2=\mathbf{q}+\delta/\|\mathbf{v}^\star\|_2$, with component $\approx\mathbf{Q}\delta_1/\|\mathbf{v}^\star\|_2$ in the span of $\mathbf{Q}$. Since $\|\delta_1\|_2/\|\mathbf{v}^\star\|_2\lesssim n\,\sqrt{\ell}\,\bar{\xi}\,(\|\mathbf{a}\|_2/\|\mathbf{v}\|_2)$, this component can be very large if $\|\mathbf{v}\|_2/\|\mathbf{a}\|_2$ $(=\sin\angle(\mathbf{Q},\mathbf{a}))$ is small. For this reason, repeated Gram–Schmidt (see Alg. 3.1) repeats the orthogonalisation if this angle is small. Since the tangent of this angle is freely available ($\approx \|\mathbf{v}\|_2/\|\vec{h}\|_2$) as a side-product of the process, this quantity is used in the repetition criterion: repeat if $\|\mathbf{v}\|_2 \leq \tau\,\|\vec{h}\|_2$. Note that, if $\tau\,\|\vec{h}\|_2 < \|\mathbf{v}\|_2 \leq \frac{1}{2}\sqrt{2}\|\mathbf{a}\|_2$, then $\|\vec{h}\|_2 \geq \frac{1}{2}\sqrt{2}\|\mathbf{a}\|_2$ and $\|\mathbf{a}\|_2/\|\mathbf{v}\|_2 \leq \sqrt{2}/\tau$. If $\|\mathbf{v}\|_2\,\frac{1}{2}\sqrt{2}\|\mathbf{a}\|_2$ then $\|\mathbf{a}\|_2/\|\mathbf{v}\|_2 \leq \sqrt{2}$.

Repetition of the orthogonalisation removes the error components in the span of $\mathbf{Q}$. However, it also introduces new error components. These might require another repeated orthogonalisation, which would make the procedure quit costly. Fortunately, the following theorem guarantees that we have to repeat at most once: two classical Gram-Schmidt steps is enough. The trick is that, we should not repeat if $\|\mathbf{v}\|_2$ is of order machine precision times $\|\mathbf{a}\|_2$. Since errors of this size can not be avoided anyway (see the discussion in the paragraph containing (1.30)), a vector $\mathbf{v}$ of this size is numerically equal to $\mathbf{0}$ and can be replaced by $\mathbf{0}$.

Consider the practical variant of repeated Gram–Schmidt of Alg. 3.1, where now $\nu > 0$, that is, $\|\mathbf{v}\|_2 > 0$, is replaced by $\nu > \bar{\xi}\,(\ell + n)\,\sqrt{\ell}\,\|\mathbf{a}\|_2$. Then, for the computed quantities $\mathbf{q}^\star$ and $\vec{h}^\star$, we have

Theorem 3.6 (Twice is enough) *Assume* $\mathbf{Q}$ *is exactly orthonormal:* $\mathbf{Q}^*\mathbf{Q} = I_\ell$.
Then, the number of repetitions is at most one. Moreover, we obtain maximal numerical accuracy, and deviation of orthogonality is bounded by (order) ξ/τ. *To be more precise,*

$$\|\mathbf{a} - [\mathbf{Q}, \mathbf{q}^\star]\vec{h}^\star\|_2 \leq \bar{\xi}\,(\ell + n)\,\sqrt{\ell}\,\|\mathbf{a}\|_2, \qquad \|\mathbf{Q}^*\mathbf{q}^\star\|_2 \leq \frac{\bar{\xi}}{\tau}\,\sqrt{2}\,(\ell + n)\,\sqrt{\ell}.$$

With, say $\tau = 10^{-3}$, we limit the number of repetitions and, as compared to machine precision, we loose only a few additional digits in orthogonality (say, 10^{-11} instead of 10^{-14}). Unfortunately, in practice, the matrix $\mathbf{Q}$ that is available will not be exactly orthonormal. It will have been constructed with Gram-Schmidt (as in Alg. 3.2) and local rounding errors will have been propagated (and amplified). To limit the effects of this, a larger value of τ (a τ of order 1) appears to be required. This situation with inexaxt $\mathbf{Q}$ can be analysed with results from Exercise 3.6. It turns out that a value as $\tau = 0.7$ works well in practice. Here, also the number of repetitions seems to be at most one (no proof). But, of course, a repetition is required more often than with small τ.

We now generalise the results in Lecture 2.E on lower triangular solves.

Round-off in Triangular systems. Let $\mathbf{L}$ be a $k \times k$ lower triangular system with non-zero diagonal entries and let $\mathbf{A}$ be a $k \times n$ matrix.
We follow the algorithm in Exercise 2.2 to solve $\mathbf{U}$ from $\mathbf{LU} = \mathbf{A}$: with $\ell_j \equiv \mathbf{L}\mathbf{e}_j - \mathbf{e}_j$ and $\mathbf{L}_j \equiv \mathbf{I} + \ell_j\mathbf{e}_j^*$, consider the sequence $(\mathbf{A}_j)$ of $n \times k$ matrices for which

$$\mathbf{A}_0 \equiv \mathbf{A}, \quad (\mathbf{I} + \ell_j\mathbf{e}_j^*)\mathbf{A}_j = \mathbf{A}_{j-1} \quad (j = 1, \ldots, k)$$

($\mathbf{A}_j$ is to be solved). Put $\mathbf{U} \equiv \mathbf{A}_k$. Let $\widehat{\mathbf{U}}$ be the matrix $\mathbf{U}$ that we actually obtain (in rounded arithmetic) by this process.

Theorem 3.7 *There is an* $k \times n$ *matrix* Δ *such that*

$$\mathbf{L}\widehat{\mathbf{U}} = \mathbf{A} + \Delta \quad \textit{with} \quad |\Delta| \leq k\mathbf{u}\,|\mathbf{L}|\,|\widehat{\mathbf{U}}|. \tag{3.19}$$

Exercise 3.22. ***Proof of Theorem*** **3.7.** Let $(\widehat{\mathbf{A}}_j)$ be the sequence of computed version of $(\mathbf{A}_j)$ obtained by solving $\mathbf{A}_j$ from $\mathbf{L}_j\mathbf{A}_j = \widehat{\mathbf{A}}_{j-1}$.

(a) Prove that $\widehat{\mathbf{U}} = \widehat{\mathbf{A}}_k$.

(b) Prove that $\widehat{\mathbf{A}}_{j-1} = \mathbf{L}_j\widehat{\mathbf{A}}_j + \Delta_j$ with $|\Delta_j| \leq \mathbf{u}\,|\mathbf{L}_j|\,|\widehat{\mathbf{A}}_j|$.

(c) Prove (3.19).

Discussion. (i) Note that the upper triangularity of $\mathbf{U}$ does not play a role.
(ii) Note also that errors on $\mathbf{L}$ need not be discussed: the exact $\mathbf{L}$ will be different from the computed one. However in the analysis we simple use the $\mathbf{L}$ that becomes available.
(iii) From this analysis, we learn that we actually compute an exact LU-decomposition from a slightly perturbed matrix. The actual size of the perturbation depends on the size of $|\mathbf{L}|$ and $|\widehat{\mathbf{U}}|$ in relation to the size of $\mathbf{A}$: see Exercise 3.1.

QR-decomposition using Householder reflections. Let $\mathbf{A}$ be an $n \times k$ matrix. We analyse the effect of rounding errors in the construction of the QR-decomposition using Householder reflections as explained in Exercise 3.16.
Let $\widehat{\mathbf{R}}$ the resulting computed $n \times k$ upper triangular matrix, and let $\mathbf{a}^{(1)}, \mathbf{a}^{(2)}, \ldots, \mathbf{a}^{(k)}$ be a sequence of n-vectors of machine numbers for the Householder reflections $\mathbf{H}_1, \ldots, \mathbf{H}_k$ as actually used in the computation. Let $\mathbf{Q} \equiv \mathbf{H}_1 \cdot \ldots \cdot \mathbf{H}_k$.

Theorem 3.8 *There is an $n \times k$ matrix Δ_A such that*

$$(\mathbf{A} + \Delta_A) = \mathbf{Q}\widehat{\mathbf{R}} \quad \textit{with} \quad \|\Delta_A\|_{\mathrm{F}} \leq k\, c\, n\, \mathbf{u}\, \|\mathbf{A}\|_{\mathrm{F}}. \tag{3.20}$$

Here c is some moderate constant.

Note that the matrix $\mathbf{Q}$ is the exact product of exact Householder reflections based on *computed* vectors of appropriate columns of the computed $\widehat{\mathbf{A}}^{(j)}$. This observation explains the remarkable stability of Householder QR decomposition: the rounding errors in the preceding steps are not reflected in a perturbed unitarity (the Householder reflection is not the one that we would have obtained in exact arithmetic, but it is nevertheless an Householder reflection and therefore a unitary map: the perturbation from unitarity is only from local errors from actually applying the Householder reflection [as explained in the introduction of Exercise 3.23 below] and not from rounding errors in preceding steps).

Exercise 3.23. ***Proof of Theorem*** 3.8**.** Let $\mathbf{a} = (a_1, \ldots, a_n)^{\mathrm{T}}$ and $\mathbf{b}$ be n-vectors of machine numbers. Define

$$\mathbf{H}\mathbf{b} \equiv \mathbf{b} - \mathbf{v}\beta \quad \text{where} \quad \mathbf{v} \equiv \mathbf{a} + \mathrm{sign}(a_1)\|\mathbf{a}\|_2\mathbf{e}_1, \quad \beta \equiv 2\frac{\mathbf{v}^*\mathbf{b}}{\mathbf{v}^*\mathbf{v}}.$$

$\mathbf{H} = \mathbf{I} - 2\frac{\mathbf{v}\mathbf{v}^*}{\mathbf{v}^*\mathbf{v}}$ is the Householder reflection that maps $\mathbf{a}$ to a multiple of the first standard basis vector $\mathbf{e}_1$. In rounded arithmetic we have that

$$(\mathbf{H}\mathbf{b})\widehat{\ } = (\mathbf{H} + \Delta)\mathbf{b} = \mathbf{H}\mathbf{b} + \delta_b \quad \text{with} \quad \|\Delta\|_{\mathrm{F}} \leq cn\mathbf{u}, \quad \|\delta_b\|_2 \leq cn\mathbf{u}\|\mathbf{b}\|_2.$$

Here, c is a moderate constant. The exact value of c may be different at different locations. In the result here, it is used that $\mathbf{v}$ has been computed, $\widehat{\mathbf{v}}$, as well as $\widehat{\mathbf{v}}^*\mathbf{b}$, $\widehat{\mathbf{v}}^*\widehat{\mathbf{v}}$, $\widehat{\beta}$ and $(\mathbf{b} - \widehat{\mathbf{v}}\widehat{\beta})\widehat{\ }$. You do not have to prove this result here, but you can used it in this exercise.

Let $\mathbf{a}^{(1)}, \mathbf{a}^{(2)}, \ldots, \mathbf{a}^{(k)}$ be a sequence of n-vectors of machine numbers and let $\mathbf{H}_j$ be the corresponding Householder reflection that maps $\mathbf{a}^{(j)}$ to $\mathbf{e}_j$.

(a) Prove that

$$(\mathbf{H}_k \cdot \ldots \cdot \mathbf{H}_1\mathbf{b})\widehat{\ } = \mathbf{H}_k \cdot \ldots \cdot \mathbf{H}_1\mathbf{b} + \delta_b \quad \text{with} \quad \|\delta_b\|_2 \leq kcn\mathbf{u}\|\mathbf{b}\|_2$$

(b) Let $\mathbf{A}^{(1)} \equiv \mathbf{A}$ be an $n \times k$ matrix. Let $\mathbf{A}^{(j+1)} \equiv \mathbf{H}_j\mathbf{A}^{(j)}$ $(j = 1, \ldots, k)$. Prove that

$$\widehat{\mathbf{A}}^{(j+1)} = \mathbf{H}_j \cdot \ldots \cdot \mathbf{H}_1\mathbf{A} + \Delta_A \quad \text{with} \quad \|\Delta_A\|_{\mathrm{F}} \leq jcn\mathbf{u}\|\mathbf{A}\|_{\mathrm{F}}.$$

(c) Put $\widehat{\mathbf{R}} \equiv \widehat{\mathbf{A}}^{(k+1)}$ and $\mathbf{Q} \equiv \mathbf{H}_1 \cdot \ldots \cdot \mathbf{H}_k$. Prove (3.20).

(d) Prove that the Householder QR factorisation leads to a computed upper triangular matrix $\widehat{\mathbf{R}}$ such that, for some $n \times k$ matrix Δ_A and some $n \times n$ unitary $\mathbf{Q}$ we have (3.20).

(e) Suppose $k = n$ and $\mathbf{A}$ is non-singular. We solve the system $\mathbf{A}\mathbf{x} = \mathbf{b}$ with Householder QR: solve $\mathbf{x}$ from $\mathbf{R}\mathbf{x} = \mathbf{H}_k \cdot \ldots \cdot \mathbf{H}_1\mathbf{b}$ (i.e., $\mathbf{x} = \mathbf{R}^{-1}\mathbf{Q}^*\mathbf{b}$). Let $\widehat{\mathbf{x}}$ be the computed solution. Assume that we can solve systems involving $\widehat{\mathbf{R}}$ exactly. Prove that

$$(\mathbf{A} + \Delta_A)\widehat{\mathbf{x}} = \mathbf{b} + \delta_b \quad \text{with} \quad \|\Delta_A\| \leq cn^2\mathbf{u}\|\mathbf{A}\|_{\mathrm{F}}, \quad \|\delta_b\|_2 \leq cn^2\mathbf{u}\|\mathbf{b}\|_2.$$

QR-decomposition with modified Gram–Schmidt.
Let $\mathbf{A}$ be an $n \times k$ matrix of full rank. Let $\widehat{\mathbf{Q}}$ be the $n \times k$ 'orthonormal' matrix and $\widehat{R}$ the $k \times k$ upper triangular matrix as computed by applying the modified Gram-Schmidt process to the columns of $\mathbf{A}$.

Theorem 3.9 *There is an $n \times k$ matrix Δ_1 such that*

$$\mathbf{A} = \widehat{\mathbf{Q}}\widehat{R} + \Delta_1 \quad \textit{with} \quad |\Delta_1| \leq k\,\mathbf{u}\,|\widehat{\mathbf{Q}}|\,|\widehat{R}|. \tag{3.21}$$

With $\mathcal{C}_2(\mathbf{A})$ the 2-norm condition number of $\mathbf{A}$, we have

$$\|\widehat{\mathbf{Q}}^*\widehat{\mathbf{Q}} - I\|_2 \leq 4k^2\,\mathbf{u}\,\mathcal{C}_2(\mathbf{A}). \tag{3.22}$$

The loss of orthogonality as expressed in (3.22) when applying modified Gram-Schmidt to the first k columns of a matrix $\mathbf{A}$ is bounded by a modest multiple of the machine precision times the conditioning of these k columns.

The result in (3.21) also holds for classical Gram-Schmidt. However, there is no result as in (3.22) for classical Gram-Schmidt: the loss of orthogonality can be much worse. It is conjectured that the loss of orthogonality in classical Gram-Schmidt can be bounded by a modest multiple of $k^2\mathbf{u}\,\mathcal{C}_2^2(\mathbf{A})$: the *square* of the conditioning of $\mathbf{A}$ is involved.

Exercise 3.24. ***Proof of Theorem*** **3.9.**

(a) Assume the computed $\widehat{R}$ matrix is available.
With $r_j^* \equiv e_j^*\widehat{R} - e_j^*$, put $\widehat{R}_j \equiv I + e_j\widehat{R}_j^*$. Put $\mathbf{A}_1 \equiv \mathbf{A}$.
Show that $\widehat{\mathbf{Q}} = \widehat{\mathbf{A}}_{k+1}$, where $\widehat{\mathbf{A}}_{j+1}$ is obtained by solving $\mathbf{A}_{j+1}$ from $\mathbf{A}_{j+1}\widehat{R}_j = \widehat{\mathbf{A}}_j$ ($\widehat{\mathbf{A}}_{j+1}$ is the computed solution), ($j = 1, \ldots, k$). (Note that orthonormalisation does not play a role here).

(b) Use Exercise 3.22 to show (3.21).

(c) When using modified Gram-Schmidt, then Exercise 3.23 can be used to prove that there is an orthonormal matrix $\mathbf{Q}$, such that for some Δ_2 we have that

$$\mathbf{A} = \mathbf{Q}\widehat{R} + \Delta_2 \quad \text{with} \quad |\Delta_1| \leq k\,\mathbf{u}\,|\mathbf{Q}|\,|\widehat{R}|$$

(You do not have to prove that here). Note that $\widehat{R}$ is the upper triangular matrix as computed by the Modified Gram-Schmidt process, the same as in the part (b). Prove that

$$\|(\mathbf{Q} - \widehat{\mathbf{Q}})\widehat{R}\|_2 \leq k\,\mathbf{u}(\|\mathbf{Q}\|_{\mathrm{F}} + \|\widehat{\mathbf{Q}}\|_{\mathrm{F}})\|\widehat{R}\|_{\mathrm{F}}.$$

Argue that this (roughly) implies that (why roughly?)

$$\|\mathbf{Q} - \widehat{\mathbf{Q}}\|_2 \leq 2k^2\,\mathbf{u}\,\mathcal{C}_2(\mathbf{A}).$$

(d) Prove (3.22).

References

[1] Nicholas J. Higham, *Accuracy and stability of numerical algorithms*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1996. MR 97a:65047