Certified Roundoff Error Bounds using Bernstein Expansions and Sparse Krivine-Stengle Representations

Alexandre Rocca, Victor Magron, Thao Dang

Certified Roundoff Error Bounds using Bernstein Expansions and Sparse Krivine-Stengle Representations

Alexandre Rocca[1,2], Victor Magron[1], and Thao Dang[1]

[1] VERIMAG/CNRS
700 avenue Centrale, 38400 Saint Martin D'Hères, France
[2] UJF-Grenoble 1/CNRS, TIMC-IMAG,
UMR 5525, Grenoble, F-38041, France

Abstract. Floating point error is an inevitable drawback of embedded systems implementation. Computing rigorous upper bounds of roundoff errors is absolutely necessary for the validation of critical software. This problem of computing rigorous upper bounds is even more challenging when addressing non-linear programs. In this paper, we propose and compare two new methods based on Bernstein expansions and sparse Krivine-Stengle representations, adapted from the field of the global optimization, to compute upper bounds of roundoff errors for programs implementing polynomial functions. We release two related software package FPBern and FPKiSten, and compare them with state of the art tools.
We show that these two methods achieve competitive performance, while computing accurate upper bounds by comparison with other tools.

Keywords: Polynomial Optimization, Floating Point Arithmetic, Roundoff Error Bounds, Linear Programming Relaxations, Bernstein Expansions, Krivine-Stengle Representations

1 Introduction

Theoretical models, algorithms, and programs are often reasoned and designed in real algebra. However, their implementation on computers uses floating point algebra: this conversion from real numbers and their operations to floating point is not without errors. Indeed, due to finite memory and binary encoding in computers, real numbers cannot be exactly represented by floating point numbers. Moreover, numerous properties of the real algebra are not conserved such as commutativity or associativity.

The consequences of such imprecisions become particularly significant in safety-critical systems, especially in embedded systems which often include control components implemented as computer programs. When implementing an algorithm designed in real algebra, and initially tested on computers with single or double floating point precision, one would like to ensure that the roundoff error is not too large on more limited platforms (small processor, low memory capacity) by computing their accurate upper bounds.

For programs implementing linear functions, SAT/SMT solvers, as well as affine arithmetic, are efficient tools to obtain good upper bounds. When extending to programs with non-linear polynomial functions, the problem of determining a precise upper bound becomes substantially more difficult, since polynomial optimization problems are in general NP-hard [17]. We can cite at least three closely related and recent frameworks designed to provide upper bounds of roundoff errors for non-linear programs. `FPTaylor` [25] is a tool based on Taylor-interval methods, while `Rosa` [5] combines SMT with interval arithmetic. `Real2Float` [19] relies on Putinar representations of positive polynomials while exploiting sparsity in a similar way as the second method that we propose in this paper.

We introduce two methods, coming from the field of polynomial optimization, to compute upper bounds on roundoff errors of polynomial programs. The first method is based on Bernstein expansions of polynomials, while the second relies on sparse Krivine-Stengle certificates for positive polynomials. In practice, these methods (presented in Section 3) provide accurate bounds at a reasonable computational cost. Indeed, the size of the Bernstein expansion used in the first method as well as the size of the LP relaxation problems considered in the second method are both linear w.r.t. the number of roundoff error variables.

1.1 Overview

Before explaining in detail each method, let us first illustrate the addressed problem on an example. Let f be the degree two polynomial defined by:

$$f(x) := x^2 - x \ , \quad \forall x \in X = [0,1].$$

When approximating the value of f at a given real number x, one actually computes the floating point result $\hat{f} = \hat{x} \otimes \hat{x} \ominus \hat{x}$, with all the real operators $+,-,\times$ being substituted by their associated floating point operators $\oplus$, $\ominus$, $\otimes$, and x being represented by the floating point number $\hat{x}$ (see Section 2.1 for more details on floating point arithmetics). A first simple rounding model consists of introducing an error term e_i for each floating point operation, as well as for each floating point variable. For instance, $\hat{x} \otimes \hat{x}$ corresponds to $((1+e_1)\,x\,(1+e_1)\,x)\,(1+e_2)$, where e_1 is the error term between x and $\hat{x}$, and e_2 is the one associated to the operation $\otimes$. Let $\mathbf{e}$ be the vector of all error terms e_i. Given $e_i \in [-\varepsilon, \varepsilon]$ for all i, with ε being the machine precision, we can write the floating point approximation $\hat{f}$ of f as follows:

$$\hat{f}(x,\mathbf{e}) = (((1+e_1)x(1+e_1)x)(1+e_2) - x(1+e_1))(1+e_3).$$

Then, the absolute roundoff error is defined by:

$$r(x,\mathbf{e}) := \max_{\substack{x\in[0,1]\\ \mathbf{e}\in[-\varepsilon,\varepsilon]^3}} \ (|\hat{f}(x,\mathbf{e}) - f(x)|) \ .$$

However, we can make this computation easier with a slight approximation: $|\hat{f}(x,\mathbf{e}) - f(x)| \leq |l(x,\mathbf{e})| + |h(x,\mathbf{e})|$ with $l(x,\mathbf{e})$ being the sum of the terms of

$(\hat{f}(x,\mathbf{e}) - f(x))$ which are linear in $\mathbf{e}$, and $h(x,\mathbf{e})$ the sum of the terms which are non-linear in $\mathbf{e}$. The term $|h(x,\mathbf{e})|$ can then be over-approximated by $O(|\mathbf{e}|^2)$ which is *in general* negligible compared to $|l(x,\mathbf{e})|$, and can be bounded using standard interval arithmetic. For this reason, we focus on computing an upper bound of $|l(x,\mathbf{e})|$. In the context of our example, $l(x,\mathbf{e})$ is given by:

$$l(x,\mathbf{e}) = (2x^2 - x)e_1 + x^2 e_2 + (x^2 - x)e_3. \tag{1}$$

We divide each error term e_j by ε, and then consider the (scaled) linear part $l' := \frac{l}{\varepsilon}$ of the roundoff error with the error terms $\mathbf{e} \in [-1,1]^3$. For all $x \in [0,1]$, and $\mathbf{e} \in [-1,1]^3$, one can easily compute a valid upper bound of $|l'(x,\mathbf{e})|$ with interval arithmetic. Surcharging the notation for elementary operations $+, -, \times$ in interval arithmetic, one has $l'(x,\mathbf{e}) \in ([-0.125,1] \times [-1,1] + [0,1] \times [-1,1] + [-0.25,0] \times [-1,1]) = [-2.25, 2.25]$, yielding $|l(x,\mathbf{e})| \leq 2.25\varepsilon$.
Using the first method based on Bernstein expansion detailed in Section 3.1, we compute 2ε as an upper bound of $|l(x,\mathbf{e})|$ in 0.23s using `FPBern(b)` a rational arithmetic implementation. With the second method based on sparse Krivine-Stengle representation detailed in Section 3.2, we also compute an upper bound of 2ε in 0.03s.
Although on this particular example, the method based the sparse Krivine-Stengle representation appears to be more time-efficient, in general the computational cost of the method based on Bernstein expansions is lower. For this example, the bounds provided by both methods are tighter than the ones determined by interval arithmetic. We emphasize the fact that the bounds provided by our two methods can be certified. Indeed, in the first case, the Bernstein coefficients (see Sections 2.2 and 3.1) can be computed either with rational arithmetic or certified interval arithmetic to ensure guaranteed values of upper bounds. In the second case, the positivity certificates are directly provided by sparse Krivine-Stengle representations.

1.2 Related Works

We first mention two tools, based on positivity certificates, to compute roundoff error bounds. The first tool, related to [3], relies on a similar approach to our second method. It uses dense Krivine-Stengle representations of positive polynomials to cast the initial problem as a finite dimensional LP problem. To reduce the size of this possibly large LP, [3] provides heuristics to eliminate some variables and constraints involved in the dense representation. However, this approach has the main drawback of loosing the property of convergence toward optimal solutions of the initial problem. Our second method uses sparse representations and is based on the previous works by [11] and [27], allowing to ensure the convergence towards optimal solutions while greatly reducing the computational cost of LP problems. Another tool, `Real2Float` [19], exploits sparsity in the same way while using Putinar representations of positive polynomials, leading to solving semidefinite (SDP) problems. Bounds provided by such SDP relaxations are in general more precise than LP relaxations [15], but the solving

cost is higher.
Several other tools are available to compute floating point roundoff errors. SMT solvers are efficient when handling linear programs, but often provide coarse bounds for non-linear programs, e.g. when the analysis is done in isolation [5]. The `Rosa` [5] tool is a solver mixing SMT and interval arithmetic which compiles functional `SCALA` programs implementing non-linear functions (involving $/, \sqrt{}$, and polynomials) as well as conditional statements. SMT solvers are theoretically able to output certificates which can be validated externally afterwards. `FPTaylor` tool [25], relies on *Symbolic Taylor expansion* method, which consists of a branch and bound algorithm based on interval arithmetic. Bernstein expansions have been extensively used to handle systems of polynomial equations [21,23] as well as systems of polynomial inequalities (including polynomial optimization), see for example [23,8,22]. Yet, to the best of our knowledge, there is no tool based on Bernstein expansions in the context of roundoff error computation. The `Gappa` tool provides certified bounds with elaborated interval arithmetic procedure relying on multiple-precision dyadic fractions. The static analysis tool `FLUCTUAT` [7] performs forward computation (by contrast with optimization) to analyze floating point `C` programs. Both `FLUCTUAT` and `Gappa` use a different rounding model (see Section 2.1), also available in `FPTaylor`, that we do not handle in our current implementation. Some tools also allow formal validation of certified bounds. `FPTaylor`, `Real2Float` [19], as well as `Gappa` [6] provide formal proof certificates, with `HOL-Light` [12] for the first case, and `Coq` [4] for the two other ones.

1.3 Key Contributions

Here is a summary of our key contributions:

- We present two new methods to compute upper bounds of floating point roundoff errors for programs implementing multivariate polynomial functions with input variables constrained to boxes. The first one is based on Bernstein expansions and the second one relies on sparse Krivine-Stengle representations. We also propose a theoretical framework to guarantee the validity of upper bounds computed with both methods (see Section 3). In addition, we give an alternative shorter proof in Section 2.3 for the existence of Krivine-Stengle representations for sparse positive polynomials (proof of Theorem 4).
- We release two software packages based on each method. The first one, called `FPBern`[3], computes the bounds using the Bernstein expansions, with two modules built on top of the software related to [8]: `FPBern(a)` is a `C++` module using double precision floating point arithmetic while `FPBern(b)` is a `Matlab` module using rational arithmetic. The second one `FPKriSten`[4] computes the bounds using Krivine-Stengle representations in `Matlab`. `FPKriSten` is built on top of the implementation related to [27].

[3] `https://github.com/roccaa/FPBern`
[4] `https://github.com/roccaa/FPKriSten`

- We compare our two methods implemented in `FPBern` and `FPKriSten` to three state-of-the-art methods. Our new methods have similar precision with the compared tools (`Real2Float, Rosa, FPTaylor`). At the same time, `FPBern(a)` shows an important time performance improvement, while `FPBern(b)` and `FPKriSten` has similar time performances compared with the other tools, yielding promising results.

The rest of the paper is organized as follows: in Section 2, we give basic background on floating point arithmetic, the Bernstein expansions and Krivine-Stengle representations. In Section 3 we give the main contributions, that is the computation of roundoff error bounds using Bernstein expansions and sparse Krivine-Stengle representations. Finally, in Section 4 we compare the performance and precision of our two methods with the existing tools, and show the advantages of our tools.

2 Preliminaries

We first recall useful notation on multivariate calculus. For $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ and the multi-index $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}^n$, we denote by $\mathbf{x}^{\boldsymbol{\alpha}}$ the product $\prod_{i=1}^n x_i^{\alpha_i}$. We also define $|\boldsymbol{\alpha}| = |\alpha_1| + \ldots + |\alpha_n|$, $\mathbf{0} = (0, \ldots, 0)$ and $\mathbf{1} = (1, \ldots, 1)$. The notation $\sum_{\boldsymbol{\alpha}}$ is the nested sum $\sum_{\alpha_1} \ldots \sum_{\alpha_n}$. Equivalently we have $\prod_{\boldsymbol{\alpha}}$ which is equal to the nested product $\prod_{\alpha_1} \ldots \prod_{\alpha_n}$.
Given another multi-index $\mathbf{d} = (d_1, \ldots, d_n) \in \mathbb{N}^n$, the inequality $\boldsymbol{\alpha} < \mathbf{d}$ (resp. $\boldsymbol{\alpha} \leq \mathbf{d}$) means that the inequality holds for each sub-index: $\alpha_1 < d_1, \ldots, \alpha_n < d_n$ (resp. $\alpha_1 \leq d_1, \ldots, \alpha_n \leq d_n$). Moreover, the binomial coefficient $\binom{\mathbf{d}}{\boldsymbol{\alpha}}$ is the product $\prod_{i=1}^n \binom{d_i}{\alpha_i}$.
Let $\mathbb{R}[\mathbf{x}]$ be the vector space of multivariate polynomials. Given $f \in \mathbb{R}[\mathbf{x}]$, we associate a *multi-degree* $\mathbf{d} = (d_1, \ldots, d_n)$ to f, with each d_i standing for the degree of f with respect to the variable x_i. Then, we can write $f(\mathbf{x}) = \sum_{\boldsymbol{\gamma} \leq \mathbf{d}} a_{\boldsymbol{\gamma}} \mathbf{x}^{\boldsymbol{\gamma}}$, with $a_{\boldsymbol{\gamma}}$ (also noted $(f)_{\boldsymbol{\gamma}}$) being the coefficients of f in the monomial basis and each $\boldsymbol{\gamma} \in \mathbb{N}^n$ is a multi-index. The degree d of f is given by $d := \max_{\{\boldsymbol{\gamma} : a_{\boldsymbol{\gamma}} \neq 0\}} |\boldsymbol{\gamma}|$. As an example, if $f(x_1, x_2) = x_1^4 x_2 + x_1^1 x_2^3$ then $\mathbf{d} = (4, 3)$ and $d = 5$. For the polynomial l used in Section 1.1, one has $\mathbf{d} = (2, 1, 1, 1)$ and $d = 3$.

2.1 Floating Point arithmetic

This section gives background on floating point arithmetic, inspired from material available in [25, Section 3]. The IEEE754 standard [28] defines a binary floating point number as a triple significant, sign, and exponent (denoted by sig, sgn, exp) which represents the numerical value of $(-1)^{sgn} \times sig \times 2^{exp}$. The standard describes 3 formats (32, 64, and 128 bits) which vary by the size of the significant and the exponent, as well as special values (such as NaN, the infinities). Denoting by $\mathbb{F}$ the set of floating point numbers, we call rounding operator, the function $\mathrm{rnd} : \mathbb{R} \to \mathbb{F}$ which takes a real number and returns the

closest floating point number rounded to the nearest, toward zero, or toward $\pm\infty$. A simple model of rounding is given by the following formula:

$$\mathrm{rnd}(x) = x(1+e) + u,$$

with $|e| \leq \varepsilon$, $|u| \leq \mu$ and $eu = 0$. The value ε is the maximal relative error (given by the machine precision [28]), and μ is the maximal absolute error for numbers very close to 0. For example, in the single (32 bits) format, ε is equal to 2^{-24} while μ equals 2^{-150}. It is clear that in general μ is negligible compared to ε, thus we neglect terms depending on u in the remainder of this paper.
Given an operation $\mathrm{op} : \mathbb{R}^n \to \mathbb{R}$, let $\mathrm{op}_{\mathrm{FP}}$ be the corresponding floating point operation. An operation is exactly rounded when $\mathrm{op}_{\mathrm{FP}}(\mathbf{x}) = \mathrm{rdn}(\mathrm{op}(\mathbf{x}))$, for all $\mathbf{x} \in \mathbb{R}^n$.
In the IEEE754 standard the following operations are defined as exactly rounded: $+, -, \times, /, \sqrt{}$, and the `fma` operation[5]. It follows that for those operations we have the continuation of the simple rounding model $\mathrm{op}_{\mathrm{FP}}(\mathbf{x}) = \mathrm{op}(\mathbf{x})(1+e)$.
The previous rounding model is called "simple" in contrast with more improved rounding model. Given the function $\mathrm{pc}(x) = \max_{k\in\mathbb{Z}}\{2^k : 2^k < x\}$, then the improved rounding model is defined by: $\mathrm{op}_{\mathrm{FP}}(\mathbf{x}) = \mathrm{op}(\mathbf{x}) + \mathrm{pc}(\mathrm{op}(\mathbf{x}))$, for all $\mathbf{x} \in \mathbb{R}^n$. As the function pc is piecewise constant, this rounding model needs design of algorithms based on successive subdivisions, which is not currently handled in our methods. Combining branch and bound algorithms with interval arithmetic is adapted to roundoff error computation with such rounding model, which is the case with `FLUCTUAT`[7], `Gappa`[6], and `FPTaylor` [25].

2.2 Bernstein Expansion of Polynomials

In this section we give mandatory background on the Bernstein expansion for the contribution detailed in Section 3.1. Given a multivariate polynomial $f \in \mathbb{R}[\mathbf{x}]$, we recall how to compute a lower bound of $\underline{f}^* := \min_{\mathbf{x}\in[0,1]^n} f(\mathbf{x})$. The next result can be retrieved in [9, Theorem 2]:

Theorem 1 (Multivariate Bernstein expansion). *Given a multivariate polynomial f and a degree $\mathbf{k} \geq \mathbf{d}$ with $\mathbf{d}$ the multi-degree of f, then the Bernstein expansion of multi-degree $\mathbf{k}$ of f is given by:*

$$f(\mathbf{x}) = \sum_{\boldsymbol{\gamma}} a_{\boldsymbol{\gamma}} \mathbf{x}^{\boldsymbol{\gamma}} = \sum_{\boldsymbol{\alpha}\leq\mathbf{k}} b_{\boldsymbol{\alpha}}^{(f)} \mathbf{B}_{\mathbf{k},\boldsymbol{\alpha}}(\mathbf{x}). \tag{2}$$

where $b_{\boldsymbol{\alpha}}^{(f)}$ (also denoted by $b_{\boldsymbol{\alpha}}$ when there is no confusion) are the Bernstein coefficients (of multi-degree $\mathbf{k}$) of f, and $\mathbf{B}_{\mathbf{k},\boldsymbol{\alpha}}(\mathbf{x})$ are the Bernstein basis polynomials defined by $\mathbf{B}_{\mathbf{k},\boldsymbol{\alpha}}(\mathbf{x}) := \prod_{i=1}^n B_{k_i,\alpha_i}(x_i)$ and $B_{k_i,\alpha_i}(x_i) := \binom{k_i}{\alpha_i} x_i^{\alpha_i}(1-x_i)^{k_i-\alpha_i}$. The Bernstein coefficients are given by the following formulas:

$$b_{\boldsymbol{\alpha}} = \sum_{\boldsymbol{\beta}<\boldsymbol{\alpha}} \frac{\binom{\boldsymbol{\alpha}}{\boldsymbol{\beta}}}{\binom{\mathbf{k}}{\boldsymbol{\beta}}} a_{\boldsymbol{\beta}}, \quad \mathbf{0} \leq \boldsymbol{\alpha} \leq \mathbf{k}. \tag{3}$$

[5] The `fma` operator is defined by $\mathtt{fma}(x,y,z)$=$x \times y + z$.

The Bernstein expansion having numerous properties, we give only four of them which are useful for Section 3.1. For a more exhaustive introduction to Bernstein expansion, as well as some proof of the basic properties, we refer the interested reader to [23].

Property 1 (Cardinality [23, (3.14)]). The number of Bernstein coefficients in the Bernstein expansion (of multi-degree $\mathbf{k}$) is equal to $(\mathbf{k}+\mathbf{1})^{\mathbf{1}} = \prod_{i=1}^{n}(k_i+1)$.

Property 2 (Linearity [23, (3.2.3)]). Given two polynomials p_1 and p_2, one has:

$$b_{\boldsymbol{\alpha}}^{(cp_1+p_2)} = cb_{\boldsymbol{\alpha}}^{(p_1)} + b_{\boldsymbol{\alpha}}^{(p_2)} \ , \quad \forall c \in \mathbb{R},$$

where Bernstein expansions with same multi-degrees are considered.

Property 3 (Enclosure [23, (3.2.4)]). The minimum (resp. maximum) of a polynomial f over $[0,1]^n$ can be lower bounded (resp. upper bounded) by the minimum (resp. maximum) of its Bernstein coefficients:

$$\min_{\boldsymbol{\alpha}\leq\mathbf{k}} b_{\boldsymbol{\alpha}} \leq f(\mathbf{x}) \leq \max_{\boldsymbol{\alpha}\leq\mathbf{k}} b_{\boldsymbol{\alpha}}, \quad \forall \mathbf{x} \in [0,1]^n \quad .$$

Property 4 (Sharpness [23, (3.2.5)]). If the minimum (resp. maximum) of the $b_{\boldsymbol{\alpha}}$ is reached for $\boldsymbol{\alpha}$ in a corner of the box $[0,k_1]\times\cdots\times[0,k_n]$, then $b_{\boldsymbol{\alpha}}$ is the minimum (resp. maximum) of f over $[0,1]^n$.

Property 1 gives the maximal computational cost needed to find a lower bound of f^* for a Bernstein expansion of fixed multi-degree $\mathbf{k}$. Property 3 is used to bound from below optimal values, while Property 4 allows to determine if the lower bound is optimal.

2.3 Dense and Sparse Krivine-Stengle Representations

In this section, we first give the necessary background on Krivine-Stengle representations, used in the context of polynomial optimization. Then, we present a sparse version based on [11]. These notions are applied later in Section 3.2.

Dense Krivine-Stengle representations. Krivine-Stengle certificates for positive polynomials can first be found in [14,26] (see also [16, Theorem 1(b)]). Such certificates give representations of positive polynomials over a set $\mathbf{K} = \{\mathbf{x}\in\mathbb{R}^n : 0 \leq g_i(\mathbf{x}) \leq 1,\ i=1,\ldots,p\}$, with $g_1,\ldots,g_p \in \mathbb{R}[\mathbf{x}]$. The compact set $\mathbf{K}$ is a basic semialgebraic set, since it is defined as a conjunction of polynomial inequalities.
Given $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_p)$ and $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_p)$, let us define the polynomial $h_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbf{x}) = \mathbf{g}^{\boldsymbol{\alpha}}(\mathbf{1}-\mathbf{g})^{\boldsymbol{\beta}} = \prod_{i=1}^{p} g_i^{\alpha_i}(1-g_i)^{\beta_i}$.
For instance on the two-dimensional unit box, one has $n=p=2$, $\mathbf{K} = [0,1]^2 = \{\mathbf{x}\in\mathbb{R}^2 : 0\leq x_1\leq 1\,,\ 0\leq x_2\leq 1\}$. With $\boldsymbol{\alpha}=(2,1)$ and $\boldsymbol{\beta}=(1,3)$, one has $h_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbf{x}) = x_1^2x_2(1-x_1)(1-x_2)^3$.

Theorem 2 (Dense Krivine-Stengle representations). *Let $\psi \in \mathbb{R}[\mathbf{x}]$ be a positive polynomial over $\mathbf{K}$. Then there exist $k \in \mathbb{N}$ and a finite number of nonnegative weights $\lambda_{\boldsymbol{\alpha},\boldsymbol{\beta}} \geq 0$ such that:*

$$\psi(\mathbf{x}) = \sum_{|\boldsymbol{\alpha}+\boldsymbol{\beta}|\leq k} \lambda_{\boldsymbol{\alpha},\boldsymbol{\beta}} h_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbf{x}), \quad \forall \mathbf{x} \in \mathbb{R}^n. \tag{4}$$

It is possible to compute the weights $\lambda_{\boldsymbol{\alpha},\boldsymbol{\beta}}$ by identifying in the monomial basis the coefficients of the polynomials in the left and right sides of (4). Denoting by $(\psi)_{\boldsymbol{\gamma}}$ the monomial coefficients of ψ, with $\boldsymbol{\gamma} \in \mathbb{N}^n_k := \{\boldsymbol{\gamma} \in \mathbb{N}^n : |\boldsymbol{\gamma}| \leq k\}$, the $\lambda_{\boldsymbol{\alpha},\boldsymbol{\beta}}$ fulfill the following equalities:

$$\psi_{\boldsymbol{\gamma}} = \sum_{|\boldsymbol{\alpha}+\boldsymbol{\beta}|\leq k} \lambda_{\boldsymbol{\alpha},\boldsymbol{\beta}} (h_{\boldsymbol{\alpha},\boldsymbol{\beta}})_{\boldsymbol{\gamma}}, \quad \forall \boldsymbol{\gamma} \in \mathbb{N}^n_k. \tag{5}$$

Global optimization using the dense Krivine-Stengle representations. Here we consider the polynomial minimization problem $\underline{f}^* := \min_{\mathbf{x}\in\mathbf{K}} f(\mathbf{x})$, with f a polynomial of degree d. We can rewrite this problem as the following infinite dimensional problem:

$$\begin{aligned} \underline{f}^* := \max_{t\in\mathbb{R}} \quad & t, \\ \text{s.t.} \quad & f(\mathbf{x}) - t \geq 0\,, \quad \forall \mathbf{x} \in \mathbf{K}. \end{aligned} \tag{6}$$

The idea is to look for a hierarchy of finite dimensional linear programming (LP) relaxations by using Krivine-Stengle representations of the positive polynomial $\psi = f - t$ involved in Problem (6). Applying Theorem 2 to this polynomial, we obtain the following LP problem for each $k \geq d$:

$$\begin{aligned} p^*_k := \max_{t,\lambda_{\boldsymbol{\alpha},\boldsymbol{\beta}}} \quad & t, \\ \text{s.t} \quad & (f-t)_{\boldsymbol{\gamma}} = \sum_{|\boldsymbol{\alpha}+\boldsymbol{\beta}|\leq k} \lambda_{\boldsymbol{\alpha},\boldsymbol{\beta}} (h_{\boldsymbol{\alpha},\boldsymbol{\beta}})_{\boldsymbol{\gamma}}\,, \quad \forall \boldsymbol{\gamma} \in \mathbb{N}^n_k\,, \\ & \lambda_{\boldsymbol{\alpha},\boldsymbol{\beta}} \geq 0. \end{aligned} \tag{7}$$

As in [16, (4)], one has:

Theorem 3 (Dense Krivine-Stengle LP relaxations). *The sequence of optimal values (p^*_k) satisfies $p^*_k \to \underline{f}^*$ as $k \to +\infty$. Moreover each p^*_k is a lower bound of $\underline{f}^*$.*

At fixed k, the total number of variables of Problem (7) is given by the number of $\lambda_{\boldsymbol{\alpha},\boldsymbol{\beta}}$ and t, that is $\binom{2n+k}{k} + 1$. The number of constraints is equal to the cardinality of $\mathbb{N}^n_k$, which is $\binom{n+k}{k}$.

Sparse Krivine-Stengle representations. We now explain how to derive less computationally expensive LP relaxations, by relying on sparse Krivine-Stengle representations. For $I \subseteq \{1,\ldots,n\}$, let $\mathbb{R}[\mathbf{x}, I]$ be the ring of polynomials restricted to the variables $\{x_i \ : \ i \in I\}$. We borrow the notion of a sparsity pattern from [27, Assumption 1]:

Definition 1 (Sparsity Pattern). *Given $m \in \mathbb{N}$, $I_j \subseteq \{1,\ldots,n\}$, and $J_j \subseteq \{1,\ldots,p\}$ for all $j = 1,\ldots,m$, a sparsity pattern is defined by the four following conditions:*

- *f can be written as: $f = \sum_{j=1}^m f_j$ with $f_j \in \mathbb{R}[\mathbf{x}, I_j]$,*
- *$g_i \in \mathbb{R}[\mathbf{x}, I_j]$ for all $i \in J_j$, for all $j = 1,\ldots,m$,*
- *$\bigcup_{j=1}^m I_j = \{1,\ldots,n\}$ and $\bigcup_{j=1}^m J_j = \{1,\ldots,p\}$,*
- *(Running Intersection Property) for all $j = 1,\ldots,m-1$, there exists $s \leq j$ s.t. $I_{j+1} \cap \bigcup_{i=1}^j I_i \subseteq I_s$.*

As an example, the four conditions stated in Definition 1 are satisfied while considering $f(\mathbf{x}) = x_1x_2 + x_1^2x_3$ on the hypercube $\mathbf{K} = [0,1]^3$. Indeed, one has $f_1(\mathbf{x}) = x_1x_2 \in \mathbb{R}[\mathbf{x}, I_1]$, $f_2(\mathbf{x}) = x_1^2x_3 \in \mathbb{R}[\mathbf{x}, I_2]$ with $I_1 = \{1,2\}$, $I_2 = \{1,3\}$. Taking $J_1 = I_1$ and $J_2 = I_2$, one has $g_i = x_i \in \mathbb{R}[\mathbf{x}, I_j]$ for all $i \in I_j$, $j = 1,2$.

Let us consider a given sparsity pattern as stated above. By noting $n_j = |I_j|$, $p_j = |J_j|$, then the set $\mathbf{K} = \{\mathbf{x} \in \mathbb{R}^n : 0 \leq g_i(\mathbf{x}) \leq 1,\ i = 1,\ldots,p\}$ yields subsets $\mathbf{K}_j = \{\mathbf{x} \in \mathbb{R}^{n_j} : 0 \leq g_i(\mathbf{x}) \leq 1,\ i \in J_j\}$, with $j = 1,\ldots,m$. If $\mathbf{K}$ is a compact subset of $\mathbb{R}^n$ then each $\mathbf{K}_j$ is a compact subset of $\mathbb{R}^{n_j}$. As in the dense case, let us note $h_{\boldsymbol{\alpha}_j,\boldsymbol{\beta}_j} := \mathbf{g}^{\boldsymbol{\alpha}_j}(\mathbf{1}-\mathbf{g})^{\boldsymbol{\beta}_j}$, for given $\boldsymbol{\alpha}_j, \boldsymbol{\beta}_j \in \mathbb{N}^{n_j}$.

The following result, a sparse variant of Theorem 2, can be retrieved from [27, Theorem 1] but we also provide here a shorter alternative proof by using [11].

Theorem 4 (Sparse Krivine-Stengle representations). *Let $f, g_1, \ldots, g_p \in \mathbb{R}[\mathbf{x}]$ be given and assume that there exist I_j and J_j, $j = 1,\ldots,m$, which satisfy the four conditions stated in Definition 1. If f is positive over $\mathbf{K}$, then there exist $\phi_j \in \mathbb{R}[\mathbf{x}, I_j]$, $j = 1,\ldots,m$ such that $f = \sum_{j=1}^m \phi_j$ and $\phi_j > 0$ over $\mathbf{K}_j$. In addition, there exist $k \in \mathbb{N}$ and finitely many nonnegative weights $\lambda_{\boldsymbol{\alpha}_j,\boldsymbol{\beta}_j}$, $j = 1,\ldots,m$, such that:*

$$\phi_j = \sum_{|\boldsymbol{\alpha}_j+\boldsymbol{\beta}_j| \leq k} \lambda_{\boldsymbol{\alpha}_j,\boldsymbol{\beta}_j} h_{\boldsymbol{\alpha}_j,\boldsymbol{\beta}_j}\ , \quad j = 1,\ldots,m. \tag{8}$$

Proof. From [11, Lemma 3], there exist $\phi_j \in \mathbb{R}[\mathbf{x}, I_j]$ such that $f = \sum_{j=1}^m \phi_j$ and $\phi_j > 0$ on $\mathbf{K}_j$. Applying Theorem 2 on each ϕ_j, there exist $k_j \in \mathbb{N}$ and finitely many nonnegative weights $\lambda_{\boldsymbol{\alpha}_j,\boldsymbol{\beta}_j}$ such that $\phi_j = \sum_{|\boldsymbol{\alpha}_j+\boldsymbol{\beta}_j| \leq k_j} \lambda_{\boldsymbol{\alpha}_j,\boldsymbol{\beta}_j} h_{\boldsymbol{\alpha}_j,\boldsymbol{\beta}_j}$. With $k = \max_{1\leq j\leq m}\{k_j\}$, we complete the representations with as many zero λ as necessary, we obtain the desired result. □

In Theorem 4, one assumes that f can be written as the sum $f = \sum_{j=1}^m f_j$, where each f_j is not necessarily positive. The first result of the theorem states that that f can be written as another sum $f = \sum_{j=1}^m \phi^j$, where each ϕ_j is now positive. As in the dense case, the $\lambda_{\boldsymbol{\alpha}_j,\boldsymbol{\beta}_j}$ can be computed by equalizing the coefficients in the monomial basis. We also obtain a hierarchy of LP relaxations to approximate the solution of polynomial optimization problems. For the sake of conciseness, we only provide these relaxations as well as their computational costs in the particular context of roundoff error bounds in Section 3.2.

3 Two new methods to compute roundoff errors bounds

This section is dedicated to our main contributions. We provide two new methods to compute absolute roundoff error bounds using either Bernstein expansions or sparse Krivine-Stengle representations. Here we consider a given program which implements a polynomial expression f with input variables $\mathbf{x}$ satisfying a set of input constraints encoded by $\mathbf{X}$. We restrict ourselves to the case where $\mathbf{X}$ is the unit box $[0,1]^n$.

Following the simple rounding model described in Section 2.1, we note $\hat{f}(\mathbf{x},\mathbf{e})$ the rounded expression of f after introduction of the rounding variables $\mathbf{e}$ (one additional variable is introduced for each real variable x_i or constant as well as for each arithmetic operation $+,\times$ or $-$). For a given machine epsilon ε, these error variables also satisfy a set of constraints encoded by the box $[-\varepsilon,\varepsilon]^m$. As explained in [19, Section 3.1], we can decompose the roundoff error as follows: $r(\mathbf{x},\mathbf{e}) := \hat{f}(\mathbf{x},\mathbf{e}) - f(\mathbf{x}) = l(\mathbf{x},\mathbf{e}) + h(\mathbf{x},\mathbf{e})$, where $l(\mathbf{x},\mathbf{e}) := \sum_{j=1}^m \frac{\partial r(\mathbf{x},\mathbf{e})}{\partial e_j}(\mathbf{x},0)e_j = \sum_{j=1}^m s_j(\mathbf{x})e_j$. One obtains an enclosure of h using interval arithmetic to bound second-order error terms in the Taylor expansion of r w.r.t. $\mathbf{e}$ (as in [25,19]).

We note d the degree of l. After dividing each error variable e_j by ε, we now consider the optimization of the (scaled) linear part $l' := l/\varepsilon$ of the roundoff error. In other words, we focus on computing upper bounds of the maximal absolute value $l'^* := \max_{(\mathbf{x},\mathbf{e})\in\mathbf{X}\times\mathbf{E}} |l'(\mathbf{x},\mathbf{e})|$ where $\mathbf{E} = [-1,1]^m$.

3.1 Bernstein expansions of roundoff errors

The first method is the approximation of l'^* with the Bernstein expansions. Let $\mathbf{d}$ be the multi-degree of $l'_{\mathbf{e}}$. From the above definition of l, note that $\mathbf{d}$ is also the multi-degree of f. For each $\mathbf{k} \geq \mathbf{d}$, let us note $\overline{l'_{\mathbf{k}}} := \max_{\boldsymbol{\alpha}\leq\mathbf{k}} \sum_{j=1}^m |b_{\boldsymbol{\alpha}}^{(s_j)}|$ and $\underline{l'_{\mathbf{k}}} := -\overline{l'_{\mathbf{k}}}$. Our procedure is based on the following lemma:

Lemma 1. *For each $\mathbf{k} \geq \mathbf{d}$, the polynomial $l'(\mathbf{x},\mathbf{e})$ can be bounded as follows:*

$$\underline{l'_{\mathbf{k}}} \leq l'(\mathbf{x},\mathbf{e}) \leq \overline{l'_{\mathbf{k}}}\,, \quad \forall (\mathbf{x},\mathbf{e}) \in \mathbf{X}\times\mathbf{E}\,. \tag{9}$$

Proof. We write $l'_{\mathbf{e}} \in \mathbb{R}[\mathbf{x}]$ the polynomial $l'(\mathbf{x},\mathbf{e})$ for a given $\mathbf{e}\in\mathbf{E}$. Property 3 provides the enclosure of $l'_{\mathbf{e}}(\mathbf{x})$ w.r.t. $\mathbf{x}$ for a given $\mathbf{e}\in\mathbf{E}$:

$$\min_{\boldsymbol{\alpha}\leq\mathbf{k}} b_{\boldsymbol{\alpha}}^{(l'_{\mathbf{e}})} \leq l'_{\mathbf{e}}(\mathbf{x}) \leq \max_{\boldsymbol{\alpha}\leq\mathbf{k}} b_{\boldsymbol{\alpha}}^{(l'_{\mathbf{e}})}\,, \quad \forall \mathbf{x}\in[0,1]^n\,, \tag{10}$$

where each Bernstein coefficient satisfies $b_{\boldsymbol{\alpha}}^{(l'_{\mathbf{e}})} = \sum_{j=1}^m e_j b_{\boldsymbol{\alpha}}^{(s_j)}$ by Property 2 (each e_j being a scalar in $[-1,1]$). The proof of the left inequality comes from:

$$\begin{aligned}\min_{\mathbf{e}\in[-1,1]^m} \big(\min_{\boldsymbol{\alpha}\leq\mathbf{k}} (\sum_{j=1}^m e_j b_{\boldsymbol{\alpha}}^{(s_j)})\big) &= \min_{\boldsymbol{\alpha}\leq\mathbf{k}} \big(\min_{\mathbf{e}\in[-1,1]^m} (\sum_{j=1}^m e_j b_{\boldsymbol{\alpha}}^{(s_j)})\big) \\ &= \min_{\boldsymbol{\alpha}\leq\mathbf{k}} \sum_{j=1}^m -|b_{\boldsymbol{\alpha}}^{(s_j)}| = -\max_{\boldsymbol{\alpha}\leq\mathbf{k}} \sum_{j=1}^m |b_{\boldsymbol{\alpha}}^{(s_j)}|\quad .\end{aligned}$$

The proof of the right inequality is similar. □

Remark 1. The computational cost of $\underline{l'_{\mathbf{k}}}$ is $m(\mathbf{k}+\mathbf{1})^{\mathbf{1}}$ since we need to compute the Bernstein coefficients for each $s_j(\mathbf{x})$. This cost is polynomial in the degree and exponential in n but is linear in m. In the implementation described in Section 4, we first compute each $b_{\boldsymbol{\alpha}}^{(l'_{\mathbf{e}})}$ as a function of $\mathbf{e}$ and then optimize afterwards over $[-1,1]^m$.

Example 1. For the polynomial l defined in (1) (Section 1.1), one has $l(x,\mathbf{e}) = (2x^2-x)e_1 + x^2e_2 + (x^2-x)e_3$. Applying the above method with $\mathbf{k}=\mathbf{d}=2$, one considers the following Bernstein coefficients:

$$b_0^{(l'_{\mathbf{e}})} = 0, \quad b_1^{(l'_{\mathbf{e}})} = -\frac{e_1}{2} - \frac{e_3}{2}, \quad b_2^{(l'_{\mathbf{e}})} = e_1 + e_2.$$

The number of Bernstein coefficients w.r.t. x is 3, which is much lower than the one w.r.t. $(x,\mathbf{e})$, which is equal to 24. One can obtain an upper bound (resp. lower bound) by taking the maximum (resp. minimum) of the Bernstein coefficients. In this case, $\max_{\mathbf{e}\in[-1,1]^3} b_1^{(l'_{\mathbf{e}})} = 0$, $\max_{\mathbf{e}\in[-1,1]^3} b_2^{(l'_{\mathbf{e}})} = 1$ and $\max_{\mathbf{e}\in[-1,1]^3} b_3^{(l'_{\mathbf{e}})} = 2$. Thus, one obtains $\overline{l'_{\mathbf{k}}} = 2$ as an upper bound of l'^* yielding $l^* \leq 2\varepsilon$.

3.2 Sparse Krivine-Stengle representations of roundoff errors

Here we explain how to compute lower bounds of $\underline{l'} := \min_{(\mathbf{x},\mathbf{e})\in\mathbf{X}\times\mathbf{E}} l'(\mathbf{x},\mathbf{e})$ by using sparse Krivine-Stengle representations. We obtain upper bounds of $\overline{l'} := \max_{(\mathbf{x},\mathbf{e})\in\mathbf{X}\times\mathbf{E}} l'(\mathbf{x},\mathbf{e})$ in a similar way.

For the sake of consistency with Section 2.3, we introduce the variable $\mathbf{y} \in \mathbb{R}^{n+m}$ defined by $y_j := x_j$, $j = 1,\ldots,n$ and $y_j := e_{j-n}$, $j = n+1,\ldots,n+m$. Then, one can write the set $\mathbf{K} = \mathbf{X}\times\mathbf{E}$ as follows:

$$\mathbf{K} = \{\mathbf{y}\in\mathbb{R}^{n+m} : 0 \leq g_j(\mathbf{y}) \leq 1\,, \quad j = 1,\ldots,n+m\}\,, \tag{11}$$

with $g_j(\mathbf{y}) := x_j$, for each $j = 1,\ldots,n$ and $g_j(\mathbf{y}) := \frac{1}{2} + \frac{e_j}{2}$, for each $j = n+1,\ldots,n+m$.

Lemma 2. *For each $j = 1,\ldots,m$, let us define $I_j := \{1,\ldots,n,n+j\}$ and $J_j := I_j$. Then the sets I_j and J_j satisfy the four conditions stated in Definition 1.*

Proof. The first condition holds as $l'(\mathbf{y}) = l'(\mathbf{x},\mathbf{e}) = \sum_{j=1}^m s_j(\mathbf{x},\mathbf{e})e_j = \sum_{j=1}^m s_j(\mathbf{y})e_j$, with $s_j(\mathbf{y}) \in \mathbb{R}[\mathbf{y}, I_j]$. The second and third condition are obvious. The running intersection property comes from $I_{j+1}\cap I_j = \{1,\ldots,n\} \subseteq I_j$. □

Given $\boldsymbol{\alpha},\boldsymbol{\beta} \in \mathbb{N}^{n+1}$, one can write $\boldsymbol{\alpha} = (\boldsymbol{\alpha}',\gamma)$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}',\delta)$, for $\boldsymbol{\alpha}',\boldsymbol{\beta}' \in \mathbb{N}^n$, $\gamma,\delta \in \mathbb{N}$. In our case, this gives the following formulation for the polynomial $h_{\boldsymbol{\alpha}_j,\boldsymbol{\beta}_j}(\mathbf{y}) = \mathbf{g}^{\boldsymbol{\alpha}_j}(\mathbf{1}-\mathbf{g})^{\boldsymbol{\beta}_j}$:

$$h_{\boldsymbol{\alpha}_j,\boldsymbol{\beta}_j}(\mathbf{y}) = h_{\boldsymbol{\alpha}'_j,\boldsymbol{\beta}'_j,\gamma_j,\delta_j}(\mathbf{x},\mathbf{e}) = \mathbf{x}^{\boldsymbol{\alpha}'_j}(\mathbf{1}-\mathbf{x})^{\boldsymbol{\beta}'_j}(\frac{1}{2}+\frac{e_j}{2})^{\gamma_j}(\frac{1}{2}-\frac{e_j}{2})^{\delta_j}\,.$$

For instance, with the polynomial l' considered in Section 1.1 and depending on x, e_1, e_2, e_3, one can consider the multi-indices $\boldsymbol{\alpha}_1 = (1,2), \boldsymbol{\beta}_1 = (2,3)$ associated to the roundoff variable e_1. Then $h_{\boldsymbol{\alpha}_1,\boldsymbol{\beta}_1}(\mathbf{y}) = x(1-x)^2(\frac{1}{2}+\frac{e_1}{2})^2(\frac{1}{2}-\frac{e_1}{2})^3$.

Now, we consider the following hierarchy of LP relaxations, for each $k \geq d$:

$$\begin{aligned} \underline{l'_k} := \max_{t,\lambda_{\boldsymbol{\alpha}_j,\boldsymbol{\beta}_j}} \quad & t\,, \\ \text{s.t} \quad & l' - t = \sum_{j=1}^{m} \phi_j\,, \\ & \phi_j = \sum_{|\boldsymbol{\alpha}_j+\boldsymbol{\beta}_j|\leq k} \lambda_{\boldsymbol{\alpha}_j,\boldsymbol{\beta}_j} h_{\boldsymbol{\alpha}_j,\boldsymbol{\beta}_j}\,, \quad j = 1,\ldots,m\,, \\ & \lambda_{\boldsymbol{\alpha}_j,\boldsymbol{\beta}_j} \geq 0\,, \quad j = 1,\ldots,m\,. \end{aligned} \tag{12}$$

Similarly, we obtain $\overline{l'_k}$ while replacing max by min and $l'-t$ by $t-l'$ in LP (12).

Lemma 3. *The sequence of optimal values $(\underline{l'_k})$ (resp. $(\overline{l'_k})$) satisfies $\underline{l'_k} \uparrow \underline{l'}$ (resp. $\overline{l'_k} \downarrow \overline{l'}$) as $k \to +\infty$. In addition, $l'_k := \max\{|\underline{l'_k}|, |\overline{l'_k}|\} \to l'^*$ as $k \to +\infty$.*

Proof. By construction $(\underline{l'_k})$ is monotone nondecreasing. For a given arbitrary $\varepsilon' > 0$, the polynomial $l' - \underline{l'} + \varepsilon'$ is positive over $\mathbf{K}$. By Lemma 2, the subsets I_j and J_j satisfy the four conditions stated in Definition 1, so we can apply Theorem 4 to $l' - \underline{l'} + \varepsilon'$. This yields the existence of ϕ_j, $j = 1,\ldots,m$, such that $l' - \underline{l'} + \varepsilon' = \sum_{j=1}^m \phi_j$ and $\phi_j = \sum_{|\boldsymbol{\alpha}_j+\boldsymbol{\beta}_j|\leq k} \lambda_{\boldsymbol{\alpha}_j,\boldsymbol{\beta}_j} h_{\boldsymbol{\alpha}_j,\boldsymbol{\beta}_j}$, $j = 1,\ldots,m$. Hence, $(\underline{l'} - \varepsilon', \phi_j, \lambda_{\boldsymbol{\alpha}_j,\boldsymbol{\beta}_j})$ is feasible for LP (12). It follows that there exists k such that $\underline{l'_k} \geq l' - \varepsilon'$. Since $\underline{l'_k} \leq l'$, and ε' has been arbitrary chosen, we obtain the convergence result for the sequence $(\underline{l'_k})$. The proof is analogous for $(\overline{l'_k})$ and yields $\max\{|\underline{l'_k}|, |\overline{l'_k}|\} \to \max\{|\underline{l'}|, |\overline{l'}|\} = l'^*$ as $k \to +\infty$, the desired result. □

Remark 2. In the special case of roundoff error computation, one can prove that the number of variables of LP (12) is $m\binom{2(n+1)+k}{k} + 1$ with a number of constraints equal to $[\frac{mk}{n+1}+1]\binom{n+k}{k}$. This is in contrast with the dense case where the number of LP variables is $\binom{2(n+m)+k}{k} + 1$ with a number of constraints equal to $\binom{n+m+k}{k}$.

Proof of Remark 2. As we replace a function ϕ of dimension $(n+m)$ by a sum of m functions ϕ_j of dimension $(n+1)$, the number of coefficients $\lambda_{\boldsymbol{\alpha}_j,\boldsymbol{\beta}_j}$ is $m\binom{2(n+1)+k}{k}$. This leads to a total of $m\binom{2(n+1)+k}{k} + 1$ variables when adding t.

The number of equality constraints is the number of monomials involved in $\sum_{j=1}^m \phi_j$. Each ϕ_j has $\binom{(n+1)+k}{k}$ monomials. However there are redundant monomials between all the ϕ_j: the ones depending of only $\mathbf{x}$, and not $\mathbf{e}$. These $\binom{n+k}{k}$ monomials should appear only once. This leads to a final number of $m\binom{(n+1)+k}{k} - (m-1)\binom{n+k}{k}$ monomials which is equal to $[\frac{mk}{n+1}+1]\binom{n+k}{k}$. □

Example 2. Continuing Example 1, for the polynomial l defined in (1) (Section 1.1), we consider LP (12) at the relaxation order $k = d = 3$. This problem involves $3\binom{2\times(1+1)+3}{3} + 1 = 106$ variables and $[\frac{3\times 3}{2} + 1]\binom{4}{3} = 22$ constraints. This is in contrast with a dense Krivine-Stengle representation, where the corresponding LP involves 35 linear equalities and 166 variables. Computing the values of $\underline{l'_k}$ and $\overline{l'_k}$ provides an upper bound of 2 for l'^*, yielding $l^* \leq 2\varepsilon$.

4 Implementation & Results

The `FPBern` and `FPKriSten` software packages. We provide two distinct software packages to compute certified error bounds of roundoff errors for programs implementing polynomial functions with floating point precision. The first tool `FPBern` relies on the method from Section 3.1 and the second tool `FPKriSten` on the method from Section 3.2.

`FPBern` is built on top of the software presented in [8] to manipulate Bernstein expansions, which includes a `C++` module `FPBern(a)` and a `Matlab` module `FPBern(b)`. Their main difference is that Bernstein coefficients are computed with double precision floating point arithmetic in `FPBern(a)` and with rational arithmetic in `FPBern(b)`. Polynomial operations are handled with GiNaC [2] in `FPBern(a)` and with `Matlab Symbolic Toolbox` in `FPBern(b)`. Note that the Bernstein coefficient computations are not fully certified with `FPBern(a)` yet. We plan to obtain verified upper bounds by using the framework in [10].

`FPKriSten` is built on top of the SBSOS software related to [27] which handles sparse polynomial optimization problems by solving a hierarchy of convex relaxations. This hierarchy is obtained by mixing Krivine-Stengle and Putinar representations of positive polynomials. To improve the overall performance in our particular case, we only consider the former representation yielding the hierarchy of LP relaxations (12). Among several LP solvers, Cplex [13] yields the best performance in our case (see also [1] for more comparisons). Polynomials are handled with the Yalmip [18] toolbox available for `Matlab`. Even though the semantics of programs considered in this paper is actually much simpler than that considered by other tools such as `Rosa` [5] or Fluctuat [7], we emphasize that those tools may be combined with external non-linear solvers to solve specific sub-problems, a task that either `FPBern` or `FPKriSten` can fulfill.

Experimental results. We tested our two software packages with 20 programs (see Appendix A) where 12 are existing benchmarks coming from biology, space control and optimization fields, and 8 are generated as follows, with $\mathbf{x} = (x_1, \ldots, x_n) \in [-1, 1]^n$.

$$\texttt{ex-n-nSum-deg}(\mathbf{x}) := \sum_{j=0}^{\texttt{nSum}} \Big(\prod_{k=1}^{\texttt{deg}} \Big(\sum_{i=1}^{\texttt{n}} x_i\Big)\Big)\,. \tag{13}$$

The first 9 programs are used for similar comparison in [19, Section 4.1], the following 3 come from [24]. Eventually the 8 generated benchmarks allow to

evaluate *independently* the performance of the tools w.r.t. either the number of input variables (through the variable `n`), the degree (through `deg`) or the number of error variables (through `nSum`). Taking $\mathbf{x} \in [-1,1]^n$ allows avoiding monotonicity of the polynomial (which could be exploited by the Bernstein techniques).
We recall that each program implements a polynomial function $f(\mathbf{x})$ with box constrained input variables. To provide an upper bound of the absolute roundoff error $|f(\mathbf{x}) - \hat{f}(\mathbf{x},\mathbf{e})| = |l(\mathbf{x},\mathbf{e}) + h(\mathbf{x},\mathbf{e})|$, we rely on `Real2Float` to generate l and to bound h (see [19, Section 3.1]). Then the optimization methods of Section 3 are applied to bound a function l', obtained after linear transformation of l, over the unit box.
At a given multi-degree $\mathbf{k}$, `FPBern` computes the bound $\overline{l'_{\mathbf{k}}}$ (see Lemma 1). Similarly, at a given relaxation order k, `FPKriSten` computes the bound l'_k (see Lemma 3). To achieve fast computations, the default value of $\mathbf{k}$ is the multi-degree $\mathbf{d}$ of $l'_{\mathbf{e}}$ (equal to the multi-degree of the input polynomial f) and the default value of k is the degree d of l' (equal to the successor of the degree of f).
The experiments were carried out on an Intel Core i7-5600U (2.60Ghz, 16GB) with Ubuntu 14.04LTS, `Matlab 2015a`, GiNaC 1.7.1, and Cplex 12.63. Our benchmark settings are similar to [19, Section 4] as we compare the accuracy and execution times of our two tools with `Rosa real compiler` [5](version from May 2014), `Real2Float` [19](version from July 2016) and `FPTaylor` [25] (version from May 2016) on programs implemented in double precision while considering input variables as real variables. All these tools use a simple rounding model (see Section 2.1) and have been executed with their default parameters.

Table 1 shows the result of the absolute roundoff error while Table 2 displays execution times obtained through averaging over 5 runs. For each benchmark, we indicate the number n (resp. m) of input (resp. error) variables as well as the degree d of l'. For `FPKriSten` the Cplex solving time in Table 2 is given between parentheses. Note that the overall efficiency of the tool could be improved by constructing the hierarchy of LP (12) with a `C++` implementation.

Our two methods yield more accurate bounds for the 3 benchmarks `kepler1`, `sineTaylor` and `kepler2`, which is the program involving the largest number of error variables.
For `kepler1`, `FPBern(a)` and `FPKriSten` are less precise than `FPBern(b)` but are still 6% more precise than `Real2Float` and `FPTaylor` and 53% more precise than `Rosa`. For `kepler2`, our two tools are 3% (resp. 42%) more precise than `FPTaylor` and `Real2Float` (resp. `Rosa`). In addition, Property 4 holds for these three programs with `FPBern(b)`, which ensures bound optimality. For all other benchmarks `FPTaylor` provides the most accurate upper bounds. Our tools are more accurate than `Real2Float` except for `sineOrder3` and `himmilbeau`. In particular, for `himmilbeau`, `FPBern` and `FPKriSten` are 40% (resp. 50%) less precise than `Real2Float` (resp. `FPTaylor`). One way to obtain better bounds would be to increase the degree $\mathbf{k}$ (resp. relaxation order k) within `FPBern`

Table 1. Comparison results of upper bounds for absolute roundoff errors. The best results are emphasized using **bold fonts**.

Benchmark	n	m	d	FPBern(a)	FPBern(b)	FPKriSten	Real2Float	Rosa	FPTaylor
rigidBody1	3	10	3	5.33e–13	5.33e–13	5.33e–13	5.33e–13	5.08e–13	**3.87e–13**
rigidBody2	3	15	5	6.48e–11	6.48e–11	6.48e–11	6.48e–11	6.48e–11	**5.24e–11**
kepler0	6	21	3	1.08e–13	1.08e–13	1.08e–13	1.18e–13	1.16e–13	**1.05e–13**
kepler1	4	28	4	4.23e–13	**4.04e–13**	4.23e–13	4.47**e**–13	6.49e–13	4.49e–13
kepler2	6	42	4	**2.03e–12**	**2.03e–12**	**2.03e–12**	2.09**e**–12	2.89e–12	2.10e–12
sineTaylor	1	13	8	5.51e–16	**5.48e–16**	5.51e–16	6.03**e**–16	9.56e–16	6.75e–16
sineOrder3	1	6	4	1.35e–15	1.35e–15	1.25e–15	1.19e–15	1.11e–15	**9.97e–16**
sqroot	1	15	5	1.29e–15	1.29e–15	1.29e–15	1.29e–15	8.41e–16	**7.13e–16**
himmilbeau	2	11	5	2.00e–12	2.00e–12	1.97e–12	1.43e–12	1.43e–12	**1.32e–12**
schwefel	3	15	5	1.48e–11	1.48e–11	1.48e–11	1.49e–11	1.49e–11	**1.03e–11**
magnetism	7	27	3	1.27e–14	1.27e–14	1.27e–14	1.27e–14	1.27e–14	**7.61e–15**
caprasse	4	34	5	4.49e–15	4.49e–15	4.49e–15	5.63e–15	5.96e–15	**3.04e–15**
ex-2-2-5	2	9	3	2.23e–14	2.23e–14	2.23e–14	2.23e–14	2.23e–14	**1.96e–14**
ex-2-2-10	2	14	3	5.33e–14	5.33e–14	5.33e–14	5.33e–15	5.33e–14	**4.85e–14**
ex-2-2-15	2	19	3	9.55e–14	9.55e–14	9.55e–14	9.55e–14	9.55e–14	**8.84e–14**
ex-2-2-20	2	24	3	1.49e–13	1.49e–13	1.49e–13	TIMEOUT	1.49e–13	**1.40e–13**
ex-2-5-2	2	9	6	1.67e–13	1.67e–13	1.67e–13	1.67e–13	1.67e–13	**1.41e–13**
ex-2-10-2	2	14	11	1.05e–11	1.05e–11	1.34e–11	1.05e–11	1.05e–11	**8.76e–12**
ex-5-2-2	5	12	3	8.55e–14	8.55e–14	8.55e–14	8.55e–14	8.55e–14	**7.72e–14**
ex-10-2-2	10	22	3	5.16e–13	TIMEOUT	5.16e–13	5.16e–13	5.16e–13	**4.82e–13**

(resp. FPKriSten). Preliminary experiments indicate modest accuracy improvement at the expense of performance.

FPBern(a) is the fastest for almost all benchmarks (except program ex-10-2-2 where Rosa yields best performance). FPBern(b) is much slower due to its Matlab implementation, and the use of certified rational arithmetic. We plan to implement a similar certification scheme within FPBern(a).
On the first 12 benchmarks, FPBern(a) is always the fastest while having a similar precision to Real2Float or Rosa.
The results obtained with the 8 generated benchmarks emphasize the limitations of each method. The Bernstein method performs very well when the number of input variables is low, even if the degree increases, as shown in the results for the 6 programs from ex-2-2-5 to ex-2-10-2. This is related to the polynomial dependancy w.r.t. the degree when fixing the number of input variables. However, for the last 2 programs ex-5-2-2 and ex-10-2-2 where the dimension increases, the computation time increases exponentially. This confirms the theoretical result stated in Remark 1 as the number of Bernstein coefficients is exponential w.r.t. the dimension at fixed degree.

On the same programs, the method based on Krivine-Stengle representations performs better when the dimension increases, at fixed degree. This confirms the constraint dependency w.r.t. $[\frac{mk}{n+1}+1]\binom{n+k}{k}$ stated in Remark 2.

Table 2. Comparison of execution times (in seconds) for absolute roundoff error bounds. For each model, the best results are emphasized using **bold fonts**.

Benchmark	n	m	d	FPBern(a)	FPBern(b)	FPKriSten	Real2Float	Rosa	FPTaylor
rigidBody1	3	10	3	**5e–4**	0.88	0.22(0.02)	0.58	0.13	1.84
rigidBody2	3	15	5	**2e–3**	1.87	2.78(0.47)	0.26	2.17	3.01
kepler0	6	21	3	**4e–3**	9.62	1.93(0.18)	0.22	3.78	4.93
kepler1	4	28	4	**6e–3**	6.91	3.93(0.53)	17.6	63.1	9.33
kepler2	6	42	4	**5e–2**	64.9	20.5(3.75)	16.5	106	19.1
sineTaylor	1	13	8	**6e–4**	0.50	0.92(0.27)	1.05	3.50	2.91
sineOrder3	1	6	4	**2e–4**	0.27	0.08(0.01)	0.40	0.48	1.90
sqroot	1	15	5	**2e–4**	0.34	0.24(0.02)	0.14	0.77	2.70
himmilbeau	2	11	5	**1e–3**	1.72	0.77(0.22)	0.20	2.51	3.28
schwefel	3	15	5	**2e–3**	3.04	2.90(0.56)	0.23	3.91	0.53
magnetism	7	27	3	**9e–2**	176	3.07(0.26)	0.29	1.95	5.91
caprasse	4	34	5	**6e–3**	6.03	18.8(4.89)	3.63	17.6	12.2
ex-2-2-5	2	9	3	**4e–4**	0.69	0.12(0.01)	0.07	4.20	2.30
ex-2-2-10	2	14	3	**5e–4**	0.71	0.17(0.01)	0.35	4.75	3.42
ex-2-2-15	2	19	3	**6e–4**	0.72	0.23(0.02)	9.75	5.33	4.91
ex-2-2-20	2	24	3	**8e–4**	0.73	0.28(0.02)	TIMEOUT	6.28	6.27
ex-2-5-2	2	9	6	**2e–2**	2.34	1.23(0.26)	0.27	4.26	2.53
ex-2-10-2	2	14	11	**2e–2**	7.34	96.9(58.5)	49.2	9.37	5.07
ex-5-2-2	5	12	3	**8e–3**	18.3	0.70(0.08)	0.21	4.45	12.3
ex-10-2-2	10	22	3	39.5	TIMEOUT	6.11(0.6)	30.7	**5.34**	34.6

Results for the 4 programs from ex-2-2-5 to ex-2-2-20 also indicate that our methods are the least sensible to an increase of error variables. We note that FPKriSten is often the second fastest tool.

Let us now provide an overall evaluation of our tools. Our tools are comparable with Real2Float (resp. Rosa) in terms of accuracy and faster than them. In comparison with FPTaylor, our tools are in general less precise but still very competitive in accuracy, and they outperform FPTaylor in computation time. A salient advantage of our tools, in particular FPKriSten, over FPTaylor is a good trade-off between computation time and accuracy for large polynomials. As we can see from the experimental results, for ex-10-2-2, FPKriSten took only 6.11s while FPTaylor took 34.6s for comparable precisions. Note that the experimentations were done with FPBern(b) and FPKriSten implemented in Matlab; their C++ implementations would allow a significant speed-up.

The good time performances of our tools come from the exploitation of sparsity. Indeed, a direct Bernstein expansion of the polynomial l associated to kepler2 leads to compute $3^6 \times 2^{42}$ coefficients against 42×3^6 with FPBern. Similarly, dense Krivine-Stengle representations yield an LP with $\binom{100}{4} + 1 = 3\,921\,226$ variables while LP (12) involves $42\binom{18}{4} + 1 = 128\,521$ variables.

5 Conclusion and Future Works

We propose two new methods to compute upper bounds of absolute roundoff errors occurring while executing polynomials programs with floating point precision. The first method uses symbolic Bernstein expansions of polynomials while the second one relies on a hierarchy of LP relaxations derived from sparse Krivine-Stengle representations. The overall computational cost is drastically reduced compared to the dense problem, thanks to a specific exploitation of the sparsity pattern between input and error variables, yielding promising experimental results.

Our approach is currently limited to programs implementing polynomials with box constrained variables. First, a direction of further research investigation is an extension to handle more complicated input sets. Extending to semialgebraic sets is theoretically possible with the hierarchy of LP relaxations based on sparse Krivine-Stengle representations but requires careful implementation in order not to compromise efficiency. For our method based on Bernstein expansions, it would be worth adapting the techniques described in [20] to obtain polygonal approximations of semialgebraic sets. Second, we intend to aim at formal verification of bounds by interfacing either `FPBern` with the PVS libraries [22] related to Bernstein expansions, or `FPKirSten` with the `Coq` libraries available in `Real2Float` [19]. Finally, a delicate but important open problem is to apply such optimization techniques in order to handle roundoff errors of programs implementing finite or infinite loops as well as conditional statements.

References

1. Decision tree for optimization software. `http://plato.la.asu.edu/bench.html`. Accessed: 2016-10-18.
2. Bauer, C., Frink, A., and Kreckel, R. Introduction to the ginac framework for symbolic computation within the c++ programming language. *J. Symb. Comput. 33*, 1 (Jan. 2002), 1–12.
3. Boland, D., and Constantinides, G. A. Automated precision analysis: A polynomial algebraic approach. In *FCCM'10* (2010), pp. 157–164.
4. The Coq Proof Assistant, 2016. `http://coq.inria.fr/`.
5. Darulova, E., and Kuncak, V. Towards a Compiler for Reals. Tech. rep., Ecole Polytechnique Federale de Lausanne, 2016.
6. Daumas, M., and Melquiond, G. Certification of Bounds on Expressions Involving Rounded Operators. *ACM Trans. Math. Softw. 37*, 1 (Jan. 2010), 2:1–2:20.
7. Delmas, D., Goubault, E., Putot, S., Souyris, J., Tekkal, K., and Vdrine, F. Towards an industrial use of fluctuat on safety-critical avionics software. In *Formal Methods for Industrial Critical Systems*, M. Alpuente, B. Cook, and C. Joubert, Eds., vol. 5825 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2009, pp. 53–69.
8. Dreossi, T., and Dang, T. Parameter synthesis for polynomial biological models. In *Proceedings of the 17th international conference on Hybrid systems: computation and control* (2014), ACM, pp. 233–242.
9. Garloff, J. Convergent bounds for the range of multivariate polynomials. In *Interval Mathematics 1985*. Springer, 1986, pp. 37–56.
10. Garloff, J., and Smith, A. P. Rigorous affine lower bound functions for multivariate polynomials and their use in global optimisation. 2008.
11. Grimm, D., Netzer, T., and Schweighofer, M. A note on the representation of positive polynomials with structured sparsity. *Archiv der Mathematik 89*, 5 (2007), 399–403.
12. Harrison, J. HOL Light: A Tutorial Introduction. In *FMCAD* (1996), M. K. Srivas and A. J. Camilleri, Eds., vol. 1166 of *Lecture Notes in Computer Science*, Springer, pp. 265–269.
13. ILOG, Inc. ILOG CPLEX: High-performance software for mathematical programming and optimization, 2006. See `http://www.ilog.com/products/cplex/`.
14. Krivine, J.-L. Anneaux préordonnés. *Journal d'analyse mathématique 12*, 1 (1964), 307–326.
15. Lasserre, J. *Moments, Positive Polynomials and Their Applications*. Imperial College Press optimization series. Imperial College Press, 2009.
16. Lasserre, J. B., Toh, K.-C., and Yang, S. A bounded degree sos hierarchy for polynomial optimization. *EURO Journal on Computational Optimization* (2015), 1–31.
17. Laurent, M. Sums of squares, moment matrices and optimization over polynomials. In *Emerging applications of algebraic geometry*. Springer, 2009, pp. 157–270.
18. Lfberg, J. Yalmip : A toolbox for modeling and optimization in MATLAB. In *Proceedings of the CACSD Conference* (Taipei, Taiwan, 2004).
19. Magron, V., Constantinides, G., and Donaldson, A. Certified roundoff error bounds using semidefinite programming. *arXiv preprint arXiv:1507.03331* (2015).
20. Mantzaflaris, A., and Mourrain, B. *A Subdivision Approach to Planar Semialgebraic Sets*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010, pp. 104–123.

21. Mourrain, B., and Pavone, J. P. Subdivision methods for solving polynomial equations. *Journal of Symbolic Computation 44*, 3 (2009), 292–306.
22. Muñoz, C., and Narkawicz, A. Formalization of a representation of Bernstein polynomials and applications to global optimization. *Journal of Automated Reasoning 51*, 2 (August 2013), 151–196.
23. Smith, A. P. *Enclosure methods for systems of polynomial equations and inequalities*. PhD thesis, 2012.
24. Solovyev, A., and Hales, T. C. Formal verification of nonlinear inequalities with taylor interval approximations. In *NASA Formal Methods, 5th International Symposium, NFM 2013, Moffett Field, CA, USA, May 14-16, 2013. Proceedings* (2013), G. Brat, N. Rungta, and A. Venet, Eds., vol. 7871 of *Lecture Notes in Computer Science*, Springer, pp. 383–397.
25. Solovyev, A., Jacobsen, C., Rakamarić, Z., and Gopalakrishnan, G. Rigorous Estimation of Floating-Point Round-off Errors with Symbolic Taylor Expansions. In *Proceedings of the 20th International Symposium on Formal Methods (FM)* (2015), N. Bjørner and F. de Boer, Eds., vol. 9109 of *Lecture Notes in Computer Science*, Springer, pp. 532–550.
26. Stengle, G. A nullstellensatz and a positivstellensatz in semialgebraic geometry. *Mathematische Annalen 207*, 2 (1974), 87–97.
27. Weisser, T., Lasserre, J.-B., and Toh, K.-C. A bounded degree sos hierarchy for large scale polynomial optimization with sparsity. *arXiv preprint arXiv:1607.01151* (2016).
28. Zuras, D., Cowlishaw, M., Aiken, A., Applegate, M., Bailey, D., Bass, S., Bhandarkar, D., Bhat, M., Bindel, D., Boldo, S., et al. Ieee standard for floating-point arithmetic. *IEEE Std 754-2008* (2008), 1–70.

A Polynomial Program Benchmarks

- rigibody1 : $(x_1, x_2, x_3) \mapsto -x_1x_2 - 2x_2x_3 - x_1 - x_3$ defined on $[-15, 15]^3$.
- rigibody2 : $(x_1, x_2, x_3) \mapsto 2x_1x_2x_3 + 6x_3^2 - x_2^2x_1x_3 - x_2$ defined on $[-15, 15]^3$.
- kepler0 : $(x_1, x_2, x_3, x_4, x_5, x_6) \mapsto x_2x_5 + x_3x_6 - x_2x_3 - x_5x_6 + x_1(-x_1 + x_2 + x_3 - x_4 + x_5 + x_6)$ defined on $[4, 6.36]^6$.
- kepler1 : $(x_1, x_2, x_3, x_4) \mapsto x_1x_4(-x_1 + x_2 + x_3 - x_4) + x_2(x_1 - x_2 + x_3 + x_4) + x_3(x_1 + x_2 - x_3 + x_4) - x_2x_3x_4 - x_1x_3 - x_1x_2 - x_4$ defined on $[4, 6.36]^4$.
- kepler2 : $(x_1, x_2, x_3, x_4, x_5, x_6) \mapsto x_1x_4(-x_1 + x_2 + x_3 - x_4 + x_5 + x_6) + x_2x_5(x_1 - x_2 + x_3 + x_4 - x_5 + x_6) + x_3x_6(x_1 + x_2 - x_3 + x_4 + x_5 - x_6) - x_2x_3x_4 - x_1x_3x_5 - x_1x_2x_6 - x_4x_5x_6$ defined on $[4, 6.36]^6$.
- sineTaylor : $x \mapsto x - \frac{x^3}{6.0} + \frac{x^5}{120.0} - \frac{x^7}{5040.0}$ defined on $[-1.57079632679, 1.57079632679]$.
- sineOrder3 : $x \mapsto 0.954929658551372x - 0.12900613773279798x^3$ defined on $[-2, 2]$.
- sqroot : $x \mapsto 1.0 + 0.5x - 0.125x^2 + 0.0625x^3 - 0.0390625x^4$ defined on $[0, 1]$.
- himmilbeau : $(x_1, x_2) \mapsto (x_1^2 + x_2 - 11)^2 + (x_1 + x_2^2 - 7)^2$ defined on $[-5, 5]^2$.
- schwefel : $(x_1, x_2, x_3) \mapsto (x_1 - x_2)^2 + (x_2 - 1)^2 + (x_1 - x_3^2)^2 + (x_3 - 1)^2$ defined on $[-10, 10]^3$.
- magnetism : $(x_1, x_2, x_3, x_4, x_5, x_6, x_7) \mapsto x_1^2 + 2x_2^2 + 2x_3^2 + 2x_4^2 + 2x_5^2 + 2x_6^2 + 2x_7^2 - x_1$ defined on $[-1, 1]^7$.
- caprasse : $(x_1, x_2, x_3, x_4) \mapsto x_1x_3^3 + 4x_2x_3^2x_4 + 4x_1x_3x_4^2 + 2x_2x_4^3 + 4x_1x_3 + 4x_3^2 - 10x_2x_4 - 10x_4^2 + 2$ defined on $[-0.5, 0.5]^4$.