Capítulo 1

Introducción

1. Vector Spaces

Definición 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

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

Definición 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.

Definición 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.

Propiedad 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 $A = (a_i j)$ 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.

Definición 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.

Definición 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

Definición 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.

Propiedad 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.

Observación 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.

Definición 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.

Observación 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).

Teorema 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 neightborhood 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.

Definición 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

1. Numerical Differentiation

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.

1.1. Formulas de tres puntos

Fórmula de tres puntos hacia delante

$$f'(x) = \frac{-3f(x)+4f(x+h)-f(x+2h)}{2h} + O(h^2) \tag{2.4}$$

Fórmula de tres puntos hacia atrás

$$f'(x) = \frac{f(x-2h)-4f(x-h)+3f(x)}{2h} + O(h^2) \tag{2.5}$$

Fórmula de tres puntos centrada

$$f'(x) = \frac{-f(x-h)+f(x+h)}{2h} - O(h^2) \tag{2.6}$$

1.2. Five-Point Formulas

One commo five-point formula is used to determine approximations for the derivative at the midpoint.

Five-Point Midpoint Formula

$$f'(x) = \frac{1}{12h}[f(x-2h)-8f(x-h)+8f(x+h) - f(x+2h)] + O(h^4) \tag{2.7}$$

Five-Point Endpoint Formula

$$f'(x) = \frac{1}{12h}[-25f(x)+48f(x+h)-36f(x+2h) + 16f(x+3h)-3f(x+4h)] + O(h^4) \tag{2.8}$$

1.3. Round-Off Error Instability

It is particularly important to pay attention to round-off when approximating derivatives. To illustrate the situation, let us examine the three-point midpoint formula,

$$f'(x_0) = \frac{1}{2h}[f(x_0+h)-f(x_0-h)] - \frac{h^2}{6}f^{(3)}(\xi_1)$$

more closely. Suppose that in evaluating $f(x_0+h)$ and $f(x_0-h)$ we encounter round-off errors $e(x_0+h)$ and $e(x_0-h)$. Then our computations actually use the values $\tilde{f}(x_0+h)$ and $\tilde{f}(x_0-h)$, which are related to the true values $f(x_0+h)$ and $f(x_0-h)$ by $f(x_0+h) = \tilde{f}(x_0+h)+e(x_0+h)$ and $f(x_0-h) = \tilde{f}(x_0-h)+e(x_0-h)$.

The total error in the approximation,

$$f'(x_0) - \frac{\tilde{f}(x_0+h)-\tilde{f}(x_0-h)}{2h} =$$

$$= \frac{e(x_0+h)-e(x_0-h)}{2h} - \frac{h^2}{6}f^{(3)}(\xi_1)$$

is due bot to round-off error, the first part, and to truncation error. If we assume that the round-off errors $e(x_0 \pm h)$ are bounded by some number $\epsilon > 0$ and

that the third derivative of f is bounded by a number $M > 0$, then

$$\left| f'(x_0) - \frac{\tilde{f}(x_0 + h) - \tilde{f}(x_0 - h)}{2h} \right| \leq \frac{\epsilon}{h} + \frac{h^2}{6} M$$

To reduce the truncation error, $h^2 M/6$, we need to reduce h. But h is reduced, the round-off error ϵ/h grows. In practice, then, it is seldom advantageous to let h be too small, because in that case the round-off error will dominate the calculations.

Capítulo 3

Número de Condición de una Matriz

1. Matrix Norms

Definición 1.1. A matrix norm is a mapping $\|\cdot\| : \mathbb{R}^{m\times n} \to \mathbb{R}$ such that:

1. $\|A\| \geq 0 \forall A \in \mathbb{R}^{m\times n}$ and $\|A\| = 0$ if and only if $A = 0$;
2. $\|\alpha A\| = |\alpha|\|A\| \forall \alpha \in \mathbb{R}, \forall A \in \mathbb{R}^{m\times n}$ (homogeneity);
3. $\|A + B\| \leq \|A\| + \|B\| \forall A, B \in \mathbb{R}^{m\times n}$ (triangular inequality).

Unless otherwise specified we shall employ the same symbol $\|\cdot\|$, to denote matrix norms and vector norms.

We can better characterize the matriz norms by introducing the concepts of compatible norm and norm induced by a vector norm.

Definición 1.2. We say that a matrix norm $\|\cdot\|$ is compatible or consistent with a vector norm $\|\cdot\|$ if

$$\|Ax\| \leq \|A\|\|x\|, \quad \forall x \in \mathbb{R}^n \tag{3.1}$$

More generally, given three norms, all denoted by $\|\cdot\|$, albeit defined on $\mathbb{R}^m$, $\mathbb{R}^n$ and $\mathbb{R}^{m\times n}$, respectively, we say that they are consistent if $\forall x \in \mathbb{R}^n$, $Ax = y \in \mathbb{R}^m$, $A \in \mathbb{R}^{m\times n}$, we hace that $\|y\| \leq \|A\|\|x\|$

In order to single out matrix norms of practical interest, the followinf property is in general required

Definición 1.3. We say that a matriz norm $\|\cdot\|$ is sub-multiplicative if $\forall A \in \mathbb{R}^{m\times n}$, $\forall B \in \mathbb{R}^{m\times q}$

$$\|AB\| \leq \|A\|\|B\| \tag{3.2}$$

This property is not satisfied by any matrix norm. For example, the norm $\|A\|_\Delta = \text{máx}\,|a_i j|$ for $i = 1, ..., n$, $j = 1, ..., m$ does nor satisfy (3.2) if applied to the matrices

$$A = B = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

since $2 = \|AB\|_\Delta < \|A\|_\Delta\|B\|_\Delta = 1$.

Notice that, given a certain sub-multiplicative matrix norm $\|\cdot\|_\alpha$, there always exists a consistent vector norm. For instance, given any fixed vector $y \neq 0$ in $\mathbb{C}^n$, it suffices to define the consistent vector norm as

$$\|x\| = \|xy^H\|_\alpha \quad x \in \mathbb{C}^n$$

As a consequence, in the case of sub-multiplicative matrix norms it is no longer necessary to explicitly specify the vector norm with respect to the matrix norm is consistent.

In view of the definition of a natural norm, we recall the following theorem.

Teorema 1.1. Let $\|\cdot\|$ be a vector norm. The function

$$\|A\| = \sup_{x\neq 0} \frac{\|Ax\|}{\|x\|} \tag{3.3}$$

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

Demostración. We start by noticing that (3.3) is equivalent to

$$\|A\| = \sup_{\|x\|=1} \|Ax\| \tag{3.4}$$

Indeed, one can define for any $x \neq 0$ the unit vector $u = x/\|x\|$, so that (3.3) becomes

$$\|A\| = \sup_{\|u\|=1} \|Au\| = \|Aw\| \quad \|w\| = 1$$

This being taken as given, let us check that (3.3) (or, equivalently, (3.4)) is actually a norm, making direct use of Definition 1.1

1. If $\|Ax\| \geq 0$, then it follows that $\|A\| = \sup_{\|x\|=1} \|Ax\| \geq 0$. Moreover

$$\|A\| = \sup_{\|x\|\neq 0} \frac{\|Ax\|}{\|x\|} = 0 \Leftrightarrow \|Ax\| = 0 \forall x \neq 0$$

and $Ax = 0 \forall x \neq 0$ if and only if $A = 0$; therefore $\|A\| = 0 \Leftrightarrow A = 0$.

2. Given a scalar α

$$\|\alpha A\| = \sup_{\|x\|=1} \|\alpha Ax\| = |\alpha| \sup_{\|x\|=1} \|Ax\| = |\alpha| \|A\|$$

3. Finally, triangular inequality holds. Indeed, by definition of suupremum, if $x \neq 0$ then

$$\frac{\|Ax\|}{\|x\|} \leq \|A\| \Rightarrow \|Ax\| \leq \|A\| \|x\|$$

so that, taking x with unit norm, one gets

$$\|(A+B)x\| =\leq \|Ax\| + \|Bx\| \leq \|A\| + \|B\|$$

from which it follows that $\|A + B\| = \sup_{\|x\|=1} \|(A+B)x\| \leq \|A\| + \|B\|$

□

Relevant instances of induces matrix norms are the so-called p-norms defined as

$$\|A\|_p = \sup_{x\neq 0} \frac{\|Ax\|_p}{\|x\|_p}$$

The 1-nrom and the infinity norm are easily computable since

$$\|A\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^{m} |a_{ij}|, \quad \|A\|_\infty = \max_{i=1,\ldots,m} \sum_{j=1}^{n} |a_{ij}|$$

and they are called the column sum norm and the row sum norm, respectively.

Moreover, we have $\|A\|_1 = \|A^T\|_\infty$ and, if A is self-adjoint or real symetric, $\|A\|_1 = \|A\|_\infty$.

A special discussion is deserved by the 2-norm or spectral norm for which the following theorem holds.

Teorema 1.2. Let $\sigma_1(A)$ be the largest sigular value of A. Then

$$\|A\|_2 = \sqrt{\rho(A^H A)} = \sqrt{\rho(AA^H)} = \sigma_1(A) \tag{3.5}$$

In particular, if A is hermitian (or real symmetric), then

$$\|A\|_2 = \rho(A) \tag{3.6}$$

while, if A is unitary, $\|A\|_2 = 1$.

Demostración. Since $A^H A$ is hermitian, there exists an unitary matrix U such that

$$U^H A^H A U = diag(\mu_1, ..., \mu_n)$$

where μ_i, are the (positive) eigenvalues of $A^H A$. Let $y = U^H x$, then

$$\|A\|_2 = \sup_{x\neq 0} \sqrt{\frac{(A^H Ax, x)}{(x,x)}} = \sup_{y\neq 0} \sqrt{\frac{(U^H A^H AUy, y)}{(y,y)}}$$

$$= \sup_{y\neq 0} \sqrt{\sum_{i=1}^{n} \mu_i |y_i|^2 / \sum_{i=1}^{n} |y_i|^2} = \sqrt{\max_{i=1,\ldots,n} |\mu_i|}$$

from which (3.5) follows, thanks to

If A is hermitian, the same considerations as above apply directly to A. Finally if A is unitary, we have

$$\|Ax\|_2^2 = (Ax, Ax) = (x, A^H Ax) = \|x\|_2^2$$

so that $\|A\|_2 = 1$. □

As a consequence, the computation of $\|A\|_2$ is much more extensive thant that of $\|A\|_\infty$ or $\|A\|_1$. However, if only an estimate of $\|A\|_2$ is required, the following relations can be profitably employed in the case of square matrices

$$\max_{i,j} |a_{ij}| \leq \|A\|_2 \leq n \max_{i,j} a_{ij}$$

$$\frac{1}{\sqrt{n}} \|A\|_\infty \leq \|A\|_2 \leq \sqrt{n} \|A\|_\infty$$

$$\frac{1}{\sqrt{n}} \|A\|_1 \leq \|A\|_2 \leq \sqrt{n} \|A\|_1$$

$$\|A\|_2 \leq \sqrt{\|A\|_1 \|A\|_\infty}$$

Moreover, if A is normal then $\|A\|_2 \leq \|A\|_p$ for any n and all $p \geq 2$.

Teorema 1.3. Let $|||\cdot|||$ be a matrix norm induced by a vector norm $\|\cdot\|$. Then, the following relations hold:

1. $\|Ax\| \leq |||A||| \|x\|$, that is $|||\cdot|||$ is a norm compatible with $\|\cdot\|$;

2. $|||I||| = 1$

3. $|||AB||| \leq |||A||| |||B|||$, that is, $|||\cdot|||$ is sub-multiplicative.

Demostración. Part 1 of the theorem is already contained in the proof of Theorem 1.1, while part 2 follows from the fact that $|||I||| = \sup_{x\neq 0} \|Ix\|/\|x\| = 1$. Part 3 is simple to check. □

Notice that the p-norms are sub-multiplicative. Moreover, we remark that the sub-multiplicativity property by itself would only allow us to conclude that $|||I||| \geq 1$. Indeed, $|||I||| = |||I \cdot I||| \leq |||I|||^2$.

2. Stability Analysis of Linear Systems

2.1. The Condition Number of a Matrix

The condition number of a matrix $A \in \mathbb{C}^{m\times n}$ is defined as

$$K(A) = \|A\|\|A^{-1}\| \tag{3.7}$$

where $\|\cdot\|$ is an induced matrix norm.

Capítulo 4

Interpolación

Capítulo 5

Integración

Capítulo 6

Raíces de Ecuaciones y Sistemas No Lineales

1. Soluciones de las ecuaciones en una variable

1.1. El método de bisección

Este proceso implica encontrar una raíz, o solución, para una ecuación de la forma $f(x) = 0$, para una función f dada. Una raíz de esta ecuación también recibe el nombre de cero de la función f.

La primera técnica, basada en el teorema del valor intermedio, recibe el nombre de bisección, o método de búsqueda binaria.

Suponga que f es una función continua definida dentro del intervalo $[a, b]$ con $f(a)$ y $f(b)$ de signos opuestos. El teorema del valor intermedio implica que existe un número p en (a, b) con $f(p) = 0$. A pesar de que el procedimiento operará cuando haya más de una raíz en el intervalo (a, b), para simplicidad, nosotros asumimos que la raíz en este intervalo es única. El método realiza repetidamente una reducción a la mitad (o bisección) de los subintervalos de $[a, b]$ y, en cada paso, localizar la mitad que contiene p.

Para comenzar, sea $a_1 = a$ y $b_1 = b$ y sea p_1 es el punto medio de $[a, b]$, es decir,

$$p_1 = a_1 + \frac{b_1 - a_1}{2} = \frac{a_1 + b_1}{2}$$

- Si $f(p_1) = 0$, entonces $p = p_1$ y terminamos.
- Si $f(p_1) \neq 0$, entonces $f(p_1)$ tiene el mismo signo que ya sea $f(a_1)$ o $f(b_1)$.
  - Si $f(p_1)$ y $f(a_1)$ tienen el mismo signo, $p \in (p_1, b_1)$. Sea $a_2 = p_1$ y $b_2 = b_1$
  - Si $f(p_1)$ y $f(a_1)$ tienen signos opuestos, $p \in (a_1, p_1)$. Sea $a_2 = a_1$ y $b_2 = p_1$.

Entonces, volvemos a aplicar el proceso al intervalo $[a_2, b_2]$.

Se pueden aplicar otros procedimientos de parada, por ejemplo, podemos seleccionar una tolerancia $\epsilon > 0$ y generar $p_1, ..., p_N$ hasta que se cumpla una de las siguientes condiciones:

$$|f(p_N)| < \epsilon \tag{6.1}$$

$$\frac{|p_N - p_{N-1}|}{|p_N|} < \epsilon, \quad p_N \neq 0 \tag{6.2}$$

$$|p_N - p_{N-1}| < \epsilon \tag{6.3}$$

El método de bisección, a pesar de que está conceptualmente claro, tiene desventajas significativas. Su velocidad de convergencia es más lenta y se podría descartar inadvertidamente una buena aproximación intermedia. Sin embargo, el método tiene la importante propiedad de que siempre converge a una solución y por esta razón con frecuencia se utiliza como iniciador para los métodos más eficientes que veremos más adelante en este capítulo.

Teorema 1.1. Suponga que $f \in C[a, b]$ y $f(a) \cdot f(b) < 0$. El método de bisección genera una sucesión $p_n{}_{n=1}^{\infty}$ que se aproxima a cero p de f con

$$|p_n - p| \leq \frac{b - a}{2^n}, \text{ cuando } n \geq 1$$

1.2. Iteración de punto fijo

Un punto fijo para una función es un número en el que el valor de la función no cambia cuando se aplica la función.

Definición 1.1. El número p es un punto fijo para la función dada g si $g(p) = p$.

Teorema 1.2.
- Si $g \in C[a, b]$ y $g(x) \in [a, b]$ para todas $x \in [a, b]$, entonces g tiene por lo menos un punto fijo en $[a, b]$.

- Si, además, $g'(x)$ existe en (a, b) y hay una constante positiva $k < 1$ con

$$|g'(x)| \leq k, \forall x \in (a, b)$$

entonces, existe exactamente un punto fijo en $[a, b]$. (Figura 6.1)

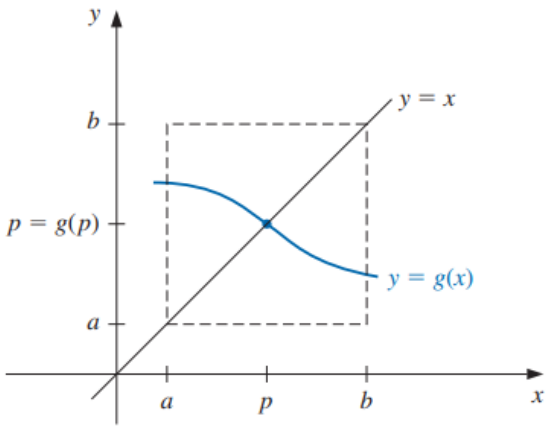

Figura 6.1: Teorema del Punto Fijo

Demostración. Si $g(a) = a$ o $g(b) = b$, entonces g tiene un punto fijo en un extremo. De lo contrario, entonces $g(a) > a$ y $g(b) < b$. La función $h(x) = g(x) - x$ es continua en $[a, b]$, con

$$h(a) = g(a) - a > 0 \quad \text{y} \quad h(b) = g(b) - b > 0$$

El teorema de valor intermedio implica que existe $p \in (a, b)$ para la cual $h(p) = 0$. Este número p es un punto fijo para g porque

$$0 = h(p) = g(p) - p \quad \text{implica que} \quad g(p) = p$$

Suponga, además, que $|g'(x)| \leq k < 1$ y que p y q son puntos fijos en $[a, b]$. Si $p \neq q$, entonces el teorema de valor medio implica que existe un número ξ entre p y q y por lo tanto en $[a, b]$ con

$$\frac{g(p) - g(p)}{p - q} = g'(\xi)$$

Por lo tanto

$$|p-q| = |g(p)-g(q)| = |g'(\xi)||p-q| \leq k|p-q| < |p-q$$

lo cual es una contradicción. Esta contradicción debe provenir de la única suposición $p \neq q$. Por lo tanto, $p = q$ y el punto fijo en $[a, b]$ es único. □

Iteración de punto fijo

Para aproximar el punto fijo de una función g, elegimos una aproximación inicial p_0 y generamos la sucesión $p_n{}_{n=0}^{\infty}$ al permitir $p_n = g(p_{n-1})$, para cada $n \geq 1$. Si la sucesión converge a p y g es continua, entonces

$$p = \lim_{n\to\infty} p_n = \lim_{n\to\infty} g(p_{n-1}) = g\left(\lim_{n\to\infty} p_{n-1}\right) = g(p)$$

y se obtiene una solución para $x = g(x)$. Esta técnica recibe el nombre de punto fijo o iteración funcional (Figura 6.2).

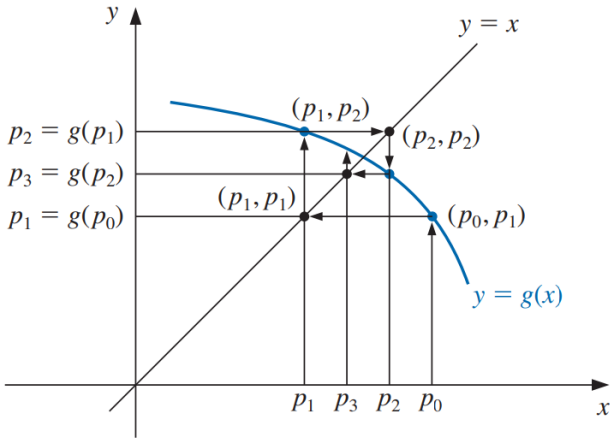

Figura 6.2: Iteración del Punto Fijo

Teorema 1.3. (Teorema de punto fijo) Sea $g \in C[a, b]$ tal que $g(x) \in [a, b]$ para todas las $x \in [a, b]$. Suponga, además, que existe g' en (a, b) y que existe una constante $0 < k < 1$ con

$$|g'(x)| \leq k, \text{ para todas } x \in (a, b)$$

Entonces, para cualquier número p_0 en $[a, b]$, la sucesión definida por

$$p_n = g(p_{n-1}), \quad n \geq 1$$

converge al único punto fijo p en $[a, b]$.

Demostración. El teorema 1.3 implica que existe un único punto p en $[a, b]$ con $g(p) = p$. Ya que g mapea $[a, b]$ en sí mismo, la sucesión $p_n{}_{n=0}^{\infty}$ se define para todas las $n \geq 0$ y $p_n \in [a, b]$ para todas las n. Al utilizar el hecho de que $|g'(x)| \leq k|$ y el teorema de valor medio **??**, tenemos, para cada n,

$$|p_n-p| = |g(p_{n-1})-g(p)| = |g'(\xi_n)||p_{n-1}-p| \leq k|p_{n-1}-p|$$

donde $\xi_n \in (a,b)$. Al aplicar esta desigualdad de manera inductiva obtenemos

$$|p_n - p| \leq k|p_{n-1} - p| \leq k^2|p_{n-2} - p| \leq ... \leq k^n|p_0 - p| \tag{6.4}$$

Ya que $0 < k < 1$, tenemos que $\lim_{n\to\infty} k^n = 0$ y

$$\lim_{n\to\infty} |p_n - p| \leq \lim_{n\to\infty} k^n|p_0 - p| = 0$$

Por lo tanto, $p_n{}_{n=0}^{\infty}$ converge a p. □

Observación 1.1. Si g satisface las hipótesis del teorema 1.3, entonces las octas del error relacionado con el uso de p_n para aproximar p, están dadas por

$$|p_n - p| \leq k^n \max\{p_0 - a, b - p_0\} \tag{6.5}$$

y

$$|p_n - p| \leq \frac{k^n}{1-k}|p_1 - p_0|, \text{ para toda } n \geq 1 \tag{6.6}$$

1.3. Método de Newton

El método de Newton (o de Newton-Raphson) es uno de los métodos numéricos más poderosos y reconocidos para resolver un problema de encontrar la raíz. Existen muchas formas de presentar el método de Newton.

Supona que $f \in C^2[a,b]$. Si $p_0 \in [a,b]$ es una aproximación para p, de tal forma que $f'(p_0 \neq 0)$ y $|p - p_0|$ es "pequeño". Considere que el primer polinomio de Taylor para $f(x)$ expandido alrededor de p_0 y evaluado en $x = p$:

$$f(p) = f(p_0) + (p - p_0)f'(p_0) + \frac{(p-p_0)^2}{2}f''(\xi(p))$$

donde $\xi(p)$ se encuentra entre p y p_0. Puesto que $f(p) = 0$, esta ecuación nos da

$$0 = f(p_0) + (p - p_0)f'(p_0) + \frac{(p-p_0)^2}{2}f''(\xi(p))$$

El método de Newton se deriva al suponer que como $|p - p_0|$ es pequeño, el término relacionado con $(p - p_0)^2$ es mucho más pequeño, entonces

$$0 \approx f(p_0) + (p - p_0)f'(p_0)$$

Al resolver para p obtenemos

$$p \approx p_0 - \frac{f(p_0)}{f'(p_0)} \equiv p_1$$

Esto constituye la base para el método de Newton, que empieza con una aproximación inicial p_0 y genera la sucesión $p_n{}_{n=0}^{\infty}$, mediante

$$p_n = p_{n-1} - \frac{f(p_{n-1})}{f'(p_{n-1})}, \text{ para } n \geq 1 \tag{6.7}$$

La figura 6.3 ilustra cómo se obtienen las aproximaciones usando tangentes sucesivas.

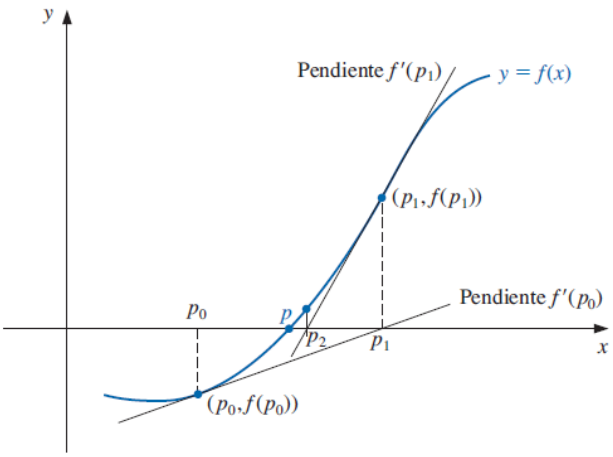

Figura 6.3: Método de Newton

Convergencia con el método de Newton

La derivación del método de Newton por medio de la serie de Taylor al inicio de la sección señala la importancia de una aproximación inicial precisa. La suposición crucial es que el término relacionado con $(p - p_0)^2$ es, en comparación con $|p - p_0|$, tan pequeño que se puede eliminar. Claramente esto será falso a menos que p_0 sea una buena aproximación para p. Si p_0 no está suficientemente cerca de la raíz real, existen pocas razones para sospechar que el método de Newton convergerá a la raíz. Sin embargo, en algunos casos, incluso las malas aproximaciones iniciales producirán convergencia.

Teorema 1.4. Teorema de Ostrowski. Sea $f \in C^2[a,b]$. Si $p \in (a,b)$ es tal que $f(p) = 0$ y $f'(p) \neq 0$, entonces existe una $\delta > 0$ tal que el método de Newton genera una sucesión $p_n{}_{n=0}^{\infty}$ que converge a p para cualquier aproximación inicial $p_0 \in [p - \delta, p + \delta]$.

El método de la secante

El método de Newton es una técnica en extremo poderosa, pero tiene una debilidad importante: la necesidad de conocer el valor de la derivada de f en

cada aproximación. Con frecuencia, $f'(x)$ es mucho más dificil y necesita más operaciones aritméticas para calcular $f(x)$.

Para evitar el problema de la evaluación de la derivada en el método de Newton, presentamos una ligera variación. Por definición,

$$f'(p_{n-1}) = \lim_{x \to p_{n-1}} \frac{f(x) - f(p_{n-1})}{x - p_{n-1}}$$

Si p_{n-2} está cerca de p_{n-1}, Entonces

$$f'(p_{n-1}) \approx \frac{f(p_{n-2}) - f(p_{n-1})}{p_{n-2} - p_{n-1}} = \frac{f(p_{n-1}) - f(p_{n-2})}{p_{n-1} - p_{n-2}}$$

Usando esta aproximación para $f'(p_{n-1})$ en la fórmula de Newton obtenemos

$$p_n = p_{n-1} - \frac{f(p_{n-1})(p_{n-1} - p_{n-2})}{f(p_{n-1}) - f(p_{n-2})} \tag{6.8}$$

Esta técnica recibe el nombre de método de la secante y se ilustra en la figura 6.4.

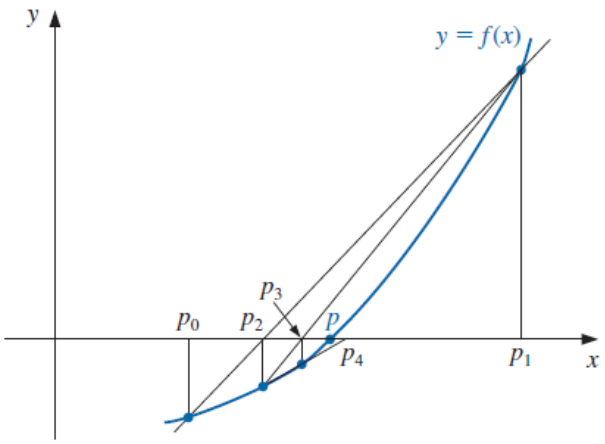

Figura 6.4: Método de la Secanta

El método de posición falsa

El método de posición falsa (también llamado Regula Falsi) genera aproximaciones de la misma manera que el método de la secante, pero incluye una prueba para garantizar que la raíz simepre se agrupa entre iteraciones sucesivas.

En primer lugar, seleccionamos las aproximaciones iniciales p_0 y p_1 con $f(p_0) \cdot f(p_1) < 0$. La aproximación p_2 se selecciona de la misma forma que en el método de la secante como la intersección en x de la recta que une $(p_0, f(p_0))$ y $(p_1, f(p_1))$. Para decidir cuál línea secante se usa para calcular p_3, considere $f(p_2) \cdot f(p_1)$ o, más concretamente, $sgnf(p_2) \cdot sgnf(p_1)$.

- Si $sgnf(p_2) \cdot sgnf(p_1) < 0$, entonces p_1 y p_2 agrupan una raíz. Seleccione p_3 como la intersección en x de la recta que une $(p_1, f(p_1))$ y $(p_2, f(p_2))$.
- Si no, seleccionamos p_3 como la intersección en x de la recta que une $(p_0, f(p_0))$ y $(p_2, f(p_2))$ y, a continuación intercambia los índices en p_0 y p_1.

De manera similar, una vez se encuentra p_3, el signo de $f(p_3) \cdot f(p_2)$ determina si usamos p_2 y p_3 o p_3 y p_1 para calcular p_4. En el último caso, se vuelve a etiquetar p_2 y p_1. Reetiquetar garantiza que la raíz se agrupa entre iteraciones sucesivas.

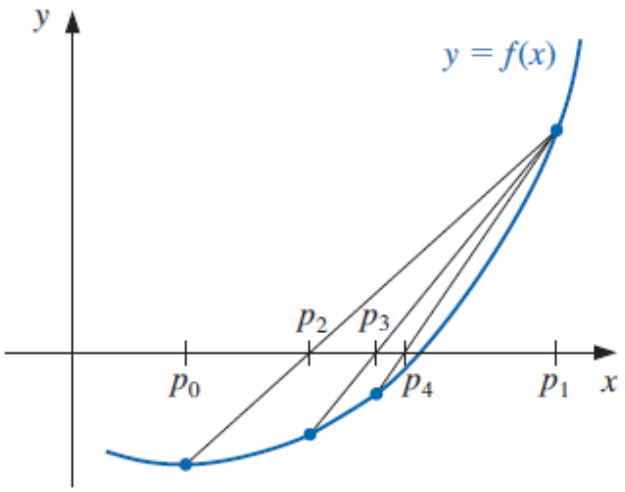

Figura 6.5: Método de la Posición Falsa

1.4. Análisis de error para métodos iterativos

En esta sección inverstigamos el orden de convergencia de esquemas de iteración funcional y, con el propósito de obtener convergencia rápida, redescubrimo el método de Newton. También consideramos consideramos formas para acelerar la convergencia del método de Newton en circunstancias especiales. Primero, sin embargo, necesitamos un nuevo procedimiento para medir qué tan rápido converge una sucesión.

Orden de convergencia

Definición 1.2. Suponga $p_n{}_{n=0}^{\infty}$ es una sucesión que converge a p, con $p_n \neq p$ para todas las n. Si existen constantes positivas λ y α con

$$\lim_{n \to \infty} \frac{|p_{n+1} - p|}{|p_n - p|^{\alpha}} = \lambda$$

Entonces $p_n{}_{n=0}^{\infty}$ converge a p de orden α, con constante de error asintótica λ

Se dice que una técnica iterativa de la forma $p_n = g(p_{n-1})$ es de orden α si la sucesión $p_n{}_{n=0}^{\infty}$ converge a la solución $p = g(p)$ de orden α.

En general, una sucesión con un alto orden converge más rápidamente uqe una sucesión con un orden más bajo. La constante asintótica afecta la velocidad de concergencia pero no el grado del orden. Se presta atención especial a dos casos:

1. Si $\alpha = 1$ (y $\lambda < 1$), la sucesión es linealmente convergente.
2. Si $\alpha = 2$, la sucesión es cuadráticamente convergente.

Teorema 1.5. Sea $g \in [a, b]$ tal que $g(x) \in [a, b]$ para todas las $x \in [a, b]$. Suponga además que g' es continua en (a, b) y que existe una constante positiva $k < 1$ con

$$|g'(x)| \leq k, \text{ para toda } x \in (a, b)$$

Si $g'(p) \neq 0$, entonces para cualquier número $np_0 \neq p$ en $[a, b]$, la sucesión

$$p_n = g(p_{n-1}), \text{ para } n \geq 1,$$

Converge sólo linealmente para el único punto fijo p en $[a, b]$.

Demostración. Sabemos que, a partir del teorema del punto fijo, la sucesión converge a p. Puesto que existe g' en (a, b), podemos aplicar el teorema del valor medio para g para demostrar que para cualquier n,

$$p_{n+1} - p = g(p_n) - g(p) = g'(\xi_n)(p_n - p)$$

donde ξ_n está entre p_n y p. Ya que $p_n{}_{n=0}^{\infty}$ converge a p, también tenemos que $\xi_n{}_{n=0}^{\infty}$ converge a p. Puesto que g_0 es continua en (a, b), tenemos

$$\lim_{n\to\infty} g'(\xi_n) = g'(p)$$

Por lo tanto

$$\lim_{n\to\infty} \frac{p_{n+1} - p}{p_n - p} = \lim_{n\to\infty} g'(\xi_n) = g'(p)$$

y

$$\lim_{n\to\infty} \frac{|p_{n+1} - p|}{|p_n - p|} = |g'(p)|$$

De este modo, si $|g'(p)| \neq 0$, la iteración de punto fijo muestra convergencia lineal con error asíntótico constante $|g'(p)|$. □

Teorema 1.6. Sea p una solución de la ecuación $x = g(x)$. Suponga que $g'(p) = 0$ y que g'' es continua con $|g''(x)| < M$ en un intervalo abierto I que contiene a p. Entonces existe $\delta > 0$ talo que para $p_0 \in [p - \delta, p + \delta]$, la sucesión definida por $p_n = g(p_{n-1})$, cuando $n \geq 1$, converge, por lo menos cuadráticamente a p. Además, con valores suficientemente grandes de n,

$$|p_{n+1} - p| < \frac{M}{2}|p_n - p|^2$$

Demostración. Seleccione k en $(0, 1)$ y $\delta > 0$ tal que el intervalo $[p - \delta, p + \delta]$, contenido en I, tenemos $|g'(x) \leq k$ y g'' continua.l Puesto que $|g'(x) \leq k < 1|$. Al expandir $g(x)$ en un polinomio lineal de Taylor, para $x \in [p - \delta, p + \delta]$ obtenemos

$$g(x) = g(p) + g'(p)(x - p) + \frac{g''(\xi)}{2}(x - p)^2$$

donde ξ se encuentra entre x y p. Las hipótesis $g(p) = p$ y $g'(p) = 0$ implican que

$$g(x) = p + \frac{g''(\xi)}{2}(x - p)^2$$

En especial, cuando $x = p_n$,

$$p_{n+1} = g(p_n) = p + \frac{g''(\xi_n)}{2}(p_n - p)^2$$

con ξ_n entre p_n y p. Por lo tanto,

$$p_{n+1} - p = \frac{g''(\xi_n)}{2}(p_n - p)^2$$

Puesto que $|g'(x)| \leq k < 1$ en $[p - \delta, p + \delta]$ en sí mismo, por el teorema de putno fijo se sigue que $p_n{}_{n=0}^{\infty}$ converge a p. Pero como ξ_n se encuentra entre p y p_n para cada n, entonces $\xi_n{}_{n=0}^{\infty}$ también converge a p y

$$\lim_{n\to\infty} \frac{|p_{n+1} - p|}{|p_n - p|^2} = \frac{|g''(p)|}{2}$$

Este reseultado implica que la sucesión $p_n{}_{n=0}^{\infty}$ es cuadráticamente convergente si $g''(p) \neq 0$ y de convergencia de orden superior si $g''(p) = 0$.

Puesto que g'' es continua y está estrictamente acotada por M en el intervalo $[p-\delta, p+\delta]$, esto también implica que, para los valores suficientemente grandes de n,

$$|p_{n+1} - p| < \frac{M}{2}|p_n - p|^2$$

□

Raíces múltiples

Definición 1.3. Una solución p de $f(x) = 0$ es un cero de multiplicidad m de f si para $x \neq p$, podemos escribir $f(x) = (x-p)^m g(x)$, donde $\lim_{x\to p} g(x) \neq 0$.

Teorema 1.7. La función $f \in C^1[a,b]$ tiene un cero simple en p en (a,b) si y sólo si $f(p) = 0$, pero $f'(p) \neq 0$.

Teorema 1.8. La función $f \in C^m[a,b]$ tiene un cero de multiplicidad m en p en (a,b) si y sólo si

$$0 = f(p) = f'(p) = f''(p) = ... = f^{(m-1)(p)}$$

pero

$$f^{(m)}(p) \neq 0$$

2. Soluciones numéricas de sistemas de ecuaciones no lineales

2.1. Puntos fijos para funciones de varias variables

Un sistema de ecuaciones no lineales tiene la forma

$$\begin{cases} f_1 = (x_1, x_2, ..., x_n) = 0 \\ f_2 = (x_1, x_2, ..., x_n) = 0 \\ \quad \vdots \\ f_2 = (x_1, x_2, ..., x_n) = 0 \end{cases} \tag{6.9}$$

donde cada función f_i se puede pensar como un mapeo de un vector $x = (x_1, x_2, ..., x_n)^t$ del espacio n dimensional $\mathbb{R}^n$ en la recta real $\mathbb{R}$

Este sistema de n ecuaciones no lineales en n variables también se puede representar al definir una función F de mapeo $\mathbb{R}^n$ en $\mathbb{R}^n$.

Si se utiliza notacion vectorial para representar las variables $x_a, x_2, ...x_n$, entonces el sistema asume la forma

$$F(x) = 0 \tag{6.10}$$

Las funciones $f_1, f_2, ..., f_n$ reciben el nombre de funciones coordenadas de F.

Definición 2.1. Sea f una función definida en un conjunto $D \subset \mathbb{R}^n$ en $\mathbb{R}$ y rango en $\mathbb{R}$. Se dice que la función f tiene límite L en x_0, escrito

$$\lim_{x\to x_0} f(x) = L$$

si, dado cualquier número $\epsilon > 0$, existe un número $\delta > 0$ con

$$|f(x) - L| < \epsilon$$

siempre que $x \in D$, y

$$0 < \|x - x_0\| < \delta$$

Definición 2.2. Sea f una función del conjunto $D \subset \mathbb{R}^n$. La función f es continua en $x_0 \in D$ siempre que exista $\lim_{x\to x_0} f(x_0)$ y

$$\lim_{x\to x_0} f(x) = f(x_0)$$

Además, f es continua en un conjunto D si f es continua en cada punto de D. Este concepto se expresa al escribir $f \in C(D)$

Definición 2.3. Sea F una función desde $D \subset \mathbb{R}^n$ a $\mathbb{R}^n$ de la forma

$$F(x) = (f_1(x), f_2(x), ..., f_n(x))^t$$

donde fi es un mapeo de $\mathbb{R}^n$ hasta $\mathbb{R}$ para cada i. Definimos

$$\lim_{x-x_0} F(x) = L = (L_1, L_2, ...L_n)^t$$

si y sólo si $\lim_{x\to x_0} f_i(x) = L_i$ para cada $i = 1, 2, ..., n$.

Teorema 2.1. Sea f una función de $D \subset \mathbb{R}^n$ a $\mathbb{R}$ y $x_0 \in D$. Suponga que existen todas las derivadas parciales de f y las constantes $\delta > 0$ y $K > 0$, de tal forma que siempre que $\|x - x_0\| < \delta$ y $x \in D$, tenemos

$$\left|\frac{\partial f(x)}{\partial x_j}\right| \leq K, \text{ para cada } j = 1, 2, ..., n$$

Entonces f es continua en x_0.

Puntos fijos en $\mathbb{R}^n$

Definición 2.4. Una función G desde $D \subset \mathbb{R}^n$ hasta $\mathbb{R}^n$ tiene un punto fijo en $p \in D$ si $G(p) = p$.

Teorema 2.2. Sea $D = (x_1, x_2, ..., x_n)^t | a_i \leq x_i \leq b_i, \text{ para cada } i = 1, 2, ..., n$ para algún conjunto de constantes $a_1, a_2, ..., a_n$ y $b_1, b_2, ..., b_n$. Suponga que G es una función continua

en $D \subset \mathbb{R}^n$ a $\mathbb{R}$ con la propiedad de que $G(x) \in D$, siempre que $x \in D$. Entonces G tiene un punto fijo en D.

Además, suponga que todas las funciones componentes de G tienen derivadas parciales continuas y que existe una constante $K < 1$ con

$$\left|\frac{\partial g_i(x)}{\partial x_j}\right| \leq G(x^{(k-1)}), \quad \text{siempre que } x \in D$$

para cada $j = 1, 2, ..., n$ y cada funcikón componente g_i. Entonces, la sucesión de punto fijo $x^{(k)}{}_{k=0}^{\infty}$ definida por $x^{(0)}$ seleccionada arbitrariamente en D y generada por medio de

$$x^{(k)} = G(x^{(k-1)}) \quad \text{para cada } k \geq 1$$

converge al único punto fijo $p \in D$ y

$$\|x^{(k)} - p\|_\infty \leq \frac{K^k}{1-K}\|x^{(1)} - x^{(0)}\|_\infty \tag{6.11}$$

2.2. Método de Newton

Para construir el algoritmo que conduce a un método de punto fijo adecuado en el caso unidemnsional, encontramos una función ϕ con la proopiedad de que

$$g(x) = x - \phi(x)f(x)$$

da convergencia cuadrática para el punto fijo p de la función g. A partir de esta condición el método de Newton evolucionó al seleccionar $\phi(x) = 1/f'(x)$ suponiendo que $f'(x) \neq 0$.

Un enfoque similar en el caso n-dimensional implica una matriz

$$A(x) = \begin{bmatrix} a_{11}(x) & a_{12}(x) & \dots & a_{1n}(x) \\ a_{21}(x) & a_{22}(x) & \dots & a_{2n}(x) \\ \vdots & \vdots & & \vdots \\ a_{n1}(x) & a_{n2}(x) & \dots & a_{nn}(x) \end{bmatrix} \tag{6.12}$$

donde cada una de las entradas $a_{ij}(x)$ es una función de $\mathbb{R}^n$ a $\mathbb{R}$. Esto requiere encontrar $A(x)$ de tal forma que

$$G(x) = x - A(x)^{-1}F(x)$$

da convergencia cuadrática para la solución de $F(x) = 0$, suponiendo que $A(x)$ es no singular en el punto fijo p de G.

Teorema 2.3. Si p es la solución de $G(x) = x$. Suponga que existe un número $\delta > 0$ con las propiedades:

- $\partial g_I/\partial x_j$ es continua en $N_\delta = x|\|x - p\| < \delta$, para cada $i = 1, 2, ..., n$ y $j = 1, 2, ..., n$
- $\partial^2 g_i(x)/(\partial x_j \partial x_k)$ es continua y $|\partial^2 g_i(x)/(\partial x_j \partial x_k)| \leq M$ para algunas constantes M, siempre que $x \in N_\delta$, para cada $i = 1, 2, ..., n$, $j = 1, 2, ..., n$ y $k = 1, 2, ..., n$.
- $\partial g_i(p)/\partial x_k = 0$, para cada $i = 1, 2, ..., n$ y $k = 1, 2, ..., n$.

Entonces, un número $\hat{\delta} \leq \delta$ existe de tal forma que la sucesión generada por $x^{(k)} = G(x^{(k-1)})$ converge de forma cuadrática en p para cualquier selección de $x^{(0)}$, siempre y cuando $\|x^{(0)} - p\| < \hat{\delta}$. Además,

$$\|x^{(k)} - p\|_\infty \leq \frac{n^2 M}{2}\|x^{(k-1)} - p\|_\infty^2, \text{ para cada } k \geq 1$$

La matriz jacobiana

Defina la matriz $J(x)$ mediante

$$J(x) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(x) & \frac{\partial f_1}{\partial x_2}(x) & \dots & \frac{\partial f_1}{\partial x_n}(x) \\ \frac{\partial f_2}{\partial x_1}(x) & \frac{\partial f_2}{\partial x_2}(x) & \dots & \frac{\partial f_2}{\partial x_n}(x) \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_n}{\partial x_1}(x) & \frac{\partial f_n}{\partial x_2}(x) & \dots & \frac{\partial f_n}{\partial x_n}(x) \end{bmatrix} \tag{6.13}$$

Una seleccion adecuada para $A(x)$ es $A(x) = J(x)$. La función G se define mediante

$$G(x) = x - J(x)^{-1}F(x)$$

y el procedimiento de iteración de punto fijo evoluciona al seleccionar $x^{(0)}$ y generar, para $k \geq 1$,

$$x^{(k)} = G(x^{(k-1)}) = x^{(k-1)} - J(x^{(k-1)})^{-1}F(x^{(k-1)}) \tag{6.14}$$

Esto recibe el nombre de método de Newton para sistemas no lineales y en general se espera que proporcione convergencia cuadrática, siempre y cuando se conozca un valor inicial suficientemente preciso y que $J(p)^{-1}$ exista.

Una debilidad en el método de Newton surge de la necesidad de calcular e invertir la matriz $J(x)$ en cada paso. En la práctica, el cálculo explícito de $J(x)^{-1}$ se evita al realizar la operación en una forma de dos pasos. Primero se encuentra un vector y que satisface $J(x^{(k-1)})y = -F(x^{(k-1)})$. Entonces, la nueva aproximación, $x^{(k)}$, se obtiene sumando y a $x^{(k-1)}$.

Capítulo 7

Optimización Sin Restricciones y Mínimos Cuadrados

1. Optimización Sin Restricciones

El punto $x^* \in \mathbb{R}^n$ es un minimizador global de f, mientras que x^* es un minimizador local de f si existe $R > 0$ tal que

$$f(x^*) \leq f(x), \quad \forall x \in B(x^*; R)$$

A lo largo de esta sección, siempre asumiremos que $f \in C^1(\mathbb{R}^n)$. Denotaremos por

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}(x), ..., \frac{\partial f}{\partial x_n}(x) \right)^T$$

el gradiente de f en un punto x. Si d es un vector distinto de cero en $\mathbb{R}^n$ entonces la derivada direccional de f respecto a d es

$$\frac{\partial f}{\partial d}(x) = \lim_{\alpha \to \infty} \frac{f(x + \alpha f) - f(x)}{\alpha}$$

y satisface

$$\frac{\partial f(x)}{\partial d} = \nabla f(x)^T d$$

Además, denotando por $(x, x + \alpha d)$ el segmento en $\mathbb{R}^n$ que une los puntos x y $x + \alpha d$, con $\alpha \in \mathbb{R}$, la expansión de Taylor asegura que existe un $\xi \in (x, x + \alpha d)$ tal que

$$f(x + \alpha d) - f(x) = \alpha \nabla f(\xi)^T d \tag{7.1}$$

Si $f \in C^2(\mathbb{R}^n)$, denotaremos por $H(x)$ (o $\nabla^2 f(x)$) la matriz Hessiana de f evaluada en un punto x, cuyos elementos son

$$h_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}, \quad i, j = 1, ..., n$$

En tal caso, se puede demostrar que, si $d \neq 0$, existe la derivada direccional de segundo orden y tenemos

$$\frac{\partial^2 f}{\partial d^2}(x) = d^T H(x) d$$

Para un ξ adecuado en $(x, x + d)$, también tenemos

$$f(x + d) - f(x) = \nabla f(x)^T d + \frac{1}{2} d^T H(\xi) d$$

Propiedad 1.1. Sea $x^* \in \mathbb{R}^n$ un minimizador local de f y supongamos que $f \in C^1(B(x^*; R))$ para un $R > 0$ adecuado. Entonces $\nabla f(x^*) = 0$. Además, si $f \in C^2(B(x^*; R))$, entonces $h(x^*)$ es semidefinida positiva. Por el contrario, si $\nabla f(x^*) = 0$ y $H(x^*)$ es definida positiva, entonces x^* es un minimizador local de f en $B(x^*; R)$.

Un punto x^* tal que $\nabla f(x^*) = 0$ se dice que es un punto crítico de f. Esta condición es necesaria para que se cumpla la optimalidad. Sin embargo, esta condición también se vuelve suficiente si f es una función convexa en $\mathbb{R}^n$, es decir, tal que para todo $x, y \in \mathbb{R}^n$ y para cualquier $\alpha \in [0, 1]$,

$$f[\alpha x + (1 - \alpha) y] \leq \alpha f(x) + (1 - \alpha) f(y) \tag{7.2}$$

1.1. Métodos de descenso

Dado un vector inicial $x^{(0)} \in \mathbb{R}^n$, calcular para $k \geq 0$ hasta la convergencia

$$x^{(k+1)} = x^{(k)} + \alpha_k d^{(k)} \tag{7.3}$$

donde $d(k)$ es una dirección elegida de manera apropiada y α_k es un parámetro positivo (llamado tamaño de paso) que mide el paso a lo largo de la dirección $d(k)$. Esta dirección $d(k)$ es una dirección de descenso si

$$\begin{aligned} d^{(k)} \nabla f(x^{(k)}) < 0 \quad \text{si} \quad \nabla f(x^{(k)}) \neq 0, \\ d^{(k)} = 0 \quad \text{si} \quad \nabla f(x^{(k)}) = 0. \end{aligned} \tag{7.4}$$

Un método de descenso es un método como 7.3, en el que los vectores $d(k)$ son direcciones de descenso.

Si $d_k \in \mathbb{R}^n$ es una dirección de descenso, entonces existe $\alpha_k > 0$, suficientemente pequeña, tal que

$$f(x^{(k)} + \alpha_k d^{(k)}) < f(x^{(k)}) \tag{7.5}$$

siempre que f sea diferenciable de manera continua. De hecho, tomando en 7.1 $\xi = x^{(k)} + \vartheta\alpha_k d^{(k)}$ con $\vartheta \in (0,1)$, y empleando la continuidad de $\bigtriangledown f$ obtenemos

$$f(x^{(k)} + \alpha_k d^{(k)}) - f(x^{(k)}) = \alpha_k \bigtriangledown f(x^{(k)})^T d^{(k)} + \epsilon \tag{7.6}$$

donde ϵ tiende a cero cuando α_k tiende a cero. Como consecuencia, si $\alpha_k > 0$ es suficientemente pequeño, el signo del lado izquierdo de 7.6 coincide con el signo de $\bigtriangledown f(x^{(k)})^T d^{(k)}$, de modo que 7.5 se satisface si $d^{(k)}$ es una dirección de descenso.

Diferentes elecciones de $d^{(k)}$ corresponden a diferentes métodos. En particular, recordamos los siguientes:

- El método de Newton, en el que
$$d^{(k)} = -H^{-1}(x^{(k)}) \bigtriangledown f(x^{(k)})$$
siempre que H sea definida positiva dentro de un vecindario suficientemente grande del punto x^*.
- Métodos de Newton inexactos, en lo que
$$d^{(k)} = -B_k^{-1} \bigtriangledown f(x^{(k)})$$
donde B_k es una aproximación adecuada de $H(x^{(k)})$.
- El método del gradiente o método del máximo descenso, que corresponde a establecer $d^{(k)} = -\bigtriangledown f(x^{(k)})$. Este método es, por lo tanto, un método de Newton inexacto, en el que $B_k = I$. También puede considerearse un método tipo gradiente, ya que ${d^{(k)}}^T \bigtriangledown f(x^{(k)}) = -\| \bigtriangledown f(x^{(k)})\|_2^2$
- El método del gradiente conjugado, para el cual
$$d^{(k)} = - \bigtriangledown f(x^{(k)}) + \beta_k d^{(k-1)}$$
donde β_k es un escalar que se selecciona adecuadamente de manera que las direcciones $d^{(k)}$ resulten ser mutuamente ortogonales con respecto a un producto escalar adecuado.

Un método para calcular α_k consiste en resolver el siguiente problema de minimización en una dimensión:

$$\text{enc. } \alpha \text{ tal que } \phi(\alpha) = f(x^{(k)} + \alpha d^{(k)}) \text{ sea minimizado} \tag{7.7}$$

En tal caso, tenemos el siguiente resultado.

Teorema 1.1. Considerando el método de descenso 7.3, si en el paso genérico k, el parámetro α_k se establece igual a la solución exacta de 7.7, entonces se cumple la siguiente propiedad de ortogonalidad:

$$\nabla f(x^{(k+1)})^T d^{(k)} = 0$$

1.2. Técnicas de búsqueda en línea

Los métodos con los que vamos a tratar en esta sección son técnicas iterativas que terminan tan pronto como se satisface algún criterio de detención basado en la precisión de α_k.

La experiencia práctica revela que no es necesario resolver con precisión el problema (7.29) para desarrollar métodos eficientes. En cambio, es crucial imponer alguna limitación sobre las longitudes de paso (y, por lo tanto, sobre los valores admisibles para α_k). De hecho, sin introducir ninguna limitación, una solicitud razonable sobre α_k sería que el nuevo iterado $x^{(k+1)}$ cumpla la desigualdad

$$f(x^{(k+1)}) < f(x^{(k)}) \tag{7.8}$$

donde $x^{(k)}$ y $d^{(k)}$ se han fijado. Con este propósito, el procedimiento basado en comenzar con un valor (suficientemente grande) de la longitud de paso α_k y reducir este valor a la mitad hasta que se cumpla 7.8, puede dar lugar a resultados completamente erróneos.

1.3. Métodos tipo Newton para la minimización de funciones

Capítulo 8

Optimización Con Restricciones

1. Introducción

El caso más simple de optimización con restricciones se puede formular de la siguiente manera. Dada una función $f : \mathbb{R}^n \to R$,

$$minimize \quad f(x), \quad \text{con } x \in \Omega \subset \mathbb{R}^n \tag{8.1}$$

Más precisamente, se dice que un punto x^* es un minimizador global en Ω si satisface 8.1, mientras que es un minimizador local si $\exists R > 0$ tal que

$$f(x^*) \leq f(x), \quad \forall x \in B(x^*; R) \subset \Omega$$

La existencia de soluciones para el problema 8.1 está, por ejemplo, asegurada por el Teorema de Weiertrass, en el caso en que f sea continua y Ω sea un conjunto cerrado y acotado. Bajo la suposición de que Ω es un conjunto convexo, se cumplen las siguientes condiciones de optimalidad.

Propiedad 1.1. Sea $\Omega \subset \mathbb{R}^n$ y $f \in C^1(B(x^*; R))$, para una $R > 0$ adecuado. Entonces:

1. Si x^* es un minimizador local de f, entonces:

$$\bigtriangledown f(x^*)^T(x - x^*) \geq 0, \quad \forall x \in \Omega \tag{8.2}$$

2. Además, si f es convexa en Ω, entonces x^* es un minimizador global de f.

Propiedad 1.2. Sea $\Omega \subset \mathbb{R}^n$ un conjunto cerrado y convexo, y sea f una función fuertemente convexa en Ω. Entonces, existe un único minimizador local $x^* \in \Omega$

Un caso notable de 8.1 es el siguiente problema: dada $f : \mathbb{R}^n \to \mathbb{R}$

$$minimize \quad f(x), \text{ bajo la restricción de que } h(x) = 0 \tag{8.3}$$

donde $h : \mathbb{R}^n \to \mathbb{R}$, con $m \leq n$, es una función dada con componentes $h_1, ..., h_m$. Los analogos de los puntos críticos en el problema 8.3 se denominan puntos regulares.

Definición 1.1. Un punto $x^* \in \mathbb{R}^n$, tal que $h(x^*) = 0$, se dice que es regular si los vectores de la matriz Jacobiana $J_h(x^*)$ son linealmente independientes, asumiendo que $h_i \in C^2(B(x^*; R))$, para un $R > 0$ adecuado y $i = 1, ..., m$

Nuestro objetivo ahora es convertir el problema 8.3 en un problema de minimización sin restricciones.

Por ese motivo, introducimos la función del Lagrangiano $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^{\geqslant} \to \mathbb{R}$

$$\mathcal{L}(x, \lambda) = f(x) + \lambda^T h(x)$$

donde el vector λ se conoce como multiplicador de Lagrange. Además, denotemos por $J_{\mathcal{L}}$ la matriz Jacobiana asociada a $\mathcal{L}$, pero donde las derivadas parciales se toman únicamente con respecto a las variables $x_1, ..., x_n$.

Propiedad 1.3. Sea x^* un minimizador local para 8.3 y suponiendo que, para un $R > 0$ adecuado, $f, h_i \in C^1(B(x^*; R))$, para $i = 1, ..., m$. Entonces, existe un vector único $\lambda^* \in \mathbb{R}^m$ tal que $J_{\mathcal{L}}(c^*, \lambda^*) = 0$.

Por el contrario, supongamos que $x^* \in \mathbb{R}^n$ satisface $h(x^*) = 0$ y que, para un $R > 0$ adecuado y $i = 1, ..., m$, $f, h_i \in X^2(B(x^*; R))$. Sea $H_{\mathcal{L}}$ la matriz de entradas $\frac{\partial^2 \mathcal{L}}{\partial x_i \partial x_j}$ para $i, j = 1, ..., n$. Si existe un vector $\lambda^* \in \mathbb{R}^m$ tal que $J_{\mathcal{L}}(x^*, \lambda^*) = 0$ y

$$z^T H_{\mathcal{L}}(x^*, \lambda^*) z > 0 \quad \forall z \neq 0, \quad \text{con } \bigtriangledown h(x^*)^T z = 0$$

entonces x^* es un minimizador estricto local de 8.3.

La última clase de problemas que vamos a tratar incluye el caso en el que también están presentes restricciones de desigualdad, es decir: dado $f : \mathbb{R}^n \to \mathbb{R}$,

$$minimize f(x), \text{ bajo la restricción } h(x) = 0 \text{ y } g(x) \leq 0 \tag{8.4}$$

donde $h : \mathbb{R}^n \to \mathbb{R}^m$, con $m \leq n$, y $g : \mathbb{R}^n \to \mathbb{R}^r$ son dos funciones dadas. Se entiende que $g(x) \leq 0$ significa que $g_i(x) \leq 0$ para $i = 1, ..., r$. Las restricciones de desigualdad generan una complicación formal adicional con respecto al caso examinado previamente, pero no impiden convertir la solución de 8.4 en la minimización de una función Lagangiana adecuada.

Definición 1.2. Supongamos que $h_i, g_j \in C^1(B(x^*; R))$ para un $R > 0$ adecuado, con $i = 1, ..., m$ y $j = 1, ..., r$, y denotemos por $J(x^*)$ el conjunto de índices j tales que $g_j(x^*) = 0$. Un punto $x^* \in \mathbb{R}^n$ tal que $h(x^*) = 0$ y $g(x^*) \leq 0$ se dice que es regular si los vectores columna de la matriz Jacobiana $J_h(x^*)$ junto con los vectores $\bigtriangledown g_j(x^*), j \in \mathcal{J}(x^*)$, forman un conjunto de vectores linealmente independientes.

2. Condiciones necesarias de Kuhn-Tucker para la programación no lineal

En esta sección recordamos algunos resultados, conocidos como las condiciones de Kuhn-Tucker, que aseguran en general la existencia de una solución local para el problema de programación no lineal. Bajo suposiciones adecuadad, también garantizan la existencia de una solución global. A lo largo de esta sección suponemos que un problema de minimización siempre se puede reformular como un problema de maximización.

Consideremos el problema general de programación no lineal:

$$\begin{aligned} &\text{dada } f : \mathbb{R}^n \to \mathbb{R} \\ &maximize \quad f(x), \text{ sujeto a} \\ &g_i(x) \leq b_i, \quad i = 1, ..., l \\ &g_i(x) \geq b_i, \quad i = l+1, ..., k \\ &g_i(x) = b_i, i = k+1, ..., m \\ &x \geq 0 \end{aligned} \tag{8.5}$$

Un vector x que satisface las restricciones anteriores se llama una solución factible de 8.5 y el conjunto de las soluciones factibles se llama la región factible. Suponemos en adelante que $f, g_i \in C^1(\mathbb{R}^n)$, con $i = 1, ..., m$ y definimos conjuntos $I_= = i : g_i(x^*) = b_i$, $I_{/==} \; i : g_i(x^*) \neq b_i$, $J_= = i : x_i^* = 0$, $J_> = i : x_i^* > 0$, denotando por x^* un maximizador local de f. Asociamos con 8.5 el siguiente Lagrangiano

$$\mathcal{L}(x, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i [b_i - g_i(x)] - \sum_{i=m+1}^{m+n} \lambda_i x_{i-m}$$

Propiedad 2.1. Si f tiene un máximo local restringido en el punto $x = x^*$, es necesario que el vector $\lambda^* \in \mathbb{R}^{m+1}$ exista tal que (primera condición de Kuhn-Tucker):

$$\bigtriangledown_x \mathcal{L}(x^*, \lambda^*) \leq 0$$

donde se cumple igualdad estricta para componente $i \in J_>$. Además (segunda condición de Kuhn-Tucker):

$$\bigtriangledown_x \mathcal{L}(x^*, \lambda^*)^T x^* = 0$$

La tercera condición de Kuhn-Tucker requiere que:

$$\begin{aligned} \bigtriangledown_\lambda \mathcal{L}(x^*, \lambda^*) \geq 0 \quad i = 1, ...l \\ \bigtriangledown_\lambda \mathcal{L}(x^*, \lambda^*) \leq 0 \quad i = l+1, ...k \\ \bigtriangledown_\lambda \mathcal{L}(x^*, \lambda^*) = 0 \quad i = k+1, ...m \end{aligned}$$

Además (cuarta condición de Kuhn-Tucker):

$$\bigtriangledown_x \mathcal{L}(x^*, \lambda^*)^T x^* = 0$$

Propiedad 2.2. Supongamos que la función f en 8.5 es una función cóncava (es decir, $-f$ es convexa) en la región factible. Supongamos también que el punto (x^*, λ^*) satisface todas las condiciones necesarias de Kuhn-Tucker y que las funciones g_i para las cuales $\lambda_i^* > 0$ son convexas, mientras que aquellas para las cuales $\lambda_i^* < 0$ son cóncavas. Entonces, $f(x^*)$ es el máximo global restringido de f para el problema 8.5.

3. Método del Penalti

La idea básica de este método es eliminar, parcial o completamente, las restricciones para transformar el problema restringido en uno no restringido. Este nuevo problema se caracteriza por la presencia de un parámetro que proporciona una medida de la precisión con la que se impone efectivamente la restricción.

Consideremos el problema restringido 8.3, suponiendo que buscamos la solución x^* únicamente en $\Omega \subset \mathbb{R}^n$. Supongamos que dicho problema admite al menos una solución en Ω y lo escribimos en la siguiente forma penalizada:

$$minimize \quad \mathcal{L}_\alpha(x) \quad \text{para } x \in \Omega \tag{8.6}$$

donde

$$\mathcal{L}_\alpha(x) = f(x) + \frac{1}{2}\alpha \|h(x)\|_2^2$$

La función $\mathcal{L}_\alpha : \mathbb{R}^n \to \mathbb{R}$ se llama el Lagrangiano penalizado, y α se llama el parámetro de penalización. Es claro que si la restricción se satisficiera exactamente, entonces minimizar f sería equivalente a minimizar $\mathcal{L}_\alpha$.

El método de penalización es una técnica iterativa para resolver 8.6.

Para $k = 0, 1, ...$, hasta la convergencia, se debe resolver la secuancia de problemas:

$$minimize \quad \mathcal{L}_{\alpha_k}(x) \quad \text{con } x \in \Omega \tag{8.7}$$

donde α_k es una secuencia monótamente creciente de parámetros de penalización positivos, tal que $\alpha_k \to \infty$ cuando $k \to \infty$.

Como consecuencia, después de elegir α_k, en cada paso del proceso de penalizaciónd ebemos resolver un problema de minimización con respecto a la variable x, lo que da lugar a una seguencia de valores x_k^*, soluciones de 8.7. Al hacer esto, la función objetivo $\mathcal{L}_{\alpha_k}(x)$ tiende a infinito, a menos que $h(x)$ sea igual a cero.

Los problemas de minimización pueden ser resueltos luego mediante uno de los métodos introducidos en el capítulo anterior. La siguiente propiedad garantiza la convergencia del método de penalización en la forma 8.6.

Propiedad 3.1. Supongamos que $f : \mathbb{R}^\kappa \to \mathbb{R}$ y $h : \mathbb{R}^n \to \mathbb{R}^m$, con $m \leq n$, son funciones continuas en un conjunto cerrado $\Omega \subset \mathbb{R}^n$ y supongamos que la secuencia de parámetros de penalización $\alpha_k > 0$ es monótamente divergente. Finalmente, sea x_k^* el minimizador global del problema 8.7 en el paso k. Entonces, al tomar el límite cuando $k \to \infty$, la secuencia x_k^* converge a x^*, que es un minimizador global de f en Ω y satisface la restricción $h(x^*) = 0$.

En cuanto a la selección de parámetros α_k, se puede demostrar que valores grandes de α_k hacen que el problema de minimización 8.7 esté mal condicionado, lo que hace que su solución sea bastante costosa, a menos que la suposición inicial esté particularmente cerca de x^*. Por otro lado, la secuencia α_k no debe crecer demasiado lentamente, ya que esto afectaría negativamente la convergencia global del método.

Una elección que se hace comúnmente en la práctica es seleccionar un valor no demasiado grande α_0 y luego establecer $\alpha_k = \beta\alpha_{k-1}$ para $k > 0$ donde β es un número entero entre 4 y 10.

4. Método de los multiplicadores de Lagrange

Una variante del método de penalización hace uso de la función Lagrangiana aumentada $\mathcal{G}_\alpha : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ dada por

$$\mathcal{G}_\alpha(x, \lambda) = f(x) + \lambda^T h(x) + \frac{1}{2}\alpha\|h(x)\|_2^2 \tag{8.8}$$

donde $\lambda \in \mathbb{R}^m$ es un multiplicador de Lagrange. Claramente, si x^* es una solución del problema 8.3, entonces también será solución de 8.8, pero con la ventaja, respecto a 8.6, de disponeer de un grado de libertad adicional λ. El método de penalización aplicado a 8.8 se formula de la siguiente manera: para $k = 0, 1...$, hasta la convergencia, resolver la secuencia de problemas

$$minimize \quad \mathcal{G}_{\alpha_k}(x, \lambda_k) \text{ para } x \in \Omega \tag{8.9}$$

donde α_k es una secuencia acotada de vectores desconocidos en $\mathbb{R}^m$, y los parámetros α_k se definen como antes.

En cuanto a la elección de los multiplicadores, la secuencia de vectores λ_k se asigna típicamente mediante la siguiente fórmula:

$$\lambda_{k+1} = \lambda_k + \alpha_k h(x^{(k)})$$

donde λ_0 es un valor dado, mientras que la secuencia de α_k puede establecerse a priori o modificarse durante la ejecución.

En lo que respecta a las propiedades de convergencia del método de los multiplicadores de Lagrange, se cumple el siguiente resultado local.

Propiedad 4.1. Supongamos que x^* es un minimizador local estricto regular del problema 8.3 y que:

1. $f, h_i \in C^2(B(x^*; R))$ con $i = 1, ..., m$ y para un $R > 0$ adecuado.
2. El par (x^*, λ^*) satisface $z^T H_{\mathcal{G}_0}(x^*, \lambda^*)z > 0, \forall z \neq 0$ tal que $J_h(x^*)^T z = 0$
3. Existe un $\overline{\alpha} > 0$ tal que $H_{\mathcal{G}_{\overline{\alpha}}}(x^*, \lambda^*) > 0$

Entonces, existen tres escalares positivos δ, γ y M tal que, para cualquier par $(\lambda, \alpha) \in V = \{(\lambda, \alpha) \in \mathbb{R}^{m+1} : \|\lambda - \lambda^*\|_2 < \delta\alpha, \alpha \geq \overline{\alpha}\}$, el problema

$$minimize \quad \mathcal{G}_\alpha(x, \lambda), \text{ para } x \in B(x^*; \gamma)$$

admite una solución única $x(\lambda, \alpha)$, diferenciable con respecto a sus argumentos. Además, para todo $(\lambda, \alpha) \in V$

$$\|x(\lambda, \alpha) - x^*\|_2 \leq M\|\lambda - \lambda^*\|_2$$

Bajo suposiciones adicionales se puede demostrar que el método de los multiplicadores de Lagrange converge. Además, si $\alpha_k \to \infty$ cuando $k \to \infty$, entonces

$$\lim_{k\to\infty} \frac{\|\lambda_{k+1} - \lambda^*\|_2}{\|\lambda_k - \lambda^*\|_2} = 0$$

y la convergencia del método es más que lineal.

En el caso en que la secuencia α_k tenga una cota superior, el método converge linealmente.

Finalmente, cabe señalar que, a diferencia del método de penalización, ya no es necesario que la secuencia α_k tienda a infinito. Esto, a su vez, limita el mal condicionamiento del problema 8.9 a medida que α_k crece. Otra ventaja se refiere a la tasa de convergencia del método, que resulta ser independiente de la tasa de crecimiento del parámetro de penalización en el caso de la técnica de los multiplicadores de Lagrange. Esto, por supuesto, implica una reducción considerable del coste computacional.

Capítulo 9

Métodos para Ecuaciones Diferenciales Ordinarias

En este capítulo estudiamos problemas de la forma

$$y'(t) = f(t, y(t)), \quad t \geq t_0$$
$$y(t_0) \equiv y_0 \tag{9.1}$$

donde la variable independiente t normalmente representa tiempo, $y \equiv y(t)$ es la función desconocida que buscamos, y y_0 es un valor inicial conocido. Esto se llama problema del valor inicial (PVI) para la ecuación diferencial ordinaria (EDO) $y' = f(t, y)$.

1. Existencia y unicidad de las soluciones

Al estudiar cuestiones sobre la existencia y unicidad de soluciones para la ecuación 9.1, consideramos que la función del lado derecho es una función de dos varibales independientes t y y, y hacemos suposiciones sobre su comportamiento como función de cada una de estas variables. El siguiente teorema proporciona condiciones suficientes para que el problema de valor inicial tenga una solución localmente.

Teorema 1.1. Si f es continua en un rectángulo R centrado en (t_0, y_0),

$$R = (t, y) : |t - t_0| \leq \alpha, |y - y_0| \leq \beta$$

entonces el PVI 9.1 tiene una solución $y(t)$ para $|t - t_0| \leq min(\alpha, \beta/M)$, donde $M = max_R|f(t, y)|$.

Aunque exista una solución, esta puede no ser única. El siguiente teorema proporciona condiciones suficientes para la existencia y unicidad locales.

Teorema 1.2. Si f y $\frac{\partial f}{\partial y}$ son continuas en el rectángulo R, entonces el problema de valor inicial tiene una única solución $y(t)$ para $|t - t_0| \leq min(\alpha, \beta/M)$ donde $M = max_R|f(t, y)|$.

1.1. Teoría elemental de problemas de valor inicial

Definición 1.1. Se dice que una función $f(t, y)$ satisface la condición de Lipschitz en la variable y en un conjunto $D \subset \mathbb{R}^n$ si existe una constante $L > 0$ con

$$|f(t, y_1) - f(t, y_2)| \leq L|y_1 - y_2|$$

siempre que (t, y_1) y (t, y_2) estén en D. La constante L recibe el nombre de constante de Lipschitz para f.

Definición 1.2. Se dice que un conjunto $D \subset \mathbb{R}^n$ es convexo siempre que (t_1, y_1) y (t_2, y_2) pertenezcan a D, entonces $((1 - \lambda)t_1 + \lambda t_2, (1 - \lambda)t_1 + \lambda t_2)$ también pertenece a D para cada λ en $[0, 1]$.

Teorema 1.3. Suponga que $f(t, y)$ se define sobre un conjunto convexo $D \subset \mathbb{R}^n$. Si existe una constante $L > 0$ con

$$\left|\frac{\partial f}{\partial y}(t, y)\right| \leq L, \text{ para todo } (t, y) \in D \tag{9.2}$$

entonces f satisface la condición de Lipschitz en D en la variable y con constante L de Lipschitz.

Definición 1.3. El problema de valor inicial

$$\frac{dy}{dt} = f(t, y), \quad a \leq t \leq b, \quad y(a) = \alpha \tag{9.3}$$

se dice que es un problema bien planteado si:

- Existe una única solución $y(t)$
- Existen constantes $\epsilon > 0$ y $k > 0$, tales que para cualquier ϵ en $(0, \epsilon_0)$, siempre que $\delta(t)$ es continua con $|\delta(t)| < \epsilon$ para toda t en $[a, b]$, y cuando $|\delta_0| < \epsilon$, el problema de valor inicial

$$\frac{dz}{dt} = f(t, z) + \delta(t), \quad a \leq t \leq b, \quad z(a) = \alpha + \delta_0 \tag{9.4}$$

tiene una única solución $z(t)$ que satisface

$$|z(t) - y(t)| < k\epsilon \text{ para toda } t \in [a, b]$$

El problema especificado por la ecuación 9.4 recibe el nombre de problema perturbado relacionado con el problema original en la ecuación 9.3. Suponge la posibilidad de un error introducido en la declaración de la ecuación diferencial, así como un error ∂_0 presente en la condición inicial.

Teorema 1.4. Suponga que $D = (t, y =)|a \leq t \leq b$ y $-\infty < y < \infty$. Si f es continua y satisface la condición de Lipschitz en la variable y sobre el conjunto D, entonces el problema de valor inicial

$$\frac{dy}{dt} = f(t, y), a \leq t \leq b, y(a) = \alpha$$

está bien planteado.

2. Métodos de un paso

Suponiendo que el problema de valor inicial 9.1 estébien planteado, aproximaremos su solución en el tiempo T dividiendo el intervalo $[t_0, T]$ en pequeños subintervalos y reemplazando la derivada temporal sobre cada subintervalo por un cociente de diferencias finitas.

Sea $t_0, t_1, ..., T_N$ los puntos finales de los subintervalos (llamados nodos o puntos de malla), donde $t_N = T$, y sea la solución aproximada en el tiempo t_j denotada como y_j. Un método de un paso es aquel en el que la solución aproximada en el tiempo t_{k+1} se determina a partir de la solución en el tiempo t_k

2.1. Método de Euler

El método de Euler es la técnica de aproximación más básica para resolver problemas de valor inicial.

El objetivo del método de Euler es obtener aproximaciones para el problema de valor inicial bien planteado

$$\frac{dy}{dt} = f(t, y), \quad a \leq t \leq b, \quad y(a) = \alpha \tag{9.5}$$

No se obtendrá una aproximación continua a la solución $y(t)$; en su lugarm las aproximaciones para y se generarán en varios valores, llamados puntos de malla, en el intervalo $[a, b]$. Una vez que se obtiene la solución aproximada en los puntos, la solución aproximada en otros puntos en el intervalo se puede encontrar a través de interpolación.

Primero estipulamos que los puntos de malla estén igualmente expaciados a lo largo del intervalo $[a, b]$. Esta condición se garantiza al seleccionar un entero positivo N, al establecer $h = (b - a)/N$, y seleccionar los puntos de malla

$$t_i = a + ih, \quad \forall i = 0, 1, 2, ..., N$$

La distancia común entre los puntos $h = t_{i+1} - t_i$ recibe el nombre de tamaño de paso.

Usaremos el teorema de Taylor para deducir el método de Euler. Suponga que $y(t)$, la única solución para 9.5, tiene dos derivadas continuas en $[a, b]$, de tal forma que cada $i = 0, 1, 2, ..., N - 1$

$$y(t_{i+1}) = y(t_i) + (t_{i+1} - t_i)y'(t_i) + \frac{(t_{i+1} - t_i)^2}{2}y''(\xi_i)$$

para algún número ξ_i en (t_i, t_{i+1}). Puesto que $h = t_{i+1} - t_i$, tenemos

$$y(t_{i+1}) = y(t_i) + hy'(t_i) + \frac{h^2}{2}y''(\xi_i)$$

y ya que $y(t)$ satisface la ecuació diferencial 9.5,

$$y(t_{i+1}) = y(t_i) + hf(t_i, y(t_i)) + \frac{h^2}{2}y''(\xi_i) \tag{9.6}$$

El método de Euler construye $w_i \approx y(t_i)$, para cada $i = 1, 2, ..., N$, al borrar el término restante. Por lo tanto, el método de Euler es

$$w_0 = \alpha$$

$$w_{i+1} = w_i + hf(t_i, w_i), \quad \forall i = 0, 1, ..., N - 1 \tag{9.7}$$

La ecuación 9.7 recibe el nombre de ecuación de diferencia relacionada con el método de Euler.

Cotas del error para el método de Euler

A pesar de que el método de Euler no es por completo apropiado para garantizar su uso en la práctica, es suficientemente básico para analizar el error producido a partir de esta aplicación.

Para derivar una cota del error para el método de Euler, necesitamos dos lemas de cálculo.

Lema 2.1. Para toda $x \geq -1$ y cualquier m positiva, tenemos $0 \leq (1+x)^m \leq e^{mx}$

Lema 2.2. Si s y t son números reales positivos, $a_{i\,i=0}^{k}$ es una sucesión que satisface $a_0 \geq -t/s$, y

$$a_{i+1} \leq (1+s)a_i + t, \quad \forall i = 0, 1, 2, ..., k-1 \tag{9.8}$$

entonces

$$a_{i+1} \leq e^{(i+1)s}\left(a_0 + \frac{t}{s}\right) - \frac{t}{s}$$

Teorema 2.1. Suponga que f es continua y satisface la condición de Lipschitz con constante L en

$$D = (t, y)|a \leq t \leq b \text{ y } -\infty < y < \infty$$

y que existe una constante M con

$$|y''(t)| \leq M, \quad \forall t \in [a, b]$$

donde $y(t)$ denota la única solución para el problema de valor inicial

$$y' = f(t, y), \quad a \leq t \leq b, \quad y(a) = \alpha$$

Sean $w_0, w_1, ..., w_N$ las aproximaciones generadas por el método de Euler para un entero positivo N. Entonces, para cada $i = 0, 1, 2, ..., N$

$$|y(t_i) - w_i| \leq \frac{hM}{2L}\left[e^{L(t_i - a)} - 1\right] \tag{9.9}$$

La principal importancia de la fórmula de la cota de error determinada en el teorema 2.1 es que la cota depende linealmente del tamaño de paso h. Por consiguiente, disminuir el tamaño de paso debería proporcionar mayor precisión para las aproximaciones en la misma medida.

Olvidado en el resultado del teorema 2.1 está el efecto que el error de redondeo representa en la selección del tamaño de paso. Conforme h se vuelve más pequeño, se necesitan más cálculos y se espera más error de redondeo. Entonces, en la actualidad, la forma de la ecuación de diferencia

$$w_0 = \alpha$$

$$w_{i+1} = w_i + hf(t_i, w_i), \quad \forall i = 0, 1, ..., N-1$$

no se utiliza para calcular la aproximación a la solución y_i en un punto de malla t_i. En su lugar, usamos una ecuación de la forma

$$u_0 = \alpha + \delta_0$$

$$u_{i+1} = u_i + hf(t_i, y_i) + \delta_{i+1}, \quad \forall i = 0, 1, ..., N-1 \tag{9.10}$$

donde δ_i denota el error de redondeo asociado con u_i.

Teorema 2.2. Si $y(t)$ denota la única solución para el problema de valor inicial

$$y' = f(t, y), \quad a \leq t \leq b, \quad y(a) = \alpha \tag{9.11}$$

y $u_0, u_1, ..., u_N$ son las aproximaciones obtenidas de la ecuación 9.10. Si $|\delta_i| < \delta$ para cada $i = 0, 1, ..., N$ y la hipótesis del teorema 2.1 son aplicables a la ecuación 9.11, entonces

$$|y(t_i - u_i)| \leq \frac{1}{L}\left(\frac{hM}{2} + \frac{\delta}{h}\right)\left[e^{L(t_i - a)} - 1\right] + |\delta_0|e^{L(t_i - a)} \tag{9.12}$$

para cada $i = 0, 1, ..., N$.

La cota de error 9.12 ya no es lineal en h. De hecho, puesto que

$$\lim_{h \to 0}\left(\frac{hM}{2} + \frac{\delta}{h}\right) = \infty$$

se esperaría que el error se vuelva más grande para los valores suficientemente pequeños de h. El cálculo se puede usar para determinar una cota inferior para el tamaño de paso h. Si $E(h) = (hM/2) + (\delta/h)$ implica que $E'(h) = (M/2) - (\delta/h^2)$:

- Si $h < \sqrt{2\delta/M}$, entonces $E'(h) < 0$ y $E(h)$ disminuye.
- Si $h > \sqrt{2\delta/M}$, entonces $E'(h) > 0$ y $E(h)$ aumenta.

El valor mínimo de $E(h)$ se presenta cuando

$$h = \sqrt{\frac{2\delta}{M}} \tag{9.13}$$

La disminución de h más allá de este valor tiende a incrementar el error total en la aproximación. Por lo general, sin embargo, el valor de δ es suficientemente pequeño para que esta cota inferior para h no afecte la operación del método de Euler.

2.2. Método del medio punto

El método del punto medio se define tomando un medio paso con el método de Euler para aproximar la solución en el tiempo $t_{k+1/2} \equiv (t_k + t_{k+1})/2$, y luego tomando un paso completo utilizando el valor de la solución en $t_{k+1/2}$ y la solución aproximada $y_{k+1/2}$:

$$y_{k+1/2} = y_k + \frac{h}{2}f(t_k, y_k) \tag{9.14}$$

$$y_{k+1} = y_k + hf(t_{k+1/2}, y_{k+1/2}) \tag{9.15}$$

Para determinar el error de truncamiento local de este método, expandimos la solución exacta en

una serie de Taylor alrededor de $t_{k+1/2} = t_k + h/2$:
$y(t_{k+1}) = y(t_{k+1/2}) + \frac{h}{2} f(t_{k+1/2}, y(t_{k+1/2})) + \frac{(h/2)^2}{2} y''(t_{k+1/2}) + O(h^3)$
$y(t_k = y(t_{k+1/2})) - \frac{h}{2} f(t_{k+1/2}, y(t_{k+1/2})) + \frac{(h/2)^2}{2} y''(t_{k+1/2}) + O(h^3)$

Restando estas dos ecuaciones se obtiene

$$y(t_{k+1}) - y(t_k) = hf(t_{k+1/2}, y(t_{k+1/2})) + O(h^3)$$

Ahora, expandiendo $y_(t_{k+1/2})$ alrededor de t_k se obtiene:

$$y(t_{k+1/2}) = y(t_k) + \frac{h}{2} f(t_k, y(t_k)) + O(h^2)$$

Y al hacer esta sustitución, tenemos:

$$\begin{aligned} &y(t_{k+1}) - y(t_k) = \\ &= hf(t_{k+1/2}, y(t_k) + \frac{h}{2} f(t_k, y(t_k)) + O(h^2)) + O(h^3) = \\ &hf(t_{k+1/2}, y(t_k) + \frac{h}{2} f(t_k, y(t_k))) + O(h^3) \quad (9.16) \end{aligned}$$

donde la segunda igualdad se sigue porque f es Lipschitz en su segundo argumento, de modo que $hf(t, y + O(h^2)) = hf(t, y) + O(h^3)$. Dado que a partir de 9.14 y 9.15, la solución aproximada satiface:

$$y_{k+1} = y_k + hf(t_{k+1/2}, y_k + \frac{h}{2} f(t_k, y_k))$$

o, de manera equivalente

$$\frac{y_{k+1} - y_k}{h} = f(t_{k+1/2}, y_k + \frac{h}{2} f(t_k, y_k))$$

y, a partir de 9.16, la solución exacta satisface: $\frac{y(t_{k+1}) - y(t_k)}{h} = f(t_{k+1/2}, y(t_k) + \frac{h}{2} f(t_k, y(t_k))) + O(h^2)$ el error de truncamiento local en el método del punto medio es $O(h^2)$; es decir, el método es de precisión de segundo orden.

2.3. Métodos basados en fórmulas de cuadratura

Integrando la ecuación diferencial 9.1 de t a $t+h$ obtenemos

$$y(t+h) = y(t) + \int_t^{t+h} f(s, y(s)) ds \qquad (9.17)$$

La integral al lado derecho de la ecuación puede aproximarse con cualquier fórmula de cuadratura vista anteriormente. Por ejemplo, usando la regla del trapecio para aproximar la integral,

$$\int_t^{t+h} f(s, y(s)) ds = \frac{h}{2} [f(t, y(t)) + f(t+h, y(t+h))] + O(h^3)$$

conduce al método del trapecio para resolver el PVI 9.1,

$$y_{k+1} = y_k + \frac{h}{2} [f(t_k, y_k) + f(t_{k+1}, y_{k+1})] \qquad (9.18)$$

Esto se llama un método implícito, ya que el nuevo valor y_{k+1} aparece en ambos lados de la ecuación. Para determinar y_{k+1}, se debe resolver una ecuación no lineal. Dado que el error en la aproximación de la regla del trapecio para la integral es $O(h^3)$, el error de truncamiento local para este método es $O(h^2)$.

Para evitar resolver la ecuación no lineal en el método del trapecio, se puede utilizar el método de Heun, que primero estima y_{k+1} utilizando el método de Euler y luego usa esta estimación en el lado derecho de 9.18:

$$\tilde{y}_{k+1} = y_k + hf(t_k, y_k)$$
$$y_{k+1} = y_k + \frac{h}{2} [f(t_k, y_k) + f(t_{k+1}, \tilde{y}_{k+1})]$$

El método de Heun sigue una línea cuya pendiente es el promedio de la pendiente de la curva solución en (t_k, y_k) y la pendiente de la curva solución en $(t_{k+1}, \tilde{y}_{k+1})$, donde $\tilde{y}_{k+1}$ es el resultado de un paso con el método de Euler.

2.4. Método de Runge - Kutta

Los métodos Runge-Kutta tienen el error de truncamiento local de orden superior a los métodos de Taylor, pero eliminan la necesidad de calcular y evaluar las derivadas de $f(t, y)$.

Teorema 2.3. Suponga que $f(t, y)$ y todas sus derivadas parciales de orden menor o igual a $n+1$ son continuas en $D = (t, y) | a \le t \le b, c \le y \le d$ y si $(t_0, y_0) \in D$. Para cada $(t, y) \in D$, existe ξ entre t y t_0 y μ entre y y y_0 con

$$f(t, y) = P_n(t, y) + R_n(t, y)$$

donde

$$\begin{aligned} P_n(t,y) = f(t_0,y_0) &+ \left[(t-t_0) \frac{\partial f}{\partial t}(t_0,y_0) + (y-y_0) \frac{\partial f}{\partial y}(t_0,y_0) \right] \\ &+ \left[\frac{(t-t_0)^2}{2} \frac{\partial^2 f}{\partial t^2}(t_0,y_0) + (t-t_0)(y-y_0) \frac{\partial^2 f}{\partial t \partial y}(t_0,y_0) \right. \\ &\left. + \frac{(y-y_0)^2}{2} \frac{\partial^2 f}{\partial y^2}(t_0,y_0) \right] + \dots \\ &+ \frac{1}{n!} \sum_{j=0}^{n} \binom{n}{j} (t-t_0)^{n-j} (y-y_0)^j \frac{\partial^n f}{\partial t^{n-j} \partial y^j}(t_0,y_0). \end{aligned}$$

y

$$R_n(t,y) = \frac{1}{(n+1)!} \sum_{j=0}^{n+1} \binom{n+1}{j} (t-t_0)^{n+1-j} (y-y_0)^j \frac{\partial^{n+1} f}{\partial t^{n+1-j} \partial y^j}(\xi, \mu)$$

La función $P_n(t, y)$ recibe el nombre del emésimo polinomio de Taylor en dos variables para la función f cerca de (t_0, y_0), y $R_n(t, y)$ es el término restante asociado con $P_n(t, y)$.

Métodos de Runge-Kutta de orden 2

El primer paso para deducir un método Runge-Kutta es determinar los valores para a_i, α_1, β_1 con la propiedad de que $a_1 f(t`\alpha_1, y + \beta_1)$ se aproxima a

$$T^{(2)}(t,y) = f(t,y) + \frac{h}{2} f'(t,y)$$

con error no mayor a $O(h^2)$, que es igual al orden del error de truncamiento local para el método de Taylor de orden 2. Ya que

$$f'(t,y) = \frac{df}{dt}(t,y) = \frac{\partial f}{\partial t}(t,y) + \frac{\partial f}{\partial y}(t,y) \cdot y'(t)$$

$$y'(t) = f(t,y)$$

tenemos

$$T^{(2)}(t,y) = f(t,y) + \frac{h}{2}\frac{\partial f}{\partial t}(t,y) + \frac{h}{2}\frac{\partial f}{\partial y}(t,y) \cdot f(t,y) \tag{9.19}$$

Al expandir $f(t+\alpha_1, y+\beta_1)$ en su polinomio de Taylor de grado 1, cerca de (t,y) obtenemos

$$\alpha_1 f(t+\alpha_1, y+\beta_1) = \alpha_1 f(t,y) + a_1\alpha_1 \frac{\partial f}{\partial t}(t,y) + a_1\beta_1 \frac{\partial f}{\partial y}(t,y) + a_1 \cdot R_1(t+\alpha_1, y+\beta_1) \tag{9.20}$$

donde

$$R_1(t+\alpha_1, y+\beta_1) = \frac{\alpha_1^2}{2}\frac{\partial^2 f}{\partial t^2}(\xi,\mu) + \alpha_1\beta_1\frac{\partial^2 f}{\partial t \partial y}(\xi,\mu) + \frac{\beta_1^2}{2}\frac{\partial^2 f}{\partial y^2}(\xi,\mu) \tag{9.21}$$

para algunas ξ entre t y $t+\alpha_1$ y μ entre y y $y+\beta_1$.

Al ajustar los coeficientes de ff y sus derivadas en las ecuaciones 9.19 y 9.20 obtenemos las tres ecuaciones

$$f(t,y) : a_1 = 1$$

$$\frac{\partial f}{\partial t}(t,y) : a_1\alpha_1 = \frac{h}{2}$$

$$\frac{\partial f}{\partial y}(t,y) : a_1\beta_1 = \frac{h}{2} f(t,y)$$

Los parámetos son por tanto: $a_1 = 1$, $\alpha_1 = \frac{h}{2}$ y $\beta_1 = \frac{h}{2} f(t,y)$; por lo que $T^{(2)}(t,y) = f\left(t+\frac{h}{2}, y+\frac{h}{2}f(t,y)\right) - R_1\left(t+\frac{h}{2}, y+\frac{h}{2}f(t,y)\right)$

y, a partir de la ecuación 9.21

$$R_1\left(t+\frac{h}{2}, y+\frac{h}{2}f(t,y)\right) = \frac{h^2}{8}\frac{\partial^2 f}{\partial t^2}(\xi,\mu) + \frac{h^2}{4}f(t,y)\frac{\partial^2 f}{\partial t \partial y}(\xi,\mu) + \frac{h^2}{8}(f(t,y))^2\frac{\partial^2 f}{\partial y^2}(\xi,\mu)$$

Si todas las derivadas parciales de segundo orden de f están acotadas, entonces

$$R_1\left(t+\frac{h}{2}, y+\frac{h}{2}f(t,y)\right)$$

es $O(h^2)$. En consecuencia, el orden de error para este nuevo método es igual al del método de Taylor de orden 2.

El método de ecuación de diferencia que resulta de reemplazar $T^{(2)}(t,y)$ en el método de Taylor de orden 2 por $f(t+(h/2), y+(h/2)f(t,y))$ es un método Runge-Kutta específico, conocido como método de punto medio.

Método de punto medio

$$w_0 = \alpha$$

$$w_{i+1} = w_i + hf(t_i, w_i), \quad \forall i = 0, 1, ..., N-1$$

Solamente se encuentran tres parámetros en $a_1 f(t+\alpha_1, y+\beta_1)$, y todos son necesarios para ajustar $T^{(2)}$. Por lo que se requiere una forma más complicada para satisfaces las condiciones para cualquier de los métodos de Taylor de orden superior.

La forma de cuatro parámetros más adecuada para aproximar

$$T^{(3)}(t,y) = f(t,y) + \frac{h}{2}f'(t,y) + \frac{h}{6}f''(t,y)$$

es

$$a_1 f(t,y) + a_2 f(t+\alpha_2, y+\delta_2 f(t,y)) \tag{9.22}$$

e incluso con esto, no hay suficiente flexibilidad para ajustar el término

$$\frac{h^2}{6}\left[\frac{\partial f}{\partial y}(t,y)\right]^2 f(t,y)$$

lo cual resulta en la expansión de $(h^2/6)f''(t,y)$. Por consiguiente, lo mejor que se puede obtener al usar 9.22 son métodos con error de truncamiento local $O(h^2)$.

Sin embargo, el hecho de que 9.22 tenga cuatro parámetos proporciona una flexibilidad en su elección, por lo que se puede derivar una serie de métodos $O(h^2)$. Uno de los más importantes es el método modificado de Euler, que corresponde a seleccionar $a_1 = a_2 = \frac{1}{2}$ y $\alpha_2 = \delta_2 = h$. Éste tiene la siguiente forma de ecuación de diferencia.

Méotodo modificado de Euler

$$w_0 = \alpha$$

$$w_{i+1} = w_i + \frac{h}{2}\left[f(t_i, w_i) + f(t_{i+1}, w_i + hf(t_i, w_i))\right]$$

$$\forall i = 0, 1, ..., N-1$$

Métodos de Runge-Kutta de orden superior

El término $T^{(3)}(t,y)$ se puede aproximar con error $O(h^3)$ mediante una expresión de la forma

$$f(t+\alpha_1, y+\delta_1 f(t+\alpha 2, y+\delta_2 f(t,y)))$$

relacionada con cuatro parámetros y el álgebra implicada en la determinaciónd de α_1, δ_1, α_2 y δ_2 es bastante tediosa. El método $O(h^3)$ más común es el de Heun, dado por

$$w_0 = \alpha$$

$$w_{i+1} = w_i + \tfrac{h}{4}\left[f(t_i,w_i) + 3\left(f\left(t_i+\tfrac{2h}{3}, w_i+\tfrac{2h}{3}f\left(t_i+\tfrac{h}{3}, w_i+\tfrac{h}{3}f(t_i,w_i)\right)\right)\right)\right]$$

$$\forall i = 0,1,...,N-1$$

En general, los métodos de Runge-Kutta de orden 3 no se usan. El método Runge-Kutta que se usa de manera común es de orden 4 en forma de ecuación de diferencia, dado como sigue.

Runge-Kutta de orden 4

$$w_0 = \alpha$$
$$k_1 = hf(t_i, y_i)$$
$$k_2 = hf\left(t_i + \frac{h}{2}, w_i + \frac{1}{2}k_1\right)$$
$$k_3 = hf\left(t_i + \frac{h}{2}, w_i + \frac{1}{2}k_2\right)$$
$$k_4 = hf(t_{i+1}, w_i + k_3)$$
$$w_{i+1} = w_i + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)$$

para cada $i = 0,1,...,N-1$. Este método tiene error de truncamiento local $O(h^4)$, siempre y cuando la solución $y(t)$ tenga cinco derivadas continuas. Introducimos en el método la notación k_1, k_2, k_3 y k_4 para eliminar la necesidad de anidado sucesivo en la segunda variable de $f(t,y)$.

2.5. Analisis de métodos de un paso

Un método general explícito de un solo paso se puede escribir de la forma:

$$y_{k+1} = y_k + h\psi(t_k, y_k, h) \tag{9.23}$$

- Método de Euler

$$y_{k+1} = y_k + hf(t_k, y_k) \to \psi(t,y,h) = f(t,y)$$

- Método de Heun

$$y_{k+1} = y_k + \frac{h}{2}[f(t_k,y_k) + f(t_{k+1}, y_k + hf(t_k,y_k))] \to$$

$$\psi(t,y,h) = \frac{1}{2}[f(t,y) + f(t+h, y+hf(t,y))]$$

- Método del Medio Punto

$$y_{k+1} = y_k + hf(t_{k+1/2}, y_k + \frac{h}{2}f(t_k,y_k)) \to$$

$$\psi(t,y,h) = f(t+\frac{h}{2}, y+\frac{h}{2}f(t,y))$$

Definición 2.1. El método de un paso 9.23 es consistente si $\lim_{h\to 0}\psi(t,y,h) = f(t,y)$

Definición 2.2. El método de un paso 9.23 es estable si existe una constante K y un tamaño de paso $h_0 > 0$ tal que la diferencia entre las dos soluciones y_n y $\tilde{y}_n$ con valores iniciales y_0 y $\tilde{y}_0$, respectivamente, satiface

$$|y_n - \tilde{y}_n| \le K|y_0 - \tilde{y}_0|$$

para $h \le h_0$ y $nh \le T - t_0$

Definición 2.3. El error de truncamiento es

$$\tau(t,h) = \frac{y(t+h) - y(t)}{h} - \psi(t, y(t), h)$$

Hemos visto que para el método de Euler, el error local de truncamiento es $O(h)$ y para el método de Heun y el método del medio punto el error local de truncamiento es $O(h^2)$

Teorema 2.4. Si un método de un paso de la forma 9.23 es estable y consistente y $|\tau(t,h)| \le Ch^p$, entonces el error global está acotado por

$$\max_{k:t_0+kh\le T}|y_k - y(t_k)| \le Ch^p\frac{e^{L(T-t_0)}-1}{L} + e^{L(T-t_0)}|y_0 - y(t_0)|$$

donde L es la constante de Lipschitz de ψ.

Definición 2.4. El método de un paso 9.23 es convergente si, para todo PVI bien planteado, $\max_{k:t_k\in[t_o,T]}|y(t_k) - y_k| \to 0$ con $y_0 \to y(t_0)$ y $h \to 0$

3. Ecuaciones Stiff

Hasta ahora, nos hemos enfocado en lo que ocurre en el límite cuando $h \to 0$. Para que un método sea útil, ciertamente debe converger a la solución exacta conforme $h \to 0$. Sin embargo, en la práctica, utilizamos un tamaño de paso h fijo y no nulo, o al menos existe un límite inferior para h basado en el tiempo permitido para el cálculo y, posiblemente, en la precisión de la máquina, ya que, por debajo de cierto punto, los errores de redondeo comenzarán a causar inexactitudes. Por lo tanto, nos gustaría

entender cómo se comportan los diferentes métodos con un tamaño de paso h fijo. Esto, por supuesto, dependerá del problema, pero deseamos usar métodos que proporcionen resultados razonables con un tamaño de paso h fijo para la mayor cantidad posible de clases de problemas.

3.1. Estabilidad Absoluta

Para analizar el comportamiento de un método con un tamaño de paso h particular, se podría considerar una ecuación de prueba muy simple:

$$y' = \lambda y \tag{9.24}$$

donde λ es una constante compleja. La solución es $y(t) = e^{\lambda t} y(0)$. Ejemplos de soluciones se muestran en la figura 11.9 para los casos en que la parte real de λ es mayor que 0, igual a 0 y menor que 0. Nótese que $y(t) \to 0$ cuando $t \to \infty$ si y solo si $\mathrm{Re}(\lambda) < 0$, donde $\mathrm{Re}(\cdot)$ denota la parte real.

Definición 3.1. La región de estabilidad absoluta de un método es el conjunto de todos los números $h\lambda \in \mathbb{C}$ tales que $y_k \to 0$ cuando $k \to \infty$, al aplicar el método a la ecuación de prueba 9.24 usando un tamaño de paso h.

Definición 3.2. Un método es A-estable si su región de estabilidad absoluta contiene todo el semiplano izquierdo.

3.2. Métodos de Runge–Kutta Implícitos (IRK)

Otra clase de métodos útiles para ecuaciones rígidas son los métodos de Runge–Kutta implícitos (IRK). Estos tienen la forma:

$$\xi_j = y_k + h \sum_{i=1}^{v} a_{ij} f(t_k + c_i h, \xi_i) \quad j = 1, \ldots, v \tag{9.25}$$

$$y_{k+1} = y_k + h \sum_{j=1}^{v} b_j f(t_k + c_j h, \xi_j), \tag{9.26}$$

donde los valores a_{ij}, b_j y c_j pueden elegirse arbitrariamente, pero para garantizar la consistencia se requiere que:

$$\sum_{i=1}^{v} a_{ji} = c_j, \quad j = 1, ..., v$$

Se puede demostrar que para cada $\nu \geq 1$, existe un único método IRK de orden 2ν, y este es A-estable. El método IRK de orden 2ν se obtiene tomando los valores $c_1, \ldots, c_\nu$ como las raíces del ν-ésimo polinomio ortogonal en $[0, 1]$.

El método IRK de orden 2 ($\nu = 1$) es el método del trapecio 9.18. El método IRK de orden 4 ($\nu = 2$) tiene $c_1 = 1/2 - \sqrt{3}/6$ y $c_2 = 1/2 + \sqrt{3}/6$ como las raíces del segundo polinomio ortogonal en $[0, 1]$, y los demás parámetros se derivan para garantizar la precisión de cuarto orden.