Clase_01-Notes

February 8, 2022

1 Session 1: Principles of numerical mathematics

In this first session we are going to explain the basis of the numerical mathematics.

They are like the pilars of a cathedral, maybe they are not the most attractive part of it, but everything will collapse without them.

1.1 Well-posedness and condition number of a problem

We are going to start defining a classification of the problems that we might find.

Consider the following expression:

$$F(x,d) = 0 \tag{1}$$

in which we call x the unknown, d data and F is the relation between x and d.

If F and d are known, finding x will be called the "direct problem".

If F and x are known, finding d will be called the "inverse problem".

If x and d are known, finding F will be called the "identification problem".

In this course we will, mostly, study the direct problem.

Example: Calculate the money you have to pay for n kg of some fruit that costs 4 euros per kg.

Why could the inverse problem be ill conditioned?

Example: Let us suppose we have this expression: $A\vec{v} - \vec{b} = \vec{0}$

We can think about it as a linear system

$$\begin{bmatrix} 1 & 3 \\ 5 & 7 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$$

in which the unknown is the $\vec{v}$ and then the direct problem will be solving the linear system, or as a matrix product

$$\begin{bmatrix} 1 & 3 \\ 5 & 7 \end{bmatrix} \begin{bmatrix} 3 \\ 4 \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix}$$

in which calculating the product would be the direct problem.

Sometimes you define what is direct and what is inverse, but often, the history of the problem will tell which is which.

Now, we give a definition that tells us whether or not a problem is well-posed (Jacques Hadamard), that is, if it is a reasonable problem from the mathematical perspective.

Definition: We say that a problem is well-posed if it admits a unique solution x which depends with continuity on the data.

Definition: We say that a problem is ill-posed if it is not well-posed.

Definition: We say that x depends with continuity on the data if a little change δd in the data produces a small change in the solution δx. Mathematically:

If $F(d + \delta d, x + \delta x) = 0$ then:

$$\forall \eta > 0, \exists K(\eta, d) : \|\delta d\| < \eta \Rightarrow \|\delta x\| \leq K(\eta, d)\|\delta d\| \tag{2}$$

where K is a constant that depends on η and d.

Example: The solution of the equation $3x + d = 0$, where d is the data, is $x = d/3$ which is a continous function on d and is unique, so the problem is well-posed.

Example: Find the number of roots of the polynomial $p(x) = x^4 - (2a-1)x^2 + a(a-1)$ (a is the data of the problem). Is easy to check that we have four real roots if $a \geq 1$, two is $a \in [0, 1)$ and no real roots if $a < 0$. This is an ill posed problem because the solution does not depend continuously from the data.

Number of roots as a function of a

Notice that the problem is not calculating the roots, but the number of roots as a function of the parameter a.

Example: Another kind of example of an ill-posed problem is given by the equation y' $= \frac{3}{2}y^{1/3}$ with $y(0) = 0$. Since the solution is $y(t) = \pm t^{3/2}$, the solution is not unique (it could be plus $t^{3/2}$ or it could be minus $t^{3/2}$). As uniqueness is required for well-posedness, the problem is ill-posed.

Two different solutions of the ODE y' $= \frac{3}{2}y^{1/3}$ with $y(0) = 0$

Most problems are not so clearly ill-posed as the previous examples, that is, there is a range of problems that even if they are well posed they will be incresingly hard to solve. To quantify how well or ill posed is a problem we define the condition numbers:

Definition: Relative condition number

$$K(d) = \underset{\delta d \in D}{sup} \frac{\|\delta x\|/\|x\|}{\|\delta d\|/\|d\|} \tag{3}$$

Definition: Absolute condition number

$$K_{abs}(d) = \underset{\delta d \in D}{sup} \frac{\|\delta x\|}{\|\delta d\|} \tag{4}$$

D is a neighborhood of the origin that denotes the admissible perturbations of the data.

Definition: We say a problem is “ill-conditioned” if K is “big” where the definition of big depends on the problem.

It is important to understand that the conditioning of a problem does not depend on the algorithm used to solve it. You can develop stable and unstable algorithms for well-posed problems. The concept of stability for algorithms will be defined later on.

Having a “big” or even infinite condition number does not imply that the problem is ill-posed. Some ill-posed problems can be reformulated as an equivalent problem (that is, one that has the same solution) which are well-posed.

If a problem admits a unique solution, then there exists a mapping G, called the resolvent, between the data and the solutions sets such that:

$$x = G(d), \text{ that is, } F(G(d), d) = 0 \tag{5}$$

According to this, and assuming G is differentiable in d ($G'(d)$ exist), the Taylor expansion of G is

$$G(d + \delta d) - G(d) = G'(d)\delta d + o(\|\delta d\|) \text{ for } \delta d \to 0$$

This let us redefine the condition numbers in terms of the resolvent G:

$$K(d) \approx \|G'(d)\| \frac{\|d\|}{\|G(d)\|} \qquad \text{and} \qquad K_{abs} \approx \|G'(d)\|$$

Exercise: Prove that the expressions of the contition numbers using the resolvent are equivalent to the original ones.

Example of ill-conditioning: Algebraic second degree equation.

We want to calculate the solutions of $x^2 - 2px + 1$ with $p \geq 1$. Obviously $x_\pm = p \pm \sqrt{p^2 - 1}$.

We can formulate this problem as $F(x, p) = x^2 - 2px + 1$ where p is the data and $x_\pm = (x_+, x_-)$ the solution. The resolvent $G(p) = (p + \sqrt{p^2 - 1}, p - \sqrt{p^2 - 1})$ and its derivative $G'(p) = (1 + p/\sqrt{p^2 - 1}, 1 - p/\sqrt{p^2 - 1})$.

Then:

$$K(d) \approx \|G'(d)\| \frac{\|d\|}{\|G(d)\|} = \frac{((4p^2 - 2)/(p^2 - 1))^{1/2}}{(4p^2 - 2)^{1/2}} \|p\| = \frac{|p|}{p^2 - 1}$$

$$K_{abs}(d) \approx \|G'(d)\| = \sqrt{2} \frac{p}{\sqrt{p^2 - 1}}$$

If $p >> 1$ then the problem is well-conditioned (two distinct roots). If $p = 1$ (one double root), then G is not differentiable but in the limit $p \to 1^+$ the problem is ill conditioned as $\lim_{p \to 1^+} \|G'(p)\| = \infty$.

However, the problem is not ill-posed. We can reformulate it as $F(x, t) = x^2 - ((1 + t^2)/t)x + 1$ with $t = p + \sqrt{p^2 - 1}$. In this case $x_+ = t$ and $x_- = 1/t$ are the same for $t = 1$, and $K(t) \approx 1 \ \forall t \in \mathbb{R}$

1.2 Stability of numerical methods

Let's assume the the problem $F(x,d)=0$ is well-posed. Then, a numerical method to approximate its solution will consist, in general, of a sequence of approximate problems

$$F_n(x_n, d_n) = 0 \quad n \geq 1$$

We would expect that $x_n \underset{n\to\infty}{\to} x$. For that it is necessary that $d_n \to d$ and that F_n approximates F when $n \to \infty$.

Definition: We say that $F_n(x_n, d_n) = 0$ is consistent if

$$F_n(x,d) = F_n(x,d) - F(x,d) \underset{n\to\infty}{\to} 0$$

where x is the solution of $F(x,d)=0$ for the datum d.

Definition: We say that a method is strongly consistent if $F_n(x,d) = 0 \quad \forall n$.

In some cases when iterative methods are used, we can write them as

$$F(x_n, x_{n-1}, \cdots, x_{n-q}, d_n) = 0$$

where $x_n, x_{n-1}, \cdots, x_{n-1}$ are given. In this case the property of strong consistency becomes $F_x(x, x, \cdots, x, d) = 0 \ \forall n \geq q$.

Examples:

- Newton's method: $x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})}$ is strongly consistent.
- Composite midpoint rule: If $x = \int_a^b f(t)dt$, $x_n = H\sum_{k=1}^n f(\frac{t_k + t_{k+1}}{2}) \quad n \geq 1$ with $H = (b-a)/n$ and $t_k = a + (k-1)H$. This method to calculate the integral is consistent, but only strongly consistent if f is a piecewise linear polynomial.

In general, numerical methods obtained from the mathematical problem by truncation of limit operations (like integrals, derivatives, series,...) are not strongly consistent.

Exercise: Prove the previous assertions.

Definition: We say that a numerical method $F_n(x_n, d_n) = 0$ is well-posed (or **stable**) if for any fixed n there exists a unique solution x_n corresponding to the datum d_n, that the computation of x_n as a function of d_n is unique, and that x_n depends continuously on the data, i.e:

$$\forall \eta > 0, \exists K_n(\eta, d_n) : \|\delta d_n\| < \eta \Rightarrow \|\delta x - n\| \leq K_n(\eta, d_n)\|\delta d_n\| \tag{6}$$

We can also define:

$$K_n(d_n) = \underset{\delta d_n \in D_n}{sup} \frac{\|\delta x_n\|/\|x_n\|}{\|\delta d_n\|/\|d_n\|} \qquad \text{and} \qquad K_{abs,n}(d_n) = \underset{\delta d_n \in D_n}{sup} \frac{\|\delta x_n\|}{\|\delta d_n\|} \tag{7}$$

and from these:

$$K^{num}(d_n) = \lim_{n\to\infty} \underset{n\geq k}{sup}\, K_n(d_n) \quad \text{and} \quad K^{num}_{abs}(d_n) = \lim_{n\to\infty} \underset{n\geq k}{sup}\, K_{abs,n}(d_n) \tag{8}$$

K_{num} is the relative asymptotic condition number and K^{num}_{abs} is the absolute asymptotic condition number of the numerical method corresponding to the datum d_n.

The numerical method is said to be well-conditioned if the condition number K^{num} is "small" for any admissible datum d_n and ill-conditioned otherwise.

We can also define the resolvent G_n for the numerical method:

$ x_n=G(d_n)$, that is $F(G_n(d_n),d_n)=0 $

Assuming it is differentiable:

$$K_n(d_n) \approx \|G'_n(d_n)\| \frac{\|d_n\|}{\|G_n(d_n)\|} \qquad \text{and} \qquad K_{abs} \approx \|G'_n(d_n)\|$$

Examples:

- Sum and subtraction. The sum defined as

  $$f : \underset{(a,b)}{\mathbb{R}^2} \to \underset{a+b}{\mathbb{R}}$$

  has derivative $f'(a,b) = (1,1)^T$, and thus, its condition number $K((a,b)) \approx \frac{|a|+|b|}{|a+b|} \approx 1$ The subtraction defined as

  $$f : \underset{(a,b)}{\mathbb{R}^2} \to \underset{a-b}{\mathbb{R}}$$

  has derivative $f'(a,b) = (1,-1)^T$, and thus, its condition number $K((a,b)) \approx \frac{|a|+|b|}{|a-b|}$ which can be very big if $a \approx b$.

- Finding the roots of $x^2 - 2px + 1 = 0$ is well-conditioned, but we can develop an unstable algorithm: $x_- = p - \sqrt{p^2-1}$ because this formula is subject to errors due to numerical cancellation of digits in the subtraction. The Newton's method could be a stable algorithm to solve this problem:

  $$x_n = x_{n-1} - \frac{x_{n-1}^2 - 2px_{n-1} + 1}{2x_{n-1} - 2p}$$

  The method's condition number is $K_n(p) = \frac{|p|}{|x_n - p|}$. To compute $K^{num}_n(p)$ we notice that if the algorithm converges, then $x_n \to x_+$ or x_-, therefore, $|x_n - p| \to \sqrt{p^2-1}$ and $k_n(p) \to K^{num}_n(p) \approx \frac{|p|}{\sqrt{p^2-1}}$ which is similar to the condition number of the exact problem. Then, if $p \approx 1$ the problem is ill-conditioned.

Definition: We say that the numerical method $F_n(x_n, d_n) = 0$ is convergent if and only if

$$\forall \epsilon > 0\ \exists n_0(\epsilon), \exists \delta(n_0, \epsilon) : \forall n > n_0, \forall \|\delta d_n\| < \delta(n_0, \epsilon) \Rightarrow \|x(d) - x_n(d + \delta d_n)\| < \epsilon$$

where d is an admissible datum, $x(d)$ the corresponding solution, and $x(d + \delta d_n)$ is the solution of the numerical problem $(F_n(x_n, d_n))$ with datum $d + \delta d_n$.

Definition: Absolute and relative errors:

$$E(x_n) = |x - x_n| \qquad E_{rel}(x_n) = \frac{|x - x_n|}{|x|} \quad (x \neq 0)$$

Definition: Error by component:

$$E^c_{rel}(x_n) = \max_{i,j} \frac{|(x - x_n)_{i,j}|}{|x_{i,j}|} \quad (x_{i,j} \neq 0)$$

1.2.1 Relations between stability and convergence

The concepts of stability and convergence are strongly connected. If a (numerical) problem is well-posed, stability is a necessary condition for convergence. Moreover, if the numerical problem is consistent, stability is a sufficient condition for convergence. This is known as "equivalence" or "Lax-Richtmyer" theorem: *For a consistent numerical method, stability is equivalent to convergence.*

1.2.2 Sources of errors in computational models

Whenever the numerical problem (NP) is an approximation of a mathematical problem (MP) and this latter is in turn a model of a physical problem (PP), we say that NP ($F_n(x_n, d_n) = 0$) is a computational model for PP.

The global error $e = |x_{ph} - x_n|$ can be interpreted as the sum of the MP error $e_m = x - x_{ph}$ and the computational problem error $e_c = \hat{x} - x$ ($e = e_m + e_c$).

Sources of error

e_a: Error induced by the numerical algorithm and the rounding errors.

In general, we can enumerate the following sources of error:
1. Errors due to the model, that can be **reduced** by using a proper model. 2. Errors due to data, that can be **reduced** improving the measurement's accuracy. 3. Truncation errors, arising from the approximation (truncation) of limit operations (integrals, derivatives,…). 4. Rounding errors.

Type 3 and 4 errors give rise to the computational error. A numerical method will be convergent if this error can be made arbitrarily small increasing the computational effort. Although convergence is the primary goal of a numerical method, there are also the accuracy, the reliability and the efficiency.

Accuracy means that the errors are small with respect to a fixed tolerance. It is usually quantified by the infinitesimal order of the error e_n with respect to the discretization characteristic parameter (for example the largest grid spacing).

Note: Machine precision does not limit, theoretically, the accuracy.

Reliability means that it is very likely that the global error is below a certain tolerance.

Efficiency means that the computational (effort) complexity needed to control the error (number of operations and memory) is as small as possible.

Definition: Algorithm is a directive that indicates, through elementary operations, all the passages needed to solve a problem. It should finish after a finite number of steps, and as a consequence the executor (man or machine) must find within the algorithm itself all the instructions to completely solve the problem.

Complexity of an algorithm is a measure of its executing time.

Complexity of a problem is the complexity of the algorithm with smallest complexity capable of solving the problem.

Interesting links:

- Well and ill-posed problems
- Condition number (Wikipedia)