Exact Real Arithmetic with Perturbation Analysis and Proof of Correctness

Sarmen Keshishzadeh and Jan Friso Groote

Department of Mathematics and Computer Science
Eindhoven University of Technology
Den Dolech 2, 5612 AZ Eindhoven, The Netherlands

Abstract

In this article, we consider a simple representation for real numbers and propose top-down procedures to approximate various algebraic and transcendental operations with arbitrary precision. Detailed algorithms and proofs are provided to guarantee the correctness of the approximations. Moreover, we develop and apply a perturbation analysis method to show that our approximation procedures only recompute expressions when unavoidable.

In the last decade, various theories have been developed and implemented to realize real computations with arbitrary precision. Proof of correctness for existing approaches typically consider basic algebraic operations, whereas detailed arguments about transcendental operations are not available. Another important observation is that in each approach some expressions might require iterative computations to guarantee the desired precision. However, no formal reasoning is provided to prove that such iterative calculations are essential in the approximation procedures. In our approximations of real functions, we explicitly relate the precision of the inputs to the guaranteed precision of the output, provide full proofs and a precise analysis of the necessity of iterations.

1 Introduction

Various scientific disciplines use computations involving real numbers to model and reason about different phenomena in the world. Real numbers are typically approximated by floating point numbers in scientific calculations. Round-off errors are inevitable in such approximations and they might build up into catastrophic errors in some cases. Exact real arithmetic approaches address this issue by devising computation procedures that given an expression and a precision requested by the user produce an output that is guaranteed to meet the precision requirement.

Several approaches [16, 13] to exact real arithmetic are based on iterative bottom-up calculations. Given an expression and a desired precision, bottom-up approaches typically start with calculating the inputs with an arbitrary precision higher than the requested precision. Then, the sub-expressions are evaluated in a bottom-up way. After evaluating the sub-expressions of each level, the guaranteed precision is passed to the higher level. These calculations proceed until the main expression is calculated and its guaranteed precision is determined. If the precision obtained for the expression is not adequate, the computation restarts with increased precisions for the inputs.

In contrast, various authors [3, 15] have proposed top-down approaches to exact real arithmetic. In top-down approaches, the required precision of each sub-expression is determined based on the precision required for its immediate parent expression. For certain types of expressions, the required precision of the sub-expressions can be calculated immediately. However, some expressions may require to first obtain additional information about the magnitude of the values of their sub-expressions before determining their required precision. Thus, in general, it might be necessary to recompute certain expressions.

The main benefit of top-down approaches is that they exploit the structure of a given expression to estimate the required precision of the sub-expressions. In this context, one would ideally like to have top-down approximations for algebraic and transcendental functions such that 1) the approximations are proven to be correct and 2) iterative calculations are avoided unless they are proven to be necessary.

In several studies [3, 17], proofs of correctness for algebraic operations are available. However, the arguments about transcendental functions provide little insight about the correctness of the approximations and the effect of these operations on precision. Taylor expansions are the most prominent way to approximate transcendental functions. Calculations with Taylor expansions are typically restricted to a base interval; range reduction identities are used to extend the computations to the complete domain of a function. Proofs of correctness for transcendental functions are limited to the base interval [18, 3, 17], whereas little attention is given to the general case where the computations introduced by range reduction identities influence the output precision.

The second desired property for a top-down approach is related to the iterative nature of the computations. As discussed above, bottom-up and top-down approaches rely on iterative computation schemes. However, no formal reasoning is provided to prove that such iterative calculations are essential in the approximation procedures.

In this article, we consider a simple representation for real numbers and propose a top-down approach to approximate various algebraic and transcendental functions with arbitrary precision. For each operation, we describe an approximation procedure and relate the precision of the inputs to the guaranteed precision for the output. To guarantee the correctness of our top-down approach, we provide detailed proofs of correctness for the proposed approximations.

To identify computational problems that require iterative calculations in our

top-down approach, we have developed a perturbation analysis method. Our analysis describes the influence of errors in the inputs of a computational problem on the output precision. We apply perturbation analysis to show that our approximation procedures only recompute expressions when this is unavoidable.

Overview We discuss different approaches to defining computability of functions in Section 2. In Section 3 we introduce a representation for real numbers and specify the syntax of the expressions that we consider in our computations. To analyze computational problems in this setting, a perturbation analysis method is introduced in Section 4. In Section 5 we discuss our approximations of algebraic operations. We approximate transcendental functions using Riemann sums and Taylor expansions in Section 6 and 7, respectively. Section 8 contains discussions about related work. In Section 9 we draw some conclusions and suggest directions for future research.

2   Computable Real Functions

Real arithmetic is concerned with performing computations on real numbers. In order to do calculations with real numbers, it is necessary to define what it means for an operation to be computable. In this section we briefly discuss different approaches to defining computability.

Since real numbers are infinite objects, one can use infinite streams from a finite alphabet Σ to represent them. This gives rise to a definition of computability called Type-2 Theory of Effectivity (TTE, [21]). In TTE computable operations are defined in terms of functions $f : \Sigma^\omega \to \Sigma^\omega$ that receive infinite words as input and produce infinite words as output. An essential property of a Type-2 computable function is the finiteness property [21]. This property indicates that for a computable function f, any finite prefix of the output $f(x)$ is computable by some finite portion of the input x.

An alternative definition of computability has been introduced by the Russian school of constructive analysis [14, 12]. In this definition, computable operations are defined in terms of functions $f : \Sigma^* \to \Sigma^*$ that receive finite words as input and produce finite words as output. This approach is sometimes referred to as Type-1 computability. A function f is Type-1 computable if there is a Turing machine that transforms any finite input string $x \in \Sigma^*$ to the finite output strings $f(x) \in \Sigma^*$. Type-1 machines provide a natural way to define computability on, for instance, rational numbers and finite graphs.

In both Type-1 and Type-2 approaches, Σ depends on the concrete representation that we use for input/output objects. For instance, one can use the binary signed-digit representation to represent real numbers in the inputs/outputs of computations.

The relation between Type-1 and Type-2 computable functions has been investigated in various studies. It is known that restricting the domain of a Type-2 computable function to finite streams results in a Type-1 computable function [4, 19]. However, not every Type-1 function can be obtained by restricting some

Type-2 computable function [19, 11]. To illustrate this, we consider a function $f_1 : \{0,1\}^* \to \{0,1\}^*$ defined as follows:

$$f_1(s) = \begin{cases} 0 & \text{if } s = 0^k \\ 1 & \text{if } s = 0^k 1 s' \end{cases}$$

where $k \in \mathbb{N}$ is the length of the longest prefix of zeros in s and $s' \in \Sigma^*$ is a finite suffix of s in the second case. The function f_1 performs computations on finite strings and one can construct a Type-1 Turing machine to compute this function. By extending the domain of f_1 to infinite strings, we obtain $f_2 : \Sigma^\omega \to \Sigma^\omega$ such that:

$$f_2(s) = \begin{cases} 0 & \text{if } s = 0^\omega \\ 1 & \text{if } s = 0^k 1 s' \end{cases}$$

The function f_2 is not computable with a Type-2 Turing machine; it is not possible to write 0 in the output after reading a finite prefix from the input. The interested reader can refer to [4] for more details about the relation between Type-1 and Type-2 computable functions.

In addition to Type-1 and Type-2, one can also consider a third approach to defining computability based on certain finite structures that provide precise descriptions for specific classes of real numbers. For instance, Lagrange's theorem on continued fractions indicates that the real numbers whose continued fraction is periodic are the quadratic irrationals. Based on this observation, one can define computability in terms of functions f that given a finite and precise representation of x produce a finite and precise representation of $f(x)$.

The exact real arithmetic approach that we introduce in this article is based on Type-2 computability. In Section 3 we discuss a representation for real numbers in terms of rational numbers that are coupled with a notion of precision. Our approximations for arithmetic operations rely on the finiteness property of computable functions. Thus, for a given computational problem, a desired precision for the output is obtained based on sufficiently good approximations of the inputs.

3 Real Numbers: Representation & Operations

In this section we first discuss our representation of real numbers and then describe the syntax of the expressions that we aim to calculate in our setting.

Since real numbers are infinite objects, a finite representation of an arbitrary real number x can only approximate x with a certain precision. In scientific measurements and calculations, the amount of error that we commit in approximations is measured by an absolute or relative error. In practice, an absolute error is of little use. Since numbers tend to have very different orders of magnitude, it is the relative error that shows the significance of the lost digits in measurements or calculations. Hence, in our setting we use a representation based on the relative error.

Definition 1. *A real number x is represented by a tuple (m, n, p) such that:*

$$|x - \frac{m}{n}| < |\frac{m}{n}|\frac{1}{2^p}$$

where $m, n \in \mathbb{Z} \setminus \{0\}, p \in \mathbb{N}$.

The representation (m, n, p) for x means that $\frac{m}{n}$ approximates x and the relative error of this approximation does not exceed $\frac{1}{2^p}$.

In this article, we focus on calculating expressions that can be described with the following grammar:

$$E ::= c \mid -E \mid E \cdot E \mid \frac{1}{E} \mid E + E \mid \sqrt{E} \mid e^E \mid \tag{1}$$
$$\ln(E) \mid \arctan(E) \mid \cos(E) \mid \sin(E)$$

where c represents a rational constant.

The following identities show that other interesting operations can be described in terms of the operations of this grammar:

$$\tan(x) = \frac{\sin(x)}{\cos(x)}$$
$$\cot(x) = \frac{1}{\tan(x)}$$
$$\arcsin(x) = \arctan(\frac{x}{\sqrt{1 - x^2}})$$
$$\arccos(x) = \arctan(\frac{\sqrt{1 - x^2}}{x})$$
$$\text{arccot}(x) = \arccos(\frac{x}{\sqrt{1 + x^2}})$$

4 Sensitivity of Operations to Perturbations in the Arguments

Our goal is to develop a top-down exact real arithmetic approach based on the representation of Definition 1. Thus, for a given computational problem, it is essential to estimate the required precision of the inputs based on the desired precision in the output. Moreover, we would like to investigate to what extend the operations of grammar (1) can be calculated in a top-down manner without iterations.

To analyze the operations of grammar (1), we introduce a pertubation analysis method for measuring the sensitivity of the operations to perturbations in their arguments. We consider two general cases in our analysis. First, we consider a function $f(x)$ in one variable and show how errors in the input influence the output (Section 4.1). Then, we consider a function $f(x, y)$ with two arguments and investigate the effect of errors in the inputs on the output (Section 4.2).

4.1 Perturbation Analysis for Unary Functions

Let $f(x)$ be a differentiable function that we want to calculate in point $x = a$. Suppose that Δa is a perturbation in the argument a. The relative error in the calculation of $f(a)$ caused by this perturbation is:

$$|\frac{f(a + \Delta a) - f(a)}{f(a)}|$$

We want to relate this relative error to the relative error of the argument, namely $|\frac{\Delta a}{a}|$. To this end, we use the following approximation of the function f in point $x = a + \Delta a$:

$$f(a + \Delta a) \approx f(a) + f'(a)\Delta a$$

We can approximate the relative error of f as follows:

$$|\frac{f(a + \Delta a) - f(a)}{f(a)}| \approx |\frac{f'(a)\Delta a}{f(a)}| = |\frac{af'(a)}{f(a)}||\frac{\Delta a}{a}| \tag{2}$$

From equality (2) one can see that the quantity $|\frac{af'(a)}{f(a)}|$ determines the effect of the relative error $|\frac{\Delta a}{a}|$ on the output. In numerical analysis and linear algebra the quantity $|\frac{xf'(x)}{f(x)}|$ is usually referred to as the *condition number* of $f(x)$ [5, 6].

4.2 Perturbation Analysis for Binary Functions

Let $f(x, y)$ be a differentiable function that we want to calculate in point $(x, y) = (a, b)$. Suppose Δa and Δb are perturbations in the arguments a and b, respectively. The relative error of $f(a, b)$ caused by these perturbations can be calculated as follows:

$$|\frac{f(a + \Delta a, b + \Delta b) - f(a, b)}{f(a, b)}| \tag{3}$$

To find an upper-bound for (3), we use the following first-order approximation of the function f in point $(x, y) = (a + \Delta a, b + \Delta b)$:

$$f(a + \Delta a, b + \Delta b) \approx f(a, b) + f_x(a, b)\Delta a + f_y(a, b)\Delta b$$

where $f_x(a, b)$ and $f_y(a, b)$ are the partial derivatives of f with respect to x and y in (a, b), respectively. We relate the relative error in the calculation of f to the relative errors $|\frac{\Delta a}{a}|$ and $|\frac{\Delta b}{b}|$ as follows:

$$|\frac{f(a + \Delta a, b + \Delta b) - f(a, b)}{f(a, b)}| \approx |\frac{f_x(a, b)\Delta a + f_y(a, b)\Delta b}{f(a, b)}| \leq$$

$$|\frac{af_x(a, b)}{f(a, b)}||\frac{\Delta a}{a}| + |\frac{bf_y(a, b)}{f(a, b)}||\frac{\Delta b}{b}| \leq$$

$$(|\frac{af_x(a, b)}{f(a, b)}| + |\frac{bf_y(a, b)}{f(a, b)}|)\max\{|\frac{\Delta a}{a}|, |\frac{\Delta b}{b}|\} \tag{4}$$

The quantity $|\frac{af_x(a,b)}{f(a,b)}| + |\frac{bf_y(a,b)}{f(a,b)}|$ determines the upper-bound calculated in inequality (4) and we use this quantity to measure the effect of erroneous arguments on the output. It should be noted that in inequality (4) we have considered $|\frac{\Delta a}{a}|$ and $|\frac{\Delta b}{b}|$ as independent factors that can influence the relative error of f. This way of reasoning about the sensitivity of $f(x,y)$ is related to *componentwise* analysis of perturbation in numerical analysis and linear algebra [7], which we use in this article.

Another possibility is to relate the relative error of (3) to the quantity:

$$\frac{\left\|\begin{bmatrix}\Delta a\\\Delta b\end{bmatrix}\right\|}{\left\|\begin{bmatrix}a\\b\end{bmatrix}\right\|}$$

This type of analysis is usually referred to as *normwise* analysis of perturbation [7].

In the following sections, we provide a top-down approach for approximating various algebraic and transcendental functions. Perturbation analysis will be used to show that in our approximations we only recompute expressions when essential.

5   Approximating Algebraic Operations

In this section we calculate the algebraic operations of grammar (1) using a top-down approach. We formulate and prove theorems that allow us to calculate expressions involving unary negation (Section 5.1), multiplication (Section 5.2), inverse (Section 5.3), addition (Section 5.4), and square root (Section 5.5). Based on the theorems, we provide different implementations of COMPUTE($expr, p$) to calculate algebraic operations. These implementations receive an algebraic expression $expr$ and a desired precision p and produce an output with the desired precision. In each case, we also apply the perturbation analysis of Section 4 and show that we avoid unnecessary iterations in our approximations.

5.1   Unary Negation

Theorem 5.1. *Let x be a real number represented by (m, n, p). Then $-x$ can be represented by $(-m, n, p)$.*

Proof. Since x is represented by (m, n, p) we can write:

$$\frac{m}{n} - |\frac{m}{n}|\frac{1}{2^p} < x < \frac{m}{n} + |\frac{m}{n}|\frac{1}{2^p}$$
$$-\frac{m}{n} - |\frac{m}{n}|\frac{1}{2^p} < -x < -\frac{m}{n} + |\frac{m}{n}|\frac{1}{2^p}$$

Thus, we can represent $-x$ by $(-m, n, p)$.

$\square$

Algorithm 1 applies Theorem 5.1 and approximates $-x$ based on a representation (m, n, p) of x. To confirm that $-x$ can be approximated with arbitrary precision in one pass, we calculate $|\frac{xf'(x)}{f(x)}|$ for the function $f(x) = -x$:

$$|\frac{xf'(x)}{f(x)}| = |\frac{(x)(-1)}{-x}| = 1$$

The quantity $|\frac{xf'(x)}{f(x)}|$ is small and independent of the argument x and hence the amount of precision that we lose in unary negation (which is 0 in the case of Theorem 5.1) can be calculated independently of x.

Algorithm 1 Unary Negation

Require: *expr* has the shape $-x$
  1: **procedure** COMPUTE(*expr*, *p*)
  2:     $\frac{m}{n} \leftarrow$ COMPUTE(x, p)
  3:     **return** $\frac{-m}{n}$

5.2 Multiplication $x \cdot y$

Theorem 5.2. *Let x and y be two real numbers represented by (m, n, p) and (m', n', p), respectively. Then $x \cdot y$ can be represented by $(mm', nn', p - 2)$.*

Proof. From Definition 1 we can write:

$$\frac{m}{n} - |\frac{m}{n}|\frac{1}{2^p} < x < \frac{m}{n} + |\frac{m}{n}|\frac{1}{2^p} \tag{5}$$

$$\frac{m'}{n'} - |\frac{m'}{n'}|\frac{1}{2^p} < y < \frac{m'}{n'} + |\frac{m'}{n'}|\frac{1}{2^p} \tag{6}$$

We consider three cases:

1. Suppose $\frac{mm'}{nn'} > 0$. We can multiply inequalities (5) and (6) as follows:

$$\frac{mm'}{nn'}(1 - \frac{1}{2^p})^2 < x \cdot y < \frac{mm'}{nn'}(1 + \frac{1}{2^p})^2$$

If $x \cdot y$ can be represented by $(mm', nn', p - 2)$ then it must be the case that:

$$\frac{mm'}{nn'}(1 - \frac{1}{2^{p-2}}) < x \cdot y < \frac{mm'}{nn'}(1 + \frac{1}{2^{p-2}})$$

To see that this is valid, we need to show that:

$$\frac{mm'}{nn'}(1 + \frac{1}{2^p})^2 \leq \frac{mm'}{nn'}(1 + \frac{1}{2^{p-2}})$$

$$\frac{mm'}{nn'}(1 - \frac{1}{2^{p-2}}) \leq \frac{mm'}{nn'}(1 - \frac{1}{2^p})^2$$

But $\frac{mm'}{nn'} > 0$ and from Proposition 1 (see Appendix A) we know that both inequalities hold.

2. Suppose $\frac{m}{n} > 0$, $\frac{m'}{n'} < 0$. We can rewrite inequalities (5) and (6) as follows:

$$\frac{m}{n}(1 - \frac{1}{2^p}) < x < \frac{m}{n}(1 + \frac{1}{2^p}) \qquad (7)$$

$$\frac{m'}{n'}(1 + \frac{1}{2^p}) < y < \frac{m'}{n'}(1 - \frac{1}{2^p}) \qquad (8)$$

Multiplying inequalities (7) and (8) we get:

$$\frac{mm'}{nn'}(1 + \frac{1}{2^p})^2 < x \cdot y < \frac{mm'}{nn'}(1 - \frac{1}{2^p})^2$$

If $x \cdot y$ is representable by $(mm', nn', p - 2)$ then it must be the case that:

$$\frac{mm'}{nn'}(1 + \frac{1}{2^{p-2}}) < x \cdot y < \frac{mm'}{nn'}(1 - \frac{1}{2^{p-2}})$$

To show that this is valid, it suffices to prove that:

$$\frac{mm'}{nn'}(1 - \frac{1}{2^p})^2 \leq \frac{mm'}{nn'}(1 - \frac{1}{2^{p-2}})$$

$$\frac{mm'}{nn'}(1 + \frac{1}{2^{p-2}}) \leq \frac{mm'}{nn'}(1 + \frac{1}{2^p})^2$$

But $\frac{mm'}{nn'} < 0$ and from Proposition 1 (see Appendix A) we know that both inequalities hold.

3. Suppose $\frac{m}{n} < 0$, $\frac{m'}{n'} > 0$. This case can be proved similarly to the second case.

$\square$

Algorithm 2 Multiplication

Require: *expr* has the shape $x \cdot y$
 1: **procedure** COMPUTE(*expr*, p)
 2: $\quad \frac{m}{n} \leftarrow$ COMPUTE($x, p + 2$)
 3: $\quad \frac{m'}{n'} \leftarrow$ COMPUTE($y, p + 2$)
 4: $\quad$ **return** $\frac{mm'}{nn'}$

Algorithm 2 depicts an approximation of $x \cdot y$ based on Theorem 5.2. Given approximations of x and y, this algorithm approximates $x \cdot y$ in one pass; the loss of precision is predictable without additional information about x and y. To confirm this claim, we apply the perturbation analysis of Section 4 on the function $f(x, y) = x \cdot y$:

$$|\frac{xf_x(x, y)}{f(x, y)}| + |\frac{yf_y(x, y)}{f(x, y)}| = |\frac{x.y}{x.y}| + |\frac{y.x}{x.y}| = 2$$

The sensitivity measure is small and independent of the arguments. Thus, in a top-down approach, $x \cdot y$ can be approximated in one pass without iterative computations of x and y.

5.3 Inverse

Theorem 5.3. *Let x be a real number represented by (m, n, p). Then $\frac{1}{x}$ can be represented by $(n, m, p - 1)$.*

Proof. We consider two cases:

1. Suppose $\frac{m}{n} > 0$. Since x is represented by (m, n, p) we can write:

$$\frac{m}{n}(1 - \frac{1}{2^p}) < x < \frac{m}{n}(1 + \frac{1}{2^p})$$
$$\frac{n}{m}(\frac{2^p}{2^p + 1}) < \frac{1}{x} < \frac{n}{m}(\frac{2^p}{2^p - 1})$$

   If $\frac{1}{x}$ is representable by $(n, m, p - 1)$, then it must be the case that:

$$\frac{n}{m}(1 - \frac{1}{2^{p-1}}) < \frac{1}{x} < \frac{n}{m}(1 + \frac{1}{2^{p-1}})$$

   To see that this is valid, it suffices to show:

$$\frac{n}{m}(\frac{2^p}{2^p - 1}) \leq \frac{n}{m}(1 + \frac{1}{2^{p-1}})$$
$$\frac{n}{m}(1 - \frac{1}{2^{p-1}}) \leq \frac{n}{m}(\frac{2^p}{2^p + 1})$$

   But $\frac{n}{m} > 0$ and hence both inequalities follow from Proposition 2 (see Appendix A).

2. Suppose $\frac{m}{n} < 0$. Since x is representable by (m, n, p) we have:

$$\frac{m}{n}(1 + \frac{1}{2^p}) < x < \frac{m}{n}(1 - \frac{1}{2^p})$$
$$\frac{n}{m}(\frac{2^p}{2^p - 1}) < \frac{1}{x} < \frac{n}{m}(\frac{2^p}{2^p + 1})$$

   If $\frac{1}{x}$ is representable by $(n, m, p - 1)$ then it must be the case that:

$$\frac{n}{m}(1 + \frac{1}{2^{p-1}}) < \frac{1}{x} < \frac{n}{m}(1 - \frac{1}{2^{p-1}})$$

   To see that this is valid, we need to show:

$$\frac{n}{m}(\frac{2^p}{2^p + 1}) \leq \frac{n}{m}(1 - \frac{1}{2^{p-1}})$$
$$\frac{n}{m}(1 + \frac{1}{2^{p-1}}) \leq \frac{n}{m}(\frac{2^p}{2^p - 1})$$

   But $\frac{n}{m} < 0$ and hence the inequalities follow from Proposition 2 (see Appendix A).

$\square$

Algorithm 3 Inverse

Require: *expr* has the shape $\frac{1}{x}$
1: **procedure** COMPUTE($expr, p$)
2:     $\frac{m}{n} \leftarrow$ COMPUTE($x, p + 1$)
3:         **return** $\frac{n}{m}$

Algorithm 3 approximates $\frac{1}{x}$ with precision p based on Theorem 5.3. Given an approximation of x with precision p, the algorithm allows us to approximate $\frac{1}{x}$ in one pass. We use perturbation analysis and calculate the quantity $|\frac{xf'(x)}{f(x)}|$ for $f(x) = \frac{1}{x}$ to show that loss of precision in the inverse can be estimated independently of the argument:

$$|\frac{xf'(x)}{f(x)}| = |\frac{(x)(\frac{-1}{x^2})}{(\frac{1}{x})}| = 1$$

The quantity $|\frac{xf'(x)}{f(x)}|$ is a constant and hence iterative computations can be avoided when calculating the inverse.

5.4   Addition

Theorem 5.4. *Let x and y be two real numbers represented by (m, n, p) and (m', n', p), respectively. The value of $x + y$ can be approximated as follows:*

i. If $\frac{mm'}{nn'} > 0$, then $x + y$ can be represented by $(mn' + m'n, nn', p)$.

ii. If $\frac{mm'}{nn'} < 0$ and $i \in \mathbb{N}^+$ is the smallest natural number such that $i \geq \log_2(\frac{1 + \frac{\min(|\frac{m}{n}|, |\frac{m'}{n'}|)}{\max(|\frac{m}{n}|, |\frac{m'}{n'}|)}}{1 - \frac{\min(|\frac{m}{n}|, |\frac{m'}{n'}|)}{\max(|\frac{m}{n}|, |\frac{m'}{n'}|)}})$, then $x + y$ can be represented by $(mn' + m'n, nn', p - i)$.

Proof.

i. For numbers x and y we can write:

$$\frac{m}{n} - |\frac{m}{n}|\frac{1}{2^p} < x < \frac{m}{n} + |\frac{m}{n}|\frac{1}{2^p} \tag{9}$$

$$\frac{m'}{n'} - |\frac{m'}{n'}|\frac{1}{2^p} < y < \frac{m'}{n'} + |\frac{m'}{n'}|\frac{1}{2^p} \tag{10}$$

From inequalities (9) and (10) we can write:

$$(\frac{m}{n} + \frac{m'}{n'}) - \frac{1}{2^p}(|\frac{m}{n}| + |\frac{m'}{n'}|) < x + y < (\frac{m}{n} + \frac{m'}{n'}) + \frac{1}{2^p}(|\frac{m}{n}| + |\frac{m'}{n'}|) \tag{11}$$

11

If $x + y$ is representable by $(mn' + m'n, nn', p)$, then it must be the case that:

$$(\frac{m}{n} + \frac{m'}{n'}) - \frac{1}{2^p}|\frac{m}{n} + \frac{m'}{n'}| < x + y < (\frac{m}{n} + \frac{m'}{n'}) + \frac{1}{2^p}|\frac{m}{n} + \frac{m'}{n'}|$$

To show that this is valid, we need to prove that:

$$(\frac{m}{n} + \frac{m'}{n'}) + \frac{1}{2^p}(|\frac{m}{n}| + |\frac{m'}{n'}|) \le (\frac{m}{n} + \frac{m'}{n'}) + \frac{1}{2^p}|\frac{m}{n} + \frac{m'}{n'}| \quad (12)$$

$$(\frac{m}{n} + \frac{m'}{n'}) - \frac{1}{2^p}|\frac{m}{n} + \frac{m'}{n'}| \le (\frac{m}{n} + \frac{m'}{n'}) - \frac{1}{2^p}(|\frac{m}{n}| + |\frac{m'}{n'}|) \quad (13)$$

The rational numbers $\frac{m}{n}$ and $\frac{m'}{n'}$ have the same sign. Therefore, $|\frac{m}{n} + \frac{m'}{n'}| = |\frac{m}{n}| + |\frac{m'}{n'}|$ holds and inequalities (12) and (13) are valid.

ii. If $x + y$ is representable by $(mn' + m'n, nn', p - i)$, then it must be the case that:

$$(\frac{m}{n} + \frac{m'}{n'}) - \frac{1}{2^{p-i}}|\frac{m}{n} + \frac{m'}{n'}| < x + y < (\frac{m}{n} + \frac{m'}{n'}) + \frac{1}{2^{p-i}}|\frac{m}{n} + \frac{m'}{n'}|$$

To show that this holds, we need to prove the following (see inequality (11)):

$$(\frac{m}{n} + \frac{m'}{n'}) + \frac{1}{2^p}(|\frac{m}{n}| + |\frac{m'}{n'}|) \le (\frac{m}{n} + \frac{m'}{n'}) + \frac{1}{2^{p-i}}|\frac{m}{n} + \frac{m'}{n'}|$$

$$(\frac{m}{n} + \frac{m'}{n'}) - \frac{1}{2^{p-i}}|\frac{m}{n} + \frac{m'}{n'}| \le (\frac{m}{n} + \frac{m'}{n'}) - \frac{1}{2^p}(|\frac{m}{n}| + |\frac{m'}{n'}|)$$

For both inequalities, it boils down to proving the following:

$$\frac{1}{2^i}(|\frac{m}{n}| + |\frac{m'}{n'}|) \le |\frac{m}{n} + \frac{m'}{n'}| \quad (14)$$

To prove inequality (14), we consider the following two cases:

1. Suppose $\frac{m}{n} > 0$ and $\frac{m'}{n'} < 0$. We can rewrite inequality (14) as follows:

$$|\frac{m}{n} + \frac{m'}{n'}| \ge \frac{1}{2^i}(\frac{m}{n} - \frac{m'}{n'}) \Leftrightarrow (\frac{m}{n} + \frac{m'}{n'}) \ge \frac{1}{2^i}(\frac{m}{n} - \frac{m'}{n'}) \vee$$
$$(\frac{m}{n} + \frac{m'}{n'}) \le \frac{1}{2^i}(\frac{m'}{n'} - \frac{m}{n})$$

In other words, we should show $\frac{\frac{m}{n}}{-\frac{m'}{n'}} \ge \frac{2^i+1}{2^i-1}$ or $\frac{-\frac{m'}{n'}}{\frac{m}{n}} \ge \frac{2^i+1}{2^i-1}$. Depending on the values of $|\frac{m}{n}|$ and $|\frac{m'}{n'}|$, both cases follow from $\frac{\max(|\frac{m}{n}|,|\frac{m'}{n'}|)}{\min(|\frac{m}{n}|,|\frac{m'}{n'}|)} \ge$

$\frac{2^i+1}{2^i-1}$. Equivalently, we should have $i \ge \log_2(\frac{1+\frac{\min(|\frac{m}{n}|,|\frac{m'}{n'}|)}{\max(|\frac{m}{n}|,|\frac{m'}{n'}|)}}{1-\frac{\min(|\frac{m}{n}|,|\frac{m'}{n'}|)}{\max(|\frac{m}{n}|,|\frac{m'}{n'}|)}})$.

12

2. Suppose $\frac{m}{n} < 0$ and $\frac{m'}{n'} > 0$. We can rewrite inequality (14) as follows:

$$|\frac{m}{n} + \frac{m'}{n'}| \geq \frac{1}{2^i}(\frac{m'}{n'} - \frac{m}{n}) \Leftrightarrow (\frac{m}{n} + \frac{m'}{n'}) \geq \frac{1}{2^i}(\frac{m'}{n'} - \frac{m}{n}) \vee$$
$$(\frac{m}{n} + \frac{m'}{n'}) \leq \frac{1}{2^i}(\frac{m}{n} - \frac{m'}{n'})$$

In other word, we need to prove $\frac{\frac{m'}{n'}}{-\frac{m}{n}} \geq \frac{2^i+1}{2^i-1}$ or $\frac{-\frac{m}{n}}{\frac{m'}{n'}} \geq \frac{2^i+1}{2^i-1}$. Depending

on the values of $|\frac{m}{n}|$ and $|\frac{m'}{n'}|$, both cases follow from $\frac{\max(|\frac{m}{n}|,|\frac{m'}{n'}|)}{\min(|\frac{m}{n}|,|\frac{m'}{n'}|)} \geq$

$\frac{2^i+1}{2^i-1}$. Equivalently, we should have $i \geq \log_2(\frac{1+\frac{\min(|\frac{m}{n}|,|\frac{m'}{n'}|)}{\max(|\frac{m}{n}|,|\frac{m'}{n'}|)}}{1-\frac{\min(|\frac{m}{n}|,|\frac{m'}{n'}|)}{\max(|\frac{m}{n}|,|\frac{m'}{n'}|)}})$.

$\square$

Algorithm 4 Addition

Require: *expr* has the shape $x + y$
1: **procedure** COMPUTE($expr, p$)
2:     $dp \leftarrow p$
3:     **repeat**
4:         $\frac{m}{n} \leftarrow$ COMPUTE(x, dp)
5:         $\frac{m'}{n'} \leftarrow$ COMPUTE(y, dp)
6:         **if** $\frac{mm'}{nn'} > 0$ **then**   ▷ Theorem 5.4.i
7:             **return** $\frac{mn'+m'n}{nn'}$
8:         **else**   ▷ Theorem 5.4.ii
9:             $i \leftarrow \lceil \log_2(\frac{1+\frac{\min(|\frac{m}{n}|,|\frac{m'}{n'}|)}{\max(|\frac{m}{n}|,|\frac{m'}{n'}|)}}{1-\frac{\min(|\frac{m}{n}|,|\frac{m'}{n'}|)}{\max(|\frac{m}{n}|,|\frac{m'}{n'}|)}}) \rceil$
10:             **if** $dp - i \geq p$ **then**
11:                 **return** $\frac{mn'+m'n}{nn'}$
12:             **else**
13:                 $dp \leftarrow dp + 1$
14:     **until** *true*

Algorithm 4 applies Theorem 5.4 to approximate $x + y$ with precision p. If approximations $\frac{m}{n}$ and $\frac{m'}{n'}$ have the same sign, we do not lose precision by calculating $x + y$. On the other hand, if $\frac{m}{n}$ and $\frac{m'}{n'}$ have different signs, the amount of precision that is lost depends on the magnitude of $\frac{\min(|\frac{m}{n}|,|\frac{m'}{n'}|)}{\max(|\frac{m}{n}|,|\frac{m'}{n'}|)}$. This indicates that if $\frac{mm'}{nn'} < 0$ and $|\frac{m}{n}| \approx |\frac{m'}{n'}|$, a significant amount of precision can be lost in $x+y$. Thus, if the guaranteed precision for $x+y$ (i.e., $p-i$) is not sufficient, x and y must be recomputed with higher precisions (see Line 3,13 in Algorithm 4).

To confirm this observation, we apply the perturbation analysis of Section 4 on $f(x, y) = x + y$:

$$|\frac{xf_x(x,y)}{f(x,y)}| + |\frac{yf_y(x,y)}{f(x,y)}| = |\frac{x}{x+y}| + |\frac{y}{x+y}|$$

The quantity $|\frac{x}{x+y}| + |\frac{y}{x+y}|$ is 1 when $xy > 0$. Thus, we can estimate the loss of precision in $x + y$ independently of the arguments when $xy > 0$.

However, the quantity $|\frac{x}{x+y}| + |\frac{y}{x+y}|$ can be arbitrarily large when $xy < 0$ and $|x+y| \approx 0$. This confirms that a significant amount of precision can be lost in $x + y$.

Perturbation analysis shows that in general, we cannot estimate the amount of precision that is lost in $x + y$ independently of the arguments. Hence, if approximation $\frac{m}{n}$ and $\frac{m'}{n'}$ have different signs, recomputing x and y might be essential to obtain the desired precision for $x + y$. Loss of precision in $x + y$ is sometimes referred to as loss of significance [9] or catastrophic cancellation [2].

It should be noted that Theorem 5.4 does not imply that $x + y$ is always fundamentally problematic when $xy < 0$ and $|x + y| \approx 0$. In certain cases, the calculation can be adjusted in such a way that loss of significance can be avoided and the expression can be calculated in one pass.

Suppose we want to calculate $\sqrt{x+1} - \sqrt{x}$ for a relatively large x. Since $\sqrt{x+1} \approx \sqrt{x}$, we will lose a significant amount of precision if we directly calculate $\sqrt{x+1} - \sqrt{x}$. However, we can change the calculation algorithm by rewriting the expression as follows:

$$\sqrt{x+1} - \sqrt{x} = (\sqrt{x+1} - \sqrt{x}) \times \frac{\sqrt{x+1} + \sqrt{x}}{\sqrt{x+1} + \sqrt{x}} = \frac{1}{\sqrt{x+1} + \sqrt{x}}$$

In the new expression, all the operations can be approximated with a desired precision in one pass (see Section 5.5 on calculating square root). Hence, we can approximate the new expression without recomputing the sub-expressions with higher precisions. Applying the perturbation analysis of Section 4 also shows that $f(x) = \sqrt{x+1} - \sqrt{x}$ is not fundamentally problematic:

$$|\frac{xf'(x)}{f(x)}| = |\frac{x(\frac{1}{2\sqrt{x+1}} - \frac{1}{2\sqrt{x}})}{\sqrt{x+1} - \sqrt{x}}| = \frac{1}{2}\sqrt{\frac{x}{x+1}} < \frac{1}{2}$$

Using perturbation analysis, we can identify instances of $x+y$ where adjustments in the algorithm can avoid loss of significance. However, to our knowledge a general scheme for making such adjustments does not exist.

5.5 Square Root

In this section, we calculate $\sqrt{x}$ by approximating the root of $f(y) = y^2 - x$ using the Newton-Raphson method [9]. The Newton-Raphson method starts with an initial approximation y_0 for $\sqrt{x}$ and iteratively generates a sequence of

approximations. Assuming that the precise value of x is available, the sequence of approximations is generated by:

$$y_{n+1} = \frac{y_n^2 + x}{2y_n}$$

In what follows, we prove a theorem for approximating $\sqrt{x}$ by the Newton-Raphson method when an approximation (m, n, p) of x is available.

Theorem 5.5. *Let x be a real number represented by (m, n, p) such that:*

$$\frac{m}{n} = 0.b_1 \dots b_k \times 2^a$$

where $b_i \in \{0, 1\}$ for $1 \leq i \leq k$ and $b_1 = 1, a \in \mathbb{Z}$. Then $\sqrt{x}$ can be represented by $(m', n', p - 4N)$ where $N = \lceil \log_2(\log_2(2^{p+3} + 1)) \rceil + 1$ and $\frac{m'}{n'}$ is the N-th term of the following sequence:

$$y_{n+1} = \frac{y_n^2 + \frac{m}{n}}{2y_n}, \quad y_0 = 2^{\lceil \frac{a}{2} \rceil + 1}$$

Proof. From Definition 1 we write:

$$\frac{m}{n}\left(1 - \frac{1}{2^p}\right) < x < \frac{m}{n}\left(1 + \frac{1}{2^p}\right)$$

To prove the theorem, it suffices to show:

$$|\sqrt{x} - y_N| < \frac{1}{2^{p-4N}}|y_N| \tag{15}$$

We rewrite the left hand side of inequality (15):

$$|\sqrt{x} - y_N| = |\sqrt{x} - z_N + z_N - y_N| \leq |\sqrt{x} - z_N| + |z_N - y_N| \tag{16}$$

In inequality (16), z_N is the N-th term of the following sequence:

$$z_{n+1} = \frac{z_n^2 + x}{2z_n}, \quad z_0 = 2^{\lceil \frac{a}{2} \rceil + 1}$$

To prove inequality (15), it suffices to show that the following inequalities hold:

$$|\sqrt{x} - z_N| < \frac{1}{2^{p-4N+1}}|y_N| \tag{17}$$

$$|z_N - y_N| < \frac{1}{2^{p-4N+1}}|y_N| \tag{18}$$

Proof for inequality (17): First, we show that $y_n > \sqrt{\frac{x}{2}}$ for all $n \in \mathbb{N}$:

$$
\begin{aligned}
y_n - \sqrt{\frac{x}{2}} = \frac{y_{n-1}^2 + \frac{m}{n}}{2y_{n-1}} - \sqrt{\frac{x}{2}} &> \frac{y_{n-1}^2 + \frac{m}{n}}{2y_{n-1}} - \sqrt{\frac{m}{2n}\left(1 + \frac{1}{2^p}\right)} \\
&\geq \frac{y_{n-1}^2 + \frac{m}{n}}{2y_{n-1}} - \sqrt{\frac{m}{n}} \\
&= \frac{y_{n-1}^2 + \frac{m}{n} - 2y_{n-1}\sqrt{\frac{m}{n}}}{2y_{n-1}} = \frac{(y_{n-1} - \sqrt{\frac{m}{n}})^2}{2y_{n-1}} \geq 0
\end{aligned}
$$

15

Observe that as $y_n > \sqrt{\frac{x}{2}}$, inequality (17) is valid, if the following inequality holds:

$$|\sqrt{x} - z_N| < \frac{1}{2^{p-4N+\frac{3}{2}}}\sqrt{x} \tag{19}$$

To prove inequality (19), we find N such that:

$$z_N - \sqrt{x} = \frac{1}{2^{p+2}}\sqrt{x} \tag{20}$$

We calculate the quantity $\frac{z_N+\sqrt{x}}{z_N-\sqrt{x}}$:

$$\frac{z_N + \sqrt{x}}{z_N - \sqrt{x}} = \frac{\frac{z_{N-1}^2+x}{2z_{N-1}} + \sqrt{x}}{\frac{z_{N-1}^2+x}{2z_{N-1}} - \sqrt{x}} = \frac{z_{N-1}^2 + x + 2z_{N-1}\sqrt{x}}{z_{N-1}^2 + x - 2z_{N-1}\sqrt{x}}$$

$$= \left(\frac{z_{N-1} + \sqrt{x}}{z_{N-1} - \sqrt{x}}\right)^2 = \ldots = \left(\frac{z_0 + \sqrt{x}}{z_0 - \sqrt{x}}\right)^{2^N} \tag{21}$$

Suppose equality (20) holds for N. We can rewrite the right hand side of equality (21) as follows:

$$\left(\frac{z_0 + \sqrt{x}}{z_0 - \sqrt{x}}\right)^{2^N} = \frac{z_N + \sqrt{x}}{z_N - \sqrt{x}} = \frac{(2 + \frac{1}{2^{p+2}})\sqrt{x}}{(\frac{1}{2^{p+2}})\sqrt{x}} = 2^{p+3} + 1 \tag{22}$$

The index N that guarantees the required precision of equality (20) can be calculated from equality (22):

$$2^N \log_2\left(\frac{z_0 + \sqrt{x}}{z_0 - \sqrt{x}}\right) = \log_2(2^{p+3} + 1)$$

$$N = \log_2(\log_2(2^{p+3} + 1)) - \log_2\left(\log_2\left(\frac{z_0 + \sqrt{x}}{z_0 - \sqrt{x}}\right)\right) \tag{23}$$

To guarantee that N is well-defined, we show that $z_0 > x$. To prove the inequality, we use the assumptions $\frac{m}{n} = 0.b_1\ldots b_k \times 2^a, b_i \in \{0,1\}$ for $i = 1, \ldots, k$ and $b_0 = 1, z_0 = 2^{\lceil\frac{a}{2}\rceil+1}$:

$$\sqrt{x} < \sqrt{\frac{m}{n}\left(1 + \frac{1}{2^p}\right)} \leq \sqrt{0.b_1\ldots b_k} \times 2^{\frac{a}{2}} \times \sqrt{2} < 2^{\frac{a}{2}} \times \sqrt{2} < z_0$$

To obtain an estimation for N from equality (23), we calculate an upper bound for $\frac{z_0}{\sqrt{x}}$:

$$\frac{z_0}{\sqrt{x}} < \frac{2^{\lceil\frac{a}{2}\rceil+1}}{\sqrt{\frac{m}{n}(1 - \frac{1}{2^p})}} < \frac{2^{\frac{a+1}{2}+1}}{\sqrt{0.b_1\ldots b_k} \times 2^{\frac{a}{2}} \times \sqrt{\frac{1}{2}}} \leq \frac{2\sqrt{2}}{\sqrt[4]{2} \times \sqrt{\frac{1}{2}}} = \frac{4}{\sqrt[4]{2}}$$

In the worst case, the initial approximation z_0 differs from $\sqrt{x}$ by a factor $\frac{4}{\sqrt[4]{2}}$. We use this estimation in equality (23) to calculate the number of iterations for the Newton-Raphson method:

$$N = \log_2(\log_2(2^{p+3}+1)) - \log_2(\log_2(\frac{\frac{4}{\sqrt[4]{2}}+1}{\frac{4}{\sqrt[4]{2}}-1}))$$

$$< \lceil \log_2(\log_2(2^{p+3}+1)) \rceil + 1 \tag{24}$$

Proof for inequality (18)**:** To prove inequality (18), we consider the calculations in $z_N = \frac{z_{N-1}^2+x}{2z_{N-1}}$ and estimate the amount of error that we commit in the approximation $y_N = \frac{y_{N-1}^2+\frac{m}{n}}{2y_{N-1}}$.

Let $P(k)$ denote the amount of precision that we lose when we approximate z_k by y_k. Thus, we lose $P(N-1)$ units of precision if we approximate z_{N-1} by y_{N-1}:

$$|z_{N-1} - y_{N-1}| < \frac{1}{2^{p-P(N-1)}}|y_{N-1}|$$

The precision is reduced by 2 units when z_{N-1}^2 is approximated by y_{N-1}^2 (see Theorem 5.2):

$$|z_{N-1}^2 - y_{N-1}^2| < \frac{1}{2^{p-P(N-1)-2}}|y_{N-1}^2|$$

We approximate $z_{N-1}^2 + x$ by $y_{N-1}^2 + \frac{m}{n}$. We do not lose precision in this approximation (see Theorem 5.4.i):

$$|(z_{N-1}^2+x)-(y_{N-1}^2+\frac{m}{n})| < \frac{1}{2^{p-P(N-1)-2}}|y_{N-1}^2+\frac{m}{n}| \tag{25}$$

Given the approximation y_{N-1} of z_{N-1}, one unit of precision is lost in the approximation of $\frac{1}{z_{N-1}}$ (see Theorem 5.3):

$$|\frac{1}{z_{N-1}} - \frac{1}{y_{N-1}}| < \frac{1}{2^{p-P(N-1)-1}}|\frac{1}{y_{N-1}}|$$

$$|\frac{1}{2z_{N-1}} - \frac{1}{2y_{N-1}}| < \frac{1}{2^{p-P(N-1)-1}}|\frac{1}{2y_{N-1}}| \tag{26}$$

Finally, we approximate $\frac{z_{N-1}^2+x}{2z_{N-1}}$ based on the approximations described in inequality (25) and (26) (see Theorem 5.2):

$$|z_N - y_N| = |\frac{z_{N-1}^2+x}{2z_{N-1}} - \frac{y_{N-1}^2+\frac{m}{n}}{2y_{N-1}}| < \frac{1}{2^{p-P(N-1)-4}}|\frac{y_{N-1}^2+\frac{m}{n}}{2y_{N-1}}| \tag{27}$$

From inequality (27) we obtain the following recursive formula:

$$P(N) = P(N-1)+4$$

Since $y_0 = z_0 = 2^{\lceil \frac{a}{2} \rceil + 1}$, we lose $P(N) = P(0) + 4N = 4N$ units of precision in our approximation of z_N. We apply the number of iterations calculated in inequality (24) and obtain:

$$P(N) = 4N < 4\lceil \log_2(\log_2(2^{p+3} + 1)) \rceil + 4$$

$\square$

Algorithm 5 Square Root

Require: *expr* has the shape $\sqrt{x}$
1: **procedure** COMPUTE(*expr, p*)
2:    Choose p_x such that $p_x \geq p + 4\lceil \log_2 \left(\log_2(2^{p_x+3} + 1) \right) \rceil + 4$
3:    $N \leftarrow \lceil \log_2 \left(\log_2(2^{p_x+3} + 1) \right) \rceil + 1$
4:    $\frac{m}{n} \leftarrow$ COMPUTE(*x, p_x*)
5:    **if** $\frac{m}{n} < 0$ **then**
6:        "Undefined operation"
7:    **else**
8:        $\triangleright$ $\frac{m}{n}$ can be represented as $0.b_1 \ldots b_k \times 2^a$
9:        $\triangleright$ $b_i \in \{0,1\}$ for $1 \leq i \leq k, b_1 = 1$ and $a \in \mathbb{Z}$
10:        $a \leftarrow \lfloor \log_2(\frac{m}{n}) \rfloor + 1$
11:        $y_0 \leftarrow 2^{\lceil \frac{a}{2} \rceil + 1}$
12:        **for** $i = 1$ to N **do**
13:            $y_i \leftarrow \frac{y_{i-1}^2 + \frac{m}{n}}{2y_{i-1}}$
14:        $\frac{m'}{n'} \leftarrow y_N$
15:        **return** $\frac{m'}{n'}$

Algorithm 5 applies Theorem 5.5 to approximate $\sqrt{x}$ with precision p. As indicated in Theorem 5.5, loss of precision in the square root can be estimated independently of the argument x and hence Algorithm 5 approximates $\sqrt{x}$ in one pass. We apply perturbation analysis on $f(x) = \sqrt{x}$ to show this:

$$\left| \frac{xf'(x)}{f(x)} \right| = \left| \frac{(x)(\frac{1}{2\sqrt{x}})}{\sqrt{x}} \right| = \frac{1}{2}$$

The quantity $|\frac{xf'(x)}{f(x)}|$ is a small constant. Thus, in our top down approach, we can approximate $\sqrt{x}$ without iterative computations.

6 Approximating Transcendental Functions by Riemann Sums

In this section we introduce approximations for e^x (Section 6.1), $\ln(x)$ (Section 6.2), and $\arctan(x)$ (Section 6.3). We use Riemann sums to approximate these functions.
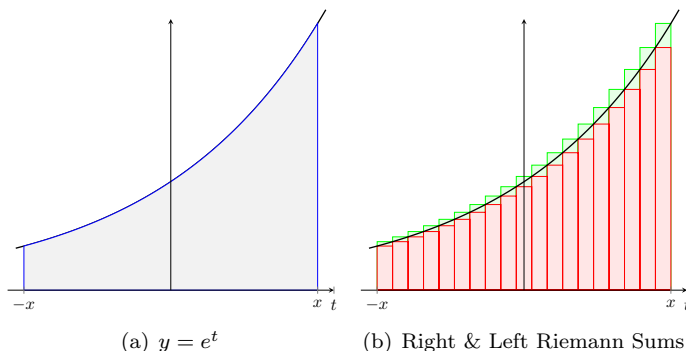
(a) $y = e^t$ (b) Right & Left Riemann Sums

Figure 1: Approximating e^x

The COMPUTE($expr, p$) function introduced in Section 5 will be extended to allow the approximation of $e^x, \ln(x)$ and $\arctan(x)$ with a given precision p. Perturbation analysis will also be used to identify computational problems in which iterative computations are unavoidable.

6.1 Exponential Function

To approximate the exponential function by Riemann sums, we first provide a simple approximation for e^x where we assume that x is precise. Then we extend this calculation to approximate e^x where x is represented by (m, n, p).

Suppose $x > 0$. To calculate e^x we consider the curve $y = e^t$ and calculate the area enclosed by this curve and the t-axis between $t = -x$ and $t = x$ as follows (see Fig. 1(a)):

$$\int_{-x}^{x} e^t dt = e^x - e^{-x}$$

We use Riemann sums to approximate this area; Fig. 1(b) shows two approximations from above and below using rectangles. Thus, we get the following inequalities for N rectangles:

$$\sum_{i=0}^{N-1} \frac{2x}{N} e^{-x + \frac{2x}{N}i} \leq e^x - e^{-x} \leq \sum_{i=1}^{N} \frac{2x}{N} e^{-x + \frac{2x}{N}i} \tag{28}$$

We rewrite inequality (28) as follows:

$$\sum_{i=0}^{N-1} \left(\frac{2x}{N}\right) e^{\frac{2x}{N}i} \leq e^{2x} - 1 \leq \sum_{i=1}^{N} \left(\frac{2x}{N}\right) e^{\frac{2x}{N}i}$$

$$\left(\frac{2x}{N}\right)\left(\frac{e^{2x} - 1}{e^{\frac{2x}{N}} - 1}\right) \leq e^{2x} - 1 \leq \left(\frac{2x}{N}\right)\left(\frac{e^{\frac{2x}{N}}(e^{2x} - 1)}{e^{\frac{2x}{N}} - 1}\right)$$

$$\frac{2x}{N} \leq e^{\frac{2x}{N}} - 1 \leq \frac{2x}{N} e^{\frac{2x}{N}} \tag{29}$$

19

We assume that $N > 2x$ and calculate an upper bound and a lower bound for e^x from inequality (29):

$$(1 + \frac{2x}{N})^{\frac{N}{2}} \leq e^x \leq (\frac{N}{N - 2x})^{\frac{N}{2}} \tag{30}$$

We can estimate the precision of the approximations calculated in inequality (30). For example, we can approximate e^x by $(\frac{N}{N-2x})^{\frac{N}{2}}$ and the absolute error of this approximation can be calculated as follows:

$$|(\frac{N}{N - 2x})^{\frac{N}{2}} - e^x| \leq |(\frac{N}{N - 2x})^{\frac{N}{2}} - (1 + \frac{2x}{N})^{\frac{N}{2}}|$$

$$= ((\frac{N}{N - 2x}) - (\frac{N + 2x}{N})) \sum_{i=0}^{\frac{N}{2}-1} (\frac{N}{N - 2x})^i (\frac{N + 2x}{N})^{\frac{N}{2}-i-1}$$

$$= \frac{4x^2}{N(N - 2x)} \sum_{i=0}^{\frac{N}{2}-1} (\frac{N}{N - 2x})^i (\frac{N + 2x}{N})^{\frac{N}{2}-i-1}$$

$$\leq \frac{4x^2}{N(N - 2x)} \sum_{i=0}^{\frac{N}{2}-1} (\frac{N}{N - 2x})^{\frac{N}{2}-1} = \frac{2x^2}{N} (\frac{N}{N - 2x})^{\frac{N}{2}} \tag{31}$$

For the last inequality we apply $\frac{N}{N-2x} \geq \frac{N+2x}{N}$.

In the discussion above, we have treated x as a precise value. In the following theorem, we extend this calculation and describe an approximation of e^x that relies on a representation (m, n, p) of x. To simplify our approximations, we first assume that $|\frac{m}{n}| < 1$. Afterwards, we extend our approximations to an arbitrary (m, n, p).

Theorem 6.1. *Let x be a real number represented by (m, n, p) and $|\frac{m}{n}| < 1$. Suppose N is a natural number such that $N > 2^{\frac{p+11}{3}}$. The value of e^x can be approximated as follows:*

i. If $0 < \frac{m}{n} < 1$ then e^x can be represented by $(m', n', p - 2\lceil\log_2(\frac{N}{2})\rceil - 3)$ where $\frac{m'}{n'} = (\frac{N}{N - \frac{2m}{n}})^{\frac{N}{2}}$.

ii. If $-1 < \frac{m}{n} < 0$ then e^x can be represented by $(m', n', p - 2\lceil\log_2(\frac{N}{2})\rceil - 4)$ where $\frac{m'}{n'} = \frac{1}{(\frac{N}{N + \frac{2m}{n}})^{\frac{N}{2}}}$.

Proof.

i. Suppose $0 < \frac{m}{n} < 1$. Since x is represented by (m, n, p), we have:

$$0 < \frac{m}{n}(1 - \frac{1}{2^p}) < x < \frac{m}{n}(1 + \frac{1}{2^p}) \leq \frac{2m}{n} < 2 \tag{32}$$

To prove the theorem, it suffices to show that:

$$|e^x - (\frac{N}{N - \frac{2m}{n}})^{\frac{N}{2}}| < \frac{1}{2^{p-2\lceil \log_2(\frac{N}{2})\rceil - 3}}|(\frac{N}{N - \frac{2m}{n}})^{\frac{N}{2}}| \qquad (33)$$

We rewrite the left hand side of inequality (33) as follows:

$$|e^x - (\frac{N}{N - \frac{2m}{n}})^{\frac{N}{2}}| = |e^x - (\frac{N}{N - 2x})^{\frac{N}{2}} + (\frac{N}{N - 2x})^{\frac{N}{2}} - (\frac{N}{N - \frac{2m}{n}})^{\frac{N}{2}}|$$

$$\leq |e^x - (\frac{N}{N - 2x})^{\frac{N}{2}}| + |(\frac{N}{N - 2x})^{\frac{N}{2}} - (\frac{N}{N - \frac{2m}{n}})^{\frac{N}{2}}|$$

To prove inequality (33), it suffices to show that the following inequalities hold:

$$|e^x - (\frac{N}{N - 2x})^{\frac{N}{2}}| < \frac{1}{2^{p-2\lceil \log_2(\frac{N}{2})\rceil - 2}}|(\frac{N}{N - \frac{2m}{n}})^{\frac{N}{2}}| \qquad (34)$$

$$|(\frac{N}{N - 2x})^{\frac{N}{2}} - (\frac{N}{N - \frac{2m}{n}})^{\frac{N}{2}}| < \frac{1}{2^{p-2\lceil \log_2(\frac{N}{2})\rceil - 2}}|(\frac{N}{N - \frac{2m}{n}})^{\frac{N}{2}}| \qquad (35)$$

Proof for inequality (34): From inequality (31), we obtain an upper-bound for the left hand side of inequality (34):

$$|e^x - (\frac{N}{N - 2x})^{\frac{N}{2}}| \leq \frac{2x^2}{N}(\frac{N}{N - 2x})^{\frac{N}{2}} \qquad (36)$$

To calculate an upper bound for the right hand side of inequality (36), we consider the function $f(x) = \frac{2x^2}{N}(\frac{N}{N-2x})^{\frac{N}{2}}$ and calculate its derivative:

$$f'(x) = \frac{4x}{N}(\frac{N}{N - 2x})^{\frac{N}{2}} + x^2(\frac{N}{N - 2x})^{\frac{N}{2} - 1}(\frac{2N}{(N - 2x)^2})$$

From inequality (32) we obtain $x \in (0, 2)$. We choose:

$$N > 4 > 2x \qquad (37)$$

to ensure that $f(x)$ is increasing for $x \in (0, 2)$, i.e., $f'(x) > 0$. We rewrite inequality (36) as follows:

$$|e^x - (\frac{N}{N - 2x})^{\frac{N}{2}}| \leq f(x) \leq f(2) = (\frac{8}{N})(\frac{N}{N - 4})^{\frac{N}{2}} \qquad (38)$$

We calculate a lower bound for the right hand side of inequality (34) as follows:

$$\frac{1}{2^{p-2\lceil \log_2(\frac{N}{2})\rceil - 2}}|(\frac{N}{N - \frac{2m}{n}})^{\frac{N}{2}}| > \frac{1}{2^{p-2 \log_2(\frac{N}{2}) - 2}} \qquad (39)$$

21

Thus, to prove inequality (34) it suffices to show that the following inequality holds (see inequality (38), (39)):

$$(\frac{8}{N})(\frac{N}{N-4})^{\frac{N}{2}} < \frac{1}{2^{p-2\log_2(\frac{N}{2})-2}}$$

This is equivalent to the following:

$$3 - \log_2(N) + \frac{N}{2}\left(\log_2(1 + \frac{4}{N-4})\right) < -p + 2\log_2(\frac{N}{2}) + 2 \qquad (40)$$

We choose $N > 8$ and apply Proposition 3 (see Appendix A) to obtain an upper bound for $\log_2(1 + \frac{4}{N-4})$:

$$\log_2(1 + \frac{4}{N-4}) < \frac{4}{(N-4)\ln(2)} < \frac{8}{N-4} \qquad (41)$$

Based on inequality (40),(41), it is sufficient to find an $N > 8$ satisfying:

$$3 - \log_2(N) + (\frac{N}{2})(\frac{8}{N-4}) < -p + 2\log_2(\frac{N}{2}) + 2 \qquad (42)$$

Inequality (42) is equivalent to the following:

$$-\log_2(\frac{N^3}{4}) + \frac{16}{N-4} < -p - 5$$

From $N > 8$, we conclude $\frac{16}{N-4} < 4$. Thus, we choose N such that $N > \max(2^{\frac{p+11}{3}}, 8) = 2^{\frac{p+11}{3}}$.

Proof for inequality (35)**:** To prove the inequality, we estimate the amount of precision that is lost when we approximate $(\frac{N}{N-2x})^{\frac{N}{2}}$ by $(\frac{N}{N-\frac{2m}{n}})^{\frac{N}{2}}$.

The number x is represented by (m, n, p). Thus, we have:

$$|x - \frac{m}{n}| < \frac{1}{2^p}|\frac{m}{n}|$$
$$|2x - \frac{2m}{n}| < \frac{1}{2^p}|\frac{2m}{n}|$$

Since $N > 8$, one unit of precision is lost when we approximate $N - 2x$ by $N - \frac{2m}{n}$ (see Theorem 5.4.ii):

$$|(N - 2x) - (N - \frac{2m}{n})| < \frac{1}{2^{p-1}}|N - \frac{2m}{n}|$$

Approximating $\frac{1}{N-2x}$ by $\frac{1}{N-\frac{2m}{n}}$ reduces the precision by one unit (see Theorem 5.3):

$$|\frac{1}{N-2x} - \frac{1}{N-\frac{2m}{n}}| < \frac{1}{2^{p-2}}|\frac{1}{N-\frac{2m}{n}}|$$
$$|\frac{N}{N-2x} - \frac{N}{N-\frac{2m}{n}}| < \frac{1}{2^{p-2}}|\frac{N}{N-\frac{2m}{n}}|$$

22

Finally, approximating $(\frac{N}{N-2x})^{\frac{N}{2}}$ reduces the precision by $2\lceil\log_2(\frac{N}{2})\rceil$ units (see Lemma 1 in Appendix A):

$$|(\frac{N}{N-2x})^{\frac{N}{2}} - (\frac{N}{N-\frac{2m}{n}})^{\frac{N}{2}}| < \frac{1}{2^{p-2\lceil\log_2(\frac{N}{2})\rceil-2}}|(\frac{N}{N-\frac{2m}{n}})^{\frac{N}{2}}|$$

ii. Suppose $\frac{m}{n} < 0$. We use the following identity to calculate e^x:

$$e^x = \frac{1}{e^{-x}}$$

We represent $-x$ by $(-m, n, p)$ (see Theorem 5.1). Then, we apply the first part of the theorem and Theorem 5.3 to approximate e^{-x} and $\frac{1}{e^{-x}}$, respectively.

$\square$

In what follows, we extend the approximations of Theorem 6.1 and calculate the exponential function for x represented by (m, n, p) where $|\frac{m}{n}| \geq 1$.

Theorem 6.2. *Let x be a real number represented by (m, n, p) and $|\frac{m}{n}| \geq 1$. Suppose k and N are natural numbers such that:*

$$|\frac{m}{2^k n}| < 1 \ , \ N > 2^{\frac{p+11}{3}}$$

The value of e^x can be approximated as follows:

i. *If $\frac{m}{n} > 0$ then e^x can be represented by $(m', n', p - 2\lceil\log_2(\frac{N}{2})\rceil - 2k - 3)$ where $\frac{m'}{n'} = (\frac{N}{N-\frac{2m}{n}})^{N\cdot2^{k-1}}$.*

ii. *If $\frac{m}{n} < 0$ then e^x can be represented by $(m', n', p - 2\lceil\log_2(\frac{N}{2})\rceil - 2k - 4)$ where $\frac{m'}{n'} = \frac{1}{(\frac{N}{N+\frac{2m}{n}})^{N\cdot2^{k-1}}}$.*

Proof. Since $|\frac{m}{n}| \geq 1$, we choose $k \in \mathbb{N}$ such that $|\frac{m}{2^k n}| < 1$. We use the following identity to calculate e^x:

$$e^x = (e^{\frac{x}{2^k}})^{2^k} \tag{43}$$

We approximate $\frac{x}{2^k}$ by $\frac{m}{2^k n}$. Since 2^k is a constant, we do not lose precision in this approximation. We apply Theorem 7.2 to approximate $e^{\frac{x}{2^k}}$. This approximation reduces the precision by:

- $2\lceil\log_2(\frac{N}{2})\rceil + 3$ units, if $0 < \frac{m}{2^k n} < 1$;

- $2\lceil\log_2(\frac{N}{2})\rceil + 4$ units, if $-1 < \frac{m}{2^k n} < 0$.

Suppose $\frac{m'}{n'}$ is the approximation obtained for $e^{\frac{x}{2^k}}$ from Theorem 6.1. We approximate $(e^{\frac{x}{2^k}})^{2^k}$ by $(\frac{m'}{n'})^{2^k}$; we lose $2k$ units of precision in this calculation (see Lemma 1 in Appendix A).

$\square$

23

Algorithm 6 Exponential Function

Require: *expr* has the shape e^x

1: **procedure** COMPUTE(*expr*, *p*)
2:    Choose N such that $N > 2^{\frac{p+11}{3}}$
3:    $p_x \leftarrow p + 2\lceil \log_2(\frac{N}{2})\rceil + 4$
4:    **repeat**
5:       $\frac{m}{n} \leftarrow$ COMPUTE(x, p_x)
6:       **if** $0 < \frac{m}{n} < 1$ **then** ▷ Theorem 6.1.i
7:          $\frac{m'}{n'} \leftarrow \left(\frac{N}{N - \frac{2m}{n}}\right)^{\frac{N}{2}}$
8:          **return** $\frac{m'}{n'}$
9:       **else if** $-1 < \frac{m}{n} < 0$ **then** ▷ Theorem 6.1.ii
10:          $\frac{m'}{n'} \leftarrow \frac{1}{\left(\frac{N}{N + \frac{2m}{n}}\right)^{\frac{N}{2}}}$
11:          **return** $\frac{m'}{n'}$
12:       **else**
13:          Choose $k \in \mathbb{N}$ such that $|\frac{m}{2^k n}| < 1$
14:          **if** $(\frac{m}{n} > 0) \wedge$ ▷ Theorem 6.2.i
15:             $(p_x - 2\lceil \log_2(\frac{N}{2})\rceil - 2k - 3 \geq p)$ **then**
16:             $\frac{m'}{n'} \leftarrow \left(\frac{N}{N - \frac{2m}{n}}\right)^{N \cdot 2^{k-1}}$
17:             **return** $\frac{m'}{n'}$
18:          **else if** $(\frac{m}{n} < 0) \wedge$ ▷ Theorem 6.2.ii
19:             $(p_x - 2\lceil \log_2(\frac{N}{2})\rceil - 2k - 4 \geq p)$ **then**
20:             $\frac{m'}{n'} \leftarrow \frac{1}{\left(\frac{N}{N + \frac{2m}{n}}\right)^{N \cdot 2^{k-1}}}$
21:             **return** $\frac{m'}{n'}$
22:          **else**
23:             $p_x \leftarrow p_x + 1$
24:    **until** *true*

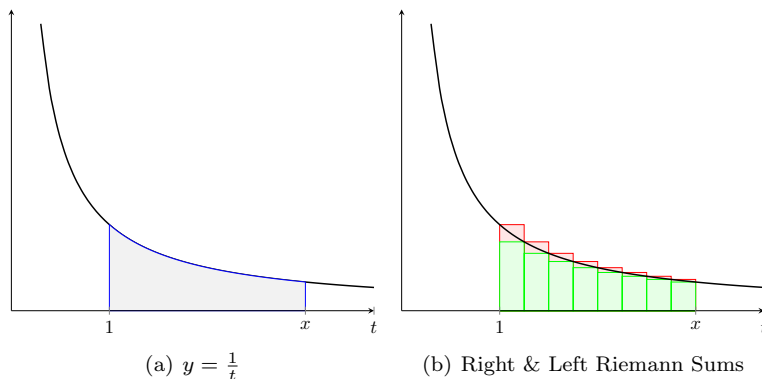(a) $y = \frac{1}{t}$ (b) Right & Left Riemann Sums

Figure 2: Approximating $\ln(x)$

Algorithm 6 implements the approximations described by Theorem 6.1 and 6.2 to calculate e^x with arbitrary precision. Observe that when $|\frac{m}{n}| < 1$, e^x can be approximated in one pass. However, when $|\frac{m}{n}| \geq 1$, loss of precision depends on the magnitude of $|\frac{m}{n}|$. Thus, recomputing x with higher precisions might be necessary to compensate for the loss of precision caused by applying equality (43) (see Line 4,23 in Algorithm 6).

To confirm that iterative computations are unavoidable, we apply perturbation analysis on $f(x) = e^x$:

$$|\frac{xf'(x)}{f(x)}| = |\frac{xe^x}{e^x}| = |x|$$

The quantity $|x|$ can become arbitrarily large and hence approximating e^x with precision p in one pass is not always possible.

6.2 Natural Logarithm

In this section, we first discuss an approximation for $\ln(x)$ based on Riemann sums where we assume that x is precise. Then we extend this calculation to approximate $\ln(x)$ where x is represented by (m, n, p).

Suppose $x > 1$ is a real number and we want to approximate $\ln(x)$. We consider the curve $y = \frac{1}{t}$ (see Fig. 2(a)) and calculate the area enclosed by this curve and the t-axis between $t = 1$ and $t = x$. This area can be calculated as follows:

$$\int_1^x \frac{dt}{t} = \ln(x)$$

We use Riemann sums to approximate this area; Fig. 2(b) shows how the area can be approximated from below and above using rectangles. Thus, we get the

following inequalities for N rectangles:

$$\frac{x-1}{N}\sum_{i=1}^{N}\frac{1}{1+\frac{i}{N}(x-1)} \leq \ln(x) \leq \frac{x-1}{N}\sum_{i=0}^{N-1}\frac{1}{1+\frac{i}{N}(x-1)}$$

This gives us an upper bound and a lower bound for $\ln(x)$ and by increasing N we get more precise approximations.

We can estimate the precision of our approximations. For instance, if we approximate $\ln(x)$ by $\frac{x-1}{N}\sum_{i=0}^{N-1}\frac{1}{1+\frac{i}{N}(x-1)}$ the absolute error can be estimated as follows:

$$|\frac{x-1}{N}\sum_{i=0}^{N-1}\frac{1}{1+\frac{i}{N}(x-1)}-\ln(x)|$$

$$\leq |\frac{x-1}{N}\sum_{i=0}^{N-1}\frac{1}{1+\frac{i}{N}(x-1)}-\frac{x-1}{N}\sum_{i=1}^{N}\frac{1}{1+\frac{i}{N}(x-1)}|$$

$$=\frac{x-1}{N}(1-\frac{1}{x})=\frac{(x-1)^2}{Nx} \tag{44}$$

Up to this point, we have assumed that the precise value of x is available. In what follows, we formulate a theorem to describe an approximation of $\ln(x)$ based on a representation (m,n,p) of x.

Theorem 6.3. *Let x be a real number represented by (m,n,p) such that $p \geq 1$.*

i. If $\frac{m}{n} > 1$, then $\ln(x)$ can be represented by $(m',n',p-j-4)$ where $\frac{m'}{n'} = \frac{(\frac{m}{n})-1}{N}\sum_{i=0}^{N-1}\frac{1}{1+\frac{i}{N}((\frac{m}{n})-1)}$, $N=\lceil 2^{p-2}\frac{(\frac{m}{n})^2}{\frac{m}{n}-1}\rceil$, and j is the smallest natural number such that $j \geq \log_2(\frac{1+\frac{n}{m}}{1-\frac{n}{m}})$ holds.

ii. If $0 < \frac{m}{n} < 1$, then $\ln(x)$ can be represented by $(m',n',p-j-5)$ where $\frac{m'}{n'} = -\frac{(\frac{n}{m})-1}{N}\sum_{i=0}^{N-1}\frac{1}{1+\frac{i}{N}((\frac{n}{m})-1)}$, $N=\lceil 2^{p-2}\frac{(\frac{n}{m})^2}{\frac{n}{m}-1}\rceil$, and j is the smallest natural number such that $j \geq \log_2(\frac{1+\frac{m}{n}}{1-\frac{m}{n}})$ holds.

Proof.

i. Suppose $\frac{m}{n} > 1$. The number x is represented by (m,n,p) and hence we can write:

$$\frac{1}{2} < \frac{m}{2n} \leq \frac{m}{n}(1-\frac{1}{2^p}) < x < \frac{m}{n}(1+\frac{1}{2^p}) \leq \frac{3m}{2n} \tag{45}$$

To prove the theorem, we need to prove the following inequality:

$$|\ln(x)-\frac{\frac{m}{n}-1}{N}\sum_{i=0}^{N-1}\frac{1}{1+\frac{i}{N}(\frac{m}{n}-1)}|$$

$$< \frac{1}{2^{p-j-4}}|\frac{\frac{m}{n}-1}{N}\sum_{i=0}^{N-1}\frac{1}{1+\frac{i}{N}(\frac{m}{n}-1)}| \tag{46}$$

We rewrite the left hand side of inequality (46) as follows:

$$| \ln(x) - \frac{\frac{m}{n} - 1}{N} \sum_{i=0}^{N-1} \frac{1}{1 + \frac{i}{N}(\frac{m}{n} - 1)} | =$$

$$| \ln(x) - \frac{x - 1}{N} \sum_{i=0}^{N-1} \frac{1}{1 + \frac{i}{N}(x - 1)}$$

$$+ \frac{x - 1}{N} \sum_{i=0}^{N-1} \frac{1}{1 + \frac{i}{N}(x - 1)} - \frac{\frac{m}{n} - 1}{N} \sum_{i=0}^{N-1} \frac{1}{1 + \frac{i}{N}(\frac{m}{n} - 1)} | \leq$$

$$| \ln(x) - \frac{x - 1}{N} \sum_{i=0}^{N-1} \frac{1}{1 + \frac{i}{N}(x - 1)} |$$

$$+ | \frac{x - 1}{N} \sum_{i=0}^{N-1} \frac{1}{1 + \frac{i}{N}(x - 1)} - \frac{\frac{m}{n} - 1}{N} \sum_{i=0}^{N-1} \frac{1}{1 + \frac{i}{N}(\frac{m}{n} - 1)} |$$

To prove inequality (46), it suffices to show that the following inequalities hold:

$$| \ln(x) - \frac{x - 1}{N} \sum_{i=0}^{N-1} \frac{1}{1 + \frac{i}{N}(x - 1)} | < \frac{1}{2^{p-j-3}} | \frac{\frac{m}{n} - 1}{N} \sum_{i=0}^{N-1} \frac{1}{1 + \frac{i}{N}(\frac{m}{n} - 1)} | \tag{47}$$

$$| \frac{x - 1}{N} \sum_{i=0}^{N-1} \frac{1}{1 + \frac{i}{N}(x - 1)} - \frac{\frac{m}{n} - 1}{N} \sum_{i=0}^{N-1} \frac{1}{1 + \frac{i}{N}(\frac{m}{n} - 1)} |$$

$$< \frac{1}{2^{p-j-3}} | \frac{\frac{m}{n} - 1}{N} \sum_{i=0}^{N-1} \frac{1}{1 + \frac{i}{N}(\frac{m}{n} - 1)} | \tag{48}$$

Proof for inequality (47): We use inequality (44) and calculate an upper bound for the left hand side of inequality (47):

$$| \ln(x) - \frac{x - 1}{N} \sum_{i=0}^{N-1} \frac{1}{1 + \frac{i}{N}(x - 1)} | \leq \frac{(x - 1)^2}{Nx} \tag{49}$$

To calculate an upper bound for $\frac{(x-1)^2}{Nx}$, we consider the function $f(x) = \frac{(x-1)^2}{Nx}$ and calculate its derivative:

$$f'(x) = \frac{x^2 - 1}{Nx^2}$$

From inequality (45) we conclude that $x \in [\frac{1}{2}, \frac{3m}{2n}]$. The function $f(x)$ is decreasing ($f'(x) \leq 0$) in the interval $[\frac{1}{2}, 1]$ and increasing ($f'(x) \geq 0$)

in the interval $[1, \frac{3m}{2n}]$. Thus, the maximum of $f(x)$ for $x \in [\frac{1}{2}, \frac{3m}{2n}]$ is $\max(f(\frac{1}{2}), f(\frac{3m}{2n}))$. We use this to rewrite inequality (49):

$$|\ln(x) - \frac{x-1}{N} \sum_{i=0}^{N-1} \frac{1}{1 + \frac{i}{N}(x-1)}| \leq f(x) \leq \max(f(\frac{1}{2}), f(\frac{3m}{2n}))$$

$$= \max(\frac{1}{2N}, \frac{(\frac{3m}{2n} - 1)^2}{N(\frac{3m}{2n})})$$

$$\leq \max(\frac{1}{2N}, \frac{(\frac{3m}{2n})^2}{N(\frac{3m}{2n})}) = \frac{3m}{2Nn} \qquad (50)$$

We also calculate a lower bound for the right hand side of inequality (47):

$$\frac{1}{2^{p-j-3}}|\frac{\frac{m}{n}-1}{N} \sum_{i=0}^{N-1} \frac{1}{1 + \frac{i}{N}(\frac{m}{n} - 1)}| > \frac{1}{2^{p-3}}|\frac{(\frac{m}{n} - 1)}{N} \sum_{i=0}^{N-1} \frac{1}{1 + (\frac{N-1}{N})(\frac{m}{n} - 1)}|$$

$$= \frac{1}{2^{p-3}} \cdot (\frac{m}{n} - 1) \cdot \frac{N}{N + (N-1)(\frac{m}{n} - 1)}$$

$$> \frac{1}{2^{p-3}} \cdot (\frac{m}{n} - 1) \cdot \frac{N}{N(1 + \frac{m}{n} - 1)}$$

$$= \frac{(\frac{m}{n} - 1)}{2^{p-3}(\frac{m}{n})} \qquad (51)$$

To show that inequality (47) holds, it suffices to prove the following inequality (see inequality (50),(51)):

$$\frac{3m}{2Nn} < \frac{(\frac{m}{n} - 1)}{2^{p-3}(\frac{m}{n})}$$

Thus, it suffices to choose $N \geq 2^{p-2} \frac{(\frac{m}{n})^2}{\frac{m}{n}-1}$.

Proof for inequality (48)**:** To prove inequality (48) we consider the calculations in $\frac{x-1}{N} \sum_{i=0}^{N-1} \frac{1}{1 + \frac{i}{N}(x-1)}$ and estimate the amount of precision that we lose in the approximation $\frac{\frac{m}{n}-1}{N} \sum_{i=0}^{N-1} \frac{1}{1 + \frac{i}{N}(\frac{m}{n}-1)}$.

From Theorem 5.4.ii we conclude that j units of precision is lost by subtracting 1 from x where $j \geq \log_2(\frac{1 + \frac{n}{m}}{1 - \frac{n}{m}})$. We obtain the following inequalities:

$$|(x - 1) - (\frac{m}{n} - 1)| < \frac{1}{2^{p-j}}|\frac{m}{n} - 1|$$

$$|\frac{i}{N}(x - 1) - \frac{i}{N}(\frac{m}{n} - 1)| < \frac{1}{2^{p-j}}|\frac{i}{N}(\frac{m}{n} - 1)|$$

The approximation $\frac{i}{N}(\frac{m}{n} - 1)$ of $\frac{i}{N}(x - 1)$ is positive. Hence, we do not lose precision by adding $\frac{i}{N}(x - 1)$ and 1 (see Theorem 5.4.i).

$$|(1 + \frac{i}{N}(x - 1)) - (1 + \frac{i}{N}(\frac{m}{n} - 1))| < \frac{1}{2^{p-j}}|1 + \frac{i}{N}(\frac{m}{n} - 1)|$$

By approximating the inverse of $1 + \frac{i}{N}(x-1)$, we lose 1 unit of precision (see Theorem 5.3):

$$\left| \frac{1}{1 + \frac{i}{N}(x-1)} - \frac{1}{1 + \frac{i}{N}(\frac{m}{n}-1)} \right| < \frac{1}{2^{p-j-1}} \left| \frac{1}{1 + \frac{i}{N}(\frac{m}{n}-1)} \right|$$

We lose 2 units of precision by multiplying $\frac{x-1}{N}$ and $\frac{1}{1+\frac{i}{N}(x-1)}$ (see Theorem 5.2):

$$\left| \frac{x-1}{N} \cdot \frac{1}{1 + \frac{i}{N}(x-1)} - \frac{(\frac{m}{n}-1)}{N} \cdot \frac{1}{1 + \frac{i}{N}(\frac{m}{n}-1)} \right|$$
$$< \frac{1}{2^{p-j-3}} \left| \frac{(\frac{m}{n}-1)}{N} \cdot \frac{1}{1 + \frac{i}{N}(\frac{m}{n}-1)} \right|$$

Since the numbers $\frac{x-1}{N} \cdot \frac{1}{1+\frac{i}{N}(x-1)}$ are approximated by the positive numbers $\frac{\frac{m}{n}-1}{N} \cdot \frac{1}{1+\frac{i}{N}(\frac{m}{n}-1)}$, calculating the summation $\frac{x-1}{N} \sum_{i=0}^{N-1} \frac{1}{1+\frac{i}{N}(x-1)}$ does not affect the precision:

$$\left| \frac{x-1}{N} \sum_{i=0}^{N-1} \frac{1}{1 + \frac{i}{N}(x-1)} - \frac{\frac{m}{n}-1}{N} \sum_{i=0}^{N-1} \frac{1}{1 + \frac{i}{N}(\frac{m}{n}-1)} \right|$$
$$< \frac{1}{2^{p-j-3}} \left| \frac{\frac{m}{n}-1}{N} \sum_{i=0}^{N-1} \frac{1}{1 + \frac{i}{N}(\frac{m}{n}-1)} \right|$$

ii. Suppose $0 < \frac{m}{n} < 1$. We use the following identity to approximate $\ln(x)$:

$$\ln(x) = -\ln(\frac{1}{x})$$

We approximate $\frac{1}{x}$ by $(n, m, p-1)$ (see Theorem 5.3). Then, we apply the first part of the theorem and Theorem 5.1 to approximate $-\ln(\frac{1}{x})$.

$\square$

Algorithm 7 applies Theorem 6.3 to approximate $\ln(x)$ with arbitrary precision. Note that when the approximation $\frac{m}{n}$ is close to 1 the amount of precision that we lose in the calculations depends on the magnitude of $\frac{m}{n}$. Loss of precision for $x \approx 1$ in Algorithm 7 is due to our approximation formula, $\frac{x-1}{N} \sum_{i=0}^{N-1} \frac{1}{1+\frac{i}{N}(x-1)}$. We divide the interval between 1 and x into N subintervals and approximate the area under the curve $f(t) = \frac{1}{t}$. The length of the interval $[1, x]$ is crucial in our approximation. Thus, recomputing x with higher precisions is necessary when a significant amount of precision is lost in $x - 1$ (see Line 3,22 in Algorithm 7).

To show that approximating $\ln(x)$ for $x \approx 1$ is fundamentally problematic we apply perturbation analysis on $f(x) = \ln(x)$:

$$\left| \frac{x f'(x)}{f(x)} \right| = \left| \frac{x(\frac{1}{x})}{\ln(x)} \right| = \left| \frac{1}{\ln(x)} \right|$$

29

Algorithm 7 Natural Logarithm

Require: *expr* has the shape $\ln(x)$
1: **procedure** Compute($expr, p$)
2:     $p_x \leftarrow p + 5$
3:     **repeat**
4:         $\frac{m}{n} \leftarrow$ Compute(x, p_x)
5:         **if** $\frac{m}{n} > 1$ **then**
6:             $arg \leftarrow \frac{m}{n}$
7:             $\ell \leftarrow 4$
8:         **else if** $0 < \frac{m}{n} < 1$ **then**
9:             $arg \leftarrow \frac{n}{m}$
10:            $\ell \leftarrow 5$
11:         **else**
12:            "Undefined operation"
13:         $j \leftarrow \lceil \log_2(\frac{1+\frac{1}{arg}}{1-\frac{1}{arg}}) \rceil$
14:         **if** $p_x - j - \ell \geq p$ **then**
15:             $N \leftarrow \lceil 2^{p_x - 2} \frac{arg^2}{arg-1} \rceil$
16:             $\frac{m'}{n'} \leftarrow \frac{arg-1}{N} \sum_{i=0}^{N-1} \frac{1}{1+\frac{i}{N}(arg-1)}$
17:             **if** $\frac{m}{n} > 1$ **then**
18:                 **return** $\frac{m'}{n'}$
19:             **else**
20:                 **return** $\frac{-m'}{n'}$
21:         **else**
22:            $p_x \leftarrow p_x + 1$
23:     **until** *true*

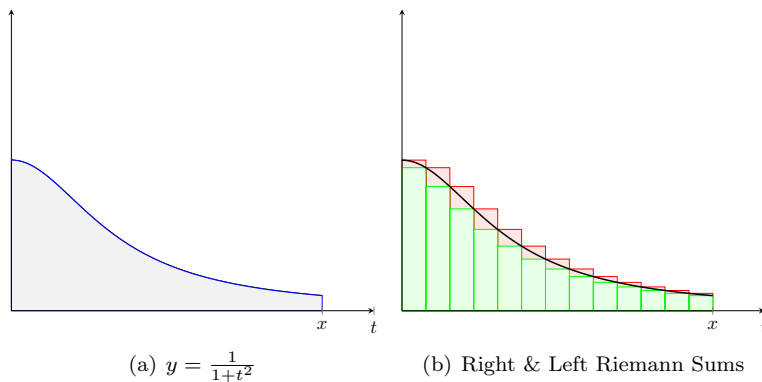(a) $y = \frac{1}{1+t^2}$ (b) Right & Left Riemann Sums

Figure 3: Approximating $\arctan(x)$

When $x \approx 1$ the quantity $\ln(x)$ is a small value and hence committing a small error in the approximation of x causes a significant error in calculating $\ln(x)$. Thus, for $x \approx 1$, iterative computations are unavoidable.

6.3 Arctangent

We first introduce an approximation of $\arctan(x)$ using Riemann sums. The assumption is that the precise value of x is available. Afterwards, we extend our calculations to introduce an approximation of $\arctan(x)$ based on a representation (m, n, p) of x.

Suppose that $x > 0$ is a real number and we want to approximate $\arctan(x)$. We consider the curve $y = \frac{1}{1+t^2}$; see Fig. 3(a). We calculate the area enclosed by this curve and the t-axis between $t = 0$ and $t = x$:

$$\int_0^x \frac{dt}{1 + t^2} = \arctan(x)$$

We approximate this area using Riemann sums. Fig. 3(b) shows approximations from above and below for the integral using rectangles. From Fig. 3(b) we can derive the following inequalities for N rectangles:

$$\frac{x}{N} \sum_{i=1}^{N} \frac{1}{1 + (\frac{i}{N})^2 x^2} \le \arctan(x) \le \frac{x}{N} \sum_{i=0}^{N-1} \frac{1}{1 + (\frac{i}{N})^2 x^2}$$

We want to estimate the precision of our approximations. Suppose we approximate $\arctan(x)$ by $\frac{x}{N} \sum_{i=0}^{N-1} \frac{1}{1+(\frac{i}{N})^2 x^2}$. The absolute error of this approxi-

mation can be estimated as follows:

$$|\frac{x}{N}\sum_{i=0}^{N-1}\frac{1}{1+(\frac{i}{N})^2 x^2} - \arctan(x)| \leq |\frac{x}{N}\sum_{i=0}^{N-1}\frac{1}{1+(\frac{i}{N})^2 x^2} - \frac{x}{N}\sum_{i=1}^{N}\frac{1}{1+(\frac{i}{N})^2 x^2}|$$

$$= \frac{x}{N}(1 - \frac{1}{1+x^2}) = \frac{x^3}{N(1+x^2)} \tag{52}$$

Up to this point, we have assumed that x is precisely calculated. In what follows, we assume that x is approximated by (m, n, p). We extend the Riemann sum calculation to compute $\arctan(x)$ using the given approximation of x.

Theorem 6.4. *Let x be a real number represented by (m, n, p). Then $\arctan(x)$ can be represented by $(m', n', p - 6)$ where $\frac{m'}{n'} = \frac{(\frac{m}{n})}{N}\sum_{i=0}^{N-1}\frac{1}{1+(\frac{i}{N})^2(\frac{m}{n})^2}$ and $N = 2^{p-2}\lceil(\frac{m}{n})^2\rceil$.*

Proof. We consider two cases:

1. Suppose $\frac{m}{n} > 0$. Since x is represented by (m, n, p) we can write:

$$\frac{m}{n}(1 - \frac{1}{2^p}) < x < \frac{m}{n}(1 + \frac{1}{2^p}) \leq \frac{2m}{n} \tag{53}$$

To prove the theorem, we should show that:

$$|\arctan(x) - \frac{(\frac{m}{n})}{N}\sum_{i=0}^{N-1}\frac{1}{1+(\frac{i}{N})^2(\frac{m}{n})^2}| < \frac{1}{2^{p-6}}|\frac{(\frac{m}{n})}{N}\sum_{i=0}^{N-1}\frac{1}{1+(\frac{i}{N})^2(\frac{m}{n})^2}| \tag{54}$$

We rewrite the left hand side of inequality (54) as follows:

$$|\arctan(x) - \frac{(\frac{m}{n})}{N}\sum_{i=0}^{N-1}\frac{1}{1+(\frac{i}{N})^2(\frac{m}{n})^2}| =$$

$$|\arctan(x) - \frac{x}{N}\sum_{i=0}^{N-1}\frac{1}{1+(\frac{i}{N})^2 x^2}$$

$$+ \frac{x}{N}\sum_{i=0}^{N-1}\frac{1}{1+(\frac{i}{N})^2 x^2} - \frac{(\frac{m}{n})}{N}\sum_{i=0}^{N-1}\frac{1}{1+(\frac{i}{N})^2(\frac{m}{n})^2}| \leq$$

$$|\arctan(x) - \frac{x}{N}\sum_{i=0}^{N-1}\frac{1}{1+(\frac{i}{N})^2 x^2}|$$

$$+ |\frac{x}{N}\sum_{i=0}^{N-1}\frac{1}{1+(\frac{i}{N})^2 x^2} - \frac{(\frac{m}{n})}{N}\sum_{i=0}^{N-1}\frac{1}{1+(\frac{i}{N})^2(\frac{m}{n})^2}|$$

Thus, to prove inequality (54), it suffices to prove the following inequalities:

$$|\arctan(x) - \frac{x}{N}\sum_{i=0}^{N-1}\frac{1}{1+(\frac{i}{N})^2 x^2}| < \frac{1}{2^{p-5}}|\frac{(\frac{m}{n})}{N}\sum_{i=0}^{N-1}\frac{1}{1+(\frac{i}{N})^2(\frac{m}{n})^2}| \quad (55)$$

$$|\frac{x}{N}\sum_{i=0}^{N-1}\frac{1}{1+(\frac{i}{N})^2 x^2} - \frac{(\frac{m}{n})}{N}\sum_{i=0}^{N-1}\frac{1}{1+(\frac{i}{N})^2(\frac{m}{n})^2}|$$

$$< \frac{1}{2^{p-5}}|\frac{(\frac{m}{n})}{N}\sum_{i=0}^{N-1}\frac{1}{1+(\frac{i}{N})^2(\frac{m}{n})^2}| \quad (56)$$

Proof for inequality (55): We first consider the left hand side of inequality (55) and calculate an upper bound for it. From inequality (52) we can write:

$$|\arctan(x) - \frac{x}{N}\sum_{i=0}^{N-1}\frac{1}{1+(\frac{i}{N})^2 x^2}| \leq \frac{x^3}{N(1+x^2)} \quad (57)$$

To obtain an upper-bound for $\frac{x^3}{N(1+x^2)}$, we consider the function $f(x) = \frac{x^3}{N(1+x^2)}$ and calculate its derivative:

$$f'(x) = \frac{3x^2 + x^4}{N(1+x^2)^2} > 0$$

Thus, $f(x)$ is an increasing function and its maximum occurs when x gets its maximum value. Inequality (53) implies that $\frac{2m}{n}$ is an upper bound for x and hence we can rewrite inequality (57) as follows:

$$|\arctan(x) - \frac{x}{N}\sum_{i=0}^{N-1}\frac{1}{1+(\frac{i}{N})^2 x^2}| \leq \frac{x^3}{N(1+x^2)} \leq \frac{8(\frac{m}{n})^3}{N(1+4(\frac{m}{n})^2)}$$

$$< \frac{8(\frac{m}{n})^3}{N(1+(\frac{m}{n})^2)} \quad (58)$$

We also calculate a lower bound for the right hand side of inequality (55):

$$\frac{1}{2^{p-5}}|\frac{(\frac{m}{n})}{N}\sum_{i=0}^{N-1}\frac{1}{1+(\frac{i}{N})^2(\frac{m}{n})^2}| > (\frac{1}{2^{p-5}})\frac{(\frac{m}{n})}{N}\sum_{i=0}^{N-1}\frac{1}{1+(\frac{N-1}{N})^2(\frac{m}{n})^2}$$

$$= (\frac{1}{2^{p-5}}).(\frac{m}{n}).(\frac{N^2}{N^2 + (N-1)^2(\frac{m}{n})^2})$$

$$> (\frac{1}{2^{p-5}}).(\frac{m}{n}).(\frac{N^2}{N^2(1+(\frac{m}{n})^2)})$$

$$= \frac{(\frac{m}{n})}{2^{p-5}(1+(\frac{m}{n})^2)} \quad (59)$$

To prove inequality (55), it suffices to show that the following inequality holds (see inequality (58),(59)):

$$\frac{8(\frac{m}{n})^3}{N(1+(\frac{m}{n})^2)} < \frac{(\frac{m}{n})}{2^{p-5}(1+(\frac{m}{n})^2)}$$

We divide all the components by $\frac{\frac{m}{n}}{(1+(\frac{m}{n})^2)}$, obtaining:

$$\frac{8(\frac{m}{n})^2}{N} < \frac{1}{2^{p-5}}$$

Thus, it suffices to take $N = 2^{p-2}\lceil(\frac{m}{n})^2\rceil$.

Proof for inequality (56): To prove the inequality, we consider the calculations in $\frac{x}{N}\sum_{i=0}^{N-1}\frac{1}{1+(\frac{i}{N})^2 x^2}$ and estimate the amount of error that we commit in the approximation $\frac{(\frac{m}{n})}{N}\sum_{i=0}^{N-1}\frac{1}{1+(\frac{i}{N})^2(\frac{m}{n})^2}$.

From Theorem 5.2 and inequality (53), we conclude that 2 units of precision is lost by approximating x^2 by $(\frac{m}{n})^2$ and hence we obtain:

$$|x^2 - (\frac{m}{n})^2| < \frac{1}{2^{p-2}}|(\frac{m}{n})^2|$$

$$|(\frac{i}{N})^2 x^2 - (\frac{i}{N})^2(\frac{m}{n})^2| < \frac{1}{2^{p-2}}|(\frac{i}{N})^2(\frac{m}{n})^2|$$

The approximation $(\frac{i}{N})^2(\frac{m}{n})^2$ of $(\frac{i}{N})^2 x^2$ is positive. Thus, we do not lose precision by adding $(\frac{i}{N})^2 x^2$ and 1 (see Theorem 5.4.i):

$$|(1 + (\frac{i}{N})^2 x^2) - (1 + (\frac{i}{N})^2(\frac{m}{n})^2)| < \frac{1}{2^{p-2}}|1 + (\frac{i}{N})^2(\frac{m}{n})^2|$$

We lose 1 unit of precision by approximating the inverse of $1 + (\frac{i}{N})^2 x^2$ (see Theorem 5.3):

$$|\frac{1}{1 + (\frac{i}{N})^2 x^2} - \frac{1}{1 + (\frac{i}{N})^2(\frac{m}{n})^2}| < \frac{1}{2^{p-3}}|\frac{1}{1 + (\frac{i}{N})^2(\frac{m}{n})^2}|$$

We lose 2 units of precision in the approximation of $\frac{x}{N} \cdot \frac{1}{1+(\frac{i}{N})^2 x^2}$ (see Theorem 5.2):

$$|\frac{x}{N} \cdot \frac{1}{1 + (\frac{i}{N})^2 x^2} - \frac{(\frac{m}{n})}{N} \cdot \frac{1}{1 + (\frac{i}{N})^2(\frac{m}{n})^2}| < \frac{1}{2^{p-5}}|\frac{(\frac{m}{n})}{N} \cdot \frac{1}{1 + (\frac{i}{N})^2(\frac{m}{n})^2}|$$

The approximations $\frac{\frac{m}{n}}{N} \cdot \frac{1}{1+(\frac{i}{N})^2(\frac{m}{n})^2}$ are positive for $0 \leq i \leq N-1$. Thus, we do not lose precision by calculating the summation $\frac{x}{N}\sum_{i=0}^{N-1}\frac{1}{1+(\frac{i}{N})^2 x^2}$

(see Theorem 5.4.i):

$$\left| \frac{x}{N} \sum_{i=0}^{N-1} \frac{1}{1 + (\frac{i}{N})^2 x^2} - \frac{(\frac{m}{n})}{N} \sum_{i=0}^{N-1} \frac{1}{1 + (\frac{i}{N})^2 (\frac{m}{n})^2} \right|$$

$$< \frac{1}{2^{p-5}} \left| \frac{(\frac{m}{n})}{N} \sum_{i=0}^{N-1} \frac{1}{1 + (\frac{i}{N})^2 (\frac{m}{n})^2} \right|$$

2. Suppose $\frac{m}{n} < 0$. We can write:

$$\frac{m}{n}\left(1 + \frac{1}{2^p}\right) < x < \frac{m}{n}\left(1 - \frac{1}{2^p}\right)$$

Observe that $x < 0$; we can use the following identity for $\arctan(x)$:

$$\arctan(x) = -\arctan(-x)$$

where $-x > 0$. We first approximate $\arctan(-x)$ using the first part of the proof. Afterwards, we apply Theorem 5.1 to approximate $-\arctan(-x)$. Unary negation does not influence the precision.

$\square$

Algorithm 8 Arctangent

Require: *expr* has the shape $\arctan(x)$
1: **procedure** Compute($expr, p$)
2:     $\frac{m}{n} \leftarrow$ Compute($x, p + 6$)
3:     $N \leftarrow 2^{p+4}\lceil (\frac{m}{n})^2 \rceil$
4:     $\frac{m'}{n'} \leftarrow \frac{(\frac{m}{n})}{N} \sum_{i=0}^{N-1} \frac{1}{1 + (\frac{i}{N})^2 (\frac{m}{n})^2}$
5:     **return** $\frac{m'}{n'}$

Algorithm 8 applies Theorem 6.4 to approximate $\arctan(x)$ with arbitrary precision. Note that Theorem 6.4 predicts the amount of precision that is lost by calculating $\arctan(x)$ independently of the argument x and hence Algorithm 8 calculates $\arctan(x)$ in one pass.

To confirm that $\arctan(x)$ is computable in one pass, we apply perturbation analysis and calculate the quantity $|\frac{xf'(x)}{f(x)}|$ for $f(x) = \arctan(x)$:

$$\left|\frac{xf'(x)}{f(x)}\right| = \left|\frac{x(\frac{1}{1+x^2})}{\arctan(x)}\right| = \left|\frac{x}{(1+x^2)\arctan(x)}\right|$$

Proposition 4 (see Appendix A) shows that $|\frac{x}{(1+x^2)\arctan(x)}| < 1$; iterative computations are not required for approximating $\arctan(x)$.

7 Approximating Transcendental Functions by Taylor Expansions

In this section we first briefly discuss the basics of approximating functions using Taylor expansions. Afterwards, we use Taylor expansions to approximate $\sin(x)$ and $\cos(x)$.

Suppose $f : D \to R$ is a function and $I = (a, b) \subseteq D$ such that:

- f has n continuous derivatives on I (denoted by $f^{(i)}(x)$ for $1 \leq i \leq n$);

- $f^{(n+1)}$ exists on I;

- $x_0 \in I$.

Taylor's theorem states that for every $x \in I$ there is a number c_x between x and x_0 such that $f(x) = P_n(x) + R_n(x)$ where:

$$P_n(x) = \sum_{i=0}^{n} \frac{f^{(i)}(x_0)}{i!}(x - x_0)^i \ , \ R_n(x) = \frac{f^{(n+1)}(c_x)}{(n+1)!}(x - x_0)^{n+1} \tag{60}$$

The formula $R_n(x)$ is called the Lagrange form of the remainder [20].

In the remaining of this section, we discuss approximations for $\sin(x)$ and $\cos(x)$. Our approximations are based on the following Taylor expansions around the point $x_0 = 0$:

$$\sin(x) = \sum_{i=0}^{\infty} \frac{(-1)^i}{(2i+1)!} x^{2i+1} \tag{61}$$

$$\cos(x) = \sum_{i=0}^{\infty} \frac{(-1)^i}{(2i)!} x^{2i} \tag{62}$$

For each function, we first describe an approximation that is applicable to the base interval $I = (-1, 1)$ (see Section 7.1). Afterwards, we extend our calculations to the complete domain of the functions using range reduction identities (see Section 7.2). Proof of correctness for the base and general cases are provided. Moreover, the COMPUTE($expr, p$) function is extended to approximate $\sin(x)$ and $\cos(x)$ with arbitrary precision. We use perturbation analysis to confirm the observations obtained from the approximations.

7.1 Approximating Functions in the Base Interval

7.1.1 Sine

In this section we use the Taylor expansion from equality (61) to approximate $\sin(x)$. We assume that the input argument x is represented by (m, n, p) and $|\frac{m}{n}| < 1$.

Theorem 7.1. *Let x be a real number represented by (m, n, p) such that $-1 < \frac{m}{n} < 1$. Then $\sin(x)$ can be represented by $(m', n', p - 2\lceil \log_2(2N-1) \rceil - 3)$ where $\frac{m'}{n'} = \sum_{i=0}^{N} \frac{(-1)^i}{(2i+1)!} (\frac{m}{n})^{2i+1}$ and $N \in \mathbb{N}$ is an odd number such that $(\frac{5}{6})(\frac{(2N+2)!(2N-1)^2}{2^{2N}}) > 2^p$.*

Proof. The real number x is given with precision p:

$$\frac{m}{n} - |\frac{m}{n}| \leq \frac{m}{n} - \frac{1}{2^p}|\frac{m}{n}| < x < \frac{m}{n} + \frac{1}{2^p}|\frac{m}{n}| \leq \frac{m}{n} + |\frac{m}{n}| \tag{63}$$

To prove the theorem, we should show that the following inequality holds:

$$|\sin(x) - \sum_{i=0}^{N} \frac{(-1)^i}{(2i+1)!}(\frac{m}{n})^{2i+1}| < \frac{1}{2^{p-2\lceil \log_2(2N-1) \rceil - 3}}|\sum_{i=0}^{N} \frac{(-1)^i}{(2i+1)!}(\frac{m}{n})^{2i+1}| \tag{64}$$

We rewrite the left hand side of inequality (64) as follows:

$$|\sin(x) - \sum_{i=0}^{N} \frac{(-1)^i}{(2i+1)!}(\frac{m}{n})^{2i+1}| =$$

$$|\sin(x) - \sum_{i=0}^{N} \frac{(-1)^i}{(2i+1)!}x^{2i+1} + \sum_{i=0}^{N} \frac{(-1)^i}{(2i+1)!}x^{2i+1} - \sum_{i=0}^{N} \frac{(-1)^i}{(2i+1)!}(\frac{m}{n})^{2i+1}| \leq$$

$$|\sin(x) - \sum_{i=0}^{N} \frac{(-1)^i}{(2i+1)!}x^{2i+1}| + |\sum_{i=0}^{N} \frac{(-1)^i}{(2i+1)!}x^{2i+1} - \sum_{i=0}^{N} \frac{(-1)^i}{(2i+1)!}(\frac{m}{n})^{2i+1}|$$

Thus, we prove inequality (64) by showing that the following inequalities hold:

$$|\sin(x) - \sum_{i=0}^{N} \frac{(-1)^i}{(2i+1)!}x^{2i+1}| < \frac{1}{2^{p-2\lceil \log_2(2N-1) \rceil - 2}}|\sum_{i=0}^{N} \frac{(-1)^i}{(2i+1)!}(\frac{m}{n})^{2i+1}| \tag{65}$$

$$|\sum_{i=0}^{N} \frac{(-1)^i}{(2i+1)!}x^{2i+1} - \sum_{i=0}^{N} \frac{(-1)^i}{(2i+1)!}(\frac{m}{n})^{2i+1}|$$

$$< \frac{1}{2^{p-2\lceil \log_2(2N-1) \rceil - 2}}|\sum_{i=0}^{N} \frac{(-1)^i}{(2i+1)!}(\frac{m}{n})^{2i+1}| \tag{66}$$

Proof for inequality (65)**:** Based on Taylor's theorem (see equality (60)) and inequality (63), we can rewrite the left hand side of inequality (65) as follows:

$$|\sin(x) - \sum_{i=0}^{N} \frac{(-1)^i}{(2i+1)!}x^{2i+1}| = |\frac{\sin^{(2N+2)}(c_x)x^{2N+2}}{(2N+2)!}| \leq \frac{|\frac{m}{n}|^{2N+2}2^{2N+2}}{(2N+2)!} \tag{67}$$

37

We choose $N = 2k + 1 \geq 1$. Applying Proposition 5 (see Appendix A) we can rewrite the right hand side of inequality (65) as follows:

$$\frac{1}{2^{p-2\lceil \log_2(2N-1)\rceil-2}}|\sum_{i=0}^{N}\frac{(-1)^i}{(2i+1)!}(\frac{m}{n})^{2i+1}| = \left(\frac{sgn(\frac{m}{n})}{2^{p-2\lceil \log_2(2N-1)\rceil-2}}\right)$$

$$\cdot\left(\sum_{i=0}^{N}\frac{(-1)^i}{(2i+1)!}(\frac{m}{n})^{2i+1}\right) \quad (68)$$

where sgn is the sign function. To prove inequality (65), it suffices to show that the following inequality holds (see (67),(68)):

$$\frac{|\frac{m}{n}|^{2N+2}2^{2N+2}}{(2N+2)!} < \frac{sgn(\frac{m}{n})}{2^{p-2\lceil \log_2(2N-1)\rceil-2}}\sum_{i=0}^{N}\frac{(-1)^i}{(2i+1)!}(\frac{m}{n})^{2i+1}$$

This is equivalent to:

$$\frac{sgn(\frac{m}{n})}{2^{p-2\lceil \log_2(2N-1)\rceil-2}}\sum_{i=0}^{N}\frac{(-1)^i}{(2i+1)!}(\frac{m}{n})^{2i+1} - \frac{|\frac{m}{n}|^{2N+2}2^{2N+2}}{(2N+2)!} =$$

$$\frac{sgn(\frac{m}{n})}{2^{p-2\lceil \log_2(2N-1)\rceil-2}}\left(\frac{m}{n} - (\frac{1}{3!})(\frac{m}{n})^3 + \sum_{i=2}^{N}\frac{(-1)^i}{(2i+1)!}(\frac{m}{n})^{2i+1}\right)$$

$$- \frac{|\frac{m}{n}|^{2N+2}2^{2N+2}}{(2N+2)!} > 0 \quad (69)$$

The quantity $sgn(\frac{m}{n})\sum_{i=2}^{N}\frac{(-1)^i}{(2i+1)!}(\frac{m}{n})^{2i+1}$ is positive (see Proposition 5 in Appendix A). Moreover, the approximation $\frac{m}{n}$ satisfies $|\frac{m}{n}| < 1$. To prove inequality (69) it is sufficient to show:

$$\frac{sgn(\frac{m}{n})}{2^{p-2\lceil \log_2(2N-1)\rceil-2}}(\frac{m}{n} - (\frac{1}{3!})(\frac{m}{n})^3) - \frac{|\frac{m}{n}|^{2N+2}2^{2N+2}}{(2N+2)!} >$$

$$\frac{sgn(\frac{m}{n})(2N-1)^2}{2^{p-2}}(\frac{m}{n} - (\frac{1}{3!})(\frac{m}{n})^3) - \frac{|\frac{m}{n}|^3 2^{2N+2}}{(2N+2)!} > 0 \quad (70)$$

Inequality (70) is equivalent to:

$$sgn(\frac{m}{n})(\frac{m}{n})(1 - (\frac{1}{3!})(\frac{m}{n})^2 - \frac{2^{p+2N}}{(2N+2)!(2N-1)^2}(\frac{m}{n})^2) > 0$$

Since $sgn(\frac{m}{n})(\frac{m}{n}) > 0$ and $(\frac{m}{n})^2 < 1$, we should choose an N such that:

$$\frac{1}{6} + \frac{2^{p+2N-2\lceil \log_2(2N-1)\rceil}}{(2N+2)!(2N-1)^2} < 1$$

Thus, N should satisfy $(\frac{5}{6})\frac{(2N+2)!(2N-1)^2}{2^{2N}} > 2^p$.

38

Proof for inequality (66): We show that given an approximation $\frac{m}{n}$ of x with precision p we can approximate $\sum_{i=0}^{N} \frac{(-1)^i}{(2i+1)!} x^{2i+1}$ by $\sum_{i=0}^{N} \frac{(-1)^i}{(2i+1)!} (\frac{m}{n})^{2i+1}$ with precision $p - 2\lceil \log_2(2N-1) \rceil - 2$.

The sign of the terms $\frac{(-1)^i}{(2i+1)!} x^{2i+1}$ alternates between positive and negative. Thus, adding two arbitrary terms with different signs from the expansion can significantly reduce the precision (see Theorem 5.4.ii). To avoid this, we first consider specific pairs of terms for which loss of precision due to addition is bounded. Then, we calculate the summation of these pairs. The following identity shows the way we calculate $\sum_{i=0}^{N} \frac{(-1)^i}{(2i+1)!} x^{2i+1}$:

$$\sum_{i=0}^{N} \frac{(-1)^i}{(2i+1)!} x^{2i+1} = \sum_{i=0}^{k} \frac{x^{4i+1}}{(4i+1)!}\left(1 - \frac{x^2}{(4i+2)(4i+3)}\right)$$

Choosing $N = 2k+1$ allows us to make pairs of terms.

The number x is given with precision p. Thus, we can approximate x^2 with precision $p - 2$ (see Theorem 5.2):

$$|x^2 - (\frac{m}{n})^2| < \frac{1}{2^{p-2}}|(\frac{m}{n})^2|$$

and hence

$$\left|\frac{x^2}{(4i+2)(4i+3)} - \frac{(\frac{m}{n})^2}{(4i+2)(4i+3)}\right| < \frac{1}{2^{p-2}}\left|\frac{(\frac{m}{n})^2}{(4i+2)(4i+3)}\right|$$

We approximate $1 - \frac{x^2}{(4i+2)(4i+3)}$ by $1 - \frac{(\frac{m}{n})^2}{(4i+2)(4i+3)}$. Loss of precision in the approximation can be estimated by calculating the quantity $\log_2\left(\frac{1+\frac{(\frac{m}{n})^2}{(4i+2)(4i+3)}}{1-\frac{(\frac{m}{n})^2}{(4i+2)(4i+3)}}\right)$ (see Theorem 5.4.ii):

$$\log_2\left(\frac{1+\frac{(\frac{m}{n})^2}{(4i+2)(4i+3)}}{1-\frac{(\frac{m}{n})^2}{(4i+2)(4i+3)}}\right) \le \log_2\left(\frac{1+\frac{1}{6}}{1-\frac{1}{6}}\right) = \log_2(\frac{7}{5}) < 1$$

Thus, we lose at most 1 unit of precision:

$$\left|(1 - \frac{x^2}{(4i+2)(4i+3)}) - (1 - \frac{(\frac{m}{n})^2}{(4i+2)(4i+3)})\right| < \frac{1}{2^{p-3}}\left|1 - \frac{(\frac{m}{n})^2}{(4i+2)(4i+3)}\right|$$

The powers x^{4i+1} are approximated by $(\frac{m}{n})^{4i+1}$ for $0 \le i \le k$ and $2\lceil \log_2(4i+1) \rceil$ units of precision is lost in this operation. Thus, in the worst case, the precision is reduced by $2\lceil \log_2(4k+1) \rceil$ units:

$$|x^{4i+1} - (\frac{m}{n})^{4i+1}| < \frac{1}{2^{p-2\lceil \log_2(4k+1) \rceil}}|(\frac{m}{n})^{4i+1}| = \frac{1}{2^{p-2\lceil \log_2(2N-1) \rceil}}|(\frac{m}{n})^{4i+1}|$$

Multiplying approximations of x^{4i+1} by a constant does not influence the precision:

$$\left|\frac{x^{4i+1}}{(4i+1)!} - \frac{(\frac{m}{n})^{4i+1}}{(4i+1)!}\right| < \frac{1}{2^{p-2\lceil\log_2(2N-1)\rceil}}\left|\frac{(\frac{m}{n})^{4i+1}}{(4i+1)!}\right|$$

The multiplication $\frac{x^{4i+1}}{(4i+1)!}(1 - \frac{x^2}{(4i+2)(4i+3)})$ can be approximated by multiplying the approximations $\frac{(\frac{m}{n})^{4i+1}}{(4i+1)!}$ and $1 - \frac{(\frac{m}{n})^2}{(4i+2)(4i+3)}$ (see Theorem 5.2):

$$\left|\frac{x^{4i+1}}{(4i+1)!}\left(1 - \frac{x^2}{(4i+2)(4i+3)}\right) - \frac{(\frac{m}{n})^{4i+1}}{(4i+1)!}\left(1 - \frac{(\frac{m}{n})^2}{(4i+2)(4i+3)}\right)\right|$$
$$< \frac{1}{2^{p-\max\{2\lceil\log_2(2N-1)\rceil,3\}-2}}\left|\frac{(\frac{m}{n})^{4i+1}}{(4i+1)!}\left(1 - \frac{(\frac{m}{n})^2}{(4i+2)(4i+3)}\right)\right|$$
$$= \frac{1}{2^{p-2\lceil\log_2(2N-1)\rceil-2}}\left|\frac{(\frac{m}{n})^{4i+1}}{(4i+1)!}\left(1 - \frac{(\frac{m}{n})^2}{(4i+2)(4i+3)}\right)\right|$$

For the last equality, we use the assumptions $(\frac{5}{6})(\frac{(2N+2)!(N-1)^2}{2^{2N-2\lceil\log_2(2N-1)\rceil}}) > 2^p$ and $N = 2k+1$; we can conclude that $N \geq 3$.

For $0 \leq i \leq k$ the terms $\frac{(\frac{m}{n})^{4i+1}}{(4i+1)!}(1 - \frac{(\frac{m}{n})^2}{(4i+2)(4i+3)})$ have the same sign which is determined by the sign of $\frac{m}{n}$ (see Proposition 5 in Appendix A). Thus, we can approximate $\sum_{i=0}^{k}\frac{x^{4i+1}}{(4i+1)!}(1 - \frac{x^2}{(4i+2)(4i+3)})$ by $\sum_{i=0}^{k}\frac{(\frac{m}{n})^{4i+1}}{(4i+1)!}(1 - \frac{(\frac{m}{n})^2}{(4i+2)(4i+3)})$ without losing precision (see Theorem 5.4.i):

$$\left|\sum_{i=0}^{k}\frac{x^{4i+1}}{(4i+1)!}\left(1 - \frac{x^2}{(4i+2)(4i+3)}\right) - \sum_{i=0}^{k}\frac{(\frac{m}{n})^{4i+1}}{(4i+1)!}\left(1 - \frac{(\frac{m}{n})^2}{(4i+2)(4i+3)}\right)\right| <$$
$$\frac{1}{2^{p-2\lceil\log_2(2N-1)\rceil-2}}\left|\sum_{i=0}^{k}\frac{(\frac{m}{n})^{4i+1}}{(4i+1)!}\left(1 - \frac{(\frac{m}{n})^2}{(4i+2)(4i+3)}\right)\right|$$

$\square$

Theorem 7.1 provides a top-down approximation for $\sin(x)$ in the base interval. Loss of precision in this approximation is estimated independently of the argument x. To show that iterative calculations can be avoided in the base interval, we calculate $|\frac{xf'(x)}{f(x)}|$ for $f(x) = \sin(x)$:

$$\left|\frac{xf'(x)}{f(x)}\right| = \left|\frac{x \cdot \cos(x)}{\sin(x)}\right| = |x \cdot \cot(x)| \tag{71}$$

Proposition 7 (see Appendix A) shows that $|x \cdot \cot(x)| < 1$ for $x \in (-1,1)$. Thus, we can approximate $\sin(x)$ in the base interval in one pass.

7.1.2 Cosine

In this section we introduce an approximation for $\cos(x)$ where x is represented by (m, n, p) and $|\frac{m}{n}| < 1$. Our approximation is based on the Taylor expansion from equality (62).

Theorem 7.2. *Let x be a real number represented by (m, n, p) such that $-1 < \frac{m}{n} < 1$. Then $\cos(x)$ can be represented by $(m', n', p - 2\lceil \log_2(2N-2) \rceil - 3)$ where $\frac{m'}{n'} = \sum_{i=0}^{N} \frac{(-1)^i}{(2i)!} (\frac{m}{n})^{2i}$ and $N \in \mathbb{N}$ is an odd number such that $\frac{(2N+1)!(2N-2)^2}{2^{2N}} > 2^p$.*

Proof. The real number x is represented by (m, n, p) and hence we can write:

$$\frac{m}{n} - |\frac{m}{n}| < \frac{m}{n} - \frac{1}{2^p}|\frac{m}{n}| < x < \frac{m}{n} + \frac{1}{2^p}|\frac{m}{n}| \leq \frac{m}{n} + |\frac{m}{n}| \tag{72}$$

To prove the theorem, we need to show:

$$|\cos(x) - \sum_{i=0}^{N} \frac{(-1)^i}{(2i)!} (\frac{m}{n})^{2i}| < \frac{1}{2^{p-2\lceil \log_2(2N-2) \rceil - 3}}|\sum_{i=0}^{N} \frac{(-1)^i}{(2i)!} (\frac{m}{n})^{2i}| \tag{73}$$

We rewrite the left hand side of inequality (73) as follows:

$$|\cos(x) - \sum_{i=0}^{N} \frac{(-1)^i}{(2i)!} (\frac{m}{n})^{2i}| =$$

$$|\cos(x) - \sum_{i=0}^{N} \frac{(-1)^i}{(2i)!} x^{2i} + \sum_{i=0}^{N} \frac{(-1)^i}{(2i)!} x^{2i} - \sum_{i=0}^{N} \frac{(-1)^i}{(2i)!} (\frac{m}{n})^{2i}| \leq$$

$$|\cos(x) - \sum_{i=0}^{N} \frac{(-1)^i}{(2i)!} x^{2i}| + |\sum_{i=0}^{N} \frac{(-1)^i}{(2i)!} x^{2i} - \sum_{i=0}^{N} \frac{(-1)^i}{(2i)!} (\frac{m}{n})^{2i}|$$

We prove inequality (73) by showing that the following inequalities are valid:

$$|\cos(x) - \sum_{i=0}^{N} \frac{(-1)^i}{(2i)!} x^{2i}| < \frac{1}{2^{p-2\lceil \log_2(2N-2) \rceil - 2}}|\sum_{i=0}^{N} \frac{(-1)^i}{(2i)!} (\frac{m}{n})^{2i}| \tag{74}$$

$$|\sum_{i=0}^{N} \frac{(-1)^i}{(2i)!} x^{2i} - \sum_{i=0}^{N} \frac{(-1)^i}{(2i)!} (\frac{m}{n})^{2i}| < \frac{1}{2^{p-2\lceil \log_2(2N-2) \rceil - 2}}|\sum_{i=0}^{N} \frac{(-1)^i}{(2i)!} (\frac{m}{n})^{2i}| \tag{75}$$

Proof for inequality (74): We use Taylor's theorem (see equality (60)) and the bounds calculated for x in inequality (72) to rewrite the left hand side of inequality (74):

$$|\cos(x) - \sum_{i=0}^{N} \frac{(-1)^i}{(2i)!} x^{2i}| = |\frac{\cos^{(2N+1)}(c_x) x^{2N+1}}{(2N+1)!}| \leq \frac{|\frac{m}{n}|^{2N+1} 2^{2N+1}}{(2N+1)!} \tag{76}$$

We choose $N = 2k+1 \geq 1$ and apply Proposition 6 (see Appendix A) to rewrite the right hand side of inequality (74):

$$\frac{1}{2^{p-2\lceil \log_2(2N-2)\rceil -2}}\left|\sum_{i=0}^{N}\frac{(-1)^i}{(2i)!}\left(\frac{m}{n}\right)^{2i}\right| = \frac{1}{2^{p-2\lceil \log_2(2N-2)\rceil -2}}\sum_{i=0}^{N}\frac{(-1)^i}{(2i)!}\left(\frac{m}{n}\right)^{2i} \tag{77}$$

To show that inequality (74) holds, it suffices to prove the following (see inequality (76),(77)):

$$\frac{|\frac{m}{n}|^{2N+1}2^{2N+1}}{(2N+1)!} < \frac{1}{2^{p-2\lceil \log_2(2N-2)\rceil -2}}\sum_{i=0}^{N}\frac{(-1)^i}{(2i)!}\left(\frac{m}{n}\right)^{2i} \tag{78}$$

Inequality (78) is equivalent to:

$$\frac{1}{2^{p-2\lceil \log_2(2N-2)\rceil -2}}\sum_{i=0}^{N}\frac{(-1)^i}{(2i)!}\left(\frac{m}{n}\right)^{2i} - \frac{|\frac{m}{n}|^{2N+1}2^{2N+1}}{(2N+1)!} =$$

$$\frac{1}{2^{p-2\lceil \log_2(2N-2)\rceil -2}}\left(1 - \frac{1}{2}\left(\frac{m}{n}\right)^2 + \sum_{i=2}^{N}\frac{(-1)^i}{(2i)!}\left(\frac{m}{n}\right)^{2i}\right) - \frac{|\frac{m}{n}|^{2N+1}2^{2N+1}}{(2N+1)!} > 0$$

The quantity $\sum_{i=2}^{N}\frac{(-1)^i}{(2i)!}\left(\frac{m}{n}\right)^{2i}$ is positive (see Proposition 6 in Appendix A). Since $|\frac{m}{n}| < 1$, it suffices to show:

$$\frac{1}{2^{p-2\lceil \log_2(2N-2)\rceil -2}}\left(1 - \frac{1}{2}\left(\frac{m}{n}\right)^2\right) - \frac{|\frac{m}{n}|^{2N+1}2^{2N+1}}{(2N+1)!} >$$

$$\frac{(2N-2)^2}{2^{p-2}}\left(1 - \frac{1}{2}\left(\frac{m}{n}\right)^2\right) - \frac{2^{2N+1}}{(2N+1)!}\left(\frac{m}{n}\right)^2 > 0 \tag{79}$$

Inequality (79) is equivalent to:

$$1 - \frac{1}{2}\left(\frac{m}{n}\right)^2 - \frac{2^{p+2N-1}}{(2N+1)!(2N-2)^2}\left(\frac{m}{n}\right)^2 > 0$$

Since $\left(\frac{m}{n}\right)^2 < 1$, it is sufficient to choose an N that satisfies the following inequality:

$$\frac{1}{2} + \frac{2^{p+2N-1}}{(2N+1)!(2N-2)^2} < 1$$

Thus, we should choose an $N = 2k+1$ that satisfies $\frac{(2N+1)!(2N-2)^2}{2^{2N}} > 2^p$.

Proof for inequality (75): To prove the inequality, we estimate the amount of precision that is lost when we approximate $\sum_{i=0}^{N}\frac{(-1)^i}{(2i)!}x^{2i}$ by $\sum_{i=0}^{N}\frac{(-1)^i}{(2i)!}\left(\frac{m}{n}\right)^{2i}$. The sign of the terms $\frac{(-1)^i}{(2i)!}x^{2i}$ alternates between positive and negative. Thus,

42

adding two arbitrary terms with different signs from the expansion can potentially cause significant loss of precision (see Theorem 5.4.ii). To avoid this, we first consider pairs of terms for which addition can be calculated with a bounded loss of precision. Afterwards, we calculate the summation of these pairs. The following identity shows our computation scheme:

$$\sum_{i=0}^{N} \frac{(-1)^i}{(2i)!} x^{2i} = \sum_{i=0}^{k} \frac{x^{4i}}{(4i)!} \left(1 - \frac{x^2}{(4i+1)(4i+2)}\right)$$

Choosing $N = 2k + 1$ allows us to pair the terms of the summation.

Since x is given with precision p, we can approximate x^2 with precision $p-2$ (see Theorem 5.2):

$$|x^2 - (\frac{m}{n})^2| < \frac{1}{2^{p-2}}|\frac{m}{n}|^2$$

Multiplying the approximation of x^2 by a constant does not influence the precision:

$$|\frac{x^2}{(4i+1)(4i+2)} - \frac{(\frac{m}{n})^2}{(4i+1)(4i+2)}| < \frac{1}{2^{p-2}}|\frac{(\frac{m}{n})^2}{(4i+1)(4i+2)}|$$

We approximate $1 - \frac{x^2}{(4i+1)(4i+2)}$ by $1 - \frac{(\frac{m}{n})^2}{(4i+1)(4i+2)}$. To estimate lose of precision in our approximation, we calculate the quantity $\log_2(\frac{1+\frac{(\frac{m}{n})^2}{(4i+1)(4i+2)}}{1-\frac{(\frac{m}{n})^2}{(4i+1)(4i+2)}})$ (see Theorem 5.4.ii):

$$\log_2(\frac{1 + \frac{(\frac{m}{n})^2}{(4i+1)(4i+2)}}{1 - \frac{(\frac{m}{n})^2}{(4i+1)(4i+2)}}) \leq \log_2(\frac{1 + \frac{1}{2}}{1 - \frac{1}{2}}) = \log_2 3 < 2$$

Thus, we lose at most 2 units of precision:

$$|(1 - \frac{x^2}{(4i+1)(4i+2)}) - (1 - \frac{(\frac{m}{n})^2}{(4i+1)(4i+2)})| < \frac{1}{2^{p-4}}|1 - \frac{(\frac{m}{n})^2}{(4i+1)(4i+2)}|$$

Approximating x^{4i} by $(\frac{m}{n})^{4i}$ reduces the precision by $2\lceil\log_2(4i)\rceil$ units (see Lemma 1 in Appendix A); in the worst case we lose $2\lceil\log_2(4k)\rceil$ units of precision:

$$|x^{4i} - (\frac{m}{n})^{4i}| < \frac{1}{2^{p-2\lceil\log_2(4k)\rceil}}|(\frac{m}{n})^{4i}| = \frac{1}{2^{p-2\lceil\log_2(2N-2)\rceil}}|(\frac{m}{n})^{4i}|$$

Multiplying $(\frac{m}{n})^{4i}$ by a constant factor does not influence the precision of the calculation:

$$|\frac{x^{4i}}{(4i)!} - \frac{(\frac{m}{n})^{4i}}{(4i)!}| < \frac{1}{2^{p-2\lceil\log_2(2N-2)\rceil}}|\frac{(\frac{m}{n})^{4i}}{(4i)!}|$$

We approximate $\frac{x^{4i}}{(4i)!}(1 - \frac{x^2}{(4i+1)(4i+2)})$ by multiplying the approximations $\frac{(\frac{m}{n})^{4i}}{(4i)!}$ and $(1 - \frac{(\frac{m}{n})^2}{(4i+1)(4i+2)})$ (see Theorem 5.2):

$$|\frac{x^{4i}}{(4i)!}(1 - \frac{x^2}{(4i+1)(4i+2)}) - \frac{(\frac{m}{n})^{4i}}{(4i)!}(1 - \frac{(\frac{m}{n})^2}{(4i+1)(4i+2)})|$$

$$< \frac{1}{2^{p - \max\{4, 2\lceil \log_2(2N-2)\rceil\} - 2}}|\frac{(\frac{m}{n})^{4i}}{(4i)!}(1 - \frac{(\frac{m}{n})^2}{(4i+1)(4i+2)})|$$

$$= \frac{1}{2^{p - 2\lceil \log_2(2N-2)\rceil - 2}}|\frac{(\frac{m}{n})^{4i}}{(4i)!}(1 - \frac{(\frac{m}{n})^2}{(4i+1)(4i+2)})|$$

To obtain the last equality, we use the assumptions $\frac{(2N+1)!(2N-2)^2}{2^{2N}} > 2^p$ and $N = 2k + 1$; we conclude that $N \geq 3$.

The terms $\frac{(\frac{m}{n})^{4i}}{(4i)!}(1 - \frac{(\frac{m}{n})^2}{(4i+1)(4i+2)})$ are positive for $0 \leq i \leq k$ (see Proposition 6 in Appendix A) and hence we can approximate the summation $\sum_{i=0}^{k} \frac{x^{4i}}{(4i)!}(1 - \frac{x^2}{(4i+1)(4i+2)})$ without losing precision (see Theorem 5.4.i):

$$|\sum_{i=0}^{k} \frac{x^{4i}}{(4i)!}(1 - \frac{x^2}{(4i+1)(4i+2)}) - \sum_{i=0}^{k} \frac{(\frac{m}{n})^{4i}}{(4i)!}(1 - \frac{(\frac{m}{n})^2}{(4i+1)(4i+2)})| <$$

$$\frac{1}{2^{p - 2\lceil \log_2(2N-2)\rceil - 2}}|\sum_{i=0}^{k} \frac{(\frac{m}{n})^{4i}}{(4i)!}(1 - \frac{(\frac{m}{n})^2}{(4i+1)(4i+2)})|$$

$\square$

Theorem 7.2 estimates loss of precision in $\cos(x)$ in the base interval independently of the argument x. To show that iterative computations can be avoided in the base interval, we calculate $|\frac{xf'(x)}{f(x)}|$ for $f(x) = \cos(x)$:

$$|\frac{xf'(x)}{f(x)}| = |\frac{-x \cdot \sin(x)}{\cos(x)}| = |x \cdot \tan(x)| \tag{80}$$

From Proposition 8 (see Appendix A) we conclude that $|x \cdot \tan(x)| < \tan(1)$ for $x \in (-1, 1)$. Thus, we can approximate $\cos(x)$ in the base interval in one pass.

7.2 Extending Base Interval Approximations

In Section 7.1.1 and 7.1.2, we discussed approximations for $\sin(x)$ and $\cos(x)$ in the base interval. In what follows, we show that range reduction identities can be used to extend these approximations to calculate sine and cosine for an argument x represented by (m, n, p) where $|\frac{m}{n}| \geq 1$.

In our calculations, we use a representation (m', n', p) of π. This representation can be obtained based on our approximation for $\arctan(x)$ (see Section 6.3) and the following identity:

$$\pi = 4\arctan(1)$$

7.2.1 Sine

Theorem 7.3. *Let x be a real number represented by (m, n, p) such that $|\frac{m}{n}| \geq 1$ and (m', n', p) be a representation for π. Suppose $\frac{\overline{m}}{\overline{n}} = \frac{m}{n} + k\frac{m'}{n'}$ where $k \in \mathbb{Z}$ and $0 < \frac{\overline{m}}{\overline{n}} < \frac{m'}{n'}$. The value of $\sin(x)$ can be approximated as follows:*

i. If $0 < \frac{\overline{m}}{\overline{n}} < 1$, then $\sin(x)$ can be represented by $(m_1, n_1, p - i_1)$ where:

$$i_1 = s_1 + t_1$$

$$\frac{m_1}{n_1} = \sum_{i=0}^{N_1} \frac{(-1)^i}{(2i+1)!} \left(\frac{\overline{m}}{\overline{n}}\right)^{2i+1}$$

ii. If $1 \leq \frac{\overline{m}}{\overline{n}} < 2$, then $\sin(x)$ can be represented by $(m_2, n_2, p - i_2)$ where:

$$i_2 = s_1 + t_2 + 2$$

$$\frac{m_2}{n_2} = 2\left(\sum_{i=0}^{N_1} \frac{(-1)^i}{(2i+1)!} \left(\frac{\overline{m}}{2\overline{n}}\right)^{2i+1}\right) \cdot \left(\sum_{i=0}^{N_2} \frac{(-1)^i}{(2i)!} \left(\frac{\overline{m}}{2\overline{n}}\right)^{2i}\right)$$

iii. If $2 \leq \frac{\overline{m}}{\overline{n}} < \frac{m'}{n'}$, then $\sin(x)$ can be represented by $(m_3, n_3, p - i_3)$ where:

$$i_3 = s_1 + s_2 + t_3 + 6$$

$$\frac{m_3}{n_3} = 8\left(\sum_{i=0}^{N_1} \frac{(-1)^i}{(2i+1)!} \left(\frac{\overline{m}}{4\overline{n}}\right)^{2i+1}\right) \cdot \left(\sum_{i=0}^{N_2} \frac{(-1)^i}{(2i)!} \left(\frac{\overline{m}}{4\overline{n}}\right)^{2i}\right)$$

$$\cdot \left(\sum_{i=0}^{N_3} \frac{(-1)^i}{(2i+1)!} \left(\frac{m'}{4n'} - \frac{\overline{m}}{4\overline{n}}\right)^{2i+1}\right) \cdot \left(\sum_{i=0}^{N_4} \frac{(-1)^i}{(2i)!} \left(\frac{m'}{4n'} - \frac{\overline{m}}{4\overline{n}}\right)^{2i}\right)$$

In the approximations above, $s_1, s_2, N_1, N_2, N_3, N_4 \in \mathbb{N}$ are the smallest natural numbers satisfying:

$$s_1 \geq \log_2\left(\frac{1 + \frac{\min(|\frac{m}{n}|, k\frac{m'}{n'})}{\max(|\frac{m}{n}|, k\frac{m'}{n'})}}{1 - \frac{\min(|\frac{m}{n}|, k\frac{m'}{n'})}{\max(|\frac{m}{n}|, k\frac{m'}{n'})}}\right), \quad s_2 \geq \log_2\left(\frac{1 + \frac{\min(\frac{m'}{4n'}, \frac{\overline{m}}{4\overline{n}})}{\max(\frac{m'}{4n'}, \frac{\overline{m}}{4\overline{n}})}}{1 - \frac{\min(\frac{m'}{4n'}, \frac{\overline{m}}{4\overline{n}})}{\max(\frac{m'}{4n'}, \frac{\overline{m}}{4\overline{n}})}}\right)$$

$$\left(\frac{5}{6}\right)\left(\frac{(2N_1 + 2)!(2N_1 - 1)^2}{2^{2N_1}}\right) > 2^{p - s_1}, \quad \frac{(2N_2 + 1)!(2N_2 - 2)^2}{2^{2N_2}} > 2^{p - s_1}$$

$$\left(\frac{5}{6}\right)\left(\frac{(2N_3 + 2)!(2N_3 - 1)^2}{2^{2N_3}}\right) > 2^{p - s_1 - s_2}, \quad \frac{(2N_4 + 1)!(2N_4 - 2)^2}{2^{2N_4}} > 2^{p - s_1 - s_2}$$

and $t_1, t_2 \in \mathbb{N}$ are defined as follows:

$$t_1 = 2\lceil \log_2(2N_1 - 1)\rceil + 3$$
$$t_2 = \max(2\lceil \log_2(2N_1 - 1)\rceil + 3, 2\lceil \log_2(2N_2 - 2)\rceil + 3)$$
$$t_3 = \max(2\lceil \log_2(2N_1 - 1)\rceil + 3, 2\lceil \log_2(2N_2 - 2)\rceil + 3,$$
$$2\lceil \log_2(2N_3 - 1)\rceil + 3, 2\lceil \log_2(2N_4 - 2)\rceil + 3)$$

Proof. Since $|\frac{m}{n}| \geq 1$, we can choose $k \in \mathbb{Z}$ such that $\frac{\overline{m}}{\overline{n}} = \frac{m}{n} + k\frac{m'}{n'}$ and $0 < \frac{\overline{m}}{\overline{n}} < \frac{m'}{n'}$. Suppose $y = x + k\pi$. We use the following identity to calculate $\sin(x)$:

$$\sin(x) = \sin(x + k\pi) = \sin(y) \tag{81}$$

We approximate y by $\frac{\overline{m}}{\overline{n}} = \frac{m}{n} + k\frac{m'}{n'}$ (see Theorem 5.4.ii). By performing this approximation, we lose $s_1 \in \mathbb{N}$ units of precision where s_1 is the smallest number satisfying:

$$s_1 \geq \log_2\left(\frac{1 + \frac{\min(|\frac{m}{n}|, k\frac{m'}{n'})}{\max(|\frac{m}{n}|, k\frac{m'}{n'})}}{1 - \frac{\min(|\frac{m}{n}|, k\frac{m'}{n'})}{\max(|\frac{m}{n}|, k\frac{m'}{n'})}}\right)$$

We consider three cases for calculating $\sin(y)$:

1. Suppose $0 < \frac{\overline{m}}{\overline{n}} < 1$. We apply Theorem 7.1. Thus, $t_1 = 2\lceil\log_2(2N_1 - 1)\rceil + 3$ units of precision is lost in the approximation of $\sin(y)$. A total of $s_1 + t_1$ units of precision is lost in the approximation of $\sin(x)$.

2. Suppose $1 \leq \frac{\overline{m}}{\overline{n}} < 2$. We use the following identity to bring the argument within the base interval:

$$\sin(y) = 2\sin(\frac{y}{2})\cos(\frac{y}{2}) \tag{82}$$

   We approximate $\frac{y}{2}$ by $\frac{\overline{m}}{2\overline{n}}$. Since $\frac{1}{2} \leq \frac{\overline{m}}{2\overline{n}} < 1$, we apply Theorem 7.1 and 7.2 to approximate $\sin(\frac{y}{2})$ and $\cos(\frac{y}{2})$, respectively. Loss of precision in these calculations is as follows:

$$t_2 = \max(2\lceil\log_2(2N_1 - 1)\rceil + 3, 2\lceil\log_2(2N_2 - 2)\rceil + 3)$$

   The multiplication in equality (82) reduces the precision by 2 units (see Theorem 5.2). A total of $s_1 + t_2 + 2$ units of precision is lost in the approximation of $\sin(x)$.

3. Suppose $2 \leq \frac{\overline{m}}{\overline{n}} < \frac{m'}{n'}$. We use the following identity to bring the argument within the base interval.

$$\begin{aligned}
\sin(y) &= 2\sin(\frac{y}{2})\cos(\frac{y}{2}) \\
&= 4\sin(\frac{y}{4})\cos(\frac{y}{4})\cos(\frac{y}{2}) \\
&= 4\sin(\frac{y}{4})\cos(\frac{y}{4})\sin(\frac{\pi}{2} - \frac{y}{2}) \\
&= 8\sin(\frac{y}{4})\cos(\frac{y}{4})\sin(\frac{\pi}{4} - \frac{y}{4})\cos(\frac{\pi}{4} - \frac{y}{4}) \tag{83}
\end{aligned}$$

We approximate $\frac{y}{4}$ and $\frac{\pi}{4} - \frac{y}{4}$ by $\frac{\overline{m}}{4\overline{n}}$ and $\frac{m'}{4n'} - \frac{\overline{m}}{4\overline{n}}$, respectively. We lose $s_2 \in \mathbb{N}$ units of precision in this approximation (see Theorem 5.4.ii) where s_2 is the smallest number satisfying:

$$s_2 \geq \log_2\left(\frac{1 + \frac{\min(\frac{m'}{4n'}, \frac{\overline{m}}{4\overline{n}})}{\max(\frac{m'}{4n'}, \frac{\overline{m}}{4\overline{n}})}}{1 - \frac{\min(\frac{m'}{4n'}, \frac{\overline{m}}{4\overline{n}})}{\max(\frac{m'}{4n'}, \frac{\overline{m}}{4\overline{n}})}}\right)$$

We apply Theorem 7.1 and 7.2 to approximate $\sin(\frac{y}{4}), \cos(\frac{y}{4}), \sin(\frac{\pi}{4} - \frac{y}{4})$, and $\cos(\frac{\pi}{4} - \frac{y}{4})$. Lose of precision in these approximations can be calculated as follows:

$$t_3 = \max(2\lceil\log_2(2N_1 - 1)\rceil + 3, 2\lceil\log_2(2N_2 - 2)\rceil + 3,$$
$$2\lceil\log_2(2N_3 - 1)\rceil + 3, 2\lceil\log_2(2N_4 - 2)\rceil + 3)$$

The three multiplications in $\sin(\frac{y}{4})\cos(\frac{y}{4})\sin(\frac{\pi}{4} - \frac{y}{4})\cos(\frac{\pi}{4} - \frac{y}{4})$ reduce the precision by 6 units (see Theorem 5.2). A total of $s_1 + s_2 + t_3 + 6$ units of precision is lost in the approximation of $\sin(x)$.

$\square$

Algorithm 9 applies Theorem 7.1 and 7.3 to approximate $\sin(x)$ with arbitrary precision. In this algorithm, initially, we calculate x with a precision that is adequate for calculations in the base interval. However, if the obtained approximation is outside the base interval, we use the half-angle formula or add rational multiples of π to the argument (see equality (81) and (83)).

Observe that an arbitrary amount of precision can be lost in the approximation of $x + k\pi$ and $\frac{\pi}{4} - \frac{y}{4}$ when $x \approx -k\pi$. Hence, iterative computations might be necessary in the approximation (see Line 4,22,29,38 in Algorithm 9). To show that this is essential for sine, we reconsider the perturbation analysis in equality (71) for $f(x) = \sin(x)$. The quantity $|x \cdot \cot(x)|$ can be arbitrary large for $x \approx -k\pi$. Thus, iterative computations are essential for approximating $\sin(x)$.

7.2.2 Cosine

Theorem 7.4. *Let x be a real number represented by (m, n, p) such that $|\frac{m}{n}| \geq 1$ and (m', n', p) be a representation for π. Suppose $\frac{\overline{m}}{\overline{n}} = \frac{m}{n} + (2k+1)\frac{m'}{2n'}$ where $k \in \mathbb{Z}$ and $0 < \frac{\overline{m}}{\overline{n}} < \frac{m'}{n'}$. The value of $\cos(x)$ can be approximated as follows:*

i. If $0 < \frac{\overline{m}}{\overline{n}} < 1$, then $\cos(x)$ can be represented by $(m_1, n_1, p - i_1)$ where:

$$i_1 = s_1 + t_1$$
$$\frac{m_1}{n_1} = (-1)^k \sum_{i=0}^{N_1} \frac{(-1)^i}{(2i+1)!}\left(\frac{\overline{m}}{\overline{n}}\right)^{2i+1}$$

47

Algorithm 9 Sine

Require: *expr* has the shape $\sin x$

1: **procedure** Compute($expr, p$)

2:     Choose an odd N such that $\left(\frac{5}{6}\right)\left(\frac{(2N+2)!(N-1)^2}{2^{2N}}\right) > 2^{p+2\lceil \log_2(2N-1)\rceil+3}$

3:     $p_x \leftarrow p + 2\lceil \log_2(2N-1)\rceil + 3$

4:     **repeat**

5:         $\frac{m}{n} \leftarrow$ Compute(x, p_x)

6:         **if** $-1 < \frac{m}{n} < 1$ **then**   ▷ Theorem 7.1

7:             $\frac{m_0}{n_0} \leftarrow \sum_{i=0}^{N} \frac{(-1)^i}{(2i+1)!}\left(\frac{m}{n}\right)^{2i+1}$

8:             **return** $\frac{m_0}{n_0}$

9:         **else**

10:             $\frac{m'}{n'} \leftarrow$ Compute($4\arctan(1), p_x$)

11:             Choose $k \in \mathbb{Z}$ such that $0 < \frac{m}{n} + k\frac{m'}{n'} < \frac{m'}{n'}$

12:             $\frac{\overline{m}}{\overline{n}} \leftarrow \frac{m}{n} + k\frac{m'}{n'}$

13:             Choose $s_1 \in \mathbb{N}$ such that $s_1 \geq \log_2\left(\frac{1+\frac{\min(|\frac{m}{n}|,k\frac{m'}{n'})}{\max(|\frac{m}{n}|,k\frac{m'}{n'})}}{1-\frac{\min(|\frac{m}{n}|,k\frac{m'}{n'})}{\max(|\frac{m}{n}|,k\frac{m'}{n'})}}\right)$

14:             Choose an odd N_1 such that $\left(\frac{5}{6}\right)\left(\frac{(2N_1+2)!(2N_1-1)^2}{2^{2N_1}}\right) > 2^{p_x-s_1}$

15:             Choose an odd N_2 such that $\frac{(2N_2+1)!(2N_2-2)^2}{2^{2N_2}} > 2^{p_x-s_1}$

16:             **if** $0 < \frac{\overline{m}}{\overline{n}} < 1$ **then**   ▷ Theorem 7.3.i

17:                 $t_1 \leftarrow 2\lceil \log_2(2N_1-1)\rceil + 3$

18:                 **if** $p_x - s_1 - t_1 \geq p$ **then**

19:                     $\frac{m_1}{n_1} = \sum_{i=0}^{N_1} \frac{(-1)^i}{(2i+1)!}\left(\frac{\overline{m}}{\overline{n}}\right)^{2i+1}$

20:                     **return** $\frac{m_1}{n_1}$

21:                 **else**

22:                     $p_x \leftarrow p_x + 1$

23:             **else if** $1 \leq \frac{\overline{m}}{\overline{n}} < 2$ **then**   ▷ Theorem 7.3.ii

24:                 $t_2 \leftarrow \max(2\lceil \log_2(2N_1-1)\rceil + 3, 2\lceil \log_2(2N_2-2)\rceil + 3)$

25:                 **if** $p_x - s_1 - t_2 - 2 \geq p$ **then**

26:                     $\frac{m_2}{n_2} = 2\left(\sum_{i=0}^{N_1} \frac{(-1)^i}{(2i+1)!}\left(\frac{\overline{m}}{2\overline{n}}\right)^{2i+1}\right)\left(\sum_{i=0}^{N_2} \frac{(-1)^i}{(2i)!}\left(\frac{\overline{m}}{2\overline{n}}\right)^{2i}\right)$

27:                     **return** $\frac{m_2}{n_2}$

28:                 **else**

29:                     $p_x \leftarrow p_x + 1$

30:             **else**   ▷ Theorem 7.3.iii

31:                 Choose an odd N_3 such that $\left(\frac{5}{6}\right)\left(\frac{(2N_3+2)!(2N_3-1)^2}{2^{2N_3}}\right) > 2^{p_x-s_1-s_2}$

32:                 Choose an odd N_4 such that $\frac{(2N_4+1)!(2N_4-2)^2}{2^{2N_4}} > 2^{p_x-s_1-s_2}$

33:                 $t_3 \leftarrow \max(2\lceil \log_2(2N_1-1)\rceil + 3, 2\lceil \log_2(2N_2-2)\rceil + 3,$
                         $2\lceil \log_2(2N_3-1)\rceil + 3, 2\lceil \log_2(2N_4-2)\rceil + 3)$

Algorithm 9 Sine (Continued)

34:	**if** $p_x - s_1 - s_2 - t_3 - 6 \geq p$ **then**
35:	$\frac{m_3}{n_3} = 8\big(\sum_{i=0}^{N_1} \frac{(-1)^i}{(2i+1)!}(\frac{\overline{m}}{4\overline{n}})^{2i+1}\big)\big(\sum_{i=0}^{N_2} \frac{(-1)^i}{(2i)!}(\frac{\overline{m}}{4\overline{n}})^{2i}\big)$
	$\big(\sum_{i=0}^{N_3} \frac{(-1)^i}{(2i+1)!}(\frac{m'}{4n'} - \frac{\overline{m}}{4\overline{n}})^{2i+1}\big)\big(\sum_{i=0}^{N_4} \frac{(-1)^i}{(2i)!}(\frac{m'}{4n'} - \frac{\overline{m}}{4\overline{n}})^{2i}\big)$
36:	**return** $\frac{m_3}{n_3}$
37:	**else**
38:	$p_x \leftarrow p_x + 1$
39:	**until** *true*

ii. If $1 \leq \frac{\overline{m}}{\overline{n}} < 2$, then $\cos(x)$ can be represented by $(m_2, n_2, p - i_2)$ where:

$$i_2 = s_1 + t_2 + 2$$

$$\frac{m_2}{n_2} = 2 \cdot (-1)^k \left(\sum_{i=0}^{N_1} \frac{(-1)^i}{(2i+1)!} (\frac{\overline{m}}{2\overline{n}})^{2i+1} \right) \cdot \left(\sum_{i=0}^{N_2} \frac{(-1)^i}{(2i)!} (\frac{\overline{m}}{2\overline{n}})^{2i} \right)$$

iii. If $2 \leq \frac{\overline{m}}{\overline{n}} < \frac{m'}{n'}$, then $\cos(x)$ can be represented by $(m_3, n_3, p - i_3)$ where:

$$i_3 = s_1 + s_2 + t_3 + 6$$

$$\frac{m_3}{n_3} = 8 \cdot (-1)^k \left(\sum_{i=0}^{N_1} \frac{(-1)^i}{(2i+1)!} (\frac{\overline{m}}{4\overline{n}})^{2i+1} \right) \cdot \left(\sum_{i=0}^{N_2} \frac{(-1)^i}{(2i)!} (\frac{\overline{m}}{4\overline{n}})^{2i} \right)$$

$$\cdot \left(\sum_{i=0}^{N_3} \frac{(-1)^i}{(2i+1)!} (\frac{m'}{4n'} - \frac{\overline{m}}{4\overline{n}})^{2i+1} \right) \cdot \left(\sum_{i=0}^{N_4} \frac{(-1)^i}{(2i)!} (\frac{m'}{4n'} - \frac{\overline{m}}{4\overline{n}})^{2i} \right)$$

In the approximations above, $s_1, s_2, N_1, N_2, N_3, N_4 \in \mathbb{N}$ are the smallest natural numbers satisfying:

$$s_1 \geq \log_2\left(\frac{1 + \frac{\min(|\frac{m}{n}|, (2k+1)\frac{m'}{2n'})}{\max(|\frac{m}{n}|, (2k+1)\frac{m'}{2n'})}}{1 - \frac{\min(|\frac{m}{n}|, (2k+1)\frac{m'}{2n'})}{\max(|\frac{m}{n}|, (2k+1)\frac{m'}{2n'})}}\right) \ , \ s_2 \geq \log_2\left(\frac{1 + \frac{\min(\frac{m'}{4n'}, \frac{\overline{m}}{4\overline{n}})}{\max(\frac{m'}{4n'}, \frac{\overline{m}}{4\overline{n}})}}{1 - \frac{\min(\frac{m'}{4n'}, \frac{\overline{m}}{4\overline{n}})}{\max(\frac{m'}{4n'}, \frac{\overline{m}}{4\overline{n}})}}\right)$$

$$(\frac{5}{6})(\frac{(2N_1 + 2)!(2N_1 - 1)^2}{2^{2N_1}}) > 2^{p-s_1} \ , \ \frac{(2N_2 + 1)!(2N_2 - 2)^2}{2^{2N_2}} > 2^{p-s_1}$$

$$(\frac{5}{6})(\frac{(2N_3 + 2)!(2N_3 - 1)^2}{2^{2N_3}}) > 2^{p-s_1-s_2} \ , \ \frac{(2N_4 + 1)!(2N_4 - 2)^2}{2^{2N_4}} > 2^{p-s_1-s_2}$$

and $t_1, t_2 \in \mathbb{N}$ are defined as follows:

$$t_1 = 2\lceil \log_2(2N_1 - 1) \rceil + 3$$
$$t_2 = \max(2\lceil \log_2(2N_1 - 1) \rceil + 3, 2\lceil \log_2(2N_2 - 2) \rceil + 3)$$
$$t_3 = \max(2\lceil \log_2(2N_1 - 1) \rceil + 3, 2\lceil \log_2(2N_2 - 2) \rceil + 3,$$
$$2\lceil \log_2(2N_3 - 1) \rceil + 3, 2\lceil \log_2(2N_4 - 2) \rceil + 3)$$

Proof. Since $|\frac{m}{n}| \geq 1$, we can choose $k \in \mathbb{Z}$ such that $\frac{\overline{m}}{\overline{n}} = \frac{m}{n} + (2k+1)\frac{m'}{2n'}$ and $0 < \frac{\overline{m}}{\overline{n}} < \frac{m'}{n'}$. Suppose $y = x + (2k+1)\frac{\pi}{2}$. We use the following identity to calculate $\cos(x)$:

$$\cos(x) = (-1)^k \sin(x + (2k+1)\frac{\pi}{2}) = (-1)^k \sin(y)$$

We approximate y by $\frac{\overline{m}}{\overline{n}} = \frac{m}{n} + (2k+1)\frac{m'}{2n'}$. From Theorem 5.4.ii, we can conclude that $s_1 \in \mathbb{N}$ units of precision will be lost in this approximation where s_1 is the smallest natural number satisfying:

$$s_1 \geq \log_2\left(\frac{1 + \frac{\min(|\frac{m}{n}|,(2k+1)\frac{m'}{2n'})}{\max(|\frac{m}{n}|,(2k+1)\frac{m'}{2n'})}}{1 - \frac{\min(|\frac{m}{n}|,(2k+1)\frac{m'}{2n'})}{\max(|\frac{m}{n}|,(2k+1)\frac{m'}{2n'})}}\right)$$

We apply Theorem 7.3 to approximate $\sin(y)$ and determine the loss of precision in the approximation. From Theorem 5.1 we can conclude that the factor $(-1)^k$ does not influence the precision of the approximation. $\square$

Algorithm 10 applies Theorem 7.2 and 7.4 to approximate $\cos(x)$ with a desired precision. The algorithm first calculates x with a precision that is sufficient for approximating $\cos(x)$ in the base interval. Similar to $\sin(x)$, we use range reduction identities to calculate $\cos(x)$ for an arbitrary x.

Observe that a significant amount of precision might be lost in the approximation when $x \approx \frac{-(2k+1)}{2}\pi$. Thus, iterative computations might be necessary in our approximation (see Line 4,22,29,38 in Algorithm 10). To show that these recomputations are essential, we reconsider the perturbation analysis of equality (80). The quantity $|x \cdot \tan(x)|$ can be arbitrary large for $x \approx \frac{-(2k+1)}{2}\pi$ and hence iterative computations are unavoidable for $\cos(x)$.

8 Related Work

An implementation for a bottom-up approach to exact real arithmetic is proposed in [16]. For a given expression, the inputs are computed with predefined precisions and a bottom-up scheme is used to determine the guaranteed precision of the output. Iterative computations are required if the obtained precision is not adequate. A formalization of a top-down approach in a theorem prover is proposed in [17]. The author first provides a definition for a metric space based on a ball relation. Afterwards, real numbers are defined as the completion of the metric space $\mathbb{Q}$. Rational operations are lifted to approximate operations on real numbers. This approach is optimized in [10].

Two closely-related top-down approaches based on absolute errors have been studied in [3, 15]. These approaches mainly differ in their approximations of the transcendental functions. In [3] the authors introduce a general way for calculating with Taylor expansions and apply this method to approximate the

Algorithm 10 Cosine

Require: *expr* has the shape $\cos x$

1: **procedure** Compute(*expr*, *p*)

2:     Choose an odd N such that $\frac{(2N+1)!(2N-2)^2}{2^{2N}} > 2^{p+2\lceil \log_2(2N-2)\rceil+2}$

3:     $p_x \leftarrow p + 2\lceil \log_2(2N-2)\rceil + 2$

4:     **repeat**

5:         $\frac{m}{n} \leftarrow$ Compute(x, p_x)

6:         **if** $-1 < \frac{m}{n} < 1$ **then**   ▷ Theorem 7.2

7:             $\frac{m_0}{n_0} = \sum_{i=0}^{N} \frac{(-1)^i}{(2i)!}\left(\frac{m}{n}\right)^{2i}$

8:             **return** $\frac{m_0}{n_0}$

9:         **else**

10:             $\frac{m'}{n'} \leftarrow$ Compute($4\arctan(1), p_x$)

11:             Choose $k \in \mathbb{Z}$ such that $0 < \frac{m}{n} + (2k+1)\frac{m'}{2n'} < \frac{m'}{n'}$

12:             $\frac{\overline{m}}{\overline{n}} \leftarrow \frac{m}{n} + (2k+1)\frac{m'}{2n'}$

13:             Choose $s_1 \in \mathbb{N}$ such that $s_1 \geq \log_2\left(\dfrac{1 + \frac{\min(|\frac{m}{n}|,(2k+1)\frac{m'}{2n'})}{\max(|\frac{m}{n}|,(2k+1)\frac{m'}{2n'})}}{1 - \frac{\min(|\frac{m}{n}|,(2k+1)\frac{m'}{2n'})}{\max(|\frac{m}{n}|,(2k+1)\frac{m'}{2n'})}}\right)$

14:             Choose an odd N_1 such that $\left(\frac{5}{6}\right)\left(\frac{(2N_1+2)!(2N_1-1)^2}{2^{2N_1}}\right) > 2^{p_x - s_1}$

15:             Choose an odd N_2 such that $\frac{(2N_2+1)!(2N_2-2)^2}{2^{2N_2}} > 2^{p_x - s_1}$

16:             **if** $0 < \frac{\overline{m}}{\overline{n}} < 1$ **then**   ▷ Theorem 7.4.i

17:                 $t_1 \leftarrow 2\lceil \log_2(2N_1 - 1)\rceil + 3$

18:                 **if** $p_x - s_1 - t_1 \geq p$ **then**

19:                     $\frac{m_1}{n_1} = (-1)^k \sum_{i=0}^{N_1} \frac{(-1)^i}{(2i+1)!}\left(\frac{\overline{m}}{\overline{n}}\right)^{2i+1}$

20:                     **return** $\frac{m_1}{n_1}$

21:                 **else**

22:                     $p_x \leftarrow p_x + 1$

23:             **else if** $1 \leq \frac{\overline{m}}{\overline{n}} < 2$ **then**   ▷ Theorem 7.4.ii

24:                 $t_2 \leftarrow \max(2\lceil \log_2(2N_1 - 1)\rceil + 3, 2\lceil \log_2(2N_2 - 2)\rceil + 3)$

25:                 **if** $p_x - s_1 - t_2 - 2 \geq p$ **then**

26:                     $\frac{m_2}{n_2} = 2 \cdot (-1)^k \left(\sum_{i=0}^{N_1} \frac{(-1)^i}{(2i+1)!}\left(\frac{\overline{m}}{2\overline{n}}\right)^{2i+1}\right)\left(\sum_{i=0}^{N_2} \frac{(-1)^i}{(2i)!}\left(\frac{\overline{m}}{2\overline{n}}\right)^{2i}\right)$

27:                     **return** $\frac{m_2}{n_2}$

28:                 **else**

29:                     $p_x \leftarrow p_x + 1$

30:             **else**   ▷ Theorem 7.4.iii

31:                 Choose an odd N_3 such that $\left(\frac{5}{6}\right)\left(\frac{(2N_3+2)!(2N_3-1)^2}{2^{2N_3}}\right) > 2^{p - s_1 - s_2}$

32:                 Choose an odd N_4 such that $\frac{(2N_4+1)!(2N_4-2)^2}{2^{2N_4}} > 2^{p - s_1 - s_2}$

33:                 $t_3 \leftarrow \max(2\lceil \log_2(2N_1 - 1)\rceil + 3, 2\lceil \log_2(2N_2 - 2)\rceil + 3,$
                    $2\lceil \log_2(2N_3 - 1)\rceil + 3, 2\lceil \log_2(2N_4 - 2)\rceil + 3)$

Algorithm 10 Cosine (Continued)

34:          **if** $p_x - s_1 - s_2 - t_3 \geq p$ **then**

35:            $\frac{m_3}{n_3} = 8 \cdot (-1)^k \big(\sum_{i=0}^{N_1} \frac{(-1)^i}{(2i+1)!} \big(\frac{\overline{m}}{4\overline{n}} \big)^{2i+1} \big) \big(\sum_{i=0}^{N_2} \frac{(-1)^i}{(2i)!} \big(\frac{\overline{m}}{4\overline{n}} \big)^{2i} \big)$

           $\big(\sum_{i=0}^{N_3} \frac{(-1)^i}{(2i+1)!} \big(\frac{m'}{4n'} - \frac{\overline{m}}{4\overline{n}} \big)^{2i+1} \big) \big(\sum_{i=0}^{N_4} \frac{(-1)^i}{(2i)!} \big(\frac{m'}{4n'} - \frac{\overline{m}}{4\overline{n}} \big)^{2i} \big)$

36:            **return** $\frac{m_3}{n_3}$

37:          **else**

38:            $p_x \leftarrow p_x + 1$

39:     **until** *true*

transcendental functions. In [15], the approximations of the transcendental functions are treated separately and in a more ad-hoc way. In contrast, our approach is based on relative errors. We provide detailed proofs of correctness for each operation and use perturbation analysis to identify essential recomputations.

In [1] the authors propose a layered framework for computations with real numbers. The lowest layer is an implementation of floating point arithmetic. In the second layer, arithmetic operations are approximated using polynomial models. The highest layer supports more advanced features such as differential operations. Proof of correctness in a theorem prover and an implementation based on [1] are also available. In this article, we have focused on approximating specific operations, whereas in [1] polynomial models are discussed in an abstract way without concrete examples from well-known arithmetic operations.

The approach introduced in [18] is based on linear fractional transformations (LFTs). Computations are encoded as trees of LFTs; various operations are defined to extract the result of a computation from the corresponding tree. However, this approach does not specify a top-down scheme to relate the desired precision in the output and the required operations on the expression tree. Expression trees are evaluated using lazy evaluation; computations terminate when adequate information is available in the root of a tree.

A symbolic approach to exact real arithmetic has been proposed in [8]. The author uses infinite binary sequences in the golden ratio base to represent real numbers. To calculate an expression, first the symbolic techniques available in Maple are applied to obtain a simplified expression. Additional Maple procedures are implemented by the author to extract binary sequences from simplified expressions. Performing operations on binary sequences is also possible. However, choosing a suitable balance between symbolic computations and direct manipulation of binary sequences depends on the given expression. As indicated in [8], using this approach to its full potentials requires expertise in Maple. Moreover, the procedure might need adaptations for each problem.

9   Conclusion

In this article, we proposed a simple representation for real numbers and discussed a top-down approach for approximating various arithmetic operations

with arbitrary precision. The focus was on:

- providing complete algorithms and proofs of correctness for the approximations, and

- perturbation analysis to identify essential iterative computations.

Existing exact real arithmetic approaches have explored different representations for real numbers; approximations for algebraic operations and transcendental functions have also been proposed based on these representations. As far as we can see, proofs of correctness for existing approaches are restricted to basic operations. Moreover, no formal reasoning is provided to prove the necessity of iterative computations.

We envisage various extensions of the presented approach. From a practical point of view, some optimizations are essential. For example, the coefficients m and n in the representation (m, n, p) can grow rapidly during computations. Thus, space efficiency is a relevant concern. One can consider an alternative representation in which large coefficients are represented in a more efficient way. Moreover, the computational efficiency of the transcendental functions can be improved by reducing the amount of required computations (i.e, number of rectangles in Riemann sums, number of terms in Taylor expansions) to guarantee the desired precision.

As discussed in Section 5.4, in certain computational problems, computing the expressions as they are would lead to loss of precision, whereas rewriting the expressions would allow us to compute them in one pass. Our top-down approach can be extended with a set of rewrite rules that transform problematic expressions into expressions that can be calculated in one pass.

References

[1] P. Collins, M. Niqui, and N. Revol. A validated real function calculus. *Mathematics in Computer Science*, 5(4):437–467, 2011.

[2] A. Feldstein and P. Turner. Overflow, underflow, and severe loss of significance in floating-point addition and subtraction. *IMA journal of numerical analysis*, 6(2):241–251, 1986.

[3] P. Gowland and D. Lester. The correctness of an implementation of exact arithmetic. In *Proceedings of RNC 2000*, volume 140, 2000.

[4] P. Hertling. Computable real functions: Type 1 computability versus type 2 computability. In *Proceedings of CCA 1996*. Mathematik/Informatik, Universität Trier, 1996.

[5] N. Higham. *Accuracy and stability of numerical algorithms*. SIAM, 2002.

[6] N. Higham. *Functions of matrices: theory and computation.* SIAM, 2008.

[7] A. Hohmann and P. Deuflhard. *Numerical Analysis in Modern Scientific Computing: An Introduction,* volume 43. Springer Science & Business Media, 2012.

[8] T. Kelsey. Exact numerical computation via symbolic computation. In *Proceedings of CCA 2000,* pages 187–197. Springer, 2000.

[9] D. Kincaid and E. Cheney. *Numerical analysis: mathematics of scientific computing,* volume 2. American Mathematical Soc., 2002.

[10] R. Krebbers and B. Spitters. Type classes for efficient exact real arithmetic in Coq. *Logical Methods in Computer Science,* 9(1:1):1–27, 2013.

[11] G. Kreisel, D. Lacombe, and J. Shoenfield. Partial recursive functionals and effective operations. *Constructivity in Mathematics, Studies in Logic and the Foundations of Mathematics,* pages 290–297, 2000.

[12] B. Kushner and L. Lefman. *Lectures on constructive mathematical analysis,* volume 60. American Mathematical Soc., 1984.

[13] T. M. library. `http://www.mpfr.org`. Visited: August 2015.

[14] A. Markov. On the continuity of constructive functions (in Russian). *Uspekhi Mat. Nauk (NS),* 9:226–230, 1954.

[15] V. Ménissier-Morain. Arbitrary precision real arithmetic: design and algorithms. *The Journal of Logic and Algebraic Programming,* 64(1):13–39, 2005.

[16] N. Müller. The iRRAM: Exact arithmetic in C++. In *Proceedings of CCA 2001,* pages 222–252. Springer, 2001.

[17] R. O'Connor. Certified exact transcendental real number computation in Coq. In *Theorem Proving in Higher Order Logics,* pages 246–261. Springer, 2008.

[18] P. Potts. *Exact real arithmetic using Mobius transformations.* PhD thesis, PhD-thesis, Imperial College London, 1998.

[19] M. Richter and K. Wong. Computable preference and utility. *Journal of Mathematical Economics,* 32(3):339–354, 1999.

[20] M. Spivak. *Calculus.* Benjamin, 1967.

[21] K. Weihrauch. *Computable analysis: an introduction.* Springer Science & Business Media, 2012.

A Useful Propositions & Lemmas

Proposition 1. *For any $p \in \mathbb{N}$ the following inequalities hold:*

$$(1 + \frac{1}{2^p})^2 \leq 1 + \frac{1}{2^{p-2}} \tag{84}$$

$$1 - \frac{1}{2^{p-2}} \leq (1 - \frac{1}{2^p})^2 \tag{85}$$

Proof. For inequality (84) we can write:

$$(1 + \frac{1}{2^p})^2 = 1 + \frac{1}{2^{2p}} + \frac{1}{2^{p-1}} \leq 1 + \frac{1}{2^{p-1}} + \frac{1}{2^{p-1}} = 1 + \frac{1}{2^{p-2}}$$

Similarly for inequality (85) we have:

$$(1 - \frac{1}{2^p})^2 = 1 + \frac{1}{2^{2p}} - \frac{1}{2^{p-1}} \geq 1 - \frac{1}{2^{p-1}} > 1 - \frac{1}{2^{p-2}}$$

$\square$

Proposition 2. *For any $p \in \mathbb{N}^+$ the following inequalities hold:*

$$\frac{2^p}{2^p - 1} \leq 1 + \frac{1}{2^{p-1}} \tag{86}$$

$$1 - \frac{1}{2^{p-1}} \leq \frac{2^p}{2^p + 1} \tag{87}$$

Proof. Since $p \in \mathbb{N}^+$ we have $2^p \geq 2$. For inequality (86) we have:

$$\frac{2^p}{2^p - 1} = \frac{2^p - 1 + 1}{2^p - 1} = 1 + \frac{1}{2^p - 1} \leq 1 + \frac{1}{2^{p-1}}$$

Similarly for inequality (87) we can write:

$$\frac{2^p}{2^p + 1} = \frac{2^p + 1 - 1}{2^p + 1} = 1 - \frac{1}{2^p + 1} \geq 1 - \frac{1}{2^{p-1}}$$

$\square$

Proposition 3. *For any $0 < y < 1$ the following inequality holds:*

$$\log_2(1 + y) \leq \frac{y}{\ln(2)} \tag{88}$$

Proof. Based on Taylor's theorem, we can write the following expansion for $\log_2(1 + y)$:

$$\log_2(1 + y) = \sum_{i=1}^{N} (-1)^{i-1} \frac{y^i}{i \ln(2)} + \frac{(-1)^N y^{N+1}}{(N + 1)(1 + c_y)^{N+1} \ln(2)}$$

where $0 < c_y < y$. For $N = 1$ we obtain:

$$\log_2(1 + y) = \frac{y}{\ln(2)} - \frac{y^2}{2(1 + c_1)^2 \ln(2)}$$

The quantity $\frac{y^2}{2(1+c_1)^2 \ln(2)}$ is positive and hence $\log_2(1 + y) \leq \frac{y}{\ln(2)}$.

$\square$

Proposition 4. *For $x \in \mathbb{R} \setminus \{0\}$ we have $\left|\frac{x}{(1+x^2)\arctan(x)}\right| < 1$.*

Proof. We analyze the derivative of $f(x) = \left|\frac{x}{(1+x^2)\arctan(x)}\right| = \frac{x}{(1+x^2)\arctan(x)}$ and find the intervals in which $f(x)$ is increasing/decreasing.

$$f'(x) = \frac{\arctan(x)(1 - x^2) - x}{(1 + x^2)^2(\arctan(x))^2}$$

We analyze the derivative in the following cases:

1. Suppose $0 < x < 1$. We use the Taylor expansion of $\arctan(x)$ to rewrite $\arctan(x)(1 - x^2) - x$:

$$\arctan(x)(1 - x^2) - x = \sum_{i=0}^{\infty} \frac{(-1)^i}{2i + 1} x^{2i+1} - x - x^2 \arctan(x)$$

$$= \sum_{i=1}^{\infty} \frac{(-1)^i}{2i + 1} x^{2i+1} - x^2 \arctan(x)$$

$$= \sum_{i=1}^{\infty} -x^{4i-1}\left(\frac{1}{4i - 1} - \frac{x^2}{4i + 1}\right) - x^2 \arctan(x)$$

The term $-x^2 \arctan(x)$ is negative and $-x^{4i-1}\left(\frac{1}{4i-1} - \frac{x^2}{4i+1}\right)$ is negative for $i \geq 1$ and $0 < x < 1$. Thus, the summation $\sum_{i=1}^{\infty} -x^{4i-1}\left(\frac{1}{4i-1} - \frac{x^2}{4i+1}\right) - x^2 \arctan(x)$ is negative and $f'(x) < 0$. The function $f(x)$ is decreasing for $0 < x < 1$.

2. Suppose $x \geq 1$. In this case, $\arctan(x)(1 - x^2) - x < 0$. Thus, $f'(x) < 0$ and $f(x)$ is decreasing for $x \geq 1$.

3. Suppose $x < 0$. Since $f'(-x) = -f'(x)$, we can conclude from the first two cases that $f(x)$ is increasing when $x < 0$.

The case analysis shows that $f(x)$ is bounded from above by $\lim_{x \to 0} f(x) = 1$.

$\square$

Proposition 5. *For any $y \in (-1, 0) \cup (0, 1)$ and $i \in \mathbb{N}$ the following inequalities hold:*

$$\frac{(-1)^{2i}}{(4i + 1)!} y^{4i+1} + \frac{(-1)^{2i+1}}{(4i + 3)!} y^{4i+3} > 0 \qquad \text{if } y > 0$$

$$\frac{(-1)^{2i}}{(4i + 1)!} y^{4i+1} + \frac{(-1)^{2i+1}}{(4i + 3)!} y^{4i+3} < 0 \qquad \text{if } y < 0$$

Proof. We can rewrite the summation as follows:

$$\frac{(-1)^{2i}}{(4i+1)!}y^{4i+1} + \frac{(-1)^{2i+1}}{(4i+3)!}y^{4i+3} = \frac{(-1)^{2i}}{(4i+1)!}y^{4i+1}\left(1 - \frac{y^2}{(4i+2)(4i+3)}\right)$$

Since $1 - \frac{y^2}{(4i+2)(4i+3)} \geq 1 - \frac{1}{6} > 0$ the sign of the summation is determined by the sign of y^{4i+1}.

$\square$

Proposition 6. *For any $y \in (-1,0) \cup (0,1)$ and $i \in \mathbb{N}$ the following inequality holds:*

$$\frac{(-1)^{2i}}{(4i)!}y^{4i} + \frac{(-1)^{2i+1}}{(4i+2)!}y^{4i+2} > 0$$

Proof. We can rewrite the left hand side of the inequality as follows:

$$\frac{(-1)^{2i}}{(4i)!}y^{4i} + \frac{(-1)^{2i+1}}{(4i+2)!}y^{4i+2} = \frac{(-1)^{2i}}{(4i)!}y^{4i}\left(1 - \frac{1}{(4i+1)(4i+2)}y^2\right)$$

Since $1 - \frac{1}{(4i+1)(4i+2)}y^2 \geq 1 - \frac{1}{2} > 0$, the inequality holds.

$\square$

Proposition 7. *For $x \in (-1,1) \setminus \{0\}$ we have $|x \cdot \cot(x)| < 1$.*

Proof. For the interval $x \in (-1,1) \setminus \{0\}$, we have $|x \cdot \cot(x)| = x \cdot \cot(x)$. We analyze the derivative of $f(x) = x \cdot \cot(x)$ in $(-1,1) \setminus \{0\}$.

$$f'(x) = \cot(x) - x(1 + \cot^2(x)) = \frac{\sin(x)\cos(x) - x}{\sin^2(x)} = \frac{\sin(2x) - 2x}{2\sin^2(x)}$$

The point $x = 0$ is a critical point for $f(x)$. Since $|\sin(2x)| \leq |2x|$, $f(x)$ is increasing in $(-1,0)$ and decreasing in $(0,1)$. Thus, $f(x)$ is bounded from above by $\lim_{x \to 0} f(x) = 1$.

$\square$

Proposition 8. *For $x \in (-1,1)$ we have $|x \cdot \tan(x)| < \tan(1)$.*

Proof. For the interval $x \in (-1,1)$, we have $|x \cdot \tan(x)| = x \cdot \tan(x)$. To determine an upper-bound for $x \cdot \tan(x)$, we analyze the derivative of $f(x) = x \cdot \tan(x)$:

$$f'(x) = \tan(x) + x(1 + \tan^2(x)) = \frac{\sin(x)\cos(x) + x}{\cos^2(x)} = \frac{\sin(2x) + 2x}{2\cos^2(x)}$$

The quantity $\sin(2x) + 2x$ is negative in $(-1,0)$ and positive in $(0,1)$. Thus, $f(x)$ is decreasing in $(-1,0)$ and increasing in $(0,1)$. We conclude that $f(x) < \tan(1)$.

$\square$

Lemma 1. *Let x be a real number represented by (m,n,p). Then x^i can be represented by $(m^i, n^i, p - 2\lceil \log_2^i \rceil)$.*

Proof. We can apply Theorem 5.2 to calculate $x^i = x^{\lceil \frac{i}{2} \rceil} \times x^{\lfloor \frac{i}{2} \rfloor}$. Let $k = \lceil \log_2^i \rceil$ and $P(i)$ denote the precision that we lose by calculating x^i. From Theorem 5.2 we can write:

$$P(i) \leq P(\lceil \frac{i}{2} \rceil) + 2 \leq P(\lceil \frac{i}{2^k} \rceil) + 2k = 2k$$

Thus, we lose $2k = 2\lceil \log_2^i \rceil$ units of precision by calculating x^i.

$\square$