Computational Techniques - COMP 50011

Viet Pham Ngoc
From materials by Peter Harrison

17/02/2022 – 03/03/2022

Contents

1 Condition Number

1.1 General Definition and Examples

The condition number is the measure of the sensibility of problem to small fluctuations in its input. for example if you are interested in:

- The stability of a system: how will it respond to some noise in its input
- The sensibility of the solutions to a parametric equation with regard to a small change in the parameters

Then one way to quantify these behaviors is to compute the condition number of these problems. Formally:

Definition 1.1: Condition number
Let P be the problem of interest, d an input, ϵ some perturbation on the input and $s(d)$, $s(d+\epsilon)$ the corresponding outputs. Then the condition number of P, $\kappa(P)$ is defined as:

$$\boxed{\kappa(P) = \max_{\epsilon} \frac{\|s(d) - s(d+\epsilon)\|}{\|\epsilon\|}}$$

According to this definition the condition number is measured by computing the difference between the original output and the perturbed one, relative to the perturbation and then taking the maximum over all the possible perturbations: the condition number measures the worst case scenario.

Remark. The value of the condition number depends on the norms that are being used and it can also be defined in terms of relative difference as will be seen for the condition number of a matrix. In this case, the relative condition number is defined by:

$$\kappa(P) = \max_{\epsilon} \frac{\|s(d) - s(d+\epsilon)\|}{\|\epsilon\|} \frac{\|d\|}{\|s(d)\|}$$

Example 1.1: An unstable system
Suppose you have a system for which the input-output function is:

$$f(x) = x^4$$

Then for example, we have $f(10) = 10000$ and $f(10.1) \approx 10406$. In this case, P is the system, $d = 10$, $\epsilon = 0.1$, $s(d) = f(10)$ and $s(d+\epsilon) = f(10.1)$ so

$$\frac{\|s(d) - s(d+\epsilon)\|}{\|\epsilon\|} = \frac{|f(10) - f(10.1)|}{|0.1|} \approx 4000$$

The condition number being the maximum of such quantities, it comes that:

$$\kappa(P) \geq 4000$$

Example 1.2: A stable system
Suppose now that your system is governed by:

$$g(x) = 0.1x$$

Then, $g(1) = 0.1$ and $g(10) = 1$. If we take P to be the system, $d = 1$, $\epsilon = 9$, $s(d) = g(1)$ and $s(d+\epsilon) = g(10)$ then:

$$\frac{\|s(d) - s(d+\epsilon)\|}{\|\epsilon\|} = \frac{|g(1) - g(10)|}{|9|} = 0.1$$

Now take arbitrary $d \in \mathbb{R}$, $\epsilon \neq 0$ then:

$$\begin{aligned}\frac{\|s(d) - s(d+\epsilon)\|}{\|\epsilon\|} &= \frac{|g(d) - g(d+\epsilon)|}{|\epsilon|} = \frac{|0.1d - 0.1(d+\epsilon)|}{|\epsilon|} \\ &= \frac{|-0.1(\epsilon)|}{|\epsilon|} = \frac{0.1|\epsilon|}{|\epsilon|} \\ &= 0.1\end{aligned}$$

So for all d, ϵ, this quantity is the same, hence:

$$\kappa(P) = \max_{\epsilon} \frac{\|s(d) - s(d+\epsilon)\|}{\|\epsilon\|} = 0.1$$

Now let's study the condition number in the context of linear equations.

Example 1.3: Linear equations
Let's take a look at this parametric equation:

$$\begin{cases} x + y = 1 \\ x + \alpha y = 0 \end{cases}$$

This equation has a solution when $\alpha \neq 1$, in this case it is given by: $x = -\frac{\alpha}{1-\alpha}$ and $y = \frac{1}{1-\alpha}$. When α is close to 1, a slight change in its value can dramatically change the value of the solutions:

- For $\alpha = 0.999$ we have $x = -999$ and $y = 1000$
- For $\alpha = 0.9999$ we have $x = -9999$ and $y = 10000$

In this example, P is the equation, $d = 0.9999$, $s(d) = (-9999, 10000)$, $\epsilon = -0.0009$ and $s(d+\epsilon) = (-999, 1000)$. Leading to:

$$\frac{\|s(d) - s(d+\epsilon)\|}{\|\epsilon\|} = \frac{\|(-9999, 10000) - (-999, 1000)\|}{|0.0009|} = \frac{\|(-9000, 9000)\|}{0.0009}$$

The numerator being the norm of a vector, we have to choose from different norms. Let us work with the l_1-norm, then:

$$\frac{\|s(d) - s(d+\epsilon)\|_1}{\|\epsilon\|} = \frac{|-9000| + |9000|}{0.0009} = 2.10^7$$

If we were to work with the l_2-norm, we would have:

$$\frac{\|s(d) - s(d+\epsilon)\|}{\|\epsilon\|} = \frac{\sqrt{(-9000)^2 + (9000)^2}}{0.0009} = \sqrt{2}.10^7$$

These examples illustrate how the condition number can be used to quantify the sensibility of a problem to small perturbations in its input:

- If the condition number is large, then a small variation of the input will lead to a "blow up" in its output. The problem is then said to be ill-conditioned.
- On the contrary, if it is small, the output will stay relatively stable against small perturbations in the input. The problem is then said to be well-conditioned

1.2 Condition Number of a Square Matrix

In this part we will make use of some properties of matrix norms. For a refresher, please refer to the first part's lecture notes (Section 1.2).

1.2.1 Square matrix

Definition 1.2: Condition number of a matrix
Let A be a non-singular matrix. Its condition number is defined by:

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

Proposition 1.1:
Suppose that the problem P we want to solve is represented by the following linear equation:

$$\mathbf{Ax} = \mathbf{b}$$

Where A is non-singular. Then the relative condition number of the problem is the condition number of the matrix A:

$$\boxed{\kappa(P) = \kappa(\mathbf{A}) = \|\mathbf{A}^{-1}\|\|\mathbf{A}\|}$$

Proof. Let $\mathbf{A}$ be in non-singular matrix and $\mathbf{x}_0$ the solution to $\mathbf{Ax} = \mathbf{b}$. We will consider two cases: either a disturbance in $\mathbf{A}$ or a disturbance in $\mathbf{b}$.

Suppose $\mathbf{A}$ is perturbed by $\delta\mathbf{A}$ then the new solution to be solved is:

$$(\mathbf{A} + \delta\mathbf{A})\mathbf{x} = \mathbf{b}$$

Let $\mathbf{x}_0 + \delta\mathbf{x}$ be the solution to that new equation, then:

$$\begin{aligned}(\mathbf{A} + \delta\mathbf{A})(\mathbf{x}_0 + \delta\mathbf{x}) &= \mathbf{b}\\ \Rightarrow \mathbf{Ax}_0 + \mathbf{A}\delta\mathbf{x} + \delta\mathbf{Ax}_0 + \delta\mathbf{A}\delta\mathbf{x} &= \mathbf{b}\end{aligned}$$

Ignoring the second degree term ($\delta\mathbf{A}\delta\mathbf{x}$) together with the fact that $\mathbf{Ax}_0 = \mathbf{b}$ lead to:

$$\begin{aligned}\mathbf{A}\delta\mathbf{x} + \delta\mathbf{Ax}_0 &= \mathbf{0}\\ \Rightarrow \delta\mathbf{x} &= -\mathbf{A}^{-1}\delta\mathbf{Ax}_0\end{aligned}$$

Taking the norm, we have:

$$\begin{aligned}\|\delta\mathbf{x}\| &= \|\mathbf{A}^{-1}\delta\mathbf{Ax}_0\|\\ &\leq \|\mathbf{A}^{-1}\|\|\delta\mathbf{A}\|\|\mathbf{x}_0\| \text{ (consistency of the norm)}\\ &\leq \|\mathbf{A}^{-1}\|\|\mathbf{A}\|\frac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|}\|\mathbf{x}_0\|\end{aligned}$$

Hence:

$$\frac{\|\delta\mathbf{x}\|/\|\mathbf{x}_0\|}{\|\delta\mathbf{A}\|/\|\mathbf{A}\|} \leq \|\mathbf{A}^{-1}\|\|\mathbf{A}\|$$

This upper bound being true for all disturbance δA and consequent δx, it comes that:

$$\max \frac{\|\delta\mathbf{x}\|/\|\mathbf{x}_0\|}{\|\delta\mathbf{A}\|/\|\mathbf{A}\|} \leq \|\mathbf{A}^{-1}\|\|\mathbf{A}\|$$

It turns out that the upper bound can be reached so:

$$\max \frac{\|\delta\mathbf{x}\|/\|\mathbf{x}_0\|}{\|\delta\mathbf{A}\|/\|\mathbf{A}\|} = \|\mathbf{A}^{-1}\|\|\mathbf{A}\|$$

The left hand side being the definition of the relative condition number of P, we do end up with:

$$\kappa(P) = \|\mathbf{A}^{-1}\|\|\mathbf{A}\| = \kappa(\mathbf{A})$$

Suppose now that $\mathbf{b}$ is perturbed by $\delta\mathbf{b}$ and $\mathbf{x}_0 + \delta\mathbf{x}$ is such that:

$$\mathbf{A}(\mathbf{x}_0 + \delta\mathbf{x}) = \mathbf{b} + \delta\mathbf{b}$$

Developing and the fact that $\mathbf{A}\mathbf{x}_0 = \mathbf{b}$:

$$\begin{aligned}
&\mathbf{A}\delta\mathbf{x} = \delta\mathbf{b} \\
\Rightarrow &\delta\mathbf{x} = \mathbf{A}^{-1}\delta\mathbf{b} \\
\Rightarrow &\|\delta\mathbf{x}\| \leq \|\mathbf{A}^{-1}\|\|\delta\mathbf{b}\| \text{ (consistency of the norm)}
\end{aligned}$$

As $\mathbf{A}\mathbf{x}_0 = \mathbf{b}$ we also have $\|\mathbf{b}\| \leq \|\mathbf{A}\|\|\mathbf{x}_0\|$, hence:

$$\frac{\|\delta\mathbf{x}\|/\|\mathbf{x}_0\|}{\|\delta\mathbf{b}\|/\|\mathbf{b}\|} \leq \|\mathbf{A}^{-1}\|\|\mathbf{A}\|$$

The same reasoning as previously leads to:

$$\kappa(P) = \|\mathbf{A}^{-1}\|\|\mathbf{A}\| = \kappa(\mathbf{A})$$

□

Now let's study some practical example of the condition number of a matrix.

Example 1.4: Return of Example 1.3
The problem studied in Example 1.3 was the following equation:

$$\begin{cases} x + y = 1 \\ x + \alpha y = 0 \end{cases}$$

With $\alpha \neq 1$. By denoting $\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & \alpha \end{bmatrix}$ and $\mathbf{b} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, this equation can be rewritten: $\mathbf{A}\mathbf{s} = \mathbf{b}$.

For $\alpha \neq 1$, $\mathbf{A}$ is non-singular and we have $\mathbf{A}^{-1} = \frac{1}{\alpha - 1}\begin{bmatrix} \alpha & -1 \\ -1 & 1 \end{bmatrix}$. As in Example 1.3 we take $\alpha = 0.9999$.
Now let's work with the l_1-norm. We have $\|\mathbf{A}\|_1 = 2$ and $\|\mathbf{A}^{-1}\| = \frac{2}{1-\alpha} = 2.10^4$.
So

$$\kappa(\mathbf{A}) = \|\mathbf{A}^{-1}\|\|\mathbf{A}\| = 4.10^4$$

Now if we look back at the value computed in Example 1.3, it is not the relative value. The relative value is:

$$\frac{\|s(d) - s(d+\epsilon)\|}{\epsilon}\frac{\|d\|}{\|s(d)\|} = 2.10^7 \times \frac{0.9999}{19999} \approx 1000$$

Example 1.5: Big norm means ill-conditioned?
Consider the matrix

$$\mathbf{A} = \begin{bmatrix} 10^8 & 0 \\ 0 & 10^{-8} \end{bmatrix}$$

Then $\|\mathbf{A}\|_1 = 10^8$. We then have:

$$\mathbf{A}^{-1} = \begin{bmatrix} 10^{-8} & 0 \\ 0 & 10^8 \end{bmatrix}$$

And $\|\mathbf{A}^{-1}\| = 10^8$ so $\kappa(\mathbf{A}) = \|\mathbf{A}^{-1}\|\|\mathbf{A}\| = 10^{16}$. This condition number is very large, meaning that the matrix is ill-conditioned. Is it always the case that a big norms means that the matrix is ill-conditioned? Now consider the matrix

$$\mathbf{B} = \begin{bmatrix} 10^8 & 0 \\ 0 & 10^8 \end{bmatrix}$$

Then as previously, we have $\|\mathbf{B}\|_1 = 10^8$. But we also have:

$$\mathbf{B}^{-1} = \begin{bmatrix} 10^{-8} & 0 \\ 0 & 10^{-8} \end{bmatrix}$$

And $\|\mathbf{B}^{-1}\| = 10^{-8}$ meaning that $\kappa(\mathbf{B}) = 1$. So $\mathbf{B}$ is well-conditioned.
All in all, do not assume that because a matrix has a large norm, it is necessary ill-conditioned.

1.2.2 Non-Square Matrix

What happens when we are confronted with a problem represented by a non-square matrix? For example you want to solve a system with m equations and n variables where $m > n$? It can be written as

$$\mathbf{Ax} = \mathbf{b}$$

Where $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{b} \in \mathbb{R}^m$. As $\mathbf{A}$ is non-square, it can not be inverted and the condition number as defined in Definition 1.2 can not be calculated. But:

$$\mathbf{Ax} = \mathbf{b} \Rightarrow \mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b}$$

Assuming the columns of $\mathbf{A}$ are linearly independent then $\mathbf{A}^T\mathbf{A} \in \mathbb{R}^{n\times n}$ can be inverted and we end up with:

$$x = \left(A^TA\right)^{-1}A^Tb$$

The matrix $\left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T$ is called the pseudo-inverse of $\mathbf{A}$ and is denoted $\mathbf{A}^\dagger$.

Definition 1.3: Condition Number of a Non-Square Matrix
Let $\mathbf{A}$ be a non-square matrix, then its condition number is given by:

$$\boxed{\kappa(\mathbf{A}) = \|\mathbf{A}^\dagger\|\|\mathbf{A}\|}$$

Remark. For a square non-singular matrix $\mathbf{B}$, we do have $\mathbf{B}^\dagger = \mathbf{B}^{-1}$ so Definition 1.3 is actually a generalisation of Definition 1.2.

Remark. In this course we will mostly consider square non-singular matrices so you can consider Definition 1.3 as optional. If need be, it will be given to you as a reminder.

1.3 About the Conditioning of a Problem

Now that we know what the condition number is and how to compute it, how do we decide whether a problem is well-conditioned or ill-conditioned?
The problem is that there is no numerical threshold to decide of the conditioning of the problem, it depends on the problem. Nonetheless there is a rule of thumb that can be useful:

Proposition 1.2: Rule of Thumb
For a condition number $\kappa(\mathbf{A})$, you loose about $\log_{10}(\kappa(\mathbf{A}))$ significant figures in accuracy.

Example 1.6:
Suppose you have a highly accurate implementation working with 12 decimal places. If the condition number is 1,000,000 then you'll loose about 6 decimal places in precision.

Another important thing to note: suppose you have this equation:

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

Solving it numerically, you get $\mathbf{x}_0$. Then the residual vector defined by $\mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x}_0$ is a reliable indicator of the accuracy of the solution only if $\mathbf{A}$ is well-conditioned.

2 Convergence

2.1 Results in $\mathbb{R}$

2.1.1 Convergence

This part is a refresher about the convergence of sequences that have value in the real space.

Definition 2.1: Convergence of a sequence of real numbers
Let $(a_n)_{n\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}}$ be a sequence of real numbers and $l \in \mathbb{R}$. The sequence (a_n) is said to converge to its limit l, noted $a_n \xrightarrow[n\to\infty]{} l$ or $\lim_{n\to\infty} a_n = l$ if and only if:

$$\boxed{\forall \epsilon > 0,\ \exists N \in \mathbb{N} \text{ such that } \forall n > N,\ |a_n - l| < \epsilon}$$

How can we interpret this definition? Let $\epsilon > 0$, the last part of the definition can be rewritten $l - \epsilon < a_n < l + \epsilon$. This means that for $n > N$, the terms a_n are within the tracks of width 2ϵ and centered around l as shown in Figure 1. The smaller ϵ is, the tighter the tracks get around l and the bigger N has to be for a_n to remain within the tracks.

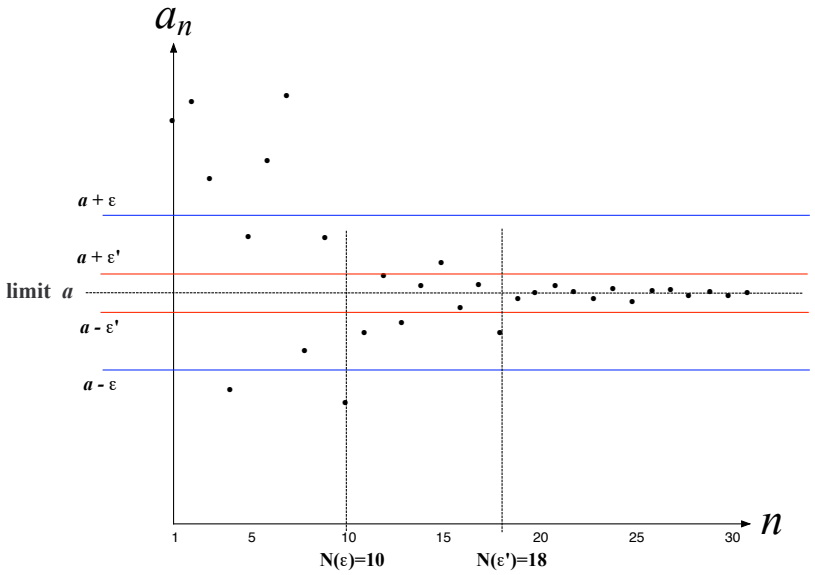

Figure 1: Illustration of convergence

Example 2.1: An easy convergence
Let (a_n) the sequence defined by: $\forall n > 0, a_n = \frac{(-1)^n}{n}$. Find its limit and show that it indeed converges to that limit.

Proof. Let's show that $a_n \xrightarrow[n\to\infty]{} 0$. Let $\epsilon > 0$, we have:

$$|a_n - 0| = \left|\frac{(-1)^n}{n}\right| = \frac{1}{n}$$

In order to have $|a_n - 0| < \epsilon$ we need to have $n > \frac{1}{\epsilon}$. So by defining $N = \left\lceil \frac{1}{\epsilon} \right\rceil$ we do have

$$\forall n > N, |a_n - 0| < \epsilon$$

□

As expected, as ϵ gets smaller, N gets larger. Although the sequence used in Example 2.1 is relatively simple, it outlines the steps to show that a sequence of real numbers (a_n) is converging:

1. Find the limit l
2. Take $\epsilon > 0$
3. Find $N \in \mathbb{N}$ such that $|a_n - l| < \epsilon$ for $n > N$, the value of N will usually depend on and decrease with ϵ

2.1.2 Cauchy Sequence

Definition 2.2: Cauchy Sequence
Let $(a_n)_{n\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}}$ a sequence of real numbers. Then (a_n) is said to be a Cauchy sequence if and only if:

$$\boxed{\forall \epsilon > 0,\ \exists N \in \mathbb{N} \text{ such that } \forall n, m > N,\ |a_n - a_m| < \epsilon}$$

Intuitively, a sequence is a Cauchy sequence if its terms gradually get closer. This concept can be used to determine the convergence of a sequence.

Theorem 2.1: Cauchy test
Let $(a_n)_{n\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}}$ a sequence of real numbers. Then (a_n) is convergent if and only if it is a Cauchy sequence

Proof. **This proof is just here for the sake of curiosity**.
Suppose (a_n) is convergent.
Let l be its limit and $\epsilon > 0$, then by definition of convergence, there exists $N \in \mathbb{N}$ such that:

$$\forall n > N, |a_n - l| < \epsilon/2$$

Now, let $n, m > N$, then:

$$|a_n - a_m| = |a_n - l + l - a_m| \leq |a_n - l| + |a_m - l| < \epsilon$$

So we have shown that:

$$\forall n, m > N, |a_n - a_m| < \epsilon$$

That is (a_n) is a Cauchy sequence.
Suppose now that (a_n) is a Cauchy sequence.
We take $\epsilon = 1$, then by definition of a Cauchy sequence, there exist $N \in \mathbb{N}$ such that:

$$\forall n > N, |a_{N+1} - a_n| < 1$$

So:

$$\forall n > N, a_{N+1} - 1 < a_n < a_{N+1} + 1$$

Let $M = \max\{|a_0|, |a_1|, \ldots, |a_N|, |a_{N+1}|, |a_{N+1} - 1|, |a_{N+1} + 1|\}$, then

$$\forall n \in \mathbb{N}, |a_n| \leq M$$

This means that (a_n) is bounded and by applying the Bolzano-Weieirstrass theorem (a bounded sequence has a convergent subsequence) we can extract a subsequence $(a_{\phi(n)})$ that is convergent ($\phi : \mathbb{N} \to \mathbb{N}$ is a strictly increasing function). We denote l its limit
Let $\epsilon > 0$, (a_n) being a Cauchy sequence, there exist $N \in \mathbb{N}$ such that:

$$\forall n, m > N, |a_n - a_m| < \epsilon/2$$

$(a_{\phi(n)})$ being convergent, there exist $N_1 \in \mathbb{N}$ such that:

$$\forall n > N_0, |a_{\phi(n)} - l| < \epsilon/2$$

ϕ being a strictly increasing function, there exist $n_0 > N_0$ such that $\phi(n_0) > N$. Take $n > N$, then:

$$|a_n - l| = |a_n - a_{\phi(n_0)} + a_{\phi(n_0)} - l| \leq |a_n - a_{\phi(n_0)}| + |a_{\phi(n_0)} - l| < \epsilon$$

So:

$$\forall n > N, |a_n - l| < \epsilon$$

That is: $\underline{(a_n) \text{ is convergent}}$ □

Remark. With Theorem 2.1 you can show that a sequence is convergent but it will not give you its limit. It is extremely useful when you just want to show that a sequence is converging but its limit is hard to find. It is particularly handy when the sequence is defined by a recurrence relation.

Remark. Theorem 2.1 being an equivalence, it can also be used to show that a sequence is not converging.

Example 2.2: A not so easy convergence
Let (a_n) be the sequence defined by $\forall n > 0,\ a_n = \sum_{i=1}^{n} \frac{1}{i(i+1)}$. Show that it is converging.
Let $n, m > 0$, then:

$$\begin{aligned} a_n - a_m &= \sum_{i=1}^{n} \frac{1}{i(i+1)} - \sum_{i=1}^{m} \frac{1}{i(i+1)} \\ &= \sum_{i=m+1}^{n} \frac{1}{i(i+1)} \\ &= \sum_{i=m+1}^{n} \frac{1}{i} - \frac{1}{i+1} \\ &= \frac{1}{m+1} - \frac{1}{n+1} \end{aligned}$$

Let $\epsilon > 0$ and $N = \left\lceil \frac{2}{\epsilon} \right\rceil$ then for $n > N$: $\frac{1}{n+1} < \frac{2}{\epsilon}$. It comes that for $n, m > N$ we have:

$$|a_n - a_m| = \left| \frac{1}{m+1} - \frac{1}{n+1} \right| \leq \left| \frac{1}{m+1} \right| + \left| \frac{1}{n+1} \right| < \epsilon$$

So $\underline{(a_n) \text{ is a Cauchy sequence of real numbers so it converges.}}$
Actually, we can simplify the expression of a_n to $a_n = 1 - \frac{1}{n+1}$ so it is easy to show that it converges and its limit is 1.

Example 2.3: A non converging sequence
One way to prove that a real valued sequence is not converging is by proving that it is not a Cauchy sequence.
Consider the sequence defined by $a_n = (-1)^n$ for $n \in \mathbb{N}$. Let $\epsilon = 1$, then for $n \in \mathbb{N}$, we have:

$$|a_{2n} - a_{2n+1}| = 2 > 1$$

This means that we can not find and N such that $|a_n - a_m| < 1$ for $n, m > N$. So (a_n) is not a Cauchy sequence, hence it is not converging.

2.2 Generalisation to Metric Spaces

2.2.1 Metric Space

Definition 2.3: Metric Space
A metric space is a tuple (S, d) where S is an non-empty set and d is a metric over S meaning:
A function $d : S \times S \to \mathbb{R}$ such that:

1. $\forall x, y \in S,\ d(x, y) \geq 0$
2. $\forall x, y \in S,\ d(x, y) = 0 \iff x = y$
3. $\forall x, y \in S,\ d(x, y) = d(y, x)$
4. $\forall x, y, z \in S,\ d(x, y) \leq d(x, z) + d(z, y)$

Remark. The first property can be derived from the 3 others: let $x, y \in S$, then

$$\begin{aligned} & d(x,x) \leq d(x,y) + d(y,x) \text{ (property 4)} \\ \Rightarrow & 0 \leq d(x,y) + d(y,x) \text{ (property 2)} \\ \Rightarrow & 0 \leq d(x,y) + d(x,y) \text{ (property 3)} \\ \Rightarrow & 0 \leq d(x,y) \end{aligned}$$

However, within this course, when asked to show that a function is metric, you'll have to show that it has the 4 listed properties.

Example 2.4: The SNCF distance
Let V be a normed vector space equipped with the norm $\|.\|$. We define the function $d : V \times V \to \mathbb{R}$ by:

- $\forall \mathbf{x}, \mathbf{y} \in V, \mathbf{x} \neq \mathbf{y} \Rightarrow d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x}\| + \|\mathbf{y}\|$
- $\forall \mathbf{x} \in V, d(\mathbf{x}, \mathbf{x}) = 0$

Then (V, d) is a metric space. This metric is inspired by the French railway system where most of the train lines go through Paris. Let's show that d is indeed a metric:

Proof. Let $\mathbf{x}, \mathbf{y} \in V$ then if $\mathbf{x} \neq \mathbf{y}$ we have $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x}\| + \|\mathbf{y}\| \geq 0$ by property of the norm. Now if $\mathbf{x} = \mathbf{y}$ then $d(\mathbf{x}, \mathbf{y}) = 0$ so all in all: $\underline{\forall \mathbf{x}, \mathbf{y} \in V, d(\mathbf{x}, \mathbf{y}) \geq 0}$
Let $\mathbf{x}, \mathbf{y} \in V$, if $\mathbf{x} = \mathbf{y}$ then $d(\mathbf{x}, \mathbf{y}) = 0$. Now suppose that $d(\mathbf{x}, \mathbf{y}) = 0$ and $\mathbf{x} \neq \mathbf{y}$. Then $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x}\| + \|\mathbf{y}\| = 0$ but the norms being non-negative, it comes that: $\|\mathbf{x}\| = \|\mathbf{y}\| = 0$ thus $\mathbf{x} = \mathbf{y} = 0$ which is a contradiction. Hence: $\underline{\forall \mathbf{x}, \mathbf{y} \in V,\ d(\mathbf{x}, \mathbf{y}) = 0 \iff \mathbf{x} = \mathbf{y}}$
For $\mathbf{x}, \mathbf{y} \in V$ we have $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x}\| + \|\mathbf{y}\| = \|\mathbf{y}\| + \|\mathbf{x}\| = d(\mathbf{y}, \mathbf{x})$: $\underline{\forall \mathbf{x}, \mathbf{y} \in V,\ d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})}$.
Finally, let $\mathbf{x}, \mathbf{y}, \mathbf{z} \in V$, if $\mathbf{x} = \mathbf{y}$ then $d(\mathbf{x}, \mathbf{y}) = 0 \leq d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y})$. If $\mathbf{x} = \mathbf{z}$ then $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{z}, \mathbf{y}) \leq d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y})$, same if $\mathbf{y} = \mathbf{z}$. Now if none of these equalities is true, we have $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x}\| + \|\mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\| + 2\|\mathbf{z}\| = \|\mathbf{x}\| + \|\mathbf{z}\| + \|\mathbf{z}\| + \|\mathbf{y}\| = d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y})$. So $\underline{\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in V,\ d(\mathbf{x}, \mathbf{y}) \leq d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y})}$.
And we have shown that (V, d) is a metric space □

In Example 2.4 we have defined a metric in a normed vector space by using a convoluted definition involving the norm. It turns out that there is a more straightforward definition of a metric in a normed vector space, this definition being commonly used.

Proposition 2.1: Metric in a normed vector space
Let V be a vector space equipped with the norm $\|.\|$. Let the d be the function defined by:

$$d : \begin{cases} V \times V & \to \mathbb{R} \\ (\mathbf{x}, \mathbf{y}) & \mapsto \|\mathbf{x} - \mathbf{y}\| \end{cases}$$

Then (V, d) is a metric space.

Proof. I leave the proof as an exercise, you just have to show that the 4 properties of a metric hold for d. □

2.2.2 Convergence

From now on, we will consider the concepts seen in $\mathbb{R}$ but generalised to a metric space (S, d). The generalisations will be pretty straightforward, mostly consisting in swapping the absolute value of the difference with the distance.

Definition 2.4: Convergence in a metric space
Let (S, d) be a metric space and (a_n) a sequence in S. (a_n) is said to converge to a limit $l \in S$ if and only if:

$$\boxed{\forall \epsilon > 0,\ \exists N \in \mathbb{N} \text{ such that } \forall n > N,\ d(a_n, l) < \epsilon}$$

As you can see, the only difference between Definition 2.1 and Definition 2.4 is that the absolute value of the difference has been replaced by the distance of the metric space.

Example 2.5:
Consider $\mathbb{R}^2$ equipped with the l_2-norm. Then according to Proposition 2.1, if d_2 is the distance defined from the l_2-norm, $(\mathbb{R}^2, d_2)$ is a metric space. Now let (a_n) the sequence in $\mathbb{R}^2$ defined by:

$$\forall n \in \mathbb{N},\ a_n = \begin{pmatrix} 1 - \frac{1}{n+1} \\ \frac{1}{n+2} \end{pmatrix}$$

Show that (a_n) converges and give its limit.
Let's show that its limit is $l = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$. Let $\epsilon > 0$, for $n \in \mathbb{N}$, we have

$$d_2(a_n, l) = \left\| \begin{pmatrix} 1 - \frac{1}{n+1} \\ \frac{1}{n+2} \end{pmatrix} - \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right\|_2 = \left\| \begin{pmatrix} -\frac{1}{n+1} \\ \frac{1}{n+2} \end{pmatrix} \right\|_2 = \sqrt{\left(\frac{1}{n+1}\right)^2 + \left(\frac{1}{n+2}\right)^2}$$

Let $N = \left\lceil \frac{\epsilon}{\sqrt{2}} \right\rceil$, then for $n > N$, we have

$$d_2(a_n, l) < \epsilon$$

So (a_n) converges to l in $(\mathbb{R}^2, d_2)$

When working in $\mathbb{R}$ there is only one distance but for some space S there might be multiple possible distance. For example you have seen that $\mathbb{R}^2$ can be equipped with a lot of different norms and according to Proposition 2.1 there are as many possible different distances. One legitimate question that arises is: does convergence with one distance means convergence with another? The answer is no but to show that, we have to consider spaces that are outside the scope of this course.
However when working with just one metric, some properties you saw concerning the limit in $\mathbb{R}$ also apply in a metric space. Among them:

Proposition 2.2: Uniqueness of the limit
Let (S, d) a metric space and (a_n) a sequence in S. If (a_n) is converging, then its limit is unique.

Proof. Suppose that (a_n) is converging and has two limits l_1 and l_2. Let $\epsilon > 0$, by definition of convergence, there there are $N_1, N_2 \in \mathbb{N}$ such that:

$$\forall n > N_1, d(a_n, l_1) < \frac{\epsilon}{2}$$

And

$$\forall n > N_2, d(a_n, l_2) < \frac{\epsilon}{2}$$

Let $n > \max(N_1, N_2)$, then we have:

$$d(l_1, l_2) \leq d(l_1, x_n) + d(x_n, l_2) < \epsilon$$

That being true for all $\epsilon > 0$, it comes $\underline{l_1 = l_2}$ □

2.2.3 Cauchy Sequence

Definition 2.5: Cauchy sequence in a metric space
Let (S, d) a metric space and (a_n) a sequence in S. Then (a_n) is said to be a Cauchy sequence if and only if:

$$\boxed{\forall \epsilon > 0,\ \exists N \in \mathbb{N} \text{ such that } \forall n, m > N,\ d(a_n, a_m) < \epsilon}$$

As for the convergence, the definition of a Cauchy sequence in a metric space is similar to the one in $\mathbb{R}$. But that does not mean that everything that is valid in $\mathbb{R}$ is also valid in any metric space. For example the Cauchy test is no longer valid, instead we have:

Theorem 2.2:
Let (S, d) a metric space and (a_n) a sequence in S: if (a_n) is convergent, then it is a Cauchy sequence

Remark. While in $\mathbb{R}$ we have an equivalence, in any metric space we are only assured of one implication, this implication being proven in a similar fashion as in $\mathbb{R}$. The bast way to show that it is an implication is to have a counter example to the converse way.

Example 2.6: Example of non convergent Cauchy sequence in a metric space
Consider $\mathbb{Q}$ the space of rational numbers with the metric d defined by the absolute value of the difference, then $(\mathbb{Q}, d)$ is a metric space. Now consider the sequence (a_n) defined by:

$$\forall n > 0,\ a_n = \sum_{k=1}^{n} \frac{1}{k^2}$$

Then (a_n) is a sequence of rational numbers. For $n, m > 0$, we can assume $n \geq m$ and we have:

$$a_n - a_m = \sum_{k=m+1}^{n} \frac{1}{k^2}$$

So

$$|a_n - a_m| = \left| \sum_{k=m+1}^{n} \frac{1}{k^2} \right| = \sum_{k=m+1}^{n} \frac{1}{k^2} \leq \sum_{k=m+1}^{n} \frac{1}{k(k+1)} = \frac{1}{m+1} - \frac{1}{n+1}$$

In a similar way as in Example 2.2 we can show that (a_n) is a Cauchy sequence. But it can be shown that:

$$\lim_{n \to \infty} \sum_{k=1}^{n} \frac{1}{k^2} = \frac{\pi^2}{6}$$

$\frac{\pi^2}{6}$ being not in $\mathbb{Q}$, (a_n) does not converge in $(\mathbb{Q}, d)$.

The relation between Cauchy sequences and convergence is actually quite important as it is used to define a type of space: the complete space:

Definition 2.6: Complete space
Let (S, d) be a metric space. Then it is said to be a complete space if and only if every Cauchy sequence in S is also converging in S.

Intuitively, a complete space is a space without "holes". For example $\mathbb{Q}$ equipped with the usual distance is not complete as $\frac{\pi^2}{6} \notin \mathbb{Q}$ but we have found a sequence converging of elements of $\mathbb{Q}$ that converges to this point so $\frac{\pi^2}{6}$ is a "hole" in $\mathbb{Q}$

Example 2.7:
For any $k > 0$, $\mathbb{R}^k$ equipped with a metric is complete.

From this Definition 2.6 and Theorem 2.2 it comes:

Theorem 2.3:
Let (S, d) be a complete metric space and (a_n) a sequence in S. Then (a_n) is converging if and only if it is a Cauchy sequence.

2.3 Fixed Point Equations

Definition 2.7: Fixed point
Let S be a non-empty set and $f : S \to S$ a function form S to itself. Then $p \in S$ is called a fixed point if:

$$\boxed{f(p) = p}$$

Example 2.8:
Let's find the fixed points of $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^2$. If p is a fixed point, then we have:

$$f(p) = p \iff p^2 = p \iff p^2 - p = 0 \iff p(p-1) = 0$$

So $p = 0$ and $p = 1$ or fixed points of f. Graphically, the fixed points are located at the intersection between the line of equation $y = x$ and the plot of equation $y = f(x)$ as represented in Figure 2.

Definition 2.8: Contraction
Let (S, d) a metric space and $f : S \to S$. f is called a contraction of S if there exists $0 < \alpha < 1$ called the contraction constant such that:

$$\boxed{\forall x, y \in S,\ d(f(x), f(y)) \leq \alpha d(x, y)}$$

Example 2.9:
We consider the metric space $(\mathbb{R}^2, d_1)$ where d_1 is the metric induced by the l_1-norm. Let $f : \mathbb{R}^2 \to \mathbb{R}^2$ the function defined by $f(x, y) = (0.25x, -0.5y)$. Then by taking $\alpha = 0.5$, we have:

$$\forall x, y \in \mathbb{R}^2, \|f(x) - f(y)\|_1 \leq 0.5\|x - y\|_1$$

These preliminary definitions allow us to introduce the following theorem:

Theorem 2.4: Fixed point theorem
Let (S, d) a complete metric space and f a continuous contraction of S. Then f has a unique fixed point.

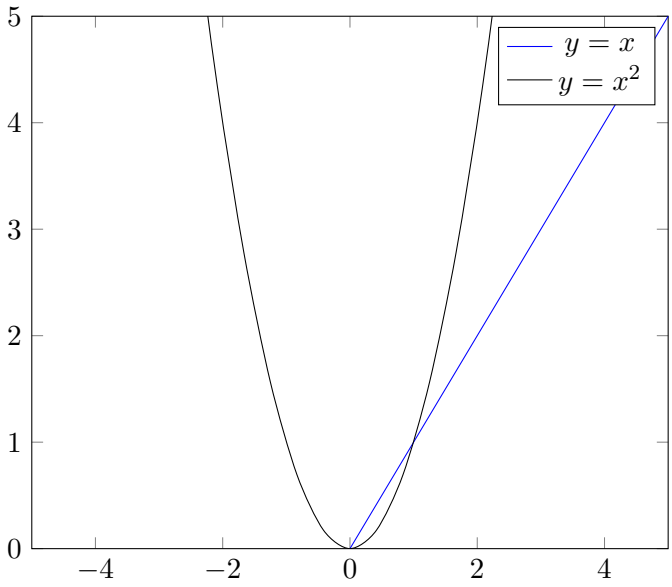

Figure 2: Graphical representation of fixed points

Proof. Let (S, d) a complete metric space and f a continuous contraction of S with contraction constant $0 < \alpha < 1$.
Let $x \in S$ and consider the sequence (x_n) in S defined by:

$$\begin{cases} x_0 = x \\ x_{n+1} = f(x_n) \end{cases}$$

We then have for $k \in \mathbb{N}$:

$$d(x_{k+1}, x_k) = d(f(x_k), f(x_{k-1})) \leq \alpha d(x_k, x_{k-1}) \leq \ldots \leq \alpha^k d(x_1, x_0)$$

Now for $m > n$:

$$d(x_m, x_n) \leq \sum_{k=n}^{m-1} d(x_{k+1}, x_k) \leq \sum_{k=n}^{m-1} \alpha^k d(x_1, x_0) = \alpha^n d(x_1, x_0) \frac{1 - \alpha^{m-n}}{1 - \alpha} < \frac{\alpha^n d(x_1, x_0)}{1 - \alpha}$$

So for $\epsilon > 0$, if we take $N = \left\lceil \frac{1}{\ln(\alpha)} \ln\left(\frac{\epsilon(1-\alpha)}{d(x_1 - x_0)}\right) \right\rceil$, for $m, n > N$ we have:

$$d(x_m, x_n) < \epsilon$$

So (x_n) is a Cauchy sequence and S being complete, (x_n) is converging in S. Let $l \in S$ be its limit. f being continuous in S, we have:

$$\lim_{n \to \infty} f(x_n) = f(\lim_{n \to \infty} x_n)$$

But $f(\lim_{n\to\infty} x_n) = f(l)$ and $\lim_{n\to\infty} f(x_n) = \lim_{n\to\infty} x_{n+1} = l$. So $f(l) = l$ and f has a fixed point.
Now suppose that $p_1, p_2 \in S$ are two fixed points of S, we then have:

$$d(p_1, p_2) = d(f(p_1), f(p_2)) \leq \alpha d(p_1, p_2)$$

As $0 < \alpha < 1$ it comes that $d(p_1, p_2) = 0$ and $p_1 = p_2$ □

Example 2.10:
Take $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = -0.5x + 1$ for $x \in \mathbb{R}$. $\mathbb{R}$ equipped with the usual distance is complete. f is continuous on $\mathbb{R}$ and for $x, y \in \mathbb{R}$ we have:

$$|f(x) - f(y)| \leq 0.5|x - y|$$

So f is a contraction of $\mathbb{R}$ and according to Theorem 2.4 it has a unique fixed point.

3 Iterative Solutions of Linear Equations

3.1 The Problem

In this part we will consider real matrices but it can easily be generalised to complex matrices.
Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{b} \in \mathbb{R}^n$. We wish to solve $\mathbf{Ax} = \mathbf{b}$. If solved via Gaussian elimination, it would take $O(n^3)$ operations which is impractical for large n. Luckily, there exist method that can approximate the solution by approaching it iteratively.

3.2 General Method

3.2.1 Presentation

Let $\mathbf{A}, \mathbf{R}, \mathbf{G} \in \mathbb{R}^{n\times n}$ be square matrices, $\mathbf{G}$ being a non-singular matrix. Then a splitting of $\mathbf{A}$ is defined as:

$$\boxed{\mathbf{A} = \mathbf{G} + \mathbf{R}}$$

Now suppose we want to solve the equation $\mathbf{Ax} = \mathbf{b}$, then we can write:

$$\begin{aligned} &\mathbf{Ax} = \mathbf{b} \\ \iff &(\mathbf{G} + \mathbf{R})\mathbf{x} = \mathbf{b} \\ \iff &\mathbf{Gx} = \mathbf{b} - \mathbf{Rx} \\ \iff &\mathbf{x} = \mathbf{G}^{-1}\mathbf{b} - \mathbf{G}^{-1}\mathbf{Rx} \end{aligned}$$

If we note $\mathbf{M} = -\mathbf{G}^{-1}\mathbf{R}$ and $\mathbf{c} = \mathbf{G}^{-1}\mathbf{b}$ then the equation to solve is now:

$$\mathbf{x} = \mathbf{Mx} + \mathbf{c}$$

So if we define $f : \mathbb{R}^n \to \mathbb{R}^n$ by $f(\mathbf{x}) = \mathbf{Mx} + \mathbf{c}$ then a solution to the equation is a fixed point of f. So if we look back at the fixed point theory one way to find such point, and hence a solution to the equation, is to define a sequence $(\mathbf{x}_k)$ together with some starting point $\mathbf{x}_0$ such that:

$$\mathbf{x}_{k+1} = \mathbf{Mx}_k + \mathbf{c}$$

These considerations lead to the following theorem:

Theorem 3.1:
Let $\mathbf{M} \in \mathbb{R}^{n\times n}$ and $c \in \mathbb{R}^n$. Let $\|.\|$ be a consistent norm on $\mathbb{R}^{n\times n}$, if $\|\mathbf{M}\| < 1$ then the sequence $(\mathbf{x}_k)$ defined by $\mathbf{x}_{k+1} = \mathbf{Mx}_k + \mathbf{c}$ converges for any starting point $\mathbf{x}_0$.

Proof. $\mathbb{R}^n$ equipped with the distance derived form a norm $\|.\|$ is complete. We have:

$$\|f(\mathbf{x}) - f(\mathbf{y})\| = \|\mathbf{Mx} + \mathbf{c} - (\mathbf{My} + \mathbf{c})\| = \|\mathbf{Mx} - \mathbf{My}\| = \|\mathbf{M}(\mathbf{x} - \mathbf{y})\| \leq \|\mathbf{M}\|\|\mathbf{x} - \mathbf{y}\|$$

By consistency of the matrix norm. So f is a contraction of $\mathbb{R}^n$, plus being a linear mapping, it is continuous. So according to Theorem 2.4 it has a unique fixed point $\mathbf{x}$.
Now for $n \in \mathbb{N}$:

$$\mathbf{x}_k - \mathbf{x} = (\mathbf{Mx}_{k-1} + \mathbf{c}) - (\mathbf{Mx} + \mathbf{c}) = \mathbf{M}(\mathbf{x}_{k-1} - \mathbf{x}) = \mathbf{M}^k(\mathbf{x}_0 - \mathbf{x})$$

So

$$\|\mathbf{x}_k - \mathbf{x}\| \leq \|\mathbf{M}\|^k\|\mathbf{x}_0 - \mathbf{x}\|$$

By consistency of the matrix norm. As $\|\mathbf{M}\| < 1$, $\|\mathbf{M}\|^k\|\mathbf{x}_0 - \mathbf{x}\| \to 0$ so $\|\mathbf{x}_k - \mathbf{x}\| \to 0$, that is $(\mathbf{x}_k)$ converges to $\mathbf{x}$. □

Example 3.1:
Let $\mathbf{A} = \begin{bmatrix} 5 & 2 \\ 3 & 4 \end{bmatrix} \in \mathbb{R}^{2\times 2}$. We work in $\mathbb{R}^2$ equipped with the l_1-norm. Let $\mathbf{G} = \begin{bmatrix} 5 & 0 \\ 0 & 4 \end{bmatrix}$ and $\mathbf{R} = \begin{bmatrix} 0 & 2 \\ 3 & 0 \end{bmatrix}$, then $\mathbf{A} = \mathbf{G} + \mathbf{R}$ and $\mathbf{G}$ is non singular. Let $\mathbf{b} \in \mathbb{R}^2$ we want to solve $\mathbf{Ax} = \mathbf{b}$. Let $\mathbf{M} = -\mathbf{G}^{-1}\mathbf{R} = -\begin{bmatrix} 0 & 2/5 \\ 3/4 & 0 \end{bmatrix}$. Then $\|\mathbf{M}\|_1 = \frac{3}{4} < 1$ and it is possible to find a solution through iterative method.

3.2.2 Efficient Choice of Splitting

Rule of thumb: the rate of convergence r is such that $r \propto -\log_{10}\|\mathbf{M}\|$. The rate of convergence is inversely proportional to the number of steps needed to reach a given precision so the smaller $\|\mathbf{M}\|$ is, the quicker the algorithm is going to converge to the solution. For this reason, we want to choose a splitting such that:

1. $-\mathbf{G}^{-1}\mathbf{Rx}$ and $\mathbf{G}^{-1}\mathbf{b}$ are easy to compute
2. $\|\mathbf{M}\|$ is small

3.3 Common Methods

3.3.1 Common Splitting

In the following sections, we will assume without loss of generality that $\mathbf{A}$ has no zeros on the diagonal (if that's not the case it is always possible to perform a change of basis to achieve that). Then we can write:

$$\mathbf{A} = \mathbf{D} + \mathbf{L} + \mathbf{U}$$

Where $\mathbf{D}$ is the diagonal of $\mathbf{A}$ and $\mathbf{L}, \mathbf{U}$ are the strict lower and upper triangular parts of $\mathbf{A}$ respectively.

Example 3.2:
Let $\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$ then we have the following splitting:

$$\mathbf{D} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 9 \end{bmatrix}, \ \mathbf{L} = \begin{bmatrix} 0 & 0 & 0 \\ 4 & 0 & 0 \\ 7 & 8 & 0 \end{bmatrix} \text{ and } \mathbf{U} = \begin{bmatrix} 0 & 2 & 3 \\ 0 & 0 & 6 \\ 0 & 0 & 0 \end{bmatrix}$$

Example 3.3:
Suppose now that we want to solve $\mathbf{Ax} = \mathbf{b}$ with $\mathbf{A} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$. The diagonal elements of A being zeros, we can not directly apply the splitting. Consider the change of basis represented by the matrix $\mathbf{C} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$. Then $\mathbf{C}^{-1} = \mathbf{C}$ and:

$$\begin{aligned} & \mathbf{Ax} = \mathbf{b} \\ \iff & \mathbf{C}^{-1}\mathbf{Ax} = \mathbf{C}^{-1}\mathbf{b} \\ \iff & (\mathbf{C}^{-1}\mathbf{AC})\mathbf{C}^{-1}\mathbf{x} = \mathbf{C}^{-1}\mathbf{b} \end{aligned}$$

But $\mathbf{C}^{-1}\mathbf{AC} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} = \mathbf{B}$. By denoting $\mathbf{C}^{-1}\mathbf{x} = \mathbf{y}$ and $\mathbf{C}^{-1}\mathbf{b} = \mathbf{c}$ it is equivalent to solve

$$\mathbf{By} = \mathbf{c}$$

Where $\mathbf{B}$ can be split according to the given method. Solving this equation will give $\mathbf{y}$ and $\mathbf{x}$ can be retrieved as $\mathbf{x} = \mathbf{Cy}$

3.3.2 Jacobi Method

Let $A \in \mathbb{R}^{n\times n}$ with the splitting seen previously:

$$\mathbf{A} = \mathbf{D} + \mathbf{L} + \mathbf{U}$$

$\mathbf{D}$ being a diagonal matrix with only non zero diagonal element, it is invertible. We denote $\mathbf{R} = \mathbf{L} + \mathbf{U}$. Now suppose that for $\mathbf{b} \in \mathbb{R}^n$ we want to solve the equation $\mathbf{Ax} = \mathbf{b}$. Then:

$$\begin{aligned} & \mathbf{Ax} = \mathbf{b} \\ \iff & (\mathbf{D} + \mathbf{R})\mathbf{x} = \mathbf{b} \\ \iff & \mathbf{x} = -\mathbf{D}^{-1}\mathbf{Rx} + \mathbf{D}^{-1}\mathbf{b} \\ \iff & \mathbf{x} = \mathbf{Mx} + \mathbf{c} \end{aligned}$$

Where $\mathbf{M} = -\mathbf{D}^{-1}\mathbf{R}$ and $\mathbf{c} = \mathbf{D}^{-1}\mathbf{b}$. As previously we consider the sequence $(\mathbf{x}^{(k)})$ in $\mathbb{R}^n$ defined by $\mathbf{x}^{(k+1)} = \mathbf{Mx}^{(k)} + \mathbf{c}$ and $\mathbf{x}^{(0)} \in \mathbb{R}^n$. What is interesting with this splitting is that the elements of $\mathbf{x}^{(k+1)}$ can easily be computed. We have

$$\mathbf{D} = \begin{bmatrix} a_{11} & & \\ & \ddots & \\ & & a_{nn} \end{bmatrix}$$

So

$$\mathbf{D}^{-1} = \begin{bmatrix} 1/a_{11} & & \\ & \ddots & \\ & & 1/a_{nn} \end{bmatrix}$$

We also have:

$$\mathbf{R} = \begin{bmatrix} 0 & a_{1,2} & \dots & a_{1,n} \\ a_{2,1} & \ddots & & \vdots \\ \vdots & & \ddots & a_{n-1,n} \\ a_{n,n} & \dots & a_{n,n-1} & 0 \end{bmatrix}$$

For $\mathbf{x} \in \mathbb{R}^n$, we thus have:

$$\mathbf{Mx} = -\mathbf{D}^1\mathbf{Rx} = -\mathbf{D}^{-1} \begin{bmatrix} \sum_{j\neq 1} a_{1,j}x_j \\ \sum_{j\neq 2} a_{2,j}x_j \\ \vdots \\ \sum_{j\neq n} a_{n,j}x_j \end{bmatrix} = \begin{bmatrix} -\frac{1}{a_{1,1}} \sum_{j\neq 1} a_{1,j}x_j \\ -\frac{1}{a_{2,2}} \sum_{j\neq 2} a_{2,j}x_j \\ \vdots \\ -\frac{1}{a_{n,n}} \sum_{j\neq n} a_{n,j}x_j \end{bmatrix}$$

As $\mathbf{x}^{(k+1)} = \mathbf{Mx}^{(k)} + \mathbf{c}$, it comes that:

$$x_i^{(k+1)} = \frac{1}{a_{i,i}} \left(b_i - \sum_{j\neq i} a_{i,j}x_j^{(k)} \right)$$

So in order to compute the i^{th} element of $x^{(k+1)}$ we only need the $b, x^{(k)}$ and the i^{th} row of A. This is great for parallelisation as each row of A can be distributed to different and independent processes so that the elements of $x^{(k+1)}$ can be computed at the same time.

3.3.3 Gauss-Seidel Method

For this method we use the following splitting:

$$\mathbf{A} = (\mathbf{D} + \mathbf{L}) + \mathbf{U}$$

As previously,$\mathbf{D}$ being a diagonal matrix with only non zero diagonal elements and $\mathbf{L}$ being a strictly lower triangle matrix, $\mathbf{D} + \mathbf{L}$ is non singular. We then have:

$$\begin{aligned} & \mathbf{Ax} = \mathbf{b} \\ \iff & (\mathbf{D} + \mathbf{L})\mathbf{x} = \mathbf{b} - \mathbf{Ux} \\ \iff & \mathbf{x} = \mathbf{Mx} + \mathbf{c} \end{aligned}$$

Where $\mathbf{M} = -(\mathbf{D} + \mathbf{L})^{-1}\mathbf{U}$ and $\mathbf{c} = (\mathbf{D} + \mathbf{L})^{-1}\mathbf{b}$. As previously, we consider the sequence $\mathbf{x}^{(k)}$ defined by $\mathbf{x}^{(k+1)} = \mathbf{Mx}^{(k)} + \mathbf{c}$ and $\mathbf{x}^{(0)} \in \mathbb{R}^n$. In this case, we take advantage of the shape of the matrix $\mathbf{D} + \mathbf{L}$ to compute the elements of $\mathbf{x}^{(k+1)}$. Indeed, we have:

$$\mathbf{D} + \mathbf{L} = \begin{pmatrix} a_{1,1} & 0 & \dots & 0 \\ a_{2,1} & a_{2,2} & \ddots & \vdots \\ \vdots & & \ddots & 0 \\ a_{n,1} & \dots & \dots & a_{n,n} \end{pmatrix}$$

And

$$\mathbf{U} = \begin{pmatrix} 0 & a_{1,2} & \dots & a_{1,n} \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & a_{n-1,n} \\ 0 & \dots & \dots & 0 \end{pmatrix}$$

Retracing our steps, we have:

$$\mathbf{x}^{(k+1)} = \mathbf{Mx}^{(k)} + \mathbf{c} \iff (\mathbf{D} + \mathbf{L})x^{(k+1)} = -\mathbf{Ux}^{(k)} + \mathbf{b}$$

For $1 \leq i \leq n$, it comes:

$$\sum_{j \leq i} a_{i,j} x_j^{(k+1)} = -\sum_{j > i} a_{i,j} x_j^{(k)} + b_i$$

So the i^{th} element of $x^{(n+1)}$ can be computed thanks to $\mathbf{A}$, $\mathbf{x}^{(n)}$, $\mathbf{b}$ and the k^{th} elements of $\mathbf{x}^{(n+1)}$ for $k < i$. The update being computed with quantities that are more recent, the convergence is faster.

3.3.4 Convergence of these Methods

Definition 3.1:
Let $\mathbf{A} \in \mathbb{R}^{n \times n}$. It is said to be strictly row diagonally dominant if:

$$\boxed{\forall i, |a_{ii}| > \sum_{j \neq i} |a_{ij}|}$$

This allows us to give sufficient condition for the Jacobi and Gauss-Seidel methods to converge:

Theorem 3.2:
Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ and we want to solve the equation $\mathbf{Ax} = \mathbf{b}$. If $\mathbf{A}$ is strictly row diagonally dominant then the Jacobi and Gauss-Seidel methods will converge. G-S is faster.

Proof. We show this for the Jacobi method. In this case, we consider the sequence $(\mathbf{x}^{(k)})$ defined by $\mathbf{x}^{(k+1)} = \mathbf{M}\mathbf{x}^{(k)} + \mathbf{c}$ with $\mathbf{M} = \mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})$.
$\mathbf{A}$ being strictly row diagonally dominant, we have for $1 \leq i \leq n$: $|a_{ii}| > \sum_{j \neq i} |a_{ij}|$, that is:

$$\frac{1}{|a_{ii}|} \sum_{j \neq i} |a_{ij}| < 1$$

That being the case for all $1 \leq i \leq n$, we have:

$$\max_{1 \leq i \leq n} \frac{1}{|a_{ii}|} \sum_{j \neq i} |a_{ij}| < 1$$

Or:

$$\|\mathbf{M}\|_\infty < 1$$

According to Theorem 3.1: the sequence $(\mathbf{x}^{(k)})$ converges and its limit is the solution to the equation.
Now for the Gauss-Seidel method: InA11 this case $\mathbf{M} = -(\mathbf{D} + \mathbf{L})^{-1}\mathbf{U}$. Let $\lambda \in \mathbb{C}$ an eigenvalue of $\mathbf{M}$ and $\mathbf{v} \in \mathbb{C}^n$ a corresponding eigenvector. Without loss of generality, we assume that $\|\mathbf{v}\|_\infty = 1$. Let i such that $|v_i| = 1$, then for $j \neq i$: $|v_j| \leq 1$. We have:

$$\mathbf{M}\mathbf{v} = \lambda\mathbf{v} \iff -(\mathbf{D} + \mathbf{L})^{-1}\mathbf{U}\mathbf{v} = \lambda\mathbf{v} \iff -\mathbf{U}\mathbf{v} = \lambda(\mathbf{D} + \mathbf{L})\mathbf{v}$$

This leads to:

$$-\sum_{j > i_0} a_{i_0,j} v_j = \lambda \sum_{j \leq i_0} a_{i_0,j} v_j \iff -\lambda a_{i_0,i_0} v_{i_0} = \lambda \sum_{j < i_0} a_{i_0,j} v_j + \sum_{j > i_0} a_{i_0,j} v_j$$

Taking the modulus:

$$|\lambda||a_{i_0,i_0}| \leq |\lambda| \sum_{j < i_0} |a_{i_0,j}| + \sum_{j > i_0} |a_{i_0,j}| \iff |\lambda| \leq \frac{\sum_{j > i_0} |a_{i_0,j}|}{|a_{i_0,i_0}| - \sum_{j < i_0} |a_{i_0,j}|}$$

But A being strictly row diagonally dominant, we have

$$\frac{\sum_{j > i_0} |a_{i_0,j}|}{|a_{i_0,i_0}| - \sum_{j < i_0} |a_{i_0,j}|} < 1$$

Hence:

$$|\lambda| < 1$$

That being true for all eigenvalues of $\mathbf{M}$, we have $\rho(\mathbf{A}) < 1$. Let $\epsilon > 0$ such that $\rho(\mathbf{A}) + \epsilon < 1$ then according to Lemma 3.1 there exist an induced norm such that $\|A\| < \rho(\mathbf{A}) + \epsilon < 1$. An induced norm being also consistent, Theorem 3.1 then shows the convergence of the sequence. □

This Lemma and its proof are absolutely out of the scope of this course!!!!

Lemma 3.1:
Let $\mathbf{A} \in \mathbb{R}^{n \times n}$. We call $\rho(\mathbf{A}) = \max_{\lambda \in Sp(\mathbf{A})} |\lambda|$ the spectral radius of $\mathbf{A}$. Let $\epsilon > 0$, then there exists an induced norm such that $\|\mathbf{A}\| < \rho(\mathbf{A}) + \epsilon$

Proof. Let $\mathbf{A} \in \mathbb{R}^{n \times n}$, then from the Schur decomposition, there exist a unitary matrix $\mathbf{Q} \in \mathbb{C}^{n \times n}$ and an upper triangular matrix $\mathbf{T} \in \mathbb{C}^{n \times n}$ such that

$$\mathbf{A} = \mathbf{Q}\mathbf{T}\mathbf{Q}^{-1}$$

Then the eigenvalues of $\mathbf{A}$ are the diagonal elements of $\mathbf{T}$ and $\mathbf{Q}$ being unitary, its columns are orthogonal one to another. Let $(\mathbf{f_1}, \ldots, \mathbf{f_n})$ be the columns of $\mathbf{Q}$. Then for $1 \leq j \leq n$ we have:

$$\mathbf{Af_j} = \sum_{i \leq j} \mathbf{t_{ij} f_i}$$

Let $\nu \in]0,1[$, we define $\mathbf{e}_j = \nu^j \mathbf{f}_j$, then $(\mathbf{e}_1, \ldots, \mathbf{e}_n)$ is a basis of $\mathbb{C}^n$ and we have:

$$\mathbf{A}\mathbf{e}_j = \mathbf{A}(\nu^{j-1}\mathbf{f}_j) = \nu^{j-1} \sum_{i \leq j} t_{ij} \mathbf{f}_i = \sum_{i \leq j} \nu^{j-i} t_{ij} \mathbf{e}_i$$

Now take $\mathbf{x} \in \mathbb{R}^n$, as $\mathbb{R}^n \subset \mathbb{C}^n$, we also have $\mathbf{x} \in \mathbb{C}^n$ and $(\mathbf{e}_1, \ldots, \mathbf{e}_n)$ being a basis of $\mathbb{C}^n$ we can write $\mathbf{x} = \sum x_j \mathbf{e}_j$. We define the norm $\|.\|$ as $\|\mathbf{x}\| = \sqrt{\sum \|x_j\|^2}$. We have:

$$\mathbf{Ax} = \mathbf{A} \sum x_j \mathbf{e}_j = \sum x_j \mathbf{A}\mathbf{e}_j = \sum_{j=1}^{n} x_j \sum_{i \leq j} \nu^{j-i} t_{ij} \mathbf{e}_i = \sum_{i=1}^{n} \left(\sum_{j=i}^{n} x_j \nu^{j-i} t_{ij} \right) \mathbf{e}_i$$

So

$$\|\mathbf{Ax}\|^2 = \sum_{i=1}^{n} \left| \sum_{j=i}^{n} x_j \nu^{j-i} t_{ij} \right|^2$$

But

$$\left| \sum_{j=i}^{n} x_j \nu^{j-i} t_{ij} \right|^2 \leq \left(\sum_{j=i}^{n} |x_j \nu^{j-i} t_{ij}| \right)^2$$

And

$$\left(\sum_{j=i}^{n} |x_j \nu^{j-i} t_{ij}| \right)^2 = \sum_{k,l=i}^{n} |x_k \nu^{k-i} t_{ik}||x_l \nu^{l-i} t_{il}| = |x_i t_{ii}|^2 + \sum_{k,l=i,(k,l)\neq(i,i)}^{n} |x_k \nu^{k-i} t_{ik}||x_l \nu^{l-i} t_{il}|$$

Now t_{ii} being a diagonal element of $\mathbf{T}$, it is an eigenvalue of $\mathbf{A}$ and we have

$$|t_{ii}| \leq \rho(\mathbf{A})$$

So

$$|x_i t_{ii}|^2 \leq |x_i|^2 \rho(\mathbf{A})^2$$

We also have:

$$|x_k \nu^{k-i} t_{ik}||x_l \nu^{l-i} t_{il}| = \nu^{k+l-2i} |x_k||x_l||t_{ik}||t_{il}|$$

For $k, l \geq i$ with $(k,l) \neq (i,i)$ we have $k+l-2i \geq 1$ and as $\nu \in]0,1[$ it comes that $\nu^{k+l-2i} \leq \nu$. We denote $M = \max_{1 \leq a,b \leq n} |t_{ab}|$ then we have:

$$|x_k \nu^{k-i} t_{ik}||x_l \nu^{l-i} t_{il}| \leq \nu M^2 |x_k||x_l|$$

But

$$|x_k||x_l| \leq (\max_{1 \leq k \leq n} x_k)^2 \leq \sum_{k=1}^{n} |x_k|^2 = \|\mathbf{x}\|^2$$

So:

$$\left(\sum_{j=i}^{n} |x_j \nu^{j-i} t_{ij}| \right)^2 \leq |x_i|^2 \rho(\mathbf{A})^2 + n^2 \nu M^2 \|\mathbf{x}\|^2$$

This leads to:

$$\|\mathbf{Ax}\|^2 \leq \sum_{i=1}^{n} (|x_i|^2 \rho(\mathbf{A})^2 + n^2 \nu M^2 \|x\|^2) = \|\mathbf{x}\|^2 \rho(\mathbf{A})^2 + \|\mathbf{x}\|^2 n^3 \nu M^2$$

And

$$\frac{\|\mathbf{Ax}\|^2}{\|\mathbf{x}\|^2} \leq \rho(\mathbf{A})^2 + n^3 \nu M^2 \Rightarrow \frac{\|\mathbf{Ax}\|}{\|\mathbf{x}\|} \leq \rho(\mathbf{A}) \sqrt{1 + \nu \frac{n^3 M^2}{\rho(\mathbf{A})^2}}$$

This inequality being true for all $x \in \mathbb{R}^n$, it comes:

$$\|\mathbf{A}\| \leq \rho(\mathbf{A}) \sqrt{1 + \nu \frac{n^3 M^2}{\rho(\mathbf{A})^2}}$$

Now let $\epsilon > 0$, then it is possible to choose $\nu \in]0, 1[$ such that $\rho(\mathbf{A}) \sqrt{1 + \nu \frac{n^3 M^2}{\rho(\mathbf{A})^2}} < \rho(\mathbf{A}) + \epsilon$
Hence:

$$\|\mathbf{A}\| \leq \rho(\mathbf{A}) + \epsilon$$

□

3.3.5 Other Results

Proposition 3.1: Influence of the condition number on the convergence
These techniques will converge faster when the condition number of $\mathbf{A}$ is small. They will converge slower, sometimes they will even diverge when the condition number is big.

Definition 3.2: Irreducible matrix
A matrix $\mathbf{A}$ is said to be irreducible if by symmetric permutation of rows and columns it can not take the form:

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}12 \\ \mathbf{0} & \mathbf{A}_{22} \end{bmatrix}$$

Where $\mathbf{A}_{11}$ and $\mathbf{A}_{22}$ are square block matrices.

Theorem 3.3:
If $\mathbf{A}$ weakly row diagonally dominant and irreducible, both Jacobi and G-S still converge. G-S is faster.

4 Iterative Techniques to Compute Eigenvalues and Eigenvectors

4.1 The Problem

As previously, we consider real matrices but it can be generalised to complex matrices. Let $\mathbf{A} \in \mathbb{R}^{n\times n}$. We wish to find its eigenvalues and eigenspaces. In order to find the eigenvalues, one would have to find the roots of the characteristic polynomial $\det(\mathbf{A} - \lambda\mathbf{I})$ which is expensive. Then to find one eigenvector, one would have to solve a system of equation using Gaussian elimination that would take $O(n^3)$ operations. All in all, it would take $O(n^4)$ operations to compute all eigenvectors. We thus need numerical methods to compute the eigenvalues and eigenvectors.

4.2 Power Methods

Example 4.1:
Let $\mathbf{A} = \begin{bmatrix} 1 & -3 & 2 & 4 \\ -3 & 1 & 4 & 2 \\ 2 & 4 & 1 & -3 \\ 4 & 2 & -3 & 1 \end{bmatrix}$
Its eigenvalues are 2, 4, 6 and -8 with eigenspaces:
$E_2 = \text{span}\left\{\begin{bmatrix}-1\\1\\-1\\1\end{bmatrix}\right\}$, $E_4 = \text{span}\left\{\begin{bmatrix}1\\1\\1\\1\end{bmatrix}\right\}$, $E_6 = \text{span}\left\{\begin{bmatrix}-1\\1\\1\\-1\end{bmatrix}\right\}$ and $E_{-8} = \text{span}\left\{\begin{bmatrix}-1\\-1\\1\\1\end{bmatrix}\right\}$ respectively.

Let's consider the sequence $(\mathbf{x}_n)$ defined by $\mathbf{x}_{n+1} = \frac{\mathbf{A}\mathbf{x}_n}{\|\mathbf{A}\mathbf{x}_n\|_\infty}$ and $\mathbf{x}_0 = \begin{bmatrix}1\\0\\0\\0\end{bmatrix}$.

Then we have:
$\mathbf{x}_1 = \begin{bmatrix}0.25\\-0.75\\0.5\\1\end{bmatrix}$, $\mathbf{x}_2 = \begin{bmatrix}1\\0.33\\-0.67\\-0.13\end{bmatrix}$, $\mathbf{x}_3 = \begin{bmatrix}-0.29\\-0.86\\0.47\\1\end{bmatrix}$, $\mathbf{x}_4 = \begin{bmatrix}1\\0.54\\-0.90\\-0.45\end{bmatrix}$, $\mathbf{x}_5 = \begin{bmatrix}-0.58\\-0.95\\0.63\\1\end{bmatrix}$ and $\mathbf{x}_{15} = \begin{bmatrix}-0.97\\-1\\0.97\\1\end{bmatrix}$.
So our guess is that $(\mathbf{x}_n)$ will eventually end up in E_{-8} and we have $\|\mathbf{A}\mathbf{x}_n\|_\infty \xrightarrow[n\to\infty]{} 8$

This example illustrates how the power-iteration can help find an eigenvalue and corresponding eigenvector of $\mathbf{A}$, this eigenvalue being the one with the largest modulus. Let us formalise this intuition.

Definition 4.1:
Let $\mathbf{A} \in \mathbb{R}^{n\times n}$. A dominant eigenvalue of $\mathbf{A}$ is an eigenvalue with the largest modulus. A dominant eigenvector is an eigenvector corresponding to a dominant eigenvalue.

4.2.1 Power Iteration

Theorem 4.1:
Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ a diagonalisable matrix with eigenvalues of distinct modulus. Let $\lambda \in \mathbb{R}$ be the dominant eigenvalue.
We consider the sequence $(\mathbf{x}_k)$ defined by $\mathbf{x}_{k+1} = \frac{\mathbf{A}\mathbf{x}_k}{\|\mathbf{A}\mathbf{x}_k\|}$ and $x_0 \in \mathbb{R}^n \setminus \{0\}$.
Then:

$$\boxed{\mathbf{x}_k \xrightarrow[k\to\infty]{} \mathbf{v}} \text{ and } \boxed{\|\mathbf{A}\mathbf{x}_k\| \xrightarrow[k\to\infty]{} |\lambda|}$$

Where $\mathbf{v} \in \mathbb{R}^n$ is a normalised dominant eigenvector. In this case, the notion of convergence that we use for $(\mathbf{x}_k)$ is not exactly rigorous, as you can see in Example 4.1, if the dominant eigenvalue is negative, it can end up oscillating between $\mathbf{v}$ and $-\mathbf{v}$. So here it has to be understood as "the sequence converges to the corresponding eigenspace"

Proof. $\mathbf{A} \in \mathbb{R}^{n\times n}$ being diagonalisable, there exist a basis $(\mathbf{e}_1, \ldots, \mathbf{e}_n)$ of $\mathbb{R}^n$ formed with normalised eigenvectors of $\mathbf{A}$. Let λ the dominant eigenvalue of $\mathbf{A}$ and p its geometric multiplicity. Without loss of generality, we can suppose that $\mathbf{e}_1, \ldots, \mathbf{e}_p$ are the eigenvectors corresponding to λ. For $i > p$, $\mathbf{e}_i$ is an eigenvector corresponding to the eigenvalue λ_i with $|\lambda_i| < |\lambda|$. Take $\mathbf{x}_0 \in \mathbb{R}^n$, then: $\mathbf{x}_0 = \sum_{i=0}^{n} \alpha_i \mathbf{e}_i$ and for $k \in \mathbb{N}$:

$$\mathbf{A}^k \mathbf{x}_0 = \sum_{i=1}^{p} \lambda^k \alpha_i \mathbf{e}_i + \sum_{i=p+1}^{n} \lambda_i^k \alpha_i \mathbf{e}_i$$

So

$$\mathbf{x}_k = \frac{\mathbf{A}^k \mathbf{x}_0}{\|\mathbf{A}^k \mathbf{x}_0\|} = \frac{\lambda^k}{\|\mathbf{A}^k \mathbf{x}_0\|} \left(\sum_{i=1}^{p} \alpha_i \mathbf{e}_i + \sum_{i=p+1}^{n} \left(\frac{\lambda_i}{\lambda}\right)^k \alpha_i \mathbf{e}_i \right)$$

Let us denote $\mathbf{u} = \sum_{i=1}^{p} \alpha_i \mathbf{e}_i$. Because $|\lambda_i| < |\lambda|$ for $i > p$, it comes that $\left(\frac{\lambda_i}{\lambda}\right)^k \xrightarrow[k\to\infty]{} 0$ so:

$$\mathbf{x}_k \xrightarrow[k\to\infty]{} \frac{\mathbf{u}}{\|\mathbf{u}\|}$$

Which is a dominant eigenvector of A. Now by continuity of A and the norm,

$$\|\mathbf{A}\mathbf{x}_k\| \xrightarrow[k\to\infty]{} \left\| \mathbf{A} \frac{\mathbf{u}}{\|\mathbf{u}\|} \right\| = \left\| \lambda \frac{\mathbf{u}}{\|\mathbf{u}\|} \right\| = |\lambda|$$

□

Although quite simple to apply, the power iteration has some limitations:

1. x_0 being chosen at random, it is possible for it to be such that $\alpha_1 = \ldots = \alpha_p = 0$. In this case, the iteration will yield the second dominant eigenvalue and eigenvector.
   - Need to make sure that there is at least one non zero component in the corresponding eigenspace.
   - Usually not an issue in practice as rounding errors during the normalisations will introduce a non zero component.
2. We assumed that all eigenvalues have distinct modulus which is not always the case. It is possible for a matrix to have multiple eigenvalue of maximum modulus. In this case, the power iteration will converge to a linear combination of the corresponding eigenvectors.
3. Convergence may be slow if the dominant eigenvalue is not "very dominant".

4.2.2 Inverse Power Iteration

With the power iteration, we have seen that it is possible to approximate the dominant eigenvalue of a matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$. As previously, suppose that $\mathbf{A}$ has eigenvalue with different modulus and in addition, suppose that it is non-singular. This means that none of its eigenvalue are zero and μ is an eigenvalue of $\mathbf{A}$ if and only if $\frac{1}{\mu}$ is an eigenvalue of $\mathbf{A^{-1}}$ with the same eigenvectors. So if we denote λ to be the eigenvalue of $\mathbf{A}$ with the smallest modulus, then $\frac{1}{\lambda}$ is a dominant eigenvalue of $\mathbf{A^{-1}}$.

Example 4.2:
Considering the matrix from Example 4.1, it is non-singular and its inverse is:

$$\mathbf{A^{-1}} = \frac{1}{96} \begin{bmatrix} 19 & -13 & 17 & 1 \\ -13 & 19 & 1 & 17 \\ 17 & 1 & 19 & -13 \\ 1 & 17 & -13 & 19 \end{bmatrix}$$

The eigenvalues of $\mathbf{A^{-1}}$ are $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{6}$ and $-\frac{1}{8}$ with respective eigenspaces:

$$E_{\frac{1}{2}} = \text{span}\left\{\begin{bmatrix}-1\\1\\-1\\1\end{bmatrix}\right\},\ E_{\frac{1}{4}} = \text{span}\left\{\begin{bmatrix}1\\1\\1\\1\end{bmatrix}\right\},\ E_{\frac{1}{6}} = \text{span}\left\{\begin{bmatrix}-1\\1\\1\\-1\end{bmatrix}\right\} \text{ and } E_{-\frac{1}{8}} = \text{span}\left\{\begin{bmatrix}-1\\-1\\1\\1\end{bmatrix}\right\}.$$

Starting from $\mathbf{x}_0 = \begin{bmatrix}1\\0\end{bmatrix}$, similar iterations to Example 4.1 would show that the process converges to the vector $\begin{bmatrix}1\\-1\end{bmatrix}$ and $\|\mathbf{A^{-1}x}_n\|_\infty \xrightarrow[n\to\infty]{} \frac{1}{2}$

From this example we can write the following Theorem, that is similar to Theorem 4.1.

Theorem 4.2:
Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ a diagonalisable non-singular matrix with eigenvalues of distinct modulus. Let $\lambda \in \mathbb{R}$ be the eigenvalue with the smallest modulus. We consider the sequence $(\mathbf{x}_k)$ defined by $\mathbf{x}_{k+1} = \frac{\mathbf{A^{-1}x}_k}{\|\mathbf{A^{-1}x}_k\|}$ and $x_0 \in \mathbb{R}^n \setminus \{0\}$.
Then:

$$\boxed{\mathbf{x}_k \xrightarrow[k\to\infty]{} \mathbf{v}} \text{ and } \boxed{\|\mathbf{A^{-1}x}_k\| \xrightarrow[k\to\infty]{} \left|\frac{1}{\lambda}\right|}$$

Where $\mathbf{v} \in \mathbb{R}^n$ is a normalised eigenvector corresponding to λ.

Proof. The proof is similar to that of Theorem 4.1, except that you first have to show the result about the eigenvalues of $\mathbf{A^{-1}}$ being the inverse of the ones of $\mathbf{A}$, the eigenvectors being conserved. □

4.2.3 Shifts

Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $s \in \mathbb{R}$. We call the matrix $\mathbf{A} - s\mathbf{I}$ a shifted matrix and we have the following property:

Property 4.1:
Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $s \in \mathbb{R}$. $\lambda \in \mathbb{R}$ is an eigenvalue of $\mathbf{A}$ if and only if $\lambda - s$ is an eigenvalue of $\mathbf{A} - s\mathbf{I}$ with the same eigenvectors.

Thanks to the shifted matrix, it is possible to focus on a particular eigenvalue of the matrix A.

Theorem 4.3:
Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ a diagonalisable matrix with eigenvalues of distinct modulus and $s \in \mathbb{R}$. Suppose that the shifted matrix $\mathbf{A} - s\mathbf{I}$ is non-singular (it is non-singular as long as s is not an eigenvalue of A), then by performing inverse power iterations, we can find the eigenvalue of $\mathbf{A}$ that it the closest to s.

Proof. Inverse power iteration on $\mathbf{A} - s\mathbf{I}$ converges to the eigenvalue with the smallest modulus λ with corresponding eigenvector $\mathbf{v}$. We thus have:

$$(\mathbf{A} - s\mathbf{I})\,\mathbf{v} = \lambda\mathbf{v} \iff \mathbf{Av} = (s+\lambda)\mathbf{v}$$

so $\mathbf{v}$ is an eigenvector of $\mathbf{A}$ corresponding to the eigenvalue $s + \lambda$ which is the eigenvalue of $\mathbf{A}$ that is the closest to s □

4.3 Additional Techniques

4.3.1 Rayleigh Quotient

Definition 4.2:
Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{x} \in \mathbb{R}^n \setminus \{0\}$, the Rayleigh quotient $R(\mathbf{A}, \mathbf{x})$ is given by:

$$\boxed{R(\mathbf{A}, \mathbf{x}) = \frac{\mathbf{x}^T\mathbf{Ax}}{\mathbf{x}^T\mathbf{x}}}$$

When using an iterative technique to find the eigenvalue/eigenvector of a matrix, you can use the Rayleigh quotient to monitor the convergence to the eigenvalue as it gives access to an approximation of the eigenvalue directly and not its modulus.

Example 4.3:
Applying the Rayleigh quotient to the sequence found in Example 4.1, we have: $R(\mathbf{A}, \mathbf{x_0}) = 1$, $R(\mathbf{A}, \mathbf{x_1}) = -1.87$, $R(\mathbf{A}, \mathbf{x_2}) = -4.23$, $R(\mathbf{A}, \mathbf{x_3}) = -5.75$, $R(\mathbf{A}, \mathbf{x_4}) = -6.67$, $R(\mathbf{A}, \mathbf{x_5}) = -7.24$, $R(\mathbf{A}, \mathbf{x_{15}}) = -7.99$.

4.3.2 Deflation

Suppose that thanks to the power iteration, you have found the dominant eigenvalue of $\mathbf{A} \in \mathbb{R}^{n \times n}$, how do you find the second dominant eigenvalue? One way would be to shift but you would have to pick a s that is close to the this second dominant eigenvalue. Problem: you don't know its value. Another way is to deflate the matrix $\mathbf{A}$ into a matrix $\mathbf{B} \in \mathbb{R}^{(n-1)\times(n-1)}$ that has the same eigenvalues than $\mathbf{A}$ except for the dominant which is absent. How does deflation works?

Let $\lambda_1, \ldots, \lambda_n$ the eigenvalues of $\mathbf{A}$, ordered according to their magnitude with λ_1 being the dominant one with corresponding eigenvector $\mathbf{x_1}$. Define $\mathbf{H} \in \mathbb{R}^{n \times n}$ a non-singular matrix such that:

$$\mathbf{H x_1} = \alpha \mathbf{e_1}$$

Where $\alpha \in \mathbb{R} \setminus \{0\}$ and $\mathbf{e_1} = \begin{bmatrix} 1, 0, \ldots, 0 \end{bmatrix}^T$ is the first vector in the standard basis. We then have

$$\mathbf{HAH^{-1}e_1} = \mathbf{HA}\frac{\mathbf{x_1}}{\alpha} = \mathbf{H}\frac{\lambda_1}{\alpha}\mathbf{x_1} = \lambda_1 \mathbf{e_1}$$

So the first column of $\mathbf{HAH^{-1}}$ is $\begin{bmatrix} \lambda_1, 0, \ldots, 0 \end{bmatrix}^T$ thus we can write:

$$\mathbf{HAH^{-1}} = \begin{bmatrix} \lambda_1 & \mathbf{b}^T \\ \mathbf{0} & \mathbf{B} \end{bmatrix}$$

Where $\mathbf{B} \in \mathbb{R}^{(n-1)\times(n-1)}$. Now let λ be an eigenvalue of $\mathbf{A}$ and $\mathbf{x}$ a corresponding eigenvalue. We have:

$$\mathbf{Ax} = \lambda \mathbf{x} \iff \mathbf{HAx} = \lambda \mathbf{Hx} \iff \mathbf{HAH^{-1}}(\mathbf{Hx}) = \lambda \mathbf{Hx}$$

So λ is eigenvalue of $\mathbf{A}$ if and only if it is also eigenvalue of $\mathbf{HAH^{-1}}$. This means that $\mathbf{B}$ has eigenvalues $\lambda_2, \ldots, \lambda_n$.

Property 4.2:
Let λ_2 be the second dominant eigenvector of $\mathbf{A}$ with $\lambda_2 \neq \lambda_1$. An eigenvector $\mathbf{x_2}$ of $\mathbf{A}$ corresponding to eigenvalue λ_2 is:

$$x_2 = \mathbf{H^{-1}} \begin{bmatrix} \beta \\ \mathbf{z_2} \end{bmatrix}$$

With $\beta = \frac{\mathbf{b^T z_2}}{\lambda_2 - \lambda_1}$ and $\mathbf{z_2}$ is a dominant eigenvector of $\mathbf{B}$

Proof. We have seen that $\mathbf{A}$ and $\mathbf{HAH^{-1}}$ have the same eigenvalues. Let $\mathbf{y}_2$ be an eigenvector of $\mathbf{HAH^{-1}}$ corresponding to λ_2, we denote $\mathbf{y_2} = \begin{bmatrix} \beta, \mathbf{z} \end{bmatrix}^T$ and we have:

$$\mathbf{HAH^{-1}y_2} = \begin{bmatrix} \lambda_1 & \mathbf{b}^T \\ \mathbf{0} & \mathbf{B} \end{bmatrix} \begin{bmatrix} \beta \\ \mathbf{z} \end{bmatrix} = \begin{bmatrix} \lambda_1 \beta + \mathbf{b^T z} \\ \mathbf{Bz} \end{bmatrix} = \lambda_2 \begin{bmatrix} \beta \\ \mathbf{z} \end{bmatrix}$$

It follows that $\lambda_1 \beta + \mathbf{b^T z} = \lambda_2 \beta$ and $\mathbf{Bz} = \lambda_2 \mathbf{z}$. The first equation yields: $\beta = \frac{\mathbf{b^T z_2}}{\lambda_2 - \lambda_1}$ while the second equation means that $\mathbf{z}$ is an eigenvector of $\mathbf{B}$ corresponding to λ_2 which is the dominant eigenvector of $\mathbf{B}$. So $\mathbf{z}$ and λ_2 can be found through power iteration. □

Remark. Here we have assumed that $\lambda_1 \neq \lambda_2$ or equivalently that the dominant eigenvalue λ_1 had geometric multiplicity 1. Suppose it is no longer the case and the dominant eigenvalue has multiplicity p, that is $\lambda_1 = \lambda_2 = \ldots = \lambda_p$. Then the deflation would work similarly but with blocks instead of vectors and $\mathbf{B}$ would be a $(n-p) \times (n-p)$ matrix.

One integral part of deflation is the matrix $\mathbf{H}$. How can you build such matrix? One possibility is using the Householder transformation that represents the reflection through a hyperplane with normal vector $\mathbf{u}$:

$$\mathbf{H} = \mathbf{I} - \frac{\mathbf{2uu^T}}{\mathbf{u^T u}}$$

$\mathbf{H}$ is symmetric and orthogonal, that is: $\mathbf{H} = \mathbf{H}^T = \mathbf{H}^{-1}$. Now remember that we want $\mathbf{H}$ to be such that $\mathbf{Hx_1} = \alpha \mathbf{e_1}$ with $\alpha \in \mathbb{R} \setminus \{0\}$. To do that, define $\mathbf{u} = \mathbf{x_1} + \|\mathbf{x_1}\|_\mathbf{2}\mathbf{e_1}$. We then have:

$$\mathbf{Hx_1} = \mathbf{x_1} - \frac{\mathbf{2u(u^T x_1)}}{\mathbf{u^T u}}$$

But

$$\begin{aligned}
\mathbf{u^T x_1} &= (\mathbf{x_1} + \|\mathbf{x_1}\|_\mathbf{2}\mathbf{e_1})^\mathbf{T}\mathbf{x_1} \\
&= \mathbf{x_1^T x_1} + \|\mathbf{x_1}\|_\mathbf{2}\mathbf{e_1^T x_1} \\
&= \|\mathbf{x_1}\|_\mathbf{2}(\|\mathbf{x_1}\|_\mathbf{2} + x_{1,1})
\end{aligned}$$

Where $x_{1,1}$ is the first element of $\mathbf{x_1}$. We also have:

$$\begin{aligned}
\mathbf{u^T u} &= (\mathbf{x_1} + \|\mathbf{x_1}\|_\mathbf{2}\mathbf{e_1})^\mathbf{T}(\mathbf{x_1} + \|\mathbf{x_1}\|_\mathbf{2}\mathbf{e_1}) \\
&= \mathbf{x_1^T x_1} + \mathbf{2}\|\mathbf{x_1}\|_\mathbf{2}\mathbf{e_1^T x_1} + \|\mathbf{x_1}\|_\mathbf{2}^\mathbf{2}\mathbf{e_1^T e_1} \\
&= \mathbf{2}\|\mathbf{x_1}\|_\mathbf{2}(\|\mathbf{x_1}\|_\mathbf{2} + x_{1,1})
\end{aligned}$$

So all in all:

$$\mathbf{Hx_1} = \mathbf{x_1} - \mathbf{u} = -\|\mathbf{x_1}\|_\mathbf{2}\mathbf{e_1}$$

So by setting $\alpha = -\|\mathbf{x_1}\|$, we do have $\mathbf{Hx_1} = \alpha \mathbf{e_1}$ as intended.

Example 4.4:
Let's apply deflation to the matrix from Example 4.1 using the Householder matrix. We take $\mathbf{x_1} = \begin{bmatrix} 1 & 1 & -1 & -1 \end{bmatrix}^T$, then $\mathbf{u} = \begin{bmatrix} 3 & 1 & -1 & -1 \end{bmatrix}^T$ and

$$\mathbf{H} = \frac{1}{6}\begin{bmatrix} -3 & -3 & 3 & 3 \\ -3 & 5 & 1 & 1 \\ 3 & 1 & 5 & -1 \\ 3 & 1 & -1 & 5 \end{bmatrix}$$

$\mathbf{H}$ being a Householder transformation, it is its own inverse and we have:

$$\mathbf{HAH^{-1}} = \mathbf{HAH} = \begin{bmatrix} -8 & 0 & 0 & 0 \\ 0 & 4 & 4/3 & -4/3 \\ 0 & 4/3 & 10/3 & 0 \\ 0 & -4/3 & 0 & 14/3 \end{bmatrix}$$

So

$$\mathbf{B} = \begin{bmatrix} 4 & 4/3 & -4/3 \\ 4/3 & 10/3 & 0 \\ -4/3 & 0 & 14/3 \end{bmatrix}$$

And $\mathbf{b}^T = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$. Performing power iterations on $\mathbf{B}$, the algorithm converges to $\mathbf{z_2} = \begin{bmatrix} 1 & 1/2 & -1 \end{bmatrix}^T$ with corresponding eigenvalue $\lambda_2 = 6$. In this case, we have $\beta = \frac{\mathbf{b^T z_2}}{\lambda_2 - \lambda_1} = 0$ and

$$\mathbf{x_2} = \mathbf{H}^{-1} \begin{bmatrix} 0 \\ 1 \\ 1/2 \\ -1 \end{bmatrix} = \mathbf{H} \begin{bmatrix} 0 \\ 1 \\ 1/2 \\ -1 \end{bmatrix} = \frac{3}{4} \begin{bmatrix} -1 \\ 1 \\ 1 \\ -1 \end{bmatrix}$$

Which is indeed an eigenvector of $\mathbf{A}$.

4.4 QR-Algorithm

4.4.1 The Algorithm

As its name suggests, the QR-algorithm relies on the QR decomposition and the way it works is quite simple but powerful. It is one of the most used algorithm to find the eigenvalues of a matrix and works for most matrices.
Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ the matrix for which we wish to find the eigenvalues and eigenvectors. We consider the sequence $(\mathbf{A}_k)$ defined as follow:

- $\mathbf{A_0} = \mathbf{A}$
- For $k \in \mathbb{N}$, apply the QR decomposition to $\mathbf{A_k}$: $\mathbf{A_k} = \mathbf{Q_{k+1}R_{k+1}}$ where $\mathbf{Q_{k+1}}$ is an orthogonal matrix and $\mathbf{R_{k+1}}$ is an upper triangular matrix
- Set $\mathbf{A_{k+1}} = \mathbf{R_{k+1}Q_{k+1}}$
- Stop after "sufficient" iterations

Lets see how this algorithm can help us find the eigenvalues of $\mathbf{A}$

Property 4.3:
For $k \in \mathbb{N}$, $\mathbf{A_k}$ is similar to $\mathbf{A}$.

Proof. Let $k > 0$, we have: $\mathbf{A_k} = \mathbf{R_k Q_k}$. But $\mathbf{A_{k-1}} = \mathbf{Q_k R_k}$, so $\mathbf{R_k} = \mathbf{Q_k^T A_{k-1}}$, hence:

$$\mathbf{A_k} = \mathbf{Q_k^T A_{k-1} Q_k}$$

And

$$\mathbf{A_k} = \mathbf{Q_k^T Q_{k-1}^T \dots Q_1^T A Q_1 \dots Q_{k-1} Q_k}$$

By denoting $\tilde{\mathbf{Q}}_\mathbf{k} = \mathbf{Q_1 \dots Q_{k-1} Q_k}$, $\tilde{\mathbf{Q}}_\mathbf{k}$ is orthogonal and we have:

$$\mathbf{A_k} = \tilde{\mathbf{Q}}_\mathbf{k}^\mathbf{T} \mathbf{A} \tilde{\mathbf{Q}}_\mathbf{k}$$

So $\mathbf{A_k}$ and $\mathbf{A}$ are similar. □

From the similarity of $\mathbf{A}$ with $\mathbf{A_k}$ for $k \in \mathbb{N}$ comes this following property:

Property 4.4:
For $k \in \mathbb{N}$, Property 4.3 gives that $\mathbf{A_k} = \tilde{\mathbf{Q}}_\mathbf{k}^\mathbf{T} \mathbf{A} \tilde{\mathbf{Q}}_\mathbf{k}$ so $\mathbf{A_k}$ and $\mathbf{A}$ have the same eigenvalues and $\mathbf{v}$ is an eigenvector of $\mathbf{A_k}$ if and only if $\tilde{\mathbf{Q}}_\mathbf{k}\mathbf{v}$ is an eigenvector of $\mathbf{A}$

Proof. Let λ be an eigenvalue of $\mathbf{A_k}$ with corresponding eigenvector $\mathbf{v}$. We have:

$$\mathbf{A_k v} = \lambda \mathbf{v} \iff \tilde{\mathbf{Q}}_\mathbf{k} \mathbf{A_k v} = \lambda \tilde{\mathbf{Q}}_\mathbf{k} \mathbf{v} \iff \tilde{\mathbf{Q}}_\mathbf{k} \mathbf{A_k} \tilde{\mathbf{Q}}_\mathbf{k}^\mathbf{T} (\tilde{\mathbf{Q}}_\mathbf{k} \mathbf{v}) = \lambda \tilde{\mathbf{Q}}_\mathbf{k} \mathbf{v}$$

As $\mathbf{A} = \tilde{\mathbf{Q}}_\mathbf{k} \mathbf{A_k} \tilde{\mathbf{Q}}_\mathbf{k}^\mathbf{T}$, this is equivalent to:

$$\mathbf{A} \tilde{\mathbf{Q}}_\mathbf{k} \mathbf{v} = \lambda \tilde{\mathbf{Q}}_\mathbf{k} \mathbf{v}$$

So λ is eigenvalue of $\mathbf{A_k}$ with corresponding eigenvector $\mathbf{v}$ if and only if it is also eigenvalue of $\mathbf{A}$ with corresponding eigenvector $\tilde{\mathbf{Q}}_\mathbf{k} \mathbf{v}$ □

We will only state the following property which concerns the convergence of the QR-algorithm as the proof is quite complicated.

Property 4.5:
The sequence (A_k) converges to an upper triangular matrix

The following property explains why the fact that the QR-algorithm converges to an upper triangular matrix is interesting.

Property 4.6:
The eigenvalues of an upper triangular matrix are simply its diagonal elements.

Proof. Let $\mathbf{U} \in \mathbb{R}^{n \times n}$ be an upper triangular matrix. We have:

$$U = \begin{bmatrix} u_{11} & \dots & \dots & u_{1n} \\ 0 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & u_{nn} \end{bmatrix}$$

Then the characteristic polynomial is given by:

$$\det(\mathbf{U} - \lambda \mathbf{I}) = \prod_{i=1}^{n} (u_{ii} - \lambda)$$

So the diagonal elements of $\mathbf{U}$ are its eigenvalues. □

Thanks to all these properties you can see why the QR-algorithm is an interesting way of computing the eigenvalues and eigenvectors of a matrix:

- The QR decomposition of a matrix can be efficiently computed
- The sequence of matrices converges to an upper triangle matrix for which the eigenvalues can be easily extracted
- It is even possible to stop as soon as the matrix is "almost" upper triangular (there are still some non zero elements in the lower part)
- Unlike the power methods, the QR algorithm primarily gives access to the eigenvalues of the matrix. Once they are obtained, the corresponding eigenvectors still have to be computed
- The eigenvectors of an upper triangular matrix are easier to compute thanks to back substitution. Still, it might be quicker to compute them using shifted power methods as the eigenvalues are now known

4.4.2 Application to Symmetric Matrices

When applied to symmetric matrices, the QR algorithm behaves nicely:

Property 4.7:
If $\mathbf{A}$ is symmetric, so are all the $\mathbf{A_k}$.

Proof. for $k \in \mathbb{N}$, we have $\mathbf{A_k} = \tilde{\mathbf{Q}}_\mathbf{k}^\mathbf{T}\mathbf{A}\tilde{\mathbf{Q}}_\mathbf{k}$, taking the transpose, we get:

$$\mathbf{A}_\mathbf{k}^\mathbf{T} = (\tilde{\mathbf{Q}}_\mathbf{k}^\mathbf{T}\mathbf{A}\tilde{\mathbf{Q}}_\mathbf{k})^\mathbf{T} = \tilde{\mathbf{Q}}_\mathbf{k}^\mathbf{T}\mathbf{A}^\mathbf{T}(\tilde{\mathbf{Q}}_\mathbf{k}^\mathbf{T})^\mathbf{T} = \tilde{\mathbf{Q}}_\mathbf{k}^\mathbf{T}\mathbf{A}\tilde{\mathbf{Q}}_\mathbf{k} = \mathbf{A_k}$$

□

This leads to this result about the convergence of the QR-algorithm:

Property 4.8:
If $\mathbf{A}$ is symmetric, the algorithm converges to a diagonal matrix, hence the eigenvectors of $\mathbf{A}$ are the columns of $\tilde{\mathbf{Q}}_\mathbf{k}$.

Proof. From Property 4.5 the sequence $(\mathbf{A_k})$ converges to an upper triangular matrix but from Property 4.7 for $k \in \mathbb{N}$, $\mathbf{A_k}$ is symmetric. A symmetric upper triangular matrix is a diagonal matrix. Let k_0 be such that $\mathbf{A_{k_0}}$ is diagonal. In this case, we have $\mathbf{A} = \tilde{\mathbf{Q}}_{\mathbf{k_0}}^\mathbf{T}\mathbf{A}_{\mathbf{k_0}}\tilde{\mathbf{Q}}_{\mathbf{k_0}}$ which is an eigendecomposition of $\mathbf{A}$. □

So when the QR-algorithm is applied to symmetric matrices, both the eigenvalues and eigenvectors are directly found.

4.4.3 Convergence in a particular case

We will show the convergence of the QR algorithm for a particular class of matrices. But first we need some preliminary results.

Theorem 4.4: LU decomposition
A non singular matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ can be factorised as

$$\mathbf{A} = \mathbf{LU}$$

Where $\mathbf{L}$ is a lower triangular matrix and $\mathbf{U}$ is an upper triangular matrix

.

Proof. Starting from $\mathbf{A} \in \mathbb{R}^{n\times n}$ we transform it into an upper triangular matrix using elementary row operations. We have

$$\mathbf{A} = \begin{bmatrix} a_{11} & \dots & \dots \\ a_{12} & \ddots & \\ \vdots & & \ddots \end{bmatrix}$$

We multiply it to the left by:

$$\mathbf{E}_{12} = \begin{bmatrix} 1 & & & \\ -\frac{a_{12}}{a_{11}} & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix}$$

This yields:

$$\mathbf{E}_{12}\mathbf{A} = \begin{bmatrix} a_{11} & \dots & \dots \\ 0 & \ddots & \\ \vdots & & \ddots \end{bmatrix}$$

Continuing the same process, we can eliminate the elements of $\mathbf{A}$ that are below the diagonal resulting in an upper triangular matrix $\mathbf{U}$. Let's call $\mathbf{L}^{-1}$ the product of all the matrices used to transform $\mathbf{A}$ into $\mathbf{U}$. We then have

$$\mathbf{L}^{-1}\mathbf{A} = \mathbf{U}$$

The all have the same form as $\mathbf{E}_{12}$ which is lower triangular. The product of lower triangular matrices and the inverse of a lower triangular matrix being also lower triangular, $\mathbf{L}$ is lower triangular and we end up with:

$$\mathbf{A} = \mathbf{L}\mathbf{U}$$

□

The LU decomposition is generally not unique unless we impose some conditions:

Theorem 4.5: Uniqueness of the LU decomposition
If $\mathbf{A}$ is non singular and $\mathbf{A} = \mathbf{LU}$ with the diagonal elements of $\mathbf{L}$ being all one, the decomposition is unique

Proof. Suppose $\mathbf{A} = \mathbf{L_1}\mathbf{U_1} = \mathbf{L_2}\mathbf{U_2}$ where the diagonal elements of $\mathbf{L_1}, \mathbf{L_2}$ are 1's. $\mathbf{A}$ being non singular so are $\mathbf{L_1}, \mathbf{L_2}, \mathbf{U_1}, \mathbf{U_2}$ and we have:

$$\mathbf{L_2^{-1}}\mathbf{L_1} = \mathbf{U_2}\mathbf{U_1^{-1}}$$

The LHS is lower triangular and the RHS is upper triangular so $\mathbf{L_2^{-1}}\mathbf{L_1}$ is diagonal. The diagonal elements of $\mathbf{L_2^{-1}}$ are the inverse of the diagonal element of $\mathbf{L_2}$ and the diagonal elements of the product of lower triangular matrices are the product of the diagonal elements of the matrices so the diagonal elements of $\mathbf{L_2^{-1}}\mathbf{L_1}$ are $\frac{l_{ii}^{(1)}}{l_{ii}^{(2)}} = 1$ so

$$\mathbf{L_2^{-1}}\mathbf{L_1} = \mathbf{I}$$

Hence $\mathbf{L_2} = \mathbf{L_1}$ and $\mathbf{U_2} = \mathbf{U_1}$ □

We will also need this final result which is proven in Tutorial 6:

Theorem 4.6:
$\mathbf{A}^n = \mathbf{Q_1} \dots \mathbf{Q_n}\mathbf{R_1} \dots \mathbf{R_n}$

We now have all the tools we need to show this convergence:

Theorem 4.7:
Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be a symmetric positive definite matrix with distinct eigenvalues $\lambda_1 > \lambda_2 > \dots > \lambda_n > 0$ with eigendecomposition $\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q^T}$. Suppose $\mathbf{Q^T} = \mathbf{LU}$ with unit lower triangular $\mathbf{L}$ and the diagonal elements of $\mathbf{U}$ are positive. Then $\mathbf{A_k} \to \mathbf{\Lambda}$

Proof. As $\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q^T}$, we have $\mathbf{A}^n = \mathbf{Q}\mathbf{\Lambda}^n\mathbf{Q^T}$. But from Theorem 4.6 we also have $\mathbf{A}^n = \mathbf{Q_1} \dots \mathbf{Q_n}\mathbf{R_n} \dots \mathbf{R_n}$. Where we suppose that the diagonal elements of $\mathbf{R_k}$ are positive. So

$$\mathbf{Q}\mathbf{\Lambda}^n\mathbf{Q^T} = \mathbf{Q_1} \dots \mathbf{Q_n}\mathbf{R_1} \dots \mathbf{R_n}$$

Now $\mathbf{Q^T = LU}$ so

$$\begin{aligned}
&\mathbf{Q\Lambda^n LU = Q_1 \dots Q_n R_1 \dots R_n} \\
\Rightarrow &\mathbf{Q\Lambda^n L = Q_1 \dots Q_n R_1 \dots R_n U^{-1}} \\
\Rightarrow &\mathbf{Q\Lambda^n L\Lambda^{-n} = Q_1 \dots Q_n R_1 \dots R_n U^{-1}\Lambda^{-n}}
\end{aligned}$$

Looking at the elements of $\mathbf{\Lambda^n L\Lambda^{-n}}$ we have:

$$\begin{cases} l_{ij}\left(\frac{\lambda_i}{\lambda_j}\right)^n \text{ if } i > j \\ 1 \text{ if } i = j \\ 0 \text{ if } i < j \end{cases}$$

So $\mathbf{\Lambda^n L\Lambda^{-n}} \to \mathbf{I}$ and $\mathbf{Q\Lambda^n L\Lambda^{-n}} \to \mathbf{Q}$. This means that $\mathbf{Q_1 \dots Q_n R_1 \dots R_n U^{-1}\Lambda^{-n}} \to \mathbf{Q}$. We have $(\mathbf{Q_1 \dots Q_n})(\mathbf{R_1 \dots R_n U^{-1}\Lambda^{-n}})$ which is a QR decomposition and the diagonal elements of $\mathbf{R_1 \dots R_n U^{-1}\Lambda^{-n}}$ are positive so by uniqueness of the QR decomposition in this case:

$$\mathbf{Q_1 \dots Q_n} \to \mathbf{Q}$$

So

$$\mathbf{A}_k = \tilde{\mathbf{Q}}_k^T \mathbf{A} \tilde{\mathbf{Q}}_k \to \mathbf{Q^T AQ = \Lambda}$$

□

That's it for this part. Hope you enjoyed it :)