Taylor models: proof that arithmetic operations performed with floating-point arithmetic provide guaranteed results

Nathalie Revol INRIA, projet Arénaire, LIP, ENS Lyon
Nathalie.Revol@ens-lyon.fr

Taylor models mini-workshop, Miami Beach, December 16-20, 2002

Outline

Taylor models: introduction

- Taylor models: definition
- Taylor models and floating-point arithmetic

Arithmetic operations in Taylor models using FP arithmetic

- Multiplication by a scalar
- Addition
- Multiplication

Proof that arithmetic operations are correct

- Multiplication by a scalar
- Addition
- Multiplication

Conclusion and "TODO" list

Taylor models: definition

$$\begin{array}{rccc} f & : & [-1,1]^v & \rightarrow & \mathbb{R} \\ & & (x_1, \cdots, x_v) & \mapsto & f(x_1, \cdots, x_v) \end{array}$$

is represented by

$$T_o(x_1, \cdots, x_v) + I_L$$

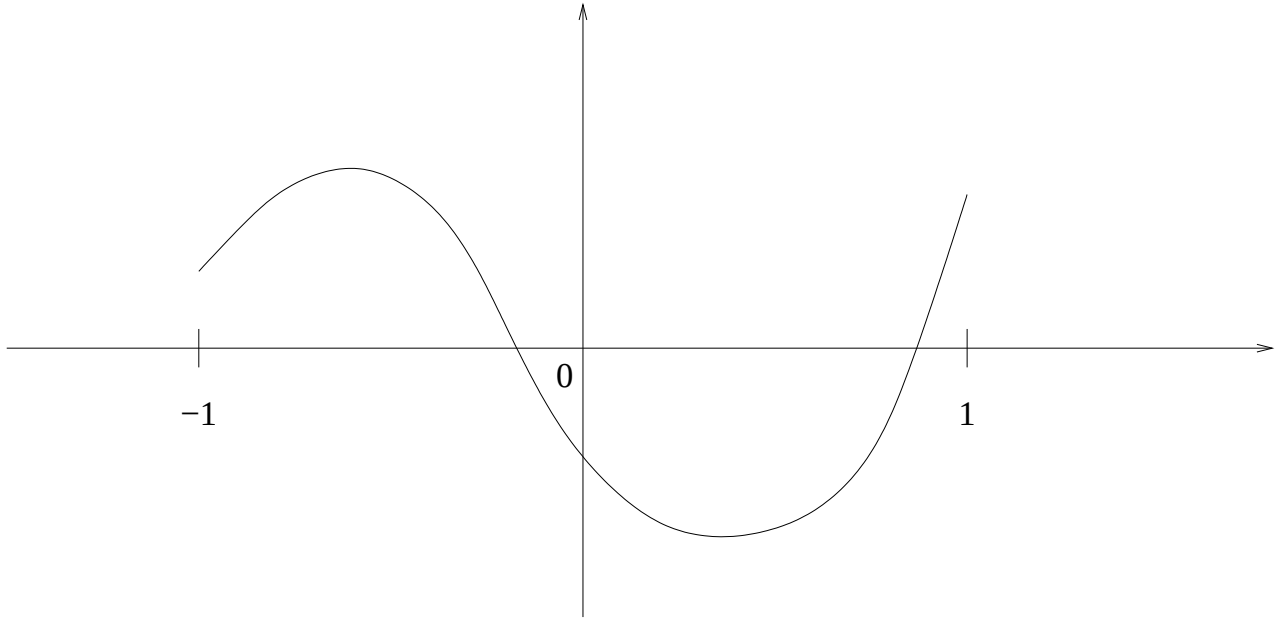
where

T_o is a polynomial of order o

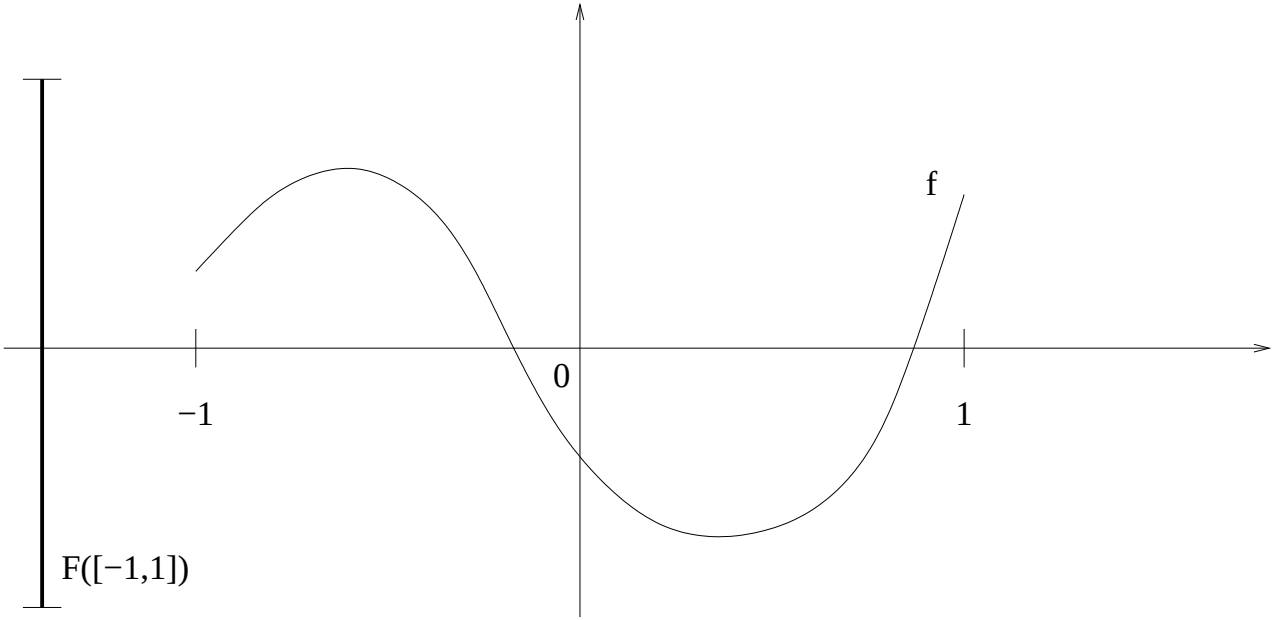
I_L is an interval enclosing the Lagrange remainder,

$$I_L = \frac{1}{(o+1)!} \| f^{(o+1)} \|_\infty \times [-1,1]$$

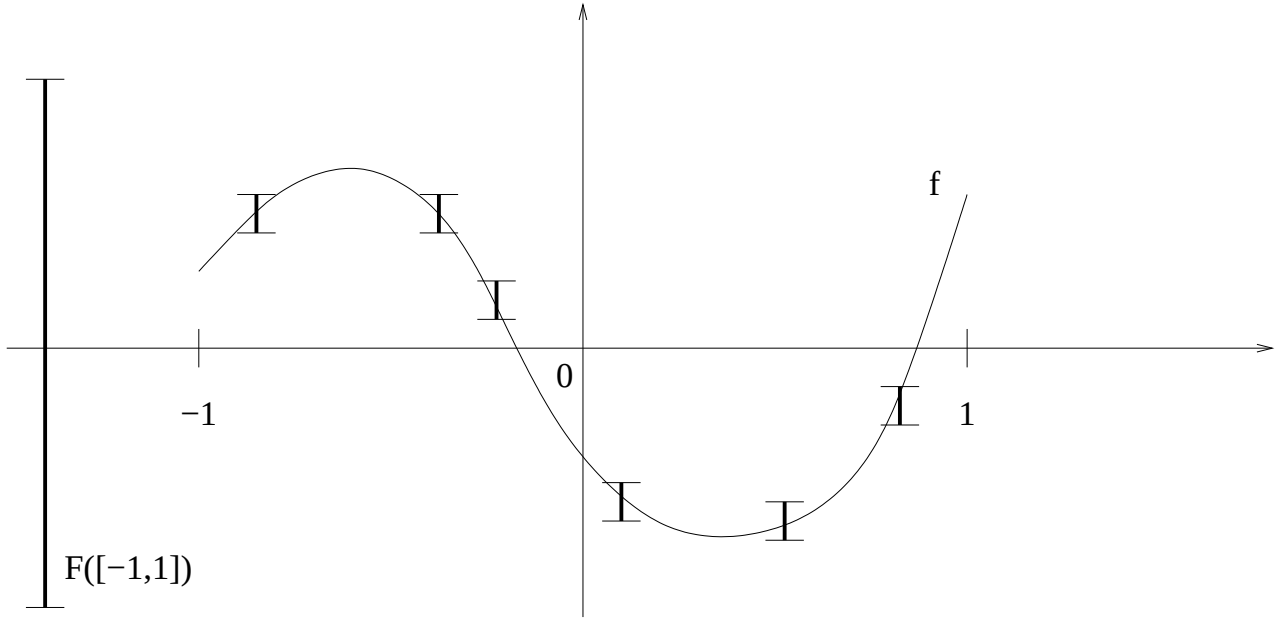
Taylor models: principle on a graphic



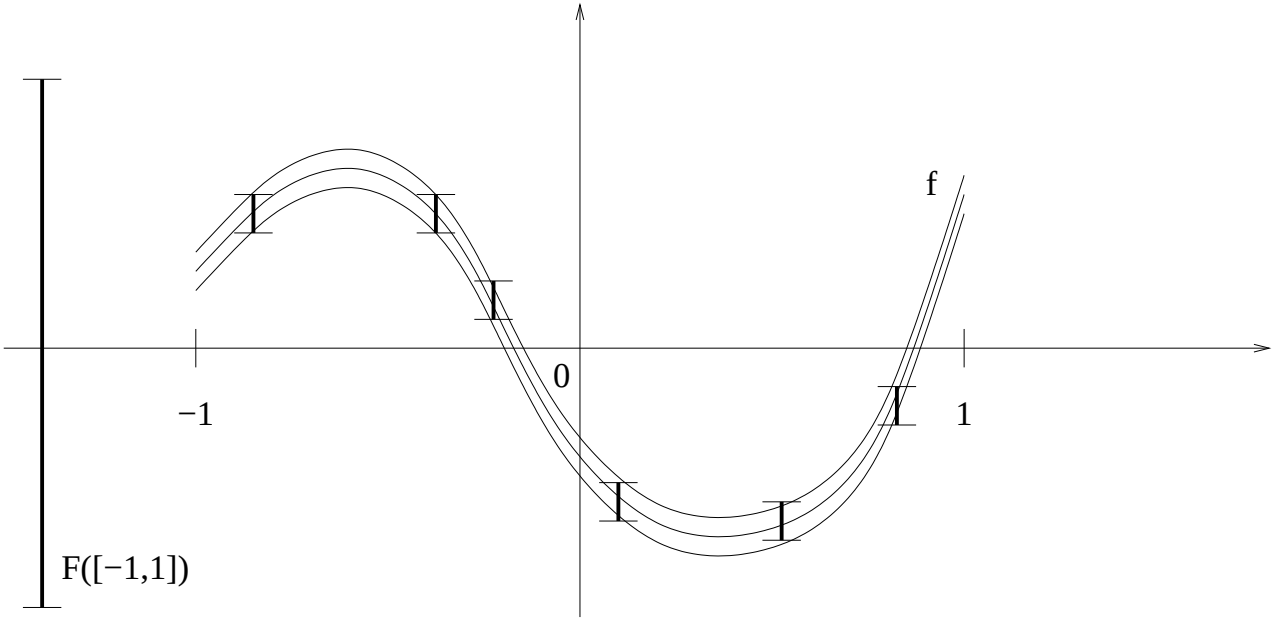
Taylor models: principle on a graphic



Taylor models: principle on a graphic



Taylor models: principle on a graphic



Taylor models and floating-point arithmetic

$$\begin{array}{rccc} f & : & [-1,1]^v & \rightarrow & \mathbb{R} \\ & & (x_1, \cdots, x_v) & \mapsto & f(x_1, \cdots, x_v) \end{array}$$

is represented by

$$T_o(x_1, \cdots, x_v) + I_{FP}$$

T_o is a polynomial **with floating-point coefficients** of order o

I_{FP} is an interval enclosing the Lagrange remainder

and **an enclosure of the rounding errors**

Taylor models, FP arithmetic and sparse representation

$$\begin{array}{rcccc} f & : & [-1,1]^v & \rightarrow & I\!R \\ & & (x_1, \cdots, x_v) & \mapsto & f(x_1, \cdots, x_v) \end{array}$$

is represented by

$$T_o(x_1, \cdots, x_v) + I$$

T_o is a polynomial with **not too small floating-point coefficients**

I_{FP} is an interval enclosing the Lagrange remainder,

an enclosure of the rounding errors and

an enclosure of the terms corresponding to small coefficients

Question: are the results guaranteed enclosures?

Recollection

[A. Benedetti] (message from March 2002)

```
In the works of Berz the coefficients of the polynomial
part are always represented by floating point numbers.  Should
these be intervals?  Since the coefficients are manipulated
every time the Taylor models are combined in arithmetic
operations or used as arguments in elementary functions,
how can I get verified result if intervals are not used
for the coefficients?
```

Question: what is the approximation order?

Definition of the approximation order

Usual analysis: x being a point, T is of order o iff

$$\forall x \in X,\ |T(x) - f(x)| = \mathcal{O}(x^o)$$

Taylor models (using exact arithmetic) are of order o.

What happens with floating-point coefficients?
Notion of $approximate\ order$?

Interval analysis: X being an interval, F is of order o iff

$$w(F(X)) = \mathcal{O}(w(X)^o)$$

Order > 2: NP-hard.

Cocnlusion: same vocabulary but not same meaning.

Outline

Taylor models: introduction

- Taylor model: definition
- Taylor models and floating-point arithmetic

Arithmetic operations in Taylor models using FP arithmetic

- Multiplication by a scalar
- Addition
- Multiplication

Proof that arithmetic operations are correct

- Multiplication by a scalar
- Addition
- Multiplication

Conclusion and "TODO" list

Taylor models: notations

coefficients of the polynomial: a_i, $1 \leq i \leq p$ (floating-point numbers)
I interval containing the Lagrange remainder, the bound on rounding errors and the swept terms.

ε_m : machine precision, $i.e.$ for every operation the relative rounding error is $\leq \varepsilon_m/2$
ε_u: machine $underflow$ threshold
ε_c: logical $underflow$ threshold: every number $< \varepsilon_c$ is replaced by 0 or swept (hyp: $\varepsilon_c^2 > \varepsilon_u$)

t : tallying variable (for rounding errors)
s : sweeping variable (to get rid of small coefficients).

Taylor models: operations using an ideal arithmetic

Multiplication by a scalar

$T' = [b_k, J] = c \times T$ with $T = [a_i, I]$:

for $k = 1$ to p do

$b_k = c \times a_k$

$J = c \times I$

Addition

$T = [b_k, J] = T_1 + T_2$ with $T_1 = [a_i^{(1)}, I^{(1)}]$ and $T_2 = [a_i^{(2)}, I^{(2)}]$:

for $k = 1$ to p do

$b_k = a_k^{(1)} + a_k^{(2)}$

$J = I^{(1)} + I^{(2)}$

Taylor models: operations using an ideal arithmetic

Multiplication

$T = [b_k, J] = T_1 \times T_2$ with $T_1 = [a_i^{(1)}, I^{(1)}]$ and $T_2 = [a_i^{(2)}, I^{(2)}]$:

$J = I^{(1)} \times I^{(2)}$

for $k = 1$ to p do

  for $j = 1$ to p do

    if $l :=$ order$(j)+$ order$(k) \leq o$ then

      $b_l = b_l + a_k^{(1)} \times a_j^{(2)}$

    else

      $J = J + [-|a_k^{(1)} \times a_j^{(2)}|, |a_k^{(1)} \times a_j^{(2)}|]$

$J = J + [-\sum_i |a_i^{(1)}|, \sum_i |a_i^{(1)}|] \times I^{(2)}$

$J = J + I^{(1)} \times [-\sum_i |a_i^{(2)}|, \sum_i |a_i^{(2)}|]$

Taylor models: operations using floating-point arithmetic

Notations: ideal operations: usual symbols,
floating-point or interval operations: circled symbols.

Multiplication by a scalar

$T' = [b_k, J] = c \times T$ with $T = [a_i, I]$:

	error upper bounded by:
for $k = 1$ to p do	
$b_k = c \otimes a_k$	$\varepsilon_m/2 \times \lvert c \times a_k \rvert$
$J = c \otimes I$	interval operation

Taylor models: operations using floating-point arithmetic

Addition

$T = [b_k, J] = T_1 + T_2$ with $T_1 = [a_i^{(1)}, I^{(1)}]$ and $T_2 = [a_i^{(2)}, I^{(2)}]$:

	error upper bounded by:
for $k = 1$ to p do	
$b_k = a_k^{(1)} \oplus a_k^{(2)}$	$\varepsilon_m/2 \times (\lvert a_k^{(1)} \rvert + \lvert a_k^{(2)} \rvert)$
$J = I^{(1)} \oplus I^{(2)}$	interval operation

Taylor models: operations using FP arithmetic

Multiplication

$T = [b_k, J] = T_1 \times T_2$ with $T_1 = [a_i^{(1)}, I^{(1)}]$ and $T_2 = [a_i^{(2)}, I^{(2)}]$:

	error upper bounded by:
$J = I^{(1)} \otimes I^{(2)}$	
for $k = 1$ to p do	
for $j = 1$ to p do	
$p = a_k^{(1)} \otimes a_j^{(2)}$	$\varepsilon_m/2 \times \lvert a_k^{(1)} \times a_k^{(2)} \rvert$
if $l :=$ order$(j)+$ order$(k) \leq o$ then	
$b_l = b_l \oplus p$	$\varepsilon_m/2 \times (\lvert b_l \rvert + \lvert p \rvert)$
else	
$J = J \oplus [-\lvert p \rvert, \lvert p \rvert]$	interval operation
$J = J + [-\sum_i \lvert a_i^{(1)} \rvert, \sum_i \lvert a_i^{(1)} \rvert] \times I^{(2)}$	interval operations
$J = J + I^{(1)} \times [-\sum_i \lvert a_i^{(2)} \rvert, \sum_i \lvert a_i^{(2)} \rvert]$	except on the sums

Outline

Taylor models: introduction

- Taylor model: definition
- Taylor models and floating-point arithmetic

Arithmetic operations in Taylor models using FP arithmetic

- Multiplication by a scalar
- Addition
- Multiplication

Proof that arithmetic operations are correct

- Multiplication by a scalar
- Addition
- Multiplication

Conclusion and "TODO" list

Taylor models: estimating floating-point errors using floating-point arithmetic

Problem (paradox): rounding errors are due to floating-point arithmetic: how to estimate them using **floating-point arithmetic?**

Starting point:

$$\text{(assumption) nb op} \times \varepsilon_m \leq 1/2$$

and

$$|(a \oplus b) - (a + b)| \leq \varepsilon_m \otimes (|a| \oplus |b|)$$

and even

$$|(a \oplus b) - (a + b)| \leq \varepsilon_m \otimes \max(|a|, |b|)$$

$$|(a \otimes b) - (a \times b)| \leq \varepsilon_m \otimes |a \otimes b|$$

Taylor models: estimating floating-point errors using floating-point arithmetic

Multiplication by a scalar

$T' = [b_k, J] = c \times T$ with $T = [a_i, I]$:

$t = 0$

$s = 0$

for $k = 1$ to p do

$b_k = c \otimes a_k$

$t = t \oplus |b_k|$

if $|b_k| < \varepsilon_c$ then

$s = s \oplus |b_k|$

$b_k = 0$

$J = c \otimes I \oplus 2 \otimes \varepsilon_m \otimes [-t, t] \oplus 2 \otimes [-s, s]$

Taylor models: estimating floating-point errors using floating-point arithmetic

Goal: either prove that is provides guaranteed results or propose an algorithm that provides guaranteed results.

1. prove that the t variable really takes into account rounding errors (i.e. the errors for $c \otimes a_k$ and the errors for the accumulation into t);
2. prove that the swept terms (put into s) and the errors for the computation of s are correctly taken into account;
3. the last operation is an interval one, thus rounding errors are taken into account properly.

Notations: ideal operations: usual symbols,
floating-point or interval operations: circled symbols.

Taylor models: multiplication by a scala proof for the computation of t

- error on $b_k = c \otimes a_k \leq \varepsilon_m \otimes |b_k| \leq \varepsilon_m \left(1 + \frac{\varepsilon_m}{2}\right) |b_k|$
$\Rightarrow \sum_k$ error on $b_k \leq \sum_k \varepsilon_m \otimes |b_k| \leq \varepsilon_m \left(1 + \frac{\varepsilon_m}{2}\right) \sum_k |b_k|$.

- computed value $= 2 \otimes \varepsilon_m \otimes \bigoplus_k |b_k| \geq 2\varepsilon_m \left(1 - \frac{\varepsilon_m}{2}\right) \bigoplus_k |b_k|$.

Does this hold: $\varepsilon_m \left(1 + \frac{\varepsilon_m}{2}\right) \sum_k |b_k| \leq 2\varepsilon_m \left(1 - \frac{\varepsilon_m}{2}\right) \bigoplus_k |b_k|$?

- **Relation between $\sum_k |b_k|$ and $\bigoplus_k |b_k|$?**

Higham (4.4): **from now on, $n =$ nb op**

$$\begin{array}{rcl} \left|\sum_k |b_k| - \bigoplus_k |b_k|\right| & \leq & n\frac{\varepsilon_m}{2} \sum_k |b_k| \\ \Rightarrow \quad \sum_k |b_k| & \leq & \frac{1}{1-n\frac{\varepsilon_m}{2}} \bigoplus_k |b_k| \end{array}$$

$$\text{We have } \varepsilon_m \left(1+\frac{\varepsilon_m}{2}\right) \sum_k |b_k| \leq \frac{\varepsilon_m \left(1+\frac{\varepsilon_m}{2}\right)}{1-n\frac{\varepsilon_m}{2}} \bigoplus_k |b_k|,$$

$$\text{dos this hold? } \frac{\varepsilon_m \left(1+\frac{\varepsilon_m}{2}\right)}{1-n\frac{\varepsilon_m}{2}} \bigoplus_k |b_k| \leq 2\varepsilon_m \left(1-\frac{\varepsilon_m}{2}\right) \bigoplus_k |b_k| ?$$

$$\text{i.e. } \frac{1+\frac{\varepsilon_m}{2}}{\left(1-\frac{\varepsilon_m}{2}\right).\left(1-n\frac{\varepsilon_m}{2}\right)} \leq 2?$$

Yes thanks to $\varepsilon_m < 1/5$ and $n\varepsilon < 1/2$.

proof for the computation of s

- Let K denote $\{k/|b_k| < \varepsilon_c\}$, let us prove that

$$2 \otimes s = 2. \bigoplus_{k \in K} |b_k| \geq \sum_{k \in K} |b_k| + \text{ error on this sum.}$$

- error on this sum $\leq \left(1 - \sharp K \frac{\varepsilon_m}{2}\right)^{-1} . \sharp K \frac{\varepsilon_m}{2} \bigoplus_{k \in K} |b_k|$
- The following holds

$$\sum_{k \in K} |b_k| + \text{ error on this sum } \leq \left(1 - \sharp K \frac{\varepsilon_m}{2}\right)^{-1} . \left(1 + \sharp K \frac{\varepsilon_m}{2}\right) \bigoplus_{k \in K} |b_k|$$

  - as $\sharp K \varepsilon_m < 1/2$, $\left(1 - \sharp K \frac{\varepsilon_m}{2}\right)^{-1} \leq \frac{4}{3}$ and $\left(1 + \sharp K \frac{\varepsilon_m}{2}\right) \leq \frac{5}{4}$

this gives the sought result.

Taylor models: proof for the addition

Addition

$T = [b_k, J] = T_1 + T_2$ with $T_1 = [a_i^{(1)}, I^{(1)}]$ and $T_2 = [a_i^{(2)}, I^{(2)}]$:

$t = 0$

$s = 0$

for $k = 1$ to p do

$b_k = a_k^{(1)} \oplus a_k^{(2)}$

$t = t \oplus |a_k^{(1)}| \oplus |a_k^{(2)}|$

if $|b_k| < \varepsilon_c$ then

$s = s \oplus |b_k|$

$b_k = 0$

$J = I^{(1)} \oplus I^{(2)} \oplus 2 \otimes \varepsilon_m \otimes [-t, t] \oplus 2 \otimes [-s, s]$

Taylor models: proof for the addition

Goal: either prove that is provides guaranteed results or propose an algorithm that provides guaranteed results.

1. prove that the t variable really takes into account rounding errors (i.e. the errors for $a_k^{(1)} \oplus a_k^{(2)}$ and the errors for the accumulation into t): cf. proof for the multiplication by a scalar;

2. prove that the swept terms (put into s) and the errors for the computation of s are correctly taken into account: cf. proof for the multiplication by a scalar;

3. the last operation is an interval one, thus rounding errors are taken into account properly.

Taylor models: proof for the multiplication

$T = [b_k, J] = T_1 \times T_2$ with $T_1 = [a_i^{(1)}, I^{(1)}]$ and $T_2 = [a_i^{(2)}, I^{(2)}]$

$t = 0,\ s = 0$

$J = I^{(1)} \otimes I^{(2)}$

for $k = 1$ to p do

  for $j = 1$ to p do

    $p = a_k^{(1)} \otimes a_j^{(2)}$

    $t = t \oplus |p|$

    if $l :=$ order$(j)+$ order$(k) \leq o$ then

      $b_l = b_l \oplus p$

      $t = t \oplus \max(|b_l|, |p|)$

    if $|b_k| < \varepsilon_c$ then

      $s = s \oplus |b_k|$ and then $b_k = 0$

$$J\oplus = 2\otimes\bigoplus_j \left(([-1,1]\otimes\bigoplus_{k>o-j}|a_k^{(1)}|)\oplus I^{(1)}\right)\otimes[-|a_j^{(2)}|,|a_j^{(2)}|]$$

$$J = J\oplus 2\otimes[-\bigoplus_k|a_k^{(1)}|,\bigoplus_k|a_k^{(1)}|]\otimes I^{(2)}$$

$$J = J\oplus 2\otimes\varepsilon_m\otimes[-t,t]\oplus 2\otimes[-s,s]$$

Taylor models: proof for the multiplication

Goal: either prove that is provides guaranteed results or propose an algorithm that provides guaranteed results.

1. prove that the t variable really takes into account rounding errors: cf. proof for the multiplication by a scalar;

2. prove that the swept terms (put into s) and the errors for the computation of s are correctly taken into account: cf. proof for the multiplication by a scalar;

3. in the last line, rounding errors are taken into account thanks to interval arithmetic and to the factor "2" for the sums.
   Remark: this is different from Cosy, where interval operations are performed at each step $\Rightarrow$ no need for this factor "2" (?).

Conclusion and future work

• **Almost done:** check that the algorithms given here are the ones implemented in Cosy.

• **Done:** prove that the rounding errors are correctly taken into account, $i.e.$ that even with FP arithmetic, results are **guaranteed**. Multiplication to be discussed. . .

• **To do:** proof that translations-homotheties are also correct with FP arithmetic (from any domain to $[-1, 1]$ and reciprocally).

• **To do:** same work on the $intrinsics$: $/$, $\sqrt{\ }$ and elementary functions (with some reasonable assumptions on the quality of FP elementary functions).