# Exercise A: Least square problems

## A1

Minimising $\|Ax-b\|_2$ with respect to x means finding x such that the derivative of $\|Ax-b\|_2$ with respect to x is equal to zero. Equivalently, we can minimise the square of the norm which can be developed as follows:

$$\begin{aligned}\|Ax-b\|_2^2 &= \|Ax\|_2^2 - 2 < Ax, b > + \|b\|_2^2 \\ &= x^T A^T Ax - 2x^T A^T b + b^T b\end{aligned}$$

The derivative with respect to x is:

$$\begin{aligned}\frac{\partial \|Ax-b\|_2^2}{\partial x} &= \frac{\partial (x^T A^T Ax - 2x^T A^T b + b^T b)}{\partial x} \\ &= \frac{\partial (x^T A^T Ax)}{\partial x} - \frac{\partial (2x^T A^T b)}{\partial x} + \frac{\partial (b^T b)}{\partial x} \\ &= 2A^T Ax - 2A^T b + 0\end{aligned}$$

This derivative has to be equal to zero. Thus:

$$\begin{aligned}\frac{\partial \|Ax-b\|_2^2}{\partial x} &= 0 \\ \Leftrightarrow 2A^T Ax - 2A^T b &= 0 \\ \Leftrightarrow 2A^T Ax &= 2A^T b \\ \Leftrightarrow A^T Ax &= A^T b\end{aligned}$$

To demonstrate that when A has full column rank, i.e. rank(A)=n, the solution is unique, we first prove that Ker($A^T A$)=Ker(A):

$$\forall x \in Ker(A) : Ax = 0 \Leftrightarrow A^T Ax = 0 \Leftrightarrow x \in Ker(A^T A) \Rightarrow Ker(A) \subset Ker(A^T A)$$

$$\begin{aligned}\forall x \in Ker(A^T A) : A^T Ax = 0 \Leftrightarrow x^T A^T Ax = 0 \Leftrightarrow \|Ax\|_2 = 0 \Leftrightarrow Ax = 0 \Leftrightarrow x \in Ker(A) \\ \Rightarrow Ker(A^T A) \subset Ker(A)\end{aligned}$$

According to the rank-nullity Theorem:

$$\begin{aligned}rank(A) &= n - dim(Ker(A)) \\ yet : rank(A) &= n \\ \Rightarrow n &= n - dim(Ker(A)) \\ \Leftrightarrow dim(Ker(A)) &= 0 = dim(Ker(A^T A)) \\ \Rightarrow Ker(A^T A) &= \{0\}\end{aligned}$$

As $Ker(A^T A) = \{0\}$ and $A^T A$ is square, we deduce $A^T A$ is invertible and it follows from Theorem 2.1 of the course notes that the solution of the system is unique.

## A2

Suppose the QR decomposition of $A$ is given by $Q\binom{R}{0}$, where $Q \in \mathbb{R}^{m\times m}$ is unitary and $R \in \mathbb{R}^{n\times n}$ is upper triangular. We will express the solution of

$$A^T Ax = A^T b \tag{1}$$

in terms of the QR decomposition of A.
Let $R_f = \binom{R}{0}$. We can rewrite equation (1) as:

$$(QR_f)^T QR_f x = (QR_f)^T b$$
$$R_f^T Q^T QR_f x = R_f^T Q^T b$$

As Q is unitary and hence $Q^T Q = I$, we have:

$$\begin{pmatrix} R^T & 0 \end{pmatrix} \begin{pmatrix} R \\ 0 \end{pmatrix} x = \begin{pmatrix} R^T & 0 \end{pmatrix} Q^T b$$

If we call $\hat{Q}$ the matrix consisting of the first n columns of Q, equation (1) becomes:

$$R^T Rx = R^T \hat{Q}^T b$$

From theorem 2.8 of the course notes, we know that every matrix $A \in \mathbb{C}^{m\times n}$ of full column-rank admits a factorization $A = Q_1 R_1$ where $Q_1 \in \mathbb{C}^{m\times n}$ is an isometry and $R_1 \in \mathbb{C}^{n\times n}$ is an upper triangular matrix with positive diagonal. The matrix $Q_1$ corresponds to $\hat{Q}$ and $R_1$ simply to $R$. Hence we deduce that $R$ is invertible. This allows us to premultiply both sides of the equation by the inverse of $R^T$:

$$R^{-T} R^T Rx = R^{-T} R^T \hat{Q}^T b$$
$$Rx = \hat{Q}^T b$$

The solution of equation (1) is therefore $x = R^{-1}\hat{Q}^T b$ and the computation of the solution is reduced to the resolution of a single triangular system of linear equations (which can be solved efficiently using backward substitution).

## Exercise B: Low-rank approximation

### B1

For every matrix $A \in \mathbb{R}^{m\times n}$, there exist unitary transformations $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ such that

$$A = U\Sigma V^*, \quad \text{where} \quad \Sigma = \left[\begin{array}{ccc|c} \sigma_1 & & 0 & \\ & \ddots & & 0_{r\times(n-r)} \\ 0 & & \sigma_r & \\ \hline & 0_{(m-r)\times r} & & 0_{(m-r)\times(n-r)} \end{array}\right],$$

with real positive singular values $\sigma_1 \geqslant \cdots \geqslant \sigma_r > 0$.

These singular values are unique: the intuition to see this is that the singular value decomposition is computed inductively, and that the unitary matrices preserve the norm. By taking the property that $\|X\|_2 = \sigma_1$, this means that at every step of the decomposition, no matter what unitary transformations are chosen, the norm (and thus the maximal singular value of the submatrix we are working on) is the same.

Next, we show that the rank of a matrix is equal to its number of nonzero singular values.

*Proof.* We know that the rank of a diagonal matrix is equal to the number of its nonzero entries. We also note that in the decomposition $A = U\Sigma V$, $U$ and $V$ are of full rank. Therefore, $\text{rank}(A) = \text{rank}(\Sigma) = r$. □

### B2

Let $x \in \mathbb{R}^{m\times n}$ be such that $|X_{ij}| \leqslant \varepsilon$ for all $i \in \{1,\ldots,m\}$ and $j \in \{1,\ldots,n\}$. Let $\|X\|_2$ be the 2-norm of $X$ and let $\|X\|_F$ be its Frobenius norm. We show that $\|X\|_2 \leqslant \|X\|_F \leqslant \sqrt{mn}\varepsilon$.

*Proof.* First, we show the first inequality. We know from the lecture notes that

$$\|X\|_2 = \sigma_{\max},$$
$$\|X\|_F = \left[\sum_i \sigma_i\right]^{1/2},$$

where $\sigma_i$ are the singular values of $X$. From this, it is immediately clear that $\|X\|_2 \leqslant \|X\|_F$.

Next, we use an equivalent form of the Frobenius norm to show the second inequality:

$$\|X\|_F = \left[\sum_{i,j} |X_{ij}|^2\right]^{1/2}.$$

Knowing that $|X_{ij}| \leqslant \varepsilon$, it is immediate that $\|X\|_F \leqslant \left[\sum_{i,j} \varepsilon^2\right]^{1/2} = \left[mn\varepsilon^2\right]^{1/2} = \sqrt{mn}\varepsilon$. This concludes the proof. □

We also give an example where these bounds are tight. Indeed, consider the matrix $X = I_1 \in \mathbb{R}^{1\times 1}$. Clearly, we have $|X_{ij}| \leqslant \varepsilon = 1$ for all $i, j$ (only one value is possible for each). We know that the only singular value of this matrix is 1, and hence

$$\|X\|_2 = \|X\|_F = \sqrt{1 \cdot 1}\varepsilon = 1\varepsilon.$$

### B3

We start by observing that if $B = A + X$, then by Theorem 3.28 in the lecture notes, we can write

$$\begin{aligned}
\sigma_{\min(m,n)-j+1}(B) &= \min_{\mathcal{S}_j} \max_{x\in\mathcal{S}_j\setminus\{0\}} \frac{\|B\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} \\
&= \min_{\mathcal{S}_j} \max_{x\in\mathcal{S}_j\setminus\{0\}} \frac{\|(A+X)\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} \\
&\leqslant \min_{\mathcal{S}_j} \max_{x\in\mathcal{S}_j\setminus\{0\}} \left(\frac{\|A\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} + \frac{\|X\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2}\right) \\
&\leqslant \sigma_{\min(m,n)-j+1}(A) + \sigma_{\min(m,n)-j+1}(X).
\end{aligned}$$

However, we know that $A$ has rank $r$, and hence by the result of B1, we find that if $\min(m,n) - j + 1 > r$, the singular value of $A$ in the expression is zero, and hence that

$$\begin{aligned}
\sigma_{\min(m,n)-j+1}(B) &\leqslant \sigma_{\min(m,n)-j+1}(X) \\
&\leqslant \sqrt{mn}\varepsilon,
\end{aligned}$$

where the last inequality is a consequence of B2.

A criterion that can then be used to estimate the rank $r$ of $A$ is then to take the smallest $r$ such that $\sigma_{r+1}(B) \leqslant \sqrt{mn}\varepsilon$. This is very similar to the description given on p. 60 of the lecture notes, concerning the numerical rank of the matrix $A$.

## Exercise C: Realization theory

In this last exercise, we are interested in finding an AR model corresponding to the data obtained during the covid pandemic. Indeed, we want to find the parameters $\alpha_i$ of :

$$y(t) = \alpha_0 + \sum_{i=1}^{p} \alpha_i y(t-i)$$

For this, we will minimimize the squared error $f = \|y - \hat{y}\|_2$. (We have N+1 outputs)

Let's rewrite our problem as a system of linear equations to have a least squares problem.

$$\begin{bmatrix} \hat{y}(p) \\ \vdots \\ \hat{y}(N) \end{bmatrix} = \begin{bmatrix} 1 & y(p-1) & y(p-2) & \dots & y(0) \\ 1 & y(p) & y(p-1) & \dots & y(1) \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & y(N-1) & y(N-2) & \dots & y(N-p) \end{bmatrix}$$

Note that we consider $\hat{y}(0) = y(0) \dots \hat{y}(p-1) = y(p-1)$ as initial conditions. Thus our problem can be written as $\min_{x\in\mathbb{R}^{p+1}} \|Ax - y\|_2$.

From there, we can use what we obtained in question A2 to solve the problem with a QR decomposition and backward substitution.

**Discussion**

**Bonus question**