Numerical Analysis

Daming Li

Department of Mathematics, Shanghai JiaoTong University,
Shanghai, 200240, China
Email: lidaming@sjtu.edu.cn

October 14, 2014

Basic Concepts

- Limit : $\lim_{x \to c} f(x) = L$
- Continuity : $\lim_{x \to c} f(x) = f(c)$
- Derivative : $f'(c) = \lim_{x \to c} \frac{f(x)-f(c)}{x-c}$
- Differentiable in R or (a, b), $[a, b]$
- $C(R)$, $C^1(R)$, $C^n(R)$:

$$C^\infty(R) \subset C^2(R) \subset C^1(R) \subset C(R)$$

The inclusion is a proper one. $f(x) = |x|$.

Taylor expansion with Lagrange's formula for remainder

If $f \in C^n[a,b]$ and if $f^{(n+1)}$ exists on (a,b), then for any points c and x in $[a,b]$ there exists a point ξ between c and x such that

$$f(x) = \sum_{k=0}^{n} \frac{1}{k!} f^{(k)}(c)(x-c)^k + E_n(x)$$

where

$$E_n(x) = \frac{1}{(n+1)!} f^{(n+1)}(\xi)(x-c)^{n+1}$$

Mean-value theorem: $f(x) = f(c) + f'(\xi)(x-c)$. A special case: $f(a) = f(b) = 0$, there exists $\xi \in (a,b)$ such that $f'(\xi) = 0$. It called Rolle's theorem.

Taylor expansion with integral form of the remainder

If $f \in C^{n+1}[a,b]$, then for any points c and x in $[a,b]$ there exists a point ξ between c and x such that

$$f(x) = \sum_{k=0}^{n} \frac{1}{k!} f^{(k)}(c)(x-c)^k + R_n(x)$$

where

$$R_n(x) = \frac{1}{n!} \int_c^x f^{(n+1)}(t)(x-t)^n dt$$

Taylor expansion for vector-valued functions of vectors

For a function $f : R^2 \to R$, the simplest expression of Taylor's formula is a symbolic one:

$$f(a+h, b+k) = \sum_{i=0}^{n} \frac{1}{i!}\left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\right)^i f(a,b) + E_n(h,k)$$

$$E_n(h,k) = \frac{1}{(n+1)!}\left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\right)^{n+1} f(a+\theta h, b+\theta k), \quad \theta \in (0,1)$$

$$\left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\right)^0 f(a,b) = f(a,b)$$

$$\left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\right)^1 f(a,b) = \left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\right) f(a,b)$$

$$\left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\right)^2 f(a,b) = \left(h^2\frac{\partial^2 f}{\partial x^2} + 2hk\frac{\partial^2 f}{\partial x \partial y} + k^2\frac{\partial^2 f}{\partial y^2}\right) f(a,b)$$

Rounding

Rounding is an important concept in scientific computing. Consider a positive decimal number x of the form $0.\cdots$ using m digits to the right of the decimal point. One rounds x to n decimal places $(n < m)$ in a manner that depends on the value of the $(n+1)$st digit. If this digit is a 0, 1,2, 3, or 4, then the nth digit is not changed and all following digits are discarded. If it is a 5, 6, 7, 8, or 9, then the nth digit is increased by one unit and the remaining digits are discarded.

$$.1735499 \rightarrow .1735$$

$$.9999500 \rightarrow 1.000$$

Rounding

If x is rounded so that $\tilde{x}$ is the n-digit approximation to it, then

$$|x - \tilde{x}| \leq \frac{1}{2} \times 10^{-n}$$

If the $(n+1)$st digit of x is 0, 1, 2, 3, or 4, then $x = \tilde{x} + \epsilon$ with $0 \leq \epsilon < \frac{1}{2} \times 10^{-n}$. If it is 5, 6, 7, 8, or 9, then $\tilde{x} = \hat{x} + 10^{-n}$ where $\hat{x}$ is a number with the same n digits as x and all digits beyond the nth are 0. Now $x = \hat{x} + \delta \times 10^{-n}$ with $1 > \delta \geq \frac{1}{2}$ and $\tilde{x} - x = (1 - \delta) \times 10^{-n}$. Since $1 - \delta \leq \frac{1}{2}$, we proves our result.
If x is a decimal number, the chopped or truncated n-digit approximation to it is the number $\hat{x}$ obtained by simply discarding all digits beyond the n-th. For it we have

$$|x - \hat{x}| < 10^{-n}$$

Let $\{x_n\}$ and $\{\alpha_n\}$ be two different sequences. We write $x_n = O(\alpha_n)$ (x_n is "big oh" of α_n) if $|x_n| \leq C|\alpha_n|$ for sufficient large n. If $\alpha_n \neq 0$ for all n, this means that the ratio x_n/α_n remains bounded (by C) as $n \to \infty$. The equation $x_n = o(\alpha_n)$ (x_n is "little oh" of α_n) means that $\lim_{n\to\infty}(x_n/\alpha_n) = 0$.

These two notions give a coarse method of comparing two sequences. They are frequently used when both sequences converge to zero. If $x_n \to 0$, $\alpha_n \to 0$, and $x_n = O(\alpha_n)$, then x_n converges to zero at least as rapidly as α_n. If $x_n = o(\alpha_n)$, then x_n converges to 0 more rapidly than α_n. Here are some examples:

$$\frac{n+1}{n^2} = O\left(\frac{1}{n}\right), \quad \frac{1}{n \ln n} = o\left(\frac{1}{n}\right),$$

$$\ln 2 - \sum_{k=1}^{n-1} (-1)^{k-1} \frac{1}{k} = O\left(\frac{1}{n}\right) \quad \textit{slow convergence}$$

$$e^x - \sum_{k=0}^{n-1} \frac{1}{k!} x^k = O\left(\frac{1}{n!}\right) \quad \textit{fast convergence}$$

The notation just introduced is used also for functions other than sequences.

$$\sin x = x - \frac{x^3}{6} + O(x^5), \quad x \to 0$$

This means that there exists a neighborhood of 0 and a constant C such that on that neighborhood

$$\left| \sin x - x + \frac{x^3}{6} \right| \leq C|x^5|$$

$f(x) = O(g(x))(x \to \infty)$ means that $|f(x)| \leq C|g(x)|$ for sufficient large $|x|$. In general we write $f(x) = O(g(x))(x \to x_0)$. Similarly, $f(x) = o(g(x))(x \to x_0)$.

Let $\{x_n\}$ be a sequence of real numbers tending to a limit x^*, We say that the rate of convergence is at least linear if there are a constant $c < 1$ and an integer N such that

$$|x_{n+1} - x^*| \leq c|x_n - x^*|, \quad (n \geq N)$$

We say that the rate of convergence is at least superlinear if there exist a sequence ϵ_n tending to 0 and an integer N such that

$$|x_{n+1} - x^*| \leq \epsilon_n|x_n - x^*|, \quad (n \geq N)$$

The convergence is at least quadratic if there are a constant C (not necessarily less than 1) and an integer N such that

$$|x_{n+1} - x^*| \leq C|x_n - x^*|^2, \quad (n \geq N)$$

In general, if there are constants C and a and an integer N such that

$$|x_{n+1} - x^*| \leq C|x_n - x^*|^\alpha, \quad (n \geq N)$$

we say that the rate of convergence is of order α, at least.

Mean-Value Theorem for Integrals

Let u and v be continuous real-valued functions on an interval $[a, b]$, and suppose that $v \geq 0$. Then there exists a point ξ in $[a, b]$ such that

$$\int_a^b u(x)v(x)dx = u(\xi) \int_a^b v(x)dx$$

Nested Multiplication

The polynomial

$$p(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots a_0$$

To evaluate the polynomial efficiently, we can group the terms using nested multiplication:

$$p(x) = a_0 + x(a_1 + x(a_2 + \cdots + x(a_{n-1} + x(a_n)))\cdots)$$

This corresponds to the following simple algorithm involving only n multiplications and n additions:

- $p \leftarrow a_n$
- for $k = n - 1, \cdots, 0$ do
- $\qquad p \leftarrow xp + a_k$
- endfor

This procedure is also know as Horner's method or synthetic division.

Upper and Lower Bounds

Two important concepts that arise frequently in numerical analysis are supremum (least upper bound) and infimum (greatest lower bound).

Let $S = \{x : x^2 < 2\}$, $\sup S = \sqrt{2}$ and $\inf S = -\sqrt{2}$.

$\sup_{0<x<\pi/6} \sin x = \frac{1}{2}$.

Let G be a function of two real variables defined and continuously differentiable in a neighborhood of (x_0, y_0). If $G(x_0, y_0) = 0$ and $\partial G/\partial y \neq 0$ at (x_0, y_0), then there is a positive δ and a continuously differentiable function f defined for $|x - x_0| < \delta$ such that $f(x_0) = y_0$ and $G(x, f(x)) = 0$ for $|x - x_0| < \delta$.
Does the equation $x^7 + 2y^8 - y^3 = 0$ define y as a continuously differentiable function of x in some neighborhood of $x = -1$?
To answer this, let $(x_0, y_0) = (-1, 1)$, and put $G(x, y) = x^7 + 2y^8 - y^3$. Then $G(x_0, y_0) = 0$ and $\partial G/\partial y = 16y^7 - 3y^2$ which is nonzero if $(x, y) = (x_0, y_0)$. By implicit function theorem, y is a continuously differentiable function of x near $x_0 = -1$.

Implicit function

If $f(x)$ is a function defined implicitly by $G(x, y) = 0$ then for x in some interval, $G(x, f(x)) = 0$.

$$\frac{\partial G}{\partial x} + \frac{\partial G}{\partial y}\frac{dy}{dx} = 0$$

from which we obtain

$$\frac{dy}{dx} = f'(x) = -\frac{\partial G}{\partial x}\Big/\frac{\partial G}{\partial y}$$

Exercises: P.17

6,7,8,2,12,13,16,22

Let V be the set of all infinite sequences of complex numbers, such as

$$x = [x_1, x_2, x_3, \cdots], \quad y = [y_1, y_2, y_3, \cdots]$$

We define two operations:

$$x + y = [x_1 + y_1, x_2 + y_2, x_3 + y_3, \cdots]$$

$$\lambda x = [\lambda x_1, \lambda x_2, \lambda x_3, \cdots]$$

A more compact way of writing these equations is

$$(x + y)_n = x_n + y_n, \quad (\lambda x)_n = \lambda x_n$$

A zero element in V is $0 = [0, 0, 0, \cdots]$. Define the displacement operator, denoted by E, from $V \to V$, defined by

$$Ex = [x_2, x_3, x_4, \cdots], \quad x = [x_1, x_2, x_3, \cdots]$$

Thus $(Ex)_n = x_{n+1}$. The power E^k of E is thus given by $(E^k)_n = x_{n+k}$.

We are particularly interested in the linear combinations of powers of E

$$L = \sum_{i=0}^{m} c_i E^i$$

where E^0 is the identify operator. The polynomial in E of L

$$p(\lambda) = \sum_{i=0}^{m} c_i \lambda^i$$

is called the characteristic polynomial of L. The null space of L is defined by the set

$$\{x \in V : Lx = 0\}$$

Difference Equations

Taking $c_0 = 2$, $c_1 = -3$, $c_2 = 1$, and all other $c_i = 0$. The resulting equation, which is known as a difference equation, can be written in three forms:

$$(E^2 - 3E^1 + 2E^0)x = 0$$

$$x_{n+2} - 3x_{n+1} + 2x_n = 0 \quad (n = 1, 2, 3....)$$

$$p(E)x = 0 \quad p(\lambda) == \lambda^2 - 3\lambda + 2$$

Choose x_1 and x_2 arbitrarily and then determine $x_3, x_4, \cdots$ by the above equality. For instance, we can obtain in this way the various solutions

$$[1, 0, -2, -6, -14, -30, \cdots]$$

$$[1, 1, 1, 1, \cdots]$$

$$[2.4, 8, 16, \cdots]$$

Putting $x_n = \lambda^n$ in the recursive relation,

$$\lambda^{n+2} - 3\lambda^{n+1} + 2\lambda^n = 0$$

$$\lambda^n(\lambda - 1)(\lambda - 2) = 0$$

The trivial solution is $[0, 0, 0, \cdots]$. The other basis solution is $u_n = 1$ and $v_n = 2^n$. We seek a solution of $x = \alpha u + \beta v$ where α and β have not determined yet. Since we have $x_n = \alpha u_n + \beta v_n$, we have

$$\begin{cases} x_1 = \alpha + 2\beta \\ x_2 = \alpha + 4\beta \end{cases}$$

which determined α and β uniquely.

Difference Equations

If p is a polynomial and λ is a zero of p, then one solution of the difference equation $p(E)x = 0$ is $[\lambda_1, \lambda_2, \lambda_3, \cdots]$. If all the zeros of p are simple and nonzero, then each solution of the difference equation is a linear combination of such special solutions.

Proof First, if λ is any complex number and $u = [\lambda_1, \lambda_2, \lambda_3, \cdots]$, then $Eu = \lambda u$

$$p(E)u = \Big(\sum_{i=0}^{m} c_i E^i\Big)u = \sum_{i=0}^{m} c_i(E^i u) = \sum_{i=0}^{m} c_i \lambda^i u = p(\lambda)u = 0$$

Let p be a polynomial all of whose zeros, $\lambda_1, \lambda_2, \cdots, \lambda_m$, are simple and nonzero. Corresponding to any zero λ_k there is a solution of the difference equation $p(E)x = 0$; namely, we have the solution $u^{(k)} = [\lambda_k, \lambda_k^2, \lambda_k^3, \cdots]$. Let x denote an arbitrary solution of the difference equation. We seek to express x in the form $x = \sum_{k=1}^{m} a_k u^{(k)}$. Taking the first m components of the sequences in this equation, we obtain

$$\sum_{i=1}^{m} b_i \lambda_k^i = 0, \quad or \quad \sum_{i=1}^{m} b_i \lambda_k^{i-1} = 0$$

The $m \times m$ matrix having elements λ_k^i is nonsingular because its singularity would imply a nontrivial equation (This last equation exhibits a polynomial of degree $m - 1$ having m zeros.). We thus determines $a_1, a_2, \cdots, a_m$ uniquely. It remains to be proven that the above equality is valid for all values of i. Put $z = \sum_{k=1}^{m} a_k u^{(k)}$. Then $P(E)z = 0$ or equivalently $\sum_{i=0}^{m} c_i z_{n+i} = 0$ for all n. In other words

$$z_{n+m} = -c_m^{-1}(c_0 z_n + c_1 z_{n+1} + \cdots + c_{m-1} z_{n+m-1}) \quad (n = 1, 2, \cdots)$$

Note that $c_m \neq 0$ because the polynomial p has m distinct zeros and is therefore of degree m. Since $z_i = 0$ for $i = 1, 2, \cdots, m$, we have $z = 0$.

There remains the problem of solving a difference equation $p(E)x = 0$ when p has multiple zeros. Define $x(\lambda) = [\lambda, \lambda^2, \lambda^3, \cdots]$. If p is any polynomial, we have seen that

$$p(E)x(\lambda) = p(\lambda)x(\lambda)$$

A differentiation with respect to λ yields

$$p(E)x'(\lambda) = p'(\lambda)x(\lambda) + p(\lambda)x'(\lambda)$$

If λ is a multiple zero of p, then $p(\lambda) = p'(\lambda) = 0$, the above two equalities show that $x(\lambda)$ and $x'(\lambda)$ are solutions of the difference equation. Thus, a solution is the sequence $x'(\lambda) = [1, 2\lambda, 3\lambda^2, \cdots]$. If $\lambda \neq 0$, it is independent of the solution $x(\lambda)$ because

$$\det \begin{pmatrix} \lambda & \lambda^2 \\ 1 & 2\lambda \end{pmatrix} \neq 0$$

By extending this reasoning, one can prove that if λ is a zero of p having multiplicity k, then the following sequences are solutions of the difference equation $p(E)x = 0$:

$$x(\lambda) = [\lambda, \lambda^2, \lambda^3, \cdots]$$

$$x'(\lambda) = [1, 2\lambda, 3\lambda^2, \cdots]$$

$$x''(\lambda) = [0, 2, 6\lambda, \cdots]$$

Let p be a polynomial satisfying $p(0) \neq 0$. Then a basis for the null space of $p(E)$ is obtained as follows. With each zero λ of p having multiplicity k, associate the k basic solutions $x(\lambda), x'(\lambda), \cdots, x^{(k-1)}(\lambda)$, where $x(\lambda) = [\lambda, \lambda^2, \lambda^3, \cdots]$.

An element $x = [x_1, x_2, \cdots]$ of V is said to be bounded if there is a constant c such that $|x_n| \leq c$ for all n. In other words, $\sup_n |x_n| < \infty$. A difference equation of the form $p(E)x = 0$ is said to be stable if all of its solutions are bounded. We now ask whether there is an easy method of identifying a stable difference equation.

For a polynomial p satisfying $p(0) \neq 0$, these properties are equivalent: (i) The difference equation $p(E)x = 0$ is stable. (ii) All zeros of p satisfy $|z| \leq 1$, and all multiple zeros satisfy $|z| < 1$.
Proof. Assume that (ii) is true of p. Let λ be a zero of p. Then one solution of the corresponding difference equation is $x(\lambda) = [\lambda, \lambda^2, \lambda^3, \cdots]$. Since $|\lambda| \leq 1$, this sequence is bounded. If λ is a multiple zero, then one or more of $x'(\lambda), x''(\lambda), \cdots$ will also be a solution of the difference equation. In this case, $|\lambda| < 1$, by (ii). Since, by elementary calculus (L'Hopital's Rule)

$$\lim_{n\to} n^k \lambda^n = 0 \quad (k = 0, 1, \cdots)$$

we see that each sequence $x'(\lambda), x''(\lambda), \cdots$ is bounded.
For the converse, suppose that (ii) is false. If p has a zero λ satisfying $|\lambda| > 1$, then the sequence $x(\lambda)$ is unbounded. If p has a multiple zero λ satisfying $|\lambda| \geq 1$, then $x'(\lambda)$ is unbounded since its general term satisfies the inequality

$$|x_n| = n|\lambda|^{n-1} \geq n$$

Exercises: P.26

2,3,5,11(a)(b),14,15,16