Contrôle dynamique de méthodes d’approximation

Fabienne Jézéquel
Laboratoire d’Informatique de Paris 6

ARINEWS, ENS Lyon, 7-8 mars 2005
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 $e_c(h)$.

If $h$ decreases, $L(h)$: 

```
<table>
<thead>
<tr>
<th>exponent</th>
<th>mantissa</th>
</tr>
</thead>
</table>
```

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

⇒ Continuous stochastic arithmetic: $X = (m, \sigma^2)$
⇒ 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 $C_{a,b} \approx 5.8$. 

F. Jézéquel
Dynamical control of approximation methods
7-8 Mar. 2005 4 / 23
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.
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.

Dynamical control of approximation methods

Theorem

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

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

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 \( L_n - L_{n+1} = \circ.0 \), the exact significant bits of \( L_{n+1} \) are those of \( L \), up to one.
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 $\circ$ be an arithmetical operator: $\circ \in \{+, -, \times, /\}$ and $s\circ$ the corresponding stochastic operator: $s\circ \in \{s+, s-, s\times, s/\}$.

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

proved for stochastic operations

used in practice for results obtained in DSA

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 $p+1$. 
\[ I = \int_{\Omega} f(\mathbf{x}) d\mathbf{x} \quad \text{with} \quad \Omega \subset \mathbb{R}^N \]

The approximation \( Q \) is called \textbf{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

\[ s = \int_a^b \int_{y_1(x)}^{y_2(x)} f(x, y) \, dx \, dy = \int_a^b g(x) \, dx \] 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 \( \delta \).

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 \( \delta \)

\[ \Rightarrow \] if the convergence zone is reached, the significant bits common to \( s_{n-1} \) and \( s_n \) are common with \( s \), up to \( \delta \)

\[ \Rightarrow \] if \( S_{n-1} = S_n = \@.0 \), the exact significant bits of \( S_n \) are those of \( s \), up to \( 2\delta \).
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 \( 0 < a < 2 \).

\[ g(5/3) \approx 4.45 \text{ (W. Harrison 1981)} \]
\[ g(5/3) \approx 4.6262911 \text{ (SIAM review 1996)} \]

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

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

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

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

where \( f_1(x) = \exp(ax) \left[ (1 + \exp(-2x))^a - 1 - \exp(-2ax) \right] \)
\( f_2(x) = \exp(ax)u(x) - \exp(-ax), \)

\[ u(x) = \lim_{n \to \infty} u_n(x) \text{ with } u_n(x) = \sum_{i=1}^{n-1} \frac{\exp(-2ix)}{i!} \prod_{j=0}^{i-1}(a - j). \]

Dynamical choice of several parameters:

- \( n \) such that \( u_n(x) \approx u(x) \)
- \( k \) such that \( \int_k^l f_2(x) \, dx \approx \int_l^\infty f_2(x) \, dx \)
- the number of iterations with Romberg’s method
Computation of an integral involved in crystallography
Theoretical and numerical results

**Proposition**

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

<table>
<thead>
<tr>
<th>$a$</th>
<th>$\delta \approx$</th>
<th>$g(a)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0.34</td>
<td>exact: $-1.694426169587958E+000$&lt;br&gt;DSA: $-1.694426169587958E+000$</td>
</tr>
<tr>
<td>5/3</td>
<td>0.39</td>
<td>exact: $4.626291111983995E+000$&lt;br&gt;DSA: $4.626291111983995E+000$</td>
</tr>
<tr>
<td>1.9999</td>
<td>3.6</td>
<td>exact: $1.999899986776092E+004$&lt;br&gt;DSA: $1.99989997358E+004$</td>
</tr>
</tbody>
</table>

The exact significant digits of $G(a)$ are in common with $g(a)$, up to $[\delta]$. 
\[ \tau(\varepsilon, \nu) = \frac{1}{\omega(\varepsilon)} \int_0^{\frac{\pi}{2}} d\theta \sin(\theta) \int_0^{\infty} dn \int_0^{\infty} dp \ h(n, p, \theta, \varepsilon, \nu) \]

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

\( \omega \) is a normalization function

\[ h(n, p, \theta, \varepsilon, \nu) = \psi(z)\Gamma(n - \varepsilon - z) + \psi(-z)\Gamma(n - \varepsilon + z) \]
\[ -\psi(z)\Gamma(n + \varepsilon - z) - \psi(z)\Gamma(n + \varepsilon + z) \]

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

Study of an integral involved in the neutron star theory

Dynamical control of the computation

The numerical problems:

- two infinite bounds
  \[ \int_0^\infty \ldots \text{ is replaced by } \sum_{j=0}^{k} \int_{jL}^{(j+1)L} \ldots \]
  \[ \Rightarrow \text{ Dynamical choice of } k \]

- \( \Gamma(x) = \frac{x}{\exp(x) - 1} \) generates cancellations if \( x \approx 0 \).
  
  a series expansion of \( \Gamma(x) \) is used: 
  \[ \Gamma(x) \approx \frac{1}{1 + \frac{x}{2} + \ldots + \frac{x^{n-1}}{n!}} \]
  \[ \Rightarrow \text{ Dynamical choice of } n \]

- With the principle of “iterated integrals”, the Gauss-Legendre method is used and generates both a truncation error and a round-off error
  \[ \Rightarrow \text{ Dynamical control of the Gauss-Legendre method} \]
The numerical problems:

- two infinite bounds
  \[ \int_0^\infty \ldots \text{is replaced by} \sum_{j=0}^k \int_{jL}^{(j+1)L} \ldots \]
  ⇒ 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: \( \Gamma(x) \approx \frac{1}{1 + \frac{x}{2} + \ldots + \frac{x^{n-1}}{n!}} \)
  ⇒ 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
  ⇒ Dynamical control of the Gauss-Legendre method
\( \tau (\varepsilon, \nu) \) has been computed using DSA in single precision for 5752 points \((\varepsilon, \nu)\) defined by:

\[
\begin{align*}
\varepsilon &= 10^a \quad \text{with} \quad a = -4.0, -3.9, -3.8, \ldots, 4.0 \\
\nu &= 10^b \quad \text{with} \quad b = -4.0, -3.9, -3.8, \ldots, 3.0.
\end{align*}
\]

The run time of the code varies from 45 s to 3347 s depending on the values of \(\varepsilon\) and \(\nu\), the average run time being 389 s.
Numerical quality of the approximations obtained

**Proposition**

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

<table>
<thead>
<tr>
<th>nb. of exact significant digits</th>
<th>occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>217</td>
</tr>
<tr>
<td>5</td>
<td>665</td>
</tr>
<tr>
<td>6</td>
<td>3347</td>
</tr>
<tr>
<td>7</td>
<td>1522</td>
</tr>
</tbody>
</table>

$\Rightarrow$ we can guarantee 1 to 5 significant digits in the results obtained.
Study of an integral involved in the neutron star theory

Numerical results

\[ \tau(\varepsilon, V) \]
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 \( u_n - u_{n+1} = 0.0 \), 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
  - Other approximation 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