Métodos Numéricos

Mercedes Román Ruiz

28 de octubre de 2024

Índice general

Capítulo 1

Introducción

1. Vector Spaces

Definition 1.1. A vector space eover the numeric field K ($K = \mathbb{R}$ or $K = \mathbb{C}$) is a nonempty set V, whose elements are called vector and in which two operations are defined, called addition and scalar multiplication, that enjoy the following properties:

1. addition is commutative and associative;
2. there exists an element $0 \in V$ (the zero vector or null vector) such that $v + 0 = v$ for each $v \in V$
3. $0 \cdot v = 0$, $1 \cdot v = v$, for each $v \in V$, where 0 and 1 are respectively the zero and the unity of K;
4. for each element $v \in V$ there exists its opposite, $-v$, in V such that $v + (-v) = 0$;
5. the following distributive properties hold

$$\forall \alpha \in K, \forall v, w \in V, \alpha(v + w) = \alpha v + \alpha w$$

$$\forall \alpha, \beta \in K, \forall v \in V, (\alpha + \beta)v = \alpha v + \beta v$$

6. the following associative property holds

$$\forall \alpha, \beta \in K, \forall v \in V, (\alpha\beta)v = \alpha(\beta v)$$

Definition 1.2. We say that a nonempty part W of V is a vector subspace of V if W is a vector space over K.

Definition 1.3. A system of vector $v_1, ..., v_n$ of a vector space V is called linearly independent if the relation

$$\alpha_1 v_1 + \alpha_2 v_2 + ... + \alpha_m v_m = 0$$

with $\alpha_1, \alpha_2, ..., \alpha_m \in K$ implies that $\alpha_1 = \alpha_2 = ... = \alpha_m = 0$. Otherwise, the system will be called linearly dependent.

We call a basis of V any system of linearly independent generator of V. If $u_1, ..., u_n$ is a basis of V, the expression $v = v_1 u_1 + ... + v_n u_n$ is called the decomposition of v with respect to the basis and the scalars $v_1, ..., v_n \in K$ are the components of v with respect to the given basis. Moreover, the following pro holds.

Property 1.1. Let V be a vector space which admits a basis of n vectors. Then every system of linearly independent vector of V hast at most n elements and any other basis of V has n elements. The number n is called the dimension of V and we write $dim(V) = n$. If, instead, for any n there always exist n linearly independent vectors of V, the vector space is called infinite dimensional.

2. Matrices

Let m and n be two positive integers. We call a matrix having m rows and n columns, or a matrix $m \times n$, or a matrix (m, n), with elements in K, a set of mn scalars $a_{ij} \in K$, with $i = 1, ..., m$ and $j = 1, ..., n$, represented in the following rectangular array

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \tag{1.1}$$

When $K = \mathbb{R}$ or $K = \mathbb{C}$ we shall respectively write $A \in \mathbb{R}^{m \times n}$ or $A \in \mathbb{C}^{m \times n}$, to explicitly outline the numerical fields which the elements of A belong to. Capital letters will be used to denote the matrices, while the lower case letters corresponding to those upper case letters will denote the matrix entries.

We shall abbreviate (1.1) as $A = (a_ij)$ with $i = 1, ..., m$ and $j = 1, ..., n$. The index i is called row index, while j is the column index. The set $(a_{i1}, a_{i2}, ..., a_{in})$ is called the i-th row of A; likewise, $(a_{1j}, a_{2j}, ..., a_{mj})$ is the j-th column of A.

If $n = m$ the matrix is called squared or having order n and the set of the entries $(a_{11}, a_{22}, ..., a_{nn})$ is called its main diagonal.

A matrix having one row or one column is called a row vector or column vector respectively. Unless otherwise specified, we shall always assume that a vector is a column vector. In the case $n = m = 1$, the matrix will simply denote a scalar of K.

Definition 2.1. Let A be a matrix $m \times n$. Let $1 \leq i_1 < i_2 < ... < i_k \leq m$ and $1 \leq j_1 < j_2 < ... < j_l \leq n$ two sets of contiguous indexes. The matrix $S(k \times l)$ of entries $s_{pq} = a_{i_p j_q}$ with $p = 1, ..., k$, $q = 1, ..., l$ is called a submatrix of A. If $k = l$ and $i_r = j_r$ for $r = 1, ..., k$, S is called a principal submatrix of A.

Definition 2.2. A matrix $A(m \times n)$ is called block partitioned or said to be partitioned into submatrices if

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1l} \\ A_{21} & A_{22} & \cdots & A_{2l} \\ \vdots & \vdots & \ddots & \vdots \\ A_{k1} & A_{k2} & \cdots & A_{kl} \end{bmatrix}$$

where A_{ij} are submatrices of A.

3. Operations with Matrices

3.1. Matrices and Linear Mappings

Definition 3.1. A linear map for $\mathbb{C}^n$ into $\mathbb{C}^m$ is a function $f : \mathbb{C}^n \to \mathbb{C}^m$ such that $f(\alpha x+\beta y) = \alpha f(x)+ \beta f(y)$, $\forall \alpha, \beta \in K$ and $\forall x, y \in \mathbb{C}^n$.

The following result links matrices and linear maps.

Property 3.1. Let $f : \mathbb{C}^n \to \mathbb{C}^m$ be a linear map. Then, ther exists a unique matrix $A_f \in \mathbb{C}^{m \times n}$ such that

$$f(x) = A_f x \quad \forall x \in \mathbb{C}^n \tag{1.2}$$

Conversely, if $A_f \in \mathbb{C}^{m \times n}$ then the function defined in (1.2) is a linear map from $\mathbb{C}^n$ into $\mathbb{C}^m$.

4. Well-posedness and Condition Number of a Problem

Consider the following problem: find x such that

$$F(x, d) = 0 \tag{1.3}$$

where d is the set of data which the solution depends on and F is the functional relation between x and d. According to the kind of problem that is represented in (1.3), the variables x and d may be real numbers, vectors or functions. Typically, (1.3) is called a direct problem if F and d are given and x is unknown, inverse problem if F and xn are known and d is the unknown, identigication problem when x and d are given while the functional relation F is the unknown.

Problem (1.3) is well posed if it admits a unique solution x which depends with continuaity on the data. We shall use the terms well posed and stable in an interchanging manner and we sall deal hecenforth only with well-posed problems.

A problem which does not enjoy the property above is called ill posed or unstable and before undertaking its numerical solution it has to be regularized, that is, it must be suitably transformed into a well-posed problem. Indeed, it is not appropriate to pretend the numerical method can cure the pathologies of an instrinsically ill-posed problem.

Let D be the set of admissible data, i.e. the set of the values of d in correspondance of which problem (1.3) admits a unique solution. Continuous dependence on the data means that small perturbations on the data d of D yield "small" changes in the solution x. Precisely, led $d \in D$ and denote by δd a perturbation admissible in the sense that $d + \delta d \in D$ and by δx the corresponding change in the solution, in such a way that

$$F(x + \delta x, d + \delta d) = 0 \tag{1.4}$$

Then, we require that

$$\exists \eta_0 = \eta_0(d) > 0, \quad \exists K_0 = K_0(d) \text{ such that if } \\ \|\delta d\| \leq \eta_0 \text{ then } \|\delta x\| \leq K_0 \|\delta d\| \tag{1.5}$$

The norms used fot the data and for the solution may not coincide, whenever d and x represent variables of different kinds.

Remark 4.1. The property of continuous dependence on the data could have been stated in the following alternative way, which is more akin to the classical form of Analysis $\forall \epsilon > 0 \ \exists \delta = \delta(\epsilon)$ such that if $\|\delta d\| \leq \delta$ then $\|\delta x\| \leq \epsilon$.

The form (1.5) is however more suitable to express in the following the concept of numerical stability, that is, the property that small perturbations on the data yield perturbations of the same order on the solution.

With the aim of making the stability analysis more quantitative, we introduce the following definition.

Definition 4.1. For problem (1.3) we define the relative condition number to be

$$K(d) = sup\{\frac{\|\delta x\|/\|x\|}{\|\delta d\|/\|d\|}, \delta d \neq 0, d + \delta d \in D\} \quad (1.6)$$

Whenever $d = 0$ or $x = 0$, it is necessary to introduce the absolute condition number, given by

$$K_{abs}(d) = sup\{\frac{\|\delta x\|}{\|\delta d\|}, \delta d \neq 0, d + \delta d \in D\} \quad (1.7)$$

Problem (1.3) is called ill-conditioned if $K(d)$ is “big” for any admissible datum d.

The property of a problem of being well-conditioned is independent of the numerical method that is being used to solve it. In fact, it is possible to generate stable as well as unstable numerical schemes for solving well-conditioned problems. The concept of stability for an algorithm or for a numerical method is analogous to that used for problem (1.3) and will be made precise in the next section.

Remark 4.2. (Ill-posed problems) Even in the case in which the condition number does not exist (formally, it is infinite), it is not necessarily true that the problem is ill-posed. In fact there exist well posed problems for which the condition number is infinite, but suuch that they can be reformulated in equivalente problems with a finite condition number.

If problem (1.3) admits a unique solution, then there necessarily exists a mappin G, that we call resolvent, between the sets of the data and of the solutions, such that

$$x = G(d), \text{ that is } F(G(d), d) = 0 \quad (1.8)$$

According to this definition, (1.4) yields $x+\delta x = G(d+\delta d)$. Assuming that G is differentiable in d and denoting formally by $G'(d)$ its derivative with respect to d (if $G : \mathbb{R}^n \to \mathbb{R}^m, G'(d)$ will be the Jacobian matrix of G evaluated at the vector d), a Taylor's expansion of G truncated at first order ensures that

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

wheere $\| \cdot \|$ is a suitable vector norm and $o(\cdot)$ is the classical infinitesimal symbol denoting an infinitesimal term of higher order with respect to its argument. Neglecting the infinitesimal of higher order with respect to $\|\|deltad\|$, from (1.6) and (1.7) we respectively deduce that

$$K(d) \approx \|G'(d)\| \frac{\|d\|}{\|G(d)\|}, \quad K_{abs}(d) \approx \|G'(d)\| \quad (1.9)$$

where the symbol $\| \cdot \|$, when applied to a matrix, denotes the induced matrix norm (1.10) associated with the vector norm introduced above. The estimates in (1.9) are of great practical usefulness in the analysis of problems in the form (1.8).

Theorem 4.1. Let $\|\cdot\|$ be a vector norm. The function

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} \quad (1.10)$$

is a matrix norm called induced matrix norm or natural matrix norm.

In view of (1.9), the quantity $\|G'(d)\|$ is an approximation of $K_{abs}(d)$ and is sometimes called first order absolute condition number. This latter represents the limit of the Lipschitz constant of G as the perturbation on the data tends to zero.

Such a number does not always provide a sound estimate of the condition number $K_{abs}(d)$. This happends, for instance, when G' vanishes at a point whilst G is nonnull in a neighborhood of the same point. For example, take $x = G(d) = cos(d) - 1$ for $d \in (-\pi/2, \pi/2)$, we have $G'(0) = 0$ while $K_{abs}(0) = 2/\pi$.

5. Stability of Numerical Methods

We shall henceforth suppose the problem (1.3) to be well posed. A numerical method for the approximate solution of (1.3) will consist, in general, of a sequence of approximate problems

$$F_n(x_n, d_n) = 0 \quad n \geqq \quad (1.11)$$

depending on a certain parameter n (to be defined case by case). The understood expectation is that $x_n \to x$ as $n \to \infty$, i.e. that the numerical solution converges to the exact solution. For that, it is necessary that $d_n \to d$ and that F_n approximates F, as $n \to \infty$. Precisely, if the datum d of problem (1.3) is admissible for F_n, we say that (1.11) is consistent if

$$F_n(x, d) = F_n(x, d) - F(x, d) \to 0 \text{ for } n \to \infty \quad (1.12)$$

where x is the solution to problem (1.3) corresponding to the datum d.

A method is said to be strongly consistent if $F_n(x,d)=0$ for any value of n and not only for $n\to\infty$.

In some cases problem (1.11) could take the following form

$$F_n(x_n, x_{n-1}, ..., x_{n-q}, d_n) = 0 \quad n \geq q \tag{1.13}$$

where $x_0, x_1, ..., x_{q-1}$ are given. In such case, the property of strong consistency becomes $F_n(x, x, ..., x, d) = 0$ for all $n \geq q$.

Recalling what has been previously state about problem (1.3), in order for the numerical method to be well posed (or stable) we require that for any fixed n, there exists a unique solution x_n corresponding to the datum d_n, that x_n depends continuously on the data. More precisely, let d_n be an arbitrary element of D_n, wehre D_n is the set of all admissible data for (1.11). Let δd_n be a perturbation admissible in the sense that $d_n + \delta d_n \in D_n$, and let δx_n denote the corresponding perturbation on the solution, that is

$$F_n(x_n + \delta x_n n, d_n + \delta d_n) = 0$$

Then we require that

$$\begin{aligned}&\exists \eta_0 = \eta_0(d_n) < 0, \exists K_0 = K_0(d_n) \text{ such that}\\ &\quad \text{if } \|\delta d_n\| \leq \eta_0 \text{ then } \|\delta x_n\| \leq K_0 \|\delta d_n\|\end{aligned} \tag{1.14}$$

As done in (1.6), we introduce for each problem in the sequence (1.11) the quantities

$$\begin{aligned}K_n(d_n) &= \sup\{\frac{\|\delta x_n\|/\|x_n\|}{\|\delta d_n\|/\|d_n\|}, \delta d_n \neq 0, d_n + \delta d_n \in D_n\},\\ K_{abs,n}(d_n) &= \sup\{\frac{\|\delta x_n\|}{\|\delta d_n\|}, \delta d_n \neq 0, d_n + \delta d_n \in D_n\}\end{aligned} \tag{1.15}$$

The numerical method is sait to be well condition if $K_n(d_n)$ is "small" for any admissible datum d_n, ill conditioned otherwise. As in (1.8), let us consider the case where, for each n, the functional relation (1.11) defines a mapping G_n between the sets of the numerical data and the solutions

$$x_n = G_n(d_n), \text{ that is } F_n(G_n(d_n)d_n) = 0 \tag{1.16}$$

Assuing that G_n is differentiable, we can obtain from (1.15)

$$K_n(d_n) \approx \|G'_n(d_n)\| \frac{\|d_n\|}{\|G_n(d_n)\|}, \quad K_{abs,n} \approx \|G'_n(d_n)\| \tag{1.17}$$

We observe that, in the case where the sets of admissible data in problems (1.3) and (1.11) coincide, we can use in (1.14) and (1.15) the quantity d instead of d_n. In such case, we can define the relative and absolute asymptotic condition number corresponding to the datum d as follows

$$K^{num}(d) = \lim_{k\to\infty} \sup_{n\geq k} K_n(d)$$

$$K^{num}_{abs}(d) = \lim_{k\to\infty} \sup_{n\geq k} K_{abs,n}(d)$$

The final foal of numerical approximation is, of course, to build, through numerical problems of type (1.11), solutions x_n that "get closer" to the solution of problem (1.3) as much as n gets larger. This concept is made precise in the next definition.

Definition 5.1. The numerical method (1.11) is convergent iff

$$\begin{aligned}&\forall \epsilon > 0 \exists n_0 = n_0(\epsilon), \exists \delta = \delta(n_0, \epsilon) > 0 \text{ such that}\\ &\forall n > n_0(\epsilon), \forall \delta d_n : \|\delta d_n\| \leq \delta \to \|x(d) - x_n(d + \delta d_n)\| \leq \epsilon\end{aligned} \tag{1.18}$$

where d is na admissible datum for the problem (1.3), $x(d)$ is the corresponding solution and $x_n(d+\delta d_n)$ is the solution of the numerical problem (1.11) with datum $d + \delta d_n$.

To verify the implication (1.18) it suffices to check that under the same assumptions

$$\|x(d + \delta d_n) - x_n(d + \delta d_n)\| \leq \frac{\epsilon}{2} \tag{1.19}$$

Indeed, thanks to (1.5) we have

$$\begin{aligned}\|x(d) - x_n(d + \delta d_n)\| &\leq \|x(d) - x(d + \delta d_n)\|\\ &+ \|x(d + \delta d_n) - x_n(d + \delta d_n)\| \leq K_0 \|\delta d_n\| + \frac{\epsilon}{2}\end{aligned}$$

Choosing $\delta = min\{\eta_0, \epsilon/(2K_0)\}$ one obtains (1.18).

Measures of the convergence of x_n to x are given by the absolute error or the relative error, respectively defined as

$$E(x_n) = |x - x_n|, \quad E_{rel}(x_n) = \frac{|x - x_n|}{|x|}(\text{ if } x, \neq 0) \tag{1.20}$$

In the cases where x and x_n are matrix or vector quantities, in addition to the definitions in (1.20) it is sometimes useful to introduce the relative error by component defined as

$$E^c_{rel}(x_n) = \max_{i,j} \frac{|x - x_n|_{ij}}{|x_{ij}|} \tag{1.21}$$

5.1. Relations between Stability and Convergence

The concepts of stability and convergence are strongly connected.

First of all, if problem (1.3) is well posed, a necessary condition in order for the numerical problem (1.11) to be convergent is that it is stable.

Let us thus assume that the method is convergent, that is, (1.18) holds for an arbitrary $\epsilon > 0$. We have

$$\|\delta x_n\| = \|x_n(d+\delta d_n) - x_n(d)\| \leq \|x_n(d) - x(d)\| \\ + \|x(d) - x(d+\delta d_n)\| + \|x(d+\delta d_n) - x_n(d+\delta d_n)\| \\ \leq K(\delta(n_0,\epsilon),d)\|\delta d_n\| + \epsilon \quad (1.22)$$

having used (1.5) and (1.19) twice. Choosing now δd_n, suchh that $\|\delta d_n\| \leq \eta_0$, we deduce that $\|\delta d_n\| \leq \eta_0$, we deduce that $\|\delta x_n\|/\|\delta d_n\|$ can be bounced by $K_0 = K(\delta(n_0,\epsilon),d) + 1$, provided that $\epsilon \leq \|\delta d_n\|$, so that the method is stable. Thus, we are interested in stable numerical methods since only these can e convergent.

The stability of a numerical method becomes a sufficient condition for the numerical problem (1.11) to converge if this latter is also consistent with problem (1.3). Indeed, under these assumptions we have

$$\|x(d+\delta d_n) - x_n(d+\delta d_n)\| \leq \|x(d+\delta d_n) - x(d)\|$$

$$+\|x(d) - x_n(d)\| + \|x_n(d) - x_n(d+\delta d_n)\|$$

Thanks to (1.5), the first term at right-hand side can be bounded by $\|\delta d_n\|$. A similar bound holds fot the third term, due to the stability property (1.14). Finally, concerning the remaining term, if F_n is differentiable with respect to the variable x, an expansion in a Taylor series gives

$$F_n(x(d),d) - F_n(x_n(d),d) = \frac{\partial F_n}{\partial x}|_{(x,d)}(x(d) - x_n(d))$$

for a suitable x "between" $x(d)$ and $x_n(d)$. Assuming also that $\partial f_n/\partial x$ is invertible, we get

$$x(d) - x_n(d) = (\frac{\partial F_n}{\partial x})^{-1}_{|(x,d)}[F_n(x(d),d) - F_n(x_n(d),d)] \quad (1.23)$$

On the other hand, replacing $F_n(x_n(d),d)$ with $F_n(x(d),d)$ and passing to the norms, we find

$$\|x(d) - x_n(d)\| \leq \|(\frac{\partial F_n}{\partial x})^{-1}_{|(x,d)}\|\|F_n(x(d),d) - F(x(d),d)\|$$

Thanks to (1.12) we can thus conclude that $\|x(d) - x_n(d)\| \to 0$ for $n \to \infty$. The result that has just been proved, although stated in qualitative terms, is a milestone in numerical analysis, knows as equivalence theorem (or LaxRichtmyer theorem): "for a consistent numerical method, stability is equivalent to convergence".

6. Sources of Error in Computational Models

Whenever the numerical problem (1.11) is an approximation to the mathematical problem (1.3) and this latter is in turn a model of a physical problem, we shall say that (1.11) is a computational model for PP.

In this process the global error, denoted by e, is expressed by the difference between the actually computed solution, $\hat{x}_n$, and the physical solution, x_{ph}, of which x provides a model. The global error e of the mathematical model, given by $x - x_{ph}$, and the error e_c of the computational model, $\hat{x}_n - x$, that is $e = e_m + e_c$.

The error e_m will in turn take into account the error of the mathematical model in strict sense and the error on the data. In the same way, e_c turns out o be the combination of the numerical discretization error $e_n = x_n - x$, the error e_a introduced by the numerical algorithm and the roundoff error introduced by the computer during the actial solution of problem (1.11).

In general, we can thus outline the following sources of error:

1. error due to the model, that can be controlled by a proper choice of the mathematical model;
2. errors in the data, taht can be reduced by enhacing the accuracy in the measurement of the data themself;
3. truncation error, arisisng form havind replaced in the numerical model limits by operations that involve a finite number of steps;
4. rounding errors.

The error at the items 3. and 4. give rise to the computational error. A numerical method will thus be convergent if this error can be made aritrarily small by increasing the computational effort. Of course, convrgence is the primary, albeit nor unique, goal of a numerical method, the others being accuracy, reliability and efficiencyy.

Accuracy meand that the errors are small with respect to a fiex tolerance. It is usually quantified by the order of infinitesimal of the error e_n with respect to the discretization characteristic parameter. By the way, we notice that machine precision does not limit, on theoretical grounds, the accuracy.

Reliability means it is likely that the global error can be gauranteed to be below a certain tolerance. Of course, a numerical model can be considered to be reliable only if suitably tested, that is,successfully applied to several test cases.

Efficiency means that the computational complexity that is needed to control the error is as small as possible.

By algorithm we mean a directive that indicates, through elementary operations, all the passages that are needed to solve a specific problem. An algorithm can in turn contain sub-algorithms and must have the feature of terminating after a finite number of elemnetary operations. As a consequence,the executor of the algorithm must find within the algorithm itseld all the instructions to completely solve the problem at hand.

Finally, the complexity of an algorithm is a measure of its executing time. Calculating the complexity of an algorithm is therefore a part of the analysis of the efficiency of a numerical method. Since several algorithms, with different complexities, can be employed to solve the same problem P, it is useful to introduce the concept of complexity of a problem, this latter meaning the complecity of the algorithm that has a minimum complexity among those solving P. The complexity of a problem is typically measured by a parameter directly associated with P.

7. Machine Representation of Numbers

Any machine operation is affected by rounding error or roundoff. They are due to the fact that on a computer only a finite subset of the set of teal numbers can be represented.

7.1. The Positional System

Let a base $\beta \in \mathbb{N}$ be fixed with $\beta \geq 2$, and let x be a real number with a finite number of digits x_k with $0 \leq x_k < \beta$ for $k = -m, ..., n$. The notation (conventionally adopted)

$$x_\beta = (-1)^s[x_n x_{n-1} ... x_1 x_0 x_{-1} x_{-2} ... x_{-m}, x_n \neq 0] \tag{1.24}$$

is called the positional representation of x with respect to the base β. The point between x_0 and x_{-1} is called decimal point if the base is 10, binary point if the base is 2, while s depends on the sign of x ($s = 0$ if x is positive, 1 if negative). Relation (1.24) actually means

$$x_\beta = (-1)^s \left(\sum_{k=-m}^{n} x_k \beta^k \right)$$

Any real number can be approximated by numbers having a finite representation. Indeed, having fixed the base β, the followin property holds

$$\forall \epsilon > 0, \forall x_\beta \in \mathbb{R}, \exists y_\beta \in \mathbb{R} \text{ such that } |y_\beta - x_\beta| < \epsilon$$

where y_β has finite positional representation.

In fact, given the positive number $x_\beta = x_n x_{n-1} ... x_1 x_0 x_{-1} x_{-2} ... x_{-m} ...$ with the number of digits, finite or infinite, for any $r \geq 1$ one can build two numbers

$$x_\beta^{(l)} = \sum_{k=0}^{r-1} x_{x-k} \beta^{n-k}, x_\beta^{(u)} = x_\beta^{(l)} = \beta^{n-r+1}$$

having r digits, such that $x_\beta^{(l)} < x_\beta < x_\beta^{(u)}$ and $x_\beta^{(u)} - x_\beta^{(l)} = \beta^{n-r+1}$. If r is chosen in such a way that $\beta^{n-r+1} < \epsilon$, then taking y_β equal to $x_\beta^{(l)}$ or $x_\beta^{(u)}$ yields the desired inequality. This result legitimates the computer representation of real numbers (and thus by a finite number of digits).

Although theoretically speaking all the bases are equivalent, in the computational practice three are the bases generally employed, base 2 in binary, base 10 or decimal and base 16 or hexadecimal. In what follows, we will assume that β is an even integer.

To simplify notations, we shall write x instead of x_β, leaving the base β understood.

8. Ejercicios

(1) Se considera la siguiente sucesión definida por recursión

$$x_0 = 1 \quad x_1 = \frac{1}{5} \quad x_{n+1} = \frac{36}{5}x_n - \frac{7}{5}x_{n-1}$$

Esta sucesión tiene como solución $x_n = \frac{1}{5^n}$. Utilizar Matlab para calcular $\frac{1}{5^n}$ utilizando la sucesión recusiva del principio, para $n \leq 32$. Hacer el estudio del error absoluto y relativo.

(2) Repetir el ejercicio anterior tomando $x_0 = 2$ y $x_1 = \frac{36}{5}$, teniendo en cuenta que la solución es ahora $x_n = \frac{1}{5^n} + 7^n$.

(3) Compara los resultados de los ejercicios 1 y 2 y dar una explicación formal de lo que está sucediendo.

Capítulo 2

Diferenciación Numérica

We have already seen one way to approximate the derivative of a function f:

$$f'(x) = \frac{f(x+h) - f(x)}{h} \tag{2.1}$$

for some small number h. To determine the accuracy of this approximation, we use Taylor's theorem, assuming that $f \in C^2$:

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2} f''(\xi) \quad \xi \in [x, x+h] \rightarrow$$

$$f'(x) = \frac{f(x+h) - f(x)}{h} - \frac{h}{2} f''(\xi)$$

The term $\frac{h}{2} f''(\xi)$ is called the truncation error, or, the discretization error, and the approximation is said to be first-order accurate since the truncation error is $O(h)$.

However, roundoff also plays a role in the evaluation of the finite difference quotient (2.1). For example, if h is so small that $x+h$ is rounded to x, then the computed difference quotient will be 0. More generally, even if the only error made is in rounding the values $f(x+h)$ and $f(x)$, then the computed difference quotient will be

$$\frac{f(x+h)(1+\delta_1) - f(x)(1+\delta_2)}{h} =$$

$$= \frac{f(x+mh) - f(x)}{h} + \frac{\delta_1 f(x+h) - \delta_2 f(x)}{h}$$

Since each $|\delta_i|$ is less than the machine precision ϵ, this implies that the rounding error is less than or equal to

$$\frac{\epsilon(|f(x)| + |f(x+h)|)}{h}$$

Since the truncation error is proportional to h and the rounding error is proportionalto $1/h$, the best accuracu is achieved when these two quantities are approximately equal. Ignoring the constants $|f''(\xi)/2|$ and $(|f(x)| + |f(x+h)|)$, this means that

$$h \approx \frac{\epsilon}{h} \rightarrow h \approx \sqrt{\epsilon}$$

and in this case the error (truncation error or rounding error) is about $\sqrt{\epsilon}$. Thus, with formula (2.1), we can approximate a derivative to only about the square root of the machine precision.

Another way to approimate the derivative is to use a centered-difference formula:

$$f'(x) = \frac{f(x+h) - f(x-h)}{2h} \tag{2.2}$$

We can again determine the truncation error by using the Taylor's theorem. Expanding $f(x+h)$ and $f(x-h)$ about the point x, we find

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2} f''(x) + \frac{h^3}{6} f'''(\xi)$$

$$\xi \in [x, x+h]$$

$$f(x-h) = f(x) - hf'(x) + \frac{h^2}{2} f''(x) - \frac{h^3}{6} f'''(\eta)$$

$$\eta \in [x-h, x]$$

Substracting the two equations and solving for $f'(x)$ gives

$$f'(x) = \frac{f(x+h) - f(x-h)}{2h} - \frac{h^2}{12}(f'''(\xi) + f'''(\eta))$$

Thus the truncation error is $O(h^2)$, and this difference formula is second-order accurate.

To study the effects of roundoff, we again make the simplyfying assumption that the only roundoff that occurs is in rounding the values $f(x+h)$ and $f(x-h)$. The the computed difference quotient is

$$\frac{f(x+h)(1+\delta_1) - f(x-h)(1+\delta_2)}{2h} =$$

$$= \frac{f(x+h) - f(x-h)}{2h} + \frac{\delta_1 f(x+h) - \delta_2 f(x-h)}{2h}$$

and the roundoff term $(\delta_1 f(x+h) - \delta_2 f(x-h))/2h$ is bounded in absolute value by $\epsilon(\delta_1 f(x+h) - \delta_2 f(x-h))/(2h)$. Once again ignoring constant terms involving f and its derivatives, the greatest accuracy is now achieved when

$$h^2 \approx \frac{\epsilon}{h} \rightarrow h \approx \epsilon^{1/3}$$

and the error (truncation error or rounding error) is $\epsilon^{2/3}$. With this formula we can obtain greater accuracy, to about the 2/3 power of the machine precision.

One can approimate higher derivatives similarly. To derive a second-order-accurate approximation to the second derivative, we again use Taylor's theorem:

$$f(x+h) = f(x)+hf'(x)+\frac{h^2}{2!}f''(x)+\frac{h^3}{3!}f'''(x)+\frac{h^4}{4!}f''''(\xi)$$

$$\xi \in [x, x+h]$$

$$f(x-h) = f(x)-hf'(x)+\frac{h^2}{2!}f''(x)-\frac{h^3}{3!}f'''(x)+\frac{h^4}{4!}f''''(\eta)$$

$$\eta \in [x-h, x]$$

Adding this two equations gives

$$f(x+h)+f(x-h) = 2f(x)+h^2f''(x)+\frac{h^4}{12}f''''(v)$$

$$v \in [\eta, \xi]$$

Solving for $f''(x)$, we obtain the formula

$$f''(x) = \frac{f(x+h)-2f(x)+f(x-h)}{h^2} - \frac{h^2}{12}f''''(v)$$

Using the approximation

$$f''(x) = \frac{f(x+h)-2f(x)+f(x-h)}{h^2} \tag{2.3}$$

the truncation error is $O(h^2)$. Note, however that a similar rounding error analysis predicts rounding error of size ϵ/h^2, so the smallest total error occurs when h is about $\epsilon^{1/4}$ and then the truncation error and the rounding error are each abour $\sqrt{\epsilon}$. With machine precision $\epsilon \approx 10^{-16}$, this means that h should not be taken to be less than about 10^{-4}. Evaluation of standard finite difference quotients for higher derivatives is even more sensitive to the effects of roundoff.

Bibliografía

[1] A. Quarteroni; R. Sacco; F. Sale. *Numerical Mathematics*. Springer, 2007.