4 Further Iterative Methods

4.1 Power Method for Matrix Eigenvalues

We discuss the problem of estimating the **dominant eigenvalue and its corresponding eigenvector** of a square matrix. Let the $n \times n$ matrix A satisfies:

(i) There is a **single eigenvalue** of maximum modulus.

Let the eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_n$ be labeled so that

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_n|.$$

(ii) To briefly discuss the idea, we assume that there is a set of n linearly independent unit eigenvectors. This means that there is a basis

$$\left\{\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \cdots, \mathbf{u}^{(n)}\right\}$$

for $\mathbb{R}^n$ such that

$$A\mathbf{u}^{(i)} = \lambda_i \mathbf{u}^{(i)}, \quad i = 1, 2, \cdots, n,$$

and $\|\mathbf{u}^{(i)}\| = 1$.

- Begin with an initial vector $\mathbf{x}^{(0)} \neq \mathbf{0}$, we write

$$\mathbf{x}^{(0)} = a_1\mathbf{u}^{(1)} + a_2\mathbf{u}^{(2)} + \cdots + a_n\mathbf{u}^{(n)}.$$

Here $\left\{\mathbf{u}^{(i)}\right\}$ is a basis (unit vector) for $\mathbb{R}^n$ and we assume that $a_1 \neq 0$.

- Now

$$\begin{aligned} A^k\mathbf{x}^{(0)} &= a_1A^k\mathbf{u}^{(1)} + \ldots + a_nA^k\mathbf{u}^{(n)} \\ &= a_1\lambda_1^k\mathbf{u}^{(1)} + \ldots + a_n\lambda_n^k\mathbf{u}^{(n)} \quad \boxed{\text{because } \mathrm{A}\mathbf{u}^{(\mathrm{i})} = \lambda_{\mathrm{i}}\mathbf{u}^{(\mathrm{i})}} \\ &= \lambda_1^k\left\{a_1\mathbf{u}^{(1)} + \left(\frac{\lambda_2}{\lambda_1}\right)^k a_2\mathbf{u}^{(2)} + \ldots + \left(\frac{\lambda_n}{\lambda_1}\right)^k a_n\mathbf{u}^{(n)}\right\}. \end{aligned}$$

- We remark that the convergent rate "**speed**" of the power method depends on the "gap" between $|\lambda_1|$ and $|\lambda_2|$. That is to say the **smaller the value of** $|\lambda_2|/|\lambda_1|$, the faster the convergence rate will be. Because one can observe that

$$1 > \left|\frac{\lambda_2}{\lambda_1}\right| \geq \left|\frac{\lambda_3}{\lambda_1}\right| \geq \cdots \geq \left|\frac{\lambda_n}{\lambda_1}\right|.$$

- Since

$$\frac{|\lambda_i|}{|\lambda_1|} < 1 \quad \text{for } i = 2, \ldots, n,$$

we have

$$\lim_{k\to\infty} \frac{|\lambda_i|^k}{|\lambda_1|^k} = 0 \quad \text{for } i = 2, \ldots, n.$$

Hence we have $A^k \mathbf{x}^{(0)} \approx a_1 \lambda_1^k \mathbf{u}^{(1)}$.

- Define

$$\mathbf{x}^{(k+1)} = \frac{A^{k+1}\mathbf{x}^{(0)}}{\|A^k \mathbf{x}^{(0)}\|} \quad \text{we have} \quad \mathbf{x}^{(k+1)} = \frac{A\mathbf{x}^{(k)}}{||\mathbf{x}^{(k)}||}.$$

We note that

$$\lim_{k\to\infty} ||\mathbf{x}^{(k+1)}|| = \lim_{k\to\infty} \frac{||a_1 \lambda_1^{k+1} \mathbf{u}^{(1)}||}{\|a_1 \lambda_1^k \mathbf{u}^{(1)}\|} = |\lambda_1|$$

where $\|\cdot\|$ can be $\|\cdot\|_1$, $\|\cdot\|_2$ or $\|\cdot\|_\infty$. Therefore we have

$$\lim_{k\to\infty} \frac{\mathbf{x}^{(k+1)}}{\|\mathbf{x}^{(k+1)}\|} = \mathbf{u}^{(1)},$$

λ_1 can be found by comparing $A\mathbf{u}^{(1)}$ and $\mathbf{u}^{(1)}$.

Example 4.1

$$A = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix} \quad \text{with initial guess} \quad \mathbf{x}^{(0)} = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

We take the vector norm $\|\cdot\|$ to be $\|\cdot\|_2$ and iterate four times to get an estimate of the largest eigenvalue and the corresponding unit eigenvector.

$$\mathbf{r}^{(1)} = A\mathbf{x}^{(0)} = [1.7321, 2.3094, 1.7321]^T, \quad \mathbf{x}^{(1)} = \frac{\mathbf{r}^{(1)}}{||\mathbf{r}^{(1)}||_2} = [0.5145, 0.6860, 0.5145];$$
$$\mathbf{r}^{(2)} = A\mathbf{x}^{(1)} = [1.7150, 2.4010, 1.7150]^T, \quad \mathbf{x}^{(2)} = \frac{\mathbf{r}^{(2)}}{||\mathbf{r}^{(2)}||_2} = [0.5025, 0.7035, 0.5025];$$
$$\mathbf{r}^{(3)} = A\mathbf{x}^{(2)} = [1.7086, 2.4121, 1.7086]^T, \quad \mathbf{x}^{(3)} = \frac{\mathbf{r}^{(3)}}{||\mathbf{r}^{(3)}||_2} = [0.5004, 0.7065, 0.5004];$$
$$\mathbf{r}^{(4)} = A\mathbf{x}^{(3)} = [1.7074, 2.4139, 1.7074]^T, \quad \mathbf{x}^{(4)} = \frac{\mathbf{r}^{(4)}}{||\mathbf{r}^{(4)}||_2} = [0.5001, 0.7070, 0.5001];$$

$$\|\mathbf{r}_1\|_2 = 3.3665, \quad \|\mathbf{r}_2\|_2 = 3.4128, \quad \|\mathbf{r}_3\|_2 = 3.4142 \quad \text{and} \quad \|\mathbf{r}_4\|_2 = 3.4142.$$

- Therefore $\lambda_1 \approx 3.4142$ and $\mathbf{u}^{(1)} \approx [0.5001, 0.7070, 0.5001]^T$.

For the stopping criteria, one may consider $||\mathbf{x}^{(n+1)} - \mathbf{x}^{(n)}||_2 < 10^{-6}$.

MATLAB CODE (Power Method)

```
A=[2 1 0;1 2 1;0 1 2];
error=1;
r=ones(3,1)/3^(0.5); x=r; y=r;
k=0;
while error > 10^(-6)
    y=A*x;
    r=y/norm(y);
    error=norm(x-r);
    x=r;
    k=k+1;
end;
```

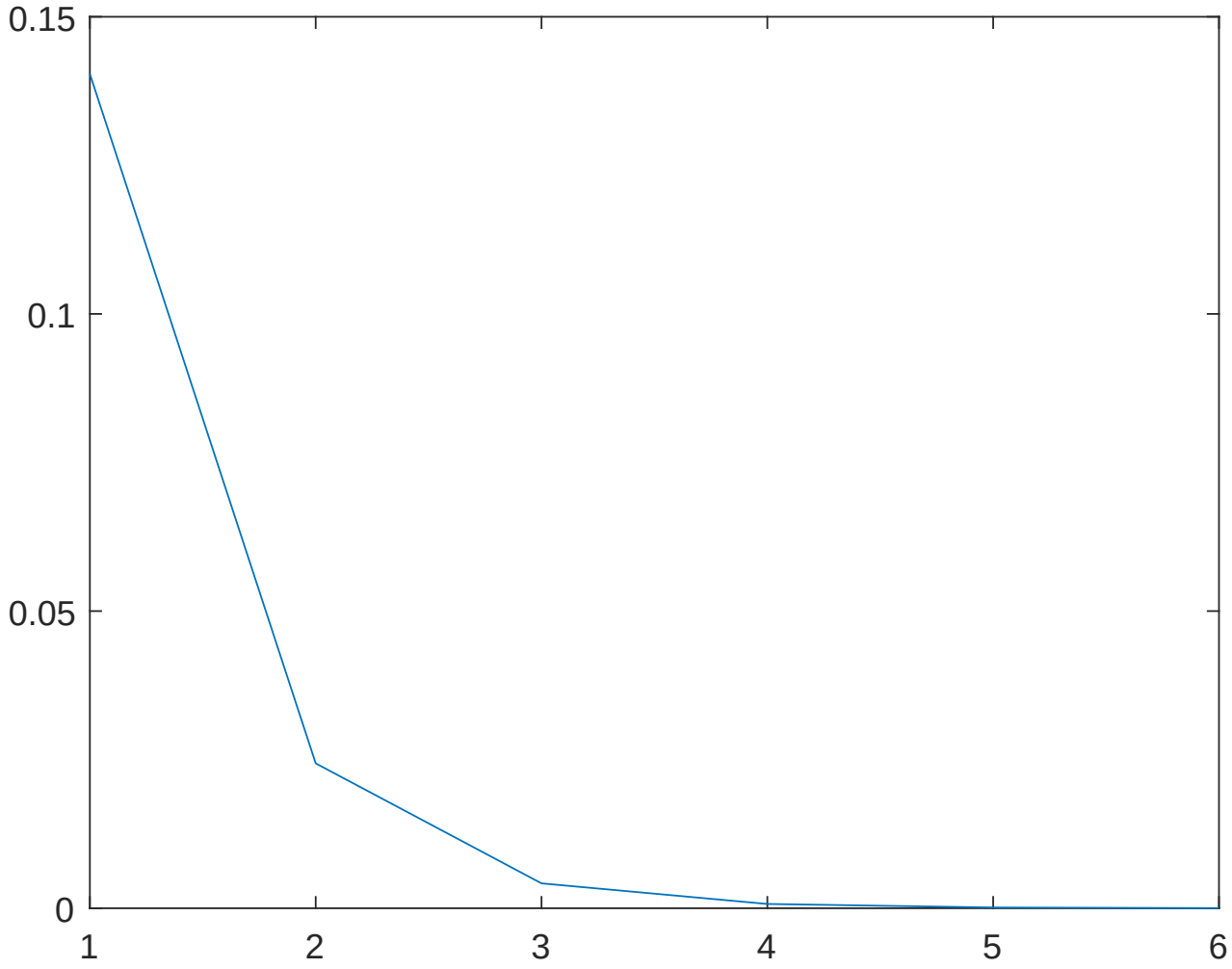

Figure 1: A Plot of $||\mathbf{x}^{(n+1)} - \mathbf{x}^{(n)}||_2$ $(n = 1, 2, 3, 4, 5)$.

• The eignevalues and eigenvectors of A are $3.4142 > 2.0000 > 0.5858$ and

$$[0.5000, 0.7071, 0.5000]^T, [0.7071, 0.0000, -0.7071]^T, [0.5000, -0.7071, 0.5000]^T,$$

respectively.

• Suppose we have obtained $\lambda_1 = 3.4142$ and $\mathbf{u}_1 = [0.5000, 0.7071, 0.5000]^T$, can we estimate λ_2 by using the power method assuming that $\lambda_2 > \lambda_3$?

• It is possible if we begin with $\mathbf{x}^{(0)} = a_1\mathbf{u}^{(1)} + a_2\mathbf{u}^{(2)} + \cdots + a_n\mathbf{u}^{(n)}$ such that $a_1 = 0$ and $a_2 \neq 0$.

• In this example, we may choose a random $\mathbf{x}_0 = [x_1, x_2, x_3]^T$ orthogonal to $\mathbf{u}_1$, i.e., $0.5x_1 + 0.7071x_2 + 0.5x_3 = 0$. We may try $\mathbf{x}_0 = [1, 0.7071, -2]^T$.

$\mathbf{r}_1 = [2.7071, 0.4142, -3.2929]^T$, $\mathbf{x}_1 = [0.6321, 0.0967, -0.7689]^T$;
$\mathbf{r}_2 = [1.3609, 0.0566, -1.4410]^T$, $\mathbf{x}_2 = [0.6863, 0.0286, -0.7267]^T$;
$\vdots$
$\mathbf{r}_7 = [1.4140, -0.0000, -1.4144]^T$, $\mathbf{x}_7 = [0.7070, 0.0000, -0.7072]^T$
$\lambda_2 \approx ||\mathbf{r}_7||_2 = 2.0000.$

4.1.1 Inverse Power Method

Suppose the eigenvalues of an $n \times n$ matrix A satisfy

$$|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_{n-1}| > |\lambda_n| > 0$$

How to calculate λ_n?

Theorem 4.1. *We note that if λ is an eigenvalue of A (A is non-singular) then λ^{-1} is an eigenvalue of A^{-1}.*

Proof Let $A\mathbf{x} = \lambda\mathbf{x}, \mathbf{x} \neq \mathbf{0}$ then $\mathbf{x} = A^{-1}(\lambda\mathbf{x})$ and therefore

$$\lambda^{-1}\mathbf{x} = A^{-1}\mathbf{x}.$$

This implies λ^{-1} is an eigenvalue of A^{-1}. Now we have

$$|\lambda_n^{-1}| > |\lambda_{n-1}^{-1}| \geq \cdots \geq |\lambda_1^{-1}|$$

and they are eigenvalues of A^{-1}.

- Therefore we may apply the power method to A^{-1} and get λ_n^{-1} and hence λ_n.

Example 4.2

Consider $A = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix}$ and $\mathbf{x}^{(0)} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$

$$\mathbf{x}^{(1)} = A^{-1}\mathbf{x}^{(0)}/||\mathbf{x}^{(0)}||_2 = [0.2887, 0.0000, 0.2887]^T, \quad ||\mathbf{x}^{(1)}||_2 = 0.4082$$
$$\mathbf{x}^{(2)} = [0.7071, -0.7071, 0.7071]^T, \quad ||\mathbf{x}^{(2)}||_2 = 1.2247$$
$$\vdots$$
$$\mathbf{x}^{(8)} = [0.8536, -1.2071, 0.8536]^T, \quad ||\mathbf{x}^{(8)}||_2 = 1.7071$$

- Hence

$$\lambda_3 \approx \frac{1}{1.7071} = 0.5858$$

and its corresponding eigenvector is

$$[0.5001, -0.7071, 0.5001]^T.$$

4.1.2 Shifted Matrix Method

Another problem is to compute the eigenvalue of A closest to a given value μ.

• We suppose that one eigenvalue of A, let say λ_k satisfies

$$\boxed{0 < |\lambda_k - \mu| < \varepsilon}$$

and all the other eigenvalues of A satisfy

$$\boxed{|\lambda_i - \mu| > \varepsilon.}$$

• The idea is to consider the matrix $(A - \mu I)$, (eigenvalues of $(A - \mu I)$ are $\lambda_i - \mu$) and apply the inverse power method to $(A - \mu I)$ and get the smallest eigenvalue

$$z = \lambda_k - \mu$$

and hence

$$\lambda_k = z + \mu.$$

4.1.3 Finding the Dominant Root of a Polynomial

Suppose it is known that the roots of the equation

$$\boxed{f(x) = x^n + c_{n-1}x^{n-1} + c_{n-2}x^{n-2} + \cdots + c_0 = 0}$$

satisfy $|r_1| > |r_2| \geq \cdots \geq |r_n|$.

- We then consider

$$A_n = \begin{bmatrix} 0 & \cdots & \cdots & \cdots & 0 & -c_0 \\ 1 & 0 & \cdots & \cdots & 0 & -c_1 \\ 0 & 1 & \ddots & & \vdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \vdots & & \ddots & \ddots & 0 & \vdots \\ 0 & \cdots & \cdots & 0 & 1 & -c_{n-1} \end{bmatrix}$$

and

$$\det(xI_n - A_n) = f(x).$$

- The power method can be used for solving the **largest** root in modulus.

4.2 Condition Number of a Matrix

• Consider the linear system

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

where A is an $n \times n$ invertible matrix. We shall analyze the **error** in the solution $\mathbf{x}$ due to a small **perturbation** of $\mathbf{b}$.

• If $\mathbf{b}$ is perturbed (change a bit) to $\widetilde{\mathbf{b}}$ and hence the solution will also perturb to $\widetilde{\mathbf{x}}$ such that

$$A\widetilde{\mathbf{x}} = \widetilde{\mathbf{b}}.$$

• The error $\mathbf{e}$ in $\mathbf{x}$ can be obtained as follows:

$$\begin{aligned} A(\mathbf{x} - \widetilde{\mathbf{x}}) &= \mathbf{b} - \widetilde{\mathbf{b}} \\ \Longrightarrow \mathbf{e} = (\mathbf{x} - \widetilde{\mathbf{x}}) &= A^{-1}(\mathbf{b} - \widetilde{\mathbf{b}}) \end{aligned}$$

- From

$$\begin{aligned}\|\mathbf{x}-\widetilde{\mathbf{x}}\| &\le \|A^{-1}\|_M\|\mathbf{b}-\widetilde{\mathbf{b}}\| \\ &= \|A^{-1}\|_M \times \boxed{\frac{\|A\mathbf{x}\|}{\|\mathbf{b}\|}} \times \|\mathbf{b}-\widetilde{\mathbf{b}}\| \\ &\le \|A^{-1}\|_M\|A\|_M\frac{\|\mathbf{x}\|}{\|\mathbf{b}\|}\|\mathbf{b}-\widetilde{\mathbf{b}}\|\end{aligned}$$

- We obtain the relative perturbation:

$$\boxed{\frac{\|\mathbf{x}-\widetilde{\mathbf{x}}\|}{\|\mathbf{x}\|}} \le \kappa(A)\boxed{\frac{\|\mathbf{b}-\widetilde{\mathbf{b}}\|}{\|\mathbf{b}\|}} \tag{4.1}$$

Relative error

where the **condition number** of A is

$$\boxed{\kappa(A) = \|A^{-1}\|_M\|A\|_M}.$$

It tells us that the relative error in $\mathbf{x}$ is no greater than $\kappa(A)$ times the relative error in $\mathbf{b}$.

• The value of condition number depends on the specific matrix norms. But the following is always true.

Theorem 4.2. $\kappa(A) \geq 1$ *for any square matrix A.*

Proof It is due to the fact that

$$\|A^{-1}\|_M \cdot \|A\|_M \geq \|A^{-1} \cdot A\|_M = \|I\|_M = 1.$$

□

• The approximate solution $\widetilde{\mathbf{x}}$ to

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

is obtained from

$$A\widetilde{\mathbf{x}} = \widetilde{\mathbf{b}}.$$

The **residual vector**:

$$\mathbf{r} = \mathbf{b} - A\widetilde{\mathbf{x}} = \mathbf{b} - \widetilde{\mathbf{b}}$$

measures how $A\widetilde{\mathbf{x}}$ is close to $\mathbf{b}$.

• The difference between the exact solution $\mathbf{x}$ and the approximate solution $\widetilde{\mathbf{x}}$ is called the **error vector**:

$$\mathbf{e} = \mathbf{x} - \widetilde{\mathbf{x}}.$$

The relationship

$$A\mathbf{e} = \mathbf{r}$$

between the error vector and the residual vector is of fundamental importance.

• The relative error in $\mathbf{x}$

$$\frac{\|\mathbf{x} - \widetilde{\mathbf{x}}\|}{\|\mathbf{x}\|} = \frac{\|\mathbf{e}\|}{\|\mathbf{x}\|}$$

can also be bounded from below, again with the help of the condition number. The following theorem is left as an exercise in the assignment.

Theorem 4.3. *In solving systems of linear equations $A\mathbf{x} = \mathbf{b}$, the following inequality holds*

$$\boxed{\frac{1}{\kappa(A)}\frac{\|\mathbf{r}\|}{\|\mathbf{b}\|} \leq \frac{\|\mathbf{e}\|}{\|\mathbf{x}\|} \leq \kappa(A)\frac{\|\mathbf{r}\|}{\|\mathbf{b}\|}}$$

• A matrix with a large condition number is said to be **ill conditioned**. In this case, the solution of a system $A\mathbf{x} = \mathbf{b}$ may be very sensitive to small changes in the vector $\mathbf{b}$. If the condition number of A is of moderate size, the matrix is said to be **well conditioned**.

• If $\kappa(A)$ is very large, then $\|\mathbf{e}\|$ can be **very large** even if $\|\mathbf{r}\|$ is **small**.

Example 4.3 For the linear system

$$\underbrace{\begin{bmatrix} 10^{10} & 0 \\ 0 & 10^{-10} \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} 0 \\ 1 \end{bmatrix}}_{\mathbf{x}} = \underbrace{\begin{bmatrix} 0 \\ 10^{-10} \end{bmatrix}}_{\mathbf{b}}.$$

we consider the approximate solution

$$\widetilde{\mathbf{x}} = \begin{bmatrix} 0 \\ 10^{k} \end{bmatrix} \quad \text{where } k > 0.$$

- We have

$$\mathbf{r} = \mathbf{b} - A\widetilde{\mathbf{x}} = \begin{bmatrix} 0 \\ 10^{-10} \end{bmatrix} - \begin{bmatrix} 0 \\ 10^{k-10} \end{bmatrix} = \begin{bmatrix} 0 \\ 10^{-10} - 10^{k-10} \end{bmatrix}$$

and hence

$$\|\mathbf{r}\|_\infty = |10^{k-10} - 10^{-10}| \leq 10^{k-10}.$$

- But

$$\mathbf{e} = \mathbf{x} - \widetilde{\mathbf{x}} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} - \begin{bmatrix} 0 \\ 10^{k} \end{bmatrix} = \begin{bmatrix} 0 \\ -10^{k} + 1 \end{bmatrix}$$

where

$$\|\mathbf{e}\|_\infty = |10^k - 1|.$$

In particular, we have for $k = 5$, we have

$$\|\mathbf{r}\|_\infty \leq 10^{-5} \quad \text{but} \quad \|\mathbf{e}\|_\infty \geq 10^5 - 1.$$

4.3 Gershgorin's Theorem

Proposition 1. (The Gershgorin's theorem) The eigenvalues of an $n \times n$ matrix A are contained in the union of the following n disks D_i where

$$\boxed{D_i = \{z \in C : |z - a_{ii}| \leq -|a_{ii}| + \sum_{j=1}^{n} |a_{ij}|\}.}$$

Proof Let λ be an eigenvalue of A and $\mathbf{x}$ be its corresponding eigenvector such that $||\mathbf{x}||_\infty = |x_i| = 1$. This can be done by dividing $\mathbf{x}$ by $\max\{|x_i|\}$. Since $A\mathbf{x} = \lambda\mathbf{x}$ we have

$$\lambda x_i = \sum_{j=1}^{n} a_{ij} x_j$$

and therefore

$$(\lambda - a_{ii}) x_i = \sum_{j=1, j \neq i}^{n} a_{ij} x_j.$$

Hence

$$|\lambda - a_{ii}| = |(\lambda - a_{ii})x_i| \leq \sum_{j=1, j\neq i}^{n} |a_{ij}x_j| \leq \sum_{j=1, j\neq i}^{n} |a_{ij}|.$$

Therefore $\lambda \in D_i$.

Example 4.4

$$A = \begin{bmatrix} \boxed{2i} & 1 & 1 \\ 1 & \boxed{4} & 1 \\ 0 & i & \boxed{3} \end{bmatrix}$$

$D_1 = \{z : |z - 2i| \leq 2\}$

$D_2 = \{z : |z - 4| \leq 2\}$

$D_3 = \{z : |z - 3| \leq 1\}$

$\lambda_1 = -0.2300 + 1.9382i$

$\lambda_2 = 4.4925 + 0.7269i$

$\lambda_3 = 2.7371 - 0.6651i$

- We observe that $\lambda_i \in D_i \quad i = 1, 2, 3.$

Proposition 2. If Q is a column stochastic matrix then $\rho(Q^k) = 1$ (non-negative matrix and all column sums are one).

Proof We note that

$$\mathbf{1}Q = \mathbf{1}$$

where $\mathbf{1} = [1, 1, \ldots, 1]$.

• Therefore

$$\mathbf{1}Q^k = \mathbf{1}.$$

This means that 1 is an eigenvalue of Q^k. Thus we conclude that $\rho(Q^k) \geq 1$.

• By using the Gershgorin's theorem and the fact that all the entries of Q^k are non-negative, all the column sums of Q^k are equal to one, we have $\rho(Q^k) \leq 1$.

• Hence we conclude that $\rho(Q^k) = 1$.

Theorem 4.4. *Let $\lambda_1, \cdots, \lambda_n$ be the eigenvalues of an $n \times n$ matrix A and there exists P such that*

$$P^{-1}AP = \textit{Diag}\,[\lambda_1\ \lambda_2 \cdots \lambda_n] = D.$$

Let B be any $n \times n$ matrix. Then the eigenvalues of $A + B$ lie in the Union of all the disks D_i

$$D_i = \{\lambda \in \mathbb{C} : |\lambda - \lambda_i| \leq \kappa(P)\|B\|_\infty\}.$$

Here $\kappa(P)$ is the condition number of P

Proof We note that $A + B$ and

$$P^{-1}(A + B)P = D + P^{-1}BP$$

have the same set of eigenvalues.

Let $C = P^{-1}BP$, then the eigenvalues of $(D + C)$ lie in the union of all the disks D_i where

$$D_i = \left\{ \lambda \in \mathbb{C} : |\lambda - \lambda_i - c_{ii}| \leq \sum_{j=1, j\neq i}^{n} |d_{ij} + c_{ij}| = \sum_{j=1, j\neq i}^{n} |c_{ij}| \right\}.$$

To show that

$$|\lambda - \lambda_i| \leq \kappa(P)\|B\|_\infty,$$

we note that

$$\begin{aligned}
|\lambda - \lambda_i| &\leq |\lambda - \lambda_i - c_{ii}| + |c_{ii}| \\
&\leq |c_{ii}| + \sum_{j=1, j\neq i}^{n} |c_{ij}| \\
&= \sum_{j=1}^{n} |c_{ij}| \\
&\leq \|C\|_\infty \\
&\leq \|P^{-1}\|_\infty \|B\|_\infty \|P\|_\infty \\
&= \kappa(P)\|B\|_\infty.
\end{aligned}$$

4.4 Steepest Descent Method

- We consider the problem of solving

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

such that

(1) A is an $n \times n$ matrix;
(2) A is **symmetric**, i.e., $A^T = A$;
(3) A is **positive definite**, i.e., $\mathbf{x}^T A\mathbf{x} > 0$ for $\mathbf{x} \neq \mathbf{0}$.

Remark Condition (3) above implies that A^{-1} exists.

- Recall the properties of the **inner product** in $\mathbb{R}^n$:

$$\boxed{< \mathbf{x}, \mathbf{y} >= \mathbf{x}^T\mathbf{y} = \sum_{i=1}^{n} x_i y_i.}$$

(i) $< \mathbf{x}, \mathbf{y} >=< \mathbf{y}, \mathbf{x} >$;

(ii) $< \alpha\mathbf{x}, \mathbf{y} >= \alpha < \mathbf{x}, \mathbf{y} >$;

(iii) $< \mathbf{x} + \mathbf{y}, \mathbf{z} >=< \mathbf{x}, \mathbf{z} > + < \mathbf{y}, \mathbf{z} >$;

(iv) $< \mathbf{x}, A\mathbf{y} >=< A^T\mathbf{x}, \mathbf{y} >$.

Proposition 3. If A is **symmetric positive definite**, then the problem of solving $A\mathbf{x} = \mathbf{b}$ is equivalent to the problem of minimizing

$$\boxed{q(\mathbf{x}) = < \mathbf{x}, A\mathbf{x} > -2 < \mathbf{x}, \mathbf{b} > .}$$

Proof Let $\mathbf{v}$ be a vector and t be a scalar. We consider the function

$$\begin{aligned} q(\mathbf{x}+t\mathbf{v}) &= < \mathbf{x}+t\mathbf{v}, A(\mathbf{x}+t\mathbf{v}) > -2 < \mathbf{x}+t\mathbf{v}, \mathbf{b} > \\ &= < \mathbf{x}, A\mathbf{x} > +t < \mathbf{x}, A\mathbf{v} > +t < \mathbf{v}, A\mathbf{x} > \\ &\quad +t^2 < \mathbf{v}, A\mathbf{v} > -2 < \mathbf{x}, \mathbf{b} > -2t < \mathbf{v}, \mathbf{b} > \\ &= q(\mathbf{x}) + 2t < \mathbf{v}, A\mathbf{x} > - 2t < \mathbf{v}, \mathbf{b} > +t^2 < \mathbf{v}, A\mathbf{v} > \\ &= q(\mathbf{x}) + 2t < \mathbf{v}, A\mathbf{x} - \mathbf{b} > +t^2 < \mathbf{v}, A\mathbf{v} > . \end{aligned}$$

- Now one can regard it as a function of t

$$q(\mathbf{x}+t\mathbf{v}) = f(t) = q(\mathbf{x}) + 2 < \mathbf{v}, A\mathbf{x}-\mathbf{b} > t + < \mathbf{v}, A\mathbf{v} > t^2.$$

- In fact, it is a quadratic function in t. Moreover $f(t)$ attains its **minimum** at t s.t. $f'(t) = 0$, i.e.,

$$2 < \mathbf{v}, A\mathbf{x}-\mathbf{b} > +2 < \mathbf{v}, A\mathbf{v} > t = 0.$$

Solving the equation we have

$$t^* = \frac{< \mathbf{v}, \mathbf{b}-A\mathbf{x} >}{< \mathbf{v}, A\mathbf{v} >}.$$

We remark that $< \mathbf{v}, A\mathbf{v} > \neq 0$ because A is positive definite.

- Therefore

$$\begin{aligned} q(\mathbf{x}+t^*\mathbf{v}) &= q(\mathbf{x}) + t^*\{2 < \mathbf{v}, A\mathbf{x}-\mathbf{b} > + < \mathbf{v}, A\mathbf{v} > t^*\} \\ &= q(\mathbf{x}) + t^*\{2 < \mathbf{v}, A\mathbf{x}-\mathbf{b} > + < \mathbf{v}, \mathbf{b}-A\mathbf{x} >\} \\ &= q(\mathbf{x}) + t^*\{< \mathbf{v}, A\mathbf{x}-\mathbf{b} >\} \\ &= q(\mathbf{x}) - \underbrace{\frac{<\mathbf{v},\mathbf{b}-A\mathbf{x}>^2}{< \mathbf{v}, A\mathbf{v} >}}_{\longleftarrow \textbf{non-negative}}. \end{aligned}$$

• We note that reduction in the value of $q(\mathbf{x})$ always occurs in passing from $\mathbf{x}$ to $\mathbf{x} + t^*\mathbf{v}$ (unless $< \mathbf{v}, \mathbf{b} - A\mathbf{x} >= 0$, in this case $\mathbf{v}$ is orthogonal to $\mathbf{b} - A\mathbf{x}$).

• This means that if

$$\mathbf{b} - A\mathbf{x} \neq \mathbf{0}$$

then we can find a vector $\widetilde{\mathbf{v}}$ such that

$$< \widetilde{\mathbf{v}}, \mathbf{b} - A\mathbf{x} >\neq 0 \quad \text{and} \quad q(\mathbf{x} + t^*\mathbf{v}) < q(\mathbf{x})$$

and $\mathbf{x}$ is **NOT** the minimizer of $q(\mathbf{x})$.

• If $\mathbf{b} - A\mathbf{x} = \mathbf{0}$ then $q(\mathbf{x} + t^*\mathbf{v}) = q(\mathbf{x})$ for any vector $\mathbf{v}$. Therefore, $\mathbf{x}$ is the minimizer.

• One may design an iterative method for solving $A\mathbf{x} = \mathbf{b}$ by using the idea in Proposition 3. Given A, an $n \times n$ symmetric positive definite matrix and $\mathbf{b}$ is an $n \times 1$ vector.

• With an $\mathbf{x}_0$, an initial guess of the solution of $A\mathbf{x} = \mathbf{b}$ we develop an iterative algorithm namely the **steepest decent method**. The iterative method reads:

Input: Max, A, $\mathbf{b}$, $\mathbf{x}_0$, Error-tol and $k = 0$,
$\mathbf{r}_0 = \mathbf{b} - A\mathbf{x}_0$, initial residual.
While $\|\mathbf{r}_k\|_2 >$ Error-tol and $k <$ Max

$$\begin{aligned} t_k &= < \mathbf{r}_k, \mathbf{r}_k > / < \mathbf{r}_k, A\mathbf{r}_k >; \\ \mathbf{x}_{k+1} &= \mathbf{x}_k + t_k \cdot \mathbf{r}_k; \\ \mathbf{r}_k &= \mathbf{b} - A\mathbf{x}_k; \\ k &= k + 1; \end{aligned}$$

end

• We remark that $\mathbf{r}_k$ is the search direction and t_k is the step size. In the iterative method, $t = t^*$ in Proposition 3 by letting

$$\mathbf{v} = \mathbf{r} = \mathbf{b} - A\mathbf{x}.$$

Example 4.5

$$A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \quad , \quad \mathbf{b} = \begin{bmatrix} 4 \\ 5 \end{bmatrix} \quad , \quad \mathbf{x}_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

k	$\mathbf{x}_k$	t	$\lVert\mathbf{r}_k\rVert_2$
1	$[1.34\ 1.68]^T$	0.3361	6.4031
2	$[0.98\ 1.97]^T$	0.9762	0.4724
3	$[1.01\ 1.99]^T$	0.3361	0.1012
4	$[0.99\ 1.99]^T$	0.9762	0.0075
5	$[1.00\ 1.99]^T$	0.3361	0.0016
$\vdots$	$\vdots$	$\vdots$	$\vdots$

• The true solution is $[1, 2]^T$. This method, "**steepest descent**" is rarely used because its convergence rate is "too slow".

4.5 Conjugate Gradient Method

Definition 4.1. (A-orthonormality.) Assuming that A is an $n \times n$ symmetric positive definite matrix, suppose that a set of vectors $\{\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_n\}$ is provided and has the following property:

$$\langle \mathbf{u}_i, A\mathbf{u}_j \rangle = \delta_{ij}$$

where

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j. \end{cases}$$

• This property is called the **A-orthonormality**. Clearly it is a generalization of the ordinary orthonormality where $A = I_n$.

Remark Here $\|\mathbf{x}\|_A = \sqrt{< \mathbf{x}, A\mathbf{x} >}$ defines a norm in $\mathbb{R}^n$.

Proposition 4. Let $U = (\mathbf{u}_1, \cdots, \mathbf{u}_n)$ then $U^T A U = I_n$.

Proof It follows from the definition.

Proposition 5. The set $\{\mathbf{u}_1, \cdots, \mathbf{u}_n\}$ forms a basis for $\mathbb{R}^n$.

Proof We only need to show that $\{\mathbf{u}_i\}$ are independent.

Suppose

$$\sum_{i=1}^{n} \alpha_i \mathbf{u}_i = \mathbf{0}$$

then

$$\begin{aligned} 0 &= \left\langle \sum_{i=1}^{n} \alpha_i \mathbf{u}_i, A\mathbf{u}_j \right\rangle \quad j = 1, \cdots, n \\ &= \sum_{i=1}^{n} \alpha_i \langle \mathbf{u}_i, A\mathbf{u}_j \rangle \\ &= \alpha_j \langle \mathbf{u}_j, A\mathbf{u}_j \rangle = \alpha_j. \end{aligned}$$

Hence $\alpha_j = 0$ for $j = 1, \cdots, n$.

- This shows that $\{\mathbf{u}_i\}$ are independent and hence form a **basis** for $\mathbb{R}^n$.

Proposition 6. Let $\{\mathbf{u}_1, \cdots, \mathbf{u}_n\}$ be an A-orthonormal system. Define the following recursive scheme:

$$\boxed{\mathbf{x}_i = \mathbf{x}_{i-1} + \langle \mathbf{b} - A\mathbf{x}_{i-1}, \mathbf{u}_i \rangle \mathbf{u}_i}$$

for $i = 1, 2, \cdots, n$ iteratively in which $\mathbf{x}_0$ is an arbitrary vector in $\mathbb{R}^n$ then we have

$$A\mathbf{x}_n = \mathbf{b}.$$

Proof Define

$$t_i = \langle \mathbf{b} - A\mathbf{x}_{i-1}, \mathbf{u}_i \rangle .$$

The iterative method reads $\mathbf{x}_i = \mathbf{x}_{i-1} + t_i\mathbf{u}_i$. We note that $A\mathbf{x}_i = A\mathbf{x}_{i-1} + t_iA\mathbf{u}_i$. Therefore

$$\begin{aligned} A\mathbf{x}_n &= A\mathbf{x}_{n-1} + t_nA\mathbf{u}_n \\ &= A\mathbf{x}_{n-2} + t_{n-1}A\mathbf{u}_{n-1} + t_nA\mathbf{u}_n \\ &\vdots \quad \vdots \end{aligned}$$

- Finally, we have

$$A\mathbf{x}_n = A\mathbf{x}_0 + t_1 A\mathbf{u}_1 + \cdots + t_n A\mathbf{u}_n.$$

- Now

$$\langle A\mathbf{x}_n - \mathbf{b}, \mathbf{u}_i \rangle = \langle A\mathbf{x}_0 - \mathbf{b}, \mathbf{u}_i \rangle + t_i.$$

Since

$$\begin{aligned}
t_i &= \langle \mathbf{b} - A\mathbf{x}_{i-1}, \mathbf{u}_i \rangle \\
&= \left\langle \mathbf{b} - \underbrace{A\mathbf{x}_0 + A\mathbf{x}_0} - \underbrace{A\mathbf{x}_1 + A\mathbf{x}_1} + \cdots - A\mathbf{x}_{i-1}, \mathbf{u}_i \right\rangle \\
&= \langle \mathbf{b} - A\mathbf{x}_0, \mathbf{u}_i \rangle + \langle A\mathbf{x}_0 - A\mathbf{x}_1, \mathbf{u}_i \rangle \\
&\quad + \langle A\mathbf{x}_1 - A\mathbf{x}_2, \mathbf{u}_i \rangle + \cdots + \langle A\mathbf{x}_{i-2} - A\mathbf{x}_{i-1}, \mathbf{u}_i \rangle \\
&= \langle \mathbf{b} - A\mathbf{x}_0, \mathbf{u}_i \rangle + \langle -t_1 A\mathbf{u}_1, \mathbf{u}_i \rangle + \cdots + \langle -t_{i-1} A\mathbf{u}_{i-1}, \mathbf{u}_i \rangle \\
&= \langle \mathbf{b} - A\mathbf{x}_0, \mathbf{u}_i \rangle .
\end{aligned}$$

Hence

$$\langle A\mathbf{x}_n - \mathbf{b}, \mathbf{u}_i \rangle = 0, \quad i = 1, \cdots, n \quad \text{and} \quad A\mathbf{x}_n - \mathbf{b} = \mathbf{0}.$$

Because $A\mathbf{x}_n - \mathbf{b}$ is orthonormal to all $\mathbf{u}_i$ and it must be the zero vector.

Definition 4.2. (A-orthogonal). Assuming A is an $n \times n$ symmetric positive definite matrix, then a set of vectors

$$\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$$

is said to be A-orthogonal if

$$\langle \mathbf{v}_i, A\mathbf{v}_j \rangle = 0 \quad \text{whenever} \quad i \neq j.$$

Proposition 6 can be extended as follows.

Theorem 4.5. *Let $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ be an A-orthogonal system of non-zero vectors for a symmetric and positive definite $n \times n$ matrix A. Define*

$$\boxed{\mathbf{x}_i = \mathbf{x}_{i-1} + \frac{\langle \mathbf{b} - A\mathbf{x}_{i-1}, \mathbf{v}_i \rangle}{\langle \mathbf{v}_i, A\mathbf{v}_i \rangle}\mathbf{v}_i}$$

in which $\mathbf{x}_0$ is arbitrary, then $A\mathbf{x}_n = \mathbf{b}$.

- We note that $\langle \mathbf{v}_i, A\mathbf{v}_i \rangle = \|\mathbf{v}_i\|_A^2$.

- The CG algorithm reads:

Given an initial guess $\mathbf{x}_0$, A, $\mathbf{b}$, Max, tol:

$\mathbf{r}_0 = \mathbf{b} - A\mathbf{x}_0$;

$\mathbf{v}_0 = \mathbf{r}_0$;

For $k = 0$ to Max-1 do

If $||\mathbf{v}_k||_2 = 0$ then stop

$t_k =< \mathbf{r}_k, \mathbf{r}_k > / < \mathbf{v}_k, A\mathbf{v}_k >$;

$\mathbf{x}_{k+1} = \mathbf{x}_k + t_k\mathbf{v}_k$;

$\mathbf{r}_{k+1} = \mathbf{r}_k - t_k A\mathbf{v}_k$;

If $||\mathbf{r}_{k+1}||_2 < \mathrm{tol}$ then stop

$s_k = \frac{<\mathbf{r}_{k+1},\mathbf{r}_{k+1}>}{<\mathbf{r}_k,\mathbf{r}_k>}$;

$\mathbf{v}_{k+1} = \mathbf{r}_{k+1} + s_k\mathbf{v}_k$;

end;

output $\mathbf{x}_{k+1}, ||\mathbf{r}_{k+1}||_2$.

Theorem 4.6. *In the conjugate gradient algorithm, for any integer* $m \leq n$ *if* $v^{(0)}, v^{(1)}, \cdots, v^{(m)}$ *are all non-zero vectors, then*

(a) $\left\langle \mathbf{r}^{(m)}, \mathbf{v}^{(i)} \right\rangle = 0 \qquad (0 \leq i < m)$

(b) $\left\langle \mathbf{r}^{(i)}, \mathbf{r}^{(i)} \right\rangle = \left\langle \mathbf{r}^{(i)}, \mathbf{v}^{(i)} \right\rangle \qquad (0 \leq i \leq m)$

(c) $\left\langle \mathbf{v}^{(m)}, A\mathbf{v}^{(i)} \right\rangle = 0 \qquad (0 \leq i < m)$

(d) $\mathbf{r}^{(i)} = \mathbf{b} - A\mathbf{x}^{(i)} \qquad (0 \leq i \leq m)$

(e) $\left\langle \mathbf{r}^{(m)}, \mathbf{r}^{(i)} \right\rangle = 0 \qquad (0 \leq i < m)$

(f) $\mathbf{r}^{(i)} \neq \mathbf{0} \qquad (0 \leq i \leq m)$

Proof We shall prove them by Mathematical Induction.

For $m = 0$, we assume that $\mathbf{v}^{(0)} \neq 0$ and therefore

$$\mathbf{r}^{(0)} = \mathbf{b} - A\mathbf{x}^{(0)} = \mathbf{v}^{(0)} \neq 0.$$

Hence the statements are true for $m = 0$.

- We assume that the theorem is true for a certain m.

We shall prove it for $m + 1$. To this end we assume

$$\mathbf{v}^{(0)}, \mathbf{v}^{(1)}, \cdots, \mathbf{v}^{(m+1)}$$

are non-zero vectors.

(a')

$$\begin{aligned}
\left\langle \mathbf{r}^{(m+1)}, \mathbf{v}^{(m)} \right\rangle &= \left\langle \mathbf{r}^{(m)} - t_m A\mathbf{v}^{(m)}, \mathbf{v}^{(m)} \right\rangle \\
&= \left\langle \mathbf{r}^{(m)}, \mathbf{v}^{(m)} \right\rangle - t_m \left\langle \mathbf{v}^{(m)}, A\mathbf{v}^{(m)} \right\rangle \\
&= \left\langle \mathbf{r}^{(m)}, \mathbf{v}^{(m)} \right\rangle - \left\langle \mathbf{r}^{(m)}, \mathbf{r}^{(m)} \right\rangle = 0.
\end{aligned}$$

Moreover $\left\langle \mathbf{r}^{(m+1)}, \mathbf{v}^{(i)} \right\rangle = \left\langle \mathbf{r}^{(m)}, \mathbf{v}^{(i)} \right\rangle - t_m \left\langle \mathbf{v}^{(m)}, A\mathbf{v}^{(i)} \right\rangle$

$= 0$ by (a) and (c).

Hence $\left\langle \mathbf{r}^{(m+1)}, \mathbf{v}^{(i)} \right\rangle = 0$ for $0 \le i \le m$.

(b′) Using (a′) and $\mathbf{v}^{(m+1)} = \mathbf{r}^{(m+1)} + S_m \mathbf{v}^{(m)}$, we have

$$\begin{aligned}
\left\langle \mathbf{r}^{(m+1)}, \mathbf{v}^{(m+1)} \right\rangle &= \left\langle \mathbf{r}^{(m+1)}, \mathbf{r}^{(m+1)} + S_m \mathbf{v}^{(m)} \right\rangle \\
&= \left\langle \mathbf{r}^{(m+1)}, \mathbf{r}^{(m+1)} \right\rangle.
\end{aligned}$$

Hence by (b) we have

$$\left\langle \mathbf{r}^{(i)}, \mathbf{r}^{(i)} \right\rangle = \left\langle \mathbf{r}^{(i)}, \mathbf{v}^{(i)} \right\rangle \quad \text{for} \quad 0 \le i \le m.$$

(c') Define $S_{-1} = 0$ and $\mathbf{v}^{(-1)} = \mathbf{0}$.

Now $\left\langle \mathbf{v}^{(m+1)}, A\mathbf{v}^{(i)} \right\rangle \quad (i < m)$

$$
\begin{aligned}
&= \left\langle \mathbf{r}^{(m+1)} + S_m \mathbf{v}^{(m)}, A\mathbf{v}^{(i)} \right\rangle = \left\langle \mathbf{r}^{(m+1)}, A\mathbf{v}^{(i)} \right\rangle + S_m \left\langle \mathbf{v}^{(m)}, A\mathbf{v}^{(i)} \right\rangle \\
&= t_i^{-1} \left\langle \mathbf{r}^{(m+1)}, \mathbf{r}^{(i)} - \mathbf{r}^{(i+1)} \right\rangle + S_m \left\langle \mathbf{v}^{(m)}, A\mathbf{v}^{(i)} \right\rangle \\
&= t_i^{-1} \left\langle \mathbf{r}^{(m+1)}, \mathbf{v}^{(i)} - S_{i-1}\mathbf{v}^{(i-1)} - \mathbf{v}^{(i+1)} + S_i \mathbf{v}^{(i)} \right\rangle + S_m \left\langle \mathbf{v}^{(m)}, A\mathbf{v}^{(i)} \right\rangle \\
&= t_i^{-1} \left\{ \left\langle \mathbf{r}^{(m+1)}, \mathbf{v}^{(i)} \right\rangle - S_{i-1} \left\langle \mathbf{r}^{(m+1)}, \mathbf{v}^{(i-1)} \right\rangle - \left\langle \mathbf{r}^{(m+1)}, \mathbf{v}^{(i+1)} \right\rangle + S_i \left\langle \mathbf{r}^{(m+1)}, \mathbf{v}^{(i)} \right\rangle \right\} \\
&\quad + S_m \left\langle \mathbf{v}^{(m)}, A\mathbf{v}^{(i)} \right\rangle .
\end{aligned}
$$

- We note if $i < m$ then by (a′) $\left\langle \mathbf{r}^{(m+1)}, \mathbf{v}^{(i)} \right\rangle = 0$ and by (c) we conclude $\left\langle \mathbf{v}^{(m+1)}, A\mathbf{v}^{(i)} \right\rangle = 0$.

- For the case when $i = m$, we have $\left\langle \mathbf{v}^{(m+1)}, A\mathbf{v}^{(m)} \right\rangle$

$$
= t_m^{-1} \left\langle \mathbf{r}^{(m+1)}, \mathbf{v}^{(m)} - S_{m-1}\mathbf{v}^{(m-1)} - \mathbf{v}^{(m+1)} + S_m \mathbf{v}^{(m)} \right\rangle + S_m \left\langle \mathbf{v}^{(m)}, A\mathbf{v}^{(m)} \right\rangle
$$

By (a′) $\left\langle \mathbf{r}^{(m+1)}, \mathbf{v}^{(m)} \right\rangle = \left\langle \mathbf{r}^{(m+1)}, \mathbf{v}^{(m-1)} \right\rangle = 0.$

Hence, $\left\langle \mathbf{v}^{(m+1)}, A\mathbf{v}^{(m)} \right\rangle = t_m^{-1} \left\langle \mathbf{r}^{(m+1)}, \mathbf{v}^{(m+1)} \right\rangle + S_m \left\langle \mathbf{v}^{(m)}, A\mathbf{v}^{(m)} \right\rangle$

$$\begin{aligned} &= -\frac{\left\langle \mathbf{v}^{(m)}, A\mathbf{v}^{(m)} \right\rangle}{\left\langle \mathbf{r}^{(m)}, \mathbf{r}^{(m)} \right\rangle} \left\langle \mathbf{r}^{(m+1)}, \mathbf{v}^{(m+1)} \right\rangle + \frac{\left\langle \mathbf{r}^{(m+1)}, \mathbf{r}^{(m+1)} \right\rangle}{\left\langle \mathbf{r}^{(m)}, \mathbf{r}^{(m)} \right\rangle} \left\langle \mathbf{v}^{(m)}, A\mathbf{v}^{(m)} \right\rangle \\ &= 0. \end{aligned}$$

- Therefore, $\left\langle \mathbf{v}^{(m)}, A\mathbf{v}^{(i)} \right\rangle = 0$ for $0 \le i < m+1$

(d')

$$\begin{aligned} \mathbf{b} - A\mathbf{x}^{(m+1)} &= \mathbf{b} - A\left(\mathbf{x}^{(m)} + t_m \mathbf{v}^{(m)}\right) \\ &= \mathbf{b} - A\mathbf{x}^{(m)} - t_m A\mathbf{v}^{(m)} \\ &= \mathbf{r}^{(m)} - \left(\mathbf{r}^{(m)} - \mathbf{r}^{(m+1)}\right) = \mathbf{r}^{(m+1)}. \end{aligned}$$

Hence, $\mathbf{b} - A\mathbf{x}^{(m+1)} = \mathbf{r}^{(m+1)} \quad 0 \le i \le m+1.$

(e')

$$\begin{aligned}\left\langle \mathbf{r}^{(m+1)}, \mathbf{r}^{(i)} \right\rangle &= \left\langle \mathbf{r}^{(m+1)}, \mathbf{v}^{(i)} - S_{i-1}\mathbf{v}^{(i-1)} \right\rangle \\ &= \left\langle \mathbf{r}^{(m+1)}, \mathbf{v}^{(i)} \right\rangle - S_{i-1}\left\langle \mathbf{r}^{(m+1)}, \mathbf{v}^{(i-1)} \right\rangle \\ &= 0.\end{aligned}$$

- Hence, $\left\langle \mathbf{r}^{(m)}, \mathbf{r}^{(i)} \right\rangle = 0$ for $0 \le i < m+1$.

(f')

$$\begin{aligned}0 &< \left\langle \mathbf{v}^{(m+1)}, A\mathbf{v}^{(m+1)} \right\rangle \\ &= \left\langle \mathbf{r}^{(m+1)} + S_m\mathbf{v}^{(m)}, A\mathbf{v}^{(m+1)} \right\rangle \\ &= \left\langle \mathbf{r}^{(m+1)}, A\mathbf{v}^{(m+1)} \right\rangle + S_m\left\langle \mathbf{v}^{(m)}, A\mathbf{v}^{(m+1)} \right\rangle \\ &= \left\langle \mathbf{r}^{(m+1)}, A\mathbf{v}^{(m+1)} \right\rangle \Rightarrow \mathbf{r}^{(m+1)} \neq 0\end{aligned}$$

Hence, $\mathbf{r}^{(i)} \neq 0$ for $0 \le i \le m+1$.

• The main computational cost in the CG algorithm comes from the matrix-vector multiplication of the form $A\mathbf{x}$. It takes at most $O(n^2)$ **operations**.

• If A is not symmetric, one can consider the **normal equation**: $A^T A\mathbf{x} = A^T\mathbf{b}$.

• The algorithm converges in at most n steps. It can be faster as the convergence rate of this method also depends on the spectrum of the matrix A_n.

• CG method can be used with a matrix called **preconditioner** to accelerate its convergence rate.

• A good preconditioner C should satisfy the following conditions.

(i) The matrix C can be constructed easily;

(ii) Given right hand side vector $\mathbf{r}$, the linear system $C\mathbf{y} = \mathbf{r}$ can be solved efficiently; and

(iii) the spectrum (or singular values) of the preconditioned system $C^{-1}A$ should be clustered around one.

• In the **Preconditioned Conjugate Gradient** (PCG) method, we solve the linear system

$$\hat{A}\hat{\mathbf{x}} = \hat{\mathbf{b}},$$

where

$$\begin{cases} \hat{A} = S^T A S \\ \hat{\mathbf{x}} = S^{-1}\mathbf{x} \\ \hat{\mathbf{b}} = S^T \mathbf{b} \end{cases}$$

such that $K(\hat{A}) < K(A)$.

• We expect the fast convergence rate of the PCG method can compensate much more than the extra cost in solving the preconditioner system in each iteration step of the PCG method.

4.6 Singular-Value Decomposition (SVD)

Theorem 4.7. *An arbitrary complex $m \times n$ matrix A can be factorized as*

$$A = PDQ$$

where P is an $m \times m$ unitary matrix, D is an $m \times n$ diagonal matrix, and Q is an $n \times n$ unitary matrix.

Proof A^*A is an $n \times n$ Hermitian matrix and positive semidefinite since

$$\mathbf{x}^*(A^*A)\mathbf{x} = (A\mathbf{x})^*(A\mathbf{x}) \geq 0$$

Denote the eigenvalues of A^*A by $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$ such that $\sigma_1^2, \sigma_2^2, \ldots, \sigma_r^2$ are positive and $\sigma_{r+1}^2, \sigma_{r+2}^2, \ldots, \sigma_n^2$ are 0.

- Let $\{\mathbf{u_1}, \mathbf{u_2}, \ldots, \mathbf{u_n}\}$ be an orthonormal set of eigenvectors for A^*A so that

$$A^*A\mathbf{u_i} = \sigma_i^2\mathbf{u_i}$$

Then

$$\|A\mathbf{u_i}\|_2^2 = \mathbf{u_i}^*A^*A\mathbf{u_i} = \mathbf{u_i}^*\sigma_i^2\mathbf{u_i} = \sigma_i^2$$

- Thus, we have $A\mathbf{u_i} = \mathbf{0}$ when $i \geq r+1$.
- Observe that

$$r = \text{rank}(A^*A) \leq \min\{\text{rank}(A^*), \text{rank}(A)\} \leq \min\{m, n\}.$$

We form an $n \times n$ matrix Q whose rows are $\{\mathbf{u_1^*}, \mathbf{u_2^*}, \ldots, \mathbf{u_n^*}\}$ and define

$$\mathbf{v_i} = \sigma_i^{-1} A\mathbf{u_i} \quad (1 \leq i \leq r).$$

- Note that

$$\mathbf{v_i^*}\mathbf{v_j} = \sigma_i^{-1}(A\mathbf{u_i})^*\sigma_j^{-1}(A\mathbf{u_j}) = (\sigma_i\sigma_j)^{-1}(\mathbf{u_i^*}A^*A\mathbf{u_j}) = \delta_{ij}$$

Thus, the v_i's form an orthonormal system for $1 \leq i, j \leq r$.

- Select additional vectors $\mathbf{v_i}$ so that $\{\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_m}\}$ is an orthonormal base for $\mathbb{C}^m$. Let P be the $m \times m$ matrix, whose columns are $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_m}$. Let D be the $m \times n$ matrix, having $\sigma_1, \sigma_2, \ldots, \sigma_r$ on its diagonal and 0's elsewhere. Then

$$(P^*AQ^*)_{ij} = \mathbf{v_i^*}A\mathbf{u_j} = D_{ij}$$

Hence $A = PDQ$.

• The numbers $\sigma_1, \sigma_2, \ldots, \sigma_n$ are called the **singular values** of A, also the non-negative square roots of the eigenvalues of A^*A.

Example 4.6 Find a singular-value decomposition for the matrix

$$\begin{bmatrix} 0 & -1.6 & 0.6 \\ 0 & 1.2 & 0.8 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Solution Following the proof of the theorem, we have

$$A^*A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Put $\sigma_1 = 1$, $\sigma_2 = 2$ and $\sigma_3 = 0$, and form the matrix Q

$$Q = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

Then

$$\mathbf{v}_1 = A\mathbf{u}_1 = [0.6, 0.8, 0, 0]^*$$
$$\mathbf{v}_2 = \frac{1}{2}A\mathbf{u}_2 = [-0.8, 0.6, 0, 0]^*$$

The simplest choices for $\mathbf{v}_3$ and $\mathbf{v}_4$ are

$$\mathbf{v}_3 = [0, 0, 1, 0]^*$$

and

$$\mathbf{v}_4 = [0, 0, 0, 1]^*,$$

respectively. The SVD of the given matrix is

$$\begin{bmatrix} 0 & -1.6 & 0.6 \\ 0 & 1.2 & 0.8 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0.6 & -0.8 & 0 & 0 \\ 0.8 & 0.6 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}.$$

4.7 Least Squares Fit

• Suppose we are given m distinct points (x_i, y_i) $(m \geq 2)$. The number of points m is much greater than the degree n of the required polynomial $p(x)$.

Then we can consider the errors in fitting the curve at the points $x_1, x_2, \ldots, x_m$:

$$|e_i| = |f(x_i) - p(x_i)| = |y_i - a_0 - a_1 x_i - \cdots - a_n x_i^n|.$$

One possible approach is to find a_i such that the **overall error** is minimized.

• The overall error to be minimized can be defined as

$$\min_{a_i} \left\{ \sum_{i=1}^{m} |e_i| \right\}$$

or

$$\min_{a_i} \left\{ \max\{|e_i|\} \right\}.$$

Both approaches result in a **Linear Programing (LP) problem**. But no closed-form solution is available in general.

- Here we shall consider the **least squares approach**:

$$\min_{a_i} \left\{ \sum_{i=1}^{m} e_i^2 \right\}$$

and we will demonstrate this by focusing on a linear function

$$\boxed{p(x) = ax + b.}$$

- The problem becomes

$$\min_{a,b} f(a,b) = \sum_{i=1}^{m} (y_i - a \cdot x_i - b)^2.$$

- We consider setting its partial derivatives to be zero:

$$\begin{cases} \frac{\partial f}{\partial a} &= -\sum_{i=1}^{m} (y_i - ax_i - b) \cdot x_i = 0 \\ \frac{\partial f}{\partial b} &= -\sum_{i=1}^{m} (y_i - ax_i - b) \cdot 1 = 0 \end{cases}$$

or equivalently the following linear system of equation:

$$J[a,b]^T = \begin{bmatrix} \sum_{i=1}^{m} x_i^2 & \sum_{i=1}^{m} x_i \\ \sum_{i=1}^{m} x_i & m \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{m} x_i y_i \\ \sum_{i=1}^{m} y_i \end{bmatrix}$$

- Solving the two equations, we get the closed-form solution as follows:

$$\boxed{a^* = \frac{m\sum_{i=1}^m x_i y_i - \sum_{i=1}^m x_i \sum_{i=1}^m y_i}{m\sum_{i=1}^m x_i^2 - (\sum_{i=1}^m x_i)^2}, \quad b^* = \frac{m\sum_{i=1}^m y_i \sum_{i=1}^m x_i^2 - \sum_{i=1}^m x_i \sum_{i=1}^m x_i y_i}{m\sum_{i=1}^m x_i^2 - (\sum_{i=1}^m x_i)^2}.}$$

- The above solution is guaranteed because

$$\det(J) = m\sum_{i=1}^m x_i^2 - \left(\sum_{i=1}^m x_i\right)^2 > 0.$$

- This can be explained by the following. Since x_i are distinct

$$\sum_{i=1}^m (y + x_i)^2 > 0$$

or

$$my^2 + 2\left(\sum_{i=1}^m x_i\right) y + \sum_{i=1}^m x_i^2 > 0.$$

The above quadratic equation in y has no real root and therefore we must have

$$4\left(\sum_{i=1}^m x_i\right)^2 < 4m\sum_{i=1}^m x_i^2.$$

- The Hessian matrix

$$H = \begin{bmatrix} \frac{\partial^2 f}{\partial a^2} & \frac{\partial^2 f}{\partial b \partial a} \\ \frac{\partial^2 f}{\partial a \partial b} & \frac{\partial^2 f}{\partial b^2} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^m x_i^2 & \sum_{i=1}^m x_i \\ \sum_{i=1}^m x_i & m \end{bmatrix} = J.$$

Since H is symmetric, its eigenvalues must be real. We observe that the product of roots is $\det J > 0$ and sum of roots is

$$\sum_{i=1}^m x_i^2 + m > 0$$

so all the eigenvalues of H are positive.

- This means that H is a **symmetric positive definite matrix**, i.e., $\mathbf{x}^T H \mathbf{x} > 0$ for all non-zero vector $\mathbf{x} \in \mathbb{R}^2$.

- By Taylor's theorem at the point (a^*, b^*), for some $\theta \in [0, 1]$ we have

$$\begin{aligned} f(a^*+c, b^*+d) &= f(a^*, b^*) + \left(c\frac{\partial}{\partial x} + d\frac{\partial}{\partial y}\right) f(a,b)|_{(a,b)=(a^*,b^*)} \\ &\quad + \frac{1}{2}\left(c\frac{\partial}{\partial x} + d\frac{\partial}{\partial y}\right)^2 f(x,y)|_{(x,y)=(a^*+\theta c, b^*+\theta d)} \\ &= f(a^*, b^*) + \frac{1}{2}\left(c^2\frac{\partial^2 f}{\partial a^2} + 2cd\frac{\partial^2 f}{\partial b \partial a} + d^2\frac{\partial^2 f}{\partial b^2}\right)|_{(x,y)=(a^*+\theta c, b^*+\theta d)} \\ &= f(a^*, b^*) + \underbrace{\frac{1}{2}[c,d]H[c,d]^T}_{non-negative}. \end{aligned}$$

- For $m = 2$, we have the straight line: $y = \frac{y_1 - y_2}{x_1 - x_2} \cdot x + \frac{y_2 x_1 - y_1 x_2}{x_1 - x_2}$.

4.8 Least Squares Problems and SVD

Let A be an $m \times n$ matrix and we are going to solve

$$\min_{\mathbf{x}} \left\{ \mathbf{r}^2 = ||A\mathbf{x} - \mathbf{b}||_2^2 \right\}$$

where $m > n$, an over-determined system and A is a full column rank matrix.

- Suppose we have the SVD of A as follows:

$$A = [P_1 \ P_2] \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} Q^T$$

where P_1 is an $m \times n$ matrix. Using the fact that, the 2-norm is invariant under orthogonal transformation, we have

$$||\mathbf{r}||_2^2 = ||\mathbf{b} - A\mathbf{x}||_2^2 = \left\| \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix} - \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} \mathbf{y} \right\|_2^2 .$$

where

$$\mathbf{b}_i = P_i^T \mathbf{b} \quad \text{and} \quad \mathbf{y} = Q^T \mathbf{x}.$$

- Now we have

$$||\mathbf{r}||_2^2 = ||\mathbf{b}_1 - \Sigma\mathbf{y}||_2^2 + ||\mathbf{b}_2||_2^2.$$

The minimum is attained when we choose $\mathbf{y}$ such that $\mathbf{b}_1 - \Sigma\mathbf{y} = \mathbf{0}$ or

$$\mathbf{y} = \Sigma^{-1}\mathbf{b}_1.$$

This means the minimizer is

$$\mathbf{x} = Q\mathbf{y} = Q\Sigma^{-1}\mathbf{b}_1 = Q\Sigma^{-1}P_1^T\mathbf{b}.$$

Since Σ is a diagonal matrix, we have

$$\Sigma^{-1} = \mathrm{diag}\left[\frac{1}{\sigma_1}, \frac{1}{\sigma_2}, \ldots, \frac{1}{\sigma_n}\right]$$

and $\sigma_i \neq 0$ as A is a full column rank matrix. The minimizer can be written as follows:

$$\boxed{\mathbf{x} = \sum_{i=1}^{n} \frac{1}{\sigma_i}\mathbf{q}_i^T\mathbf{b}\mathbf{p}_i}$$

where $\mathbf{q}_i$ and $\mathbf{p}_i$ are columns of Q and P_1, respectively.

5 Newton's Method for Solving Roots of Non-linear Equations

5.1 Non-linear Equation of One Variable

• Suppose we are to solve for $f(\alpha) = 0$ of a smooth function $f(x)$ (e.g. $f''(x)$ is differentiable) .

• Given an initial guess x_0 of the root α, the Newton's formula reads:

$$\boxed{x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \quad n = 0, 1, 2, \ldots.}$$

It gives a sequence of approximates x_n.

• The idea of Newton's method comes from the following Taylor series at x_n

$$f(x) = f(x_n) + (x - x_n)f'(x_n) + \frac{(x - x_n)^2}{2} f''(\epsilon) \tag{5.1}$$

where ϵ is between x and x_n.

 Newton's Method for Solving Roots of Non-linear Equations

- If we let $x = \alpha$ then $f(\alpha) = 0$ and we can solve α from (5.1) as follows:

$$\alpha = x_n - \frac{f(x_n)}{f'(x_n)} - \frac{(\alpha - x_n)^2}{2}\frac{f''(\epsilon_n)}{f'(x_n)}.$$

where ϵ_n is between α and x_n.

- If we drop the third term and regard it as an error term then we may have

$$\boxed{\alpha \approx x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}.}$$

In fact we have for $n = 0, 1, \ldots$

$$\alpha - x_{n+1} = -(\alpha - x_n)^2 \frac{f''(\epsilon_n)}{2f'(x_n)}. \tag{5.2}$$

Theorem 5.1. *Suppose $f(x), f'(x)$ and $f''(x)$ are continuous for all $x \in I$ where $I = (\alpha - h, \alpha + h)$ $(h > 0)$ and $f(\alpha) = 0$ and $f'(\alpha) \neq 0$. If x_0* **is close enough** *to α, the sequence of approximates x_n will converge to α.*

Proof Let us assume that h is small enough such that $f'(x) \neq 0$ for all $x \in I$. This can be done because $f'(x)$ is continuous and $f'(\alpha) \neq 0$.

- We define

$$\boxed{M = \frac{\max\limits_{x \in I} |f''(x)|}{2 \min\limits_{x \in I} |f'(x)|}.}$$

Here we pick x_0 such that $M|\alpha - x_0| < 1$.

- We note that for $n = 0, 1, 2, \ldots$, by Eq. (5.2) we have

$$|\alpha - x_{n+1}| \leq M|\alpha - x_n|^2$$

or

$$M|\alpha - x_{n+1}| \leq (M|\alpha - x_n|)^2$$

- Inductively we have

$$|\alpha - x_{n+1}| \leq \frac{1}{M}\left[M|\alpha - x_0|\right]^{2^{n+1}}.$$

Since

$$M|\alpha - x_0| < 1$$

we have

$$\lim_{n\to\infty} \frac{1}{M}\left[M|\alpha - x_0|\right]^{2^{n+1}} = 0$$

and therefore

$$\lim_{n\to\infty} x_{n+1} = \alpha.$$

 Newton's Method for Solving Roots of Non-linear Equations

Example: Solve the non-linear equation: $0.1e^x = x$.

We define $f(x) = x - 0.1e^x$. We then obtain $f'(x) = 1 - 0.1e^x$.

- The Newton's iterative scheme reads

$$x_{n+1} = x_n - \frac{x_n - 0.1e^{x_n}}{1 - 0.1e^{x_n}}.$$

Using different x_0 may result in different solutions.

x_n	$f(x_n)$	x_n	$f(x_n)$
$x_0 = 4.000$	-1.4598	$x_0 = -4.0000$	-4.0018
$x_1 = 3.673$	-0.2630	$x_1 = 0.0092$	0.0917
$x_2 = 3.583$	-0.0153	$x_2 = 0.1112$	-0.0005
$x_3 = 3.577$	-0.0000	$x_3 = 0.1118$	0.0000
$x_4 = 3.577$	-0.0000	$x_4 = 0.1118$	0.0000

- Try $x_0 = 2.300, 2.305, 2.310$ and check the behavior of convergence.

5.2 Newton's Method for Non-linear Systems

• For simplicity, we consider solving root $\mathbf{a} = [a_1, a_2]^T$ of a non-linear system of two variables $f_i(x_1, x_2)$ $(i = 1, 2)$. Recall the Taylor's theorem for function of two variables, expanding $f_i(a_1, a_2)$ at $\mathbf{b} = [b_1, b_2]^T$:

$$\boxed{0 = f_i(\mathbf{a}) = f_i(\mathbf{b}) + (a_1 - b_1)\frac{\partial f_i(\mathbf{b})}{\partial x_1} + (a_2 - b_2)\frac{\partial f_i(\mathbf{b})}{\partial x_2} + \text{second-order term.}}$$

• Dropping the second-order term, we obtain

$$\begin{cases} 0 \approx f_1(\mathbf{b}) + (a_1 - b_1)\frac{\partial f_1(\mathbf{b})}{\partial x_1} + (a_2 - b_2)\frac{\partial f_1(\mathbf{b})}{\partial x_2} \\ 0 \approx f_2(\mathbf{b}) + (a_1 - b_1)\frac{\partial f_2(\mathbf{b})}{\partial x_1} + (a_2 - b_2)\frac{\partial f_2(\mathbf{b})}{\partial x_2}. \end{cases}$$

In matrix form, we have

$$\mathbf{0} \approx \begin{bmatrix} f_1(\mathbf{b}) \\ f_2(\mathbf{b}) \end{bmatrix} + \begin{bmatrix} \frac{\partial f_1(\mathbf{b})}{\partial x_1} & \frac{\partial f_1(\mathbf{b})}{\partial x_2} \\ \frac{\partial f_2(\mathbf{b})}{\partial x_1} & \frac{\partial f_2(\mathbf{b})}{\partial x_2} \end{bmatrix} \begin{bmatrix} a_1 - b_1 \\ a_2 - b_2 \end{bmatrix} \equiv f(\mathbf{b}) + F(\mathbf{b})(\mathbf{a} - \mathbf{b})$$

or $\mathbf{a} \approx \mathbf{b} - F(\mathbf{b})^{-1} f(\mathbf{b})$. We have the **Newton's scheme** if we regard $\mathbf{a}$ as an estimate of the root given $\mathbf{b}$:

$$\boxed{\mathbf{x}_{n+1} = \mathbf{x}_n - F(\mathbf{x}_n)^{-1} f(\mathbf{x}_n).}$$

Example: Consider the following non-linear system of equations:

$$\begin{cases} f_1(x_1, x_2) &= x_1^2 + x_2^2 - 1 \\ f_2(x_1, x_2) &= x_1 - x_2 - 0.5. \end{cases}$$

Then we have the matrix

$$F(\mathbf{x}) = \begin{bmatrix} 2x_1 & 2x_2 \\ 1 & -1 \end{bmatrix} \quad \text{and} \quad F(\mathbf{x})^{-1} = \begin{bmatrix} \frac{1}{2(x_1+x_2)} & \frac{x_2}{x_1+x_2} \\ \frac{1}{2(x_1+x_2)} & \frac{-x_1}{x_1+x_2} \end{bmatrix}$$

- The Newton's iterative scheme is given by

$$\mathbf{x}_{n+1} = \mathbf{x}_n - F(\mathbf{x}_n)^{-1} f(\mathbf{x}_n).$$

or

$$\begin{bmatrix} x_1^{(n+1)} \\ x_2^{(n+1)} \end{bmatrix} = \frac{1}{2(x_1^{(n)} + x_2^{(n)})} \begin{bmatrix} (x_1^{(n)})^2 + (x_2^{(n)})^2 - x_2^{(n)} + 1 \\ (x_1^{(n)})^2 + (x_2^{(n)})^2 + x_1^{(n)} + 1 \end{bmatrix}$$

Newton's Method for Solving Roots of Non-linear Equations

- We apply Newton's iterative scheme with $\mathbf{x}_0 = [1, 1]^T$ and get

$\mathbf{x}_i$	$f(\mathbf{x}_i)$	$\|\|f(\mathbf{x}_i)\|\|_2$
$\mathbf{x}_0 = [1.0000, 1.0000]^T$	$[1.0000, -0.5000]^T$	1.1180
$\mathbf{x}_1 = [1.0000, 0.5000]^T$	$[0.2500, 0.0000]^T$	0.2500
$\mathbf{x}_2 = [0.9167, 0.4167]^T$	$[0.0139, 0.0000]^T$	0.0139
$\mathbf{x}_3 = [0.9115, 0.4115]^T$	$[0.0001, 0.0000]^T$	0.0001
$\mathbf{x}_4 = [0.9114, 0.4114]^T$	$[0.0000, 0.0000]^T$	0.0000

5.3 Contraction Mapping Method

Let

$$B = B(\mathbf{d}, r) = \{\mathbf{x} : ||\mathbf{x} - \mathbf{d}||_2 \leq r\} \subseteq \mathbb{R}^n,$$

$g(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}^n$,

$$g(\mathbf{x}) = [g_1(\mathbf{x}), \cdots, g_n(\mathbf{x})]^T$$

and all components of $g(\mathbf{x})$ are continuous differentiable in B and

$$G(\mathbf{x}) = \begin{bmatrix} \frac{\partial g_1(\mathbf{x})}{\partial x_1} & \frac{\partial g_1(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial g_1(\mathbf{x})}{\partial x_n} \\ \frac{\partial g_2(\mathbf{x})}{\partial x_1} & \frac{\partial g_2(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial g_2(\mathbf{x})}{\partial x_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial g_n(\mathbf{x})}{\partial x_1} & \frac{\partial g_n(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial g_n(\mathbf{x})}{\partial x_n} \end{bmatrix}.$$

Theorem 5.2. *Suppose* $g(B) \subseteq B$ *and* $\lambda = \max_{\mathbf{x} \in B} ||G(\mathbf{x})||_{M_\infty} < 1$ *then*

(i) The equation $\mathbf{x} = g(\mathbf{x})$ *has a unique solution* $\mathbf{a} \in B$.

(ii) For any given $x_0 \in B$, *the iteration* $\mathbf{x}_{n+1} = g(\mathbf{x}_n)$ *will converge to* $\mathbf{a}$.

Proof

• We first note by fixed point theorem, there exists at least one fixed point in B. We then show the uniqueness of the fixed point.

• Suppose $\mathbf{a}$ and $\mathbf{b}$ are two fixed points, $\mathbf{a} - \mathbf{b} = g(\mathbf{a}) - g(\mathbf{b})$. Then for each i, by the mean-value theorem, there exists $\mathbf{e} \in B$ such that

$$g_i(\mathbf{a}) - g_i(\mathbf{b}) = \left[\frac{\partial g_i(\mathbf{x})}{\partial x_1}, \cdots, \frac{\partial g_i(\mathbf{x})}{\partial x_n}\right] |_{\mathbf{x}=\mathbf{e}} (\mathbf{a} - \mathbf{b})$$

• Now we note that for $\mathbf{x} \in B$,

$$\left|\frac{\partial g_i(\mathbf{x})}{\partial x_1}\right| + \cdots + \left|\frac{\partial g_i(\mathbf{x})}{\partial x_n}\right| \leq \lambda < 1.$$

Thus

$$|g_i(\mathbf{a}) - g_i(\mathbf{b})| \leq \lambda ||(\mathbf{a} - \mathbf{b})||_\infty \tag{5.3}$$

and hence we have

$$||(\mathbf{a} - \mathbf{b})||_\infty = ||g(\mathbf{a}) - g(\mathbf{b})||_\infty \leq \lambda ||(\mathbf{a} - \mathbf{b})||_\infty < ||(\mathbf{a} - \mathbf{b})||_\infty.$$

This is impossible.

- For any $\mathbf{x}_0 \in B$, by $g(B) \subseteq B$, we have all $\mathbf{x}_n \in B$. Moreover, we have

$$\mathbf{a} - \mathbf{x}_{n+1} = g(\mathbf{a}) - g(\mathbf{x}_n).$$

By Eq. (5.3), we have

$$||\mathbf{a} - \mathbf{x}_{n+1}||_\infty = ||g(\mathbf{a}) - g(\mathbf{x}_n)||_\infty \leq \lambda ||\mathbf{a} - \mathbf{x}_n||_\infty.$$

Inductively, we have

$$||\mathbf{a} - \mathbf{x}_{n+1}||_\infty \leq \lambda^{n+1} ||\mathbf{a} - \mathbf{x}_0||_\infty.$$

- Thus

$$\lim_{n\to\infty} ||\mathbf{a} - \mathbf{x}_{n+1}||_\infty \leq \lim_{n\to\infty} \lambda^{n+1} ||\mathbf{a} - \mathbf{x}_0||_\infty = 0$$

and we conclude that

$$\lim_{n\to\infty} \mathbf{x}_{n+1} = \mathbf{a}.$$

5.4 A Brief Summary

• If A is symmetric and positive definite, then the problem of solving $Ax = b$ is equivalent to the problem of minimizing $q(\mathbf{x}) =< \mathbf{x}, A\mathbf{x} > -2 < \mathbf{x}, \mathbf{b} >$

• Assuming that A is an $n \times n$ symmetric positive definite matrix. The set of vectors $\{\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \cdots, \mathbf{u}^{(n)}\}$ is A-orthonormal if $< \mathbf{u}^{(i)}, A\mathbf{u}^{(j)} >= \delta_{ij}$.

• The conjugate gradient method converges within n steps when applied to solve a symmetric positive definite matrix system.

• The power method and the inverse power method for solving extreme eigenvalues.

• Gershgorin's Theorem: The eigenvalues of an $n \times n$ matrix A are contained in the union of the following n disks $D_i (i = 1, 2, \ldots, n)$ where

$$D_i = \left\{ z \in \mathbb{C} : |z - a_{ii}| \leq \sum_{j=1, j \neq i}^{n} |a_{ij}| \right\}.$$

• Singular-Value Decomposition (SVD): $A = PDQ$ and its relation to the least squares problem.