Numerical Methods for Ordinary Differential Equations

Numerical Methods for Ordinary Differential Equations

C. Vuik P. van Beek F. Vermolen J. van Kan

VSSD

Related titles published by VSSD:
Numerical methods in scientific computing, J. van Kan, A. Segal and F. Vermolen, xii + 279 pp., hardback, ISBN 978-90-71301-50-6
http://www.vssd.nl/hlf/a002.htm

In Dutch:
Numerieke Wiskunde voor Technici, J.J.I.M. van Kan; 128 pp
ISBN 9 78-90-407-1151-0
http://www.vssd.nl/hlf/a002.htm

Numerieke methoden voor differentiaalvergelijkingen, J. van Kan, P. van Beek, F. Vermolen, K. Vuik, x + 122 pp. (Dutch version of this volume)
http://www.vssd.nl/hlf/a018.htm

First edition 2007

Published by VSSD
Leeghwaterstraat 42, 2628 CA Delft, The Netherlands
tel. +31 15 27 82124, telefax +31 15 27 87585, e-mail: hlf@vssd.nl
internet: http://www.vssd.nl/hlf
URL about this book: **http://www.vssd.nl/hlf/a026.htm**

Printed version: ISBN-13 978-90-6562-156-6
Electronic version: ISBN-13 978-90-6562-170-2
NUR 919

Keywords: numerical analysis, ordinary differential equations

Preface

In this book we discuss several numerical methods for solving ordinary differential equations. We emphasize those aspects that play an important role in practical problems. In this introductory text we confine ourselves to ordinary differential equations with the exception of the last chapter in which we discuss the heat equation, a parabolic partial differential equation. The techniques discussed in the introductory chapters, for e.g. interpolation, numerical quadrature and the solution of nonlinear equations, may also be used outside the context of differential equations. They have been included to make the book self contained as far as the numerical aspects are concerned. Chapters, sections and exercises marked * are not part of the Delft Institutional Package.

This text is an English version of a Dutch original “Numerieke Methoden voor Differentiaalvergelijkingen”. I would like to thank Jos van Kan for translating the Dutch text into English and Hisham bin Zubair for correcting the English of this book.

Delft, July 2007
C. Vuik

Contents

1 Introduction

1.1 Some historical remarks

Modern applied mathematics started in the 17th and 18th century with scholars like Stevin, Descartes, Newton and Euler. Numerical aspects found a natural place in the analysis but the expression ”numerical mathematics” did not exist at that time. However, numerical methods invented by Newton, Euler and at a later stage by Gauss still play an important role even today.

In the 17th and the 18th century fundamental laws were formulated for various subdomains of physics, like mechanics and hydrodynamics. These took the form of simple looking mathematical equations. To the disappointment of many, these equations could be solved analytically in a few special cases only. For that reason technological development has been only loosely connected with mathematics. The introduction and availability of the modern digital computer has changed this. Using a computer it is possible to gain quantitative information with detailed and realistic mathematical models and numerical methods for a multitude of phenomena and processes in physics and technology. Application of computers and numerical methods has become ubiquitous. Statistical analysis shows that non-trivial mathematical models and methods are used in 70% of the papers appearing in the professional journals of engineering sciences.

Computations are often cheaper than experiments; experiments can be expensive, dangerous or downright impossible. Real life experiments can often be performed on a small scale only and that makes their results less reliable.

1.2 What is numerical mathematics?

Numerical mathematics is a collection of methods to approximate solutions of mathematical equations numerically by means of *finite* computational processes.

In large parts of mathematics the most important concepts are mappings and sets. In numerical mathematics we have to add the concept of computability. Computability means that the result can be obtained in a finite number of operations (so the computation time will be finite) on a finite subset of the rational numbers (because a computer has only finite memory).

In general the result will be an approximation of the analytic solution of the mathematical problem, since most mathematical equations contain operators based on infinite processes, like integrals and derivatives. Moreover, solutions are functions whose domain and image may (and usually do) contain irrational numbers.

Because, in general, numerical methods can only obtain approximate solutions, it makes sense to apply them only to problems that are insensitive to small perturbations, in other words to problems that are *stable*. The concept of stability belongs to both numerical and classical mathematics. An important instrument in studying stability is functional analysis. This discipline also plays an important role in error analysis: the difference between numerical approximation and exact solution.

Calculating with only a finite subset of the rational numbers has many consequences. For example: a computer cannot distinguish between two polynomials of sufficiently high degree. Consequently we cannot trust methods based on the main theorem of algebra (i.e. that an n-th degree polynomial has exactly n complex roots). Errors that follow from the use of finitely many digits are called *rounding errors*. We shall pay some attention to rounding errors later on in this chapter.

An important aspect of numerical mathematics is the emphasis on efficiency. Contrary to ordinary mathematics, numerical mathematics considers an increase in efficiency, i.e. a decrease of the number of operations and/or amount of storage needed, an essential improvement. Progress in this aspect is of great practical importance and the end of this development has not been reached yet. Here the creative mind will meet many challenges. On top of that, revolutions in computer architecture will overturn much conventional wisdom.

1.3 Why numerical mathematics?

A big advantage of numerical mathematics is that it can provide answers to problems that do not admit analytical solutions. Consider for example the integral

$$\int_0^\pi \sqrt{1+\cos^2 x}\,dx.$$

This is an expression for the arc length of one arc of the curve $y = \sin x$. There is no solution in closed form for this integral. A numerical method, however, can approximate this integral in a very simple way. An additional advantage is, that a numerical method only uses evaluation of standard functions and the operations: addition, subtraction, multiplication and division. Because these are just the operations a computer can perform, numerical mathematics and computers form a perfect combination.

An analytical method gives the solution as a mathematical formula, which is an advantage. From this we can gain insight in the behavior and the properties of the solution, and with a numerical solution (that gives the function as a table) this is not the case. On the other hand some form of visualization may be used to gain insight in the behavior of the solution. To draw a graph of a function with a numerical method is usually a more useful tool than to evaluate the analytical solution at a great number of points.

1.4 Rounding errors

A computer uses a finite representation of real numbers. These are stored in a computer in the form

$$\pm 0.d_1 d_2 \ldots d_n \cdot \beta^e,$$

in which $d_1 > 0$ and $0 \leq d_i < \beta$. We call this a floating point number (representation) in which $0.d_1 d_2 \ldots d_n$ is called the *mantissa*, β the *base* and e (integer) the *exponent*. Often we have $\beta = 2$ (binary representation) and $n = 24$ (*single* precision). In *double* precision we have $n = 56$. We say that the machine computes with n-bit (or n-digit) precision.

Let for $x \in \mathbb{R}$

$$0.d_1 \ldots d_n \cdot \beta^e \leq x < 0.d_1 d_2 \ldots (d_n + 1) \cdot \beta^e,$$

where for simplicity we assume that x is positive. *Rounding* x means, that x will be replaced with the floating point number closest to x, which we shall call $fl(x)$. The error caused by this process is called *rounding error*. Let us write

$$fl(x) = x(1 + \varepsilon). \tag{1.1}$$

We call $|fl(x) - x| = |\varepsilon x|$ the *absolute error* and $\frac{|fl(x)-x|}{|x|} = |\varepsilon|$ the *relative error*. The difference between the floating point numbers enclosing x is β^{e-n}. Rounding gives $|fl(x) - x| \leq \frac{1}{2}\beta^{e-n}$, so for the absolute error we have

$$|\varepsilon x| \leq \frac{1}{2}\beta^{e-n}.$$

Because $|x| \geq \beta^{e-1}$ (since $d_1 > 0$) we have for the relative error:

$$|\varepsilon| \leq eps \tag{1.2}$$

with the computer's relative precision eps defined by

$$eps = \frac{1}{2}\beta^{1-n}. \tag{1.3}$$

From $\beta = 2$ and $n = 24$ it follows that $eps \approx 6 \times 10^{-8}$, so in single precision we calculate with approximately 7 decimal digits.

Figure 1.1 shows the distribution of the floating point numbers $0.1 d_2 d_3 \cdot \beta^e$; $e = -1, 0, 1, 2$ in base 2 (binary numbers). These floating point numbers are not uniformly distributed and there is a neighborhood of 0 that contains no floating point number. A computational result lying within this neighborhood is called *underflow*. Most machines give a warning, replace the result with 0 and continue. A computational result larger than the largest floating point number that can be represented is called *overflow*. The machine warns and halts.

How do computers execute arithmetical operations in floating point arithmetic?

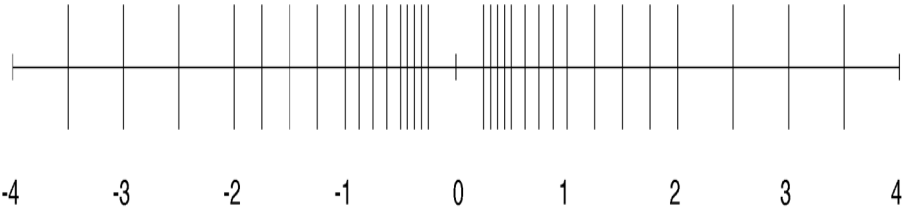

Figure 1.1 Distribution of $\pm 0.1d_2d_3 \cdot \beta^e$, $\beta = 2, e = -1, 0, 1, 2$.

Central processors are very complex and usually the following model is used to simulate reality. Let $\circ$ denote an arithmetic operation ($+, -, \times$ of /) and let x and y be floating point numbers. Then the machine result of the operation $x \circ y$ will be

$$z = fl(x \circ y). \tag{1.4}$$

The exact result of $x \circ y$ will not be a floating point number in general, hence an error results. From formula (1.1) we get

$$z = \{x \circ y\}(1 + \varepsilon), \tag{1.5}$$

for some ε satisfying (1.2) and $z \neq 0$.

Suppose x and y are numbers approximated by the floating point numbers $fl(x)$ and $fl(y)$, so $fl(x) = x(1+\varepsilon_1)$, $fl(y) = y(1+\varepsilon_2)$. We wish to calculate $x \circ y$. The absolute error in the calculated result $fl(fl(x) \circ fl(y))$ satisfies:

$$|x \circ y - fl(fl(x) \circ fl(y))| \leq |x \circ y - fl(x) \circ fl(y)| + |fl(x) \circ fl(y) - fl(fl(x) \circ fl(y))|. \tag{1.6}$$

From this expression we see that the error consists of two terms. The first term is caused by an error in the *data* and the second one by converting the result of an exact calculation to floating point form.

We shall give a few examples to show how rounding errors may affect the result of a calculation. After that we shall give general computational rules regarding the propagation of rounding errors.

Example 1.4.1

Let us take $x = \frac{5}{7}$ and $y = \frac{1}{3}$ and carry out the calculations on a system that uses $\beta = 10$ and a precision of 5 digits. In Table 1.1 you will find the results of various calculations applied to $fl(x) = 0.71429 \times 10^0$ and $fl(y) = 0.33333 \times 10^0$. We shall show how the table has been created. After normalization we find for the addition

$$fl(x) + fl(y) = (.71429 + .33333) \times 10^0 = 0.1047620000... \times 10^1$$

This result has to be rounded to 5 digits:

$$fl(fl(x) + fl(y)) = 0.10476 \times 10^1.$$

Table 1.1 Absolute and relative error for various calculations.

operation	result	exact value	absolute error	relative error
$x+y$	0.10476×10^1	$22/21$	0.190×10^{-4}	0.182×10^{-4}
$x-y$	0.38096×10^0	$8/21$	0.761×10^{-5}	0.200×10^{-4}
$x \times y$	0.23809×10^0	$5/21$	0.523×10^{-5}	0.220×10^{-4}
$x \div y$	0.21429×10^1	$15/7$	0.429×10^{-4}	0.200×10^{-4}

The exact value is $x+y = \frac{22}{21} = 1.0476190518...$ So the absolute error is $1.0476190518... - 0.10476 \times 10^1 \approx 0.190 \times 10^{-4}$ and the relative error is $\frac{0.190 \times 10^{-4}}{22/21} \approx 0.182 \times 10^{-4}$.

The error analysis of the other three operations follows the same lines.

Example 1.4.2

In this example we will use the same numbers x and y and the same precision as in the previous example. Further we use $u = 0.714251$, $v = 98765.1$ and $w = 0.111111 \times 10^{-4}$, so $fl(u) = 0.71425$, $fl(v) = 0.98765 \times 10^5$ and $w = 0.11111 \times 10^{-4}$. These numbers have been chosen in such a way that we can clearly illustrate what problems we may expect with rounding errors. In Table 1.2 $x-u$ has a small absolute error but a large relative error. If we divide $x-u$ by a small number w or multiply it with a large number v, the absolute error increases, whereas the relative error is not affected. On the other hand, adding a larger number u to a small number v results in a large absolute error but only a small relative error. We shall show how the first row has been created. The exact result

Table 1.2 Absolute and relative error for various calculations

operation	result	exact value	absolute error	relative error
$x-u$	0.40000×10^{-4}	0.34714×10^{-4}	0.528×10^{-5}	0.152
$(x-u)/w$	0.36000×10^1	0.31243×10^1	0.476	0.152
$(x-u) \times v$	0.39506×10^1	0.34287×10^1	0.522	0.152
$u+v$	0.98765×10^5	0.98766×10^5	0.814×10^0	0.824×10^{-5}

is $u = 0.714251$ and $x-u = \frac{5}{7} - .714251 = 0.3471428571.... \times 10^{-4}$, whereas $fl(u) = 0.71425 \times 10^0$ and $fl(x) - fl(u) = 0.71429 - 0.71425 = 0.0000400000 \times 10^0$. Normalization gives $fl(fl(x) - fl(u)) = 0.40000 \times 10^{-4}$. From this we obtain the absolute error: $(x-u) - fl(fl(x) - fl(u)) = (.3471428571... - .40000) \times 10^{-4} \approx 0.528 \times 10^{-5}$ and the relative error:
$\frac{0.528... \times 10^{-5}}{0.3471428... \times 10^{-4}} \approx 0.152$.

It is interesting to note, that the large relative error has nothing to do with the limitations of the floating point system (the subtraction of $fl(x)$ and $fl(u)$ is without error in this case) but is due only to the fact that the data is represented in no more than 5 decimal digits. The

zeros that remain after normalization in the single precision result $fl(fl(x) - fl(u)) = 0.40000$ have no significance, only the digit 4 is significant; the zeros that have been substituted are a mere formality and represent no information. This phenomenon is called *loss of significant digits*. The loss of significant digits has a large impact on the relative error, because of division by the small result.

A large relative error sooner or later will have some unpleasant consequences in later stages of the process, also for the absolute error. If we multiply for example $x - u$ by a large number, then we immediately also generate a large absolute error, together with the large relative error we already had. As an example we look at the third row of the table. The exact result is $(x - u) \times v = 3.4285594526000....$ Calculating $fl(fl(x) - fl(u)) \times fl(v)$ gives:

$$fl(fl(x) - fl(u)) \times fl(v) = 0.4 \times 10^{-4} \times 0.98765 \times 10^{5} = 0.3950600000 \times 10^{1}.$$

After rounding we get: $fl(fl(fl(x) - fl(u)) \times fl(v)) = 0.39506 \times 10^{1}$. This yields the absolute error: $3.42855990000460..... - 0.39506 \times 10^{1} \approx 0.522$ and the relative error: $\frac{0.522..}{3.4285....} \approx 0.152$. Suppose we add something to $(x - u) \times v$, for example: y^2; because $y = \frac{1}{3}$ and therefore $y^2 = \frac{1}{9}$, the result of this operation due to the large absolute error is indistinguishable. In other words, for the reliability of the result it does not make a difference whether we would omit the last operation and by doing that alter the numerical process. So we conclude that something is fundamentally wrong in this case.

Almost all numerical processes exhibit loss of significant digits for a certain set of input data; one might call such a set *ill conditioned*. There also are numerical processes that exhibit these phenomena for all possible input data. Such processes are called *unstable*. One of the objectives of numerical analysis is to identify unstable processes and classify them as useless. Or improve them in such a way that they become stable.

Computational rules for error propagation

In the analysis of a complete numerical process, in each subsequent step we have to interpret the accumulated error of all previous steps as a perturbation of the original data. Moreover, in the result of this step we have to take into account the propagation of these perturbations together with the floating point error. After a considerable number of steps this error source will be more important than the floating point error most of the time. (In the previous example of $(x - u) \times v$ even after two steps!) In that stage the error in a numerical process will be largely determined by the 'propagation' of the accumulated errors. The computational rules to calculate numerical error propagation are the same as those to calculate propagation of error in measurements in physical experiments. There are two rules: one for addition and subtraction and one for multiplication and division.

The approximations of x and y will be denoted by $\tilde{x}$ and $\tilde{y}$ and the (absolute) perturbations $\delta x = x - \tilde{x}$ and $\delta y = y - \tilde{y}$.

a) Addition and subtraction.
   $(x + y) - (\tilde{x} + \tilde{y}) = (x - \tilde{x}) + (y - \tilde{y}) = \delta x + \delta y$, in other words, the absolute error in the sum of two perturbed terms is equal to the sum of the absolute perturbations.

A similar rule holds for differences: $(x-y)-(\tilde{x}-\tilde{y}) = \delta x - \delta y$. Often the rule is presented in the form of an inequality (also called an *error estimate*.): $|(x \pm y) - (\tilde{x} \pm \tilde{y})| \leq |\delta x| + |\delta y|$.

b) This rule does not hold for multiplication and division. Efforts to derive a rule for *absolute* error will lead nowhere. But one may derive a similar rule for the *relative* error.

The *relative* perturbations ε_x and ε_y are defined by $\tilde{x} = x(1-\varepsilon_x)$, and similarly for y. For the relative error in a product xy we have: $\frac{xy-\tilde{x}\tilde{y}}{xy} = \frac{xy-x(1-\varepsilon_x)y(1-\varepsilon_y)}{xy} = \varepsilon_x + \varepsilon_y - \varepsilon_x\varepsilon_y \approx \varepsilon_x + \varepsilon_y$, assuming ε_x and ε_y are negligible compared to 1. In words: the relative error in a product of two perturbed factors is approximately equal to the sum of the two relative perturbations. A similar rule can be derived for division. Formulated as an error estimate we have $|\frac{xy-\tilde{x}\tilde{y}}{xy}| \leq |\varepsilon_x| + |\varepsilon_y|$.

Identification of $\tilde{x}$ with $fl(x)$ and $\tilde{y}$ with $fl(y)$ enables us to explain clearly various phenomena in floating point computations using these two simple rules.

1.5 Landau's O-symbol

In the analysis of numerical methods estimating the error is of prime importance. It is often more important to have an indication of the order of magnitude of the error than a precise expression. To save ourselves some tedious work we use Landau's O-symbol.

Definition 1.5.1 *Let f and g be given functions. We say $f(x) = O(g(x))$ ("$f(x)$ is big Oh of $g(x)$") for $x \to 0$, if there exist positive r and finite M such that*

$$|f(x)| \leq M|g(x)| \quad \textit{for all} \quad x \in [-r,r] \,.$$

To estimate errors we often use the following computational rules.

Computational rules
If $f(x) = O(x^p)$ and $g(x) = O(x^q)$ for $x \to 0$, with $p \geq 0$ and $q \geq 0$ then

a) $f(x) = O(x^s)$ for all s with $0 \leq s \leq p$.

b) $\alpha f(x) + \beta g(x) = O(x^{\min\{p,q\}})$ for all $\alpha, \beta \in \mathbb{R}$.

c) $f(x)g(x) = O(x^{p+q})$.

d) $\frac{f(x)}{|x|^s} = O(x^{p-s})$ if $0 \leq s \leq p$.

1.6 Some important theorems from analysis

In this section we recollect some important theorems from analysis that are often used in numerical analysis. In this book we use the notation $C[a,b]$ for the set of all functions continuous on the interval $[a,b]$ and $C^p[a,b]$ for the set of all functions of which all derivatives up to the p-th exist and are continuous.

Theorem 1.6.1 (Intermediate value theorem) *Assume $f \in C[a,b]$. Let $f(a) \neq f(b)$ and let F be a number between $f(a)$ and $f(b)$. Then there exists a number $c \in (a,b)$ such that $f(c) = F$.*

Theorem 1.6.2 (Rolle's theorem) *Assume $f \in C[a,b]$ and f differentiable on (a,b). If $f(a) = f(b)$, then there exists a number $c \in (a,b)$ such that $f'(c) = 0$.*

Theorem 1.6.3 (Mean value theorem) *Assume $f \in C[a,b]$ and f differentiable on (a,b), then there exists a number $c \in (a,b)$ such that $f'(c) = \frac{f(b)-f(a)}{b-a}$.*

Theorem 1.6.4 (Taylor polynomial) *Assume $f : (a,b) \to \mathbb{R}$ is $(n+1)$ times differentiable. Then for all $c, x \in (a,b)$ there exists a number ξ between c and x such that*

$$f(x) = P_n(x) + R_n(x),$$

in which the Taylor polynomial $P_n(x)$ *is given by*

$$P_n(x) = f(c) + (x-c)f'(c) + \frac{(x-c)^2}{2!}f''(c) + \ldots + \frac{(x-c)^n}{n!}f^{(n)}(c)$$

and the remainder term $R_n(x)$ *is:*

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

Proof
Take $c, x \in (a,b)$ with $c \neq x$ and let K be defined by:

$$f(x) = f(c) + (x-c)f'(c) + \frac{(x-c)^2}{2!}f''(c) + \ldots + \frac{(x-c)^n}{n!}f^{(n)}(c) + K(x-c)^{n+1}. \quad (1.7)$$

Consider the function

$$F(t) = f(t) - f(x) + (x-t)f'(t) + \frac{(x-t)^2}{2!}f''(t) + \ldots + \frac{(x-t)^n}{n!}f^{(n)}(t) + K(x-t)^{n+1}.$$

By (1.7) we have $F(c) = 0$ and $F(x) = 0$. Hence, by Rolle's theorem there exists a number ξ between c and x such that $F'(\xi) = 0$. Further elaboration gives

$$\begin{aligned} F'(\xi) &= f'(\xi) + \{f''(\xi)(x-\xi) - f'(\xi)\} + \{\frac{f'''(\xi)}{2!}(x-\xi)^2 - f''(\xi)(x-\xi)\} + \\ &\quad + \cdots + \{\frac{f^{(n+1)}(\xi)}{n!}(x-\xi)^n - \frac{f^{(n)}(\xi)}{(n-1)!}(x-\xi)^{(n-1)}\} - K(n+1)(x-\xi)^n = \\ &= \frac{f^{(n+1)}(\xi)}{n!}(x-\xi)^n - K(n+1)(x-\xi)^n = 0. \end{aligned}$$

So $K = \frac{f^{(n+1)}(\xi)}{(n+1)!}$, which proves the theorem. ⊠

Theorem 1.6.5 (Taylor polynomial of two variables) *Let $f : D \subset \mathbb{R}^2 \mapsto \mathbb{R}$ be continuous with continuous partial derivatives up to and including order $n+1$ in a ball $B \subset D$ with center $\mathbf{c} = (c_1, c_2)$ and radius ρ. Then for each $\mathbf{x} = (x_1, x_2) \in B$ there exists a $\theta \in (0,1)$, such that*

$$f(\mathbf{x}) = P_n(\mathbf{x}) + R_n(\mathbf{x}),$$

in which the Taylor polynomial $P_n(\mathbf{x})$ is given by

$$P_n(\mathbf{x}) = f(\mathbf{c}) + (x_1 - c_1)\frac{\partial f}{\partial x_1}(\mathbf{c}) + (x_2 - c_2)\frac{\partial f}{\partial x_2}(\mathbf{c}) + \frac{1}{2}\sum_{i=1}^{2}\sum_{j=1}^{2}(x_i - c_i)(x_j - c_j)\frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{c}) + \ldots$$

$$+\frac{1}{n!}\sum_{i_1=1}^{2}\sum_{i_2=1}^{2}\cdots\sum_{i_n=1}^{2}(x_{i_1} - c_{i_1})(x_{i_2} - c_{i_2})\ldots(x_{i_n} - c_{i_n})\frac{\partial^n f}{\partial x_{i_1}\partial x_{i_2}\ldots\partial x_{i_n}}(\mathbf{c})$$

and the remainder term is

$$R_n(\mathbf{x}) = \frac{1}{(n+1)!}\sum_{i_1=1}^{2}\sum_{i_2=1}^{2}\cdots$$

$$\sum_{i_{n+1}=1}^{2}(x_{i_1} - c_{i_1})(x_{i_2} - c_{i_2})\ldots(x_{i_{n+1}} - c_{i_{n+1}})\frac{\partial^{n+1} f}{\partial x_{i_1}\partial x_{i_2}\ldots\partial x_{i_{n+1}}}(\mathbf{c} + \theta(\mathbf{x} - \mathbf{c}))$$

Proof
Let for fixed $\mathbf{x}$ and $\mathbf{h}$ with $\|\mathbf{h}\| < \rho$, the function $F : (-1,1) \mapsto \mathbb{R}$ be defined by:

$$F(s) = f(\mathbf{x} + s\mathbf{h}).$$

Because of the differentiability conditions satisfied by f in the ball B, F is $(n+1)$ times continuously differentiable on the interval $(-1,1)$ and $F^k(s)$ is given by (check this!)

$$F^k(s) = \sum_{i_1=1}^{2}\sum_{i_2=1}^{2}\cdots\sum_{i_k=1}^{2}\frac{\partial^k f(\mathbf{x} + s\mathbf{h})}{\partial x_{i_1}\partial x_{i_2}\ldots\partial x_{i_k}}h_{i_1}h_{i_2}\ldots h_{i_k}.$$

Expand F into a Taylor polynomial about 0. This yields:

$$F(s) = F(0) + sF'(0) + \ldots + \frac{s^n}{n!}F^n(0) + \frac{s^{n+1}}{(n+1)!}F^{n+1}(\theta s),$$

for some $\theta \in (0,1)$. Now substitute $s = 1$ into this expression and into the expressions for the derivatives of F and the result follows. ⊠

Example
For $n = 1$ we get:

$$P_1(\mathbf{x}) = f(c_1, c_2) + (x_1 - c_1)\frac{\partial f}{\partial x_1}(c_1, c_2) + (x_2 - c_2)\frac{\partial f}{\partial x_2}(c_1, c_2),$$

and for the remainder term: $R_1(\mathbf{x})$ is $O(\|\mathbf{x} - \mathbf{c}\|^2)$.

Theorem 1.6.6 (Power series of $\frac{1}{1-x}$) *Let $x \in \mathbb{R}$ with $|x| < 1$. Then:*

$$\frac{1}{1-x} = \sum_{k=0}^{\infty} x^k.$$

Theorem 1.6.7 (Power series of e^x) *Let $x \in \mathbb{R}$. Then:*

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}.$$

1.7 Summary

In this chapter the following subjects have been discussed

- Numerical mathematics
- Rounding errors
- Landau's O-symbol
- Some important theorems from analysis

1.8 Exercises

1. Let $f(x) = x^3$. Determine the second order Taylor polynomial of f about the point $x = 1$. Compute the value of this polynomial in $x = 0.5$. Give an error estimate and compare this with the actual error.

2. Let $f(x) = e^x$. Give the n-th order Taylor polynomial about the point $x = 0$ and also give the remainder term. How large should n be chosen in order to make the error less than 10^{-6} in the interval $[0, 0.5]$?

3. We use the polynomial $P_2(x) = 1 - \frac{1}{2}x^2$ to approximate $f(x) = \cos(x)$ in the interval $[-\frac{1}{2}, \frac{1}{2}]$. Give an upper bound for the error in this approximation.

4. Let $x = \frac{1}{3}$, $y = \frac{5}{7}$. We calculate with a precision of 3 (decimal) digits. Express x and y as floating point numbers. Compute $fl(fl(x) \circ fl(y))$, $x \circ y$ and the rounding error taking $\circ = +, -, *, /$ respectively.