Numerical software with result verification, Dagstuhl, 19-24 january 2003

Dynamical control of a multiple integral using Discrete Stochastic Arithmetic

M. Charrikhi, J.-M. Chesneaux, F. Jézéquel, F. Rico, L. Villain

Laboratoire d'Informatique de Paris 6, France
LUTH, Observatoire de Paris-Meudon, France

Neutron stars

Massive stars collapse with an explosion ejecting all the matter, except a central residue:

- a neutron star (NS)
- or a black hole.

NS have a radius of 10 km and gather about 1.4 M ($M \approx 2\ 10^{33}\ g$ is the mass of the sun).

NS are mainly done of neutrons with a typical proportion of protons smaller than 10%.

With time, equilibrium between protons and neutrons takes place.

Time to reach this equilibrium is obtained by calculating high order integrals involving several variables for any particle appearing in the reactions.

The integral to compute

$$I\left(\varepsilon, v\right)=\frac{1}{g\left(\varepsilon\right)}\int_0^{\frac{\pi}{2}} d\theta \sin\left(\theta\right)\int_0^{\infty} dn\ n^2 \int_0^{\infty} dp\ H\left(n,p,\theta,\varepsilon,v\right)$$

(ε, v) belong to $\left[10^{-6}, 10^{+6}\right]$

g is a normalization function

$$\begin{aligned} H\left(n,p,\theta,\varepsilon,v\right) = \quad & f(z)\Gamma(n-\varepsilon-z)+f(-z)\Gamma(n-\varepsilon+z) \\ & -f(z)\Gamma(n+\varepsilon-z)-f(z)\Gamma(n+\varepsilon+z) \end{aligned}$$

with $z=\sqrt{p^2+\left(v\sin(\theta)\right)^2}$, $f(x)=\frac{1}{e^x+1}$, $\Gamma(x)=\frac{x}{e^x-1}$

Numerical problems

- The integral is improper for the bound $\theta=0$.
- two infinite bounds
  $\int_0^{\infty}$ ... is replaced by $\sum_{j=0}^{k}\int_{jL}^{(j+1)L}$ ...
  How to choose k ?
- $\Gamma(x)=\frac{x}{e^x-1}$ generates cancellations if $x\approx 0$.
  $\Rightarrow$ a series expansion of $\Gamma(x)$ is used: $\Gamma(x)\approx\frac{1}{1+\frac{x}{2}+...+\frac{x^{n-1}}{n!}}$
  How to choose n ?
- With the principle of "iterated integrals", Gauss-Legendre method is used and generates a truncation error

Principle of Gauss-Legendre method

Gauss-Legendre method with n points:

$\int_{-1}^{1} f(x)dx$ is approximated by $\sum_{i=1}^{n} C_i f(x_i)$,

where $\{x_i\}$ are the roots of the n-degree Legendre polynomial P_n and $C_i = \frac{2}{(1-x_i^2)(P_n'(x_i))^2}$, for $i = 1, ..., n$.

To compute $I = \int_a^b g(t)dt$, from a change of variable, we get:

$$I = \frac{(b-a)}{2} \int_{-1}^{1} g\left(\frac{(b-a)x + (b+a)}{2}\right) dx.$$

Dynamical control of Gauss-Legendre method

$[a, b]$ is partitioned in 2^m sub-intervals of length $\frac{(b-a)}{2^m}$.

The integral on each sub-interval is approximated using the classical Gauss-Legendre method with n points.

Let I_m be the approximation of I obtained.

$I_m - I = \frac{K}{4^{nm}} \left[g^{(2n-1)}(b) - g^{(2n-1)}(a)\right] + \mathcal{O}(\frac{1}{4^{nm}})$,
where K does not depend on m.

$$\boxed{C_{I_{m-1},I_m} = C_{I_{m-1},I} + \log_{10}\left(\frac{4^n}{4^n - 1}\right) + \mathcal{O}\left(\frac{1}{4^{n(m-1)}}\right)}$$

where $C_{a,b}$ is the number of digits common to a and b.

The bits common to I_{m-1} and I_m are those of I, up to one.

Control of sequences having a linear convergence

Theorem: let us consider a sequence (I_n) which converges linearly to I, $I_n - I = K\alpha^n + o(\alpha^n)$ where $K \in$ IR and $0 < \alpha < 1$.

$$\boxed{C_{I_{n-1},I_n} = C_{I_{n-1},I} + \log_{10}\left(\frac{1}{1-\alpha}\right) + \mathcal{O}\left(\alpha^{n-1}\right)}$$

If $0 < \alpha \leq \frac{1}{2}$, $\log_2\left(\frac{1}{1-\alpha}\right)$ represents one bit: the bits common to I_{n-1} and I_n are those of I up to one.

This assertion is valid if $\mathcal{O}\left(\alpha^{n-1}\right)$ is negligible: this condition is satisfied when the convergence zone is reached.

The CESTAC method

M. La Porte and J. Vignes, 1972

- each arithmetical operation is performed N times using the random rounding mode

  $\Rightarrow$ for each result of any arithmetical operation, N results R_i are computed.
- computed result: $\overline{R} = \frac{1}{N}\sum_{i=1}^{N} R_i$.
- the number C of exact significant digits is determined by applying Student's test to the different R_i:

$$C = log_{10}\left(\frac{\sqrt{N}\ \left|\overline{R}\right|}{s\ \tau_\beta}\right) \quad \text{with} \quad s^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(R_i - \overline{R}\right)^2$$

Stochastic arithmetic

Using the CESTAC method, the results of each arithmetical operation can be considered as realizations of a Gaussian random variable.

A **stochastic number** X is a Gaussian random variable denoted by $\left(m, \sigma^2\right)$, where m is the mean value of X and σ its standard deviation.

A **computational zero** (denoted by @.0) is a computed result, which is not significant or which is the mathematical zero.

With this concept, the stochastic arithmetic has been developed with new definitions for order and equality relations.

Stochastic arithmetical operations

We define the elementary operations on two stochastic numbers $X_1 = \left(m_1, \sigma_1^2\right)$ and $X_2 = \left(m_2, \sigma_2^2\right)$ by:

$$X_1 \quad s+ \quad X_2 \stackrel{def}{=} \left(m_1 + m_2, \sigma_1^2 + \sigma_2^2\right),$$

$$X_1 \quad s- \quad X_2 \stackrel{def}{=} \left(m_1 - m_2, \sigma_1^2 + \sigma_2^2\right),$$

$$X_1 \quad s\times \quad X_2 \stackrel{def}{=} \left(m_1 * m_2, m_2^2\sigma_1^2 + m_1^2\sigma_2^2\right),$$

$$X_1 \quad s/ \quad X_2 \stackrel{def}{=} \left(m_1/m_2, \left(\frac{\sigma_1}{m_2}\right)^2 + \left(\frac{m_1\sigma_2}{m_2^2}\right)^2\right), \ \textit{with } m_2 \neq 0.$$

These definitions correspond to terms to the first order in $\frac{\sigma}{m}$ between two independent Gaussian random variables.

Control of stochastic operations

Theorem: Let X_i be an approximation of the exact value x_i in stochastic arithmetic. Let us assume that the exact significant bits of X_i 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 exact value $x_1 \bigcirc x_2$, up to $max(p_1, p_2)$.

Decomposition in one-dimensional integrals

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

$$I(\varepsilon, v) = \frac{1}{g(\varepsilon)} \int_0^{\frac{\pi}{2}} f_1(\theta, \varepsilon, v) d\theta$$

$$\text{with} \quad f_1(\theta, \varepsilon, v) = sin(\theta) \int_0^{\infty} f_2(n, \theta, \varepsilon, v) dn$$

$$\text{and} \quad f_2(n, \theta, \varepsilon, v) = n^2 \int_0^{\infty} H(n, p, \theta, \varepsilon, v)\, dp$$

By composing the previous theorems, the exact significant bits of the approximation obtained are those of $I(\varepsilon, v)$, up to 6.

Computation of $s = \int_0^\infty H(n, p, \theta, \varepsilon, v)\, dp$

Let $a_j = \int_{jL}^{(j+1)L} H(n, p, \theta, \varepsilon, v)\ dp$

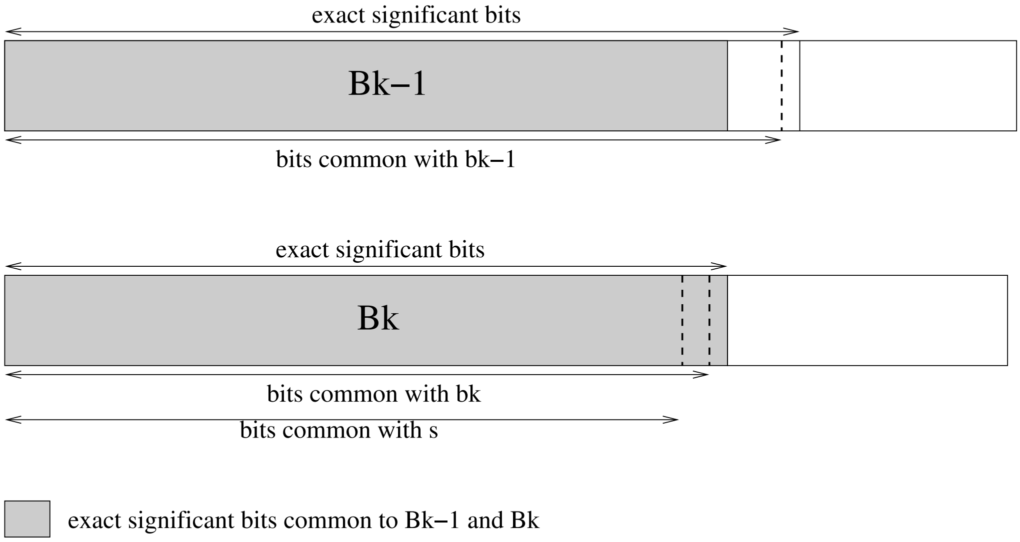
and A_j^p the approximation of a_j using Gauss-Legendre method on 2^p sub-intervals of $[jL, (j+1)L]$.

If $A_j^{p-1} - A_j^p = @.0$, the exact significant bits of A_j^p are those of a_j, up to 1.

The exact significant bits of $B_k = \sum_{j=0}^{k} A_j^p$ are those of $b_k = \sum_{j=0}^{k} a_j$, up to 1.

$b_k = \int_0^{(k+1)L} H(n, p, \theta, \varepsilon, v)\ dp$. As (b_k) converges linearly to s, the bits common to b_{k-1} and b_k are those of s up to 1.

If $B_{k-1} - B_k = @.0$,
the exact significant bits of B_k are those of s, up to 2.



Computation of $t = \int_0^\infty n^2 \left(\int_0^\infty H\left(n, p, \theta, \varepsilon, v\right) dp\right) dn$

Let $C_m = \sum_{j=0}^{m} \int_{jL}^{(j+1)L} n^2 B_k \; dn$.

The exact significant bits of C_m are those of $\int_0^{(m+1)L} n^2 \left(\int_0^\infty H\left(n, p, \theta, \varepsilon, v\right) dp\right) dn$, up to 3.

If $C_{m-1} - C_m = @.0$, the exact significant bits of C_m are those of t, up to 4.

Computation of $I\left(\varepsilon, v\right)$

Let $D_j = \int_{l_j}^{\frac{\pi}{2}} \sin\left(\theta\right) C_m \; d\theta$, with $l_j = \frac{\pi}{2.10^j}$.

The exact significant bits of D_j are those of $\int_{l_j}^{\frac{\pi}{2}} \sin\left(\theta\right) \left(\int_0^\infty n^2 \left(\int_0^\infty H\left(n, p, \theta, \varepsilon, v\right) dp\right) dn\right) d\theta$, up to 5.

If $D_{j-1} - D_j = @.0$, the exact significant bits of D_j are those of $I\left(\varepsilon, v\right)$, up to 6.

Computation with CADNA

The CADNA library allows the use on a computer of the synchronous CESTAC method and the concepts of stochastic arithmetic.

- Stability of the computation (> 12 digits)
- Computing time ($> 72h$)

The requested precision is 3 digits.

CADNA stops the computation when the **truncation error** is approximatively equal to the **round-off error**.

It is enough to stop the computation when 3 digits are guaranteed, *i.e.* when the truncation error is less than 10^{-5}.

Due to the stability of the computation, the round-off error is always below this level, and we can go back to the floating-point arithmetic.

Integration on an infinite domain

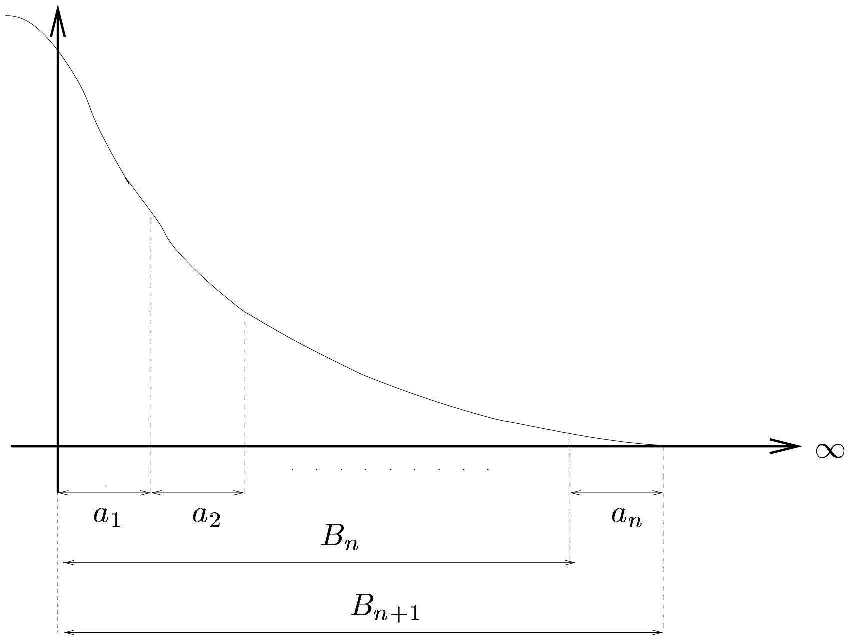
Each one-dimensional integral is evaluated by two iterative processes:

- A sequence (B_n) of integrals on a finite interval with an increasing upper bound

$$\begin{aligned} B_n &= \int_0^{nL} f(t)dt \\ B_{n+1} &= B_n + \underbrace{\int_{nL}^{(n+1)L} f(t)dt}_{a_n} \end{aligned}$$

- Each integral $a_n = lim_{p\to\infty} A_n^p$ is evaluated by Gauss-Legendre method and the stopping criterion is:

$$\left|A_n^p - A_n^{p+1}\right| < 10^{-k} \left|A_n^p\right|$$



Optimisation

If the threshold is the same for every evaluation, time may be lost in computing precisely a_n while the contribution of this value is negligible compared to B_n.

So

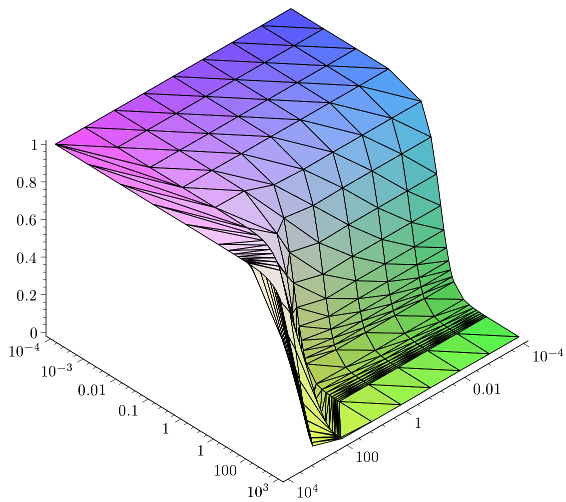
- Growing process is stopped when

$$|B_{n+1} - B_n| \quad < \quad 10^{-k}\,|B_n|$$

- The Gauss-Legendre process is stopped when

$$\left|A_n^{p+1} - A_n^{p}\right| \quad < \quad \underline{10^{-k}\,|B_n|}$$

Numerical results



For more information

at http://www.lip6.fr/cadna

- description of this study available with the source code
- description of the study, with the same approach, of an improper integral arising in a computation of energy in crystallography