Contrôle dynamique de méthodes d'approximation

Fabienne Jézéquel
Laboratoire d'Informatique de Paris 6

ARINEWS, ENS Lyon, 7-8 mars 2005

UNIVERSITE PIERRE & MARIE CURIE
LA SCIENCE A PARIS

CENTRE NATIONAL
DE LA RECHERCHE
SCIENTIFIQUE

Numerical accuracy of approximation methods

When an approximation $L(h)$ such that $\lim_{h \to 0} L(h) = L$ is computed, it is affected by:

- a truncation error $e_m(h)$
- a round-off error $e_c(h)$.

$e_m(h) \longrightarrow$

If h decreases, $L(h)$:

s	exponent	mantissa

$\longleftarrow e_c(h)$

As long as $e_c(h) < e_m(h)$, decreasing h brings reliable information to the mantissa.

The optimal step is reached when $e_c(h) \approx e_m(h)$.

1. How to determine dynamically the optimal step ?
2. Which digits in the approximation obtained are in common with L ?

Stochastic approach of round-off errors

- the CESTAC method
- the concept of computational zero

$\Rightarrow$ Continuous stochastic arithmetic: $X = (m, \sigma^2)$

$\Rightarrow$ Discrete stochastic arithmetic: $X = (X_1, X_2, ..., X_N)$

Significant digits common to two real numbers

Definition

Let a and b be two real numbers, the number of significant digits that are common to a and b can be defined in $\mathbb{R}$ by

1. for $a \neq b$, $C_{a,b} = \log_{10} \left| \frac{a+b}{2(a-b)} \right|$,
2. $\forall a \in \mathbb{R},\ C_{a,a} = +\infty$.

Example:

if $a = 2.4599976$ and $b = 2.4600012$, then $C_{a,b} \approx 5.8$.

On sequences with a linear convergence

Theorem

Let (I_n) be a sequence converging linearly to I, i.e. which satisfies $I_n - I = K\alpha^n + o(\alpha^n)$ where $K \in \mathbb{R}$ and $0 < |\alpha| < 1$, then

$$C_{I_n,I_{n+1}} = C_{I_n,I} + \log_{10}\left(\frac{1}{1-\alpha}\right) + o(1).$$

If the convergence zone is reached,
the significant decimal digits common to I_n and I_{n+1}, are those of I, up to $\log_{10}\left(\frac{1}{1-\alpha}\right)$.

If $-1 < \alpha \leq \frac{1}{2}$, then $-1 < \log_2\left(\frac{1}{1-\alpha}\right) \leq 1$.
In this case, the significant bits common to I_n and I_{n+1} are those of I, up to one.

Applications in DSA

Let us assume that the convergence zone is reached.

If $I_n - I_{n+1} = @.0$,
the difference between I_n and I_{n+1} is due to round-off errors.

Further iterations are useless.

Consequently

- the optimal iterate I_{n+1} can be dynamically determined
- if $\alpha \leq \frac{1}{2}$, the exact significant bits of I_{n+1} are those of I, up to one.

F. Jézéquel, *Dynamical control of converging sequences computation*, Applied Numerical Mathematics, 50(2): 147-164, 2004.

Dynamical control of approximation methods

Theorem

Let $L(h)$ be an approximation of order p of L, i.e.

$$L(h) - L = Kh^p + \mathcal{O}\left(h^q\right) \text{ with } 1 \leq p < q,\ K \in \mathbb{R}.$$

If L_n is the approximation computed with the step $\frac{h_0}{2^n}$, then

$$C_{L_n,L_{n+1}} = C_{L_n,L} + \log_{10}\left(\frac{2^p}{2^p - 1}\right) + \mathcal{O}\left(2^{n(p-q)}\right).$$

If the convergence zone is reached and $L_n - L_{n+1} = @.0$, the exact significant bits of L_{n+1} are those of L, up to one.

Control of stochastic operations

Theorem

Let X_i be the approximation in stochastic arithmetic of a mathematical value x_i such that its exact significant bits are those of x_i up to p_i ($i = 1, 2$).

Let $\bigcirc$ be an arithmetical operator: $\bigcirc \in \{+, -, \times, /\}$ and $s\bigcirc$ the corresponding stochastic operator: $s\bigcirc \in \{s+, s-, s\times, s/\}$.

Then the exact significant bits of $X_1 \; s\bigcirc \; X_2$ are those of the mathematical value $x_1 \bigcirc x_2$, up to $\max(p_1, p_2)$.

- proved for stochastic operations
- used in practice for results obtained in DSA

F. Jézéquel, *Dynamical control of converging sequences computation*, Applied Numerical Mathematics, 50(2): 147-164, 2004.

Dynamical control of integrals on an infinite domain

Let $g = \int_0^\infty \phi(x)dx$ and $g_m = \sum_{j=0}^{m} f_j$ with $f_j = \int_{jL}^{(j+1)L} \phi(x)dx$.

We assume that (g_m) converges linearly to g.

An approximation of each integral can be computed in DSA,
such that its exact significant bits are those of f_j, up to 1.

Let G_m be the approximation of g_m computed in DSA.

$\Rightarrow$ the exact significant bits of G_m are those of g_m, up to 1.

$\Rightarrow$ if the convergence zone is reached,
the significant bits common to g_m and g_{m+1} are those of g, up to p.

$\Rightarrow$ if $G_m - G_{m+1} = @.0$,
the exact significant bits of G_{m+1} are those of g, up to $p+1$.

Dynamical control of multiple integrals computation

PhD M. Charikhi, Jan. 2005

$$I = \int_{\Omega} f(\mathbf{x}) d\mathbf{x} \text{ with } \Omega \subset \mathbb{R}^{N}$$

can be approximated by:

$$Q[f] = \sum_{j=1}^{\nu} a_j f(\mathbf{x_j}) \text{ with } a_j \in \mathbb{R} \text{ and } \mathbf{x_j} \in \Omega.$$

The approximation Q is called cubature formula if $N \geq 2$.

- polynomial-based methods
- Monte Carlo methods

Cubpack, R. Cools et al. 1992
VANI, C.-Y. Chen 1998
CLAVIS, S. Wedner 2000

Approximation using the principle of "iterated integrals"
Computation of 2-dimensional integrals

$$s = \int_a^b \int_{y_1(x)}^{y_2(x)} f(x,y)dxdy = \int_a^b g(x)dx \text{ with } g(x) = \int_{y_1(x)}^{y_2(x)} f(x,y)dy.$$

$\forall x \in [a,b]$, an approximation $G(x)$ can be computed in DSA such that its exact significant bits are those of $g(x)$, up to δ.

Let $S_n = \phi(\{G(x_i)\})$ be the approximation of s computed in DSA and $s_n = \phi(\{g(x_i)\})$.

- $\Rightarrow$ the exact significant bits of S_n are those of s_n, up to δ
- $\Rightarrow$ if the convergence zone is reached, the significant bits common to s_{n-1} and s_n are common with s, up to δ
- $\Rightarrow$ if $S_{n-1} - S_n = @.0$,
  the exact significant bits of S_n are those of s, up to 2δ.

The exact significant bits of the approximation obtained are those of the mathematical value of the integral, up to $N\delta$.

- With Romberg's method, $\delta = 0$.
- With the trapezoidal rule, $N\delta$ represents:
  - one bit if $N \leq 2$
  - one decimal digit if $N \leq 8$.
- With Simpson's rule, $N\delta$ represents one bit if $N \leq 35$.
- With the Gauss-Legendre method with 6 points, $N\delta$ represents one bit if $N \leq 2838$.

Computation of an integral involved in crystallography

$$g(a) = \int_0^{+\infty} f(x)dx,$$

with $f(x) = [\exp(x) + \exp(-x)]^a - \exp(ax) - \exp(-ax)$ and $0 < a < 2$.

$g(5/3) \approx 4.45$ (W. Harrison 1981)
$g(5/3) \approx 4.6262911$ (SIAM review 1996)

$g(a)$ can be expressed as a series expansion:

$$g(a) = \sum_{n=1}^{+\infty} \frac{\Pi_{i=0}^{n-1}(a-i)}{(n!)(2n-a)} - \frac{1}{a}.$$

F. Jézéquel, J.-M. Chesneaux, *Computation of an infi nite integral using Romberg's method*, Numerical Algorithms, 36(3): 265-283, 2004.

Several numerical problems may occur in the computation of $g(a)$:

- for high values of x, the computation of $f(x)$ may generate cancellations,
- the upper bound of the integral is infinite,
- the quadrature method used, e.g. Romberg's method, generates both a truncation error and a round-off error.

Computation of an integral involved in crystallography
Dynamical control of the computation

In order to avoid cancellations, the same expression of the integrand is not used at both bounds of the interval.

$$g(a) \approx \int_0^l f_1(x)dx + \sum_{j=1}^{k} \int_{jl}^{(j+1)l} f_2(x)dx,$$

where
$$f_1(x) = \exp(ax)\left[(1+\exp(-2x))^a - 1 - \exp(-2ax)\right]$$
$$f_2(x) = \exp(ax)u(x) - \exp(-ax),$$
$$u(x) = \lim_{n\to\infty} u_n(x) \text{ with } u_n(x) = \sum_{i=1}^{n-1} \frac{\exp(-2ix)}{i!} \Pi_{j=0}^{i-1}(a-j).$$

Dynamical choice of several parameters:

- n such that $u_n(x) \approx u(x)$
- k such that $\int_l^{kl} f_2(x)dx \approx \int_l^{\infty} f_2(x)dx$
- the number of iterations with Romberg's method

Proposition

One can compute an approximation $G(a)$ such that its exact significant digits are those of $g(a)$, up to $\delta = \log_{10}\left(\frac{2}{1-\exp^{-l\min(a,2-a)}}\right)$.

a	$\delta \approx$		$g(a)$
0.5	0.34	exact:	-1.69442616958795 8E+000
		DSA:	-1.6944261695879 5E+000
$5/3$	0.39	exact:	4.62629111198399 5E+000
		DSA:	4.626291111983E+000
1.9999	3.6	exact:	1.99989998677609 2E+004
		DSA:	1.9998999 7358E+004

The exact significant digits of $G(a)$ are in common with $g(a)$, up to $\lceil\delta\rceil$.

Study of an integral involved in the neutron star theory

$$\tau(\varepsilon, v) = \frac{1}{\omega(\varepsilon)} \int_0^{\frac{\pi}{2}} d\theta \sin(\theta) \int_0^{\infty} dn\, n^2 \int_0^{\infty} dp\, h(n, p, \theta, \varepsilon, v)$$

$(\varepsilon, v) \in [10^{-4}, 10^4] \times [10^{-4}, 10^3]$

ω is a normalization function

$$\begin{aligned} h(n, p, \theta, \varepsilon, v) = \;& \psi(z)\Gamma(n - \varepsilon - z) + \psi(-z)\Gamma(n - \varepsilon + z) \\ & -\psi(z)\Gamma(n + \varepsilon - z) - \psi(z)\Gamma(n + \varepsilon + z) \end{aligned}$$

with $z = \sqrt{p^2 + (v \sin(\theta))^2}$, $\psi(x) = \frac{1}{\exp(x)+1}$, $\Gamma(x) = \frac{x}{\exp(x)-1}$.

F. Jézéquel, F. Rico, J.-M. Chesneaux, M. Charikhi, *Reliable computation of a multiple integral involved in the neutron star theory*, submitted to "Mathematics and Computers in Simulation".

The numerical problems:

- two infinite bounds
  $\int_0^\infty$ ... is replaced by $\sum_{j=0}^{k} \int_{jL}^{(j+1)L}$ ...
  $\Rightarrow$ Dynamical choice of k
- $\Gamma(x) = \frac{x}{\exp(x)-1}$ generates cancellations if $x \approx 0$.
  a series expansion of $\Gamma(x)$ is used: $\Gamma(x) \approx \frac{1}{1+\frac{x}{2}+\ldots+\frac{x^{n-1}}{n!}}$
  $\Rightarrow$ Dynamical choice of n
- With the principle of “iterated integrals”, the Gauss-Legendre method is used and generates both a truncation error and a round-off error
  $\Rightarrow$ Dynamical control of the Gauss-Legendre method

The numerical problems:

- two infinite bounds
  $\int_0^\infty \ldots$ is replaced by $\sum_{j=0}^{k} \int_{jL}^{(j+1)L} \ldots$
  $\Rightarrow$ Dynamical choice of k
- $\Gamma(x) = \frac{x}{\exp(x)-1}$ generates cancellations if $x \approx 0$.
  a series expansion of $\Gamma(x)$ is used: $\Gamma(x) \approx \frac{1}{1+\frac{x}{2}+\ldots+\frac{x^{n-1}}{n!}}$
  $\Rightarrow$ Dynamical choice of n
- With the principle of "iterated integrals", the Gauss-Legendre method is used and generates both a truncation error and a round-off error
  $\Rightarrow$ Dynamical control of the Gauss-Legendre method

$\tau(\varepsilon, v)$ has been computed using DSA in single precision for 5752 points (ε, v) defined by:

$$\left\{\begin{array}{lllllll} \varepsilon & = & 10^a & \textit{with} & a & = & -4.0, -3.9, -3.8, \ldots, 4.0 \\ v & = & 10^b & \textit{with} & b & = & -4.0, -3.9, -3.8, \ldots, 3.0. \end{array}\right.$$

The run time of the code varies from 45 s to 3347 s depending on the values of ε and v, the average run time being 389 s.

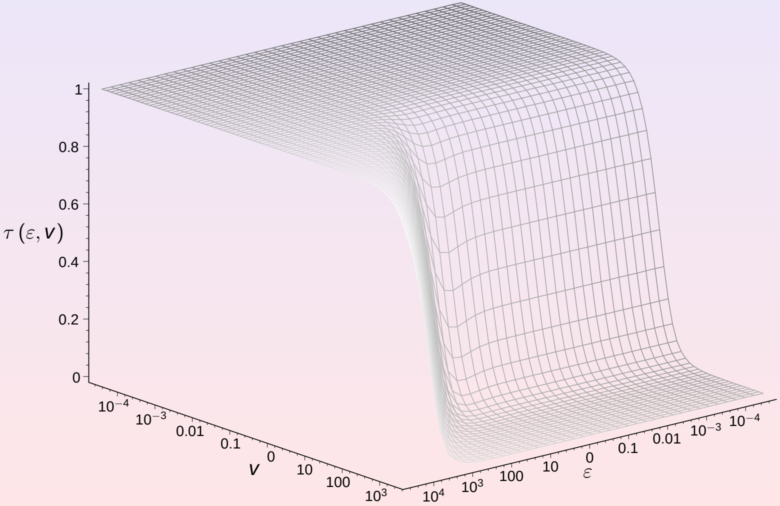
Proposition

One can compute an approximation of $\tau(\varepsilon, v)$ such that its exact significant digits are those of $\tau(\varepsilon, v)$, up to 2.

nb. of exact significant digits	occurrence
3	1
4	217
5	665
6	3347
7	1522

$\Rightarrow$ we can guarantee 1 to 5 significant digits in the results obtained.

Study of an integral involved in the neutron star theory
Numerical results



Conclusion and perspectives - 1/2

Dynamical control of converging sequences computation

Let $u = \lim_{n \to \infty} u_n$. From two iterates in the convergence zone, one can determine the first digits of u.

If $u_n - u_{n+1} = @.0$, one can determine which significant digits of u_{n+1} are in common with u.

Combination of theoretical results if several sequences are involved

For the approximation of an integral, one has to take into account:

- the dimension of the integral
- the number of improper bounds
- the possible approximation of the integrand by its series expansion
- the convergence speed of the sequences involved

Conclusion and perspectives - 2/2

- Adaptive strategies
- Other approximation methods
- Approximation of multiple integrals
  - other cubature methods
  - singular integrals
  - Monte Carlo methods
- Dynamical control of vector sequences computation
  PhD R. Adout
  - acceleration of the restarted GMRES method
  - dynamical control of the dimension of the Krylov subspace
- Automatic methods for round-off error analysis
  - DSA for MATLAB
  - compiler with DSA features
  - linear algebra library
  - grid computing: new methodologies