Scientific Computing additional chapters

Walter Mudzimbabwe
First version: 15 November 2022
This version: 9 February 2023

1 Chapter: ODEs, Boundary value problems

1.1 Boundary value problems (BVPs)

A linear second order boundary value problem (BVP) is

$$\begin{cases} y''(x) + p(x)y'(x) + q(x)y(x) = r(x) \\ y(a) = \alpha, \qquad y(b) = \beta \end{cases}$$

and a nonlinear second order boundary value problem (BVP) is

$$\begin{cases} y''(x) = f(x, y'(x), y''(x)) \\ y(a) = \alpha, \qquad y(b) = \beta \end{cases}$$

1.2 Finite differences for BVPs

Here we solve only a linear second order BVP:

$$\begin{cases} y''(x) + p(x)y'(x) + q(x)y(x) = r(x) \\ y(a) = \alpha, \qquad y(b) = \beta \end{cases}$$

First subdivide $[a, b]$ into N subintervals with size h. So

$$h = \frac{b-a}{N}, \qquad x_i = a + ih, \quad i = 0, 1, \cdots, N.$$

The finite difference method for (BVPs) consists of replacing derivatives in the BVP by difference approximations. For example:

$$y'(x_i) \approx \frac{y(x_{i+1}) - y(x_{i-1})}{2h}$$
$$y''(x_i) \approx \frac{y(x_{i+1}) - 2y(x_i) + y(x_{i-1})}{h^2}$$

Substituting these approximations in the BVP we get:

$$\left(1 - \frac{h}{2}p_i\right) y_{i-1} + (-2 + h^2 q_i) y_i + \left(1 + \frac{h}{2}p_i\right) y_{i+1} = h^2 r_i, \tag{1}$$

where $i = 1, 2, \cdots, N-1, \quad y_0 = \alpha, \quad y_N = \beta$ and

$$y_i \approx y(x_i), \quad p_i = p(x_i), \quad q_i = q(x_i), \quad r_i = r(x_i).$$

So there are $N-1$ equations in $N-1$ unknowns.

1.3 System of $N-1$ equations in $N-1$ unknowns

$$\begin{bmatrix} b_1 & c_1 & & & & \\ a_2 & b_2 & c_2 & & 0 & \\ & a_3 & b_3 & c_3 & & \\ & & \ddots & \ddots & \ddots & \\ & 0 & & a_{n-2} & b_{n-2} & c_{n-2} \\ & & & & a_{n-1} & b_{n-1} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_{N-2} \\ y_{N-1} \end{bmatrix} = \begin{bmatrix} d_1 - a_1\alpha \\ d_2 \\ d_3 \\ \vdots \\ d_{N-2} \\ d_{N-1} - c_{N-1}\beta \end{bmatrix},$$

where

$$a_i = 1 - \frac{h}{2}p_i, \quad b_i = -2 + h^2 q_i, \quad c_i = 1 + \frac{h}{2}p_i, \quad d_i = h^2 r_i,$$

for $i = 1, 2, \cdots, N-1$.
This is a tridiagonal system using Gaussian elimination, LU factorisation etc.

1.4 Example: BVP using difference approximations

Solve the second order BVP:

$$\begin{cases} y''(x) + (x+1)y'(x) - 2y(x) = (1-x^2)e^{-x} \\ y(0) = -1, \qquad y(1) = 0 \end{cases},$$

using $h = 0.2$. Compare the approximate solution with exact solution $y = (x-1)e^{-x}$.

Solution: Here

$$p(x) = (x+1), \quad q(x) = -2, \quad r(x) = (1-x^2)e^{-x}.$$

Equation (1) becomes

$$\begin{aligned}[1 - 0.1(x_i + 1)]y_{i-1} + (-2 - 0.08)y_i + [1 + 0.1(x_i + 1)]y_{i+1} \\ = 0.04(1 - x_i^2)e^{-x_i},\end{aligned}$$

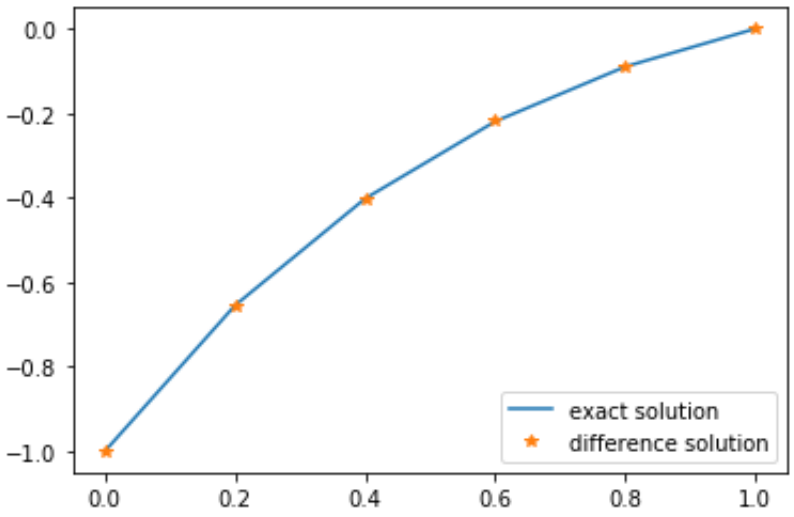
where $y_0 = -1, \quad y_5 = 0$ and $x_i = 0.2i, \quad i = 1, 2, 3, 4$

The resulting system of equations is

$$\begin{bmatrix} -2.08 & 1.12 & 0 & 0 \\ 0.86 & -2.08 & 1.14 & 0 \\ 0 & 0.84 & -2.08 & 1.16 \\ 0 & 0 & 0.82 & -2.08 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 0.91143926 \\ 0.02252275 \\ 0.01404958 \\ 0.00647034 \end{bmatrix}$$

1.5 Comparison of exact and difference approximation

x	Difference solution	Exact solution
0.0	-1.00000000	-1.0000000
0.2	-0.65413043	-0.65498460
0.4	-0.40102860	-0.40219203
0.6	-0.21847768	-0.21952465
0.8	-0.08924136	-0.08986579
1.0	0.00000000	0.00000000



2 Chapter: Matrix Computations

2.1 Matrix and Vector operations

Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ then we write

$$\mathbf{A} = [a_{ij}] = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \vdots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}$$

Addition: $\mathbb{R}^{m\times n} \times \mathbb{R}^{m\times n} \longrightarrow \mathbb{R}^{m\times n}$: $\mathbf{C} = \mathbf{A} + \mathbf{B}$ where $c_{ij} = a_{ij} + b_{ij}$
Scalar multiplication: $\mathbb{R} \times \mathbb{R}^{m\times n} \longrightarrow \mathbb{R}^{m\times n}$: $\mathbf{C} = \alpha\mathbf{A}$ where $c_{ij} = \alpha a_{ij}$
Multiplication: $\mathbb{R}^{m\times n} \times \mathbb{R}^{n\times p} \longrightarrow \mathbb{R}^{m\times p}$: $\mathbf{C} = \mathbf{AB}$ where $c_{ij} = \sum_{k=1}^{n} a_{ik}b_{kj}$
Transpose: $\mathbb{R}^{m\times n} \longrightarrow \mathbb{R}^{n\times m}$: $\mathbf{C} = \mathbf{A}^T$ where $c_{ij} = a_{ji}$

2.2 Special square matrices

Symmetric matrix: An $\mathbf{A} \in \mathbb{R}^{n\times n}$ is symmetric if $\mathbf{A}^T = \mathbf{A}$.
Examples:

$$\begin{bmatrix} 1 & 7 & 3 \\ 7 & 4 & 5 \\ 3 & 5 & 0 \end{bmatrix}, \quad \begin{bmatrix} 4 & 0 & 1 & 10 \\ 0 & -3 & 6 & -2 \\ 1 & 6 & 1 & 10 \\ 10 & -2 & 10 & 10 \end{bmatrix}.$$

Orthogonal matrix: An $\mathbf{A} \in \mathbb{R}^{n\times n}$ is orthogonal if $\mathbf{A}^T\mathbf{A} = \mathbf{I}_n$ where $\mathbf{I}_n \in \mathbb{R}^{n\times n}$ is an identity matrix.
The columns and rows are orthonormal i.e., if you take the dot product of any two columns or rows, the product is 0.
Examples:

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \quad \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

$$\begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

From $\mathbf{A}^T\mathbf{A} = \mathbf{I}_n$ we have $\mathbf{A}^{-1} = \mathbf{A}^T$ which makes finding the inverse much easier (than doing lots of row operations, imagine if $n = 1000$!)

Exercise: Show that if $\mathbf{A} \in \mathbb{R}^{n\times n}$ is symmetric and orthogonal then $\mathbf{A}^2 = \mathbf{I}_n$.
Exercise: Verify that each of the above matrices are orthogonal.
Exercise: Let A and B be orthogonal matrices prove that

(a) det(A)=1 or -1.

(b) A^{-1} is an orthogonal matrix.

(c) AB is orthogonal matrix.

(d) A^T is orthogonal matrix.

Exercise: Prove that a product of orthogonal matrices is orthogonal.
Exercise: Verify that

$$\begin{bmatrix} 2/3 & -2/3 & 1 \\ 1/3 & 2/3 & 2/3 \\ 2/3 & 1/3 & -2/3 \end{bmatrix}$$

is an orthogonal matrix.

2.3 Other Special non square matrices

Diagonal matrix: An $\mathbf{A} \in \mathbb{R}^{m\times n}$ is a diagonal matrix if $a_{ij} = 0$ when $i \neq j$.
Notation: We write $\mathbf{A}$ =diag$(\alpha_1, \alpha_2, \cdots, \alpha_k)$ where $k = \min\{m, n\}$ then

$$\mathbf{A} = [a_{ij}] \text{ is diagonal and } a_{ii} = \alpha_i \text{ for } i = 1, 2, \cdots, k$$

Examples:

$$\begin{bmatrix} 7 & 0 \\ 0 & -5 \\ 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} 3 & 0 & 0 \\ 0 & 6 & 0 \end{bmatrix}, \quad \begin{bmatrix} 4 & 0 & 0 \\ 0 & 7 & 0 \\ 0 & 0 & -13 \end{bmatrix}.$$

Not that this is not necessarily square.

2.4 Vector norms

A vector norm on $\mathbb{R}^n$ is a function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ such that:

1. $f(\mathbf{x}) \geq 0, \qquad \forall\, \mathbf{x} \in \mathbb{R}^n$
2. $f(\mathbf{x}+\mathbf{y}) = f(\mathbf{x}) + f(\mathbf{y}), \qquad \forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$
3. $f(\alpha\mathbf{x}) = \alpha f(\mathbf{x}), \qquad \forall\, \alpha \in \mathbb{R}, \forall\, \mathbf{x} \in \mathbb{R}^n.$

We denote such an $f(\mathbf{x})$ by $||\mathbf{x}||$.
Examples:
Holder/p-norms: $||\mathbf{x}||_p = (|x_1|^p + |x_2|^p + \cdots + |x_n|^p)^{1/p}$ For example:

1. $||\mathbf{x}||_1 = |x_1| + |x_2| + \cdots + |x_n|$
2. $||\mathbf{x}||_2 = (|x_1|^2 + |x_2|^2 + \cdots + |x_n|^2)^{1/2} = (\mathbf{x}^T\mathbf{x})^{1/2}$
3. $||\mathbf{x}||_\infty = \max_i |x_i|$

Exercise: Prove that the 2-norm is invariant under orthogonal transformations
Solution: By a an orthogonal transformation of $\mathbf{x}$, we mean $\mathbf{Qx}$ where $\mathbf{Q}$ is an orthogonal matrix. So the question is asking you to prove that $||\mathbf{Qx}||_2 = ||\mathbf{x}||_2$. Now

$$\begin{aligned} ||\mathbf{Qx}||_2^2 &= (\mathbf{Qx})^T\mathbf{Qx} \\ &= \mathbf{x}^T\mathbf{Q}^T\mathbf{Qx} \\ &= \mathbf{x}^T\mathbf{x}, \text{ since } \mathbf{Q}^T\mathbf{Q} = \mathbf{I}_n, \text{ because } \mathbf{Q} \text{ is orthogonal} \\ &= ||\mathbf{x}||_2^2 \end{aligned}$$

which implies $||\mathbf{Qx}||_2 = ||\mathbf{x}||_2$ since the norm can not be negative.

2.5 Matrix norms

A matrix norm on $\mathbb{R}^{m\times n}$ is a function $f : \mathbb{R}^{m\times n} \longrightarrow \mathbb{R}$ such that:

1. $f(\mathbf{A}) \geq 0, \qquad \forall\ \mathbf{A} \in \mathbb{R}^{m\times n}$ with $f(\mathbf{A}) = 0$ iff $\mathbf{A} = \mathbf{0}$
2. $f(\mathbf{A} + \mathbf{B}) = f(\mathbf{A}) + f(\mathbf{B}), \qquad \forall\ \mathbf{A}, \mathbf{B} \in \mathbb{R}^{m\times n}$
3. $f(\alpha\mathbf{A}) = \alpha f(\mathbf{A}), \qquad \forall\ \alpha \in \mathbb{R}, \forall\ \mathbf{A} \in \mathbb{R}^{m\times n}.$

Again, we denote such an $f(\mathbf{x})$ by $||\mathbf{x}||$.
Examples:
Frobenius norm: $||\mathbf{A}||_F = \left(\sum_i^m \sum_j^n |a_{ij}|^2\right)^{1/2}$.
We can also define p-norms but they are not necessary in this course.

2.6 Singular value decomposition (SVD)

Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ then there exist orthogonal matrices

$$\begin{aligned} \mathbf{U} &= [\mathbf{u}_1, \mathbf{u}_2, \cdots, \cdots, \mathbf{u}_m] \in \mathbb{R}^{m\times m} \\ \mathbf{V} &= [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n] \in \mathbb{R}^{n\times n} \end{aligned}$$

such that

$$\mathbf{U}^T\mathbf{AV} = \text{ diag}(\sigma_1, \sigma_2, \cdots, \sigma_p) \tag{2}$$

where $p = \min\{m, n\}$ and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$.
We can write (2) as

$$\mathbf{A} = \mathbf{U}\ \mathrm{diag}(\sigma_1, \sigma_2, \cdots, \sigma_p)\mathbf{V}^T$$

which is called the singular decomposition (SVD) of $\mathbf{A}$.

The σ_i's are called singular values of $\mathbf{A}$ and vectors $\mathbf{u}_i$ and $\mathbf{v}_i$ are the i^{th} left and right singular vectors respectively.

We can also verify that

$$\mathbf{A}\mathbf{v}_i = \sigma_i \mathbf{u}_i$$
$$\mathbf{A}^T\mathbf{u}_i = \sigma_i \mathbf{v}_i$$

To do this we need to verify that

$$\mathbf{A} = \sum_{j=1}^{r} \sigma_j \mathbf{u}_j \mathbf{v}_j^T$$

which implies

$$\mathbf{A}^T = \sum_{j=1}^{r} \sigma_j \mathbf{v}_j \mathbf{u}_j^T$$

Therefore

$$\begin{aligned}
\mathbf{A}\mathbf{v}_i &= \left(\sum_{j=1}^{r} \sigma_j \mathbf{u}_j \mathbf{v}_j^T \right) \mathbf{v}_i \\
&= \sum_{j=1}^{r} \sigma_j \mathbf{u}_j \mathbf{v}_j^T \mathbf{v}_i \\
&= \sigma_i \mathbf{u}_i \mathbf{I}_n \\
&= \sigma_i \mathbf{u}_i
\end{aligned}$$

Exercise: Verify that the SVD of

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 0 & 0 \end{bmatrix}$$

comprises

$$\mathbf{U} = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 1/\sqrt{2} & -1/\sqrt{2} & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{V} = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}$$

and singular values 2 and 0.
Exercise: Verify that the SVD of

$$\mathbf{A} = \begin{bmatrix} 2 & 0 \\ 0 & -3 \\ 0 & 0 \end{bmatrix}$$

comprises

$$\mathbf{U} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{V} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

and singular values 3 and 2.
Exercise: Do a search on https://numpy.org/doc/stable to find out how SVD is done in Python. Can you interpret the output? How would you verify the decomposition?
Exercise: Prove that $\mathbf{A}^T\mathbf{u}_i = \sigma_i\mathbf{v}_i$.
Exercise: Prove that $\mathbf{A}^T\mathbf{u}_i = \sigma_i\mathbf{v}_i$ assuming that $(\mathbf{A}+\mathbf{B})^T = \mathbf{A}^T + \mathbf{B}^T$
Exercise: Show $\mathbf{A} = \mathbf{U}\,\mathrm{diag}(\sigma_1, \sigma_2, \cdots, \sigma_p)\mathbf{V}^T$ using SVD.

2.7 Positive definite systems

$\mathbf{A} \in \mathbb{R}^{m\times n}$ is positive definite if

$$\mathbf{x}^T\mathbf{A}\mathbf{x} > 0, \quad \text{nonzero } \mathbf{x} \in \mathbb{R}^n.$$

Cholesky decomposition: If $\mathbf{A} \in \mathbb{R}^{m\times n}$ is symmetric and positive definite then there exists lower triangular matrix $\mathbf{G} \in \mathbb{R}^{n\times n}$ with positive entries such that

$$\mathbf{A} = \mathbf{G}\mathbf{G}^T.$$

Example: The matrix

$$\mathbf{A} = \begin{bmatrix} 2 & -2 \\ -2 & 5 \end{bmatrix}$$

is positive definite and has Cholesky decomposition

$$\mathbf{G} = \begin{bmatrix} \sqrt{2} & 0 \\ -\sqrt{2} & \sqrt{3} \end{bmatrix}.$$

Exercise: Verify that $\mathbf{A}$ in the example is positive definite.

We can construct the matrix $\mathbf{G}$ by comparing elements in the equation $\mathbf{A} = \mathbf{G}\mathbf{G}^T$.
First note that $i \geq k$ we have

$$\begin{aligned} a_{ik} &= \sum_{p=1}^{k} g_{ip}g_{kp} \\ &= \sum_{p=1}^{k-1} g_{ip}g_{kp} + g_{ik}g_{kk}, \\ \text{this implies,} \qquad g_{ik} &= \left(a_{ik} - \sum_{p=1}^{k-1} g_{ip}g_{kp}\right)/g_{kk}, \qquad i > k. \\ \text{and for } i = k, \qquad g_{kk} &= \left(a_{kk} - \sum_{p=1}^{k-1} g_{kp}^2\right)^{1/2}. \end{aligned}$$

This procedure can be summarised in the following algorithm:

Given $\mathbf{A} \in \mathbb{R}^{m\times n}$ is symmetric and positive definite then the following algorithm computes a lower triangular matrix $\mathbf{G} \in \mathbb{R}^{n\times n}$ such that $\mathbf{A} = \mathbf{G}\mathbf{G}^T$:

$$\begin{aligned}
&\text{For } k = 1, 2 \ldots, n \\
&\qquad g_{kk} = \left(a_{kk} - \sum_{p=1}^{k-1} g_{kp}^2 \right)^{1/2} \\
&\qquad \text{For } i = k+1, k+2 \ldots, n \\
&\qquad\qquad g_{ik} = \left(a_{ik} - \sum_{p=1}^{k-1} g_{ip} g_{kp} \right) / g_{kk}
\end{aligned}$$

3 Chapter: Least squares methods

Goal: To find least square solution of overdetermined systems i.e.,
$\min ||\mathbf{A}\mathbf{x} - \mathbf{b}||_2$ where $\mathbf{A} \in \mathbb{R}^{m\times n}$ $(m > n)$ and $\mathbf{b} \in \mathbb{R}^m$.
Method: Convert problem using orthogonal transformations. The idea is to use transformation to get $\mathbf{A} = \mathbf{Q}\mathbf{R}$ factorisation where $\mathbf{Q}$ is orthogonal and $\mathbf{R}$ is upper triangular. This is equivalent to applying Gram Schmidt to columns of $\mathbf{A}$.

3.1 Properties of least squares problems

Consider the least squares (LS) problem:

$$\text{Find } \mathbf{x} \in \mathbb{R}^n \text{ such that } \mathbf{A}\mathbf{x} = \mathbf{b},\ \mathbf{A} \in \mathbb{R}^{m\times n},\ (m > n) \text{ and } \mathbf{b} \in \mathbb{R}^m.$$

The system is overdetermined when there are more equations than unknowns, $m > n$. Usually there is no solution for overdetermined systems.
So we rather solve the problem

$$\min_{x\in\mathbb{R}^n} ||\mathbf{A}\mathbf{x} - \mathbf{b}||_p,$$

for a suitable p .

Mininimisation in 1-norm and ∞-norm is complicated since
$f(x) = ||\mathbf{A}\mathbf{x} - \mathbf{b}||_p$ is not differentiable for $p = 1$ and $p = \infty$. So we will use the 2-norm and this is called a least square problem.

The LS problem can be converted into an equivalent problem by using orthogonal transformations. The resulting problem is the following:

$$\min_{x\in\mathbb{R}^n} ||(\mathbf{Q}^T\mathbf{A})\mathbf{x} - (\mathbf{Q}^T\mathbf{b})||_2, \text{ where } \mathbf{A}, \mathbf{Q} \in \mathbb{R}^{m\times n},\ \mathbf{b} \in \mathbb{R}^m,\ \mathbf{Q}^T\mathbf{Q} = \mathbf{I}_m$$

Important property: If $\mathbf{x}^* \in \mathbb{R}^n$ solves

$$\min_{x\in\mathbb{R}^n} ||\mathbf{A}\mathbf{x} - \mathbf{b}||_2$$

i.e., $||\mathbf{A}\mathbf{x}^* - \mathbf{b}||_2$ is the minimum then

$$\mathbf{A}^T(\mathbf{b} - \mathbf{A}\mathbf{x}^*) = \mathbf{0}$$

Exercise Verify this formula using the matrices

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix}, \quad \mathbf{x}^* = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

Solution: To see this, we consider the minimum using calculus of the function

$$f(\mathbf{x}^*) = \frac{1}{2}||\mathbf{b} - \mathbf{A}\mathbf{x}^*||_2^2$$

Therefore

$$f(\mathbf{x}^*) = \frac{1}{2}[(b_1 - a_{11}x_1 - a_{12}x_2)^2 + (b_2 - a_{21}x_1 - a_{22}x_2)^2 + (b_3 - a_{31}x_1 - a_{32}x_2)^2]$$

and

$$\begin{aligned} \bigtriangledown f(\mathbf{x}) &= \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{bmatrix} \\ &= \begin{bmatrix} -a_{11}(b_1 - a_{11}x_1 + a_{12}x_2) - a_{21}(b_2 - a_{21}x_1 - a_{22}x_2) - a_{31}(b_3 - a_{31}x_1 - a_{32}x_2) \\ -a_{12}(b_1 - a_{11}x_1 - a_{22}x_2) - a_{22}(b_2 - a_{21}x_1 - a_{22}x_2) - a_{32}(b_3 - a_{31}x_1 - a_{32}x_2) \end{bmatrix} \\ &= -\begin{bmatrix} a_{11} & a_{21} & a_{31} \\ a_{12} & a_{22} & a_{32} \end{bmatrix} \begin{bmatrix} (b_1 - a_{11}x_1 + a_{12}x_2) \\ (b_2 - a_{21}x_1 + a_{22}x_2) \\ (b_3 - a_{31}x_1 + a_{32}x_2) \end{bmatrix} \\ &= -\mathbf{A}^T(\mathbf{b} - \mathbf{A}\mathbf{x}^*) \end{aligned}$$

We know that from calculus to miminise a $f(\mathbf{x})$ we equate it's derivative to zero and solve for $\mathbf{x}$, i.e.,

$$\mathbf{A}^T(\mathbf{b} - \mathbf{A}\mathbf{x}^*) = \mathbf{0}.$$

If $\mathbf{x} \in \mathbb{R}^n$, $r = \mathbf{b} - \mathbf{A}\mathbf{x}$ is called the residual.
The equations $\mathbf{A}^T(\mathbf{b} - \mathbf{A}\mathbf{x}^*) = \mathbf{0}$ are called normal equations.
The solution to the least square problem, denoted by $\mathbf{x}_{LS}$ is unique, with minimum 2-norm and minimum sum of squares ρ_{LS} such that

$$\rho_{LS}^2 = ||\mathbf{A}\mathbf{x}_{LS} - \mathbf{b}||_2^2,$$

3.2 Solving LS using SVD

Let $\mathbf{A} = \sum_{j=1}^{r} \sigma_j \mathbf{u}_j \mathbf{v}_j^T$ where $r = \text{rank}(\mathbf{A})$ and

$$\begin{aligned} \mathbf{U} &= [\mathbf{u}_1, \mathbf{u}_2, \cdots, \cdots, \mathbf{u}_m] \in \mathbb{R}^{m \times m} \\ \mathbf{V} &= [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n] \in \mathbb{R}^{n \times n} \end{aligned}$$

be SVD of $\mathbf{A} \in \mathbb{R}^{m\times n},\ m \geq n$. If $\mathbf{b} \in \mathbb{R}^m$ then

$$\mathbf{x}_{LS} = \sum_{j=1}^{r} \frac{1}{\sigma_j} \mathbf{u}_j^T \mathbf{b} \mathbf{v}_j$$

and

$$\rho_{LS}^2 = \sum_{j=r+1}^{m} (\mathbf{u}_j^T \mathbf{b})^2.$$

So if you know the SVD of $\mathbf{A}$ in an LS problem then you can solve it by using elements in the SVD decomposition.
Proof: Since We note since 2-norm is invariant under orthogonal transformations

$$\begin{aligned} ||\mathbf{A}\mathbf{x} - \mathbf{b}||_2^2 &= ||\mathbf{U}^T\mathbf{A}\mathbf{V}(\mathbf{V}^T\mathbf{x}) - \mathbf{U}^T\mathbf{b}||_2^2, \text{ since } \mathbf{V}\mathbf{V}^T \\ &= \sum_{j=1}^{r} (\sigma_j (\mathbf{V}^T\mathbf{x})_j - \mathbf{u}_j^T\mathbf{b})^2 + \sum_{j=r+1}^{m} (\mathbf{u}_j^T\mathbf{b})^2. \end{aligned}$$

So if $\mathbf{x}$ solves the LS problem then $||\mathbf{A}\mathbf{x} - \mathbf{b}||_2^2$ is minimum, so we can minimise the above sum by making the first sum zero. This can be done by taking

$$(\mathbf{V}^T\mathbf{x})_j = \frac{1}{\sigma_j} \mathbf{u}_j^T \mathbf{b}.$$

Therefore the minimum of the LS problem is the remaining second sum i.e.,

$$\rho_{LS}^2 = \sum_{j=r+1}^{m} (\mathbf{u}_j^T \mathbf{b})^2.$$

Example: Use SVD method to solve LS problem with

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 1 & 4 \\ 1 & 6 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

given that the SVD of $\mathbf{A}$ is

$$\mathbf{U} = \begin{bmatrix} -0.28 & 0.87 & 0.41 \\ -0.54 & 0.21 & -0.82 \\ -0.79 & -0.45 & 0.41 \end{bmatrix}, \quad \mathbf{V} = \begin{bmatrix} -0.21 & 0.98 \\ -0.98 & -0.21 \end{bmatrix}$$

and singular values are 7.65 and 0.64.
Solution:
Note that $\text{rank}(\mathbf{A}) = 2$.
The solution of the LS is

$$\begin{aligned} \mathbf{x}^* &= \sum_{j=1}^{2} \frac{1}{\sigma_j} \mathbf{u}_j^T \mathbf{b} \mathbf{v}_j = \frac{1}{7.65} \mathbf{u}_1^T \mathbf{b} \mathbf{v}_1 + \frac{1}{0.64} \mathbf{u}_2^T \mathbf{b} \mathbf{v}_2 \\ &= \frac{1}{7.65} [-0.28, -0.54, -0.79] \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} -0.21 \\ -0.98 \end{bmatrix} + \frac{1}{0.64} [0.87, 0.21, -0.45] \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} -0.98 \\ -0.21 \end{bmatrix} \\ &= \begin{bmatrix} -0.27 \\ -1.71 \end{bmatrix}. \end{aligned}$$

3.3 Solving LS problems by method of Normal equations

One of the most widely used methods for solving LS is the method of Normal equations. It involves solving the system

$$\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}$$

If $\text{rank}(\mathbf{A}) = n$ then the system is positive definite and has solution $x = x_{LS}$

Given $\mathbf{A} \in \mathbb{R}^{m\times n}$ with $\text{rank}(\mathbf{A}) = n$ and $\mathbf{b} \in \mathbb{R}^m$, the following steps can be used to solve an LS problem using Normal equations:

1. Calculate $\mathbf{C} = \mathbf{A}^T\mathbf{A}$.
2. Calculate $\mathbf{d} = \mathbf{A}^T\mathbf{b}$.
3. Compute the Cholesky factorisation of $\mathbf{C}$ i.e., find $\mathbf{G}$ such that $\mathbf{C} = \mathbf{G}\mathbf{G}^T$.
4. From $\mathbf{G}\mathbf{G}^T\mathbf{x} = \mathbf{A}^T\mathbf{b}$, let $\mathbf{y} = \mathbf{G}^T\mathbf{x}$ so that $\mathbf{G}\mathbf{y} = \mathbf{d}$.
5. Solve for $\mathbf{y}$ in $\mathbf{G}\mathbf{y} = \mathbf{d}$ using forward substitution.
6. Solve for $\mathbf{x}$ in $\mathbf{G}^T\mathbf{x} = \mathbf{y}$ using backward substitution.

Example: Solve LS problem with

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 1 & 4 \\ 1 & 6 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

using the method of Normal equations.
Solution: $\text{rank}(\mathbf{A}) = 2 = n$.

$$\mathbf{C} = \mathbf{A}^T\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 4 & 6 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 1 & 4 \\ 1 & 6 \end{bmatrix} = \begin{bmatrix} 3 & 12 \\ 12 & 56 \end{bmatrix}$$

$$\mathbf{d} = \mathbf{A}^T\mathbf{b} = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 4 & 6 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 12 \end{bmatrix}$$

The Cholesky decomposition of $\mathbf{C}$ yields

$$\mathbf{G} = \begin{bmatrix} 1.73205081 & 0. \\ 6.92820323 & 2.82842712 \end{bmatrix}.$$

Now

$$\mathbf{G}\mathbf{y} = \mathbf{d}$$

$$\begin{bmatrix} 1.73205081 & 0. \\ 6.92820323 & 2.82842712 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 12 \end{bmatrix}$$

$$\begin{aligned} y_1 &= 3/1.73205081 = 1.7320508051377546 \\ y_2 &= (12 - 1.7320508051377546 * 6.92820323)/2.82842712 \\ &= -1.25607397e - 15 \end{aligned}$$

Finally,

$$\mathbf{G}^T\mathbf{x} = \mathbf{y}$$
$$\begin{bmatrix} 1.73205081 & 6.92820323 \\ 0 & 2.82842712 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1.7320508051377546 \\ -1.25607397e-15 \end{bmatrix}$$
$$x_1 = 1$$
$$x_2 = -4.4408921e-16$$

Therefore $x_{LS} = (1, -4.4408921e-16)^T$.

3.4 Solving LS problems by classical Gram-Schmidt (CGS) algorithm

OB problem: Given independent vectors $a_1, a_2, \ldots, a_n \in \mathbb{R}^m$ find orthonormal basis for span$\{a_1, a_2, \ldots, a_n\}$.

Solution: if $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n] = \mathbf{QR}$ and $\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_m]$. From matrix multiplication, we know that

$$a_{pk} = \sum_{i=1}^{k} q_{pi} r_{ik}, \quad k = 1, 2, \ldots, n.$$

Therefore

$$\mathbf{a}_k = \sum_{i=1}^{k} r_{ik} \mathbf{q}_i, \quad k = 1, 2, \ldots, n. \tag{3}$$

By orthonormality of $\mathbf{q}_i$, $\mathbf{q}_i^T \mathbf{q}_j = 0$ when $i \neq j$ and $\mathbf{q}_i^T \mathbf{q}_i = 1$. Therefore

$$\begin{aligned} \mathbf{q}_i^T \mathbf{a}_k &= \mathbf{q}_i^T \sum_{j=1}^{k} r_{jk} \mathbf{q}_j \\ &= \sum_{j=1}^{k} r_{jk} \mathbf{q}_i^T \mathbf{q}_j \\ &= r_{ik} \mathbf{q}_i^T \mathbf{q}_i \\ &= r_{ik}, \text{ since } \mathbf{q}_i^T \mathbf{q}_i = 1 \end{aligned}$$

From (3)

$$\mathbf{a}_k = \sum_{i=1}^{k-1} r_{ik} \mathbf{q}_i + r_{kk} \mathbf{q}_k$$
$$\text{rearranging,} \quad \mathbf{q}_k = \frac{\mathbf{a}_k - \sum_{i=1}^{k-1} r_{ik} \mathbf{q}_i}{r_{kk}}$$

Let

$$\mathbf{z}_k = \mathbf{a}_k - \sum_{i=1}^{k-1} s_{ik}\mathbf{q}_i, \text{ where } s_{ik} = \mathbf{q}_i^T\mathbf{q}_k$$

and

$$r_{kk} = ||\mathbf{z}_k||_2^2 = \mathbf{z}_k^T\mathbf{z}_k.$$

Given $\mathbf{A} \in \mathbb{R}^{m\times n}$ then the following is the classical Gram-Schmidt (CGS) algorithm which computes the decomposition $\mathbf{A} = \mathbf{QR}$:

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

$$\begin{aligned}
&s_{ik} = \mathbf{q}_i^T\mathbf{q}_k, \qquad i = 1, 2 \ldots, k-1\\
&\mathbf{z}_k = \mathbf{a}_k - \sum_{i=1}^{k-1} s_{ik}\mathbf{q}_i\\
&r_{kk} = ||\mathbf{z}_k||_2^2\\
&\mathbf{q}_k = \mathbf{z}_k/r_{kk}\\
&r_{ik} = s_{ik}/r_{kk}, \qquad i = 1, 2 \ldots, k-1
\end{aligned}$$

3.5 Solving LS problems using QR decomposition

CGS computes $\mathbf{A} = \mathbf{Q}_1\mathbf{R}_1, \mathbf{Q}_1 \in \mathbb{R}^{m\times n}, \mathbf{R}_1 \in \mathbb{R}^{n\times n}$ where $\mathbf{Q}_1^T\mathbf{Q}_1 = \mathbf{I}_n$ and $\mathbf{R}_1$ is upper triangular.

Exercise: Show that the normal equations $(\mathbf{A}^T\mathbf{A})\mathbf{x} = \mathbf{A}^T\mathbf{b}$ becomes $\mathbf{R}_1\mathbf{x} = \mathbf{Q}_1^T\mathbf{b}$ using CGS.
Solution:

$$\begin{aligned}
\mathbf{A}^T\mathbf{A} &= (\mathbf{Q}_1\mathbf{R}_1)^T\mathbf{Q}_1\mathbf{R}_1\\
&= \mathbf{R}_1^T\mathbf{Q}_1^T\mathbf{Q}_1\mathbf{R}_1\\
&= \mathbf{R}_1^T\mathbf{I}_n\mathbf{R}_1\\
&= \mathbf{R}_1^T\mathbf{R}_1
\end{aligned}$$

Therefore

$$\begin{aligned}
(\mathbf{A}^T\mathbf{A})\mathbf{x} = \mathbf{A}^T\mathbf{b} \text{ implies } (\mathbf{R}_1^T\mathbf{R}_1)\mathbf{x} &= \mathbf{R}_1^T\mathbf{Q}_1^T\mathbf{b}\\
\text{implies} \qquad \mathbf{R}_1\mathbf{x} &= \mathbf{Q}_1^T\mathbf{b}
\end{aligned}$$