March 26, 2018

Lecture 9 – Least Square Problems

Consider the least square problem $\mathbf{A}\mathbf{x} = \mathbf{b} \equiv (\beta_1, \ldots, \beta_n)^{\mathrm{T}}$, where $\mathbf{A}$ is an $n \times m$ matrix.

The situation where $\mathbf{b} \in \mathcal{R}(\mathbf{A})$ is of particular interest: often there is a vectors $\mathbf{x} = \mathbf{x}_{\text{ideal}}$ that forms $\mathbf{b}$ (via $\mathbf{A}$), $\mathbf{b} = \mathbf{b}_{\text{ideal}} \equiv \mathbf{A}\mathbf{x}_{\text{ideal}}$. $\mathbf{x}_{\text{ideal}}$ is unknown while $\mathbf{b}_{\text{ideal}}$ is available (from, e.g., measurements). In practise, the 'ideal' $\mathbf{b}_{\text{ideal}}$ will be polluted by errors (from measurements, computations, noise, etc.) and the measured $\mathbf{b} = \mathbf{b}_{\text{ideal}} + \delta$ may not be in $\mathcal{R}(\mathbf{A})$.
The system $\mathbf{A}\mathbf{x} = \mathbf{b}$ is said to be **compatible** or **consistent** if $\mathbf{b} \in \mathcal{R}(\mathbf{A})$.

Notation 9.1 If $\mathbf{F}$ is a real-valued function on $\mathbb{C}^m$ and $\mathcal{V}$ is a subset of $\mathbb{C}^m$ then

$$\{\mathbf{F}(\mathbf{x}) \,\big|\, \mathbf{x} \in \mathcal{V}\} = \min$$

means $\mathbf{x} = \operatorname{argmin}\{\mathbf{F}(\widetilde{\mathbf{x}}) \,\big|\, \widetilde{\mathbf{x}} \in \mathcal{V}\}$. If $\mathcal{V} = \mathbb{C}^m$, then we simply put $\mathbf{F}(\mathbf{x}) = \min$.

Exercise 9.1. Prove the following equivalences.
(a) **Minimal residual** (**least squares**): $\|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 = \min \quad \Leftrightarrow$

$$\mathbf{A}^*\mathbf{A}\mathbf{x} = \mathbf{A}^*\mathbf{b} \quad \Leftrightarrow \quad \begin{bmatrix} \mathbf{I} & \mathbf{A} \\ \mathbf{A}^* & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{r} \\ \mathbf{x} \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ \mathbf{0} \end{bmatrix}.$$

If $m \leq n$, then the minimal residual solution is unique iff $\mathbf{A}$ has full rank.
(b) **Minimal norm**: $\{\|\mathbf{x}\|_2 \,\big|\, \mathbf{A}\mathbf{x} = \mathbf{b}\} = \min \quad \Leftrightarrow$

$$\mathbf{A}\mathbf{A}^*\mathbf{y} = \mathbf{b} \;\&\; \mathbf{x} = \mathbf{A}^*\mathbf{y} \quad \Leftrightarrow \quad \begin{bmatrix} \mathbf{0} & \mathbf{A} \\ \mathbf{A}^* & -\mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \mathbf{x} \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ \mathbf{0} \end{bmatrix}.$$

Note that the minimal norm solution exists iff $\mathbf{b} \in \mathcal{R}(\mathbf{A})$.
If $n \leq m$ then the minimal norm solution exists for all $\mathbf{b} \in \mathbb{C}^n$ iff $\mathbf{A}$ has full rank.
(c) **Minimal norm minimal residual** (or **least square minimal norm** [LSMN]):
$\mathbf{x} = \operatorname{argmin}\{\|\widetilde{\mathbf{x}}\|_2 \,\big|\, \|\mathbf{A}\widetilde{\mathbf{x}} - \mathbf{b}\|_2 = \min\}$

$$\Leftrightarrow \quad \mathbf{x} = \mathbf{A}^\dagger\mathbf{b} \quad \Leftrightarrow \quad \mathbf{A}^*\mathbf{A}\mathbf{x} = \mathbf{A}^*\mathbf{b} \quad \& \quad \mathbf{x} = \mathbf{A}^*\mathbf{y} \;\text{ for some } \mathbf{y}.$$

Here, $\mathbf{A}^\dagger$ is the Moore–Penrose pseudo-inverse of $\mathbf{A}$. In particular, the minimal norm minimal residual solution exists.
(d) **Damped least squares**: $\|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 + \tau^2\|\mathbf{x}\|_2^2 = \min \quad \Leftrightarrow$

$$\left\| \begin{bmatrix} \mathbf{b} \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{A} \\ \tau\,\mathbf{I} \end{bmatrix} \mathbf{x} \right\|_2 = \min \quad \Leftrightarrow \quad \begin{bmatrix} \mathbf{I} & \mathbf{A} \\ \mathbf{A}^* & -\tau^2\,\mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{r} \\ \mathbf{x} \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ \mathbf{0} \end{bmatrix}.$$

Exercise 9.2. Consider the damped least squares problem

$$\begin{bmatrix} \mathbf{A} \\ \tau\,\mathbf{I} \end{bmatrix} \mathbf{x} = \begin{bmatrix} \mathbf{b} \\ \mathbf{0} \end{bmatrix}.$$

(a) How are the singular values of the damped matrix related to the ones of the original matrix?
(b) Explain why the damped least squares problem is less sensitive to noise.
(c) The error in $\mathbf{x}$ in the damped least square problem (with noise on $\mathbf{b}$) has two components, one from the noise on $\mathbf{b}$ and an **approximation error** from τ (that is, with exact $\mathbf{b}$, the

difference between two LSMN solutions, one with $\tau = 0$ and the other with $\tau > 0$).
Why two components? Analyse these two components.

Exercise 9.3. **De la Garza's method.** Gauss–Seidel for the normal equations $\mathbf{A}^*\mathbf{A}\mathbf{x} = \mathbf{A}^*\mathbf{b}$ defines updates

$$\mathbf{x}_{\text{new}} = \mathbf{x}_{\text{old}} + \alpha \mathbf{e}_j,$$

where $\mathbf{e}_j$ is the jth basis vector, α is selected to minimise the 2-norm of the residual $\mathbf{r}_{\text{new}} \equiv \mathbf{b} - \mathbf{A}\mathbf{x}_{\text{new}}$ of $\mathbf{x}_{\text{new}}$.

(a) Show that, with $\mathbf{a}_j \equiv \mathbf{A}\mathbf{e}_j$, α is such that $\mathbf{a}_j^*\mathbf{r}_{\text{new}} = 0$. Conclude that,

$$\alpha = \frac{\mu}{\|\mathbf{a}_j\|_2}, \quad \text{where} \quad \mathbf{r}_{\text{old}} \equiv \mathbf{b} - \mathbf{A}\mathbf{x}_{\text{old}} \quad \text{and} \quad \mu \equiv \frac{\mathbf{a}_j^*\mathbf{r}_{\text{old}}}{\|\mathbf{a}_j\|_2}.$$

(b) The residual can be updated as $\mathbf{r}_{\text{new}} = \mathbf{r}_{\text{old}} - \alpha\mathbf{A}\mathbf{e}_j$. Write down the algorithm that arises by repeatedly cycling through all j $(j = 1, \ldots, m)$. Note that, with an appropriate implementation, the scalings by $\|\mathbf{a}_j\|_2^2$ have to be applied only once (by using the scalings as a diagonal right-preconditioner of the system $\mathbf{A}\mathbf{x} = \mathbf{b}$. How?).

This approach is called **De la Garza's method**. Gauss–Jacobi for the normal equations leads to a technique called **Simultaneous Iterative Reconstruction Technique** (**SIRT**).

(c) Show that the computational costs per cycle are equal to the costs of an update of the form $\mathbf{x} + \mathbf{A}^*\mathbf{c}$ plus an update of the form $\mathbf{r} - \mathbf{A}\mathbf{u}$.

(d) Show that, De la Garza's method is residual-norm reducing (or, to be more precise, residual norm non-increasing)

$$\|\mathbf{r}_{\text{new}}\|_2^2 = \|\mathbf{r}_{\text{old}}\|_2^2 - \|\mathbf{a}_j\|_2^2 |\alpha|^2 = \|\mathbf{r}_{\text{old}}\|_2^2 - |\mu|^2.$$

Does this last estimate implies convergence?

(e) Prove that De La Garza's method converges to the minimal residual solution (if this solution exists and is unique. Hint: Theorem 4.4).

Exercise 9.4. **Kaczmarz' method.** We denote the ith row of $\mathbf{A}$ by $\mathbf{a}_i^* \equiv \mathbf{e}_i^*\mathbf{A}$. Consider the hyperplane

$$\mathcal{L}_i \equiv \{\mathbf{y} \in \mathbb{C}^n \,\big|\, \mathbf{a}_i^*\mathbf{y} = \beta_i\} = \mathbf{y}_i + \{\mathbf{z} \,\big|\, \mathbf{a}_i^*\mathbf{z} = 0\},$$

where $\mathbf{y}_i$ is such that $\mathbf{a}_i^*\mathbf{y}_i = \beta_i$.

(a) Show that the exact solution (if it exists) is in all hyperplanes: $\mathbf{x} \in \mathcal{L}_i$ all i.

(b) Prove that $\mathbf{a}_i$ is orthogonal to the ith hyperplane $\mathcal{L}_i$.

To update an approximate solution $\mathbf{x}_{\text{old}}$ such that the updated approximation $\mathbf{x}_{\text{new}}$ satisfies the ith equation $\mathbf{a}_i^*\mathbf{x} = \beta_i$, updating in the direction $\mathbf{a}_i$ seems to be the best. Why? This suggests the following update procedure $\mathbf{x}_{\text{new}} = \mathbf{x}_{\text{old}} + \alpha\,\mathbf{a}_i$.

(c) For given $\mathbf{x}_{\text{old}}$, give an expression for the α such that $\mathbf{x}_{\text{new}}$ is in the ith hyperplane.

(d) Write down the algorithm that arises by repeatedly cycling through all rows $(i = 1, \ldots, n)$. This approach is called **Kaczmarz' method** or **Algebraic Reconstruction Technique** (**ART**). Note that, a simultaneous updating of the residual is relatively expensive. As in De la Garza's method, the scalings that are required (by $\|\mathbf{e}_i^*\mathbf{A}\|_2^2$) have to be applied only once (by using the scalings as a diagonal left-preconditioner).

(e) Show that, when omitting the update of $\mathbf{r}$, the computational costs per cycle are equal to the costs of an update of the form $\mathbf{x} + \mathbf{A}^*\mathbf{c}$ plus an update of the form $\mathbf{r} - \mathbf{A}\mathbf{u}$, as with De la Garza's method.

Suppose $\mathbf{b} \in \mathcal{R}(\mathbf{A})$. Let $\mathbf{x}$ be the minimal norm solution.

(f) Prove that Kaczmarz' method is error-norm reducing (non-increasing):

$$\|\mathbf{x} - \mathbf{x}_{\text{new}}\|_2 = \|\mathbf{x} - \mathbf{x}_{\text{old}}\|_2 - \tfrac{|\beta_i - \mathbf{a}_i^*\mathbf{x}_{\text{old}}|^2}{\|\mathbf{a}_i\|_2^2}. \tag{9.1}$$

Can convergence be concluded from this estimate (9.1)?

(g) Give a geometric proof of convergence in case $m = n = 2$ (cf., Exercise 4.5(c)).

(h) Prove that the algorithm is Gauss–Seidel for $\mathbf{A}\mathbf{A}^*\mathbf{y} = \mathbf{b}$ & $\mathbf{x} = \mathbf{A}^*\mathbf{y}$.

(i) Prove that the algorithm, with initial approximate solution $\mathbf{x}_0 = \mathbf{0}$, converges to the minimal norm solution. (Hint: Note that $\mathbf{A}\mathbf{A}^*$ maps $\mathcal{R}(\mathbf{A})$ to $\mathcal{R}(\mathbf{A})$ (Why?). Restricted to this space, $\mathbf{A}\mathbf{A}^*$ is positive definite (why?). Now use, Theorem 4.4. Why is it convenient to have a trivial initial approximate solution?)

Gauss–Seidel has been modified to SOR. A similar modification (i.e., a relaxation parameter) can be included in De La Garza's method and ART. Acceleration techniques for Gauss–Seidel, as Chebyshev iteration and Krylov subspace acceleration, can also be included, which reduces the number of iteration steps. Nevertheless, the simpler iteration methods as ART (and SIRT) have advantages in some application. Though the convergence is slower, less storage is required. For instance, in ART, there is even no need to store the matrix: whenever a 'row' $\mathbf{a}_i^*\mathbf{x} = \beta_i$ becomes available, it can be used to update the approximate solution $\mathbf{x}$. Then, it can be discarded (assuming new rows will become available), while Krylov methods need the same matrix in every step.

Exercise 9.5. **CGLS and Graig's method.** In Exercise 8.1 and Exercise 8.2, we derived a CG variant for the (minimal norm) equations $\mathbf{A}\mathbf{A}^*\mathbf{y} = \mathbf{b}$ & $\mathbf{x} = \mathbf{A}^*\mathbf{y}$ (Graig's method) and one for the (minimal residual) normal equations $\mathbf{A}^*\mathbf{A}\mathbf{x} = \mathbf{A}^*\mathbf{b}$ (CGLS), assuming $\mathbf{A}$ was square (non-singular). See also ALG. 8.1. Now, consider the general case where $\mathbf{A}$ is not (necessarily) square.

(a) Show that these variants can also be applied in this general case (if $\mathbf{b} \in \mathcal{R}(\mathbf{A})$ for Graig's method).

(b) Show that the minimisation statements in Property 8.1 also hold in this general situation (minimal errors with Graig's method and minimal residual with CGLS).

Golub-Kahan bi-diagonalisation. Let $\mathbf{A}$ be an $n \times n$ matrix. Let $\mathbf{b}$ be an n-vector.
Arnoldi's decomposition $\mathbf{A}\mathbf{V}_k = \mathbf{V}_{k+1}\underline{H}_k$ can be viewed as an iterative process to compute the Hessenberg decomposition $\mathbf{A}\mathbf{V} = \mathbf{H}\mathbf{V}$ with $\mathbf{V}$ unitary and $\mathbf{H}$ Hessenberg. The existence of this decomposition has been proved and discussed in Exercise 3.19. In Exercise 3.20, the decomposition $\mathbf{A}\mathbf{U} = \mathbf{V}\mathbf{B}$ has been proved, with $\mathbf{V}$ and $\mathbf{U}$ unitary and $\mathbf{B}$ upper bi-diagonal. The iterative process to compute this decomposition for the $(n+1) \times n$ matrix $[\mathbf{b}, \mathbf{A}]$ is the **Golub-Kahan bi-diagonalisation**. This procedure forms the basis of the LSQR method.

Exercise 9.6. **The Golub–Kahan bi-diagonalisation.** Let $\mathbf{A}$ be an $n \times (m-1)$ matrix.

(a) Show that the strategy in Exercise 3.19 applied to the $n \times m$ matrix $[\mathbf{b}, \mathbf{A}]$ leads to an $n \times n$ unitary matrix $\mathbf{V}$, an $m \times m$ unitary matrix $\mathbf{U}$, and a lower bi-diagonal $n \times m$ matrix $\mathbf{B}$ such that,

$$\rho\mathbf{v}_1 \equiv \rho\mathbf{V}\mathbf{e}_1 = \mathbf{b}, \qquad \mathbf{A}\mathbf{U} = \mathbf{V}\mathbf{B},$$

for some $\rho \in \mathbb{R}$, $|\rho| = \|\mathbf{b}\|_2$. Note that $\mathbf{A}^*\mathbf{V} = \mathbf{U}\mathbf{B}^*$ and that $\mathbf{B}^*$ is upper bi-diagonal. Conclude that $\mathbf{u}_1$ is a scalar multiple of $\mathbf{A}^*\mathbf{v}_1$.

Consider the following iterative process

$$\begin{aligned}
&\beta_1 \equiv \|\mathbf{b}\|_2, \quad \mathbf{v}_1 \equiv \mathbf{b}/\beta_1, \qquad \widetilde{\mathbf{u}}_1 = \mathbf{A}^*\mathbf{v}_1, \quad \alpha_1 \equiv \|\widetilde{\mathbf{u}}_1\|_2, \quad \mathbf{u}_1 = \widetilde{\mathbf{u}}_1/\alpha_1 \\
&\widetilde{\mathbf{v}}_k = \mathbf{A}\mathbf{u}_{k-1} - \mathbf{v}_{k-1}\alpha_{k-1}, \quad \beta_k \equiv \|\widetilde{\mathbf{v}}_k\|_2, \quad \mathbf{v}_k = \widetilde{\mathbf{v}}_k/\beta_k \qquad (k = 2, 3, \ldots) \\
&\widetilde{\mathbf{u}}_k = \mathbf{A}^*\mathbf{v}_k - \mathbf{u}_{k-1}\beta_k, \quad \alpha_k \equiv \|\widetilde{\mathbf{u}}_k\|_2, \quad \mathbf{u}_k = \widetilde{\mathbf{u}}_k/\alpha_k \qquad (k = 2, 3, \ldots)
\end{aligned}$$

(b) Show that this process leads to a decomposition as in (a):
the matrices $[\mathbf{v}_1, \ldots, \mathbf{v}_k]$ and $[\mathbf{u}_1, \ldots, \mathbf{u}_k]$ are orthonormal,

$$\mathbf{A}\mathbf{U}_{k-1} = \mathbf{V}_k\,\underline{B}_{k-1}, \qquad \mathbf{A}^*\mathbf{V}_k = \mathbf{U}_k\,B_k^*$$

with the $\alpha_1, \alpha_2, \ldots$ on the diagonal of $\underline{B}_k$ and $\beta_2, \beta_3, \ldots$ on the lower co-diagonal.

(c) Relate the Lanczos process for $\mathbf{A}\mathbf{A}^*$ and for $\mathbf{A}^*\mathbf{A}$ to this bi-diagonalisation process.

Exercise 9.7. **Krylov type of methods for structured problems.** For an $n \times k$ matrix $\mathbf{A}$ and scalars $\sigma, \mu \in \mathbb{C}$, consider the problem (cf., Exercise 9.1)

$$\begin{bmatrix} \mu\mathbf{I} & \mathbf{A} \\ \mathbf{A}^* & \sigma\mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{r} \\ \mathbf{x} \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ \mathbf{0} \end{bmatrix}$$

Consider the Golub–Kahan bi-diagonalisation in Exercise 9.6(b).

(a) Show that

$$\begin{bmatrix} \mu\mathbf{I} & \mathbf{A} \\ \mathbf{A}^* & \sigma\mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{V}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{U}_{k-1} \end{bmatrix} = \begin{bmatrix} \mathbf{V}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{U}_k \end{bmatrix} \begin{bmatrix} \mu I_k & \underline{B}_{k-1} \\ B_k^* & \sigma \underline{I}_{k-1} \end{bmatrix}, \quad \begin{bmatrix} \mu\mathbf{I} & \mathbf{A} \\ \mathbf{A}^* & \sigma\mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{V}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{U}_k \end{bmatrix} = \begin{bmatrix} \mathbf{V}_{k+1} & \mathbf{0} \\ \mathbf{0} & \mathbf{U}_k \end{bmatrix} \begin{bmatrix} \mu \underline{I}_k & \underline{B}_k \\ B_k^* & \sigma I_k \end{bmatrix}$$

(b) With

$$\widetilde{\mathbf{b}} \equiv \begin{bmatrix} \mathbf{b} \\ \mathbf{0} \end{bmatrix}, \quad \mathbb{A} \equiv \begin{bmatrix} \mu\mathbf{I} & \mathbf{A} \\ \mathbf{A}^* & \sigma\mathbf{I} \end{bmatrix}, \quad \mathbb{U}_{2k-1} \equiv \begin{bmatrix} \mathbf{V}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{U}_{k-1} \end{bmatrix}, \quad \text{and} \quad \mathbb{U}_{2k} \equiv \begin{bmatrix} \mathbf{V}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{U}_k \end{bmatrix},$$

the matrices $\mathbb{U}_j$ are orthonormal. Moreover, $\mathrm{span}(\mathbb{U}_j) = \mathcal{K}_j(\mathbb{A}, \widetilde{\mathbf{b}})$.

(c) Since $\mathbb{A}$ is Hermitian, the Lanczos process, started with $\widetilde{\mathbf{b}}$ can also be applied. Relate the resulting Lanczos vectors and tri-diagonal matrix to the quantities in part (b).

(d) Prove that

$$\begin{bmatrix} \mathbf{V}_k^* & \mathbf{0} \\ \mathbf{0} & \mathbf{U}_k^* \end{bmatrix} \begin{bmatrix} \mu\mathbf{I} & \mathbf{A}^* \\ \mathbf{A} & \sigma\mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{V}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{U}_k \end{bmatrix} = \begin{bmatrix} \mu I_k & B_k \\ B_k^* & \sigma I_k \end{bmatrix}, \quad \begin{bmatrix} \mathbf{V}_{k+1}^* & \mathbf{0} \\ \mathbf{0} & \mathbf{U}_k^* \end{bmatrix} \begin{bmatrix} \mu\mathbf{I} & \mathbf{A} \\ \mathbf{A}^* & \sigma\mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{V}_{k+1} & \mathbf{0} \\ \mathbf{0} & \mathbf{U}_k \end{bmatrix} = \begin{bmatrix} \mu I_{k+1} & \underline{B}_k \\ (\underline{B}_k)^* & \sigma I_k \end{bmatrix}$$

Note that the initial vector $\widetilde{\mathbf{b}} = (\mathbf{b}^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$ of the Krylov subspace generated by this structured block matrix $\mathbb{A}$ has only one non-zero block block. Here, we learnt that this fact leads to significant advantageous, when forming the Arnoldi (or Lanczos) decomposition: it allows

- to reduce the number of AXPYs and DOTs by a factor two (the zero blocks in the basis vectors do not have to be computed),
- to reduce the storage requirements by a factor two (the zero blocks in the basis vectors need not to be stored),
- to represent the projected matrices with the same structure as the original matrix $\mathbb{A}$.

These type of advantageous can also be exploited for some other problems with structured matrices (companion type of problems involving one matrix $\mathbf{A}$, Hamiltonian systems, etc.).

Exercise 9.8. **LSQR. Least square QR** (**LSQR**) exploits the first relation in Exercise 9.7(a).
Following the FOM approach, let $y \in \mathbb{C}^k$ and $s \in \mathbb{C}^{k+1}$ be such that

$$\begin{bmatrix} I_{k+1} & \underline{B}_k \\ (\underline{B}_k)^* & 0 \end{bmatrix} \begin{bmatrix} s \\ y \end{bmatrix} = \|\mathbf{b}\|_2 \begin{bmatrix} e_1 \\ 0 \end{bmatrix}$$

(a) Prove that y is the least square solution of the system $\underline{B}_k y = \|\mathbf{b}\|_2 \, e_1$ with residual s.

(b) Let $\mathbf{x}_k \equiv \mathbf{U}_k y$. Show that the residual $\mathbf{r}_k \equiv \mathbf{b} - \mathbf{A}\mathbf{x}_k$ equals $\mathbf{r}_k = \mathbf{V}_{k+1} s$ and $\|\mathbf{A}^*\mathbf{r}_k\|_2 = |h_{k+1,k+1} s_{k+1}|$, where h_{ij} is the (i,j)-entry of B_{k+1} and $s_{k+1} = e_{k+1}^* s$: the size of the residual $\mathbf{A}^*\mathbf{r}_k$ of the normal equations is available in low dimensional space.

(c) Note that $\mathbf{x}_k$ is of the form $\mathbf{A}^*\mathbf{y}_k$ for some n-vector $\mathbf{y}_k$. Prove that $(\mathbf{x}_k)$ converges to the minimal norm minimal residual solution if $(\|\mathbf{A}^*\mathbf{r}_k\|_2)$ converges to 0.

(d) Show that the QR-decomposition $\underline{B}_k = \underline{Q}_k R_k$ can be computed with Givens rotations and that R_k is upper triangular, bi-diagonal. This QR decomposition is used in LSQR to solve the least square system $\underline{B}_k y = \|\mathbf{b}\|_2 \, e_1$ (which explains the naming of the method): $y = R_k^{-1} z_k$ with $z_k \equiv (\underline{Q}_k^*(\|\mathbf{b}\|_2 \, e_1))$. As in MINRES, $\mathbf{x}_k$ is computed in LSQR as $\mathbf{x}_k = (\mathbf{U}_k R_k^{-1}) z_k$, rather than $\mathbf{x}_k = \mathbf{U}_k (R_k^{-1} z_k)$.

(e) Show that $\mathbf{W}_k \equiv \mathbf{U}_k R_k^{-1}$ can be updated with a two term (vector) recurrence relation and z_k by a two term (scalar) recurrence relation. Derive the LSQR algorithm.

(f) Discuss the consequences of the FOM approach for the second relation in Exercise 9.7(a).

Multishift methods. Regularisation leads to problems of the form

$$(\mathbf{A}^*\mathbf{A} + \tau^2 \mathbf{I}) \, \mathbf{x}_\tau = \mathbf{A}^*\mathbf{b}, \tag{9.2}$$

where τ is a regularisation parameter: the solution $\mathbf{x}_\tau$ of (9.2) solves

$$\|\mathbf{b} - \mathbf{A}\mathbf{x}_\tau\|_2^2 + \tau^2 \|\mathbf{x}_\tau\|_2^2 = \min .$$

To find an appropriate regularisation parameter, it is convenient to have solutions $\mathbf{x}_\tau$ for a range of values for τ. Often an appropriate τ is selected upon inspection of the so-called **L-curve**, that is the curve $\{(\|\mathbf{b} - \mathbf{A}\mathbf{x}_\tau\|_2, \|\mathbf{x}_\tau\|_2) \,\big|\, \tau \in [0, \infty)\}$. So-called **multishift methods** are variants of standard methods that are efficient in finding $\mathbf{x}_\tau$ for several τ, once a solution has been computed for one τ, typically for $\tau = 0$. Note that $\mathbf{A}^*\mathbf{A}$ is positive definite. Hence, for the problems that we discuss in this lecture, multishift methods for Hermitian matrices are the most interesting ones. For simplicity and to illustrate the main idea, we consider GMRES first.

Exercise 9.9. **Multishift GMRES.** Let $\mathbf{A}$ be an $n \times n$ matrix. Let $\mathbf{b}$ be an n-vector. Here, we discuss methods for solving

$$(\mathbf{A} + \sigma \mathbf{I})\mathbf{x}_\sigma = \mathbf{b} \tag{9.3}$$

for $\mathbf{x}_\sigma$ for several choices for $\sigma \in \mathbb{C}$.

(a) Prove that $\mathcal{K}_k(\mathbf{A}, \mathbf{b}) = \mathcal{K}_k(\mathbf{A} + \sigma\mathbf{I}, \mathbf{b})$ for all $k \in \mathbb{N}$ and all $\sigma \in \mathbb{C}$.

The fact that Krylov subspace are invariant under shifts of the matrix, forms the basis for multishift methods. The idea is the construct the Krylov subspace for one σ, say $\sigma = 0$, and to use it for all shifted problems.

Suppose $\mathbf{A}\mathbf{V}_k = \mathbf{V}_{k+1}\underline{H}_k$ is a Arnoldi decomposition with $\mathbf{v}_0 = \mathbf{b}/\|\mathbf{b}\|_2$: $\mathbf{V}_k = [\mathbf{v}_0, \ldots, \mathbf{v}_{k-1}]$.

(b) Prove that $(\mathbf{A} + \sigma\mathbf{I})\mathbf{V}_k = \mathbf{V}_{k+1}(\underline{H}_k + \sigma \underline{I}_k)$, where $\underline{I}_k$ is the $(k+1) \times k$ identity.

(c) Use the fact that the vector $\vec{\gamma}_k$ with first coordinate 1 and such that $\vec{\gamma}_k^* \underline{H}_k = \vec{0}^*$ determines the norm of the GMRES residual (see Exercise 6.1) to design a **multishift GMRES** algorithm for solving (9.3) for several σ.

From Exercise 9.9, we know it is easy to design a multishift variant of GMRES. For short recurrence methods, it is bit more complicated, since for these methods, we are not willing to to store a (large) set of basis vectors for the Krylov subspace. Nevertheless, the idea that we exploited for GMRES also extends to many short recurrence methods as CG (multishift CG), Bi-CGSTAB, etc.. As an example, we consider multishift Lanczos (cf., Exercise 7.6).

Exercise 9.10. **Multishift Lanczos method.** Consider problem (9.3) but now with $\mathbf{A}$ Hermitian.

With $\mathbf{v}_{-1} \equiv \mathbf{0}$ and $\mathbf{v}_0 \equiv \mathbf{b}/\|\mathbf{b}\|_2$, consider the Lanczos relation

$$\mathbf{A}\mathbf{v}_k = \beta_{k+1}\mathbf{v}_{k+1} + \alpha_k \mathbf{v}_k + \beta_k \mathbf{v}_{k-1} \quad (\mathbf{v}_{k+1} \perp \mathbf{v}_k, \mathbf{v}_{k-1}) \qquad (k = 0, 1, 2, \ldots)$$

with scalars β_k, β_{k+1} and α_k such that the orthogonality restrictions as indicated above hold.

$\mathbf{x}_\tau = \mathbf{0}\,,\ \mathbf{r} = \mathbf{b}$
$\mathbf{u} = \mathbf{0}\,,\ \rho = 1\,,$
$\mathbf{u}_\tau = \mathbf{0}\,,\ t_\tau = 0\,,\ \mu_\tau = 1\,,\ \gamma_\tau = 1$
`while` $\|\mathbf{r}\|_2 > tol$ `do`
  $\mathbf{s} = \mathbf{A}^*\mathbf{r}$
  $\rho' = \rho\,,\ \rho = \mathbf{s}^*\mathbf{s}\,,\ \beta = -\rho/\rho'$
  $\mathbf{u} \leftarrow \mathbf{s} - \beta\,\mathbf{u}\,,\ \mathbf{c} = \mathbf{A}\mathbf{u}$
  $\sigma = \mathbf{c}^*\mathbf{c}\,,\ \alpha = \rho/\sigma$
  $\mathbf{r} \leftarrow \mathbf{r} - \alpha\,\mathbf{c}$
  $t_\tau \leftarrow \tau^2 - (\beta/\mu_\tau)t_\tau\,,\ \mathbf{u}_\tau \leftarrow \mathbf{s} - (\beta/\mu_\tau)\,\mathbf{u}_\tau$
  $\mu_\tau = 1 + \alpha\,t_\tau\,,\ \gamma_\tau \leftarrow \gamma_\tau\mu_\tau$
  $\mathbf{x}_\tau \leftarrow \mathbf{x}_\tau + (\alpha/\gamma_\tau)\,\mathbf{u}_\tau$
`end while`

ALGORITHM 9.1. Multishift CGLS for solving $(\mathbf{A}^*\mathbf{A} + \tau^2\mathbf{I})\mathbf{x}_\tau = \mathbf{A}^*\mathbf{b}$ for $\mathbf{x}_\tau$ with residual accuracy *tol*. $\mathbf{A}$ is an $n \times m$ matrix, $\mathbf{b}$ is an n-vector, τ is a regularisation parameter. For each additional parameter τ, the solution $\mathbf{x}_\tau$ can be obtained essentially at the additional costs of one vector update per step.

Compute scalars ρ_k such that $\rho_{-1} = 0$, $\rho_0 = \|\mathbf{b}\|_2$ and

$$\frac{\beta_{k+1}}{\rho_{k+1}} + \frac{\alpha_k}{\rho_k} + \frac{\beta_k}{\rho_{k-1}} = 0 \qquad (k = 0, 1, 2, \ldots).$$

In Exercise 7.6, we learnt that $\mathbf{r}_k \equiv \rho_k\mathbf{v}_k$ are residuals for problem (9.3) with $\sigma = 0$ and we also learnt that, with $\mathbf{x}_{-1} = \mathbf{x}_0 = \mathbf{0}$,

$$\frac{\beta_{k+1}}{\rho_{k+1}}\mathbf{x}_{k+1} = -\mathbf{v}_k - \frac{\alpha_k}{\rho_k}\mathbf{x}_k - \frac{\beta_k}{\rho_{k-1}}\mathbf{x}_{k-1} \qquad (k = 0, 1, 2, \ldots)$$

generates the associated approximate solution $\mathbf{x}_k$.

(a) Show that, replacing α_k by $\alpha_k + \sigma$ leads to approximate solutions for the shifted problem.

(b) Design an efficient **multishift Lanczos method**.

As we saw in Exercise 7.5, CG can be viewed as a Lanczos variant that relies on an LU-decomposition of the Lanczos tri-diagonal matrix $\underline{T}_k$. Clearly, this insight can be used to form multishift CG. The two coupled two term recurrence relations that define CG are more stable than the three term recurrence relation that defines Lanczos. However, to maintain this stability advantage in a multishift variant, care is needed in the update of the LU-decomposition of $\underline{T}_k + \sigma\underline{I}_k$

Exercise 9.11. **Multishift CGLS.** Consider the diagonal matrix $D = \mathrm{diag}(d_0, \ldots, d_k)$ and lower bi-diagonal matrix L with all ones on the diagonal and $\ell_0, \ldots, \ell_{k-1}$ on the lower diagonal.

(a) Show that the product matrix $T \equiv LDL^*$ has $d_0, d_0\ell_0^2 + d_1, \ldots, d_{k-1}\ell_{k-1}^2 + d_k$ on the diagonal and $d_0\ell_0, d_1\ell_1, \ldots, d_{k-1}\ell_{k-1}$ on the co-diagonals:

$$T_{jj} = d_{j-1}\,\ell_{j-1}^2 + d_j, \quad T_{j+1,j} = d_{j-1}\ell_{j-1} \qquad (j = 1, 2, \ldots, k).$$

We want to compute the diagonal D_τ and lower diagonal L_τ factors of the shifted version of LDL^* from the entries of D and L:

$$LDL^* + \tau^2 I = L_\tau D_\tau L_\tau^*.$$

(b) Prove that the d_j^τ (diagonal of D_τ) and ℓ_j^τ (lower diagonals of L_τ) can be obtained by the following recursion

$$t_0^\tau = \tau^2, \quad d_j^\tau = t_j^\tau + d_j, \quad \ell_j^\tau = \ell_j \frac{d_j}{d_j^\tau}, \quad t_{j+1}^\tau = \tau^2 + \ell_j^\tau \ell_j t_j^\tau \qquad (j = 0, 1, \ldots) \tag{9.4}$$

It can be proved that this recursion is stable, in contrast to the more obvious approach, where first T is explicitly formed, shifted by $\tau^2 I$ and the standard LU-decomposition (or Cholesky) approach is followed to compute the L_τ and D_τ factors. If we apply CG, then, we actually compute the factors of the Lanczos matrix, that is, we have the D and L factors rather than T.
To be more precise, use Exercise 7.5, where it is shown that

$$\underline{T}_k = (D_\rho \underline{J}_k D_\rho^{-1}) D_\alpha^{-1} (D_\rho^{-1} J_k^* D_\rho) \tag{9.5}$$

with $D_\alpha \equiv \mathrm{diag}(\alpha_0, \alpha_1, \ldots)$ the diagonal matrix of α_j coefficients of the CG process, $D_\rho \equiv \mathrm{diag}(\sqrt{\rho_0}, \sqrt{\rho_1}, \ldots)$ the diagonal matrix of norms $\sqrt{\rho_j} \equiv \|\mathbf{r}_j^{\mathrm{CG}}\|_2$ of the CG-residuals. In the present application, $\sqrt{\rho_j}$ is the norm of the CGLS residual $\mathbf{s}_j$ of the normal equation, $\sqrt{\rho_j} \equiv \|\mathbf{s}_j\|_2$, α_j (and β_j below) of CG is precisely the α_j (and β_j) coefficient of the CGLS algorithm. The dimension of the diagonal matrices D_α and D_ρ should be clear from the context. From (9.5), and the definition of β_{j+1}, we see that

$$d_j = \frac{1}{\alpha_j}, \quad \ell_j = \frac{\sqrt{\rho_{j+1}}}{\sqrt{\rho_j}} \quad \text{and} \quad -\ell_j^2 = \beta_{j+1}.$$

(c) To show that this leads to the multishift version in Alg. 9.1 of CGLS, prove that

$$\mu_j^\tau \equiv \frac{\ell_j}{\ell_j^\tau} = \frac{d_j^\tau}{d_j} = 1 + \alpha_j t_j^\tau \quad \text{and} \quad t_{j+1}^\tau = \tau^2 - \frac{\beta_{j+1}}{\mu_j^\tau} t_j^\tau.$$

How does this rewriting of (9.4) affects the stability? Moreover, with ρ_j^τ, α_j^τ and β_j^τ the CG coefficients ρ_j, α_j and β_j, respectively, for CG aplied to the shifted normal equations ('shifted CGLS'), we have that

$$\gamma_{j+1}^\tau \equiv \frac{\sqrt{\rho_{j+1}}}{\sqrt{\rho_{j+1}^\tau}} = \mu_j^\tau \gamma_j^\tau, \quad \alpha_j^\tau = \frac{\alpha_j}{\mu_j^\tau} \quad \text{and} \quad \beta_j^\tau = \frac{\beta_j}{(\mu_j^\tau)^2}.$$

(d) Prove that $\mathbf{s}_j^\tau \equiv \mathbf{A}^*(\mathbf{b} - \mathbf{A}\mathbf{x}_j^\tau) - \tau^2 \mathbf{x}_j^\tau$ and $\mathbf{s}_j$ are co-linear, i.e., $\mathbf{s}_j^\tau$ is a scalar multiple of $\mathbf{s}_j$. To be more precise

$$\mathbf{s}_j = \gamma_j^\tau \mathbf{s}_j^\tau.$$

(e) Derive Alg. 9.1, where the $\mathbf{u}_j^\tau$ has been scaled such that $\mathbf{s}_j$ can be used to update $\mathbf{u}_j^\tau$ rather than $\mathbf{s}_j^\tau$.

Exercise 9.12. **Multishift LSQR.** Combine Exercise 9.1(d), Exercise 9.6, Exercise 9.7, to adapt the arguments in Exercise 9.8 for a derivation of a multishift variant of LSQR.

A Perturbed Least Square problems

We discuss the forward stability of the least square problem and the minimal norm problem.

Minimal residual (least square)

Let $\mathbf{A}$ be an $n \times k$ matrix with $k \leq n$ and singlar values $\sigma_1 \geq \ldots \geq \sigma_k > 0$: $\mathbf{A}$ has full column rank. Let $\mathbf{x}$ be the MR (minimal residual) solution with residual $\mathbf{r}$ of the MR problem:

$$\|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 = \min \quad \text{and} \quad \mathbf{r} \equiv \mathbf{b} - \mathbf{A}\mathbf{x}. \tag{9.6}$$

Then $\mathbf{A}^*\mathbf{r} = \mathbf{0}$, $\mathbf{x} = \mathbf{A}^\dagger\mathbf{b}$, with $\mathbf{A}^\dagger \equiv (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*$ the Moore–Penrose pseudo inverse of $\mathbf{A}$, and

$$\mathcal{C}_2(\mathbf{A}) \equiv \|\mathbf{A}\|_2 \, \|\mathbf{A}^\dagger\|_2 = \frac{\sigma_1}{\sigma_k}.$$

The matrix $\mathbf{A}$ is ill-conditioned of σ_k is relatively close to 0. In such a case regularisation as discussed in the beginning of this lecture is required. However, as we will learn below, least square problems for non-square matrices can be ill-conditioned even if σ_k is not extremely small.

Note that $\Pi_A \equiv \mathbf{A}(\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^* = \mathbf{A}\mathbf{A}^\dagger$ projects orthogonal onto $\mathcal{R}(\mathbf{A})$, while $\mathbf{I} - \Pi_A$ projects onto $\mathcal{R}(\mathbf{A})^\perp$. Put

$$\nu \equiv \|\mathbf{I} - \Pi_A\|_2.$$

Then $\nu = 1$ if $k < n$ and $\nu = 0$ if $k = n$, since $\Pi_A = \mathbf{I}$.

Consider the pertubed matrix $\mathbf{A} + \Delta$ and the perturbed input vector $\mathbf{b} + \delta_b$. Put

$$\varepsilon_A \equiv \frac{\|\Delta\|_2}{\|\mathbf{A}\|_2} \quad \text{and} \quad \varepsilon_b \equiv \frac{\|\delta_b\|_2}{\|\mathbf{b}\|_2}.$$

Theorem 9.2 *Assume $\mathcal{C}_2(\mathbf{A})\varepsilon_A < \sqrt{2} - 1$. Then $\mathbf{A} + \Delta$ has full column rank.*
Let $\widetilde{\mathbf{x}}$ and $\widetilde{\mathbf{r}}$ be the MR solution and residual, respectively, of the perturbed problem. Then

$$\widetilde{\mathbf{x}} - \mathbf{x} = \mathbf{A}^\dagger(\delta_b - \Delta\widetilde{\mathbf{x}}) + (\mathbf{A}^*\mathbf{A})^{-1}\Delta\widetilde{\mathbf{r}}, \quad \frac{\|\widetilde{\mathbf{x}} - \mathbf{x}\|_2}{\|\mathbf{x}\|_2} \lesssim \mathcal{C}_2(\mathbf{A}) \left(\varepsilon_b + \varepsilon_A + \frac{(\varepsilon_b + \mathcal{C}_2(\mathbf{A})\varepsilon_A)\|\mathbf{r}\|_2}{\|\mathbf{A}\|_2\|\mathbf{x}\|_2} \right),$$

$$\widetilde{\mathbf{r}} - \mathbf{r} = (\mathbf{I} - \Pi_A)(\delta_b - \Delta\widetilde{\mathbf{x}}) + \Pi_A\widetilde{\mathbf{r}}, \qquad \frac{\|\widetilde{\mathbf{r}} - \mathbf{r}\|_2}{\|\mathbf{b}\|_2} \lesssim \nu(\varepsilon_b + \varepsilon_A\mathcal{C}_2(\mathbf{A})) + 2\varepsilon_A\mathcal{C}_2(\mathbf{A})\frac{\|\mathbf{r}\|_2}{\|\mathbf{b}\|_2}.$$

If $n = k$ then $\mathbf{r} = \widetilde{\mathbf{r}} = \mathbf{0}$, $\nu = 0$ and the bounds in the theorem coincide with the one in Theorem 1.9. If $k < n$, then it appears that the sensitivity of the MR problem is determined by the condition number $\mathcal{C}_2(\mathbf{A})$ of the matrix $\mathbf{A}$ if the residual is zero (or small), i.e. $\mathbf{b}$ is (almost) in the range of $\mathbf{A}$, while the sensitivity depends on the square of this condition number if $\mathbf{r}$ is significant: $\|\mathbf{r}\|_2/\|\mathbf{b}\|_2$ is the sine of the angle between $\mathbf{b}$ and the range of $\mathbf{A}$, while $\|\mathbf{r}\|_2/(\|\mathbf{A}\|_2\,\|\mathbf{x}\|_2)$ is bounded by $\|\mathbf{r}\|_2/\|\mathbf{A}\mathbf{x}\|_2$, the tangent of this angle. The estimates are sharp (in order of magnitude); see Exercise 9.13. The quantity

$$\mathcal{C}_{LS}(\mathbf{A}, \mathbf{b}) = \mathcal{C}_2(\mathbf{A}) \left(1 + \mathcal{C}_2(\mathbf{A}) \frac{\|\mathbf{r}\|_2}{\|\mathbf{A}\|_2\,\|\mathbf{x}\|_2} \right) \tag{9.7}$$

appears to characterize the **conditioning of the least square problem** (with respect to perturbations on $\mathbf{A}$).

Note that, unlike for a linear system with a square non-singular matrix, for an MR problem the conditioning $\mathcal{C}_2(\mathbf{A})$ of the matrix is not the same as the conditioning $\mathcal{C}_{LS}(\mathbf{A}, \mathbf{b})$ of the problem. The solution of an MR problem can also be obtained as the solution of a non-singular square system, actually of two different non-singular square systems (see Exercise 9.1(a)). But, the conditioning of these matrices are also both very different from $\mathcal{C}_{LS}(\mathbf{A}, \mathbf{b})$. This may seem a bit strange, but apparently the fact that the perturbation is structured plays a role in the conditioning: for instance, $\mathbf{A}^*\mathbf{A}$ is perturbed as $\mathbf{A}^*\mathbf{A} + \Delta_2$ with a structured Δ_2, $\Delta_2 = \Delta^*\mathbf{A} + \mathbf{A}^*\Delta + \Delta^*\Delta$, rather than an arbitrary Δ_2. Note that the righthand side vector ($\mathbf{A}^*\mathbf{b}$) is structured as well.

We cannot expect to be able to compute the exact solution, but we may hope to be able to obtain a solution with an error of the size as indicated in the theorem (with the ε's modest multiples of the machine precision $\mathbf{u}$). We call a numerical method **backward stable** if, in rounded arithmetic, it computes a solution $\widetilde{\mathbf{x}}$ that is the exact solution of a perturbed problem with perturbations Δ and δ_b of order machine precision. Methods that compute $\mathbf{A}^*\mathbf{A}$ and $\mathbf{A}^*\mathbf{b}$ first and then solve $\mathbf{x}$ from the normal equation cannot be backward stable with respect to the MR problem. Rounding errors will at least introduce errors of the size

$$\frac{\|\widetilde{\mathbf{x}} - \mathbf{x}\|_2}{\|\mathbf{x}\|_2} \leq n\,\mathbf{u}\,\mathcal{C}_2^2(\mathbf{A}).$$

The same holds for general methods solving the 'companion form' representation of the MR-problem (cf., the last formulation of the MR problem in Exercise 9.1(a)). Note that these observation do not apply to CGLS and LSQR. They both exploit the specific structure of the problem in each step: CGLS exploits the fact that $\mathbf{A}^*\mathbf{A}$ is available in factorized form, while LSQR computes a projected system with the same structure as the original one.

For systems of moderate dimension, the QR-decomposition, $\mathbf{A} = \mathbf{QR}$, using Householder reflections can be used: $\mathbf{x} = \mathbf{R}^{-1}(\mathbf{Q}^*\mathbf{b})$. This method is **backward stable**. Note that $\mathbf{Q}^*\mathbf{b}$ can be computed without storing $\mathbf{Q}$: a Householder reflection can be applied to '$\mathbf{b}$' as soon as it is formed in a steps of the QR-decomposition.

For high dimensional problems iterative methods as CGLS and LSQR can be effective. If, for these iterative methods, at termination $\|\mathbf{r}_k - \mathbf{r}\|_2$ is a modest multiple of $\mathbf{u}\|\mathbf{A}\|_2\,\|\mathbf{x}\|_2$, then we have an approximate solution with an error that corresponds to the forward error in a solution computed by a backward stable method.

Exercise 9.13. Discuss the sharpness of the estimates in Theorem 9.2. Hint. Consider

$$\mathbf{A} + \Delta = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \\ 0 & \varepsilon \end{bmatrix} \qquad \mathbf{b} = \mathbf{b} + \delta_b = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}.$$

Exercise 9.14. ***Proof of Theorem*** 9.2***.***

(a) Show that $\mathbf{A} + \Delta$ has full rank if $2\mu + \mu^2 \leq 1$ for $\mu \equiv \varepsilon_A \mathcal{C}_s(\mathbf{A})$.

(b) Show that

$$\begin{bmatrix} \mathbf{I} & \mathbf{A} \\ \mathbf{A}^* & \mathbf{0} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{I} - \Pi_A & (\mathbf{A}^\dagger)^* \\ \mathbf{A}^\dagger & -(\mathbf{A}^*\mathbf{A})^{-1} \end{bmatrix}$$

(c) Observe that

$$\begin{bmatrix} \mathbf{I} & \mathbf{A} \\ \mathbf{A}^* & \mathbf{0} \end{bmatrix} \begin{bmatrix} \widetilde{\mathbf{r}} - \mathbf{r} \\ \widetilde{\mathbf{x}} - \mathbf{x} \end{bmatrix} = \begin{bmatrix} \delta_b - \Delta\widetilde{\mathbf{x}} \\ -\Delta^*\widetilde{\mathbf{r}} \end{bmatrix}$$

Show that $\Delta^*\widetilde{\mathbf{r}} = -\mathbf{A}^*\widetilde{\mathbf{r}}$ and prove the expresssions for $\widetilde{\mathbf{x}} - \mathbf{x}$ and $\widetilde{\mathbf{r}} - \mathbf{r}$ as given in the theorem.

(d) Prove that $\|\Pi_A(\mathbf{I} - \Pi_{\widetilde{A}})\|_2 = \|\mathbf{A}(\mathbf{A}^*\mathbf{A})^{-1}(\mathbf{A}^* - \widetilde{\mathbf{A}}^*)(\mathbf{I} - \Pi_{\widetilde{A}})\|_2 \leq \mathcal{C}_2(\mathbf{A})\varepsilon_A$.

(e) Prove the estimates of the theorem.

Minimal norm

Let $\mathbf{A}$ be an $k \times n$ with $k \leq n$ and singlar values $\sigma_1 \geq \ldots \geq \sigma_k > 0$: $\mathbf{A}$ has full row rank. Let $\mathbf{x}$ be the MN (minimal norm) solution of the MN problem:

$$\{\|\mathbf{x}\|_2 \,\big|\, \mathbf{A}\mathbf{x} = \mathbf{b}\} = \min. \tag{9.8}$$

We put $\mathbf{y} \equiv (\mathbf{A}\mathbf{A}^*)^{-1}\mathbf{b}$. Then $\mathbf{x} = \mathbf{A}^*\mathbf{y} = \mathbf{A}^\dagger\mathbf{b}$, where now the Moore–Penrose pseudo inverse of $\mathbf{A}$ equals $\mathbf{A}^\dagger \equiv \mathbf{A}^*(\mathbf{A}\mathbf{A}^*)^{-1}$. Note that $\mathbf{y} = (\mathbf{A}^\dagger)^*\mathbf{x}$. As before,

$$\mathcal{C}_2(\mathbf{A}) \equiv \|\mathbf{A}\|_2\,\|\mathbf{A}^\dagger\|_2 = \frac{\sigma_1}{\sigma_k}.$$

Note that $\Pi \equiv \mathbf{A}^*(\mathbf{A}\mathbf{A}^*)^{-1}\mathbf{A} = \mathbf{A}^\dagger\mathbf{A}$ projects orthogonal onto $\mathcal{R}(\mathbf{A}^*)$, while $\mathbf{I} - \Pi$ projects onto $\mathcal{N}(\mathbf{A}) = \mathcal{R}(\mathbf{A}^*)^\perp$. With $\nu \equiv \|\mathbf{I} - \Pi\|_2$, $\nu = 1$ if $k < n$ and $\nu = 0$ if $k = n$.

Consider the pertubed matrix $\mathbf{A} + \Delta$ and the perturbed input vector $\mathbf{b} + \delta_b$. Put

$$\varepsilon_A \equiv \frac{\|\Delta\|_2}{\|\mathbf{A}\|_2} \quad \text{and} \quad \varepsilon_b \equiv \frac{\|\delta_b\|_2}{\|\mathbf{b}\|_2}.$$

Theorem 9.3 *Assume* $\mathcal{C}_2(\mathbf{A})\varepsilon_A < \sqrt{2}-1$. *Then* $\mathbf{A}+\Delta$ *has full row rank. Let* $\widetilde{\mathbf{x}}$ *be the MN solution of the perturbed problem. Then*

$$\widetilde{\mathbf{x}}-\mathbf{x} = \mathbf{A}^{\dagger}(\delta_b - \Delta\widetilde{\mathbf{x}}) + (\mathbf{I}-\Pi)\Delta\widetilde{\mathbf{y}}, \qquad \frac{\|\widetilde{\mathbf{x}}-\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \lesssim \mathcal{C}_2(\mathbf{A})\left(\varepsilon_b + (1+\nu)\varepsilon_A\right).$$

If $n=k$, then $\mathbf{A}^{\dagger}=\mathbf{A}^{-1}$ and $\Pi=\mathbf{I}$ and the formulas coincide with the ones in Theorem 1.9. The new term $(\mathbf{I}-\Pi)\Delta\widetilde{\mathbf{y}}$ that appears if $k<n$ does not affect the conditioning:

$$\|(\mathbf{I}-\Pi)\Delta\widetilde{\mathbf{y}}\|_2 \le \nu\,\|\Delta\|_2\,\frac{\|\widetilde{\mathbf{x}}\|_2}{\sigma_k(\widetilde{\mathbf{A}})} \lesssim \nu\,\mathcal{C}_2(\mathbf{A})\,\varepsilon_A\,\|\mathbf{x}\|_2.$$

The conditioning of the MN problem equals twice (actually $1+\nu$ times) the conditioning of the matrix $\mathbf{A}$. This seems a bit surprising, since the condition number of the matrices in Exercise 9.1(b) in both formulations are proportional to $\mathcal{C}_2^2(\mathbf{A})$. However, $\widetilde{\mathbf{y}}-\mathbf{y} = (\mathbf{A}^*\mathbf{A})^{-1}(\delta_b - \Delta\widetilde{\mathbf{x}}) - (\mathbf{A}^{\dagger})^*\Delta\widetilde{\mathbf{y}}$: the ill-conditioning is reflected in $\mathbf{y}$ (in the first term). The large error from $1/\sigma_k^2$ is partially annihilated in $\mathbf{x}$ by the multiplication of $\mathbf{y}$ by $\mathbf{A}^*$. $\mathbf{x}$ 'benifits' from the fact that the error on $\mathbf{y}$ is structured. We benifit from the fact that $\mathbf{y}$ is only a intermediate quatity.

Exercise 9.15. ***Proof of Theorem*** 9.3**.** Use Exercise 9.1(b) and adapt the arguments in Exercise 9.13 to prove the theorem.