An hybrid approach to solve constraints over the floating point numbers

Claude Michel\textsuperscript{1}, Mohammed Said Belaid\textsuperscript{1}, Olivier Ponsini\textsuperscript{1}, Michel Rueher\textsuperscript{1}

\textsuperscript{1} I3S (Nice University-CNRS)
930, Route des Colles - BP 145, 06903 Sophia Antipolis Cedex (France)

January 7th, 2013
Motivations

Our aim

To *reason about floating point computation* using *constraints*

i.e., to validate or verify some piece of software, e.g.,

- automatic test case generation
- verify the conformity of a program with its specification

CP and verification

- successfully applied to programs with integer computation
- provides examples or counter examples
- needs a solver over the floating point numbers
\( \mathcal{R} \) vs \( \mathcal{F} \): the need for a dedicated solver

### Constraints over \( \mathcal{R} \) vs constraints over \( \mathcal{F} \)

- **Similarities**
  - similar syntax
  - use of floating point bounded intervals for domains

- **Differences**
  - different semantics
  - different domains

- **Examples**
  - \( x + 2.0 = 2.0 \)
    - \( 2.0 \) is a solution over \( \mathcal{R} \) and over \( \mathcal{F} \)
  - \( x + 2.0 = 2.0 \land x \neq 0 \)
    - no solution over \( \mathcal{R} \), numerous solutions over \( \mathcal{F} \)
  - \( x \times x = 2.0 \) (with a rounding to the nearest)
    - solutions over \( \mathcal{R} \), no solution over \( \mathcal{F} \)
An hybrid approach to solve constraints over the floating point numbers
Floating point computation

Properties

Somewhat counterintuitive

- Addition is non-associative: \((x \oplus y) \oplus z \neq x \oplus (y \oplus z)\)
- Multiplication is non-distributive: \(x \odot (y \oplus z) \neq x \odot y \oplus x \odot z\)
- Absorption: \(1 \oplus 10^{-10} = 1\)

Few good properties

- Commutativity is kept: \(x \oplus y = y \oplus x\)
- \(2 \otimes x = 2 \times x\) (unless overflow)
- If \(y/2 \leq x \leq 2 \times y\) then \(x \ominus y = x - y\) (Sterbenz 74)

Thus

- ... Few properties ... but ... we need to reason over \(F\)!
Projection functions over $\mathcal{F}$

**What are they?**
- solver building blocks

**Domains**
- variable domains = interval over $\mathcal{F}$, i.e.,
  \[
  \text{Notation: } x = [\underline{x}, \overline{x}] = \{x \in \mathcal{F}, \underline{x} \leq x \leq \overline{x}\}
  \]

**2 types of projection functions**
- Direct projection ($f$ increasing function + correctly rounded)
  \[
  \text{if } y = f_{\mathcal{F},+\infty}(x) \text{ then } y = y \cap [f_{\mathcal{F},+\infty}(\underline{x}), f_{\mathcal{F},+\infty}(\overline{x})]
  \]
- Inverse projection: ($f$ increasing function + correctly rounded)
  \[
  \text{if } y = f_{\mathcal{F},+\infty}(x) \text{ then } x = x \cap [f_{\mathcal{F},+\infty}^{-1}(\underline{y}), f_{\mathcal{F},+\infty}^{-1}(\overline{y})]
  \]
Constraints over $\mathcal{F}$ and limitations

**Constraints over $\mathcal{F}$**
- Direct handling of floating point computations
- All (IEEE 754) rounding modes: $-\infty$, $+\infty$, 0, to the nearest
- Basic operations (+, −, *, /) 
- ...and some other functions (sqrt, abs, ...)
- Floating point labelling

**Limitations**
- Based on a local propagation algorithm
- Suffers from multiple occurrence issues, e.g. $k \times x \times x \times x$ handled as $k \times x \times y \times z$
- Some “in house” solutions: kB consistencies ($\approx$ shaving)
- However, *need for more efficient solutions*!
Overcoming solver limitations

How to improve solver behavior?

- Introduce more *global* approach
  (though the strong limitations of floating point arithmetic)
- How? approximate over $R$ the initial problem over $F$
  to allow to reduce domain variables using solvers over $R$

Expected benefits

- More efficiency
- More flexibility
- More globality
- ...at the price of precision
- Requires cooperation of both approaches
  a so called “hybrid” approach...
An hybrid approach to solve constraints over the floating point numbers
Floating point constraints

An hybrid approach

Hybrid propagation

1. Local propagation over $\mathcal{F}$
2. Correct relaxations over $\mathcal{R}$ of the problem over $\mathcal{F}$
3. Correct linearizations over $\mathcal{R}$ of non linear terms
4. For each variable $x$
   1. $x \leftarrow \max(x, \text{Round}_{+\infty}(\text{CorrectLPMin}(x)))$
   2. $\overline{x} \leftarrow \min(\overline{x}, \text{Round}_{-\infty}(-\text{CorrectLPMin}(-x)))$
5. Local propagation over $\mathcal{F}$
An hybrid approach to solve constraints over the floating point numbers
Relaxations over $\mathbb{R}$

Relaxations

Relying on

- Correctly rounded operations: $f_\mathcal{F}(x) = o(f_\mathcal{R}(x))$
- Relative error: $\epsilon = \left| \frac{f_\mathcal{R}(x) - f_\mathcal{F}(x)}{f_\mathcal{R}(x)} \right|$

Property:

Assume that

- $z_\mathcal{F} = f_\mathcal{F}(x)$ is correctly rounded toward $-\infty$
- $z_\mathcal{F}$ is positive, normalized and $z_\mathcal{F} < \max_\mathcal{F}$,
- $z_\mathcal{R} = f_\mathcal{R}(x)$

we have

$$\frac{1}{1+z_\mathcal{R}} \leq z_\mathcal{F} \leq z_\mathcal{R}$$

with $z_\mathcal{R} \in [z_\mathcal{F}, z_\mathcal{F}^+]$
## Relaxations

<table>
<thead>
<tr>
<th></th>
<th>negative</th>
<th></th>
<th>positive</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>round</td>
<td>normal</td>
<td>denormal</td>
<td>denormal</td>
<td>normal</td>
</tr>
<tr>
<td>−∞</td>
<td>$[1^+ z_R, z_R]$</td>
<td>$[z_R - \min_F, z_R]$</td>
<td>$[z_R - \min_F, z_R]$</td>
<td>$[\frac{1}{1} z_R, z_R]$</td>
</tr>
<tr>
<td>+∞</td>
<td>$[z_R, \frac{1}{1} z_R]$</td>
<td>$[z_R, z_R + \min_F]$</td>
<td>$[z_R, z_R + \min_F]$</td>
<td>$[z_R, 1^+ z_R]$</td>
</tr>
<tr>
<td>0</td>
<td>$[z_R, \frac{1}{1} z_R]$</td>
<td>$[z_R + \min_F, z_R]$</td>
<td>$[z_R, z_R - \min_F]$</td>
<td>$[\frac{1}{1} z_R, z_R]$</td>
</tr>
<tr>
<td>near</td>
<td>$[\frac{2<em>1^+}{1+1^+} z_R, \frac{2</em>(1^-)}{1+(1^-)} z_R]$</td>
<td>$[z_R - \frac{\min_F}{2}, z_R + \frac{\min_F}{2}]$</td>
<td>$[z_R - \frac{\min_F}{2}]$</td>
<td>$[\frac{2*(1^-)}{1+(1^-)} z_R, \frac{2*1^+}{1+1^+} z_R]$</td>
</tr>
</tbody>
</table>

**Table:** Relaxations for each rounding mode with $-\max_F < z_R < \max_F$. 
An hybrid approach to solve constraints over the floating point numbers
Relaxations over \( \mathbb{R} \)

**Relaxations**

\[
\begin{align*}
\text{rounding mode: } & -\infty \\
z_F &= z_R \\
\text{max}D_F \\
1^+ z_R \\
\text{max}D_F: \text{ biggest denormalized number}
\end{align*}
\]
An hybrid approach to solve constraints over the floating point numbers
Relaxations over $\mathbb{R}$

Relaxation convex hull

rounding mode: $-\infty$

$z_F = z_R$

$z_F = \frac{1}{1+\varepsilon} z_R$

Warning

$max D_F$

$max D_F$

$(z_F, 1^+ z_F)$

Convex hull

$(z_F, z_F - min F)$
Linearizations of non linear terms

Relaxations over $\mathcal{R}$: non linear terms
- E.g, relaxation of $z_F = x \cdot y$ yields
  $$\alpha_{inf} xy + \beta_{inf} \leq z_F \leq \alpha_{sup} xy + \beta_{sup}$$
- Such non linear terms need to be linearized

Bilinear terms $xy$ linearizations
McCormick (1976): relaxations of $xy$ over $[x, \bar{x}] \times [y, \bar{y}]$:

$$[xy]_R = \begin{cases} 
(x - x)(y - y) \geq 0 \\
(x - x)(\bar{y} - y) \geq 0 \\
(\bar{x} - x)(y - y) \geq 0 \\
(\bar{x} - x)(\bar{y} - y) \geq 0
\end{cases}$$
An hybrid approach to solve constraints over the floating point numbers
Relaxations over $\mathbb{R}$

Safe use of LP

LP must be correct

Coefficients must be computed with the correct rounding direction

$$\begin{align*}
  z_R &\leq -xy -yx + \sup(\overline{xy}) \\
  -z_R &\leq +xy +yx + \sup(-\overline{xy}) \\
  -z_R &\leq +\overline{xy} +yx + \sup(-\overline{xy}) \\
  z_R &\leq -\overline{xy} -yx + \sup(\overline{xy})
\end{align*}$$

Correct minimizer

- Efficient LP solvers use floating point computation and are thus subject to rounding errors
- Neumaier & Scherbina (2004) propose a procedure to compute a correct minimizer
- Needs bounds: $b \leq \sum_{i=1}^{n} k_i x_i \leq \overline{b}$
Capture more globality

- Relaxations made on single FP operation basis
  - assume $z_F = x \times y \times v$, we get
    
    $$\alpha_{inf} xy + \beta_{inf} \leq z'_F \leq \alpha_{sup} xy + \beta_{sup}$$
    
    $$\alpha'_{inf} z'_F v + \beta'_{inf} \leq z_F \leq \alpha'_{sup} z'_F v + \beta'_{sup}$$

- Need to capture more global terms like $x \times x \times \ldots \times x \ (x^n)$ to provide tighter linearizations
  
  $$\alpha_{inf} x^n + \ldots + \beta_{inf} \leq z_F \leq \alpha_{sup} x^n + \ldots + \beta_{sup}$$

- Requires an efficient and correct procedure
The End