SC22/WG11 N462

INTERNATIONAL STANDARD

ISO/IEC 10967-2

Fourth committee draft
1999-09-30

Information technology — Language independent arithmetic —

Part 2: Elementary numerical functions

Technologies de l'information —
Arithmétique indépendante de langage —

Partie 2: Fonctions numériques élémentaires

FINAL COMMITTEE DRAFT
August 21, 2003 15:11

Editor:
Kent Karlsson
IMI, Industri-Matematik International
Kungsgatan 12
SE-411 19 Göteborg
SWEDEN
Telephone: +46-31 10 22 44
Facsimile: +46-31 13 13 25
E-mail: keka@im.se

Reference number
ISO/IEC CD 10967-2.4:1999(E)

Contents

1 Scope **1**
1.1 Inclusions 1
1.2 Exclusions 2

2 Conformity **2**

3 Normative references **3**

4 Symbols and definitions **4**
4.1 Symbols 4
4.1.1 Sets and intervals 4
4.1.2 Operators and relations 4
4.1.3 Mathematical functions 4
4.1.4 Datatypes and exceptional values 5
4.2 Definitions of terms 6

5 Specifications for the numerical functions **9**
5.1 Basic integer operations 9
5.1.1 The integer *result* and *wrap* helper functions 9
5.1.2 Integer maximum and minimum 10
5.1.3 Integer diminish 10
5.1.4 Integer power and arithmetic shift 10
5.1.5 Integer square root 11
5.1.6 Divisibility tests 11
5.1.7 Integer division and remainder 11
5.1.8 Greatest common divisor and least common positive multiple 12
5.1.9 Support operations for extended integer range 13
5.2 Basic floating point operations 13
5.2.1 The rounding and floating point *result* helper functions 14
5.2.2 Floating point maximum and minimum 15
5.2.3 Floating point diminish 17
5.2.4 Round, floor, and ceiling 17
5.2.5 Remainder after division with round to integer 18
5.2.6 Square root and reciprocal square root 18
5.2.7 Support operations for extended floating point precision 19
5.3 Elementary transcendental floating point operations 21
5.3.1 Maximum error requirements 21
5.3.2 Sign requirements 21
5.3.3 Monotonicity requirements 22
5.3.4 The *trans_result* helper function 22
5.3.5 Hypotenuse 22
5.3.6 Operations for exponentiations and logarithms 23

© ISO/IEC 1999

All rights reserved. No part of this publication may be reproduced or utilised in any form or by any means, electronic or mechanical, including photocopying and microfilm, without permission in writing from the publisher.

ISO/IEC Copyright Office • Case Postale 56 • CH-1211 Genève 20 • Switzerland

Printed in Switzerland

5.3.6.1 Integer power of argument base 23
5.3.6.2 Natural exponentiation 24
5.3.6.3 Natural exponentiation, minus one 24
5.3.6.4 Exponentiation of 2 25
5.3.6.5 Exponentiation of 10 26
5.3.6.6 Exponentiation of argument base 26
5.3.6.7 Exponentiation of one plus the argument base, minus one 27
5.3.6.8 Natural logarithm 28
5.3.6.9 Natural logarithm of one plus the argument 28
5.3.6.10 2-logarithm 28
5.3.6.11 10-logarithm 29
5.3.6.12 Argument base logarithm 29
5.3.6.13 Argument base logarithm of one plus each argument 30
5.3.7 Operations for hyperbolic elementary functions 30
5.3.7.1 Hyperbolic sine 31
5.3.7.2 Hyperbolic cosine 31
5.3.7.3 Hyperbolic tangent 32
5.3.7.4 Hyperbolic cotangent 32
5.3.7.5 Hyperbolic secant 33
5.3.7.6 Hyperbolic cosecant 33
5.3.7.7 Inverse hyperbolic sine 33
5.3.7.8 Inverse hyperbolic cosine 34
5.3.7.9 Inverse hyperbolic tangent 34
5.3.7.10 Inverse hyperbolic cotangent 35
5.3.7.11 Inverse hyperbolic secant 35
5.3.7.12 Inverse hyperbolic cosecant 35
5.3.8 Introduction to operations for trigonometric elementary functions 36
5.3.9 Operations for radian trigonometric elementary functions 36
5.3.9.1 Radian angle normalisation 37
5.3.9.2 Radian sine 38
5.3.9.3 Radian cosine 38
5.3.9.4 Radian tangent 39
5.3.9.5 Radian cotangent 39
5.3.9.6 Radian secant 39
5.3.9.7 Radian cosecant 40
5.3.9.8 Radian cosine with sine 40
5.3.9.9 Radian arc sine 40
5.3.9.10 Radian arc cosine 41
5.3.9.11 Radian arc tangent 41
5.3.9.12 Radian arc cotangent 42
5.3.9.13 Radian arc secant 43
5.3.9.14 Radian arc cosecant 44
5.3.9.15 Radian angle from Cartesian co-ordinates 44
5.3.10 Operations for trigonometrics with given angular unit 45
5.3.10.1 Argument angular-unit angle normalisation 45
5.3.10.2 Argument angular-unit sine 46
5.3.10.3 Argument angular-unit cosine 47
5.3.10.4 Argument angular-unit tangent 47
5.3.10.5 Argument angular-unit cotangent 48
5.3.10.6 Argument angular-unit secant 48
5.3.10.7 Argument angular-unit cosecant 49

5.3.10.8 Argument angular-unit cosine with sine 49
5.3.10.9 Argument angular-unit arc sine 50
5.3.10.10 Argument angular-unit arc cosine 50
5.3.10.11 Argument angular-unit arc tangent 51
5.3.10.12 Argument angular-unit arc cotangent 51
5.3.10.13 Argument angular-unit arc secant 52
5.3.10.14 Argument angular-unit arc cosecant 53
5.3.10.15 Argument angular-unit angle from Cartesian co-ordinates 53
5.3.11 Operations for angular-unit conversions 54
5.3.11.1 Converting radian angle to argument angular-unit angle 54
5.3.11.2 Converting argument angular-unit angle to radian angle 55
5.3.11.3 Converting argument angular-unit angle to (another) argument angular-unit angle 56
5.4 Conversion operations 57
5.4.1 Integer to integer conversions 58
5.4.2 Floating point to integer conversions 58
5.4.3 Integer to floating point conversions 59
5.4.4 Floating point to floating point conversions 59
5.4.5 Floating point to fixed point conversions 60
5.4.6 Fixed point to floating point conversions 61
5.5 Numerals as operations in the programming language 62
5.5.1 Numerals for integer datatypes 62
5.5.2 Numerals for floating point datatypes 62

6 Notification 63
6.1 Continuation values 63

7 Relationship with language standards 63

8 Documentation requirements 64

Annexes

A Partial conformity 67
A.1 Maximum error relaxation 67
A.2 Extra accuracy requirements relaxation 67
A.3 Relationships to other operations relaxation 68

B Rationale 69
B.1 Scope 69
B.1.1 Inclusions 69
B.1.2 Exclusions 69
B.2 Conformity 70
B.3 Normative references 70
B.4 Symbols and definitions 70
B.4.1 Symbols 70
B.4.1.1 Sets and intervals 70
B.4.1.2 Operators and relations 70
B.4.1.3 Mathematical functions 70
B.4.1.4 Datatypes and exceptional values 71
B.4.2 Definitions of terms 71
B.5 Specifications for the numerical functions 72

B.5.1 Basic integer operations . . . 72
B.5.1.1 The integer *result* and *wrap* helper functions . . . 72
B.5.1.2 Integer maximum and minimum . . . 72
B.5.1.3 Integer diminish . . . 73
B.5.1.4 Integer power and arithmetic shift . . . 73
B.5.1.5 Integer square root . . . 73
B.5.1.6 Divisibility tests . . . 73
B.5.1.7 Integer division and remainder . . . 73
B.5.1.8 Greatest common divisor and least common positive multiple . . . 74
B.5.1.9 Support operations for extended integer range . . . 74
B.5.2 Basic floating point operations . . . 74
B.5.2.1 The rounding and floating point *result* helper functions . . . 75
B.5.2.2 Floating point maximum and minimum . . . 76
B.5.2.3 Floating point diminish . . . 76
B.5.2.4 Round, floor, and ceiling . . . 76
B.5.2.5 Remainder after division and round to integer . . . 76
B.5.2.6 Square root and reciprocal square root . . . 76
B.5.2.7 Support operations for extended floating point precision . . . 77
B.5.3 Elementary transcendental floating point operations . . . 78
B.5.3.1 Maximum error requirements . . . 78
B.5.3.2 Sign requirements . . . 78
B.5.3.3 Monotonicity requirements . . . 79
B.5.3.4 The *trans_result* helper function . . . 79
B.5.3.5 Hypotenuse . . . 79
B.5.3.6 Operations for exponentiations and logarithms . . . 79
B.5.3.7 Operations for hyperbolic elementary functions . . . 80
B.5.3.8 Introduction to operations for trigonometric elementary functions 81
B.5.3.9 Operations for radian trigonometric elementary functions . . . 82
B.5.3.10 Operations for trigonometrics given angular unit . . . 84
B.5.3.11 Operations for angular-unit conversions . . . 84
B.5.4 Conversion operations . . . 85
B.5.5 Numerals as operations in the programming language . . . 85
B.5.5.1 Numerals for integer datatypes . . . 85
B.5.5.2 Numerals for floating point datatypes . . . 85
B.6 Notification . . . 86
B.6.1 Continuation values . . . 87
B.7 Relationship with language standards . . . 87
B.8 Documentation requirements . . . 87

C Example bindings for specific languages 89
C.1 General comments . . . 90
C.2 Ada . . . 90
C.3 BASIC . . . 95
C.4 C . . . 98
C.5 C++ . . . 103
C.6 Fortran . . . 108
C.7 Haskell . . . 113
C.8 Java . . . 118
C.9 Common Lisp . . . 123
C.10 ISLisp . . . 127
C.11 Modula-2 . . . 132
C.12 Pascal and Extended Pascal . . . 136

C.13 PL/I 140
C.14 SML 145

D Bibliography 151

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialised system for world-wide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organisations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work.

In the field of information technology, ISO and IEC have established a joint technical committee, ISO/IEC JTC 1, *Implementation of information technology*. Draft International Standards adopted by the joint technical committee are circulated to national bodies for voting. Publication as an International Standard requires approval by at least 75 % of the national bodies casting a vote.

International Standard ISO/IEC 10967-2 was prepared by Joint Technical Committee ISO/IEC JTC 1, Sub-Committee SC 22, *Programming languages, their environments and system software interfaces*.

ISO/IEC 10967 consists of the following parts, under the general title *Information technology — Language independent arithmetic*:

- *Part 1: Integer and floating point arithmetic*
- *Part 2: Elementary numerical functions*
- *Part 3: Complex floating point arithmetic and complex elementary numerical functions*

Additional parts will specify other arithmetic datatypes or arithmetic operations.

Introduction

Portability is a key issue for scientific and numerical software in today's heterogeneous computing environment. Such software may be required to run on systems ranging from personal computers to high performance pipelined vector processors and massively parallel systems, and the source code may be ported between several programming languages.

Part 1 of ISO/IEC 10967 specifies the basic properties of integer and floating point types that can be relied upon in writing portable software.

The aims for this Part, Part 2 of ISO/IEC 10967, are extensions of the aims for Part 1: to ensure adequate accuracy for numerical computation, predictability, notification on the production of exceptional results, and compatibility with language standards.

The content of this Part is based on Part 1, and extends Part 1's specifications to specifications for operations approximating real elementary functions, operations often required (usually without a detailed specification) by the standards for programming languages widely used for scientific software. This Part also provides specifications for conversions between the "internal" values of an arithmetic datatype, and a very close approximation in, e.g., the decimal radix. It does not cover the further transformation to decimal string format, which is usually provided by language standards. This Part also includes specifications for a number of other functions deemed useful, even though they may not be stipulated by language standards.

The numerical functions covered by this Part are computer approximations to mathematical functions of one or more real arguments. Accuracy versus performance requirements often vary with the application at hand. This is recognised by recommending that implementors support more than one library of these numerical functions. Various documentation and (program available) parameters requirements are specified to assist programmers in the selection of the library best suited to the application at hand.

Annex B is intended to be read in parallel with the standard.

Notes and annexes B to D are for information only.

Information technology — Language independent arithmetic —

Part 2: Elementary numerical functions

1 Scope

This Part of ISO/IEC 10967 defines the properties of numerical approximations for many of the real elementary numerical functions available in standard libraries for a variety of programming languages in common use for mathematical and numerical applications.

An implementor may choose any combination of hardware and software support to meet the specifications of this Part. It is the computing environment, as seen by the programmer/user, that does or does not conform to the specifications.

The term *implementation* of this Part denotes the total computing environment pertinent to this Part, including hardware, language processors, subroutine libraries, exception handling facilities, other software, and documentation.

1.1 Inclusions

The specifications of Part 1 of are included by reference in this Part.

This Part provides specifications for numerical functions for which all operand values are of integer or floating point datatypes satisfying the requirements of Part 1. Boundaries for the occurrence of exceptions and the maximum error allowed are prescribed for each specified operation. Also the result produced by giving a special value operand, such as an infinity, or a **NaN**, is prescribed for each specified floating point operation.

This Part covers most numerical functions required by the ISO/IEC standards for Ada [11], Basic [17], C [18], C++ [19], Fortran [23], ISLisp [25], Pascal [28], and PL/I [30]. In particular, specifications are provided for

a) some additional integer operations,

b) some additional non-transcendental floating point operations, including maximum and minimum operations,

c) exponentiations, logarithms, hyperbolics, and

d) trigonometrics, both in radians and for argument-given angular unit with degrees as a special case.

This Part also provides specifications for

e) conversions between integer and floating point datatypes (possibly with different radices) conforming to the requirements of Part 1, and

f) the conversion operations used, for example, in text input and output of integer and floating point numbers,

g) the results produced by an included floating point operation when one or more operand values are IEC 60559 special values, and

h) program-visible parameters that characterise certain aspects of the operations.

1.2 Exclusions

This Part provides no specifications for:

a) Numerical functions whose operands are of more than one datatype (with one exception). This standard neither requires nor excludes the presence of such "mixed operand" operations.

b) An interval datatype, or the operations on such data. This standard neither requires nor excludes such data or operations.

c) A fixed point datatype, or the operations on such data. This standard neither requires nor excludes such data or operations.

d) A rational datatype, or the operations on such data. This standard neither requires nor excludes such data or operations.

e) Complex, matrix, statistical, or symbolic operations. This standard neither requires nor excludes such data or operations.

f) The properties of arithmetic datatypes that are not related to the numerical process, such as the representation of values on physical media.

g) The properties of integer and floating point datatypes that properly belong in language standards or other spcification. Examples include

   1) the syntax of numerals and expressions in the programming language,

   2) the syntax used for parsed (input) or generated (output) character string forms for numerals by any specific programming language or library,

   3) the precedence of operators,

   4) the consequences of applying an operation to values of improper datatype, or to uninitialised data,

   5) the rules for assignment, parameter passing, and returning value,

   6) the presence or absence of automatic datatype coercions.

Furthermore, this Part does not provide specifications for:

h) how numerical functions should be implemented,

i) which algorithms are to be used for the various operations.

2 Conformity

It is expected that the provisions of this Part of ISO/IEC 10967 will be incorporated by reference and further defined in other International Standards; specifically in language standards and in language binding standards.

A binding standard specifies the correspondence between one or more operations and parameters specified in this Part and the concrete language syntax of some programming language. More generally, a binding standard specifies the correspondence between certain operations and the elements of some arbitrary computing entity. A language standard that explicitly provides such binding information can serve as a binding standard.

Conformity to this Part is always with respect to a specified set of operations. Conformity to this Part implies conformity to Part 1 for the integer and floating point datatypes used.

When a binding standard for a language exists, an implementation shall be said to conform to this Part if and only if it conforms to the binding standard. In the case of conflict between a binding standard and this Part, the specifications of the binding standard takes precedence.

When a binding standard covers only a subset of the operations defined in this Part, an implementation remains free to conform to this Part with respect to other operations independently of that binding standard.

When no binding standard for a language and some operations specified in this Part exists, an implementation conforms to this Part if and only if it provides one or more operations that together satisfy all the requirements of clauses 5 through 8 that are relevant to those operations. The implementation shall then document the binding.

An implementation is free to provide operations that do not conform to this Part, or that are beyond the scope of this Part. The implementation shall not claim or imply conformity to this Part with respect to such operations.

An implementation is permitted to have modes of operation that do not conform to this Part. A conforming implementation shall specify how to select the modes of operation that ensure conformity.

NOTES

1 Language bindings are essential. Clause 8 requires an implementation to supply a binding if no binding standard exists. See annex C for suggested language bindings.

2 A complete binding for this Part will include (explicitly or by reference) a binding for Part 1 as well, which in turn may include (explicitly or by reference) a binding for IEC 60559 as well.

3 It is not possible to conform to this Part without specifying to which set of operations conformity is claimed.

3 Normative references

The following standards contain provisions which, through reference in this text, constitute provisions of this Part. At the time of publication, the editions indicated were valid. All standards are subject to revision, and parties to agreements based on this Part are encouraged to investigate the possibility of applying the most recent edition of the standards indicated below. Members of IEC and ISO maintain registers of currently valid International Standards.

IEC 60559:1989, *Binary floating-point arithmetic for microprocessor systems.*

ISO/IEC 10967-1:1994, *Information technology — Language independent arithmetic — Part 1: Integer and floating point arithmetic.*

4 Symbols and definitions

4.1 Symbols

4.1.1 Sets and intervals

In this Part, $\mathcal{Z}$ denotes the set of mathematical integers, $\mathcal{R}$ denotes the set of classical real numbers, and $\mathcal{C}$ denotes the set of complex numbers over $\mathcal{R}$. Note that $\mathcal{Z} \subset \mathcal{R} \subset \mathcal{C}$.

$[x, z]$ designates the interval $\{y \in \mathcal{R} \mid x \leqslant y \leqslant z\}$,
$]x, z]$ designates the interval $\{y \in \mathcal{R} \mid x < y \leqslant z\}$,
$[x, z[$ designates the interval $\{y \in \mathcal{R} \mid x \leqslant y < z\}$, and
$]x, z[$ designates the interval $\{y \in \mathcal{R} \mid x < y < z\}$.

NOTE – The notation using a round bracket for an open end of an interval is not used, for the risk of confusion with the notation for pairs.

4.1.2 Operators and relations

All prefix and infix operators have their conventional (exact) mathematical meaning. The conventional notation for set definition and manipulation is also used. In particular this Part uses

$\Rightarrow$ and $\Leftrightarrow$ for logical implication and equivalence
$+$, $-$, $/$, $|x|$, $\lfloor x \rfloor$, $\lceil x \rceil$, and round(x) on reals
$\cdot$ for multiplication on reals
$<$, $\leqslant$, $=$, $\neq$, $\geqslant$, and $>$ between reals
max on non-empty upwardly closed sets of reals
min on non-empty downwardly closed sets of reals
$\cup$, $\cap$, $\times$, $\in$, $\notin$, $\subset$, $\subseteq$, $\not\subseteq$, $\neq$, and $=$ with sets
$\times$ for the Cartesian product of sets
$\rightarrow$ for a mapping between sets
$|$ for the divides relation between integers

For $x \in \mathcal{R}$, the notation $\lfloor x \rfloor$ designates the largest integer not greater than x:

$$\lfloor x \rfloor \in \mathcal{Z} \quad \text{and} \quad x - 1 < \lfloor x \rfloor \leqslant x$$

the notation $\lceil x \rceil$ designates the smallest integer not less than x:

$$\lceil x \rceil \in \mathcal{Z} \quad \text{and} \quad x \leqslant \lceil x \rceil < x + 1$$

and the notation round(x) designates the integer closest to x:

$$\mathrm{round}(x) \in \mathcal{Z} \quad \text{and} \quad x - 0.5 \leqslant \mathrm{round}(x) \leqslant x + 0.5$$

where in case x is exactly half-way between two integers, the even integer is the result.

The *divides* relation ($|$) on integers tests whether an integer i divides an integer j exactly:

$$i|j \;\Leftrightarrow\; (i \neq 0 \text{ and } i \cdot n = j \text{ for some } n \in \mathcal{Z})$$

NOTE – $i|j$ is true exactly when j/i is defined and $j/i \in \mathcal{Z}$).

4.1.3 Mathematical functions

This Part specifies properties for a number of operations numerically approximating some of the elementary functions. The following ideal mathematical functions are defined in Chapter 4 of the *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables* [48] (e is the Napierian base):

e^x, x^y, $\sqrt{x}$, ln, $\log_b$,
sinh, cosh, tanh, coth, sech, csch, arcsinh, arccosh, arctanh, arccoth, arcsech, arccsch,
sin, cos, tan, cot, sec, csc, arcsin, arccos, arctan, arccot, arcsec, arccsc.

Many of the inverses above are multi-valued. The selection of which value to return, the principal value, so as to make the inverses into functions, is done in the conventional way. The only one over which there is some difference of conventions is the arccot function. Conventions there vary for negative arguments; either a positive return value (giving a function that is continuous over zero), or a negative value (giving a sign symmetric function). In this Part arccot refers to the continuous inverse function, and arcctg refers to the sign symmetric inverse function.

$\text{arccosh}(x) \geqslant 0$, $\text{arcsech}(x) \geqslant 0$,
$\arcsin(x) \in [-\pi/2, \pi/2]$, $\arccos(x) \in [0, \pi]$, $\arctan(x) \in \,]-\pi/2, \pi/2[$,
$\text{arccot}(x) \in \,]0, \pi[$, $\text{arcctg}(x) \in \,]-\pi/2, \pi/2]$, $\text{arcsec}(x) \in [0, \pi]$, $\text{arccsc}(x) \in [-\pi/2, \pi/2]$.

NOTES

1 *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables* [48] uses the notation arccot for what is called arcctg in this Part.

2 $e = 2.71828....$ e is not in F.

4.1.4 Datatypes and exceptional values

For pairs, define:

$fst((x, y)) = x$
$snd((x, y)) = y$

Square brackets are used to write finite sequences of values. [] is the sequence containing no values. $[s]$, is the sequence of one value, s. $[s_1, s_2]$, is the sequence of two values, s_1 and then s_2, etc. The colon operator is used to prepend a value to a sequence: $x : [x_1, ..., x_n] = [x, x_1, ..., x_n]$.

$[S]$, where S is a set, denotes the set of finite sequences, where each value in each sequence is in S.

NOTE 1 – It is always clear from context, in the text of this Part, if $[X]$ is a sequence of one element, or the set of sequences with values from X. It is also clear from context if $[x_1, x_2]$ is a sequence of two values or an interval.

The datatype **Boolean** consists of the two values **true** and **false**.

Integer datatypes and floating point datatypes are defined in Part 1.

The following symbols are defined in Part 1, and used in this Part.

Exceptional values:
underflow.
Integer parameters:
$bounded_I$, $maxint_I$, and $minint_I$.
Integer helper function:
$wrap_I$.
Integer operations:
neg_I, add_I, sub_I, and mul_I.
Floating point parameters:
r_F, p_F, $emin_F$, $emax_F$, $denorm_F$, and iec_559_F.
Derived floating point constants:
$fmax_F$, $fmin_F$, $fminN_F$, $fminD_F$, and $epsilon_F$.
Floating point rounding constants:
rnd_error_F.

Floating point value sets related to F:
F^*, F_D, and F_N.
Floating point helper functions:
e_F, $result_F$, and rnd_F.
Floating point operations:
neg_F, add_F, sub_F, mul_F, div_F, and ulp_F.

Floating point datatypes that conform to Part 1 shall, for use with this Part, have a value for the parameter p_F such that $p_F \geqslant 2 \cdot \max\{1, \log_{r_F}(2 \cdot \pi)\}$, and have a value for the parameter $emin_F$ such that $emin_F \leqslant -p_F - 1$.

NOTES

2 This implies that $fminN_F < 0.5 \cdot epsilon_F / r_F$ in this Part, rather than just $fminN_F \leqslant epsilon_F$.

3 These extra requirements, which do not limit the use of any existing floating point datatype, are made 1) so that angles in radians are not too degenerate within the first two cycles, plus and minus, when represented in F, and 2) in order to justly allow avoiding the underflow notification in specifications for the $expm1_F$ and $ln1p_F$ operations.

4 F should also be such that $p_F \geqslant 2 + \log_{r_F}(1000)$, to allow for a not too coarse angle resolution anywhere in the interval $[-big\_angle\_r_F, big\_angle\_r_F]$. See clause 5.3.9.

Three new exceptional values, **overflow**, **invalid**, and **pole**, are introduced in this Part replacing tree other exceptional values used in Part 1. One new exceptional value, **absolute_precision_underflow**, is introduced in this Part with no correspondence in Part 1. **invalid** and **pole** are in this Part used instead of the **undefined** of Part 1. **overflow** is used instead of the **integer_overflow** and **floating_overflow** of Part 1. Bindings may still distinguish between **integer_overflow** and **floating_overflow**. The exceptional value **absolute_precision_underflow** is used when the given floating point angle value argument is so big that even a highly accurate result from a trigonometric operation is questionable, due to the fact that the density of floating point values has decreased significantly at these big angle values. For the exceptional values, a continuation value may be given in parenthesis after the exceptional value.

The following symbols represent floating point values defined in IEC 60559 and used in this Part:

$\mathbf{-0}$, $\mathbf{+\infty}$, $\mathbf{-\infty}$, **qNaN**, and **sNaN**.

These floating point values are not part of the set F, but if iec_559_F has the value **true**, these values are included in the floating point datatype corresponding to F.

NOTE 5 – This Part uses the above five special values for compatibility with IEC 60559. In particular, the symbol $\mathbf{-0}$ (in bold) is not the application of (mathematical) unary $-$ to the value 0, and is a value logically distinct from 0.

The specifications cover the results to be returned by an operation if given one or more of the IEC 60559 special values $\mathbf{-0}$, $\mathbf{+\infty}$, $\mathbf{-\infty}$, or **NaN**s as input values. These specifications apply only to systems which provide and support these special values. If an implementation is not capable of representing a $\mathbf{-0}$ result or continuation value, the actual result or continuation value shall be 0. If an implementation is not capable of representing a prescribed result or continuation value of the IEC 60559 special values $\mathbf{+\infty}$, $\mathbf{-\infty}$, or **qNaN**, the actual result or continuation value is binding or implementation defined.

4.2 Definitions of terms

For the purposes of this Part, the following definitions apply:

accuracy: The closeness between the true mathematical result and a computed result.

arithmetic datatype: A datatype whose non-special values are members of $\mathcal{Z}$, $\mathcal{R}$, or $\mathcal{C}$.

NOTE 1 – This standard specifies requirements for integer and floating point datatypes. Complex numbers are not covered here, but will be included in a subsequent Part of ISO/IEC 10967 [5].

continuation value: A computational value used as the result of an arithmetic operation when an exception occurs. Continuation values are intended to be used in subsequent arithmetic processing. A continuation value can be a value in F or an IEC 60559 special value. (Contrast with *exceptional value.* See 6.1.2 of Part 1.)

denormalisation loss: A larger than normal rounding error caused by the fact that subnormal values have less than full precision. (See 5.2.5 of Part 1 for a full definition.)

denormalised, denormal: The non-zero values of a floating point type F that provide less than the full precision allowed by that type. (See F_D in 5.2 of Part 1 for a full definition.)

error: (1) The difference between a computed value and the correct value. (Used in phrases like "rounding error" or "error bound".)

(2) A synonym for *exception* in phrases like "error message" or "error output". Error and exception are not synonyms in any other context.

exception: The inability of an operation to return a suitable finite numeric result from finite arguments. This might arise because no such finite result exists mathematically, or because the mathematical result cannot be represented with sufficient accuracy.

exceptional value: A non-numeric value produced by an arithmetic operation to indicate the occurrence of an exception. Exceptional values are not used in subsequent arithmetic processing. (See clause 5 of Part 1.)

NOTES

2 Exceptional values are used as part of the defining formalism only. With respect to this Part, they do not represent values of any of the datatypes described. There is no requirement that they be represented or stored in the computing system.

3 Exceptional values are not to be confused with the NaNs and infinities defined in IEC 60559. Contrast this definition with that of *continuation value* above.

helper function: A function used solely to aid in the expression of a requirement. Helper functions are not visible to the programmer, and are not required to be part of an implementation.

implementation (of this Part)**:** The total arithmetic environment presented to a programmer, including hardware, language processors, exception handling facilities, subroutine libraries, other software, and all pertinent documentation.

literal: A syntactic entity denoting a constant value without having proper sub-entities that are expressions.

monotonic approximation: An operation $op_F : ... \times F \times ... \to F$, where the other arguments are kept constant, is a monotonic approximation of a predetermined mathematical function $h : \mathcal{R} \to \mathcal{R}$ if, for every $a \in F$ and $b \in F$,

a) h is monotonic non-decreasing on $[a, b]$ implies $op_F(..., a, ...) \leqslant op_F(..., b, ...)$,

b) h is monotonic non-increasing on $[a, b]$ implies $op_F(..., a, ...) \geqslant op_F(..., b, ...)$.

monotonic non-decreasing: A function $h : \mathcal{R} \to \mathcal{R}$ is monotonic non-decreasing on a real interval $[a, b]$ if for every x and y such that $a \leqslant x \leqslant y \leqslant b$, $h(x)$ and $h(y)$ are well-defined and $h(x) \leqslant h(y)$.

monotonic non-increasing: A function $h : \mathcal{R} \rightarrow \mathcal{R}$ is monotonic non-increasing on a real interval $[a, b]$ if for every x and y such that $a \leqslant x \leqslant y \leqslant b$, $h(x)$ and $h(y)$ are well-defined and $h(x) \geqslant h(y)$.

normalised: The non-zero values of a floating point type F that provide the full precision allowed by that type. (See F_N in 5.2 of Part 1 for a full definition.)

notification: The process by which a program (or that program's end user) is informed that an arithmetic exception has occurred. For example, dividing 2 by 0 results in a notification. (See clause 6 of Part 1 for details.)

numeral: A numeric literal. It may denote a value in $\mathcal{Z}$ or $\mathcal{R}$, $-\mathbf{0}$, an infinity, or a NaN.

numerical function: A computer routine or other mechanism for the approximate evaluation of a mathematical function.

operation: A function directly available to the user/programmer, as opposed to helper functions or theoretical mathematical functions.

pole: A mathematical function f has a pole at x_0 if x_0 is finite, f is defined, finite, monotone, and continuous in at least one side of the neighbourhood of x_0, and $\lim_{x \rightarrow x_0} f(x)$ is infinite.

precision: The number of digits in the fraction of a floating point number. (See 5.2 of Part 1.)

rounding: The act of computing a representable final result for an operation that is close to the exact (but unrepresentable) result for that operation. Note that a suitable representable result may not exist (see 5.2.6 of Part 1). (See also A.5.2.6 of Part 1 for some examples.)

rounding function: Any function $rnd : \mathcal{R} \rightarrow X$ (where X is a given discrete and unlimited subset of $\mathcal{R}$) that maps each element of X to itself, and is monotonic non-decreasing. Formally, if x and y are in $\mathcal{R}$,

$$x \in X \Rightarrow rnd(x) = x$$
$$x < y \Rightarrow rnd(x) \leqslant rnd(y)$$

Note that if $u \in \mathcal{R}$ is between two adjacent values in X, $rnd(u)$ selects one of those adjacent values.

round to nearest: The property of a rounding function rnd that when $u \in \mathcal{R}$ is between two adjacent values in X, $rnd(u)$ selects the one nearest u. If the adjacent values are equidistant from u, either may be chosen deterministically.

round toward minus infinity: The property of a rounding function rnd that when $u \in \mathcal{R}$ is between two adjacent values in X, $rnd(u)$ selects the one less than u.

round toward plus infinity: The property of a rounding function rnd that when $u \in \mathcal{R}$ is between two adjacent values in X, $rnd(u)$ selects the one greater than u.

shall: A verbal form used to indicate requirements strictly to be followed in order to conform to the standard and from which no deviation is permitted. (Quoted from the directives [1].)

should: A verbal form used to indicate that among several possibilities one is recommended as particularly suitable, without mentioning or excluding others; or that (in the negative form) a certain possibility is deprecated but not prohibited. (Quoted from the directives [1].)

signature (of a function or operation)**:** A summary of information about an operation or function. A signature includes the function or operation name; a subset of allowed argument values to the operation; and a superset of results from the function or operation (including exceptional values if any), if the argument is in the subset of argument values given in the signature.

The signature

$$add_I : I \times I \to I \cup \{\mathbf{overflow}\}$$

states that the operation named add_I shall accept any pair of I values as input, and (when given such input) shall return either a single I value as its output or the exceptional value **overflow**.

A signature for an operation or function does not forbid the operation from accepting a wider range of arguments, nor does it guarantee that every value in the result range will actually be returned for some input. An operation given an argument outside the stipulated argument domain may produce a result outside the stipulated result range.

subnormal: A denormal value, the value 0, or the value $\mathbf{-0}$.

ulp: The value of one "unit in the last place" of a floating point number. This value depends on the exponent, the radix, and the precision used in representing the number. Thus, the ulp of a normalised value x (in F), with exponent t, precision p, and radix r, is r^{t-p}, and the ulp of a subnormal value is $fminD_F$. (See 5.2 of Part 1.)

5 Specifications for the numerical functions

This clause specifies a number of helper functions and operations for integer and floating point datatypes. Each operation is given a signature and is further specified by a number of cases. These cases may refer to other operations (specified in this Part or in Part 1), to mathematical functions, and to helper functions (specified in this Part or in Part 1). They also use special abstract values ($\mathbf{-\infty}, \mathbf{+\infty}, \mathbf{-0}, \mathbf{qNaN}, \mathbf{sNaN}$). For each datatype, two of these abstract values may represent several actual values each: **qNaN** and **sNaN**. Finally, the specifications may refer to exceptional values.

The signatures in the specifications in this clause specify only all non-special values as input values, and indicate as output values the superset of all non-special, special, and exceptional values that may result from these (non-special) input values. Therefore, exceptional and special values that can never result from non-special input values are not included in the signatures given. Also, signatures that, for example, include IEC 60559 special values as arguments are not given in the specifications below. This does not exclude such signatures from being valid for these operations.

5.1 Basic integer operations

Clause 5.1 of Part 1 specifies integer datatypes and a number of operations on values of an integer datatype. In this clause some additional operations on values of an integer datatype are specified.

I is the set of non-special values, $I \subseteq \mathcal{Z}$, for an integer datatype conforming to Part 1. Integer datatypes conforming to Part 1 often do not contain any **NaN** or infinity values, even though they may do so. Therefore this clause has no specifications for such values as arguments or results.

5.1.1 The integer *result* and *wrap* helper functions

The $result_I$ helper function:

$$result_I : \mathcal{Z} \to I \cup \{\mathbf{overflow}\}$$

$$\begin{array}{lll} result_I(x) & = x & \text{if } x \in I \\ & = \mathbf{overflow} & \text{if } x \in \mathcal{Z} \text{ and } x \notin I \end{array}$$

The $wrap_I$ helper function:

$$wrap_I : \mathcal{Z} \to I$$

$$\begin{array}{lll} wrap_I(x) & = x & \text{if } x \in I \\ & = x - (n \cdot (maxint_I - minint_I + 1)) & \\ & & \text{if } x \in \mathcal{Z} \text{ and } x \notin I \end{array}$$

where $n \in \mathcal{Z}$ is chosen such that the result is in I.

NOTES

1 $n = \lfloor (x - minint_I)/(maxint_I - minint_I + 1) \rfloor$ if $x \in \mathcal{Z}$ and $bounded_I = \mathbf{true}$; or equivalently $n = \lceil (x - maxint_I)/(maxint_I - minint_I + 1) \rceil$ if $x \in \mathcal{Z}$ and $bounded_I = \mathbf{true}$.

2 For some wrapping basic arithmetic operations this n is computed by the '$\_ov$' operations in clause 5.1.9.

3 The $wrap_I$ helper function is also used in Part 1.

5.1.2 Integer maximum and minimum

$$max_I : I \times I \to I$$

$$\begin{array}{lll} max_I(x, y) & = \max\{x, y\} & \text{if } x, y \in I \end{array}$$

$$min_I : I \times I \to I$$

$$\begin{array}{lll} min_I(x, y) & = \min\{x, y\} & \text{if } x, y \in I \end{array}$$

$$max\_seq_I : [I] \to I \cup \{\mathbf{pole}\}$$

$$\begin{array}{lll} max\_seq_I([x_1, ..., x_n]) & & \\ & = \mathbf{pole}(-\infty) & \text{if } n = 0 \\ & = \max\{x_1, ..., x_n\} & \text{if } n \geqslant 1 \text{ and } \{x_1, ..., x_n\} \subseteq I \end{array}$$

$$min\_seq_I : [I] \to I \cup \{\mathbf{pole}\}$$

$$\begin{array}{lll} min\_seq_I([x_1, ..., x_n]) & & \\ & = \mathbf{pole}(+\infty) & \text{if } n = 0 \\ & = \min\{x_1, ..., x_n\} & \text{if } n \geqslant 1 \text{ and } \{x_1, ..., x_n\} \subseteq I \end{array}$$

5.1.3 Integer diminish

$$dim_I : I \times I \to I \cup \{\mathbf{overflow}\}$$

$$\begin{array}{lll} dim_I(x, y) & = result_I(\max\{0, x - y\}) & \text{if } x, y \in I \end{array}$$

NOTE – dim_I cannot be implemented as $max_I(0, sub_I(x, y))$ for bounded integer types, since this latter expression has other overflow properties.

5.1.4 Integer power and arithmetic shift

$$power_I : I \times I \to I \cup \{\mathbf{overflow}, \mathbf{pole}, \mathbf{invalid}\}$$

$$\begin{array}{lll} power_I(x, y) & = result_I(x^y) & \text{if } x, y \in I \text{ and } (y > 0 \text{ or } |x| = 1) \\ & = 1 & \text{if } x \in I \text{ and } x \neq 0 \text{ and } y = 0 \\ & = \mathbf{invalid}(1) & \text{if } x = 0 \text{ and } y = 0 \\ & = \mathbf{pole}(+\infty) & \text{if } x = 0 \text{ and } y \in I \text{ and } y < 0 \\ & = \mathbf{invalid}(0) & \text{if } x, y \in I \text{ and } x \notin \{-1, 0, 1\} \text{ and } y < 0 \end{array}$$

$shift2_I : I \times I \rightarrow I \cup \{\mathbf{overflow}\}$

$$shift2_I(x, y) = result_I(\lfloor x \cdot 2^y \rfloor) \quad \text{if } x, y \in I$$

$shift10_I : I \times I \rightarrow I \cup \{\mathbf{overflow}\}$

$$shift10_I(x, y) = result_I(\lfloor x \cdot 10^y \rfloor) \quad \text{if } x, y \in I$$

5.1.5 Integer square root

$sqrt_I : I \rightarrow I \cup \{\mathbf{invalid}\}$

$$\begin{aligned} sqrt_I(x) &= \lfloor \sqrt{x} \rfloor && \text{if } x \in I \text{ and } x \geqslant 0 \\ &= \mathbf{invalid}(\mathbf{qNaN}) && \text{if } x \in I \text{ and } x < 0 \end{aligned}$$

5.1.6 Divisibility tests

$divides_I : I \times I \rightarrow \mathbf{Boolean}$

$$\begin{aligned} divides_I(x, y) &= \mathbf{true} && \text{if } x, y \in I \text{ and } x|y \\ &= \mathbf{false} && \text{if } x, y \in I \text{ and not } x|y \end{aligned}$$

NOTES

1 $divides_I(0, 0) = \mathbf{false}$, since 0 does not divide anything, not even 0.

2 $divides_I$ cannot be implemented as, e.g., $eq_I(0, moda_I(y, x))$, since the remainder functions give notifications for a zero second argument.

$even_I : I \rightarrow \mathbf{Boolean}$

$$\begin{aligned} even_I(x) &= \mathbf{true} && \text{if } x \in I \text{ and } 2|x \\ &= \mathbf{false} && \text{if } x \in I \text{ and not } 2|x \end{aligned}$$

$odd_I : I \rightarrow \mathbf{Boolean}$

$$\begin{aligned} odd_I(x) &= \mathbf{true} && \text{if } x \in I \text{ and not } 2|x \\ &= \mathbf{false} && \text{if } x \in I \text{ and } 2|x \end{aligned}$$

5.1.7 Integer division and remainder

$divf_I : I \times I \rightarrow I \cup \{\mathbf{overflow}, \mathbf{pole}, \mathbf{invalid}\}$

$$\begin{aligned} divf_I(x, y) &= result_I(\lfloor x/y \rfloor) && \text{if } x, y \in I \text{ and } y \neq 0 \\ &= \mathbf{pole}(+\infty) && \text{if } x \in I \text{ and } x > 0 \text{ and } y = 0 \\ &= \mathbf{invalid}(\mathbf{qNaN}) && \text{if } x = 0 \text{ and } y = 0 \\ &= \mathbf{pole}(-\infty) && \text{if } x \in I \text{ and } x < 0 \text{ and } y = 0 \end{aligned}$$

$moda_I : I \times I \rightarrow I \cup \{\mathbf{invalid}\}$

$$\begin{aligned} moda_I(x, y) &= x - (\lfloor x/y \rfloor \cdot y) && \text{if } x, y \in I \text{ and } y \neq 0 \\ &= \mathbf{invalid}(\mathbf{qNaN}) && \text{if } x \in I \text{ and } y = 0 \end{aligned}$$

$$group_I : I \times I \to I \cup \{\mathbf{overflow}, \mathbf{pole}, \mathbf{invalid}\}$$

$$\begin{array}{lll}
group_I(x, y) & = result_I(\lceil x/y \rceil) & \text{if } x, y \in I \text{ and } y \neq 0 \\
 & = \mathbf{pole}(\mathbf{+\infty}) & \text{if } x \in I \text{ and } x > 0 \text{ and } y = 0 \\
 & = \mathbf{invalid}(\mathbf{qNaN}) & \text{if } x = 0 \text{ and } y = 0 \\
 & = \mathbf{pole}(\mathbf{-\infty}) & \text{if } x \in I \text{ and } x < 0 \text{ and } y = 0
\end{array}$$

$$pad_I : I \times I \to I \cup \{\mathbf{invalid}\}$$

$$\begin{array}{lll}
pad_I(x, y) & = (\lceil x/y \rceil \cdot y) - x & \text{if } x, y \in I \text{ and } y \neq 0 \\
 & = \mathbf{invalid}(\mathbf{qNaN}) & \text{if } x \in I \text{ and } y = 0
\end{array}$$

$$quot_I : I \times I \to I \cup \{\mathbf{overflow}, \mathbf{pole}, \mathbf{invalid}\}$$

$$\begin{array}{lll}
quot_I(x, y) & = result_I(\mathrm{round}(x/y)) & \text{if } x, y \in I \text{ and } y \neq 0 \\
 & = \mathbf{pole}(\mathbf{+\infty}) & \text{if } x \in I \text{ and } x > 0 \text{ and } y = 0 \\
 & = \mathbf{invalid}(\mathbf{qNaN}) & \text{if } x = 0 \text{ and } y = 0 \\
 & = \mathbf{pole}(\mathbf{-\infty}) & \text{if } x \in I \text{ and } x < 0 \text{ and } y = 0
\end{array}$$

$$remr_I : I \times I \to I \cup \{\mathbf{overflow}, \mathbf{invalid}\}$$

$$\begin{array}{lll}
remr_I(x, y) & = result_I(x - (\mathrm{round}(x/y) \cdot y)) & \\
 & & \text{if } x, y \in I \text{ and } y \neq 0 \\
 & = \mathbf{invalid}(\mathbf{qNaN}) & \text{if } x \in I \text{ and } y = 0
\end{array}$$

5.1.8 Greatest common divisor and least common positive multiple

$$gcd_I : I \times I \to I \cup \{\mathbf{overflow}, \mathbf{pole}\}$$

$$\begin{array}{lll}
gcd_I(x, y) & = result_I(\max\{v \in \mathcal{Z} \mid v|x \text{ and } v|y\}) & \\
 & & \text{if } x, y \in I \text{ and } (x \neq 0 \text{ or } y \neq 0) \\
 & = \mathbf{pole}(\mathbf{+\infty}) & \text{if } x = 0 \text{ and } y = 0
\end{array}$$

$$lcm_I : I \times I \to I \cup \{\mathbf{overflow}\}$$

$$\begin{array}{lll}
lcm_I(x, y) & = result_I(\min\{v \in \mathcal{Z} \mid x|v \text{ and } y|v \text{ and } v > 0\}) & \\
 & & \text{if } x, y \in I \text{ and } x \neq 0 \text{ and } y \neq 0 \\
 & = 0 & \text{if } x, y \in I \text{ and } (x = 0 \text{ or } y = 0)
\end{array}$$

$$gcd\_seq_I : [I] \to I \cup \{\mathbf{overflow}, \mathbf{pole}\}$$

$$\begin{array}{lll}
\multicolumn{3}{l}{gcd\_seq_I([x_1, ..., x_n])} \\
 & = result_I(\max\{v \in \mathcal{Z} \mid v|x_i \text{ for all } i \in \{1, ..., n\}\}) & \\
 & & \text{if } \{x_1, ..., x_n\} \subseteq I \text{ and } \{x_1, ..., x_n\} \not\subseteq \{0\} \\
 & = \mathbf{pole}(\mathbf{+\infty}) & \text{if } \{x_1, ..., x_n\} \subseteq \{0\}
\end{array}$$

$$lcm\_seq_I : [I] \to I \cup \{\mathbf{overflow}\}$$

$$\begin{array}{lll}
\multicolumn{3}{l}{lcm\_seq_I([x_1, ..., x_n])} \\
 & = result_I(\min\{v \in \mathcal{Z} \mid x_i|v \text{ for all } i \in \{1, ..., n\} \text{ and } v > 0\}) & \\
 & & \text{if } \{x_1, ..., x_n\} \subseteq I \text{ and } 0 \notin \{x_1, ..., x_n\} \\
 & = 0 & \text{if } \{x_1, ..., x_n\} \subseteq I \text{ and } 0 \in \{x_1, ..., x_n\}
\end{array}$$

NOTE – This specification implies that $lcm\_seq_I([]) = 1$.

5.1.9 Support operations for extended integer range

These operations can be used to implement extended range integer datatypes, including unbounded integer datatypes.

$$add\_wrap_I : I \times I \to I$$

$$add\_wrap_I(x, y) = wrap_I(x + y) \qquad \text{if } x, y \in I$$

$$add\_ov_I : I \times I \to \{-1, 0, 1\}$$

$$\begin{aligned} add\_ov_I(x, y) &= ((x + y) - add\_wrap_I(x, y))/(maxint_I - minint_I + 1) \\ & \qquad \text{if } x, y \in I \text{ and } bounded_I = \mathbf{true} \\ &= 0 \qquad \text{if } x, y \in I \text{ and } bounded_I = \mathbf{false} \end{aligned}$$

$$sub\_wrap_I : I \times I \to I$$

$$sub\_wrap_I(x, y) = wrap_I(x - y) \qquad \text{if } x, y \in I$$

$$sub\_ov_I : I \times I \to \{-1, 0, 1\}$$

$$\begin{aligned} sub\_ov_I(x, y) &= ((x - y) - sub\_wrap_I(x, y))/(maxint_I - minint_I + 1) \\ & \qquad \text{if } x, y \in I \text{ and } bounded_I = \mathbf{true} \\ &= 0 \qquad \text{if } x, y \in I \text{ and } bounded_I = \mathbf{false} \end{aligned}$$

$$mul\_wrap_I : I \times I \to I$$

$$mul\_wrap_I(x, y) = wrap_I(x \cdot y) \qquad \text{if } x, y \in I$$

$$mul\_ov_I : I \times I \to I$$

$$\begin{aligned} mul\_ov_I(x, y) &= ((x \cdot y) - mul\_wrap_I(x, y))/(maxint_I - minint_I + 1) \\ & \qquad \text{if } x, y \in I \text{ and } bounded_I = \mathbf{true} \\ &= 0 \qquad \text{if } x, y \in I \text{ and } bounded_I = \mathbf{false} \end{aligned}$$

NOTE – The add_ov_I and sub_ov_I will only return -1 (for negative overflow), 0 (no overflow), and 1 (for positive overflow).

5.2 Basic floating point operations

Clause 5.2 of Part 1 specifies floating point datatypes and a number of operations on values of a floating point datatype. In this clause some additional operations on values of a floating point datatype are specified.

NOTE – Further operations on values of a floating point datatype, for elementary floating point numerical functions, are specified in clause 5.3.

F is the non-special value set, $F \subset \mathcal{R}$, for a floating point datatype conforming to Part 1. Floating point datatypes conforming to Part 1 often do contain $\mathbf{-0}$, infinity, and **NaN** values. Therefore, in this clause there are specifications for such values as arguments.

5.2.1 The rounding and floating point *result* helper functions

Floating point rounding helper functions: The floating point helper function

$$down_F : \mathcal{R} \rightarrow F^*$$

is the rounding function that rounds towards negative infinity. The floating point helper function

$$up_F : \mathcal{R} \rightarrow F^*$$

is the rounding function that rounds towards positive infinity. The floating point helper function

$$nearest_F : \mathcal{R} \rightarrow F^*$$

is the rounding function that rounds to nearest. $nearest_F$ is partially implementation defined: the handling of ties is implementation defined, but must be sign symmetric. If $iec\_559_F = \mathbf{true}$, the semantics of $nearest_F$ is completely defined by IEC 60559: in this case ties are rounded to even last digit.

$result_F$ is a helper function that is partially implementation defined.

$$result_F : \mathcal{R} \times (\mathcal{R} \rightarrow F^*) \rightarrow F \cup \{\mathbf{underflow}, \mathbf{overflow}\}$$

$$\begin{array}{lll}
result_F(x, nearest_F) & = \mathbf{overflow}(+\infty) & \text{if } x \in \mathcal{R} \text{ and } nearest_F(x) > fmax_F \\
result_F(x, nearest_F) & = \mathbf{overflow}(-\infty) & \text{if } x \in \mathcal{R} \text{ and } nearest_F(x) < -fmax_F \\
result_F(x, up_F) & = \mathbf{overflow}(+\infty) & \text{if } x \in \mathcal{R} \text{ and } up_F(x) > fmax_F \\
result_F(x, up_F) & = \mathbf{overflow}(-fmax_F) & \text{if } x \in \mathcal{R} \text{ and } up_F(x) < -fmax_F \\
result_F(x, down_F) & = \mathbf{overflow}(fmax_F) & \text{if } x \in \mathcal{R} \text{ and } down_F(x) > fmax_F \\
result_F(x, down_F) & = \mathbf{overflow}(-\infty) & \text{if } x \in \mathcal{R} \text{ and } down_F(x) < -fmax_F
\end{array}$$

otherwise:

$$\begin{array}{lll}
result_F(x, rnd) & = x & \text{if } x = 0 \\
 & = rnd(x) & \text{if } x \in \mathcal{R} \text{ and } fminN_F \leqslant |x| \text{ and } |rnd(x)| \leqslant fmax_F \\
 & = rnd(x) \text{ or } \mathbf{underflow}(c) & \\
 & & \text{if } x \in \mathcal{R} \text{ and } |x| < fminN_F \text{ and } |rnd(x)| = fminN_F \\
 & & \quad \text{and } rnd \text{ has no denormalisation loss at } x \\
 & = rnd(x) \text{ or } \mathbf{underflow}(c) & \\
 & & \text{if } x \in \mathcal{R} \text{ and } denorm_F = \mathbf{true} \text{ and} \\
 & & \quad |rnd(x)| < fminN_F \text{ and } x \neq 0 \\
 & & \quad \text{and } rnd \text{ has no denormalisation loss at } x \\
 & = \mathbf{underflow}(c) & \text{otherwise}
\end{array}$$

where

$$\begin{array}{ll}
c = rnd(x) & \text{when } denorm_F = \mathbf{true} \text{ and } (rnd(x) \neq 0 \text{ or } x > 0), \\
c = \mathbf{-0} & \text{when } denorm_F = \mathbf{true} \text{ and } rnd(x) = 0 \text{ and } x < 0, \\
c = 0 & \text{when } denorm_F = \mathbf{false} \text{ and } x > 0, \\
c = \mathbf{-0} & \text{when } denorm_F = \mathbf{false} \text{ and } x < 0
\end{array}$$

An implementation is allowed to choose between $rnd(x)$ and $\mathbf{underflow}(rnd(x))$ in the region between 0 and $fminN_F$. However, a subnormal value without underflow notification can be chosen only if $denorm_F$ is true and no denormalisation loss occurs at x.

NOTES

1 This differs from the specification of $result_F$ as given in Part 1 in the following respects: 1) the continuation values on overflow and underflow are given directly here, and 2) all instances of denormalisation loss must be accompanied with an underflow notification.

2 $denorm_F = \mathbf{false}$ implies $iec\_559_F = \mathbf{false}$, and $iec\_559_F = \mathbf{true}$ implies $denorm_F = \mathbf{true}$.

3 If $iec\_559_F = \mathbf{true}$, then subnormal results that have no denormalisation loss, e.g. are exact, do not result in an underflow notification, if the notification is by recording of indicators.

Define the $result\_NaN_F$, $result\_NaN2_F$, and $result\_NaN3_F$ helper functions:

$$result\_NaN_F : F \rightarrow \{\mathbf{invalid}\}$$

$$\begin{array}{lll} result\_NaN_F(x) & = \mathbf{qNaN} & \text{if } x \text{ is a quiet NaN} \\ & = \mathbf{invalid}(\mathbf{qNaN}) & \text{otherwise} \end{array}$$

$$result\_NaN2_F : F \times F \rightarrow \{\mathbf{invalid}\}$$

$$\begin{array}{lll} result\_NaN2_F(x, y) & & \\ & = \mathbf{qNaN} & \text{if } x \text{ is a quiet NaN and } y \text{ is not a signalling NaN} \\ & = \mathbf{qNaN} & \text{if } y \text{ is a quiet NaN and } x \text{ is not a signalling NaN} \\ & = \mathbf{invalid}(\mathbf{qNaN}) & \text{otherwise} \end{array}$$

$$result\_NaN3_F : F \times F \times F \rightarrow \{\mathbf{invalid}\}$$

$$\begin{array}{lll} result\_NaN3_F(x, y, z) & & \\ & = \mathbf{qNaN} & \text{if } x \text{ is a quiet NaN and} \\ & & \quad \text{not } y \text{ nor } z \text{ is a signalling NaN} \\ & = \mathbf{qNaN} & \text{if } y \text{ is a quiet NaN and} \\ & & \quad \text{not } x \text{ nor } z \text{ is a signalling NaN} \\ & = \mathbf{qNaN} & \text{if } z \text{ is a quiet NaN and} \\ & & \quad \text{not } x \text{ nor } y \text{ is a signalling NaN} \\ & = \mathbf{invalid}(\mathbf{qNaN}) & \text{otherwise} \end{array}$$

These helper functions are used to specify both NaN argument handling and to handle non-NaN-argument cases where **invalid**(**qNaN**) is the appropriate result.

5.2.2 Floating point maximum and minimum

The appropriate return value of the maximum and minimum operations given a quiet **NaN** (**qNaN**) as one of the input values depends on the circumstances for each point of use. Sometimes **qNaN** is the appropriate result, sometimes the non-**NaN** argument is the appropriate result. Therefore, two variants each of the floating point maximum and minimum operations are specified here, and the programmer can decide which one is appropriate to use at each particular place of usage, assuming both variants are included in the binding.

$$max_F : F \times F \rightarrow F$$

$$\begin{array}{lll} max_F(x, y) & = \max\{x, y\} & \text{if } x, y \in F \\ & = +\infty & \text{if } x = +\infty \text{ and } y \in F \cup \{-\infty, \mathbf{-0}\} \\ & = y & \text{if } x = \mathbf{-0} \text{ and } y \in F \text{ and } y \geqslant 0 \\ & = \mathbf{-0} & \text{if } x = \mathbf{-0} \text{ and } ((y \in F \text{ and } y < 0) \text{ or } y = \mathbf{-0}) \\ & = y & \text{if } x = -\infty \text{ and } y \in F \cup \{+\infty, \mathbf{-0}\} \\ & = +\infty & \text{if } y = +\infty \text{ and } x \in F \cup \{+\infty, \mathbf{-0}\} \\ & = x & \text{if } y = \mathbf{-0} \text{ and } x \in F \text{ and } x \geqslant 0 \\ & = \mathbf{-0} & \text{if } y = \mathbf{-0} \text{ and } x \in F \text{ and } x < 0 \end{array}$$

$$
\begin{array}{lll}
 & = x & \text{if } y = \mathbf{-\infty} \text{ and } x \in F \cup \{\mathbf{-\infty}, \mathbf{-0}\} \\
 & = \mathit{result\_NaN2}_F(x, y) & \text{otherwise}
\end{array}
$$

$\mathit{min}_F : F \times F \to F$

$$
\begin{array}{lll}
\mathit{min}_F(x, y) & = \min\{x, y\} & \text{if } x, y \in F \\
 & = y & \text{if } x = \mathbf{+\infty} \text{ and } y \in F \cup \{\mathbf{-\infty}, \mathbf{-0}\} \\
 & = \mathbf{-0} & \text{if } x = \mathbf{-0} \text{ and } y \in F \text{ and } y \geqslant 0 \\
 & = y & \text{if } x = \mathbf{-0} \text{ and } ((y \in F \text{ and } y < 0) \text{ or } y = \mathbf{-0}) \\
 & = \mathbf{-\infty} & \text{if } x = \mathbf{-\infty} \text{ and } y \in F \cup \{\mathbf{+\infty}, \mathbf{-0}\} \\
 & = x & \text{if } y = \mathbf{+\infty} \text{ and } x \in F \cup \{\mathbf{+\infty}, \mathbf{-0}\} \\
 & = \mathbf{-0} & \text{if } y = \mathbf{-0} \text{ and } x \in F \text{ and } x \geqslant 0 \\
 & = x & \text{if } y = \mathbf{-0} \text{ and } x \in F \text{ and } x < 0 \\
 & = \mathbf{-\infty} & \text{if } y = \mathbf{-\infty} \text{ and } x \in F \cup \{\mathbf{-\infty}, \mathbf{-0}\} \\
 & = \mathit{result\_NaN2}_F(x, y) & \text{otherwise}
\end{array}
$$

$\mathit{mmax}_F : F \times F \to F$

$$
\begin{array}{lll}
\mathit{mmax}_F(x, y) & = \mathit{max}_F(x, y) & \text{if } x, y \in F \cup \{\mathbf{+\infty}, \mathbf{-0}, \mathbf{-\infty}\} \\
 & = x & \text{if } x \in F \cup \{\mathbf{+\infty}, \mathbf{-0}, \mathbf{-\infty}\} \text{ and } y \text{ is a quiet NaN} \\
 & = y & \text{if } y \in F \cup \{\mathbf{+\infty}, \mathbf{-0}, \mathbf{-\infty}\} \text{ and } x \text{ is a quiet NaN} \\
 & = \mathit{result\_NaN2}_F(x, y) & \text{otherwise}
\end{array}
$$

$\mathit{mmin}_F : F \times F \to F$

$$
\begin{array}{lll}
\mathit{mmin}_F(x, y) & = \mathit{min}_F(x, y) & \text{if } x, y \in F \cup \{\mathbf{+\infty}, \mathbf{-0}, \mathbf{-\infty}\} \\
 & = x & \text{if } x \in F \cup \{\mathbf{+\infty}, \mathbf{-0}, \mathbf{-\infty}\} \text{ and } y \text{ is a quiet NaN} \\
 & = y & \text{if } y \in F \cup \{\mathbf{+\infty}, \mathbf{-0}, \mathbf{-\infty}\} \text{ and } x \text{ is a quiet NaN} \\
 & = \mathit{result\_NaN2}_F(x, y) & \text{otherwise}
\end{array}
$$

NOTE – If one of the arguments to mmax_F or mmin_F is a quiet NaN, that argument is ignored.

$\mathit{max\_seq}_F : [F] \to F \cup \{\mathbf{-\infty}, \mathbf{pole}\}$

$$
\begin{array}{lll}
\multicolumn{3}{l}{\mathit{max\_seq}_F([x_1, ..., x_n])} \\
 & = \mathbf{-\infty} & \text{if } n = 0 \text{ and } \mathbf{-\infty} \text{ is available} \\
 & = \mathbf{pole}(-\mathit{fmax}_F) & \text{if } n = 0 \text{ and } \mathbf{-\infty} \text{ is not available} \\
 & = \mathit{max}_F(\mathit{max\_seq}_F([x_1, ..., x_{n-1}]), x_n) & \\
 & & \text{if } n \geqslant 2 \\
 & = x_1 & \text{if } n = 1 \text{ and } x_1 \text{ is not a NaN} \\
 & = \mathit{result\_NaN}_F(x_1) & \text{otherwise}
\end{array}
$$

$\mathit{min\_seq}_F : [F] \to F \cup \{\mathbf{+\infty}, \mathbf{pole}\}$

$$
\begin{array}{lll}
\multicolumn{3}{l}{\mathit{min\_seq}_F([x_1, ..., x_n])} \\
 & = \mathbf{+\infty} & \text{if } n = 0 \text{ and } \mathbf{+\infty} \text{ is available} \\
 & = \mathbf{pole}(\mathit{fmax}_F) & \text{if } n = 0 \text{ and } \mathbf{+\infty} \text{ is not available} \\
 & = \mathit{min}_F(\mathit{min\_seq}_F([x_1, ..., x_{n-1}]), x_n) & \\
 & & \text{if } n \geqslant 2 \\
 & = x_1 & \text{if } n = 1 \text{ and } x_1 \text{ is not a NaN} \\
 & = \mathit{result\_NaN}_F(x_1) & \text{otherwise}
\end{array}
$$

$$mmax\_seq_F : [F] \rightarrow F \cup \{\mathbf{-\infty}, \mathbf{pole}\}$$

$$\begin{array}{lll}
mmax\_seq_F([x_1, ..., x_n]) & & \\
& = \mathbf{-\infty} & \text{if } n = 0 \text{ and } \mathbf{-\infty} \text{ is available} \\
& = \mathbf{pole}(-fmax_F) & \text{if } n = 0 \text{ and } \mathbf{-\infty} \text{ is not available} \\
& = mmax_F(mmax\_seq_F([x_1, ..., x_{n-1}]), x_n) & \\
& & \text{if } n \geqslant 2 \\
& = x_1 & \text{if } n = 1 \text{ and } x_1 \text{ is not a NaN} \\
& = result\_NaN_F(x_1) & \text{otherwise}
\end{array}$$

$$mmin\_seq_F : [F] \rightarrow F \cup \{\mathbf{+\infty}, \mathbf{pole}\}$$

$$\begin{array}{lll}
mmin\_seq_F([x_1, ..., x_n]) & & \\
& = \mathbf{+\infty} & \text{if } n = 0 \text{ and } \mathbf{+\infty} \text{ is available} \\
& = \mathbf{pole}(fmax_F) & \text{if } n = 0 \text{ and } \mathbf{+\infty} \text{ is not available} \\
& = mmin_F(mmin\_seq_F([x_1, ..., x_{n-1}]), x_n) & \\
& & \text{if } n \geqslant 2 \\
& = x_1 & \text{if } n = 1 \text{ and } x_1 \text{ is not a NaN} \\
& = result\_NaN_F(x_1) & \text{otherwise}
\end{array}$$

5.2.3 Floating point diminish

$$dim_F : F \times F \rightarrow F \cup \{\mathbf{overflow}, \mathbf{underflow}\}$$

$$\begin{array}{lll}
dim_F(x, y) & = result_F(\max\{0, x - y)\}, rnd_F) & \\
& & \text{if } x, y \in F \\
& = dim_F(0, y) & \text{if } x = \mathbf{-0} \text{ and } y \in F \cup \{\mathbf{-\infty}, \mathbf{-0}, \mathbf{+\infty}\} \\
& = dim_F(x, 0) & \text{if } y = \mathbf{-0} \text{ and } x \in F \cup \{\mathbf{-\infty}, \mathbf{+\infty}\} \\
& = \mathbf{+\infty} & \text{if } x = \mathbf{+\infty} \text{ and } y \in F \cup \{\mathbf{-\infty}\} \\
& = 0 & \text{if } x = \mathbf{-\infty} \text{ and } y \in F \cup \{\mathbf{+\infty}\} \\
& = 0 & \text{if } y = \mathbf{+\infty} \text{ and } x \in F \\
& = \mathbf{+\infty} & \text{if } y = \mathbf{-\infty} \text{ and } x \in F \\
& = result\_NaN2_F(x, y) & \text{otherwise}
\end{array}$$

NOTE – dim_F cannot be implemented by $max_F(0, sub_F(x, y))$, since this latter expression has other overflow properties.

5.2.4 Round, floor, and ceiling

$$rounding_F : F \rightarrow F \cup \{\mathbf{-0}\}$$

$$\begin{array}{lll}
rounding_F(x) & = \mathrm{round}(x) & \text{if } x \in F \text{ and } (x \geqslant 0 \text{ or } \mathrm{round}(x) \neq 0) \\
& = \mathbf{-0} & \text{if } x \in F \text{ and } x < 0 \text{ and } \mathrm{round}(x) = 0 \\
& = x & \text{if } x \in \{\mathbf{-\infty}, \mathbf{-0}, \mathbf{+\infty}, \} \\
& = result\_NaN_F(x) & \text{otherwise}
\end{array}$$

$$floor_F : F \rightarrow F$$

$$\begin{array}{lll}
floor_F(x) & = \lfloor x \rfloor & \text{if } x \in F \\
& = x & \text{if } x \in \{\mathbf{-\infty}, \mathbf{-0}, \mathbf{+\infty}, \} \\
& = result\_NaN_F(x) & \text{otherwise}
\end{array}$$

$ceiling_F : F \to F \cup \{\mathbf{-0}\}$

$$\begin{array}{lll}
ceiling_F(x) & = \lceil x \rceil & \text{if } x \in F \text{ and } (x \geqslant 0 \text{ or } \lceil x \rceil \neq 0) \\
 & = \mathbf{-0} & \text{if } x \in F \text{ and } x < 0 \text{ and } \lceil x \rceil = 0 \\
 & = x & \text{if } x \in \{\mathbf{-\infty}, \mathbf{-0}, \mathbf{+\infty}, \} \\
 & = result\_NaN_F(x) & \text{otherwise}
\end{array}$$

NOTE 1 – Truncate to integer is specified in Part 1, by the name $intpart_F$.

$rounding\_rest_F : F \to F$

$$\begin{array}{lll}
rounding\_rest_F(x) & & \\
 & = x - \mathrm{round}(x) & \text{if } x \in F \\
 & = 0 & \text{if } x = \mathbf{-0} \\
 & = result\_NaN_F(x) & \text{otherwise}
\end{array}$$

$floor\_rest_F : F \to F$

$$\begin{array}{lll}
floor\_rest_F(x) & = result_F(x - \lfloor x \rfloor, rnd_F) & \text{if } x \in F \\
 & = 0 & \text{if } x = \mathbf{-0} \\
 & = result\_NaN_F(x) & \text{otherwise}
\end{array}$$

$ceiling\_rest_F : F \to F$

$$\begin{array}{lll}
ceiling\_rest_F(x) & & \\
 & = result_F(x - \lceil x \rceil, rnd_F) & \text{if } x \in F \\
 & = 0 & \text{if } x = \mathbf{-0} \\
 & = result\_NaN_F(x) & \text{otherwise}
\end{array}$$

NOTE 2 – The rest after truncation is specified in Part 1, by the name $fractpart_F$.

5.2.5 Remainder after division with round to integer

$remr_F : F \times F \to F \cup \{\mathbf{-0}, \mathbf{underflow}, \mathbf{invalid}\}$

$$\begin{array}{lll}
remr_F(x, y) & = result_F(x - (\mathrm{round}(x/y) \cdot y), nearest_F) & \\
 & & \text{if } x, y \in F \text{ and } y \neq 0 \text{ and} \\
 & & \quad (x \geqslant 0 \text{ or } x - (\mathrm{round}(x/y) \cdot y) \neq 0) \\
 & = \mathbf{-0} & \text{if } x, y \in F \text{ and } y \neq 0 \text{ and} \\
 & & \quad x < 0 \text{ and } x - (\mathrm{round}(x/y) \cdot y) = 0 \\
 & = \mathbf{-0} & \text{if } x = \mathbf{-0} \text{ and } y \in F \cup \{\mathbf{-\infty}, \mathbf{+\infty}\} \text{ and } y \neq 0 \\
 & = x & \text{if } x \in F \text{ and } y \in \{\mathbf{-\infty}, \mathbf{+\infty}\} \\
 & = result\_NaN2_F(x, y) & \text{otherwise}
\end{array}$$

5.2.6 Square root and reciprocal square root

$sqrt_F : F \to F \cup \{\mathbf{invalid}\}$

$$\begin{array}{lll}
sqrt_F(x) & = nearest_F(\sqrt{x}) & \text{if } x \in F \text{ and } x \geqslant 0 \\
 & = x & \text{if } x \in \{\mathbf{-0}, \mathbf{+\infty}\} \\
 & = result\_NaN_F(x) & \text{otherwise}
\end{array}$$

$$\mathit{rec\_sqrt}_F : F \to F \cup \{\mathbf{invalid}, \mathbf{pole}\}$$

$$\begin{array}{lll}
\mathit{rec\_sqrt}_F(x) & = \mathit{rnd}_F(1/\sqrt{x}) & \text{if } x \in F \text{ and } x > 0 \\
 & = \mathbf{pole}(\mathbf{+\infty}) & \text{if } x \in \{\mathbf{-0}, 0\} \\
 & = 0 & \text{if } x = \mathbf{+\infty} \\
 & = \mathit{result\_NaN}_F(x) & \text{otherwise}
\end{array}$$

5.2.7 Support operations for extended floating point precision

These operations are useful when keeping guard digits or implementing extra precision floating point datatypes. The resulting datatypes, e.g. so-called doubled precision, do not necessarily conform to Part 1.

$$\mathit{add\_lo}_F : F \times F \to F \cup \{\mathbf{underflow}\}$$

$$\begin{array}{lll}
\mathit{add\_lo}_F(x, y) & = \mathit{result}_F((x + y) - \mathit{rnd}_F(x + y), \mathit{rnd}_F) & \\
 & & \text{if } x, y \in F \\
 & = \mathbf{-0} & \text{if } x = \mathbf{-0} \text{ and } y \in F \cup \{\mathbf{-\infty}, \mathbf{-0}, \mathbf{+\infty}\} \\
 & = \mathbf{-0} & \text{if } x \in F \cup \{\mathbf{-\infty}, \mathbf{+\infty}\} \text{ and } y = \mathbf{-0} \\
 & = y & \text{if } x = \mathbf{+\infty} \text{ and } y \in F \cup \{\mathbf{+\infty}\} \\
 & = y & \text{if } x = \mathbf{-\infty} \text{ and } y \in F \cup \{\mathbf{-\infty}\} \\
 & = x & \text{if } x \in F \text{ and } y \in \{\mathbf{-\infty}, \mathbf{+\infty}\} \\
 & = \mathit{result\_NaN2}_F(x, y) & \text{otherwise}
\end{array}$$

$$\mathit{sub\_lo}_F : F \times F \to F \cup \{\mathbf{underflow}\}$$

$$\mathit{sub\_lo}_F(x, y) \quad = \mathit{add\_lo}_F(x, \mathit{neg}_F(y))$$

NOTE 1 – If $\mathit{rnd\_style}_F = \mathit{nearest}$, then, in the absence of notifications, $\mathit{add\_lo}_F$ and $\mathit{sub\_lo}_F$ returns exact results.

$$\mathit{mul\_lo}_F : F \times F \to F \cup \{\mathbf{overflow}, \mathbf{underflow}\}$$

$$\begin{array}{lll}
\mathit{mul\_lo}_F(x, y) & = \mathit{result}_F((x \cdot y) - \mathit{rnd}_F(x \cdot y), \mathit{rnd}_F) & \\
 & & \text{if } x, y \in F \\
 & = \mathit{mul\_lo}_F(0, y) & \text{if } x = \mathbf{-0} \text{ and } y \in F \cup \{\mathbf{-\infty}, \mathbf{-0}, \mathbf{+\infty}\} \\
 & = \mathit{mul\_lo}_F(x, 0) & \text{if } x \in F \cup \{\mathbf{-\infty}, \mathbf{+\infty}\} \text{ and } y = \mathbf{-0} \\
 & = \mathit{mul}_F(x, y) & \text{if } x \in \{\mathbf{-\infty}, \mathbf{+\infty}\} \text{ and } y \in F \cup \{\mathbf{-\infty}, \mathbf{+\infty}\} \\
 & = \mathit{mul}_F(x, y) & \text{if } x \in F \text{ and } y \in \{\mathbf{-\infty}, \mathbf{+\infty}\} \\
 & = \mathit{result\_NaN2}_F(x, y) & \text{otherwise}
\end{array}$$

NOTE 2 – In the absence of notifications, $\mathit{mul\_lo}_F$ returns an exact result.

$$\mathit{div\_rest}_F : F \times F \to F \cup \{\mathbf{underflow}, \mathbf{invalid}\}$$

$$\begin{array}{lll}
\mathit{div\_rest}_F(x, y) & = \mathit{result}_F(x - (y \cdot \mathit{rnd}_F(x/y)), \mathit{rnd}_F) & \\
 & & \text{if } x, y \in F \\
 & = \mathit{div\_rest}_F(0, y) & \text{if } x = \mathbf{-0} \text{ and } y \in F \cup \{\mathbf{-\infty}, \mathbf{-0}, \mathbf{+\infty}\} \\
 & = x & \text{if } x \in F \text{ and } y \in \{\mathbf{-\infty}, \mathbf{+\infty}\} \\
 & = x & \text{if } x \in \{\mathbf{-\infty}, \mathbf{+\infty}\} \text{ and } y \in F \\
 & = \mathit{result\_NaN2}_F(x, y) & \text{otherwise}
\end{array}$$

$sqrt\_rest_F : F \to F \cup \{\textbf{underflow}, \textbf{invalid}\}$

$$
\begin{array}{lll}
sqrt\_rest_F(x) & = result_F(x - (sqrt_F(x) \cdot sqrt_F(x)), rnd_F) & \\
 & & \text{if } x \in F \text{ and } x \geqslant 0 \\
 & = \mathbf{-0} & \text{if } x = \mathbf{-0} \\
 & = \mathbf{+\infty} & \text{if } x = \mathbf{+\infty} \\
 & = result\_NaN_F(x) & \text{otherwise}
\end{array}
$$

NOTE 3 – $sqrt\_rest_F(x)$ is exact when there is no **underflow**.

For the following operation F' is a floating point type conforming to Part 1.

NOTE 4 – It is expected that $p_{F'} > p_F$, i.e. F' has higher precision than F, but that is not required.

$mul_{F \to F'} : F \times F \to F' \cup \{\mathbf{-0}, \textbf{overflow}, \textbf{underflow}\}$

$$
\begin{array}{lll}
mul_{F \to F'}(x, y) & = result_{F'}(x \cdot y, rnd_{F'}) & \text{if } x, y \in F \text{ and } x \neq 0 \text{ and } y \neq 0 \\
 & = convert_{F \to F'}(mul_F(x, y)) & \\
 & & \text{if } x \in \{\mathbf{-\infty}, \mathbf{-0}, 0, \mathbf{+\infty}\} \text{ and} \\
 & & \quad y \in F \cup \{\mathbf{-\infty}, \mathbf{-0}, \mathbf{+\infty}\} \\
 & = convert_{F \to F'}(mul_F(x, y)) & \\
 & & \text{if } y \in \{\mathbf{-\infty}, \mathbf{-0}, 0, \mathbf{+\infty}\} \text{ and } x \in F \text{ and } x \neq 0 \\
 & = result\_NaN2_{F'}(x, y) & \text{otherwise}
\end{array}
$$

5.3 Elementary transcendental floating point operations

5.3.1 Maximum error requirements

The specifications for each of the transcendental and floating point conversion operations use an approximation helper function. The approximation helper functions are ideally identical to the true mathematical functions. However, that would imply a maximum error for the corresponding operation of 0.5 ulp (i.e., the minimum value for operations that are not always exact). This Part does not require that the maximum error is only 0.5 ulp for the operations specified in clauses 5.3, 5.4, and 5.5, but allows the maximum error to be a bit bigger. To express this, the approximation helper functions need not be identical to the mathematical elementary transcendental functions, but are allowed to be approximate.

The approximation helper functions for the individual operations in these subclauses have maximum error parameters that describe the maximum *relative* error of the helper function composed with $nearest_F$, for non-subnormal results. The maximum error parameters also describe the maximum *absolute* error for subnormal results and underflow continuation values if $denorm_F = \mathbf{true}$. The relevant maximum error parameters shall be available to programs.

When the maximum error for an approximation helper function h_F, approximating f, is $max\_error\_op_F$, then for all arguments $x, ... \in F^* \times ...$ the following equation shall hold:

$$|f(x, ...) - nearest_F(h_F(x, ...))| \leqslant max\_error\_op_F \cdot r_F^{e_F(f(x,...))-p_F}$$

NOTES

1 Partially conforming implementations may have greater values for maximum error parameters than stipulated below. See annex A.

2 For most positive (and not too small) return values t, the true result is thus claimed to be in the interval $[t - (max\_error\_op_F \cdot ulp_F(t)), t + (max\_error\_op_F \cdot ulp_F(t))]$. But if the return value is exactly r_F^n for some not too small $n \in \mathcal{Z}$, then the true result is claimed to be in the interval $[t - (max\_error\_op_F \cdot ulp_F(t)/r_F), t + (max\_error\_op_F \cdot ulp_F(t))]$. Similarly for negative return values.

The results of the approximating helper functions in this clause must be exact for certain arguments as detailed below, and may be exact for all arguments. If the approximating helper function is exact for all arguments, then the corresponding maximum error parameter should be 0.5, the minimum value.

5.3.2 Sign requirements

The approximation helper functions are shall be zero exactly at the points where the approximated mathematical function is exactly zero. At points where the approximation helper functions are not zero, they are shall have the same sign as the approximated mathematical function at that point.

For the radian trigonometric helper functions, these zero and sign requirements are imposed only for arguments, x, such that $|x| \leqslant big\_angle\_r_F$ (see clause 5.3.9).

NOTE – For the operations, the continuation value after an **underflow** may be zero (or negative zero) as given by $trans\_result_F$, even though the approximation helper function is not zero at that point. Such zero results are required to be accompanied by an **underflow** notification. When appropriate, zero may also be returned for IEC 60559 infinities arguments. See the individual specifications.

5.3.3 Monotonicity requirements

Each approximation helper function in this clause shall be a monotonic approximation to the mathematical function it is approximating, except:

a) For the radian trigonometric approximation helper functions, the monotonic approximation requirement is imposed only for arguments, x, such that $|x| \leqslant \mathit{big\_angle\_r}_F$ (see clause 5.3.9).

b) The argument angular unit trigonometric and argument angular unit inverse trigonometric approximating helper functions are excepted from the monotonic approximation requirement for the angular unit argument.

5.3.4 The *trans_result* helper function

The $\mathit{trans\_result}_F$ helper function is similar to the result_F helper function (see 5.2.1), but is simplified compared to result_F concerning **underflow**: $\mathit{trans\_result}_F$ always underflows for non-zero arguments that have an absolute value less than $\mathit{fminN}_F - (\mathit{fminD}_F / r_F)$, whereas result_F does not necessarily underflow in that case. This difference from result_F is made since the argument to $\mathit{trans\_result}_F$ might not be exact. To return **underflow** or not, for a tiny result, based upon an inexact result would be misleading. For the operations specified using $\mathit{trans\_result}_F$ where the specification implies that there will be no denormalisation loss for certain tiny results, **underflow** is instead explicitly avoided.

$$\mathit{trans\_result}_F : \mathcal{R} \times (\mathcal{R} \to F^*) \to F \cup \{\mathbf{underflow}, \mathbf{overflow}\}$$

$$\begin{aligned}
&\mathit{trans\_result}_F(x, \mathit{rnd}) \\
&\qquad = \mathbf{underflow}(c) && \text{if } x \in \mathcal{R} \text{ and } \mathit{denorm}_F = \mathbf{true} \text{ and} \\
&&& \quad |\mathit{rnd}(x)| < \mathit{fminN}_F \text{ and } x \neq 0 \\
&\qquad = \mathit{result}_F(x, \mathit{rnd}) && \text{otherwise}
\end{aligned}$$

where

$$\begin{aligned}
&c = \mathit{rnd}(x) && \text{when } \mathit{rnd}(x) \neq 0 \text{ or } x > 0, \\
&c = \mathbf{-0} && \text{when } \mathit{rnd}(x) = 0 \text{ and } x < 0
\end{aligned}$$

5.3.5 Hypotenuse

There shall be a maximum error parameter for the hypot_F operation:

$$\mathit{max\_error\_hypot}_F \in F$$

The $\mathit{max\_error\_hypot}_F$ parameter shall have a value in the interval $[0.5, 1]$.

The hypot^*_F approximation helper function:

$$\mathit{hypot}^*_F : F \times F \to \mathcal{R}$$

$\mathit{hypot}^*_F(x, y)$ returns a close approximation to $\sqrt{x^2 + y^2}$ in $\mathcal{R}$, with maximum error $\mathit{max\_error\_hypot}_F$.

Further requirements on the hypot^*_F approximation helper function are:

$$\begin{aligned}
&\mathit{hypot}^*_F(x, y) = \mathit{hypot}^*_F(y, x) \\
&\mathit{hypot}^*_F(-x, y) = \mathit{hypot}^*_F(x, y) \\
&\mathit{hypot}^*_F(x, y) \geqslant \max\{|x|, |y|\} \\
&\mathit{hypot}^*_F(x, y) \leqslant |x| + |y| \\
&\mathit{hypot}^*_F(x, y) \geqslant 1 && \text{if } \sqrt{x^2 + y^2} \geqslant 1 \\
&\mathit{hypot}^*_F(x, y) \leqslant 1 && \text{if } \sqrt{x^2 + y^2} \leqslant 1
\end{aligned}$$

The hypot_F operation:

$$\mathit{hypot}_F : F \times F \to F \cup \{\mathbf{underflow}, \mathbf{overflow}\}$$

$$\begin{array}{lll}
\mathit{hypot}_F(x, y) & = \mathit{trans\_result}_F(\mathit{hypot}^*_F(x, y), \mathit{nearest}_F) & \\
& & \text{if } x, y \in F \\
& = \mathit{hypot}_F(0, y) & \text{if } x = \mathbf{-0} \text{ and } y \in F \cup \{\mathbf{-\infty}, \mathbf{-0}, \mathbf{+\infty}\} \\
& = \mathit{hypot}_F(x, 0) & \text{if } y = \mathbf{-0} \text{ and } x \in F \cup \{\mathbf{-\infty}, \mathbf{+\infty}\} \\
& = \mathbf{+\infty} & \text{if } x \in \{\mathbf{-\infty}, \mathbf{+\infty}\} \text{ and } y \in F \cup \{\mathbf{-\infty}, \mathbf{+\infty}\} \\
& = \mathbf{+\infty} & \text{if } y \in \{\mathbf{-\infty}, \mathbf{+\infty}\} \text{ and } x \in F \\
& = \mathit{result\_NaN2}_F(x, y) & \text{otherwise}
\end{array}$$

5.3.6 Operations for exponentiations and logarithms

There shall be two maximum error parameters for approximate exponentiations and logarithms:

$$\mathit{max\_error\_exp}_F \in F$$
$$\mathit{max\_error\_power}_F \in F$$

The $\mathit{max\_error\_exp}_F$ parameter shall have a value in the interval $[0.5, 1.5 \cdot \mathit{rnd\_error}_F]$. The $\mathit{max\_error\_power}_F$ parameter shall have a value in the interval $[\mathit{max\_error\_exp}_F, 2 \cdot \mathit{rnd\_error}_F]$.

NOTE – The "exp" operations are thus required to be at least as accurate as the "power" operations.

5.3.6.1 Integer power of argument base

The power^*_{FI} approximation helper function:

$$\mathit{power}^*_{FI} : F \times I \to \mathcal{R}$$

$\mathit{power}^*_{FI}(x, y)$ returns a close approximation to x^y in $\mathcal{R}$, with maximum error $\mathit{max\_error\_power}_F$.

Further requirements on the power^*_{FI} approximation helper function are:

$$\begin{array}{ll}
\mathit{power}^*_{FI}(1, y) = 1 & \text{if } y \in I \\
\mathit{power}^*_{FI}(x, 0) = 1 & \text{if } x \in F \text{ and } x \neq 0 \\
\mathit{power}^*_{FI}(x, 1) = x & \text{if } x \in F \\
\mathit{power}^*_{FI}(x, y) < \mathit{fminD}_F/2 & \text{if } x \in F \text{ and } x > 0 \text{ and } y \in I \text{ and } x^y < \mathit{fminD}_F/3 \\
\mathit{power}^*_{FI}(x, y) = \mathit{power}^*_{FI}(-x, y) & \text{if } x \in F \text{ and } x < 0 \text{ and } y \in I \text{ and } 2|y \\
\mathit{power}^*_{FI}(x, y) = -\mathit{power}^*_{FI}(-x, y) & \text{if } x \in F \text{ and } x < 0 \text{ and } y \in I \text{ and not } 2|y
\end{array}$$

The power_{FI} operation:

$$\mathit{power}_{FI} : F \times I \to F \cup \{\mathbf{underflow}, \mathbf{overflow}, \mathbf{pole}\}$$

$$\begin{array}{lll}
\mathit{power}_{FI}(x, y) & = \mathit{trans\_result}_F(\mathit{power}^*_{FI}(x, y), \mathit{nearest}_F) & \\
& & \text{if } x \in F \text{ and } x \neq 0 \text{ and } y \in I \\
& & \\
& = \mathbf{+\infty} & \text{if } x = \mathbf{-\infty} \text{ and } y \in I \text{ and } y > 0 \text{ and } 2|y \\
& = \mathbf{-\infty} & \text{if } x = \mathbf{-\infty} \text{ and } y \in I \text{ and } y > 0 \text{ and not } 2|y \\
& = 0 & \text{if } x = \mathbf{-0} \text{ and } y \in I \text{ and } y > 0 \text{ and } 2|y \\
& = \mathbf{-0} & \text{if } x = \mathbf{-0} \text{ and } y \in I \text{ and } y > 0 \text{ and not } 2|y \\
& = 0 & \text{if } x = 0 \text{ and } y \in I \text{ and } y > 0 \\
& = \mathbf{+\infty} & \text{if } x = \mathbf{+\infty} \text{ and } y \in I \text{ and } y > 0 \\
& & \\
& = 1 & \text{if } x \in \{\mathbf{-\infty}\, \mathbf{-0}, 0, \mathbf{+\infty}\} \text{ and } y = 0 \\
& & \\
& = 0 & \text{if } x = \mathbf{-\infty} \text{ and } y \in I \text{ and } y < 0 \text{ and } 2|y \\
& = \mathbf{-0} & \text{if } x = \mathbf{-\infty} \text{ and } y \in I \text{ and } y < 0 \text{ and not } 2|y
\end{array}$$

$$\begin{array}{ll}
= \mathbf{pole}(+\infty) & \text{if } x = \mathbf{-0} \text{ and } y \in I \text{ and } y < 0 \text{ and } 2|y \\
= \mathbf{pole}(-\infty) & \text{if } x = \mathbf{-0} \text{ and } y \in I \text{ and } y < 0 \text{ and not } 2|y \\
= \mathbf{pole}(+\infty) & \text{if } x = 0 \text{ and } y \in I \text{ and } y < 0 \\
= 0 & \text{if } x = +\infty \text{ and } y \in I \text{ and } y < 0 \\
 & \\
= \mathit{result\_NaN}_F(x) & \text{otherwise}
\end{array}$$

NOTES

1 $\mathit{power}_{FI}(x, y)$ will overflow approximately when $x^y > \mathit{fmax}_F$, i.e., if $x > 1$, approximately when $y > \log_x(\mathit{fmax}_F)$, and if $0 < x < 1$, approximately when $y < \log_x(\mathit{fmax}_F)$ (which is then negative). It will not overflow when $x = 0$ or when $x = 1$.

2 power_I (in clause 5.1.4) does not allow negative *exponents* since the exact result then is not in $\mathcal{Z}$. power_F (in clause 5.3.6.6) does not allow any negative *bases* since the (exact) result is not in $\mathcal{R}$ unless the exponent is integer. power_{FI} takes care of this latter case, where all exponents are ensured to be integers that have not arisen from implicit floating point rounding.

5.3.6.2 Natural exponentiation

The exp^*_F approximation helper function:

$$\mathit{exp}^*_F : F \to \mathcal{R}$$

$\mathit{exp}^*_F(x)$ returns a close approximation to e^x in $\mathcal{R}$, with maximum error $\mathit{max\_error\_exp}_F$.

Further requirements on the exp^*_F approximation helper function are:

$$\begin{array}{ll}
\mathit{exp}^*_F(1) = e & \\
\mathit{exp}^*_F(x) = 1 & \text{if } x \in F \text{ and } \mathit{exp}^*_F(x) \neq e^x \text{ and} \\
 & \quad \ln(1 - (\mathit{epsilon}_F/(2 \cdot r_F))) < x \text{ and} \\
 & \quad x < \ln(1 + (\mathit{epsilon}_F/2)) \\
 & \\
\mathit{exp}^*_F(x) < \mathit{fminD}_F/2 & \text{if } x \in F \text{ and } x < \ln(\mathit{fminD}_F) - 3
\end{array}$$

The exp_F operation:

$$\mathit{exp}_F : F \to F \cup \{\mathbf{underflow}, \mathbf{overflow}\}$$

$$\begin{array}{lll}
\mathit{exp}_F(x) & = \mathit{trans\_result}_F(\mathit{exp}^*_F(x), \mathit{nearest}_F) & \\
 & & \text{if } x \in F \\
 & = 1 & \text{if } x = \mathbf{-0} \\
 & = +\infty & \text{if } x = +\infty \\
 & = 0 & \text{if } x = -\infty \\
 & = \mathit{result\_NaN}_F(x) & \text{otherwise}
\end{array}$$

NOTES

1 $\mathit{exp}_F(1) = \mathit{nearest}_F(e)$.

2 $\mathit{exp}_F(x)$ will overflow approximately when $x > \ln(\mathit{fmax}_F)$.

5.3.6.3 Natural exponentiation, minus one

The $\mathit{expm1}^*_F$ approximation helper function:

$$\mathit{expm1}^*_F : F \to \mathcal{R}$$

$\mathit{expm1}^*_F(x)$ returns a close approximation to $e^x - 1$ in $\mathcal{R}$, with maximum error $\mathit{max\_error\_exp}_F$.

Further requirements on the $\mathit{expm1}^*_F$ approximation helper function are:

$expm1^*_F(1) = e - 1$

$expm1^*_F(x) = x$ if $x \in F$ and $expm1^*_F(x) \neq e^x - 1$ and $-epsilon_F/r_F \leqslant x < 0.5 \cdot epsilon_F/r_F$

$expm1^*_F(x) = -1$ if $x \in F$ and $expm1^*_F(x) \neq e^x - 1$ and $x < \ln(epsilon_F/(3 \cdot r_F))$

Relationship to the exp^*_F approximation helper function:

$expm1^*_F(x) \leqslant exp^*_F(x)$ if $x \in F$

The $expm1_F$ operation:

$expm1_F : F \rightarrow F \cup \{\mathbf{overflow}\}$

$$
\begin{array}{lll}
expm1_F(x) & = trans\_result_F(expm1^*_F(x), nearest_F) & \\
 & & \text{if } x \in F \text{ and } |x| \geqslant fminN_F \\
 & = x & \text{if } x \in F \text{ and } |x| < fminN_F \\
 & = \mathbf{-0} & \text{if } x = \mathbf{-0} \\
 & = \mathbf{+\infty} & \text{if } x = \mathbf{+\infty} \\
 & = -1 & \text{if } x = \mathbf{-\infty} \\
 & = result\_NaN_F(x) & \text{otherwise}
\end{array}
$$

NOTES

1 **underflow** is explicitly avoided. Part 1 requires that $fminN_F \leqslant epsilon_F$. This Part requires that $fminN_F < 0.5 \cdot epsilon_F/r_F$, so that underflow can be avoided here.

2 $expm1_F(1) = nearest_F(e - 1)$.

3 $expm1_F(x)$ will overflow approximately when $x > \ln(fmax_F)$.

5.3.6.4 Exponentiation of 2

The $exp2^*_F$ approximation helper function:

$exp2^*_F : F \rightarrow \mathcal{R}$

$exp2^*_F(x)$ returns a close approximation to 2^x in $\mathcal{R}$, with maximum error $max\_error\_exp_F$.

Further requirements on the $exp2^*_F$ approximation helper function are:

$exp2^*_F(x) = 1$ if $x \in F$ and $exp2^*_F(x) \neq 2^x$ and $\log_2(1 - (epsilon_F/(2 \cdot r_F))) < x$ and $x < \log_2(1 + (epsilon_F/2))$

$exp2^*_F(x) = 2^x$ if $x \in F \cap \mathcal{Z}$ and $2^x \in F$

$exp2^*_F(x) < fminD_F/2$ if $x \in F$ and $x < \log_2(fminD_F) - 3$

The $exp2_F$ operation:

$exp2_F : F \rightarrow F \cup \{\mathbf{underflow}, \mathbf{overflow}\}$

$$
\begin{array}{lll}
exp2_F(x) & = trans\_result_F(exp2^*_F(x), nearest_F) & \\
 & & \text{if } x \in F \\
 & = 1 & \text{if } x = \mathbf{-0} \\
 & = \mathbf{+\infty} & \text{if } x = \mathbf{+\infty} \\
 & = 0 & \text{if } x = \mathbf{-\infty} \\
 & = result\_NaN_F(x) & \text{otherwise}
\end{array}
$$

NOTE – $exp2_F(x)$ will overflow approximately when $x > \log_2(fmax_F)$.

5.3.6.5 Exponentiation of 10

The $\mathit{exp10}^*_F$ approximation helper function:

$$\mathit{exp10}^*_F : F \to \mathcal{R}$$

$\mathit{exp10}^*_F(x)$ returns a close approximation to 10^x in $\mathcal{R}$, with maximum error $\mathit{max\_error\_exp}_F$.

Further requirements on the $\mathit{exp10}^*_F$ approximation helper function are:

$$\begin{array}{ll}
\mathit{exp10}^*_F(x) = 1 & \text{if } x \in F \text{ and } \mathit{exp10}^*_F(x) \neq 10^x \text{ and} \\
 & \quad \log_{10}(1 - (\mathit{epsilon}_F/(2 \cdot r_F))) < x \text{ and} \\
 & \quad x < \log_{10}(1 + (\mathit{epsilon}_F/2)) \\
\mathit{exp10}^*_F(x) = 10^x & \text{if } x \in F \cap \mathcal{Z} \text{ and } 10^x \in F \\
\mathit{exp10}^*_F(x) < \mathit{fminD}_F/2 & \text{if } x \in F \text{ and } x < \log_{10}(\mathit{fminD}_F) - 3
\end{array}$$

The $\mathit{exp10}_F$ operation:

$$\mathit{exp10}_F : F \to F \cup \{\textbf{underflow}, \textbf{overflow}\}$$

$$\begin{array}{lll}
\mathit{exp10}_F(x) & = \mathit{trans\_result}_F(\mathit{exp10}^*_F(x), \mathit{nearest}_F) & \\
 & & \text{if } x \in F \\
 & = 1 & \text{if } x = \mathbf{-0} \\
 & = \mathbf{+\infty} & \text{if } x = \mathbf{+\infty} \\
 & = 0 & \text{if } x = \mathbf{-\infty} \\
 & = \mathit{result\_NaN}_F(x) & \text{otherwise}
\end{array}$$

NOTE – $\mathit{exp10}_F(x)$will overflow approximately when $x > \log_{10}(\mathit{fmax}_F)$.

5.3.6.6 Exponentiation of argument base

The power^*_F approximation helper function:

$$\mathit{power}^*_F : F \times F \to \mathcal{R}$$

$\mathit{power}^*_F(x, y)$ returns a close approximation to x^y in $\mathcal{R}$, with maximum error $\mathit{max\_error\_power}_F$. The power^*_F helper function need be defined only for first arguments that are greater than 0.

Further requirements on the power^*_F approximation helper function are:

$$\begin{array}{ll}
\mathit{power}^*_F(1, y) = 1 & \text{if } y \in F \\
\mathit{power}^*_F(x, 0) = 1 & \text{if } x \in F \text{ and } x > 0 \\
\mathit{power}^*_F(x, 1) = x & \text{if } x \in F \text{ and } x > 0 \\
\mathit{power}_F F^*(x, y) < \mathit{fminD}_F/2 & \text{if } x \in F \text{ and } x > 0 \text{ and } y \in F \text{ and } x^y < \mathit{fminD}_F/3
\end{array}$$

Relationship to the power^*_{FI} approximation helper function:

$$\begin{array}{ll}
\mathit{power}^*_F(x, y) = \mathit{power}^*_{FI}(x, y) & \text{if } x \in F \text{ and } x > 0 \text{ and } y \in I \cap F
\end{array}$$

The power_F operation:

$$\mathit{power}_F : F \times F \to F \cup \{\textbf{invalid}, \textbf{underflow}, \textbf{overflow}, \textbf{pole}\}$$

$$\begin{array}{lll}
\mathit{power}_F(x, y) & = \mathit{trans\_result}_F(\mathit{power}^*_F(x, y), \mathit{nearest}_F) & \\
 & & \text{if } x \in F \text{ and } x > 0 \text{ and } y \in F \\
 & = \mathit{power}_F(0, y) & \text{if } x = \mathbf{-0} \text{ and } y \in F \cup \{\mathbf{-\infty}, \mathbf{-0}, \mathbf{+\infty}\} \\
 & = \mathit{power}_F(x, 0) & \text{if } y = \mathbf{-0} \text{ and } x \in F \cup \{\mathbf{-\infty}, \mathbf{+\infty}\} \\
 & & \\
 & = \mathbf{+\infty} & \text{if } x = \mathbf{+\infty} \text{ and } ((y \in F \text{ and } y > 0) \text{ or } y = \mathbf{+\infty}) \\
 & = \mathbf{+\infty} & \text{if } x \in F \text{ and } x > 1 \text{ and } y = \mathbf{+\infty} \\
 & = 0 & \text{if } x \in F \text{ and } 0 \leqslant x < 1 \text{ and } y = \mathbf{+\infty} \\
 & = 0 & \text{if } x = 0 \text{ and } y \in F \text{ and } y > 0
\end{array}$$

$$
\begin{aligned}
&= \mathbf{pole}(\mathbf{+\infty}) && \text{if } x = 0 \text{ and } y \in F \text{ and } y < 0\\
&= \mathbf{+\infty} && \text{if } x \in F \text{ and } 0 \leqslant x < 1 \text{ and } y = \mathbf{-\infty}\\
&= 0 && \text{if } x \in F \text{ and } x > 1 \text{ and } y = \mathbf{-\infty}\\
&= 0 && \text{if } x = \mathbf{+\infty} \text{ and } ((y \in F \text{ and } y < 0) \text{ or } y = \mathbf{-\infty})\\
&= \mathit{result\_NaN2}_F(x, y) && \text{otherwise}
\end{aligned}
$$

NOTE – $\mathit{power}_F(x, y)$ will overflow approximately when $x^y > \mathit{fmax}_F$, i.e., if $x > 1$, approximately when $y > \log_x(\mathit{fmax}_F)$, and if $0 < x < 1$, approximately when $y < \log_x(\mathit{fmax}_F)$ (which is a negative number). It will not overflow when $x = 0$ or when $x = 1$.

5.3.6.7 Exponentiation of one plus the argument base, minus one

The $\mathit{power1pm1}^*_F$ approximation helper function:

$$\mathit{power1pm1}^*_F : F \times F \to \mathcal{R}$$

$\mathit{power1pm1}^*_F(x, y)$ returns a close approximation to $(1 + x)^y - 1$ in $\mathcal{R}$, with maximum error $\mathit{max\_error\_power}_F$. The $\mathit{power1pm1}^*_F$ helper function need be defined only for first arguments that are greater than -1.

Further requirements on the $\mathit{power1pm1}^*_F$ approximation helper function are:

$$
\begin{aligned}
&\mathit{power1pm1}^*_F(-1, y) = -1 && \text{if } y \in F \text{ and } y > 0\\
&\mathit{power1pm1}^*_F(x, y) = -1 && \text{if } x \in F \text{ and } x > -1 \text{ and } y \in F \text{ and}\\
& && \quad \mathit{power1pm1}^*_F(x, y) \neq (1 + x)^y - 1 \text{ and}\\
& && \quad (1 + x)^y < \mathit{epsilon}_F/(3 \cdot r_F)\\
&\mathit{power1pm1}^*_F(x, 1) = 1 + x && \text{if } x, 1 + x \in F \text{ and } x > -1
\end{aligned}
$$

Relationship to the power^*_F approximation helper function:

$$\mathit{power1pm1}^*_F(x, y) \leqslant \mathit{power}^*_F(1 + x, y) \qquad \text{if } x, 1 + x \in F \text{ and } x > -1 \text{ and } y \in F$$

NOTE 1 – $\mathit{power1pm1}^*_F(x, y) \approx y \cdot \ln(1 + x)$ if $x \in F$ and $x > -1$ and $y \in F$ and $|y \cdot \ln(1 + x)| < \mathit{epsilon}_F/r_F$.

The $\mathit{power1pm1}_F$ operation:

$$\mathit{power1pm1}_F : F \times F \to F \cup \{\mathbf{-0}, \mathbf{invalid}, \mathbf{underflow}, \mathbf{overflow}, \mathbf{pole}\}$$

$$
\begin{aligned}
&\mathit{power1pm1}_F(x, y)\\
&= \mathit{trans\_result}_F(\mathit{power1pm1}^*_F(x, y), \mathit{nearest}_F)\\
& && \text{if } x \in F \text{ and } x > -1 \text{ and } x \neq 0 \text{ and } y \in F \text{ and } y \neq 0\\
&= \mathit{mul}_F(x, y) && \text{if } x \in \{\mathbf{-0}, 0\} \text{ and } y \in F \text{ and } y \neq 0\\
&= \mathit{mul}_F(x, y) && \text{if } y \in \{\mathbf{-0}, 0\} \text{ and } x \in F \text{ and } x > -1\\
&= \mathbf{+\infty} && \text{if } x = \mathbf{+\infty} \text{ and } ((y \in F \text{ and } y > 0) \text{ or } y = \mathbf{+\infty})\\
&= \mathbf{+\infty} && \text{if } x \in F \text{ and } x > 0 \text{ and } y = \mathbf{+\infty}\\
&= -1 && \text{if } x \in F \text{ and } -1 \leqslant x < 0 \text{ and } y = \mathbf{+\infty}\\
&= -1 && \text{if } x = -1 \text{ and } y \in F \text{ and } y > 0\\
&= \mathbf{pole}(\mathbf{+\infty}) && \text{if } x = -1 \text{ and } y \in F \text{ and } y < 0\\
&= \mathbf{+\infty} && \text{if } x \in F \text{ and } -1 \leqslant x < 0 \text{ and } y = \mathbf{-\infty}\\
&= -1 && \text{if } x \in F \text{ and } x > 0 \text{ and } y = \mathbf{-\infty}\\
&= -1 && \text{if } x = \mathbf{+\infty} \text{ and } ((y \in F \text{ and } y < 0) \text{ or } y = \mathbf{-\infty})\\
&= \mathit{result\_NaN2}_F(x, y) && \text{otherwise}
\end{aligned}
$$

NOTE 2 – $\mathit{power1pm1}_F(x, y)$ will overflow approximately when $(1 + x)^y > \mathit{fmax}_F$, i.e., if $x > 0$, approximately when $y > \log_{1+x}(\mathit{fmax}_F)$, and if $-1 < x < 0$, approximately when $y < \log_{1+x}(\mathit{fmax}_F)$. It will not overflow when $x \in \{-1, 0\}$.

5.3.6.8 Natural logarithm

The ln^*_F approximation helper function:

$$ln^*_F : \mathcal{R} \rightarrow \mathcal{R}$$

$ln^*_F(x)$ returns a close approximation to $\ln(x)$ in $\mathcal{R}$, with maximum error $max\_error\_exp_F$.

A further requirement on the ln^*_F approximation helper function is:

$$ln^*_F(e) = 1$$

The ln_F operation:

$$ln_F : F \rightarrow F \cup \{\mathbf{invalid}, \mathbf{pole}\}$$

$$\begin{aligned}
ln_F(x) &= trans\_result_F(ln^*_F(x), nearest_F) && \\
&&& \text{if } x \in F \text{ and } x > 0 \\
&= \mathbf{pole}(-\infty) && \text{if } x \in \{\mathbf{-0}, 0\} \\
&= \mathbf{+\infty} && \text{if } x = \mathbf{+\infty} \\
&= result\_NaN_F(x) && \text{otherwise}
\end{aligned}$$

5.3.6.9 Natural logarithm of one plus the argument

The $ln1p^*_F$ approximation helper function:

$$ln1p^*_F : \mathcal{R} \rightarrow \mathcal{R}$$

$ln1p^*_F(x)$ returns a close approximation to $\ln(1+x)$ in $\mathcal{R}$, with maximum error $max\_error\_exp_F$.

Further requirements on the $ln1p^*_F$ approximation helper function are:

$$\begin{aligned}
&ln1p^*_F(e-1) = 1 && \\
&ln1p^*_F(x) = x && \text{if } x \in F \text{ and } ln1p^*_F(x) \neq \ln(1+x) \text{ and} \\
&&& \quad -0.5 \cdot epsilon_F/r_F < x \leqslant epsilon_F/r_F
\end{aligned}$$

Relationship to the ln^*_F approximation helper function:

$$ln1p^*_F(x) \geqslant ln^*_F(x) \qquad \text{if } x \in F \text{ and } x > 0$$

The $ln1p_F$ operation:

$$ln1p_F : F \rightarrow F \cup \{\mathbf{invalid}, \mathbf{pole}\}$$

$$\begin{aligned}
ln1p_F(x) &= trans\_result_F(ln1p^*_F(x), nearest_F) && \\
&&& \text{if } x \in F \text{ and } x > -1 \text{ and } |x| \geqslant fminN_F \\
&= x && \text{if } x \in F \text{ and } |x| < fminN_F \\
&= \mathbf{-0} && \text{if } x = \mathbf{-0} \\
&= \mathbf{pole}(-\infty) && \text{if } x = -1 \\
&= \mathbf{+\infty} && \text{if } x = \mathbf{+\infty} \\
&= result\_NaN_F(x) && \text{otherwise}
\end{aligned}$$

NOTE – **underflow** is explicitly avoided. Part 1 requires that $fminN_F \leqslant epsilon_F$. This Part requires that $fminN_F < 0.5 \cdot epsilon_F/r_F$, so that underflow can be avoided here.

5.3.6.10 2-logarithm

The $log2^*_F$ approximation helper function:

$$log2^*_F : F \rightarrow \mathcal{R}$$

$log2^*_F(x)$ returns a close approximation to $\log_2(x)$ in $\mathcal{R}$, with maximum error $max\_error\_exp_F$.

A further requirement on the $log2^*_F$ approximation helper function is:

$$log2^*_F(x) = \log_2(x) \qquad \text{if } x \in F \text{ and } \log_2(x) \in \mathcal{Z}$$

The $log2_F$ operation:

$$log2_F : F \to F \cup \{\mathbf{invalid}, \mathbf{pole}\}$$

$$\begin{aligned}
log2_F(x) &= trans\_result_F(log2^*_F(x), nearest_F) && \\
&&& \text{if } x \in F \text{ and } x > 0\\
&= \mathbf{pole}(-\infty) && \text{if } x \in \{\mathbf{-0}, 0\}\\
&= \mathbf{+\infty} && \text{if } x = \mathbf{+\infty}\\
&= result\_NaN_F(x) && \text{otherwise}
\end{aligned}$$

5.3.6.11 10-logarithm

The $log10^*_F$ approximation helper function:

$$log10^*_F : F \to \mathcal{R}$$

$log10^*_F(x)$ returns a close approximation to $\log_{10}(x)$ in $\mathcal{R}$, with maximum error $max\_error\_exp_F$.

A further requirement on the $log10^*_F$ approximation helper function is:

$$log10^*_F(x) = \log_{10}(x) \qquad \text{if } x \in F \text{ and } \log_{10}(x) \in \mathcal{Z}$$

The $log10_F$ operation:

$$log10_F : F \to F \cup \{\mathbf{invalid}, \mathbf{pole}\}$$

$$\begin{aligned}
log10_F(x) &= trans\_result_F(log10^*_F(x), nearest_F) && \\
&&& \text{if } x \in F \text{ and } x > 0\\
&= \mathbf{pole}(-\infty) && \text{if } x \in \{\mathbf{-0}, 0\}\\
&= \mathbf{+\infty} && \text{if } x = \mathbf{+\infty}\\
&= result\_NaN_F(x) && \text{otherwise}
\end{aligned}$$

5.3.6.12 Argument base logarithm

The $logbase^*_F$ approximation helper function:

$$logbase^*_F : F \times F \to \mathcal{R}$$

$logbase^*_F(x, y)$ returns a close approximation to $\log_x(y)$ in $\mathcal{R}$, with maximum error $max\_error\_power_F$.

A further requirement on the $logbase^*_F$ approximation helper function is:

$$logbase^*_F(x, x) = 1 \qquad \text{if } x \in F \text{ and } x > 0 \text{ and } x \neq 1$$

The $logbase_F$ operation:

$$logbase_F : F \times F \to F \cup \{\mathbf{invalid}, \mathbf{pole}\}$$

$$\begin{aligned}
logbase_F(x, y) &= trans\_result_F(logbase^*_F(x, y), nearest_F) && \\
&&& \text{if } x \in F \text{ and } x > 0 \text{ and } x \neq 1 \text{ and } y \in F \text{ and } y > 0\\
&= logbase_F(0, y) && \text{if } x = \mathbf{-0} \text{ and } y \in F \cup \{\mathbf{-\infty}, \mathbf{-0}, \mathbf{+\infty}\}\\
&= logbase_F(x, 0) && \text{if } y = \mathbf{-0} \text{ and } x \in F \cup \{\mathbf{-\infty}, \mathbf{+\infty}\}\\
&&&\\
&= \mathbf{pole}(\mathbf{+\infty}) && \text{if } x = 1 \text{ and } y \in F \text{ and } y > 1\\
&= \mathbf{pole}(-\infty) && \text{if } x = 1 \text{ and } y \in F \text{ and } 0 \leqslant y < 1\\
&&&\\
&= 0 && \text{if } x = \mathbf{+\infty} \text{ and } y \in F \text{ and } y \geqslant 1\\
&= \mathbf{+\infty} && \text{if } x \in F \text{ and } 1 \leqslant x \text{ and } y = \mathbf{+\infty}\\
&= \mathbf{-\infty} && \text{if } x \in F \text{ and } 0 < x < 1 \text{ and } y = \mathbf{+\infty}\\
&= \mathbf{-0} && \text{if } x = 0 \text{ and } y \in F \text{ and } y \geqslant 1\\
&= 0 && \text{if } x = 0 \text{ and } y \in F \text{ and } 0 < y < 1
\end{aligned}$$

$$
\begin{array}{ll}
= \mathbf{pole}(\mathbf{+\infty}) & \text{if } x \in F \text{ and } 0 < x < 1 \text{ and } y = 0 \\
= \mathbf{pole}(\mathbf{-\infty}) & \text{if } x \in F \text{ and } 1 < x \text{ and } y = 0 \\
= \mathbf{-0} & \text{if } x = \mathbf{+\infty} \text{ and } y \in F \text{ and } 0 < y < 1 \\
 & \\
= \mathit{result\_NaN2}_F(x, y) & \text{otherwise}
\end{array}
$$

5.3.6.13 Argument base logarithm of one plus each argument

The $\mathit{logbase1p1p}^*_F$ approximation helper function:

$\mathit{logbase1p1p}^*_F : F \times F \to \mathcal{R}$

$\mathit{logbase1p1p}^*_F(x, y)$ returns a close approximation to $\log_{(1+x)}(1 + y)$ in $\mathcal{R}$, with maximum error $\mathit{max\_error\_power}_F$.

A further requirements on $\mathit{logbase1p1p}^*_F$ approximation helper function is:

$\mathit{logbase1p1p}^*_F(x, x) = 1$ if $x \in F$ and $x > -1$ and $x \neq 0$

The $\mathit{logbase1p1p}_F$ operation:

$\mathit{logbase1p1p}_F : F \times F \to F \cup \{\mathbf{-0}, \mathbf{invalid}, \mathbf{underflow}, \mathbf{pole}\}$

$\mathit{logbase1p1p}_F(x, y)$

$$
\begin{array}{ll}
\multicolumn{2}{l}{= \mathit{trans\_result}_F(\mathit{logbase1p1p}^*_F(x, y), \mathit{nearest}_F)} \\
 & \text{if } x \in F \text{ and } x > -1 \text{ and } x \neq 0 \text{ and} \\
 & \quad y \in F \text{ and } y > -1 \text{ and } y \neq 0 \\
= \mathit{div}_F(y, x) & \text{if } x \in \{\mathbf{-0}, 0\} \text{ and} \\
 & \quad ((y \in F \text{ and } y > -1 \text{ and } y \neq 0) \text{ or } y = \mathbf{+\infty}) \\
= \mathit{div}_F(y, x) & \text{if } y \in \{\mathbf{-0}, 0\} \text{ and} \\
 & \quad ((x \in F \text{ and } x > -1) \text{ or } x = \mathbf{+\infty}) \\
 & \\
= 0 & \text{if } x = \mathbf{+\infty} \text{ and } y \in F \text{ and } y \geqslant 0 \\
= \mathbf{+\infty} & \text{if } x \in F \text{ and } 0 < x \text{ and } y = \mathbf{+\infty} \\
= \mathbf{-\infty} & \text{if } x \in F \text{ and } -1 < x < 0 \text{ and } y = \mathbf{+\infty} \\
= \mathbf{-0} & \text{if } x = -1 \text{ and } y \in F \text{ and } y \geqslant 0 \\
= 0 & \text{if } x = -1 \text{ and } y \in F \text{ and } -1 < y < 0 \\
= \mathbf{pole}(\mathbf{+\infty}) & \text{if } x \in F \text{ and } -1 < x < 0 \text{ and } y = -1 \\
= \mathbf{pole}(\mathbf{-\infty}) & \text{if } x \in F \text{ and } 0 < x \text{ and } y = -1 \\
= \mathbf{-0} & \text{if } x = \mathbf{+\infty} \text{ and } y \in F \text{ and } -1 < y < 0 \\
 & \\
= \mathit{result\_NaN2}_F(x, y) & \text{otherwise}
\end{array}
$$

5.3.7 Operations for hyperbolic elementary functions

There shall be two maximum error parameters for operations corresponding to the hyperbolic and inverse hyperbolic functions:

$\mathit{max\_error\_sinh}_F \in F$
$\mathit{max\_error\_tanh}_F \in F$

The $\mathit{max\_error\_sinh}_F$ parameter shall have a value in the interval $[0.5, 2 \cdot \mathit{rnd\_error}_F]$. The $\mathit{max\_error\_tanh}_F$ parameter shall have a value in the interval $[\mathit{max\_error\_sinh}_F, 2 \cdot \mathit{rnd\_error}_F]$.

5.3.7.1 Hyperbolic sine

The $sinh^*_F$ approximation helper function:

$sinh^*_F : F \to \mathcal{R}$

$sinh^*_F(x)$ returns a close approximation to $\sinh(x)$ in $\mathcal{R}$, with maximum error $max\_error\_sinh_F$.

Further requirements on the $sinh^*_F$ approximation helper function are:

$$
\begin{array}{ll}
sinh^*_F(x) = x & \text{if } x \in F \text{ and } sinh^*_F(x) \neq \sinh(x) \text{ and} \\
 & \quad |x| < \sqrt{2 \cdot epsilon_F / r_F} \\
sinh^*_F(-x) = -sinh^*_F(x) & \text{if } x \in F
\end{array}
$$

The $sinh_F$ operation:

$sinh_F : F \to F \cup \{\mathbf{overflow}\}$

$$
\begin{array}{lll}
sinh_F(x) & = trans\_result_F(sinh^*_F(x), nearest_F) & \\
 & & \text{if } x \in F \text{ and } |x| > fminN_F \\
 & = x & \text{if } x \in F \text{ and } |x| \leqslant fminN_F \\
 & = x & \text{if } x \in \{\mathbf{-\infty}, \mathbf{-0}, \mathbf{+\infty}\} \\
 & = result\_NaN_F(x) & \text{otherwise}
\end{array}
$$

NOTES

1 **underflow** is explicitly avoided.

2 $sinh_F(x)$ will overflow approximately when $|x| > \ln(2 \cdot fmax_F)$.

5.3.7.2 Hyperbolic cosine

The $cosh^*_F$ approximation helper function:

$cosh^*_F : F \to \mathcal{R}$

$cosh^*_F(x)$ returns a close approximation to $\cosh(x)$ in $\mathcal{R}$, with maximum error $max\_error\_sinh_F$.

Further requirements on the $cosh^*_F$ approximation helper function are:

$$
\begin{array}{ll}
cosh^*_F(x) = 1 & \text{if } x \in F \text{ and } cosh^*_F(x) \neq \cosh(x) \text{ and} \\
 & \quad |x| < \sqrt{epsilon_F} \\
cosh^*_F(-x) = cosh^*_F(x) & \text{if } x \in F
\end{array}
$$

Relationship to the $sinh^*_F$ approximation helper function:

$$
\begin{array}{ll}
cosh^*_F(x) \geqslant sinh^*_F(x) & \text{if } x \in F
\end{array}
$$

The $cosh_F$ operation:

$cosh_F : F \to F \cup \{\mathbf{overflow}\}$

$$
\begin{array}{lll}
cosh_F(x) & = trans\_result_F(cosh^*_F(x), nearest_F) & \\
 & & \text{if } x \in F \\
 & = 1 & \text{if } x = \mathbf{-0} \\
 & = \mathbf{+\infty} & \text{if } x \in \{\mathbf{-\infty}, \mathbf{+\infty}\} \\
 & = result\_NaN_F(x) & \text{otherwise}
\end{array}
$$

NOTE – $cosh_F(x)$ overflows approximately when $|x| > \ln(2 \cdot fmax_F)$.

5.3.7.3 Hyperbolic tangent

The $tanh^*_F$ approximation helper function:

$$tanh^*_F : F \rightarrow \mathcal{R}$$

$tanh^*_F(x)$ returns a close approximation to $\tanh(x)$ in $\mathcal{R}$, with maximum error $max\_error\_tanh_F$.

Further requirements on the $tanh^*_F$ approximation helper function are:

$$\begin{array}{ll}
tanh^*_F(x) = x & \text{if } x \in F \text{ and } tanh^*_F(x) \neq \tanh(x) \text{ and} \\
 & \quad |x| \leqslant \sqrt{1.5 \cdot epsilon_F / r_F} \\
tanh^*_F(x) = 1 & \text{if } x \in F \text{ and } tanh^*_F(x) \neq \tanh(x) \text{ and} \\
 & \quad x > \text{arctanh}(1 - (epsilon_F/(3 \cdot r_F))) \\
tanh^*_F(-x) = -tanh^*_F(x) & \text{if } x \in F
\end{array}$$

The $tanh_F$ operation:

$$tanh_F : F \rightarrow F$$

$$\begin{array}{lll}
tanh_F(x) & = trans\_result_F(tanh^*_F(x), nearest_F) & \\
 & & \text{if } x \in F \text{ and } |x| > fminN_F \\
 & = x & \text{if } x \in F \text{ and } |x| \leqslant fminN_F \\
 & = \mathbf{-0} & \text{if } x = \mathbf{-0} \\
 & = -1 & \text{if } x = \mathbf{-\infty} \\
 & = 1 & \text{if } x = \mathbf{+\infty} \\
 & = result\_NaN_F(x) & \text{otherwise}
\end{array}$$

NOTE – **underflow** is explicitly avoided.

5.3.7.4 Hyperbolic cotangent

The $coth^*_F$ approximation helper function:

$$coth^*_F : F \rightarrow \mathcal{R}$$

$coth^*_F(x)$ returns a close approximation to $\coth(x)$ in $\mathcal{R}$, with maximum error $max\_error\_tanh_F$.

Further requirements on the $coth^*_F$ approximation helper function are:

$$\begin{array}{ll}
coth^*_F(x) = 1 & \text{if } x \in F \text{ and } coth^*_F(x) \neq \coth(x) \text{ and} \\
 & \quad x > \text{arccoth}(1 + (epsilon_F/4)) \\
coth^*_F(-x) = -coth^*_F(x) & \text{if } x \in F
\end{array}$$

The $coth_F$ operation:

$$coth_F : F \rightarrow F \cup \{\mathbf{pole}, \mathbf{overflow}\}$$

$$\begin{array}{lll}
coth_F(x) & = trans\_result_F(coth^*_F(x), nearest_F) & \\
 & & \text{if } x \in F \text{ and } x \neq 0 \\
 & = \mathbf{pole}(\mathbf{+\infty}) & \text{if } x = 0 \\
 & = \mathbf{pole}(\mathbf{-\infty}) & \text{if } x = \mathbf{-0} \\
 & = -1 & \text{if } x = \mathbf{-\infty} \\
 & = 1 & \text{if } x = \mathbf{+\infty} \\
 & = result\_NaN_F(x) & \text{otherwise}
\end{array}$$

NOTE – $coth_F(x)$ overflows approximately when $|1/x| > fmax_F$.

5.3.7.5 Hyperbolic secant

The $sech^*_F$ approximation helper function:

$$sech^*_F : F \to \mathcal{R}$$

$sech^*_F(x)$ returns a close approximation to $\mathrm{sech}(x)$ in $\mathcal{R}$, with maximum error $max\_error\_tanh_F$.

Further requirements on the $sech^*_F$ approximation helper function are:

$$\begin{array}{ll} sech^*_F(x) = 1 & \text{if } x \in F \text{ and } sech^*_F(x) \neq \mathrm{sech}(x) \text{ and} \\ & \quad |x| < \sqrt{epsilon_F/r_F} \\ sech^*_F(-x) = sech^*_F(x) & \text{if } x \in F \\ sech^*_F(x) < fminD_F/2 & \text{if } x \in F \text{ and } x > 2 - \ln(fminD_F/4) \end{array}$$

The $sech_F$ operation:

$$sech_F : F \to F \cup \{\mathbf{underflow}\}$$

$$\begin{array}{lll} sech_F(x) & = trans\_result_F(sech^*_F(x), nearest_F) & \\ & & \text{if } x \in F \\ & = 1 & \text{if } x = \mathbf{-0} \\ & = 0 & \text{if } x \in \{\mathbf{-\infty}, \mathbf{+\infty}\} \\ & = result\_NaN_F(x) & \text{otherwise} \end{array}$$

5.3.7.6 Hyperbolic cosecant

The $csch^*_F$ approximation helper function:

$$csch^*_F : F \to \mathcal{R}$$

$csch^*_F(x)$ returns a close approximation to $\mathrm{csch}(x)$ in $\mathcal{R}$, with maximum error $max\_error\_tanh_F$.

Further requirements on the $csch^*_F$ approximation helper function are:

$$\begin{array}{ll} csch^*_F(-x) = -csch^*_F(x) & \text{if } x \in F \\ csch^*_F(x) < fminD_F/2 & \text{if } x \in F \text{ and } x > 2 - \ln(fminD_F/4) \end{array}$$

Relationship to the $sech^*_F$ approximation helper function:

$$\begin{array}{ll} csch^*_F(x) \geqslant sech^*_F(x) & \text{if } x \in F \text{ and } x > 0 \end{array}$$

The $csch_F$ operation:

$$csch_F : F \to F \cup \{\mathbf{underflow}, \mathbf{overflow}, \mathbf{pole}\}$$

$$\begin{array}{lll} csch_F(x) & = trans\_result_F(csch^*_F(x), nearest_F) & \\ & & \text{if } x \in F \text{ and } x \neq 0 \\ & = div_F(1, x) & \text{if } x \in \{\mathbf{-\infty}, \mathbf{-0}, 0, \mathbf{+\infty}\} \\ & = result\_NaN_F(x) & \text{otherwise} \end{array}$$

NOTE – $csch_F(x)$ overflows approximately when $|1/x| > fmax_F$.

5.3.7.7 Inverse hyperbolic sine

The $arcsinh^*_F$ approximation helper function:

$$arcsinh^*_F : F \to \mathcal{R}$$

$arcsinh^*_F(x)$ returns a close approximation to $\mathrm{arcsinh}(x)$ in $\mathcal{R}$, with maximum error $max\_error\_sinh_F$.

Further requirements on the $arcsinh^*_F$ approximation helper function are:

$$\begin{aligned}
&\mathit{arcsinh}^*_F(x) = x && \text{if } x \in F \text{ and } \mathit{arcsinh}^*_F(x) \neq \operatorname{arcsinh}(x) \text{ and} \\
& && \quad |x| \leqslant \sqrt{3 \cdot \mathit{epsilon}_F / r_F} \\
&\mathit{arcsinh}^*_F(-x) = -\mathit{arcsinh}^*_F(x) && \text{if } x \in F
\end{aligned}$$

The $\mathit{arcsinh}_F$ operation:

$\mathit{arcsinh}_F : F \to F$

$$\begin{aligned}
\mathit{arcsinh}_F(x) &= \mathit{trans\_result}_F(\mathit{arcsinh}^*_F(x), \mathit{nearest}_F) \\
& \qquad\qquad\qquad \text{if } x \in F \text{ and } |x| > \mathit{fminN}_F \\
&= x \qquad\qquad \text{if } x \in F \text{ and } |x| \leqslant \mathit{fminN}_F \\
&= x \qquad\qquad \text{if } x \in \{\mathbf{-\infty}, \mathbf{-0}, \mathbf{+\infty}\} \\
&= \mathit{result\_NaN}_F(x) \qquad \text{otherwise}
\end{aligned}$$

NOTE – **underflow** is explicitly avoided.

5.3.7.8 Inverse hyperbolic cosine

The $\mathit{arccosh}^*_F$ approximation helper function:

$\mathit{arccosh}^*_F : F \to \mathcal{R}$

$\mathit{arccosh}^*_F(x)$ returns a close approximation to $\operatorname{arccosh}(x)$ in $\mathcal{R}$, with maximum error $\mathit{max\_error\_sinh}_F$.

Relationship to the $\mathit{arcsinh}^*_F$ approximation helper function:

$\mathit{arccosh}^*_F(x) \leqslant \mathit{arcsinh}^*_F(x)$

The $\mathit{arccosh}_F$ operation:

$\mathit{arccosh}_F : F \to F \cup \{\mathbf{invalid}\}$

$$\begin{aligned}
\mathit{arccosh}_F(x) &= \mathit{trans\_result}_F(\mathit{arccosh}^*_F(x), \mathit{nearest}_F) \\
& \qquad\qquad\qquad \text{if } x \in F \text{ and } x \geqslant 1 \\
&= \mathbf{+\infty} \qquad\qquad \text{if } x = \mathbf{+\infty} \\
&= \mathit{result\_NaN}_F(x) \qquad \text{otherwise}
\end{aligned}$$

5.3.7.9 Inverse hyperbolic tangent

The $\mathit{arctanh}^*_F$ approximation helper function:

$\mathit{arctanh}^*_F : F \to \mathcal{R}$

$\mathit{arctanh}^*_F(x)$ returns a close approximation to $\operatorname{arctanh}(x)$ in $\mathcal{R}$, with maximum error $\mathit{max\_error\_tanh}_F$.

Further requirements on the $\mathit{arctanh}^*_F$ approximation helper function are:

$$\begin{aligned}
&\mathit{arctanh}^*_F(x) = x && \text{if } x \in F \text{ and } \mathit{arctanh}^*_F(x) \neq \operatorname{arctanh}(x) \text{ and} \\
& && \quad |x| < \sqrt{\mathit{epsilon}_F / r_F} \\
&\mathit{arctanh}^*_F(-x) = -\mathit{arctanh}^*_F(x) && \text{if } x \in F
\end{aligned}$$

The $\mathit{arctanh}_F$ operation:

$\mathit{arctanh}_F : F \to F \cup \{\mathbf{invalid}, \mathbf{pole}\}$

$$\begin{aligned}
\mathit{arctanh}_F(x) &= \mathit{trans\_result}_F(\mathit{arctanh}^*_F(x), \mathit{nearest}_F) \\
& \qquad\qquad\qquad \text{if } x \in F \text{ and } \mathit{fminN}_F < |x| < 1 \\
&= x \qquad\qquad \text{if } x \in F \text{ and } |x| \leqslant \mathit{fminN}_F \\
&= \mathbf{-0} \qquad\qquad \text{if } x = \mathbf{-0} \\
&= \mathbf{pole}(\mathbf{+\infty}) \qquad \text{if } x = 1 \\
&= \mathbf{pole}(\mathbf{-\infty}) \qquad \text{if } x = -1 \\
&= \mathit{result\_NaN}_F(x) \qquad \text{otherwise}
\end{aligned}$$

NOTE – **underflow** is explicitly avoided.

5.3.7.10 Inverse hyperbolic cotangent

The $arccoth^*_F$ approximation helper function:

$arccoth^*_F : F \rightarrow \mathcal{R}$

$arccoth^*_F(x)$ returns a close approximation to $\mathrm{arccoth}(x)$ in $\mathcal{R}$, with maximum error $max\_error\_tanh_F$.

A further requirements on the $arccoth^*_F$ approximation helper function is:

$$arccoth^*_F(-x) = -arccoth^*_F(x) \qquad \text{if } x \in F$$

The $arccoth_F$ operation:

$arccoth_F : F \rightarrow F \cup \{\textbf{invalid}, \textbf{underflow}, \textbf{pole}\}$

$$\begin{aligned} arccoth_F(x) &= trans\_result_F(arccoth^*_F(x), nearest_F) \\ & \qquad\qquad \text{if } x \in F \text{ and } |x| > 1 \\ &= \textbf{pole}(\mathbf{+\infty}) && \text{if } x = 1 \\ &= \textbf{pole}(\mathbf{-\infty}) && \text{if } x = -1 \\ &= \mathbf{-0} && \text{if } x = \mathbf{-\infty} \\ &= 0 && \text{if } x = \mathbf{+\infty} \\ &= result\_NaN_F(x) && \text{otherwise} \end{aligned}$$

NOTE – There is no **underflow** for this operation for most kinds of floating point types, e.g. IEC 60559 ones.

5.3.7.11 Inverse hyperbolic secant

The $arcsech^*_F$ approximation helper function:

$arcsech^*_F : F \rightarrow \mathcal{R}$

$arcsech^*_F(x)$ returns a close approximation to $\mathrm{arcsech}(x)$ in $\mathcal{R}$, with maximum error $max\_error\_tanh_F$.

The $arcsech_F$ operation:

$arcsech_F : F \rightarrow F \cup \{\textbf{invalid}, \textbf{pole}\}$

$$\begin{aligned} arcsech_F(x) &= trans\_result_F(arcsech^*_F(x), nearest_F) \\ & \qquad\qquad \text{if } x \in F \text{ and } 0 < x \leqslant 1 \\ &= \textbf{pole}(\mathbf{+\infty}) && \text{if } x \in \{\mathbf{-0}, 0\} \\ &= result\_NaN_F(x) && \text{otherwise} \end{aligned}$$

5.3.7.12 Inverse hyperbolic cosecant

The $arccsch^*_F$ approximation helper function:

$arccsch^*_F : F \rightarrow \mathcal{R}$

$arccsch^*_F(x)$ returns a close approximation to $\mathrm{arccsch}(x)$ in $\mathcal{R}$, with maximum error $max\_error\_tanh_F$.

A further requirements on the $arccsch^*_F$ approximation helper function is:

$$arccsch^*_F(-x) = -arccsch^*_F(x) \qquad \text{if } x \in F$$

Relationship to the $arcsinh^*_F$ approximation helper function:

$arccsch^*_F(1) = arcsinh^*_F(1)$

The $arccsch_F$ operation:

$arccsch_F : F \rightarrow F \cup \{\textbf{underflow}, \textbf{pole}\}$

$$
\begin{aligned}
arccsch_F(x) &= trans\_result_F(arccsch^*_F(x), nearest_F) && \text{if } x \in F \text{ and } x \neq 0\\
&= div_F(1, x) && \text{if } x \in \{\mathbf{-\infty}, \mathbf{-0}, 0, \mathbf{+\infty}\}\\
&= result\_NaN_F(x) && \text{otherwise}
\end{aligned}
$$

NOTE – There is no **underflow** for this operation for most kinds of floating point types, e.g. IEC 60559 ones.

5.3.8 Introduction to operations for trigonometric elementary functions

Two different operations for each of sin, cos, tan, cot, sec, csc, arcsin, arccos, arctan, arccot, arcctg, arcsec, and arccsc are specified. One version for radians and one version where the angular unit is given as a parameter.

For use in the specifications below, define the following mathematical functions:

$rad : \mathcal{R} \rightarrow \mathcal{R}$
$axis\_rad : \mathcal{R} \rightarrow \{(1,0),(0,1),(-1,0),(0,-1)\} \times \mathcal{R}$
$arc : \mathcal{R} \times \mathcal{R} \rightarrow \mathcal{R}$

The *rad*, angular value normalisation, function is defined by

$$rad(x) = x - \mathrm{round}(x/(2 \cdot \pi)) \cdot 2 \cdot \pi$$

The *axis_rad* function is defined by

$$
\begin{aligned}
axis\_rad(x) &= ((1,0), \arcsin(\sin(x))) && \text{if } \cos(x) \geqslant 1/\sqrt{2}\\
&= ((0,1), \arcsin(\cos(x))) && \text{if } \sin(x) > 1/\sqrt{2}\\
&= ((-1,0), \arcsin(\sin(x))) && \text{if } \cos(x) \leqslant -1/\sqrt{2}\\
&= ((0,-1), \arcsin(\cos(x))) && \text{if } \sin(x) < -1/\sqrt{2}
\end{aligned}
$$

The *arc*, angle, function is defined by

$$
\begin{aligned}
arc(x, y) &= -\arccos(x/\sqrt{x^2+y^2}) && \text{if } y < 0\\
&= \arccos(x/\sqrt{x^2+y^2}) && \text{if } y \geqslant 0
\end{aligned}
$$

5.3.9 Operations for radian trigonometric elementary functions

There shall be one radian big-angle parameter:

$big\_angle\_r_F \in F$

It should have the following default value:

$big\_angle\_r_F = r_F^{\lceil p_F/2 \rceil}$

A binding or implementation can include a method to change the value the radian big-angle parameter. This method should only allow the value of this parameter to be set to a value greater than $2 \cdot \pi$ and such that $ulp_F(big\_angle\_r_F) < \pi/1000$.

NOTES

1 Part 1 requires that $p_F \geqslant 2$, but see also A.5.2.0.2 in Part 1.

2 This Part requires that $p_F \geqslant 2 \cdot \max\{1, \lceil \log_{r_F}(2 \cdot \pi) \rceil\}$, in order to allow at least the first two cycles to be in the interval $[-big\_angle\_r_F, big\_angle\_r_F]$.

3 In order to allow $ulp_F(big\_angle\_r_F) < \pi/1000$, $p_F \geqslant 2 + \log_{r_F}(1000)$ should hold.

There shall be two maximum error parameters for radian trigonometric operations:

$max\_error\_sin_F \in F$
$max\_error\_tan_F \in F$

The $max\_error\_sin_F$ parameter shall have a value in the interval $[0.5, 1.5 \cdot rnd\_error_F]$. The $max\_error\_tan_F$ parameter shall have a value in the interval $[max\_error\_sin_F, 2 \cdot rnd\_error_F]$.

5.3.9.1 Radian angle normalisation

The rad^*_F approximation helper function:

$$rad^*_F : \mathcal{R} \to \mathcal{R}$$

$rad^*_F(x)$ returns a close approximation to $rad(x)$ in $\mathcal{R}$, if $|x| \leqslant big\_angle\_r_F$, with maximum error $max\_error\_sin_F$.

The $axis\_rad^*_F$ approximation helper function:

$$axis\_rad^*_F : \mathcal{R} \to \{(1,0),(0,1),(-1,0),(0,-1)\} \times \mathcal{R}$$

$axis\_rad^*_F(x)$ returns a close approximation to $axis\_rad(x)$, if $x \leqslant big\_angle\_r_F$. The approximation consists of that the second part of the result (the offset from the indicated axis) is approximate.

Further requirements on the rad^*_F and $axis\_rad^*_F$ approximation helper functions are:

$$\begin{array}{ll} rad^*_F(x) = x & \text{if } |x| < \pi \\ snd(axis\_rad^*_F(x)) = rad^*_F(x) & \text{if } fst(axis\_rad^*_F(x)) = (1,0) \end{array}$$

The rad_F operation:

$$rad_F : F \to F \cup \{\mathbf{underflow}, \mathbf{absolute\_precision\_underflow}\}$$

$$\begin{array}{lll} rad_F(x) & = trans\_result_F(rad^*_F(x), nearest_F) & \\ & & \text{if } x \in F \text{ and } |x| > fminN_F \text{ and } |x| \leqslant big\_angle\_r_F \\ & = x & \text{if } (x \in F \text{ and } |x| \leqslant fminN_F) \text{ or } x = \mathbf{-0} \\ & & \\ & = \mathbf{absolute\_precision\_underflow}(\mathbf{qNaN}) & \\ & & \text{if } x \in F \text{ and } |x| > big\_angle\_r_F \\ & = result\_NaN_F(x) & \text{otherwise} \end{array}$$

The $axis\_rad_F$ operation:

$$axis\_rad_F : F \to ((F \times F) \times F) \cup \{\mathbf{absolute\_precision\_underflow}\}$$

$$\begin{array}{lll} axis\_rad_F(x) & = (fst(axis\_rad^*_F(x)), trans\_result_F(snd(axis\_rad^*_F(x)), nearest_F)) & \\ & & \text{if } x \in F \text{ and } |x| > fminN_F \text{ and } |x| \leqslant big\_angle\_r_F \\ & = ((1,0), x) & \text{if } (x \in F \text{ and } |x| \leqslant fminN_F) \text{ or } x = \mathbf{-0} \\ & & \\ & = \mathbf{absolute\_precision\_underflow}((\mathbf{qNaN}, \mathbf{qNaN}), \mathbf{qNaN}) & \\ & & \text{if } x \in F \text{ and } |x| > big\_angle\_r_F \\ & = ((\mathbf{qNaN}, \mathbf{qNaN}), \mathbf{qNaN}) & \\ & & \text{if } x \text{ is a quiet NaN} \\ & = \mathbf{invalid}((\mathbf{qNaN}, \mathbf{qNaN}), \mathbf{qNaN}) & \\ & & \text{otherwise} \end{array}$$

NOTE – rad_F is simpler, easier to use, but less accurate than $axis\_rad_F$. The latter may still not be sufficient for implementing the radian trigonometric operations to less than the maximum error stated by the parameters.

5.3.9.2 Radian sine

The sin^*_F approximation helper function:

$$sin^*_F : \mathcal{R} \to \mathcal{R}$$

$sin^*_F(x)$ returns a close approximation to $\sin(x)$ in $\mathcal{R}$ if $|x| \leqslant big\_angle\_r_F$, with maximum error $max\_error\_sin_F$.

Further requirements on the sin^*_F approximation helper function are:

$$\begin{array}{ll}
sin^*_F(n \cdot 2 \cdot \pi + \pi/6) = 1/2 & \text{if } n \in \mathcal{Z} \text{ and } |n \cdot 2 \cdot \pi + \pi/6| \leqslant big\_angle\_r_F \\
sin^*_F(n \cdot 2 \cdot \pi + \pi/2) = 1 & \text{if } n \in \mathcal{Z} \text{ and } |n \cdot 2 \cdot \pi + \pi/2| \leqslant big\_angle\_r_F \\
sin^*_F(n \cdot 2 \cdot \pi + 5 \cdot \pi/6) = 1/2 & \text{if } n \in \mathcal{Z} \text{ and } |n \cdot 2 \cdot \pi + 5 \cdot \pi/6| \leqslant big\_angle\_r_F \\
sin^*_F(x) = x & \text{if } sin^*_F(x) \neq \sin(x) \text{ and } |x| \leqslant \sqrt{3 \cdot epsilon_F / r_F} \\
sin^*_F(-x) = -sin^*_F(x) &
\end{array}$$

The sin_F operation:

$$sin_F : F \to F \cup \{\mathbf{underflow}, \mathbf{absolute\_precision\_underflow}\}$$

$$\begin{array}{lll}
sin_F(x) & = trans\_result_F(sin^*_F(x), nearest_F) & \\
 & & \text{if } x \in F \text{ and } fminN_F < |x| \text{ and } |x| \leqslant big\_angle\_r_F \\
 & = rad_F(x) & \text{otherwise}
\end{array}$$

NOTE – **underflow** is here explicitly avoided for denormal arguments, but the operation may **underflow** for other arguments.

5.3.9.3 Radian cosine

The cos^*_F approximation helper function:

$$cos^*_F : \mathcal{R} \to \mathcal{R}$$

$cos^*_F(x)$ returns a close approximation to $\cos(x)$ in $\mathcal{R}$ if $|x| \leqslant big\_angle\_r_F$, with maximum error $max\_error\_sin_F$.

Further requirements on the cos^*_F approximation helper function are:

$$\begin{array}{ll}
cos^*_F(n \cdot 2 \cdot \pi) = 1 & \text{if } n \in \mathcal{Z} \text{ and } |n \cdot 2 \cdot \pi| \leqslant big\_angle\_r_F \\
cos^*_F(n \cdot 2 \cdot \pi + \pi/3) = 1/2 & \text{if } n \in \mathcal{Z} \text{ and } |n \cdot 2 \cdot \pi + \pi/3| \leqslant big\_angle\_r_F \\
cos^*_F(n \cdot 2 \cdot \pi + 2 \cdot \pi/3) = -1/2 & \text{if } n \in \mathcal{Z} \text{ and } |n \cdot 2 \cdot \pi + 2 \cdot \pi/3| \leqslant big\_angle\_r_F \\
cos^*_F(n \cdot 2 \cdot \pi + \pi) = -1 & \text{if } n \in \mathcal{Z} \text{ and } |n \cdot 2 \cdot \pi + \pi| \leqslant big\_angle\_r_F \\
cos^*_F(x) = 1 & \text{if } cos^*_F(x) \neq \cos(x) \text{ and } |x| < \sqrt{epsilon_F / r_F} \\
cos^*_F(-x) = cos^*_F(x) &
\end{array}$$

The cos_F operation:

$$cos_F : F \to F \cup \{\mathbf{underflow}, \mathbf{absolute\_precision\_underflow}\}$$

$$\begin{array}{lll}
cos_F(x) & = trans\_result_F(cos^*_F(x), nearest_F) & \\
 & & \text{if } x \in F \text{ and } |x| \leqslant big\_angle\_r_F \\
 & = 1 & \text{if } x = \mathbf{-0} \\
 & = rad_F(x) & \text{otherwise}
\end{array}$$

5.3.9.4 Radian tangent

The tan^*_F approximation helper function:

$$tan^*_F : \mathcal{R} \rightarrow \mathcal{R}$$

$tan^*_F(x)$ returns a close approximation to $\tan(x)$ in $\mathcal{R}$ if $|x| \leqslant big\_angle\_r_F$, with maximum error $max\_error\_tan_F$.

Further requirements on the tan^*_F approximation helper function are:

$$\begin{array}{ll}
tan^*_F(n \cdot 2 \cdot \pi + \pi/4) = 1 & \text{if } n \in \mathcal{Z} \text{ and } |n \cdot 2 \cdot \pi + \pi/4| \leqslant big\_angle\_r_F \\
tan^*_F(n \cdot 2 \cdot \pi + 3 \cdot \pi/4) = -1 & \text{if } n \in \mathcal{Z} \text{ and } |n \cdot 2 \cdot \pi + 3 \cdot \pi/4| \leqslant big\_angle\_r_F \\
tan^*_F(x) = x & \text{if } tan^*_F(x) \neq \tan(x) \text{ and } |x| < \sqrt{epsilon_F/r_F} \\
tan^*_F(-x) = -tan^*_F(x) &
\end{array}$$

The tan_F operation:

$$tan_F : F \rightarrow F \cup \{\mathbf{underflow}, \mathbf{overflow}, \mathbf{absolute\_precision\_underflow}\}$$

$$\begin{array}{lll}
tan_F(x) & = trans\_result_F(tan^*_F(x), nearest_F) & \\
 & & \text{if } x \in F \text{ and } fminN_F < |x| \text{ and } |x| \leqslant big\_angle\_r_F \\
 & = rad_F(x) & \text{otherwise}
\end{array}$$

NOTE – **underflow** is explicitly avoided for denormal arguments, but the operation may **underflow** for other arguments.

5.3.9.5 Radian cotangent

The cot^*_F approximation helper function:

$$cot^*_F : \mathcal{R} \rightarrow \mathcal{R}$$

$cot^*_F(x)$ returns a close approximation to $\cot(x)$ in $\mathcal{R}$ if $|x| \leqslant big\_angle\_r_F$, with maximum error $max\_error\_tan_F$.

Further requirements on the cot^*_F approximation helper function are:

$$\begin{array}{ll}
cot^*_F(n \cdot 2 \cdot \pi + \pi/4) = 1 & \text{if } n \in \mathcal{Z} \text{ and } |n \cdot 2 \cdot \pi + \pi/4| \leqslant big\_angle\_r_F \\
cot^*_F(n \cdot 2 \cdot \pi + 3 \cdot \pi/4) = -1 & \text{if } n \in \mathcal{Z} \text{ and } |n \cdot 2 \cdot \pi + 3 \cdot \pi/4| \leqslant big\_angle\_r_F \\
cot^*_F(-x) = -cot^*_F(x) &
\end{array}$$

The cot_F operation:

$$cot_F : F \rightarrow F \cup \{\mathbf{underflow}, \mathbf{overflow}, \mathbf{pole}, \mathbf{absolute\_precision\_underflow}\}$$

$$\begin{array}{lll}
cot_F(x) & = trans\_result_F(cot^*_F(x), nearest_F) & \\
 & & \text{if } x \in F \text{ and } x \neq 0 \text{ and } |x| \leqslant big\_angle\_r_F \\
 & = \mathbf{pole}(\mathbf{+\infty}) & \text{if } x = 0 \\
 & = \mathbf{pole}(\mathbf{-\infty}) & \text{if } x = \mathbf{-0} \\
 & = rad_F(x) & \text{otherwise}
\end{array}$$

5.3.9.6 Radian secant

The sec^*_F approximation helper function:

$$sec^*_F : \mathcal{R} \rightarrow \mathcal{R}$$

$sec^*_F(x)$ returns a close approximation to $\sec(x)$ in $\mathcal{R}$ if $|x| \leqslant big\_angle\_r_F$, with maximum error $max\_error\_tan_F$.

Further requirements on the sec^*_F approximation helper function are:

$$\begin{array}{ll}
sec^*_F(n \cdot 2 \cdot \pi) = 1 & \text{if } n \in \mathcal{Z} \text{ and } |n \cdot 2 \cdot \pi| \leqslant big\_angle\_r_F \\
sec^*_F(n \cdot 2 \cdot \pi + \pi/3) = 2 & \text{if } n \in \mathcal{Z} \text{ and } |n \cdot 2 \cdot \pi + \pi/3| \leqslant big\_angle\_r_F \\
sec^*_F(n \cdot 2 \cdot \pi + 2 \cdot \pi/3) = -2 & \text{if } n \in \mathcal{Z} \text{ and } |n \cdot 2 \cdot \pi + 2 \cdot \pi/3| \leqslant big\_angle\_r_F \\
sec^*_F(n \cdot 2 \cdot \pi + \pi) = -1 & \text{if } n \in \mathcal{Z} \text{ and } |n \cdot 2 \cdot \pi + \pi| \leqslant big\_angle\_r_F \\
sec^*_F(x) = 1 & \text{if } sec^*_F(x) \neq \sec(x) \text{ and } |x| < \sqrt{epsilon_F} \\
sec^*_F(-x) = sec^*_F(x) &
\end{array}$$

The sec_F operation:

$$sec_F : F \rightarrow F \cup \{\mathbf{overflow}, \mathbf{absolute\_precision\_underflow}\}$$

$$\begin{array}{lll}
sec_F(x) & = trans\_result_F(sec^*_F(x), nearest_F) & \\
& & \text{if } x \in F \text{ and } |x| \leqslant big\_angle\_r_F \\
& = 1 & \text{if } x = \mathbf{-0} \\
& = rad_F(x) & \text{otherwise}
\end{array}$$

5.3.9.7 Radian cosecant

The csc^*_F approximation helper function:

$$csc^*_F : \mathcal{R} \rightarrow \mathcal{R}$$

$csc^*_F(x)$ returns a close approximation to $\csc(x)$ in $\mathcal{R}$ if $|x| \leqslant big\_angle\_r_F$, with maximum error $max\_error\_tan_F$.

Further requirements on the csc^*_F approximation helper function are:

$$\begin{array}{ll}
csc^*_F(n \cdot 2 \cdot \pi + \pi/6) = 2 & \text{if } n \in \mathcal{Z} \text{ and } |n \cdot 2 \cdot \pi + \pi/6| \leqslant big\_angle\_r_F \\
csc^*_F(n \cdot 2 \cdot \pi + \pi/2) = 1 & \text{if } n \in \mathcal{Z} \text{ and } |n \cdot 2 \cdot \pi + \pi/2| \leqslant big\_angle\_r_F \\
csc^*_F(n \cdot 2 \cdot \pi + 5 \cdot \pi/6) = 2 & \text{if } n \in \mathcal{Z} \text{ and } |n \cdot 2 \cdot \pi + 5 \cdot \pi/6| \leqslant big\_angle\_r_F \\
csc^*_F(-x) = -csc^*_F(x) &
\end{array}$$

The csc_F operation:

$$csc_F : F \rightarrow F \cup \{\mathbf{overflow}, \mathbf{pole}, \mathbf{absolute\_precision\_underflow}\}$$

$$\begin{array}{lll}
csc_F(x) & = trans\_result_F(csc^*_F(x), nearest_F) & \\
& & \text{if } x \in F \text{ and } x \neq 0 \text{ and } |x| \leqslant big\_angle\_r_F \\
& = \mathbf{pole}(+\infty) & \text{if } x = 0 \\
& = \mathbf{pole}(-\infty) & \text{if } x = \mathbf{-0} \\
& = rad_F(x) & \text{otherwise}
\end{array}$$

5.3.9.8 Radian cosine with sine

$$\begin{array}{ll}
cossin_F : F \rightarrow (F \times F) \cup \{\mathbf{underflow}, \mathbf{absolute\_precision\_underflow}\} & \\
cossin_F(x) \quad = (cos_F(x), sin_F(x)) &
\end{array}$$

5.3.9.9 Radian arc sine

The $arcsin^*_F$ approximation helper function:

$$arcsin^*_F : F \rightarrow \mathcal{R}$$

$arcsin^*_F(x)$ returns a close approximation to $\arcsin(x)$ in $\mathcal{R}$, with maximum error $max\_error\_sin_F$.

Further requirements on the $arcsin^*_F$ approximation helper function are:

$arcsin^*_F(1/2) = \pi/6$
$arcsin^*_F(1) = \pi/2$
$arcsin^*_F(x) = x$ if $arcsin^*_F(x) \neq \arcsin(x)$ and $|x| < \sqrt{2 \cdot epsilon_F/r_F}$
$arcsin^*_F(-x) = -arcsin^*_F(x)$

Range limitation:

$$arcsin^{\#}_F(x) = \max\{up_F(-\pi/2), \min\{arcsin^*_F(x), down_F(\pi/2)\}\}$$

The $arcsin_F$ operation:

$$arcsin_F : F \to F \cup \{\textbf{invalid}\}$$

$$\begin{array}{lll} arcsin_F(x) & = trans\_result_F(arcsin^{\#}_F(x), nearest_F) & \\ & & \text{if } x \in F \text{ and } fminN_F < |x| \leqslant 1 \\ & = x & \text{if } (x \in F \text{ and } |x| \leqslant fminN_F) \text{ or } x = \mathbf{-0} \\ & = result\_NaN_F(x) & \text{otherwise} \end{array}$$

NOTE – **underflow** is explicitly avoided.

5.3.9.10 Radian arc cosine

The $arccos^*_F$ approximation helper function:

$$arccos^*_F : F \to \mathcal{R}$$

$arccos^*_F(x)$ returns a close approximation to $\arccos(x)$ in $\mathcal{R}$, with maximum error $max\_error\_sin_F$.

Further requirements on the $arccos^*_F$ approximation helper function are:

$arccos^*_F(1/2) = \pi/3$
$arccos^*_F(0) = \pi/2$
$arccos^*_F(-1/2) = 2 \cdot \pi/3$
$arccos^*_F(-1) = \pi$

Range limitation:

$$arccos^{\#}_F(x) = \min\{arccos^*_F(x), down_F(\pi)\}$$

The $arccos_F$ operation:

$$arccos_F : F \to F \cup \{\textbf{invalid}\}$$

$$\begin{array}{lll} arccos_F(x) & = trans\_result_F(arccos^{\#}_F(x), nearest_F) & \\ & & \text{if } x \in F \text{ and } -1 \leqslant x \leqslant 1 \\ & = arccos_F(0) & \text{if } x = \mathbf{-0} \\ & = result\_NaN_F(x) & \text{otherwise} \end{array}$$

5.3.9.11 Radian arc tangent

The $arctan^*_F$ approximation helper function:

$$arctan^*_F : F \to \mathcal{R}$$

$arctan^*_F(x)$ returns a close approximation to $\arctan(x)$ in $\mathcal{R}$, with maximum error $max\_error\_tan_F$.

Further requirements on the $arctan^*_F$ approximation helper function are:

$$\begin{aligned}
&\mathit{arctan}^*_F(1) = \pi/4 && \\
&\mathit{arctan}^*_F(x) = x && \text{if } \mathit{arctan}^*_F(x) \neq \arctan(x) \text{ and} \\
& && \quad |x| \leqslant \sqrt{1.5 \cdot \mathit{epsilon}_F / r_F} \\
&\mathit{arctan}^*_F(x) = \pi/2 && \text{if } \mathit{arctan}^*_F(x) \neq \arctan(x) \text{ and } x > 3 \cdot r_F/\mathit{epsilon}_F \\
&\mathit{arctan}^*_F(-x) = -\mathit{arctan}^*_F(x) &&
\end{aligned}$$

Range limitation:

$$\mathit{arctan}^{\#}_F(x) = \max\{\mathit{up}_F(-\pi/2), \min\{\mathit{arctan}^*_F(x), \mathit{down}_F(\pi/2)\}\}$$

The arctan_F operation:

$$\mathit{arctan}_F : F \rightarrow F$$

$$\begin{aligned}
\mathit{arctan}_F(x) \quad &= \mathit{trans\_result}_F(\mathit{arctan}^{\#}_F(x), \mathit{nearest}_F) && \\
& && \text{if } x \in F \text{ and } \mathit{fminN}_F < |x| \\
&= x && \text{if } (x \in F \text{ and } |x| \leqslant \mathit{fminN}_F) \text{ or } x = \mathbf{-0} \\
&= \mathit{up}_F(-\pi/2) && \text{if } x = \mathbf{-\infty} \\
&= \mathit{down}_F(\pi/2) && \text{if } x = \mathbf{+\infty} \\
&= \mathit{result\_NaN}_F(x) && \text{otherwise}
\end{aligned}$$

NOTES

1 $\mathit{arctan}_F(x) \approx \mathit{arc}_F(1, x)$

2 **underflow** is explicitly avoided.

5.3.9.12 Radian arc cotangent

This clause specifies two inverse cotangent operations. One approximating the continuous (but not sign symmetric) arccot, the other approximating the sign symmetric (but discontinuous at 0) arcctg.

The arccot^*_F approximation helper function:

$$\mathit{arccot}^*_F : F \rightarrow \mathcal{R}$$

$\mathit{arccot}^*_F(x)$ returns a close approximation to $\operatorname{arccot}(x)$ in $\mathcal{R}$, with maximum error $\mathit{max\_error\_tan}_F$.

The arcctg^*_F approximation helper function:

$$\mathit{arcctg}^*_F : F \rightarrow \mathcal{R}$$

$\mathit{arcctg}^*_F(x)$ returns a close approximation to $\operatorname{arcctg}(x)$ in $\mathcal{R}$, with maximum error $\mathit{max\_error\_tan}_F$.

Further requirements on the arccot^*_F and arcctg^*_F approximation helper functions are:

$$\begin{aligned}
&\mathit{arccot}^*_F(1) = \pi/4 && \\
&\mathit{arccot}^*_F(0) = \pi/2 && \\
&\mathit{arccot}^*_F(-1) = 3 \cdot \pi/4 && \\
&\mathit{arccot}^*_F(x) = \pi && \text{if } \mathit{arccot}^*_F(x) \neq \operatorname{arccot}(x) \text{ and } x < -3 \cdot r_F/\mathit{epsilon}_F \\
& && \\
&\mathit{arcctg}^*_F(x) = \mathit{arccot}^*_F(x) && \text{if } x \geqslant 0 \\
&\mathit{arcctg}^*_F(-x) = -\mathit{arcctg}^*_F(x) &&
\end{aligned}$$

Range limitation:

$$\begin{aligned}
&\mathit{arccot}^{\#}_F(x) = \min\{\mathit{arccot}^*_F(x), \mathit{down}_F(\pi)\} \\
&\mathit{arcctg}^{\#}_F(x) = \max\{\mathit{up}_F(-\pi/2), \min\{\mathit{arcctg}^*_F(x), \mathit{down}_F(\pi/2)\}\}
\end{aligned}$$

The arccot_F operation:

$$\mathit{arccot}_F : F \rightarrow F \cup \{\textbf{underflow}\}$$

$$
\begin{array}{lll}
arccot_F(x) & = trans\_result_F(arccot^{\#}_F(x)) & \\
 & & \text{if } x \in F \\
 & = nearest_F(\pi/2) & \text{if } x = \mathbf{-0} \\
 & = down_F(\pi) & \text{if } x = \mathbf{-\infty} \\
 & = 0 & \text{if } x = \mathbf{+\infty} \\
 & = result\_NaN_F(x) & \text{otherwise}
\end{array}
$$

NOTES

1 $arccot_F(x) \approx arc_F(x, 1)$.

2 There is no “jump” at zero for $arccot_F$.

The $arcctg_F$ operation:

$arcctg_F : F \to F \cup \{\textbf{underflow}\}$

$$
\begin{array}{lll}
arcctg_F(x) & = trans\_result_F(arcctg^{\#}_F(x), nearest_F) & \\
 & & \text{if } x \in F \\
 & = up_F(-\pi/2) & \text{if } x = \mathbf{-0} \\
 & = \mathbf{-0} & \text{if } x = \mathbf{-\infty} \\
 & = 0 & \text{if } x = \mathbf{+\infty} \\
 & = result\_NaN_F(x) & \text{otherwise}
\end{array}
$$

NOTE 3 – $arcctg_F(neg_F(x)) = neg_F(arcctg_F(x))$.

5.3.9.13 Radian arc secant

The $arcsec^*_F$ approximation helper function:

$arcsec^*_F : F \to \mathcal{R}$

$arcsec^*_F(x)$ returns a close approximation to $\text{arcsec}(x)$ in $\mathcal{R}$, with maximum error $max\_error\_tan_F$.

Further requirements on the $arcsec^*_F$ approximation helper function are:

$$
\begin{array}{ll}
arcsec^*_F(2) = \pi/3 & \\
arcsec^*_F(-2) = 2 \cdot \pi/3 & \\
arcsec^*_F(-1) = \pi & \\
arcsec^*_F(x) \leqslant \pi/2 & \text{if } x > 0 \\
arcsec^*_F(x) \geqslant \pi/2 & \text{if } x < 0 \\
arcsec^*_F(x) = \pi/2 & \text{if } arcsec^*_F(x) \neq \text{arcsec}(x) \text{ and } |x| > 3 \cdot r_F/epsilon_F
\end{array}
$$

Range limitation:

$$
\begin{array}{lll}
arcsec^{\#}_F(x) & = \min\{arcsec^*_F(x), down_F(\pi/2)\} & \\
 & & \text{if } x \geqslant 1 \\
 & = \max\{up_F(\pi/2), \min\{arcsec^*_F(x), down_F(\pi)\}\} & \\
 & & \text{if } x \leqslant -1
\end{array}
$$

The $arcsec_F$ operation:

$arcsec_F : F \to F \cup \{\textbf{invalid}\}$

$$
\begin{array}{lll}
arcsec_F(x) & = trans\_result_F(arcsec^{\#}_F(x), nearest_F) & \\
 & & \text{if } x \in F \text{ and } |x| \geqslant 1 \\
 & = up_F(\pi/2) & \text{if } x = \mathbf{-\infty} \\
 & = down_F(\pi/2) & \text{if } x = \mathbf{+\infty} \\
 & = result\_NaN_F(x) & \text{otherwise}
\end{array}
$$

5.3.9.14 Radian arc cosecant

The $arccsc^*_F$ approximation helper function:

$$arccsc^*_F : F \to \mathcal{R}$$

$arccsc^*_F(x)$ returns a close approximation to $\operatorname{arccsc}(x)$ in $\mathcal{R}$, with maximum error $max\_error\_tan_F$.

Further requirements on the $arccsc^*_F$ approximation helper function are:

$$arccsc^*_F(2) = \pi/6$$
$$arccsc^*_F(1) = \pi/2$$
$$arccsc^*_F(-x) = -arccsc^*_F(x)$$

Range limitation:

$$arccsc^\#_F(x) = \max\{up_F(-\pi/2), \min\{arccsc^*_F(x), down_F(\pi/2)\}\}$$

The $arccsc_F$ operation:

$$arccsc_F : F \to F \cup \{\mathbf{underflow}, \mathbf{invalid}\}$$

$$\begin{array}{lll}
arccsc_F(x) & = trans\_result_F(arccsc^\#_F(x), nearest_F) & \\
 & & \text{if } x \in F \text{ and } |x| \geqslant 1 \\
 & = \mathbf{-0} & \text{if } x = \mathbf{-\infty} \\
 & = 0 & \text{if } x = \mathbf{+\infty} \\
 & = result\_NaN_F(x) & \text{otherwise}
\end{array}$$

5.3.9.15 Radian angle from Cartesian co-ordinates

The arc^*_F approximation helper function:

$$arc^*_F : F \times F \to \mathcal{R}$$

$arc^*_F(x, y)$ returns a close approximation to $arc(x, y)$ in $\mathcal{R}$, with maximum error $max\_error\_tan_F$.

NOTE – The arc operations are often called `arctan2` (with the co-ordinate arguments swapped), or `arccot2`.

Further requirements on the arc^*_F approximation helper function are:

$$\begin{array}{ll}
arc^*_F(x, 0) = 0 & \text{if } x > 0 \\
arc^*_F(x, x) = \pi/4 & \text{if } x > 0 \\
arc^*_F(0, y) = \pi/2 & \text{if } y > 0 \\
arc^*_F(x, -x) = 3 \cdot \pi/4 & \text{if } x < 0 \\
arc^*_F(x, 0) = \pi & \text{if } x < 0 \\
arc^*_F(x, -y) = -arc^*_F(x, y) & \text{if } y \neq 0 \text{ or } x > 0
\end{array}$$

Range limitation:

$$arc^\#_F(x, y) = \max\{up_F(-\pi), \min\{arc^*_F(x, y), down_F(\pi)\}\}$$

The arc_F operation:

$$arc_F : F \times F \to F \cup \{\mathbf{underflow}\}$$

$$\begin{array}{lll}
arc_F(x, y) & = trans\_result_F(arc^\#_F(x, y), nearest_F) & \\
 & & \text{if } x, y \in F \text{ and } (x \neq 0 \text{ or } y \neq 0) \\
 & = 0 & \text{if } x = 0 \text{ and } y = 0 \\
 & = down_F(\pi) & \text{if } x = \mathbf{-0} \text{ and } y = 0 \\
 & = arc_F(0, y) & \text{if } x = \mathbf{-0} \text{ and } y \in F \cup \{\mathbf{-\infty}, \mathbf{+\infty}\} \text{ and } y \neq 0 \\
 & = neg_F(arc_F(x, 0)) & \text{if } y = \mathbf{-0} \text{ and } x \in F \cup \{\mathbf{-\infty}, \mathbf{-0}, \mathbf{+\infty}\} \\
 & & \\
 & = 0 & \text{if } x = \mathbf{+\infty} \text{ and } y \in F \text{ and } y \geqslant 0 \\
 & = \mathbf{-0} & \text{if } x = \mathbf{+\infty} \text{ and } y \in F \text{ and } y < 0
\end{array}$$

$$
\begin{array}{ll}
= nearest_F(\pi/4) & \text{if } x = +\infty \text{ and } y = +\infty \\
= nearest_F(\pi/2) & \text{if } x \in F \text{ and } y = +\infty \\
= nearest_F(3 \cdot \pi/4) & \text{if } x = -\infty \text{ and } y = +\infty \\
= down_F(\pi) & \text{if } x = -\infty \text{ and } y \in F \text{ and } y \geqslant 0 \\
= up_F(-\pi) & \text{if } x = -\infty \text{ and } y \in F \text{ and } y < 0 \\
= nearest_F(-3 \cdot \pi/4) & \text{if } x = -\infty \text{ and } y = -\infty \\
= nearest_F(-\pi/2) & \text{if } x \in F \text{ and } y = -\infty \\
= nearest_F(-\pi/4) & \text{if } x = +\infty \text{ and } y = -\infty \\
\\
= result\_NaN2_F(x, y) & \text{otherwise}
\end{array}
$$

5.3.10 Operations for trigonometrics with given angular unit

There shall be one big-angle parameter for argument angular-unit trigonometric operations:

$big\_angle\_u_F \in F$

It should have the following default value:

$big\_angle\_u_F = \lceil r_F^{\lceil p_F/2 \rceil}/6 \rceil$

A binding or implementation can include a method to change the value for this parameter. This method should only allow the value of this parameter to be set to a value greater than or equal to 1 and such that $ulp_F(big\_angle_F) \leqslant 1/2000$.

NOTE 1 – In order to allow $ulp_F(big\_angle_F) \leqslant 1/2000$, $p_F \geqslant 2 + \log_{r_F}(1000)$ should hold.

There shall be a derived parameter signifying the minimum allowed angular unit:

$min\_angular\_unit_F = r_F \cdot fminN_F/epsilon_F$

NOTE 2 – That is,$min\_angular\_unit_F = r_F^{(emin_F - 1 + p_F)}$

To make the specifications below a bit easier to express, let

$G_F = \{x \in F \quad | \quad min\_angular\_unit_F \leqslant |x|\}$.

Let $T = \{1, 2, 360, 400, 6400\}$. T consists of angle values for exactly one revolution for some common non-radian angular units: cycles, half-cycles, arc degrees, grades, and mils.

There shall be two parameterised maximum error parameters for argument angular-unit trigonometric operations:

$max\_error\_sinu_F : F \rightarrow F \cup \{\textbf{invalid}\}$
$max\_error\_tanu_F : F \rightarrow F \cup \{\textbf{invalid}\}$

For $u \in G_F$, the $max\_error\_sinu_F(u)$ parameter shall have a value in the interval $[max\_error\_sin_F, 2]$. The $max\_error\_sinu_F(u)$ parameter shall have the value of $max\_error\_sin_F$ if $|u| \in T$. For $u \in G_F$, the $max\_error\_tanu_F(u)$ parameter shall have a value in the interval $[max\_error\_tan_F, 4]$. The $max\_error\_tanu_F(u)$ parameter shall have the value of $max\_error\_tan_F$ if $|u| \in T$. The $max\_error\_sinu_F(u)$ and $max\_error\_tanu_F(u)$ parameters return **invalid** if $u \notin G_F$.

5.3.10.1 Argument angular-unit angle normalisation

The argument angular-unit normalisation computes exactly $rad(2 \cdot \pi \cdot x/u) \cdot u/(2 \cdot \pi)$, where x is the angular value, and u is the angular unit.

The $cycle_F$ operation:

$cycle_F : F \times F \rightarrow F \cup \{\mathbf{-0}, \textbf{absolute\_precision\_underflow}, \textbf{invalid}\}$

$$\begin{array}{lll}
cycle_F(u,x) & = remr_F(x,u) & \text{if } u \in G_F \text{ and } (x = \mathbf{-0} \text{ or} \\
 & & \quad (x \in F \text{ and } |x/u| \leqslant big\_angle\_u_F)) \\
 & = \mathbf{absolute\_precision\_underflow}(\mathbf{qNaN}) & \\
 & & \text{if } u \in G_F \text{ and } x \in F \text{ and } |x/u| > big\_angle\_u_F \\
 & = result\_NaN2_F(u,x) & \text{otherwise}
\end{array}$$

The $axis\_cycle_F$ operation:

$$axis\_cycle_F : F \times F \rightarrow ((F \times F) \times (F \cup \{\mathbf{-0}\})) \cup \{\mathbf{absolute\_precision\_underflow}, \mathbf{invalid}\}$$

$$\begin{array}{lll}
axis\_cycle_F(u,x) & & \\
 & = (axis(u,x), result_F(x - (\mathrm{round}(x \cdot 4/u) \cdot u/4), rnd_F)) & \\
 & & \text{if } u \in G_F \text{ and } x \in F \text{ and } |x/u| \leqslant big\_angle\_u_F \text{ and} \\
 & & \quad (x/u \geqslant 0 \text{ or } x - (\mathrm{round}(x \cdot 4/u) \cdot u/4) \neq 0) \\
 & = (axis(u,x), \mathbf{-0}) & \text{if } u \in G_F \text{ and } x \in F \text{ and } |x/u| \leqslant big\_angle\_u_F \text{ and} \\
 & & \quad x/u < 0 \text{ and } x - (\mathrm{round}(x \cdot 4/u) \cdot u/4) = 0 \text{ and} \\
 & = ((1,0), \mathbf{-0}) & \text{if } u \in G_F \text{ and } x = \mathbf{-0} \\
 & = \mathbf{absolute\_precision\_underflow}((\mathbf{qNaN}, \mathbf{qNaN}), \mathbf{qNaN}) & \\
 & & \text{if } u \in G_F \text{ and } x \in F \text{ and } |x/u| > big\_angle\_u_F \\
 & = ((\mathbf{qNaN}, \mathbf{qNaN}), \mathbf{qNaN}) & \\
 & & \text{if } x \text{ is a quiet NaN and } u \text{ is not a signalling NaN} \\
 & = ((\mathbf{qNaN}, \mathbf{qNaN}), \mathbf{qNaN}) & \\
 & & \text{if } u \text{ is a quiet NaN and } x \text{ is not a signalling NaN} \\
 & = \mathbf{invalid}((\mathbf{qNaN}, \mathbf{qNaN}), \mathbf{qNaN}) & \\
 & & \text{otherwise}
\end{array}$$

where

$$\begin{array}{lll}
axis(u,x) & = (1,0) & \text{if } \mathrm{round}(x \cdot 4/u) = 4 \cdot n \\
 & = (0,1) & \text{if } \mathrm{round}(x \cdot 4/u) = 4 \cdot n + 1 \\
 & = (-1,0) & \text{if } \mathrm{round}(x \cdot 4/u) = 4 \cdot n + 2 \\
 & = (0,-1) & \text{if } \mathrm{round}(x \cdot 4/u) = 4 \cdot n + 3
\end{array}$$

for some $n \in \mathcal{Z}$.

NOTES

1 $axis\_cycle_F(u,x)$ is exact when $div_F(u,4) = u/4$.

2 $cycle_F$ is an exact operation.

3 $cycle_F(u,x)$ is $\mathbf{-0}$ or has a result in the interval $[-|u/2|, |u/2|]$ if there is no notification.

4 A zero resulting angle is negative if the original angle value is negative.

5 The $cycle_F$ operation is used also in the specifications of the unit argument trigonometric operations. This does *not* imply that the implementation has to use the cycle operation, when implementing the operations. Just that the *results* (including notifications) must be *as if* it did.

5.3.10.2 Argument angular-unit sine

The $sinu^*_F$ approximation helper function:

$$sinu^*_F : F \times \mathcal{R} \rightarrow \mathcal{R}$$

$sinu^*_F(u,x)$ returns a close approximation to $\sin(x \cdot 2 \cdot \pi/u)$ in $\mathcal{R}$ if $u \neq 0$, with maximum error $max\_error\_sinu_F(u)$.

Further requirements on the $sinu^*_F$ approximation helper function:

$$
\begin{array}{ll}
\mathit{sinu}^*_F(u, n \cdot u + x) = \mathit{sinu}^*_F(u, x) & \text{if } n \in \mathcal{Z} \text{ and } u \in F \text{ and } u \neq 0 \\
\mathit{sinu}^*_F(u, u/12) = 1/2 & \text{if } u \in F \text{ and } u \neq 0 \\
\mathit{sinu}^*_F(u, u/4) = 1 & \text{if } u \in F \text{ and } u \neq 0 \\
\mathit{sinu}^*_F(u, 5 \cdot u/12) = 1/2 & \text{if } u \in F \text{ and } u \neq 0 \\
\mathit{sinu}^*_F(u, -x) = -\mathit{sinu}^*_F(u, x) & \text{if } u \in F \text{ and } u \neq 0 \\
\mathit{sinu}^*_F(-u, x) = -\mathit{sinu}^*_F(u, x) & \text{if } u \in F \text{ and } u \neq 0
\end{array}
$$

NOTE – $\mathit{sinu}^*_F(u, x) \approx x \cdot 2 \cdot \pi/u$ if $|x \cdot 2 \cdot \pi/u| < \mathit{fminN}_F$.

The sinu_F operation:

$$\mathit{sinu}_F : F \times F \to F \cup \{\mathbf{-0}, \mathbf{underflow}, \mathbf{invalid}, \mathbf{absolute\_precision\_underflow}\}$$

$$
\begin{array}{lll}
\mathit{sinu}_F(u, x) & = \mathit{trans\_result}_F(\mathit{sinu}^*_F(u, x), \mathit{nearest}_F) & \\
 & & \text{if } \mathit{cycle}_F(u, x) \in F \text{ and } \mathit{cycle}_F(u, x) \not\in \{-u/2, 0, u/2\} \\
 & = \mathit{div}_F(0, u) & \text{if } \mathit{cycle}_F(u, x) \in \{0, u/2\} \\
 & = \mathit{div}_F(\mathbf{-0}, u) & \text{if } \mathit{cycle}_F(u, x) \in \{-u/2, \mathbf{-0}\} \\
 & = \mathit{cycle}_F(u, x) & \text{otherwise}
\end{array}
$$

5.3.10.3 Argument angular-unit cosine

The cosu^*_F approximation helper function:

$$\mathit{cosu}^*_F : F \times \mathcal{R} \to \mathcal{R}$$

$\mathit{cosu}^*_F(u, x)$ returns a close approximation to $\cos(x \cdot 2 \cdot \pi/u)$ in $\mathcal{R}$ if $u \neq 0$, with maximum error $\mathit{max\_error\_sinu}_F(u)$.

Further requirements on the cosu^*_F approximation helper function:

$$
\begin{array}{ll}
\mathit{cosu}^*_F(u, n \cdot u + x) = \mathit{cosu}^*_F(u, x) & \text{if } n \in \mathcal{Z} \text{ and } u \in F \text{ and } u \neq 0 \\
\mathit{cosu}^*_F(u, 0) = 1 & \text{if } u \in F \text{ and } u \neq 0 \\
\mathit{cosu}^*_F(u, u/6) = 1/2 & \text{if } u \in F \text{ and } u \neq 0 \\
\mathit{cosu}^*_F(u, u/3) = -1/2 & \text{if } u \in F \text{ and } u \neq 0 \\
\mathit{cosu}^*_F(u, u/2) = -1 & \text{if } u \in F \text{ and } u \neq 0 \\
\mathit{cosu}^*_F(u, -x) = \mathit{cosu}^*_F(u, x) & \text{if } u \in F \text{ and } u \neq 0 \\
\mathit{cosu}^*_F(-u, x) = \mathit{cosu}^*_F(u, x) & \text{if } u \in F \text{ and } u \neq 0
\end{array}
$$

NOTE – $\mathit{cosu}^*_F(u, x) = 1$ should hold if $|x \cdot 2 \cdot \pi/u| < \sqrt{\mathit{epsilon}_F/r_F}$

The cosu_F operation:

$$\mathit{cosu}_F : F \times F \to F \cup \{\mathbf{underflow}, \mathbf{invalid}, \mathbf{absolute\_precision\_underflow}\}$$

$$
\begin{array}{lll}
\mathit{cosu}_F(u, x) & = \mathit{trans\_result}_F(\mathit{cosu}^*_F(u, x), \mathit{nearest}_F) & \\
 & & \text{if } \mathit{cycle}_F(u, x) \in F \\
 & = 1 & \text{if } \mathit{cycle}_F(u, x) = \mathbf{-0} \\
 & = \mathit{cycle}_F(u, x) & \text{otherwise}
\end{array}
$$

5.3.10.4 Argument angular-unit tangent

The tanu^*_F approximation helper function:

$$\mathit{tanu}^*_F : F \times \mathcal{R} \to \mathcal{R}$$

$\mathit{tanu}^*_F(u, x)$ returns a close approximation to $\tan(x \cdot 2 \cdot \pi/u)$ in $\mathcal{R}$ if $u \neq 0$, with maximum error $\mathit{max\_error\_tanu}_F(u)$.

Further requirements on the tanu^*_F approximation helper function:

$$\begin{array}{ll}
tanu^*_F(u, n \cdot u + x) = tanu^*_F(u, x) & \text{if } n \in \mathcal{Z} \text{ and } u \in F \text{ and } u \neq 0 \\
tanu^*_F(u, u/8) = 1 & \text{if } u \in F \text{ and } u \neq 0 \\
tanu^*_F(u, 3 \cdot u/8) = -1 & \text{if } u \in F \text{ and } u \neq 0 \\
tanu^*_F(u, -x) = -tanu^*_F(u, x) & \text{if } u \in F \text{ and } u \neq 0 \\
tanu^*_F(-u, x) = -tanu^*_F(u, x) & \text{if } u \in F \text{ and } u \neq 0
\end{array}$$

NOTE 1 – $tanu^*_F(u, x) \approx x \cdot 2 \cdot \pi/u$ if $|x \cdot 2 \cdot \pi/u| < fminN_F$.

The $tanu_F$ operation:

$$\begin{array}{ll}
tanu_F : F \times F \to F \cup \{\mathbf{-0}, \mathbf{pole}, \mathbf{overflow}, \mathbf{underflow}, \mathbf{invalid}, \\
\qquad\qquad \mathbf{absolute\_precision\_underflow}\}
\end{array}$$

$$\begin{array}{lll}
tanu_F(u, x) & = trans\_result_F(tanu^*_F(u, x), nearest_F) & \\
& & \text{if } cycle_F(u, x) \in F \text{ and} \\
& & \quad cycle_F(u, x) \notin \{-u/2, -u/4, 0, u/4, u/2\} \\
& = div_F(0, u) & \text{if } cycle_F(u, x) \in \{-u/2, 0\} \\
& = div_F(\mathbf{-0}, u) & \text{if } cycle_F(u, x) \in \{\mathbf{-0}, u/2\} \\
& = \mathbf{pole}(+\infty) & \text{if } cycle_F(u, x) = u/4 \\
& = \mathbf{pole}(-\infty) & \text{if } cycle_F(u, x) = -u/4 \\
& = cycle_F(u, x) & \text{otherwise}
\end{array}$$

NOTE 2 – The **pole** notification can arise for $tanu_F(u, x)$ only when $u/4$ is in F.

5.3.10.5 Argument angular-unit cotangent

The $cotu^*_F$ approximation helper function:

$$cotu^*_F : F \times \mathcal{R} \to \mathcal{R}$$

$cotu^*_F(u, x)$ returns a close approximation to $\cot(x \cdot 2 \cdot \pi/u)$ in $\mathcal{R}$ if $u \neq 0$, with maximum error $max\_error\_tanu_F(u)$.

Further requirements on the $cotu^*_F$ approximation helper function:

$$\begin{array}{ll}
cotu^*_F(u, n \cdot u + x) = cotu^*_F(u, x) & \text{if } n \in \mathcal{Z} \text{ and } u \in F \text{ and } u \neq 0 \\
cotu^*_F(u, u/8) = 1 & \text{if } u \in F \text{ and } u \neq 0 \\
cotu^*_F(u, 3 \cdot u/8) = -1 & \text{if } u \in F \text{ and } u \neq 0 \\
cotu^*_F(u, -x) = -cotu^*_F(u, x) & \text{if } u \in F \text{ and } u \neq 0 \\
cotu^*_F(-u, x) = -cotu^*_F(u, x) & \text{if } u \in F \text{ and } u \neq 0
\end{array}$$

The $cotu_F$ operation:

$$\begin{array}{ll}
cotu_F : F \times F \to F \cup \{\mathbf{-0}, \mathbf{pole}, \mathbf{overflow}, \mathbf{underflow}, \mathbf{invalid}, \\
\qquad\qquad \mathbf{absolute\_precision\_underflow}\}
\end{array}$$

$$\begin{array}{lll}
cotu_F(u, x) & = trans\_result_F(cotu^*_F(u, x), nearest_F) & \\
& & \text{if } cycle_F(u, x) \in F \text{ and} \\
& & \quad cycle_F(u, x) \notin \{-u/2, -u/4, 0, u/2\} \\
& = \mathbf{-0} & \text{if } cycle_F(u, x) = -u/4 \\
& = div_F(u, tanu_F(u, x)) & \text{if } cycle_F(u, x) \in \{-u/2, \mathbf{-0}, 0, u/2\} \\
& = cycle_F(u, x) & \text{otherwise}
\end{array}$$

5.3.10.6 Argument angular-unit secant

The $secu^*_F$ approximation helper function:

$$secu^*_F : F \times \mathcal{R} \to \mathcal{R}$$

$secu^*_F(u, x)$ returns a close approximation to $\sec(x \cdot 2 \cdot \pi/u)$ in $\mathcal{R}$ if $u \neq 0$, with maximum error $max\_error\_tanu_F(u)$.

Further requirements on the $secu^*_F$ approximation helper function:

$$\begin{array}{ll}
secu^*_F(u, n \cdot u + x) = secu^*_F(u, x) & \text{if } n \in \mathcal{Z} \text{ and } u \in F \text{ and } u \neq 0 \\
secu^*_F(u, 0) = 1 & \text{if } u \in F \text{ and } u \neq 0 \\
secu^*_F(u, u/6) = 2 & \text{if } u \in F \text{ and } u \neq 0 \\
secu^*_F(u, u/3) = -2 & \text{if } u \in F \text{ and } u \neq 0 \\
secu^*_F(u, u/2) = -1 & \text{if } u \in F \text{ and } u \neq 0 \\
secu^*_F(u, -x) = secu^*_F(u, x) & \text{if } u \in F \text{ and } u \neq 0 \\
secu^*_F(-u, x) = secu^*_F(u, x) & \text{if } u \in F \text{ and } u \neq 0 \\
secu^*_F(u, x) = 1 & \text{if } |x \cdot 2 \cdot \pi/u| < 0.5 \cdot \sqrt{epsilon_F}
\end{array}$$

The $secu_F$ operation:

$$secu_F : F \times F \rightarrow F \cup \{\mathbf{pole}, \mathbf{overflow}, \mathbf{invalid}, \mathbf{absolute\_precision\_underflow}\}$$

$$\begin{array}{lll}
secu_F(u, x) & = trans\_result_F(secu^*_F(u, x), nearest_F) & \\
 & & \text{if } cycle_F(u, x) \in F \text{ and } cycle_F(u, x) \notin \{-u/4, u/4\} \\
 & = div_F(1, cosu_F(u, x)) & \text{if } cycle_F(u, x) \in \{-u/4, \mathbf{-0}, u/4\} \\
 & = cycle_F(u, x) & \text{otherwise}
\end{array}$$

5.3.10.7 Argument angular-unit cosecant

The $cscu^*_F$ approximation helper function:

$$cscu^*_F : F \times \mathcal{R} \rightarrow \mathcal{R}$$

$cscu^*_F(u, x)$ returns a close approximation to $\csc(x \cdot 2 \cdot \pi/u)$ in $\mathcal{R}$ if $u \neq 0$, with maximum error $max\_error\_tanu_F(u)$.

Further requirements on the $cscu^*_F$ approximation helper function:

$$\begin{array}{ll}
cscu^*_F(u, n \cdot u + x) = cscu^*_F(u, x) & \text{if } n \in \mathcal{Z} \text{ and } u \in F \text{ and } u \in 0 \\
cscu^*_F(u, u/12) = 2 & \text{if } u \in F \text{ and } u \neq 0 \\
cscu^*_F(u, u/4) = 1 & \text{if } u \in F \text{ and } u \neq 0 \\
cscu^*_F(u, 5 \cdot u/12) = 2 & \text{if } u \in F \text{ and } u \neq 0 \\
cscu^*_F(u, -x) = -cscu^*_F(u, x) & \text{if } u \in F \text{ and } u \neq 0 \\
cscu^*_F(-u, x) = -cscu^*_F(u, x) & \text{if } u \in F \text{ and } u \neq 0
\end{array}$$

The $cscu_F$ operation:

$$cscu_F : F \times F \rightarrow F \cup \{\mathbf{pole}, \mathbf{overflow}, \mathbf{invalid}, \mathbf{absolute\_precision\_underflow}\}$$

$$\begin{array}{lll}
cscu_F(u, x) & = trans\_result_F(cscu^*_F(u, x), nearest_F) & \\
 & & \text{if } cycle_F(u, x) \in F \text{ and } cycle_F(u, x) \notin \{-u/2, 0, u/2\} \\
 & = div_F(1, sinu_F(u, x)) & \text{if } cycle_F(u, x) \in \{-u/2, \mathbf{-0}, 0, u/2\} \\
 & = cycle_F(u, x) & \text{otherwise}
\end{array}$$

5.3.10.8 Argument angular-unit cosine with sine

$$cossinu_F : F \times F \rightarrow (F \times (F \cup \{\mathbf{-0}\})) \cup \{\mathbf{underflow}, \mathbf{invalid}, \mathbf{absolute\_precision\_underflow}\}$$

$$cossinu_F(u, x) \quad = (cosu_F(u, x), sinu_F(u, x))$$

5.3.10.9 Argument angular-unit arc sine

The $arcsinu^*_F$ approximation helper function:

$$arcsinu^*_F : F \times F \to \mathcal{R}$$

$arcsinu^*_F(u, x)$ returns a close approximation to $\arcsin(x) \cdot u/(2 \cdot \pi)$ in $\mathcal{R}$, with maximum error $max\_error\_sinu_F(u)$.

Further requirements on the $arcsinu^*_F$ approximation helper function:

$$arcsinu^*_F(u, 1/2) = u/12$$
$$arcsinu^*_F(u, 1) = u/4$$
$$arcsinu^*_F(u, -x) = -arcsinu^*_F(u, x)$$
$$arcsinu^*_F(-u, x) = -arcsinu^*_F(u, x)$$

NOTE – $arcsinu^*_F(u, x) \approx u/(2 \cdot \pi)$ if $|x| < fminN_F$.

Range limitation:

$$arcsinu^\#_F(u, x) = \max\{up_F(-|u/4|), \min\{arcsinu^*_F(u, x), down_F(|u/4|)\}\}$$

The $arcsinu_F$ operation:

$$arcsinu_F : F \times F \to F \cup \{\mathbf{-0}, \mathbf{underflow}, \mathbf{invalid}\}$$

$$\begin{aligned} arcsinu_F(u, x) &= trans\_result_F(arcsinu^\#_F(u, x), nearest_F) \\ &\qquad\qquad \text{if } u \in G_F \text{ and } x \in F \text{ and } |x| \leqslant 1 \text{ and } x \neq 0 \\ &= mul_F(u, x) \qquad \text{if } u \in G_F \text{ and } x \in \{\mathbf{-0}, 0\} \\ &= result\_NaN2_F(u, x) \qquad \text{otherwise} \end{aligned}$$

5.3.10.10 Argument angular-unit arc cosine

The $arccosu^*_F$ approximation helper function:

$$arccosu^*_F : F \times F \to \mathcal{R}$$

$arccosu^*_F(u, x)$ returns a close approximation to $\arccos(x) \cdot u/(2 \cdot \pi)$ in $\mathcal{R}$, with maximum error $max\_error\_sinu_F(u)$.

Further requirements on the $arccosu^*_F$ approximation helper function:

$$arccosu^*_F(u, 1/2) = u/6$$
$$arccosu^*_F(u, 0) = u/4$$
$$arccosu^*_F(u, -1/2) = u/3$$
$$arccosu^*_F(u, -1) = u/2$$
$$arccosu^*_F(-u, x) = -arccosu^*_F(u, x)$$

Range limitation:

$$arccosu^\#_F(u, x) = \max\{up_F(-|u/2|), \min\{arccosu^*_F(u, x), down_F(|u/2|)\}\}$$

The $arccosu_F$ operation:

$$arccosu_F : F \times F \to F \cup \{\mathbf{underflow}, \mathbf{invalid}\}$$

$$\begin{aligned} arccosu_F(u, x) &= trans\_result_F(arccosu^\#_F(u, x), nearest_F) \\ &\qquad\qquad \text{if } u \in G_F \text{ and } x \in F \text{ and } |x| \leqslant 1 \\ &= nearest_F(u/4) \qquad \text{if } u \in G_F \text{ and } x = \mathbf{-0} \\ &= result\_NaN2_F(u, x) \qquad \text{otherwise} \end{aligned}$$

5.3.10.11 Argument angular-unit arc tangent

The $arctanu^*_F$ approximation helper function:

$$arctanu^*_F : F \times F \to \mathcal{R}$$

$arctanu^*_F(u, x)$ returns a close approximation to $\arctan(x) \cdot u/(2 \cdot \pi)$ in $\mathcal{R}$, with maximum error $max\_error\_tanu_F(u)$.

Further requirements on the $arctanu^*_F$ approximation helper function:

$$\begin{array}{ll} arctanu^*_F(u, 1) = u/8 & \\ arctanu^*_F(u, x) = u/4 & \text{if } arctanu^*_F(u, x) \neq \arctan(x) \cdot u/(2 \cdot \pi) \text{ and} \\ & \quad x > 3 \cdot r_F/epsilon_F \\ arctanu^*_F(u, -x) = -arctanu^*_F(u, x) & \\ arctanu^*_F(-u, x) = -arctanu^*_F(u, x) & \end{array}$$

NOTE 1 – $arctanu^*_F(u, x) \approx u/(2 \cdot \pi)$ if $|x| < fminN_F$

Range limitation:

$$arctanu^\#_F(u, x) = \max\{up_F(-|u/4|), \min\{arctanu^*_F(u, x), down_F(|u/4|)\}\}$$

The $arctanu_F$ operation:

$$arctanu_F : F \times F \to F \cup \{\mathbf{-0}, \mathbf{invalid}, \mathbf{underflow}\}$$

$$\begin{array}{lll} arctanu_F(u, x) & = trans\_result_F(arctanu^\#_F(u, x), nearest_F) & \\ & & \text{if } u \in G_F \text{ and } x \in F \text{ and } x \neq 0 \\ & = mul_F(x, u) & \text{if } u \in G_F \text{ and } x \in \{\mathbf{-0}, 0\} \\ & = up_F(-u/4) & \text{if } u \in G_F \text{ and } x = \mathbf{-\infty} \text{ and } u > 0 \\ & = down_F(u/4) & \text{if } u \in G_F \text{ and } x = \mathbf{+\infty} \text{ and } u > 0 \\ & = down_F(-u/4) & \text{if } u \in G_F \text{ and } x = \mathbf{-\infty} \text{ and } u < 0 \\ & = up_F(u/4) & \text{if } u \in G_F \text{ and } x = \mathbf{+\infty} \text{ and } u < 0 \\ & = result\_NaN2_F(u, x) & \text{otherwise} \end{array}$$

NOTE 2 – $arctanu_F(u, x) \approx arcu_F(u, 1, x)$.

5.3.10.12 Argument angular-unit arc cotangent

This clause specifies two inverse cotangent operations. One approximating the continuous (but not sign symmetric) arccot, the other approximating the sign symmetric (but discontinuous at 0) arcctg.

The $arccotu^*_F$ approximation helper function:

$$arccotu^*_F : F \times F \to \mathcal{R}$$

$arccotu^*_F(u, x)$ returns a close approximation to $\text{arccot}(x) \cdot u/(2 \cdot \pi)$ in $\mathcal{R}$, with maximum error $max\_error\_tanu_F(u)$.

The $arcctgu^*_F$ approximation helper function:

$$arcctgu^*_F : F \times F \to \mathcal{R}$$

$arcctgu^*_F(u, x)$ returns a close approximation to $\text{arcctg}(x) \cdot u/(2 \cdot \pi)$ in $\mathcal{R}$, with maximum error $max\_error\_tanu_F(u)$.

Further requirements on the $arccotu^*_F$ and $arcctgu^*_F$ approximation helper functions:

$$\begin{array}{ll}
arccotu^*_F(u, 1) = u/8 & \\
arccotu^*_F(u, 0) = u/4 & \\
arccotu^*_F(u, -1) = 3 \cdot u/8 & \\
arccotu^*_F(u, x) \leqslant u/2 & \text{if } u > 0 \\
arccotu^*_F(u, x) \geqslant u/2 & \text{if } u < 0 \\
arccotu^*_F(u, x) = u/2 & \text{if } arccotu^*_F(u, x) \neq \operatorname{arccot}(x) \cdot u/(2 \cdot \pi) \text{ and} \\
 & \quad x < -3 \cdot r_F/epsilon_F \\
arccotu^*_F(-u, x) = -arccotu^*_F(u, x) & \\
arcctgu^*_F(u, x) = arccotu^*_F(u, x) & \text{if } x \geqslant 0 \\
arcctgu^*_F(u, -x) = -arcctgu^*_F(u, x) &
\end{array}$$

Range limitation:

$$arccotu^{\#}_F(u, x) = \max\{up_F(-|u/2|), \min\{arccotu^*_F(u, x), down_F(|u/2|)\}\}$$
$$arcctgu^{\#}_F(u, x) = \max\{up_F(-|u/4|), \min\{arcctgu^*_F(u, x), down_F(|u/4|)\}\}$$

The $arccotu_F$ operation:

$$arccotu_F : F \times F \to F \cup \{\textbf{invalid}, \textbf{underflow}\}$$

$$\begin{array}{lll}
arccotu_F(u, x) & = trans\_result_F(arccotu^{\#}_F(u, x), nearest_F) & \\
 & & \text{if } u \in G_F \text{ and } x \in F \\
 & = nearest_F(u/4) & \text{if } u \in G_F \text{ and } x = \mathbf{-0} \\
 & = down_F(u/2) & \text{if } u \in G_F \text{ and } x = \mathbf{-\infty} \text{ and } u > 0 \\
 & = up_F(u/2) & \text{if } u \in G_F \text{ and } x = \mathbf{-\infty} \text{ and } u < 0 \\
 & = div_F(u, x) & \text{if } u \in G_F \text{ and } x = \mathbf{+\infty} \\
 & = result\_NaN2_F(u, x) & \text{otherwise}
\end{array}$$

NOTE – $arccotu_F(u, x) \approx arcu_F(u, x, 1)$.

The $arcctgu_F$ operation:

$$arcctgu_F : F \times F \to F \cup \{\textbf{invalid}, \textbf{underflow}\}$$

$$\begin{array}{lll}
arcctgu_F(u, x) & = trans\_result_F(arcctgu^{\#}_F(u, x), nearest_F) & \\
 & & \text{if } u \in G_F \text{ and } x \in F \\
 & = neg_F(arcctgu_F(u, 0)) & \text{if } u \in G_F \text{ and } x = \mathbf{-0} \\
 & = div_F(u, x) & \text{if } u \in G_F \text{ and } x \in \{\mathbf{-\infty}, \mathbf{+\infty}\} \\
 & = result\_NaN2_F(u, x) & \text{otherwise}
\end{array}$$

5.3.10.13 Argument angular-unit arc secant

The $arcsecu^*_F$ approximation helper function:

$$arcsecu^*_F : F \times F \to \mathcal{R}$$

$arcsecu^*_F(u, x)$ returns a close approximation to $\operatorname{arcsec}(x) \cdot u/(2 \cdot \pi)$ in $\mathcal{R}$, with maximum error $max\_error\_tanu_F(u)$.

Further requirements on the $arcsecu^*_F$ approximation helper function:

$$\begin{array}{ll}
arcsecu^*_F(u, 2) = u/6 & \\
arcsecu^*_F(u, -2) = u/3 & \\
arcsecu^*_F(u, -1) = u/2 & \\
arcsecu^*_F(u, x) \leqslant u/4 & \text{if } x > 0 \text{ and } u > 0 \\
arcsecu^*_F(u, x) \geqslant u/4 & \text{if } x < 0 \text{ and } u > 0 \\
arcsecu^*_F(u, x) = u/4 & \text{if } arcsecu^*_F(u, x) \neq \operatorname{arcsec}(x) \cdot u/(2 \cdot \pi) \text{ and} \\
 & \quad |x| > 3 \cdot r_F/epsilon_F \\
arcsecu^*_F(-u, x) = -arcsecu^*_F(u, x) &
\end{array}$$

Range limitation:

$$\begin{aligned}
arcsecu^{\#}_F(u, x) &= \max\{up_F(-|u/4|), \min\{arcsecu^*_F(u, x), down_F(|u/4|)\}\} && \text{if } x \geqslant 1\\
&= \max\{up_F(u/4), \min\{arcsecu^*_F(u, x), down_F(u/2)\}\} && \text{if } x \leqslant -1 \text{ and } u > 0\\
&= \max\{up_F(u/2), \min\{arcsecu^*_F(u, x), down_F(u/4)\}\} && \text{if } x \leqslant -1 \text{ and } u < 0
\end{aligned}$$

The $arcsecu_F$ operation:

$$arcsecu_F : F \times F \to F \cup \{\mathbf{underflow}, \mathbf{invalid}\}$$

$$\begin{aligned}
arcsecu_F(u, x) &= trans\_result_F(arcsecu^{\#}_F(u, x), nearest_F)\\
& && \text{if } u \in G_F \text{ and } x \in F \text{ and } (x \leqslant -1 \text{ or } x \geqslant 1)\\
&= down_F(u/4) && \text{if } u \in G_F \text{ and } x = \mathbf{-\infty} \text{ and } u > 0\\
&= up_F(u/4) && \text{if } u \in G_F \text{ and } x = \mathbf{+\infty} \text{ and } u > 0\\
&= up_F(u/4) && \text{if } u \in G_F \text{ and } x = \mathbf{-\infty} \text{ and } u < 0\\
&= down_F(u/4) && \text{if } u \in G_F \text{ and } x = \mathbf{+\infty} \text{ and } u < 0\\
&= result\_NaN2_F(u, x) && \text{otherwise}
\end{aligned}$$

5.3.10.14 Argument angular-unit arc cosecant

The $arccscu^*_F$ approximation helper function:

$$arccscu^*_F : F \times F \to \mathcal{R}$$

$arccscu^*_F(u, x)$ returns a close approximation to $\mathrm{arccsc}(x) \cdot u/(2 \cdot \pi)$ in $\mathcal{R}$, with maximum error $max\_error\_tanu_F(u)$.

Further requirements on the $arccscu^*_F$ approximation helper function:

$$\begin{aligned}
&arccscu^*_F(u, 2) = u/12\\
&arccscu^*_F(u, 1) = u/4\\
&arccscu^*_F(u, -x) = -arccscu^*_F(u, x)\\
&arccscu^*_F(-u, x) = -arccscu^*_F(u, x)
\end{aligned}$$

Range limitation:

$$arccscu^{\#}_F(u, x) = \max\{up_F(-|u/4|), \min\{arccscu^*_F(u, x), down_F(|u/4|)\}\}$$

The $arccscu_F$ operation:

$$arccscu_F : F \times F \to F \cup \{\mathbf{underflow}, \mathbf{invalid}\}$$

$$\begin{aligned}
arccscu_F(u, x) &= trans\_result_F(arccscu^{\#}_F(u, x), nearest_F)\\
& && \text{if } u \in G_F \text{ and } x \in F \text{ and } (x \geqslant 1 \text{ or } x \leqslant -1)\\
&= mul_F(-u, 0) && \text{if } u \in G_F \text{ and } x = \mathbf{-\infty}\\
&= mul_F(u, 0) && \text{if } u \in G_F \text{ and } x = \mathbf{+\infty}\\
&= result\_NaN2_F(u, x) && \text{otherwise}
\end{aligned}$$

5.3.10.15 Argument angular-unit angle from Cartesian co-ordinates

The $arcu^*_F$ approximation helper function:

$$arcu^*_F : F \times F \times F \to \mathcal{R}$$

$arcu^*_F(u, x, y)$ returns a close approximation to $arc(x, y) \cdot u/(2 \cdot \pi)$ in $\mathcal{R}$, with maximum error $max\_error\_tanu_F(u)$.

Further requirements on the $arcu^*_F$ approximation helper function:

$$\begin{array}{ll}
arcu^*_F(u,x,x) = u/8 & \text{if } x > 0 \\
arcu^*_F(u,0,y) = u/4 & \text{if } y > 0 \\
arcu^*_F(u,x,-x) = 3 \cdot u/8 & \text{if } x < 0 \\
arcu^*_F(u,x,0) = u/2 & \text{if } x < 0 \\
arcu^*_F(u,x,-y) = -arcu^*_F(u,x,y) & \text{if } y \neq 0 \text{ or } x > 0 \\
arcu^*_F(-u,x,y) = -arcu^*_F(u,x,y) &
\end{array}$$

Range limitation:

$$arcu^{\#}_F(u,x,y) = \max\{up_F(-|u/2|), \min\{arcu^*_F(u,x,y), down_F(|u/2|)\}\}$$

The $arcu_F$ operation:

$$arcu_F : F \times F \times F \to F \cup \{\mathbf{-0}, \mathbf{underflow}, \mathbf{invalid}\}$$

$$\begin{array}{lll}
arcu_F(u,x,y) & = trans\_result_F(arcu^{\#}_F(u,x,y), nearest_F) & \\
& & \text{if } u \in G_F \text{ and } x, y \in F \text{ and } (x < 0 \text{ or } y \neq 0) \\
& = mul_F(u,0) & \text{if } u \in G_F \text{ and } x \in F \text{ and } x > 0 \text{ and } y = 0 \\
& = 0 & \text{if } u \in G_F \text{ and } x = 0 \text{ and } y = 0 \\
& = down_F(u/2) & \text{if } u \in G_F \text{ and } x = \mathbf{-0} \text{ and } y = 0 \text{ and } u > 0 \\
& = up_F(u/2) & \text{if } u \in G_F \text{ and } x = \mathbf{-0} \text{ and } y = 0 \text{ and } u < 0 \\
& = arcu_F(u,0,y) & \text{if } u \in G_F \text{ and } x = \mathbf{-0} \text{ and } y \in F \cup \{\mathbf{-\infty}, \mathbf{+\infty}\} \text{ and} \\
& & \quad y \neq 0 \\
& = neg_F(arcu_F(u,x,0)) & \text{if } u \in G_F \text{ and } y = \mathbf{-0} \text{ and } x \in F \cup \{\mathbf{-\infty}, \mathbf{-0}, \mathbf{+\infty}\} \\
& & \\
& = mul_F(0,u) & \text{if } u \in G_F \text{ and } x = \mathbf{+\infty} \text{ and } y \in F \text{ and } y \geqslant 0 \\
& = mul_F(0,-u) & \text{if } u \in G_F \text{ and } x = \mathbf{+\infty} \text{ and } y \in F \text{ and } y < 0 \\
& = nearest_F(u/8) & \text{if } u \in G_F \text{ and } x = \mathbf{+\infty} \text{ and } y = \mathbf{+\infty} \\
& = nearest_F(u/4) & \text{if } u \in G_F \text{ and } x \in F \text{ and } y = \mathbf{+\infty} \\
& = nearest_F(3 \cdot u/8) & \text{if } u \in G_F \text{ and } x = \mathbf{-\infty} \text{ and } y = \mathbf{+\infty} \\
& = down_F(u/2) & \text{if } u \in G_F \text{ and } x = \mathbf{-\infty} \text{ and } y \in F \text{ and} \\
& & \quad y \geqslant 0 \text{ and } u > 0 \\
& = up_F(-u/2) & \text{if } u \in G_F \text{ and } x = \mathbf{-\infty} \text{ and } y \in F \text{ and} \\
& & \quad y < 0 \text{ and } u > 0 \\
& = up_F(u/2) & \text{if } u \in G_F \text{ and } x = \mathbf{-\infty} \text{ and } y \in F \text{ and} \\
& & \quad y > 0 \text{ and } u < 0 \\
& = down_F(-u/2) & \text{if } u \in G_F \text{ and } x = \mathbf{-\infty} \text{ and } y \in F \text{ and} \\
& & \quad y \leqslant 0 \text{ and } u < 0 \\
& = nearest_F(-3 \cdot u/8) & \text{if } u \in G_F \text{ and } x = \mathbf{-\infty} \text{ and } y = \mathbf{-\infty} \\
& = nearest_F(-u/4) & \text{if } u \in G_F \text{ and } x \in F \text{ and } y = \mathbf{-\infty} \\
& = nearest_F(-u/8) & \text{if } u \in G_F \text{ and } x = \mathbf{+\infty} \text{ and } y = \mathbf{-\infty} \\
& & \\
& = result\_NaN3_F(u,x,y) & \text{otherwise}
\end{array}$$

5.3.11 Operations for angular-unit conversions

5.3.11.1 Converting radian angle to argument angular-unit angle

Define the mathematical function:

$$rad\_to\_cycle : \mathcal{R} \times \mathcal{R} \to \mathcal{R}$$

$$\begin{array}{ll}
rad\_to\_cycle(x,v) & \\
\quad = \arccos(\cos(x)) \cdot v/(2 \cdot \pi) & \\
& \text{if } \sin(x) \geqslant 0 \text{ and } v \neq 0 \\
\quad = -\arccos(\cos(x)) \cdot v/(2 \cdot \pi) &
\end{array}$$

$$\text{if } \sin(x) < 0 \text{ and } v \neq 0$$

The $rad\_to\_cycle^*_F$ approximation helper function:

$$rad\_to\_cycle^*_F : \mathcal{R} \times F \rightarrow \mathcal{R}$$

$rad\_to\_cycle^*_F(x, v)$ returns a close approximation to $rad\_to\_cycle(x, v)$ in $\mathcal{R}$, with maximum error $max\_error\_sin_F$, if $|x| \leqslant big\_angle\_r_F$.

Further requirements on the $rad\_to\_cycle^*_F$ approximation helper function are:

$$\begin{aligned}
&rad\_to\_cycle^*_F(n \cdot 2 \cdot \pi + \pi/6, v) = v/12 && \text{if } n \in \mathcal{Z} \text{ and } |n \cdot 2 \cdot \pi + \pi/6| \leqslant big\_angle\_r_F\\
&rad\_to\_cycle^*_F(n \cdot 2 \cdot \pi + \pi/4, v) = v/8 && \text{if } n \in \mathcal{Z} \text{ and } |n \cdot 2 \cdot \pi + \pi/4| \leqslant big\_angle\_r_F\\
&rad\_to\_cycle^*_F(n \cdot 2 \cdot \pi + \pi/3, v) = v/6 && \text{if } n \in \mathcal{Z} \text{ and } |n \cdot 2 \cdot \pi + \pi/3| \leqslant big\_angle\_r_F\\
&rad\_to\_cycle^*_F(n \cdot 2 \cdot \pi + \pi/2, v) = v/4 && \text{if } n \in \mathcal{Z} \text{ and } |n \cdot 2 \cdot \pi + \pi/2| \leqslant big\_angle\_r_F\\
&rad\_to\_cycle^*_F(n \cdot 2 \cdot \pi + 2 \cdot \pi/3, v) = v/3 &&\\
& && \text{if } n \in \mathcal{Z} \text{ and } |n \cdot 2 \cdot \pi + 2 \cdot \pi/3| \leqslant big\_angle\_r_F\\
&rad\_to\_cycle^*_F(n \cdot 2 \cdot \pi + 3 \cdot \pi/4, v) = 3 \cdot v/8 &&\\
& && \text{if } n \in \mathcal{Z} \text{ and } |n \cdot 2 \cdot \pi + 3 \cdot \pi/4| \leqslant big\_angle\_r_F\\
&rad\_to\_cycle^*_F(n \cdot 2 \cdot \pi + 5 \cdot \pi/6, v) = 5 \cdot v/12 &&\\
& && \text{if } n \in \mathcal{Z} \text{ and } |n \cdot 2 \cdot \pi + 5 \cdot \pi/6| \leqslant big\_angle\_r_F\\
&rad\_to\_cycle^*_F(n \cdot 2 \cdot \pi + \pi, v) = v/2 && \text{if } n \in \mathcal{Z} \text{ and } |n \cdot 2 \cdot \pi + \pi| \leqslant big\_angle\_r_F\\
&rad\_to\_cycle^*_F(-x, v) = -rad\_to\_cycle^*_F(x, v) &&\\
& && \text{if } rad\_to\_cycle(x, v) \neq v/2\\
&rad\_to\_cycle^*_F(x, -v) = -rad\_to\_cycle^*_F(x, v) &&\\
& && \text{if } rad\_to\_cycle(x, v) \neq v/2
\end{aligned}$$

The $rad\_to\_cycle_F$ operation:

$$rad\_to\_cycle_F : F \times F \rightarrow F \cup \{\mathbf{underflow}, \mathbf{absolute\_precision\_underflow}, \mathbf{invalid}\}$$

$$\begin{aligned}
&rad\_to\_cycle_F(x, v)\\
&\qquad = trans\_result_F(rad\_to\_cycle^*_F(x, v), nearest_F)\\
&\qquad\qquad \text{if } v \in G_F \text{ and } x \in F \text{ and } |x| \leqslant big\_angle\_r_F \text{ and } x \neq 0\\
&\qquad = mul_F(v, x) \qquad \text{if } v \in G_F \text{ and } x \in \{\mathbf{-0}, 0\}\\
&\qquad = \mathbf{absolute\_precision\_underflow}(\mathbf{qNaN})\\
&\qquad\qquad \text{if } v \in G_F \text{ and } x \in F \text{ and } |x| > big\_angle\_r_F\\
&\qquad = result\_NaN2_F(x, v) \qquad \text{otherwise}
\end{aligned}$$

5.3.11.2 Converting argument angular-unit angle to radian angle

Define the mathematical function:

$$cycle\_to\_rad : \mathcal{R} \times \mathcal{R} \rightarrow \mathcal{R}$$

$$\begin{aligned}
&cycle\_to\_rad(u, x)\\
&\qquad = \arccos(\cos(x \cdot 2 \cdot \pi/u)) \quad \text{if } \sin(x \cdot 2 \cdot \pi/u) \geqslant 0\\
&\qquad = -\arccos(\cos(x \cdot 2 \cdot \pi/u))\\
&\qquad\qquad \text{if } \sin(x \cdot 2 \cdot \pi/u) < 0
\end{aligned}$$

The $cycle\_to\_rad^*_F$ approximation helper function:

$$cycle\_to\_rad^*_F : F \times \mathcal{R} \rightarrow \mathcal{R}$$

$cycle\_to\_rad^*_F(u, x)$ returns a close approximation to $cycle\_to\_rad(u, x)$ in $\mathcal{R}$, if $u \neq 0$, with maximum error $max\_error\_sin_F$.

Further requirements on the $cycle\_to\_rad^*_F$ approximation helper function are:

$$
\begin{aligned}
&\mathit{cycle\_to\_rad}^*_F(u, n \cdot u + x) = \mathit{cycle\_to\_rad}^*_F(u, x) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{if } n \in \mathcal{Z} \\
&\mathit{cycle\_to\_rad}^*_F(u, u/12) = \pi/6 \\
&\mathit{cycle\_to\_rad}^*_F(u, u/8) = \pi/4 \\
&\mathit{cycle\_to\_rad}^*_F(u, u/6) = \pi/3 \\
&\mathit{cycle\_to\_rad}^*_F(u, u/4) = \pi/2 \\
&\mathit{cycle\_to\_rad}^*_F(u, u/3) = 2 \cdot \pi/3 \\
&\mathit{cycle\_to\_rad}^*_F(u, 3 \cdot u/8) = 3 \cdot \pi/4 \\
&\mathit{cycle\_to\_rad}^*_F(u, 5 \cdot u/12) = 5 \cdot \pi/6 \\
&\mathit{cycle\_to\_rad}^*_F(u, u/2) = \pi \\
&\mathit{cycle\_to\_rad}^*_F(u, -x) = -\mathit{cycle\_to\_rad}^*_F(u, x) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{if } \mathit{cycle\_to\_rad}(u, x) \neq \pi
\end{aligned}
$$

The $\mathit{cycle\_to\_rad}_F$ operation:

$$\mathit{cycle\_to\_rad}_F : F \times F \to F \cup \{\mathbf{-0}, \mathbf{underflow}, \mathbf{absolute\_precision\_underflow}, \mathbf{invalid}\}$$

$$
\begin{aligned}
&\mathit{cycle\_to\_rad}_F(u, x) \\
&\qquad = \mathit{trans\_result}_F(\mathit{cycle\_to\_rad}^*_F(u, x), \mathit{nearest}_F) \\
&\qquad\qquad\qquad\qquad\qquad \text{if } \mathit{cycle}_F(u, x) \in F \text{ and } \mathit{cycle}_F(u, x) \neq 0 \\
&\qquad = \mathit{mul}_F(\mathit{cycle}_F(u, x), u) \quad \text{if } \mathit{cycle}_F(u, x) \in \{\mathbf{-0}, 0\} \\
&\qquad = \mathit{cycle}_F(u, x) \qquad\qquad\quad \text{otherwise}
\end{aligned}
$$

5.3.11.3 Converting argument angular-unit angle to (another) argument angular-unit angle

Define the mathematical function:

$$\mathit{cycle\_to\_cycle} : \mathcal{R} \times \mathcal{R} \times \mathcal{R} \to \mathcal{R}$$

$$
\begin{aligned}
&\mathit{cycle\_to\_cycle}(u, x, v) \\
&\qquad = \arccos(\cos(x \cdot 2 \cdot \pi/u)) \cdot v/(2 \cdot \pi) \\
&\qquad\qquad\qquad\qquad\qquad \text{if } u \neq 0 \text{ and } v \neq 0 \text{ and } \sin(x \cdot 2 \cdot \pi/u) \geqslant 0 \\
&\qquad = -\arccos(\cos(x \cdot 2 \cdot \pi/u)) \cdot v/(2 \cdot \pi) \\
&\qquad\qquad\qquad\qquad\qquad \text{if } u \neq 0 \text{ and } v \neq 0 \text{ and } \sin(x \cdot 2 \cdot \pi/u) < 0
\end{aligned}
$$

The $\mathit{cycle\_to\_cycle}^*_F$ approximation helper function:

$$\mathit{cycle\_to\_cycle}^*_F : F \times \mathcal{R} \times F \to \mathcal{R}$$

$\mathit{cycle\_to\_cycle}^*_F(u, x, v)$ returns a close approximation to $\mathit{cycle\_to\_cycle}(u, x, v)$ in $\mathcal{R}$ if $u \neq 0$ and $|x/u| \leqslant \mathit{big\_angle\_u}_F$, with maximum error $\mathit{max\_error\_sin}_F$.

Further requirements on the $\mathit{cycle\_to\_cycle}^*_F$ approximation helper function are:

$$
\begin{aligned}
&\mathit{cycle\_to\_cycle}^*_F(u, n \cdot u + x, v) = \mathit{cycle\_to\_cycle}^*_F(u, x, v) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{if } n \in \mathcal{Z} \\
&\mathit{cycle\_to\_cycle}^*_F(u, u/12, v) = v/12 \\
&\mathit{cycle\_to\_cycle}^*_F(u, u/8, v) = v/8 \\
&\mathit{cycle\_to\_cycle}^*_F(u, u/6, v) = v/6 \\
&\mathit{cycle\_to\_cycle}^*_F(u, u/4, v) = v/4 \\
&\mathit{cycle\_to\_cycle}^*_F(u, u/3, v) = v/3 \\
&\mathit{cycle\_to\_cycle}^*_F(u, 3 \cdot u/8, v) = 3 \cdot v/8 \\
&\mathit{cycle\_to\_cycle}^*_F(u, 5 \cdot u/12, v) = 5 \cdot v/12 \\
&\mathit{cycle\_to\_cycle}^*_F(u, u/2, v) = v/2 \\
&\mathit{cycle\_to\_cycle}^*_F(u, -x, v) = -\mathit{cycle\_to\_cycle}^*_F(u, x, v) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{if } \mathit{cycle\_to\_cycle}(u, x, v) \neq v/2 \\
&\mathit{cycle\_to\_cycle}^*_F(-u, x, v) = -\mathit{cycle\_to\_cycle}^*_F(u, x, v)
\end{aligned}
$$

$$\begin{aligned}
& && \text{if } \mathit{cycle\_to\_cycle}(u,x,v) \neq v/2 \\
\mathit{cycle\_to\_cycle}^*_F(u,x,-v) &= -\mathit{cycle\_to\_cycle}^*_F(u,x,v) && \\
& && \text{if } \mathit{cycle\_to\_cycle}(u,x,v) \neq v/2
\end{aligned}$$

The $\mathit{cycle\_to\_cycle}_F$ operation:

$$\mathit{cycle\_to\_cycle}_F : F \times F \times F \to F \cup \{\mathbf{-0}, \mathbf{underflow}, \mathbf{absolute\_precision\_underflow}, \mathbf{invalid}\}$$

$$\begin{aligned}
&\mathit{cycle\_to\_cycle}_F(u,x,v) \\
&\quad = \mathit{trans\_result}_F(\mathit{cycle\_to\_cycle}^*_F(u,x,v), \mathit{nearest}_F) \\
&\quad\qquad\qquad \text{if } v \in G_F \text{ and } \mathit{cycle}_F(u,x) \in F \text{ and } \mathit{cycle}_F(u,x) \neq 0 \\
&\quad = \mathit{mul}_F(v, \mathit{cycle}_F(u,x)) \quad \text{if } v \in G_F \text{ and } \mathit{cycle}_F(u,x) \in \{\mathbf{-0}, 0\} \\
&\quad = \mathbf{absolute\_precision\_underflow}(\mathbf{qNaN}) \\
&\quad\qquad\qquad \text{if } v \in G_F \text{ and} \\
&\quad\qquad\qquad\quad \mathit{cycle}_F(u,x) = \mathbf{absolute\_precision\_underflow} \\
&\quad = \mathit{result\_NaN3}_F(u,x,v) \quad \text{otherwise}
\end{aligned}$$

5.4 Conversion operations

Numeric conversion between different representation forms for integer and fractional values can take place under a number of different circumstances. E.g.:

a) explicit or implicit conversion between different numeric datatypes conforming to Part 1;

b) explicit or implicit conversion between different numeric datatypes only one of which conforms to Part 1;

c) explicit or implicit conversion between a character string and a numeric datatype.

The latter includes outputting a numeric value as a character string, inputting a numeric value from a character string source, and converting a numeral in the source program to a value in a numeric datatype (see 5.5). This Part covers only the cases where at least one of the source and target is a numeric datatype conforming to Part 1.

When a character string is involved as either source or target of a conversion, this Part does not specify the lexical syntax for the numerals parsed or formed. A binding standard should specify the lexical syntax or syntaxes for these numerals, and, when appropriate, how the lexical syntax for the numerals can be altered. With the exception of the radix used in numerals expressing fractional values, differences in lexical syntactic details that do not affect the value in $\mathcal{R}$ denoted by the numerals should not affect the result of the conversion.

Character string representations for integer values can include representations for $\mathbf{-0}$, $\mathbf{+\infty}$, $\mathbf{-\infty}$, and quiet NaNs. Character string representations for floating point and fixed point values should have formats for $\mathbf{-0}$, $\mathbf{+\infty}$, $\mathbf{-\infty}$, and quiet NaNs. For both integer and floating point values, character strings that are not numerals nor special values according to the lexical syntax used, shall be regarded as signalling NaNs when used as source of a numerical conversion.

For the cases where one of the datatypes involved in the conversion does not conform to Part 1, the values of some numeric datatype parameters need to be inferred. For integers, one need to infer the value for *bounded*, and if that is **true** then also values for *maxint* and *minint*. For floating point values, one need to infer the values for r, p, and *emax* or *emin*. In case a precise determination is not possible, values that are 'safe' for that instance should be used. 'Safe' values for otherwise undetermined inferred parameters are such that

a) monotonicity of the conversion function is not affected,

b) the error in the conversion does not exceed that specified by the maximum error parameter (see below),

c) if the value resulting from the conversion is converted back to the source datatype by a conversion conforming to this Part, the original value should be regenerated if possible, and

d) overflow and underflow are avoided if possible.

If, and only if, a specified infinite special value result cannot be represented in the target datatype, the infinity result shall be interpreted as **pole**. If, and only if, a specified NaN special value result cannot be represented in the target datatype, the NaN result shall be interpreted as **invalid**.

5.4.1 Integer to integer conversions

Let I and I' be non-special value sets for integer datatypes. At least one of the datatypes corresponding to I and I' conforms to Part 1.

$$convert_{I \to I'} : I \to I' \cup \{\mathbf{overflow}\}$$

$$\begin{array}{lll}
convert_{I \to I'}(x) & = result_{I'}(x) & \text{if } x \in I \\
 & = x & \text{if } x \in \{\mathbf{-\infty}, \mathbf{-0}, \mathbf{+\infty}\} \\
 & = \mathbf{qNaN} & \text{if } x \text{ is a quiet NaN} \\
 & = \mathbf{invalid}(\mathbf{qNaN}) & \text{if } x \text{ is a signalling NaN}
\end{array}$$

NOTE – If both are I and I' are conforming to Part 1, then this conversion is covered by Part 1. This operation generalises the $cvt_{I \to I'}$ of Part 1, since only one of the integer datatypes in the conversion need be conforming to Part 1.

5.4.2 Floating point to integer conversions

Let I be the non-special value set for an integer datatype conforming to Part 1. Let F be the non-special value set for a floating point datatype conforming to Part 1.

NOTE – The operations in this clause are more specific than the floating point to integer conversion in Part 1 which allows any rounding.

$$rounding_{F \to I} : F \to I \cup \{\mathbf{-0}, \mathbf{overflow}\}$$

$$\begin{array}{lll}
rounding_{F \to I}(x) & & \\
 & = result_I(\mathrm{round}(x)) & \text{if } x \in F \text{ and } (x \geqslant 0 \text{ or } \mathrm{round}(x) \neq 0) \\
 & = \mathbf{-0} & \text{if } x \in F \text{ and } x < 0 \text{ and } \mathrm{round}(x) = 0 \\
 & = x & \text{if } x \in \{\mathbf{-\infty}, \mathbf{-0}, \mathbf{+\infty}\} \\
 & = \mathbf{qNaN} & \text{if } x \text{ is a quiet NaN} \\
 & = \mathbf{invalid}(\mathbf{qNaN}) & \text{if } x \text{ is a signalling NaN}
\end{array}$$

$$floor_{F \to I} : F \to I \cup \{\mathbf{overflow}\}$$

$$\begin{array}{lll}
floor_{F \to I}(x) & = result_I(\lfloor x \rfloor) & \text{if } x \in F \\
 & = x & \text{if } x \in \{\mathbf{-\infty}, \mathbf{-0}, \mathbf{+\infty}\} \\
 & = \mathbf{qNaN} & \text{if } x \text{ is a quiet NaN} \\
 & = \mathbf{invalid}(\mathbf{qNaN}) & \text{if } x \text{ is a signalling NaN}
\end{array}$$

$$ceiling_{F \to I} : F \to I \cup \{\mathbf{-0}, \mathbf{overflow}\}$$

$$\begin{array}{lll}
ceiling_{F \to I}(x) & = result_I(\lceil x \rceil) & \text{if } x \in F \text{ and } (x \geqslant 0 \text{ or } \lceil x \rceil \neq 0) \\
 & = \mathbf{-0} & \text{if } x \in F \text{ and } x < 0 \text{ and } \lceil x \rceil = 0 \\
 & = x & \text{if } x \in \{\mathbf{-\infty}, \mathbf{-0}, \mathbf{+\infty}\} \\
 & = \mathbf{qNaN} & \text{if } x \text{ is a quiet NaN} \\
 & = \mathbf{invalid}(\mathbf{qNaN}) & \text{if } x \text{ is a signalling NaN}
\end{array}$$

5.4.3 Integer to floating point conversions

Let I be the non-special value set for an integer datatype. Let F be the non-special value set for a floating point datatype. At least one of the source and target datatypes is conforming to Part 1.

$$convert_{I \to F} : I \to F \cup \{\mathbf{overflow}\}$$

$$\begin{aligned}
convert_{I \to F}(x) &= result_F(x, nearest_F) && \text{if } x \in I \\
&= x && \text{if } x \in \{\mathbf{-\infty}, \mathbf{-0}, \mathbf{+\infty}\} \\
&= \mathbf{qNaN} && \text{if } x \text{ is a quiet NaN} \\
&= \mathbf{invalid}(\mathbf{qNaN}) && \text{if } x \text{ is a signalling NaN}
\end{aligned}$$

NOTE – When both I and F conform to Part 1, integer to nearest floating point conversions are covered by Part 1. In this case the operations $cvt_{I \to F}$ and $convert_{I \to F}$ are identical.

5.4.4 Floating point to floating point conversions

Define the least radix function, lb, defined for arguments that are greater than 0:

$$lb : \mathcal{Z} \to \mathcal{Z}$$

$$lb(r) = \min\{n \in \mathcal{Z} \mid n \geqslant 1 \text{ and } \exists m \in \mathcal{Z} : r = n^m\}$$

Let F, F', and F'' be non-special value sets for floating point datatypes. At least one of the source and target datatypes in the conversion conforms to Part 1.

There shall be a $max\_error\_convert_{F'}$ parameter that gives the maximum error when converting from F to F' and $lb(r_F) \neq lb(r_{F'})$. The $max\_error\_convert_{F'}$ parameter shall have a value in the interval $[0.5, 0.75]$. If $lb(r_F) = lb(r_{F'})$, the maximum error shall be 0.5 ulp when converting from F to F', but this is not reflected in any parameter.

The $convert^*_{F \to F'}$ approximation helper functions:

$$convert^*_{F \to F'} : \mathcal{R} \to \mathcal{R}$$

$convert^*_{F \to F'}(x)$ returns a close approximation to x in $\mathcal{R}$, with maximum error $max\_error\_convert_{F'}$.

Further requirements on the $convert^*_{F \to F'}$ approximation helper functions:

$$\begin{aligned}
&convert^*_{F \to F'}(x) = x && \text{if } x \in \mathcal{Z} \\
&convert^*_{F \to F'}(x) > 0 && \text{if } x > 0 \\
&convert^*_{F \to F'}(-x) = -convert^*_{F \to F'}(x) && \\
&convert^*_{F \to F'}(x) \leqslant convert^*_{F \to F'}(y) && \text{if } x < y
\end{aligned}$$

Relationship to other floating point to floating point conversion approximation helper functions:

$$convert^*_{F \to F'}(x) = convert^*_{F'' \to F'}(x) \qquad \text{if } lb(r_{F''}) = lb(r_F) \text{ and } x \in F \cap F''$$

The $convert_{F \to F'}$ operation:

$$convert_{F \to F'} : F \to F' \cup \{\mathbf{overflow}, \mathbf{underflow}\}$$

$$\begin{aligned}
&convert_{F \to F'}(x) \\
&\quad = result_{F'}(x, nearest_{F'}) && \text{if } x \in F \text{ and } lb(r_F) = lb(r_{F'}) \\
&\quad = trans\_result_{F'}(convert^*_{F \to F'}(x), nearest_{F'}) && \\
&\quad && \text{if } x \in F \text{ and } lb(r_F) \neq lb(r_{F'}) \\
&\quad = x && \text{if } x \in \{\mathbf{-\infty}, \mathbf{-0}, \mathbf{+\infty}\} \\
&\quad = \mathbf{qNaN} && \text{if } x \text{ is a quiet NaN} \\
&\quad = \mathbf{invalid}(\mathbf{qNaN}) && \text{if } x \text{ is a signalling NaN}
\end{aligned}$$

NOTE – When both datatypes conform to Part 1, and the radices for both of these floating point datatypes are the same, floating point to nearest floating point conversions are covered by Part 1. In this case the operations $cvt_{F \to F'}$ and $convert_{F \to F'}$ are identical.

5.4.5 Floating point to fixed point conversions

Let F be the non-special value set for a floating point datatype conforming to Part 1. Let D be the non-special value set for a fixed point datatype.

A fixed point datatype D is a subset of $\mathcal{R}$, characterised by a radix, $r_D \in \mathcal{Z}$ ($\geqslant 2$), a density, $d_D \in \mathcal{Z}$ ($\geqslant 0$), and if it is bounded, a maximum positive value, $dmax_D \in D^*$ ($\geqslant 1$). Given these values, the following sets are defined:

$$D^* = \{n/(r_D^{d_D}) \;\mid\; n \in Z\}$$

$$\begin{aligned} D &= D^* && \text{if } D \text{ is not bounded} \\ D &= D^* \cap [-dmax_D, dmax_D] && \text{if } D \text{ is bounded} \end{aligned}$$

NOTE 1 – D corresponds to **scaled(r_D, d_D)** in ISO/IEC 11404 Language independent datatypes (LID) [10]. LID has no parameter corresponding to $dmax_D$ even when the datatype is bounded.

The fixed point rounding helper function:

$$nearest_D : \mathcal{R} \to D^*$$

is the rounding function that rounds to nearest, ties round to even last digit.

The fixed point result helper function, $result_D$, is like $result_F$, but for a fixed point datatype. It will return **overflow** if the rounded result is not representable:

$$result_D : \mathcal{R} \times (\mathcal{R} \to D^*) \to D \cup \{\mathbf{overflow}\}$$

$$\begin{aligned} result_D(x, rnd) &= rnd(x) && \text{if } rnd(x) \in D \text{ and } (rnd(x) \neq 0 \text{ or } x \geqslant 0) \\ &= \mathbf{-0} && \text{if } rnd(x) = 0 \text{ and } x < 0 \\ &= \mathbf{overflow} && \text{if } x \in \mathcal{R} \text{ and } rnd(x) \notin D \end{aligned}$$

There shall be a $max\_error\_convert_D$ parameter that gives the maximum error when converting from F to D and $lb(r_F) \neq lb(r_D)$. The $max\_error\_convert_D$ parameter shall have a value in the interval $[0.5, 0.75]$. If $lb(r_F) = lb(r_D)$, the maximum error shall be 0.5 ulp when converting from F to D, but this is not reflected in any parameter.

The $convert^*_{F \to D}$ approximation helper function:

$$convert^*_{F \to D} : \mathcal{R} \to \mathcal{R}$$

$convert^*_{F \to D}(x)$ returns a close approximation to x in $\mathcal{R}$, with maximum error $max\_error\_convert_D$.

Further requirements on the $convert^*_{F \to D}$ approximation helper functions:

$$\begin{aligned} &convert^*_{F \to D}(x) = x && \text{if } x \in \mathcal{Z} \\ &convert^*_{F \to D}(x) > 0 && \text{if } x > 0 \\ &convert^*_{F \to D}(-x) = -convert^*_{F \to D}(x) \\ &convert^*_{F \to D}(x) \leqslant convert^*_{F \to D}(y) && \text{if } x < y \end{aligned}$$

Relationship to other floating point to fixed point conversion approximation helper functions:

$$convert^*_{F \to D}(x) = convert^*_{F'' \to D}(x) \qquad \text{if } lb(r_{F''}) = lb(r_F) \text{ and } x \in F \cap F''$$

The $convert_{F \to D}$ operation:

$$convert_{F \to D} : F \to D \cup \{\mathbf{-0}, \mathbf{overflow}\}$$

$$
\begin{array}{lll}
\mathit{convert}_{F\to D}(x) & = \mathit{result}_D(x, \mathit{nearest}_D) & \text{if } x \in F \text{ and } \mathit{lb}(r_F) = \mathit{lb}(r_D) \\
 & = \mathit{result}_D(\mathit{convert}^*_{F\to D}(x), \mathit{nearest}_D) & \\
 & & \text{if } x \in F \text{ and } \mathit{lb}(r_F) \neq \mathit{lb}(r_D) \\
 & = x & \text{if } x \in \{\mathbf{-\infty}, \mathbf{-0}, \mathbf{+\infty}\} \\
 & = \mathbf{qNaN} & \text{if } x \text{ is a quiet NaN} \\
 & = \mathbf{invalid}(\mathbf{qNaN}) & \text{if } x \text{ is a signalling NaN}
\end{array}
$$

NOTES

2 The datatype D need not be visible in the programming language. D may be a subtype of strings, according to some format. Even so, no datatype for strings need be present in the programming language.

3 This covers, among other things, “output” of floating point datatype values, to fixed point string formats. E.g. a binding may say that `float_to_fixed_string(`x`, `m`, `n`)` is bound to $\mathit{convert}_{F\to S_{m,n}}(x)$ where $S_{m,n}$ is strings of length m, representing fixed point values in radix 10 with n decimals. The binding should also detail how NaNs, signed zeroes and infinities are represented in $S_{m,n}$, as well as the precise format of the strings representing ordinary values. (Note that if the length of the target string is limited, the conversion may overflow.)

5.4.6 Fixed point to floating point conversions

Let F be the non-special value set for a floating point datatype conforming to Part 1. Let D and D' be the non-special value set for fixed point datatypes.

The $\mathit{convert}^*_{D\to F}$ approximation helper function:

$$\mathit{convert}^*_{D\to F} : \mathcal{R} \to \mathcal{R}$$

$\mathit{convert}^*_{D\to F}(x)$ returns a close approximation to x in $\mathcal{R}$, with maximum error $\mathit{max\_error\_convert}_F$.

Further requirements on the $\mathit{convert}^*_{D\to F}$ approximation helper functions:

$$
\begin{array}{ll}
\mathit{convert}^*_{D\to F}(x) = x & \text{if } x \in \mathcal{Z} \\
\mathit{convert}^*_{D\to F}(x) > 0 & \text{if } x > 0 \\
\mathit{convert}^*_{D\to F}(-x) = -\mathit{convert}^*_{D\to F}(x) & \\
\mathit{convert}^*_{D\to F}(x) \leqslant \mathit{convert}^*_{D\to F}(y) & \text{if } x < y
\end{array}
$$

Relationship to other floating point and fixed point to floating point conversion approximation helper functions:

$$
\begin{array}{ll}
\mathit{convert}^*_{D\to F}(x) = \mathit{convert}^*_{D'\to F}(x) & \text{if } \mathit{lb}(r_{D'}) = \mathit{lb}(r_D) \text{ and } x \in D \cap D' \\
\mathit{convert}^*_{D\to F}(x) = \mathit{convert}^*_{F'\to F}(x) & \text{if } \mathit{lb}(r_{F'}) = \mathit{lb}(r_D) \text{ and } x \in D \cap F'
\end{array}
$$

The $\mathit{convert}_{D\to F}$ operation:

$$\mathit{convert}_{D\to F} : D \to F \cup \{\mathbf{overflow}, \mathbf{underflow}\}$$

$$
\begin{array}{lll}
\mathit{convert}_{D\to F}(x) & = \mathit{result}_F(x, \mathit{nearest}_F) & \text{if } x \in D \text{ and } \mathit{lb}(r_D) = \mathit{lb}(r_F) \\
 & = \mathit{trans\_result}_F(\mathit{convert}^*_{D\to F}(x), \mathit{nearest}_F) & \\
 & & \text{if } x \in D \text{ and } \mathit{lb}(r_D) \neq \mathit{lb}(r_F) \\
 & = x & \text{if } x \in \{\mathbf{-\infty}, \mathbf{-0}, \mathbf{+\infty}\} \\
 & = \mathbf{qNaN} & \text{if } x \text{ is a quiet NaN} \\
 & = \mathbf{invalid}(\mathbf{qNaN}) & \text{if } x \text{ is a signalling NaN}
\end{array}
$$

5.5 Numerals as operations in the programming language

NOTE – Numerals as input, or in strings, is covered by the conversion operations above.

Each numeral is a parameterless operation. Thus, this clause introduces a very large number of operations, since the number of numerals is in principle infinite.

5.5.1 Numerals for integer datatypes

Let I' be a non-special value set for integer numerals for the datatype corresponding to I.

An integer numeral, denoting an abstract value n in $I' \cup \{\mathbf{-0}, \mathbf{+\infty}, \mathbf{-\infty}, \mathbf{qNaN}, \mathbf{sNaN}\}$, for an integer datatype, I, shall result in

$convert_{I' \to I}(n)$

For each integer datatype conforming to Part 1 and made directly available, with non-special value set I, there shall be integer numerals with radix 10.

For each radix for numerals made available for a bounded integer datatype I, there shall be integer numerals for all non-negative values of I.

For each radix for numerals made available for an unbounded integer datatype I, there shall be integer numerals for all non-negative values of I smaller than 10^{20}.

For each integer datatype made directly available and that has special values:

a) There should be a numeral for positive infinity.

b) There should be numerals for quiet and signalling NaNs.

5.5.2 Numerals for floating point datatypes

Let D' be a non-special value set for fixed point numerals for the datatype corresponding to F. Let F' be a non-special value set for floating point numerals for the datatype corresponding to F.

A fixed point numeral, denoting an abstract value x in $D' \cup \{\mathbf{-0}, \mathbf{+\infty}, \mathbf{-\infty}, \mathbf{qNaN}, \mathbf{sNaN}\}$, for a floating point datatype, F, shall result in

$convert_{D' \to F}(x)$

A floating point numeral, denoting an abstract value x in $F' \cup \{\mathbf{-0}, \mathbf{+\infty}, \mathbf{-\infty}, \mathbf{qNaN}, \mathbf{sNaN}\}$, for a floating point datatype, F, shall result in

$convert_{F' \to F}(x)$

For each floating point datatype conforming to Part 1 and made directly available, with non-special value set F, there should be radix 10 floating point numerals, and there shall be radix 10 fixed point numerals.

For each radix for fixed point numerals made available for a floating point datatype F, there shall be numerals for all bounded precision and bounded range expressible non-negative values of $\mathcal{R}$. At least a precision ($d_{D'}$) of 20 should be available. At least a range ($dmax_{D'}$) of 10^{20} should be available.

For each radix for floating point numerals made available for a floating point datatype F, there shall be numerals for all bounded precision and bounded range expressible non-negative values of $\mathcal{R}$. The precision and range bounds for the numerals shall be large enough to allow all non-negative values of F to be reachable.

For each floating point datatype made directly available:

a) There shall be a numeral for positive infinity.

b) There shall be numerals for quiet and signalling NaNs.

The conversion operations used for numerals as operations should be the same as those used by default for converting strings to values in conforming integer or floating point datatypes.

6 Notification

Notification is the process by which a user or program is informed that an arithmetic operation cannot return a suitable numeric result. Specifically, a notification shall occur when any arithmetic operation returns an exceptional value. Notification shall be performed according to the requirements of clause 6 of Part 1.

An implementation shall not give notifications for operations conforming to this Part, unless the specification requires that an exceptional value results for the given arguments.

The default method of notification should be recording of indicators.

6.1 Continuation values

If notifications are handled by a recording of indicators, in the event of notification the implementation shall provide a *continuation value* to be used in subsequent arithmetic operations. Continuation values may be in I or F (as appropriate), or be special values ($-\mathbf{0}$, $-\infty$, $+\infty$, or a **qNaN**).

Floating point datatypes that satisfy the requirements of IEC 60559 have special values in addition to the values in F. These are: $-\mathbf{0}$, $+\infty$, $-\infty$, *signaling* **NaN**s (**sNaN**), and *quiet* **NaN**s (**qNaN**). Such values may be passed as arguments to operations, and used as results or continuation values. Floating point types that do not fully conform to IEC 60559 can also have values corresponding to $-\mathbf{0}$, $+\infty$, $-\infty$, or **NaN**.

Continuation values of $-\mathbf{0}$, $+\infty$, $-\infty$, and **NaN** are required only if the parameter iec_559_F has the value **true**. If the implementation can represent such special values in the result datatype, they should be used according to the specifications in this Part.

7 Relationship with language standards

A computing system often provides some of the operations specified in this Part within the context of a programming language. The requirements of the present standard shall be in addition to those imposed by the relevant programming language standards.

This Part does not define the syntax of arithmetic expressions. However, programmers need to know how to reliably access the operations specified in this Part.

NOTE 1 – Providing the information required in this clause is properly the responsibility of programming language standards. An individual implementation would only need to provide details if it could not cite an appropriate clause of the language or binding standard.

An implementation shall document the notation that should be used to invoke an operation specified in this Part and made available. An implementation should document the notation that should be used to invoke an operation specified in this Part and that could be made available.

NOTE 2 – For example, the radian arc sine operation for an argument x ($arcsin_F(x)$) might be invoked as

`arcsin(`x`)`	in Pascal [28] and Ada [11]
`asin(`x`)`	in C [18] and Fortran [23]
`(asin `x`)`	in Common Lisp [43] and ISLisp [25]
`function asin(`x`)`	in COBOL [20]

with suitable expression of the argument (x).

An implementation shall document the semantics of arithmetic expressions in terms of compositions of the operations specified in clause 5 of this Part and in clause 5 of Part 1.

Compilers often "optimize" code as part of compilation. Thus, an arithmetic expression might not be executed as written. An implementation shall document the possible transformations of arithmetic expressions (or groups of expressions) that it permits. Typical transformations include

a) Insertion of operations, such as datatype conversions or changes in precision.

b) Replacing operations (or entire subexpressions) with others, such as "`cos(-x)`" → "`cos(x)`" (exactly the same result) or "`pi - arccos(x)`" → "`arccos(-x)`" (more accurate result) or "`exp(x)-1`" → "`expm1(x)`" (more accurate result if $x > -1$, less accurate result if $x < -1$, different notification behaviour).

c) Evaluating constant subexpressions.

d) Eliminating unneeded subexpressions.

Only transformations which alter the semantics of an expression (the values produced, and the notifications generated) need be documented. Only the range of permitted transformations need be documented. It is not necessary to describe the specific choice of transformations that will be applied to a particular expression.

The textual scope of such transformations shall be documented, and any mechanisms that provide programmer control over this process should be documented as well.

NOTE 3 – It is highly desirable that programming languages intended for numerical use provide means for limiting the transformations applied to particular arithmetic expressions. Control over changes of precision is particularly useful.

8 Documentation requirements

In order to conform to this Part, an implementation shall include documentation providing the following information to programmers.

NOTE 1 – Much of the documentation required in this clause is properly the responsibility of programming language or binding standards. An individual implementation would only need to provide details if it could not cite an appropriate clause of the language or binding standard.

a) A list of the provided operations that conform to this Part.

b) For each maximum error parameter, the value of that parameter or definition of that parameter function. Only maximum error parameters that are relevant to the provided operations need be given.

c) The value of the parameters $big\_angle\_r_F$ and $big\_angle\_u_F$. Only big angle parameters that are relevant to the provided operations need be given.

d) For the $nearest_F$ function, the rule used for rounding halfway cases, unless iec_559_F is fixed to true.

e) For each conforming operation, the continuation value provided for each notification condition. Specific continuation values that are required by this Part need not be documented. If the notification mechanism does not make use of continuation values (see clause 6), continuation values need not be documented.

NOTE 2 − Implementations that do not provide infinities or **NaN**s will have to document any continuation values used in place of such values.

f) For each conforming operation, how the results depend on the rounding mode, if rounding modes are provided. Operations may be insensitive to the rounding mode, or sensitive to it, but even then need not heed the rounding mode.

g) For each conforming operation, the notation to be used for invoking that operation.

h) For each maximum error parameter, the notation to be used to access that parameter.

i) The notation to be used to access the parameters $big\_angle\_r_F$ and $big\_angle\_u_F$.

Since the integer and floating point datatypes used in conforming operations shall satisfy the requirements of Part 1, the following information shall also be provided by any conforming implementation.

j) The translation of arithmetic expressions into combinations of the operations provided by any part of ISO/IEC 10967, including any use made of higher precision. (See clause 7 of Part 1.)

k) The methods used for notification, and the information made available about the notification. (See clause 6 of Part 1.)

l) The means for selecting among the notification methods, and the notification method used in the absence of a user selection. (See 6.3 of Part 1.)

m) The means for selecting the modes of operation that ensure conformity.

n) When “recording of indicators” is the method of notification, the datatype used to represent Ind, the method for denoting the values of Ind (the association of these values with the subsets of E must be clear), and the notation for invoking each of the “indicator” operations. (See 6.1.2 of Part 1.) In interpreting 6.1.2 of Part 1, the set of indicators E shall be interpreted as including all exceptional values listed in the signatures of conforming operations. In particular, E may need to contain **pole** and **absolute_precision_underflow**.

o) For each of the provided operations where this Part specifies a relation to another operation specified in this Part, the binding for that other operation.

p) For numerals conforming to this Part, which available string conversion operations, including reading from input, give exactly the same conversion results, even if the string syntaxes for ‘internal’ and ‘external’ numerals are different.

Annex A
(normative)

Partial conformity

If an implementation of an operation fulfills all relevant requirements according to the normative text in this Part, except the ones relaxed in this Annex, the implementation of that operation is said to *partially conform* to this Part.

Conformity to this Part shall not be claimed for operations that only fulfill Partial conformity.

Partial conformity shall not be claimed for operations that relax other requirements than those relaxed in this Annex.

A.1 Maximum error relaxation

This Part has the following maximum error requirements for conformity.

$max\_error\_hypot_F \in [0.5, 1]$

$max\_error\_exp_F \in [0.5, 1.5 * rnd\_error_F]$
$max\_error\_power_F \in [max\_error\_exp_F, 2 * rnd\_error_F]$

$max\_error\_sinh_F \in [0.5, 2 * rnd\_error_F]$
$max\_error\_tanh_F \in [max\_error\_sinh_F, 2 * rnd\_error_F]$

$max\_error\_sin_F \in [0.5, 1.5 * rnd\_error_F]$
$max\_error\_tan_F \in [max\_error\_sin_F, 2 * rnd\_error_F]$

$max\_error\_sinu_F : F \rightarrow F \cup \{\mathbf{invalid}\}$
$max\_error\_tanu_F : F \rightarrow F \cup \{\mathbf{invalid}\}$

$max\_error\_convert_F \in [0.5, 0.75]$

For $u \in G_F$, the $max\_error\_sinu_F(u)$ parameter shall be in the interval $[max\_error\_sin_F, 2]$, and the $max\_error\_tanu_F(u)$ parameter shall be in the interval $[max\_error\_tan_F, 4]$. For $u \in T$, the $max\_error\_sinu_F(u)$ parameter shall be equal to $max\_error\_sin_F$, and the $max\_error\_tanu_F(u)$ parameter shall be equal to $max\_error\_tan_F$.

In a Partially conforming implementation the maximum error parameters may be greater than what is specified by this Part. The maximum error parameter values given by an implementation shall still adequately reflect the accuracy of the relevant operations, if a claim of Partial conformity is made.

A Partially conforming implementation shall document which maximum error parameters have greater values than specified by this Part, and their values.

A.2 Extra accuracy requirements relaxation

This Part has a number of extra accuracy requirements. These are detailed in the paragraphs beginning "Further requirements on the op_F^* approximation helper function are:".

In a Partially conforming implementation these further requirements need not be fulfilled. The values returned must still be within the maximum error bounds that are given by the maximum error parameters, if a claim of Partial conformity is made.

A Partially conforming implementation shall document which extra accuracy requirements are not fulfilled by the implementation.

A.3 Relationships to other operations relaxation

This Part has a number of requirements giving relations to other operations. These are detailed in the paragraphs beginning "Relationship to the op^{*}_{F} approximation helper function:".

In a Partially conforming implementation these relationships need not be fulfilled. The values returned must still be within the maximum error bounds that are given by the maximum error parameters, if a claim of Partial conformity is made.

A Partially conforming implementation shall document which operation relationships are not fulfilled by the implementation.

Annex B
(informative)

Rationale

This annex explains and clarifies some of the ideas behind *Information technology – Language independent arithmetic – Part 2: Elementary numerical functions* (LIA-2).

B.1 Scope

B.1.1 Inclusions

LIA-2 is intended to define the meaning of some operations on Integer and floating point types as specified in LIA-1 (ISO/IEC 10967-1), in addition to the operations specified in LIA-1. LIA-2 does not specify operations for any additional arithmetic datatypes, though fixed point datatypes are used in some of the specifications for conversion operations.

The specifications for the operations covered by LIA-2 are given in sufficient detail to

a) support detailed and accurate numerical analysis of arithmetic algorithms,

b) enable a precise determination of conformity or non-conformity, and

c) prevent exceptions (like overflow) from going undetected.

LIA-2 does in no way prevent language standards or implementations including further arithmetic operations, other variations of included arithmetic operations, or the inclusion of further arithmetic datatypes, like rational number or fixed point datatypes. Some of these may become the topic of standardisation in other parts of LIA.

B.1.2 Exclusions

LIA-2 is not concerned with techniques for the *implementation* of numerical functions. Even when an LIA-2 specification is made in terms of other LIA-1 or LIA-2 operations, that does *not* imply a requirement that an implementation implements the operation in terms of those other operations. It is sufficient that the result (returned value or returned continuation value, and exception behaviour) is *as if* it was implemented in terms of those other operations.

LIA-2 does not provide specifications for operations which involve no arithmetic processing, like assignment and parameter passing, though any implicit conversions done in association with such operations are in scope. The implicit conversions should be made available to the programmer as explicit conversions.

LIA-2 does not cover operations for the support of domains such as linear algebra, statistics, and symbolic processing. Such domains deserve *separate* standardisation, if standardised.

LIA-2 only covers operations that involve integer or floating point datatypes, as specified in LIA-1, and in some cases also a Boolean datatype, but then only as result. The operations covered by LIA-2 are often to some extent covered by programming language standards, like the operations `sin`, `cos`, `tan`, `arctan`, and so on.

B.2 Conformity

Conformity to this standard is dependent on the existence of language binding standards. Each programming language committee (or other organisation responsible for a programming language or other specification to which LIA-1 and LIA-2 may apply) is encouraged to produce a binding standard covering at least those operations already required by the programming language (or similar) and also specified in LIA-2.

The term "programming language" is here used in a generalised sense to include other computing entities such as calculators, spread sheets, page description languages, web-script languages, and database query languages to the extent that they provide the operations covered by LIA-2.

Suggestions for bindings are provided in Annex C. Annex C has partial binding examples for a number of existing programming languages and LIA-2. In addition to the bindings for the operations in LIA-2, it is also necessary to provide bindings for the maximum error parameters and big angle parameters specified by LIA-2. Annex C contains suggestions for these bindings. To conform to this standard, in the absence of a binding standard, an implementation should create a binding, following the suggestions in Annex C.

B.3 Normative references

The referenced IEC 60559 standard is identical to the former IEC 559 and IEEE 754 standards.

B.4 Symbols and definitions

B.4.1 Symbols

B.4.1.1 Sets and intervals

The interval notation is in common use. It has been chosen over the other commonly used interval notation because the chosen notation has no risk of confusion with the pair notation.

B.4.1.2 Operators and relations

Note that all operators, relations, and other mathematical notation used in LIA-2 is used in their conventional exact mathematical sense. They are not used to stand for operations specified by IEC 60559, LIA-1, LIA-2, or, with the exception of programme excerpts which are clearly marked, any programming language. E.g. x/u stands for the mathematically exact result of dividing x by u, whether that value is representable in any floating point datatype or not, and $x/u \neq div_F(x,u)$ is often the case. Likewise, $=$ is the mathematical equality, not the eq_F operation: $0 \neq \mathbf{-0}$, while $eq_F(0, \mathbf{-0}) = \mathbf{true}$.

B.4.1.3 Mathematical functions

The elementary functions named sin, cos, etc. used in LIA-2 are the exact mathematical functions, not any approximation. The approximations to these mathematical functions are introduced in clauses 5.3 and 5.4 and are written in a way clearly distinct from the mathematical functions. E.g., sin^*_F, cos^*_F, etc., which are unspecified mathematical functions approximating the targeted exact mathematical functions to a specified degree; sin_F, cos_F, etc., which are the operations specified by LIA-2 based on the respecitive approximating function; `sin`, `cos`, etc., which are programming language names bound to LIA-2 operations. `sin` is thus very different from sin.

B.4.1.4 Datatypes and exceptional values

The sequence types $[I]$ and $[F]$ appear as input datatypes to a few operations: max_seq_I, min_seq_I, gcd_seq_I, lcm_seq_I, max_seq_F, min_seq_F, $mmax\_seq_F$, and $mmin\_seq_F$.

In effect, a sequence is a finite linearly ordered collection of elements which can be indexed from 1 to the length of the sequence. Equality of two or more elements with different indices is possible. Sequences are used in LIA-2 as an abstraction of arrays, lists, other kinds of one-dimensional sequenced collections, and even variable length argument lists. As used in LIA-2 the order of the elements and number of occurrences of each element, as long as it is more than one, does not matter, so multi-sets (bags) and sets also qualify.

LIA-2 uses a modified set of exceptional values compared to LIA-1. Instead of LIA-1's **undefined**, LIA-2 uses **invalid** and **pole**. IEC 60559 distinguishes between **invalid** and **divide_by_zero** (the latter is called **pole** by LIA-2). The distinction is valid and should be recognised, since **pole** indicates that an infinite but *exact* result is (or can be, if it were available) returned, while **invalid** indicates that a result in the target datatype (extended with infinities) cannot, or should not, be returned with adequate accuracy.

LIA-1 distinguished between **integer_overflow** and **floating_overflow**. This distinction is moot, since no distinction was made between **integer_undefined** and **floating_undefined**. In addition, continuing this distinction would force LIA to start distinguishing not only **integer_overflow** and **floating_overflow**, but also **fixed_overflow**, **complex_floating_overflow**, **complex_integer_overflow**, etc. Further, there is no general consensus that maintaining this distinction is useful, and many programming languages do not require a distinction. A binding standard can still maintain this distinction, if desired.

LIA allows for three methods for handing notifications: recording of indicators, change of control flow (returnable or not), and termination of program. The preferred method is recording of indicators. This allows the computation to continue using the continuation values. For **underflow** and **pole** notifications this course of action is strongly preferred, provided that a suitable continuation value can be represented in the result datatype.

Not all occurrences of the same exceptional value need be handled the same. There may be explicit mode changes in how notifications are handled, and there may be implicit changes. E.g., **invalid** without a specified (by LIA-2 or binding) continuation value to cause change of control flow (like an Ada [11] exception), while **invalid** with a specified continuation value use recording of indicators. This should be specified by bindings or by implementations.

The operations may return any of the exceptional values **overflow**, **underflow**, **invalid**, **pole**, or **absolute_precision_underflow**. This does *not* imply that the implemented operations are to actually *return* any of these values. When these values are returned according to the LIA specification, that means that the implementation is to perform a notification handling for that exceptional value. If the notification handling is by recording of indicators, then what is actually returned by the implemented operation is the continuation value.

B.4.2 Definitions of terms

Note the LIA distinction between exceptional values, exceptions, and exception handling (handling of notification by non-returnable change of control flow; as in e.g. Ada). LIA exceptional values are not the same as Ada `exceptions`, nor are they the same as IEC 60559 special values.

B.5 Specifications for the numerical functions

The abstract values used in the specifications are independent of datatype, just like the mathematical numbers are. That they are represented differently in, say, single precision and in double precision is out of scope for LIA-2.

The specifications in LIA-2 for floating point operations give details about certain special values (they are 'special' in that they are not in $\mathcal{R}$). These special values are commonly represented in floating point datatypes, in particular all floating point datatypes conforming to IEC 60559.

B.5.1 Basic integer operations

Integer datatypes can have infinity values as well as NaN values, and also may have a $\mathbf{-0}$. A corresponding I must, however, be a subset of $\mathcal{Z}$. $\mathbf{-0}$ is commonly available when the integer datatype is represented using radix-minus-1-complement, e.g. 1's complement. When using, e.g., 2's complement, the representation that would otherwise represent the most negative value can be used as a NaN. Especially for unbounded integer types, the inclusion of infinities is advisable, not for overflow, since these do not occur, but in order to have a smallest and a largest value in the type.

B.5.1.1 The integer *result* and *wrap* helper functions

The $result_I$ helper function notifies overflow when the result cannot be represented in I. When an overflow occurs, and recording of indicators is the method for handling (integer) overflows, a continuation value must be given. For bounded integer datatypes, $maxint_F$ and $minint_F$ can be suitable continuation values. In some instances a wrapped result, see below, may be used as continuation value on overflow. Few integer datatypes offer representations for positive and negative infinity. In case such representations are offered, they can be used as continuation values on overflow, similar to their use in floating point datatypes. LIA does not specify the continuation value in this case, that is left to bindings or implementations, but LIA does require that the continuation value(s) be documented.

The $wrap_I$ helper function wraps the result into a value that can be represented in I. The result is wrapped in such a way that the value returned can be used to implement extended range integer arithmetic.

B.5.1.2 Integer maximum and minimum

The operations for integer maximum and minimum are trivial, except taking the maximum or minimum of an empty sequence (empty array, empty list, zero number of parameters, or similar). The case for zero number of parameters is often syntactically excluded (as in Fortran, Common Lisp, and ISLisp), while an empty array or empty list given as a single argument must usually be possible handle at 'runtime'. LIA specifies a **pole** notification for this case. Since no (implied mathematical) division is involved here, **pole** is here to be interpreted as "exact infinite result from finite operands", in this case an empty list of numbers.

If infinity values are required to be available for a particular integer datatype, a binding may require the continuation values specified to be returned without any **pole** notification. When the specified continuation value, $\mathbf{+\infty}$ or $\mathbf{-\infty}$, is not available, other suitable continuation values may be used, and if so they must be documented. If the integer datatype is bounded, but without infinities, $maxint_F$ may be used in place of $\mathbf{+\infty}$ and $minint_F$ may be used instead of $\mathbf{-\infty}$.

Infinities as arguments are not specified for these operations, since infinities are rarely available in integer datatypes. However, compare the specification for max and min operations for floating point datatypes (clause 5.2.2).

B.5.1.3 Integer diminish

Integer diminish is sometimes called 'monus'. This operation computes the 'positive difference' between two numbers.

B.5.1.4 Integer power and arithmetic shift

The integer arithmetic shift operations can be used to implement integer multiplication and integer division more quickly in special cases.

The shift operations shift either 'right' or 'left' depending on the sign of the second argument.

Any continuation value used on overflow here must be documented, either by the binding standard or by the implementation.

B.5.1.5 Integer square root

B.5.1.6 Divisibility tests

Even and odd are simple special cases offered as separately named operations in several programming languages.

B.5.1.7 Integer division and remainder

When the result of a division between integers47 is not an integer, but the final result is required to be an integer, the quotient must be rounded. There are several ways of doing this; floor, ceiling, and unbiased round to nearest being the most important.

pad_I returns the *negative* of the remainder after division and ceiling. The reason for this is twofold: 1) for unsigned integer datatypes the remainder is $\leqslant 0$, and would thus often not be representable unless negated, and 2) it is intuitively easier to think of the "places left in the last unfilled group of equi-sized and packed groups" as a positive entity, a padding.

$remr_I$ can overflow only for unsigned integer datatypes ($minint_I = 0$), and does so for too many arguments, and negating it does not change this. $remr_I$ should therefore not be provided for unsigned integer datatypes. $remr_I$ rounds in the same way as $remr_F$, IEEE remainder.

When there is no exception, these operations fulfill $divf_I(x+n{\cdot}y, y) = divf_I(x, y)+n$, $group_I(x+n{\cdot}y, y) = group_I(x, y)+n$, $quot_I(x+2{\cdot}n{\cdot}y, y) = quot_I(x, y)+2{\cdot}n$, $moda_I(x+n{\cdot}y, y) = moda_I(x, y)$, $pad_I(x + n \cdot y, y) = pad_I(x, y)$, $remr_I(x + 2 \cdot n \cdot y, y) = remr_I(x, y)$, where $n \in \mathcal{Z}$.

Note that the div^t_I and rem^t_I from LIA-1 do not fulfill similar useful equalities, due to the disruption around 0 for this pair of operations.

When there is no exception, $divf_I(x, y) = -group_I(-x, y)$, $divf_I(x, y) = -group_I(x, -y)$, $quot_I(x, y) = -quot_I(-x, y)$, $quot_I(x, y) = -quot_I(x, -y)$, $moda_I(x, y) = -pad_I(x, -y)$, and $remr_I(x, y) = remr_I(x, -y)$.

B.5.1.8 Greatest common divisor and least common positive multiple

The greatest common divisor is useful in reducing a a rational number to its lowest terms. The least common multiple is useful in converting two rational numbers to have the same denominator.

Returning 0 for $gcd_I(0,0)$, as is sometimes suggested, would be incorrect, since the greatest common divisor for 0 and 0 should be the supremum (upper limit) of $\mathcal{Z}^+$, since these all divide 0, which is infinity.

gcd_I will overflow only if $bounded_I =$ **true**, $minint_I = -maxint_I - 1$, and both arguments are $minint_I$. The greatest common divisor is then $-minint_I$, which then is not in I.

Least common positive multiple, $lcm_I(x,y)$, overflows for many "large" arguments. E.g., if x and y are relative primes, then the least common multiple is $|x \cdot y|$, which may be greater than $maxint_I$.

B.5.1.9 Support operations for extended integer range

These operations would typically be used to extend the range of the highest level integer datatype supported by the underlying hardware of an implementation.

The two parts of an integer product, $mul\_ov_I(x,y)$ and $mul\_wrap_I(x,y)$ together provide the complete integer product. Similarly for addition and subtraction.

The use of $wrap_I$ guarantees that **overflow** will not occur.

B.5.2 Basic floating point operations

F must be a subset of $\mathcal{R}$. Floating point datatypes can have infinity values as well as NaN values, and also may have a $\mathbf{-0}$. These values are not in F. The special values are, however, commonly available in floating point datatypes today, thanks to the wide adoption of IEC 60599.

Note that for some operations the exceptional value **invalid** is produced only for argument values involving $\mathbf{-0}$, $\mathbf{+\infty}$, $\mathbf{-\infty}$, or **sNaN**. For these operations the signature given in LIA-2 does not contain **invalid**.

A report ([57]) issued by the ANSI X3J11 committee discusses possible ways of exploiting the IEC 60559 special values. The report identifies some of its suggestions as controversial and cites [53] as justification.

In the following paragraphs summarise the specifications of IEC 60559 on the creation and propagation of signed zeros, infinities, and **NaN**s. There is also some discussion of the material in [53, 54, 51].

IEC 60559 regards 0 and $\mathbf{-0}$ as almost indistinguishable. The sign is supposed to indicate the direction of approach to zero. The sign is reliable for a zero generated by underflow in a multiplication or division operation, and should be reliable also for operations that approximate elementary transcendental functions (see the LIA-2 specifications in clause 5.3). It is not reliable for a zero generated by an implied subtraction of two floating point numbers with the same value, for which case the zero is arbitrarily given a + sign. The phrase "implied subtraction" indicates either the addition of two oppositely signed numbers or the subtraction of two like signed numbers.

On occurrence of floating overflow or division of a non-zero number by zero, an implementation conforming to IEC 60559 sets the appropriate status flag (if trapping is not enabled) and then continues execution with a result of $\mathbf{+\infty}$ or $\mathbf{-\infty}$ if rounding is to nearest. Infinities as such do *not* indicate that an overflow or division by zero has occurred; infinities can be exact values. IEC 60559 states that the arithmetic of infinities is that associated with mathematical infinities. Thus, an infinity times, plus, minus, or divided by a non-zero finite floating point number yields an infinity for the result; no status flag is set and execution continues. These rules are not necessarily valid for infinities generated by overflow, thought they are valid if the infinitary arguments are exact.

NaNs are generated by invalid operations on infinities, 0/0, and the square root of a negative number (other than $\mathbf{-0}$). Thus **NaN**s can represent unknown real or complex values, as well as totally undefined values. IEC 60559 requires that the result of any of its basic operations with one or more **NaN** arguments shall be a **NaN**. This principle is not extended to the numerical functions by [53, 57]. The controversial specifications in [57] are based on an assumption that all of these special operands represent finite non-zero real-valued numbers; see [53, 54].

The LIA-2 policy (for clauses 5.2 and 5.3) for dealing with signed zeros, infinities, and **NaN**s is as follows:

a) The output is a quiet **NaN** for any operation for which one (or more) arguments is a quiet **NaN**, and none of the other arguments is a signalling NaN. There is then no notification.

b) If a mathematical function $h(x)$ is such that $h(0) = 0$, the corresponding operation $op_F(x)$ returns x if $x \in \{0, \mathbf{-0}\}$ and h has a positive derivative at 0, and $op_F(x)$ returns $neg_F(x)$ if $x \in \{0, \mathbf{-0}\}$ and h has a negative derivative at 0.

c) For an argument vector, $\overrightarrow{x}$, where that argument vector involves 0, $\mathbf{-0}$, $\mathbf{+\infty}$, or $\mathbf{-\infty}$, the result of the operation $op_F(\overrightarrow{x})$ is

$$\lim_{\overrightarrow{z} \to \overrightarrow{x}} h(\overrightarrow{z})$$

where an approach to zero is from the positive side if $\overrightarrow{x} = (..., 0, ...)$, and the approach is from the negative side if $\overrightarrow{x} = (..., \mathbf{-0}, ...)$. There is no notification if the limit exists, is finite, and is path independent. The returned value is $\mathbf{+\infty}$ or $\mathbf{-\infty}$ if the limiting value is unbounded, and the approach is towards a point infinitely far from the origin. The returned value is $\mathbf{pole}(\mathbf{+\infty})$ or $\mathbf{pole}(\mathbf{-\infty})$ if the limiting value is unbounded, and the approach is towards a finite point. The result is $\mathbf{-0}$ if the limit is zero and the approaching values are path independently negative. The result is 0 if the limit is zero and the approaching values are not path independently negative. If a path independent limit does not exist the value returned is **invalid**, and a notification occurs, with a continuation value of **qNaN** if appropriate.

An exception is made for the arc_F and $arcu_F$ operations, where it is found significantly more useful to return certain non-exceptional values for the origin and for the four double infinity argument cases, than to return an exceptional value, even with non-NaN continuation values.

B.5.2.1 The rounding and floating point *result* helper functions

The $result_F$ helper function notifies overflow when the result is too large to be approximated by a value in F. The $result_F$ helper function notifies underflow when there is (risk for) denormalisation loss for a tiny result. The $result_F$ helper function also ensures that a properly signed zero is the continuation value when a zero is appropriate. When an overflow or underflow occurs, and recording of indicators is the method for handling (floating point) overflow or undeflow, a continuation value must be provided. LIA-2 specifies a continuation value, and if that can be represented in the target datatype, that value should be used as continuation value.

B.5.2.2 Floating point maximum and minimum

As for the integer case, the maximum and minimum of empty sequences need be handled, but for floating point datatypes, infinities are usually available.

For floating point one also usually have negative zero available, and returning the correct sign on a zero result for the maximum and minimum operations requires more than simple comparisons to implement. The sign of zeroes may need to be inspected using *copysign* or *isnegativezero*.

B.5.2.3 Floating point diminish

As for the integer case, this operation computes the positive difference.

B.5.2.4 Round, floor, and ceiling

Since $fmax_F$ always has an integral value according to LIA-1, no overflow can occur for these operations.

Note that the sign of a zero result is maintained in accordance with IEC 60559:

$$floor_F(x) = neg_F(ceiling_F(neg_F(x)))$$
$$ceiling_F(x) = neg_F(floor_F(neg_F(x)))$$
$$rounding_F(x) = neg_F(rounding_F(neg_F(x)))$$

Negative zeroes, if available, are handled in such a way as to maintain these identities.

B.5.2.5 Remainder after division and round to integer

The remainder after division and unbiased round to integer (IEC 60559 remainder, or IEEE remainder) is an exact operation, even if the floating point datatype only conforms to LIA-1, but not to the more specific IEC 60559.

Remainder after floating point division and floor to integer cannot be exact. For a small negative nominator and a positive denominator, the resulting value looses much absolute accuracy in relation to the original value. Such an operation is therefore not included in LIA-2. Similarly for floating point division and ceiling.

See also the radian normalisation and the argument angular-unit normalisation operations (5.3.9.1, 5.3.10.1).

B.5.2.6 Square root and reciprocal square root

$\sqrt{x}$ cannot be exactly halfway between two values in F if $x \in F$. For $\sqrt{x}$ to be exactly halfway between two values in F would require that it had exactly $(p + 1)$ digits (last digit non-zero) for its exact representation. The square of such a number would require at least $(2 \cdot p + 1)$ digits with last $p + 1$ digits not all zero, which could not equal the p-digit number x.

The extensions $sqrt_F(\mathbf{+\infty}) = \mathbf{+\infty}$ and $sqrt_F(\mathbf{-0}) = \mathbf{-0}$ are mandated by IEC 60559. LIA-2 also requires that these hold for implementations which support infinities and signed zeros. However, it should be noted that while the second is harmless, the first may lead to erroneous results for a $\mathbf{+\infty}$ generated by an addition or subtraction with result just barely outside of $[-fmax_F, fmax_F]$ after rounding. Hence its square root would be well within the representable range. The possibility that LIA-2 should require that $sqrt_F(\mathbf{+\infty}) = \mathbf{invalid}(\mathbf{+\infty})$ was considered, but rejected because of the principle of regarding arguments as exact, even if they are not exact, when there is a non-degenerate neighbourhood around the argument point, for which the mathematical function on $\mathcal{R}$ is defined. In addition $sqrt_F(+\infty) = +\infty$ is already required by IEC 60559.

Note that the requirement that $sqrt_F(x) = \textbf{invalid}(\textbf{qNaN})$ for x strictly less than zero is mandated by IEC 60559. It follows that **NaN**s generated in this way represent imaginary values, which would become complex through addition and subtraction, and even imaginary infinities on multiplication by ordinary infinities.

The $rsqrt_F$ operation will increase performance for scaling a vector into a unit vector. Such an operation involves division of each component of the vector by the magnitude of the vector or, equivalently and with higher performance, multiplication by the reciprocal of the magnitude.

B.5.2.7 Support operations for extended floating point precision

These operations would typically be used to extend the precision of the highest level floating point datatype supported by the underlying hardware of an implementation. There is, however, no intent to provide a set of operations suitable for the implementation of a *complete* package for the support of calculations at an arbitrarily high level of precision.

The major motivation for including them in LIA-2 is to provide a capability for accurately evaluating residuals in an iterative algorithm. The residuals give a measure of the error in the current solution. More importantly they can be used to estimate a correction to the current solution. The accuracy of the correction depends on the accuracy of the residuals. The residuals are calculated as a difference in which the number of leading digits cancelled increases as the accuracy of the solution increases. A doubled precision calculation of the residuals is usually adequate to produce a reasonably efficient iteration.

For the basic floating point arithmetic doubled precision operations, the high parts may be calculated by the corresponding floating point operations as specified in LIA-1. Note, however, that in order to implement exact floating point addition and subtraction, rnd_F must round to nearest. If $add_F(x, y)$ rounds to *nearest* then the high and low parts represent $x + y$ exactly.

When the high parts of an addition or subtraction overflows, the low parts, as specified by LIA-2, return their results as if there was no overflow.

The product of two numbers, each with p digits of precision, is always exactly representable in at most $2 \cdot p$ digits. The high and low parts of the product will always represent the true product.

The remainder for division is more useful than a $2 \cdot p$-digit approximation. The remainder will be exactly representable if the high part differs from the true quotient by less than one *ulp*. The true quotient can be constructed p digits at a time by division of the successive remainders by the divisor.

The remainder for square root is more useful than a low part for the same reason that the remainder is more useful for division. The remainder for the square root operation will be exactly representable only if the high part is correctly rounded to nearest, as is required by the specification for $sqrt_F$.

See *Semantics for Exact Floating Point Operations* [63] for more information on exact floating point operations.

See *Proposal for Accurate Floating-Point Vector Arithmetic* [64] for more information on exact, or high accuracy, floating point summation and dot product. These operations may be the subject of an amendment to LIA-2.

B.5.3 Elementary transcendental floating point operations

B.5.3.1 Maximum error requirements

$max\_error\_op_F$ measures the discrepancy between the computed value $op_F(x)$ and the true mathematical value $f(x)$ in ulps of the true value. The magnitude of the error bound is thus available to a program from the computed value $op_F(x)$. Note that for results at an exponent boundary for F, y, the error away from zero is in terms of $ulp_F(y)$, whereas the error toward zero is in terms of $ulp_F(y)/r_F$, which is the ulp of values slightly smaller in magnitude than y.

Within limits, accuracy and performance may be varied to best meet customer needs. Note also that LIA-2 does not prevent a vendor from offering two or more implementations of the various operations.

The operation specifications define the domain and range for the operations. The computational domain and range are more limited for the operations than for the corresponding mathematical functions because the arithmetic datatypes are subsets of $\mathcal{R}$. Thus the actual domain of $exp_F(x)$ is approximately given by $x \leqslant \ln(fmax_F)$. For larger values of x, $exp_F(x)$ will overflow, though for $x = +\infty$ the exact result $+\infty$ will be returned. The actual range extends over F, although there are non-negative values, $v \in F$, for which there is no $x \in F$ satisfying $exp_F(x) = v$.

The thresholds for the **overflow** and **underflow** notifications are determined by the parameters defining the arithmetic datatypes. The threshold for an **invalid** notification is determined by the domain of arguments for which the mathematical function being approximated is defined. The **pole** notification is the operation's counterpart of a mathematical pole of the mathematical function being approximated by the operation. The threshold for **absolute_precision_underflow** is determined by the parameters $big\_angle\_r_F$ and $big\_angle\_u_F$.

LIA-2 imposes a fairly tight bound on the maximum error allowed in the implementation of each operation. The tightest possible bound is given by requiring rounding to nearest, for which the accompanying performance penalty is often unacceptably high for the operations approximating elementary transcendental functions. LIA-2 does not require round to nearest for such operations, but allows for a slightly wider error bound characterised via the $max\_error\_op_F$ parameters. The parameters $max\_error\_op_F$ must be documented by the implementation for each such parameter required by LIA-2. A comparison of the values of these parameters with the values of the specified maximum value for each such parameter will give some indication of the "quality" of the routines provided. Further, a comparison of the values of this parameter for two versions of a frequently used operation will give some indication of the accuracy sacrifice made in order to gain performance.

Language bindings are free to modify the error limits provided in the specifications for the operations to meet the expected requirements of their users.

Material on the implementation of high accuracy operations is provided in for example [51, 53, 60].

B.5.3.2 Sign requirements

The requirements imply that the sign of the result or continuation value is to be reliable, except for the sign of an infinite result or continuation value, where except for a signed zero argument, it is often the case that one cannot determine the sign of the infinity. Still for sign symmetric mathematical functions, the approximating operation is also sign symmetric, including infinitary results.

B.5.3.3 Monotonicity requirements

A maximum error of 0.5 ulp implies that an approximation helper function must be a monotonic approximation to the mathematical function. When the maximum error is greater than 0.5 ulp, and the rounding is not directed, this is not automatically the case.

There is no general requirement that the approximation helper functions are *strictly* monotone on the same intervals on which the corresponding exact function is strictly monotone, however, since such a requirement cannot be made due to the fact that all floating point types are discrete, not continuous.

B.5.3.4 The *trans_result* helper function

B.5.3.5 Hypotenuse

The $hypot_F$ operation can produce an overflow only if both arguments have magnitudes very close to the overflow threshold. Care must be taken in its implementation to either avoid or properly handle overflows and underflows which might occur in squaring the arguments. The function approximated by this operation is mathematically equivalent to complex absolute value, which is needed in the calculation of the modulus and argument of a complex number. It is important for this application that an implementation satisfy the constraint on the magnitude of the result returned.

LIA-2 does not follow the recommendations in [53] and in [54] that

$$hypot_F(+\infty, \mathbf{qNaN}) = +\infty$$
$$hypot_F(-\infty, \mathbf{qNaN}) = +\infty$$
$$hypot_F(\mathbf{qNaN}, +\infty) = +\infty$$
$$hypot_F(\mathbf{qNaN}, -\infty) = +\infty$$

which are based on the claim that a **qNaN** represents an (unknown) real valued number. This claim is not always valid, though it may sometimes be.

B.5.3.6 Operations for exponentiations and logarithms

For all of the exponentiation operations, overflow occurs for sufficiently large values of the argument(s).

There is a problem for $power_F(x, y)$ if both x and y are zero:

- Ada raises an 'exception' for the operation that is close in semantics to $power_F$ when both arguments are zero, in accordance with the fact that 0^0 is mathematically undefined.
- The X/OPEN Portability Guide, as well as C9x, specifies for `pow(0,0)` a return value of 1, and no notification. This specification agrees with the recommendations in [51, 53, 54, 57].

The specification in LIA-2 follows Ada, and returns **invalid** for $power_F(0, 0)$, because of the risks inherent in returning a result which might be inappropriate for the application at hand. Note however, that $power_{FI}(0, 0)$ is 1, without any notification. The reason is that the limiting value for the corresponding mathematical function, when following either of the only two continuous paths, is 1. This also agrees with the Ada specification for a floating point value raised to a power in an integer datatype, as well as that for other programming languages which distinguish these operations.

Along any path defined by $y = k/\ln(x)$ the mathematical function x^y has the value e^k. It follows that some of the limiting values for x^y depend on the choice of k, and hence are undefined, as indicated in the specification.

The result of the $power_F$ operation is **invalid** for negative values of the base x. The reason is that the floating point exponent y might imply an implicit extraction of an even root of x, which would have a complex value for negative x. This constraint is explicit in Ada, and is widely imposed in existing numerical packages provided by vendors, as well as several other programming languages.

The arguments of $power_F$ are floating point numbers. No special treatment is provided for integer floating point values, which may be approximate. The cases for integer values of the arguments are covered by the operations $power_{FI}$ and $power_I$. In the example binding for C a specification for pow_F is suppied. pow_F combines $power_F$ and $power_{F\mathcal{Z}}$ in a way suitable for C's `pow` operation.

For implementations of the $power_F$ operation there is an accuracy problem with an algorithm based on the following, mathematically valid, identity:

$$x^y = r_F^{y \cdot \log_{r_F}(x)}$$

The integer part of the product $y \cdot \log_{r_F}(x)$ defines the exponent of the result and the fractional part defines the reduced argument. If the exponent is large, and one calculates p_F digits of this intermediate result, there will be fewer than p_F digits for the fraction. Thus, in order to obtain a reduced argument accurately rounded to p_F digits, it may be necessary to calculate an approximation to $y \cdot \log_{r_F}(x)$ to a few more than $\log_{r_F}(emax_F) + p_F$ base r_F digits.

In Ada95 the operation most close to $power_{FI}$ is specified to be computed by successive multiplications, for which the error in the evaluation increases linearly with the size of the exponent. In a strict Ada implementation there is no way that a prescribed error limit of a few ulps can be met for large exponents.

The special exponentiation operations, corresponding to 2^x and 10^x, have specifications which are minor variations on those for $exp_F(x)$. Accuracy and performance can be increased if they are specially coded, rather than evaluated as, e.g., $exp_F(mul_F(x, ln_F(2)))$ or $power_F(2, x)$. Similar comments hold for the base 2 and base 10 logarithmic operations.

The $expm1_F$ operation has two advantages: Firstly, $expm1_F(x)$ is much more accurate than $sub_F(exp_F(x), 1)$ when the exponent argument is close to zero. Secondly, the $expm1_F$ operation does not **underflow** for "very" negative exponent arguments, something which may be advantageous if **underflow** handling is slow, and high accuracy for "very" negative arguments is not needed. Note in addition that underflow is avoided for this operation. This can be done only since LIA-2 adds requirements beyond those of LIA-1 regarding minimum precision (see clause 4). If those extra requirements were not done, underflow would not be justifiably removable for this operation. Similar argumentation applies to $ln1p_F$.

Similarly, there are two advantages with the $power1pm1_F$ operation: Firstly, $power1pm1_F(b, x)$ is much more accurate than $sub_F(power_F(add_F(1, b), x), 1)$ when the exponent argument is close to zero. Secondly, the $power1pm1_F$ operation does not **underflow** for "very" negative exponent arguments (when the base is greater than 1), something which may be advantageous if **underflow** handling is slow, and high accuracy for "very" negative arguments is not needed.

The handling of infinites and negative zero as arguments to the exponentiation and logarithm operations, like for all other LIA operations, follow the principles for dealing with these values as explained in section B.5.2.

B.5.3.7 Operations for hyperbolic elementary functions

The hyperbolic sine operation, $sinh_F(x)$, will overflow if $|x|$ is in the immediate neighbourhood of $\ln(2 * fmax)$, or greater.

The hyperbolic cosine operation, $cosh_F(x)$, will overflow if $|x|$ is in the immediate neighbourhood of $\ln(2 * fmax)$, or greater.

The hyperbolic cotangent operation, $coth_F(x)$, has a pole at $x = 0$.

The inverse of cosh is double valued, the two possible results having the same magnitude with opposite signs. The value returned by $arccosh_F$ is always greater than or equal to 1.

The inverse hyperbolic tangent operation $arctanh_F(x)$ has poles at $x = +1$ and at $x = -1$.

The inverse hyperbolic cotangent operation $arccoth_F(x)$ has poles at $x = +1$ and at $x = -1$.

B.5.3.8 Introduction to operations for trigonometric elementary functions

The real trigonometric functions sin, cos, tan, cot, sec, and csc are all periodic. The period for sin, cos, sec, and csc is $2 \cdot \pi$ radians (360 degrees). The period for tan and cot is π radians (180 degrees). The mathematical trigonometric functions are perfectly periodic. Their numerical counterparts are not that perfect, for two reasons.

Firstly, the radian normalisation cannot be exact, even though it can be made very good given very many digits for the approximation(s) of π used in the angle normalisation, returning an offset from the nearest axis, and including guard digits. The unit argument normalisation, however, can be made exact regardless of the (non-zero and, in case $denorm_F = \textbf{false}$, not too small) unit and the original angle, returning only a plain angle in F. LIA-2 requires unit argument angle normalisation to be exact.

Secondly, the length of one revolution is of course constant, but the density of floating point values gets sparser (in absolute spacing rather than relative) the larger the magnitude of the values are. This means that the number of floating point values gets sparser per revolution the larger the magnitude of the angle value is. For this reason the notification **absolute_precision_underflow** is introduced, together with two parameters (one for radians and one for other angular units). This notification is given when the magnitude of the angle value is "too big". Exactly when the representable angle values get too sparse depends upon the application at hand, but LIA-2 gives a default value for the parameters that define the cut-off. LIA-2 also includes specifications for high accuracy angle normalisation operations. The angle normalisation operations give a result within minus half a cycle to plus half a cycle, unless the argument angular value is too big (or there is some other error).

Note that the **absolute_precision_underflow** notification is unrelated to any argument reduction problems. Argument reduction is (implicitly for radians, explicitly for other angular units) required by LIA-2 to be very accurate. But no matter how accurate the argument reduction is, floating point values are still sparser in absolute terms the larger the values are. Note also that any use of trigonometric operations for non-trigonometric purposes is out of scope for LIA-2.

LIA-2 includes angle normalisation operations, both for radians and for other angular units. The angle normalisation operations return a value within minus half a cycle and plus half a cycle. These operations should be used to keep the representation of angles at a high accuracy. The trigonometric operations return a result within about an ulp, and that high accuracy is wasted if the angular argument is not kept at a high accuracy too. LIA-2 also includes angle normalisation operations that can be used to maintain an even higher degree of accuracy, giving the offset from the nearest axis (though without any extra guard digits). To use these one needs to keep track of the nearest axis, which unfortunately complicates programs that use this latter method.

Note that $rad(x) = \arccos(\cos(x))$ if $\sin(x) > 0$ and $rad(x) = -\arccos(\cos(x))$ if $\sin(x) < 0$. The first part of $axis\_rad(x)$ indicates which axis is nearest to the angle x. The second part of $axis\_rad(x)$ is an angle offset from the axis that is nearest to the angle x. The second part of

$axis\_rad(x)$ is equal to $rad(x)$ if $\cos(x) \geqslant 1/\sqrt{2}$ (i.e. if the first part of $axis\_rad(x)$ is $(1, 0)$). More generally, the second part of $axis\_rad(x)$ is equal to $rad(4 \cdot x)/4$.

$rad(x)$ returns the same angle as the angle value x, but the returned angle value is between $-\pi$ and π. The *rad* function is defined to be used as the basis for the angle normalisation operations. The *axis_rad* function is defined to be used as the basis for a numerically more accurate radian angle normalisation operation. The *arc* function is defined to be used as the basis for the arcus (angle) operations, which are used for conversion from Cartesian to polar co-ordinates.

B.5.3.9 Operations for radian trigonometric elementary functions

The radian trigonometric approximation helper functions (including those for normalisation and conversion from radians) are required to have the same zero points as the approximated mathematical function only if the absolute value of the argument is less than or equal to $big\_angle\_r_F$. Likewise, the radian trigonometric approximation helper functions are required to have the same sign as the approximated mathematical function only if the absolute value of the argument is less than or equal to $big\_angle\_r_F$.

The $big\_angle\_r_F$ parameter may be adjusted by bindings, or even by some compiler flag, or mode setting within a program. However, this method should only allow the value of this parameter to be set to a value greater than $2 \cdot \pi$, so that at least arguments within the first two (plus and minus) cycles are allowed, and such that $ulp_F(big\_angle\_r_F) < \pi/1000$, so that at least 2000 evenly distributed points within the 'last' cycle (farthest away from 0) are distinguishable. The latter gives a rather low accuracy at the far ends of the range, especially if p_F is comparatively large, so values this large for $big\_angle\_r_F$ are not recommendable unless the application is such that high accuracy trigonometric operations are not needed.

For reduction of an argument given in radians, implementations use one or several approximate value(s) of π (or of a multiple of π), valid to, say, n digits. The division implied in the argument reduction cannot be valid to more than n digits, which implies a maximum absolute angle value for which the reduction yields an accurate reduced angle value.

Regarding argument reduction for radians, there is a particular problem when the result of the trigonometric operation is very small (or very big), but the angular argument is not very small. In such cases the argument reduction must be very accurate, using an extra-precise approximation to π, relative to what is normally used for arguments of similar magnitude, so that significant digits in the result are not lost. Such loss would imply non-conformance to LIA-2 by the error in the final result being greater than that specified by LIA-2. In general, extra care has to be taken when the second part of $axis\_rad(x)$ is close to 0.

Note that if $big\_angle\_r_F$ is allowed to be increased, then, for conformity with LIA-2, the radian angle reduction may need to be more precise.

- tan and sec have *poles* at odd multiples of $\pi/2$ radians (90 degrees).
- cot and csc have *poles* at multiples of π radians (180 degrees).

All four of the corresponding operations with poles may produce **overflow** for arguments sufficiently close to the poles of the functions. The tan_F operation produces no **pole** notification. The reason is that the poles of $\tan(x)$ are at odd multiples of $\pi/2$, which are not representable in F. The mathematical cotangent function has a pole at the origin. For a system which supports signed zeros and infinities, the continuation values are $\mathbf{+\infty}$ and $\mathbf{-\infty}$ for arguments of 0 and $\mathbf{-0}$ respectively. Although the mathematical function sec has poles at odd multiples of $\pi/2$, the sec_F operation will not generate any pole notification because such arguments are not representable in F.

The **pole** notification cannot occur for any non-zero argument in radians because π is not representable in F, nor is $\pi/2$. For the angular unit argument trigonometric operations a the sign of the infinitary continuation value has been chosen arbitrarily for a **pole** which occurs for a non-zero argument. However, sign symmetry, when appropriate, is maintained.

The operations may produce **underflow** for arguments sufficiently close to their zeros. For a denormalised argument x, the sin_F, tan_F, $arcsin_F$, and $arctan_F$ return x for the result, with very high accuracy. Similarly, for a denormalised argument, cos_F and sec_F can return a result of 1.0 with very high accuracy.

The trigonometric inverses are multiple valued. They are rendered single valued by defining a principal value range. This range is closely related to a branch cut in the complex plane for the corresponding complex function. Among the floating point numerical functions this branch cut is "visible" only for the arc_F operation. The arc function has a branch cut along the negative real axis. For $x < 0$ the function has a discontinuity from $-\pi$ to $+\pi$ as y passes through zero from negative to positive values. Thus for $x < 0$, systems supporting signed zeros can handle the discontinuity as follows:

$$arc_F(x, \mathbf{-0}) = up_F(-\pi)$$
$$arc_F(x, 0) = down_F(\pi)$$

There is a problem for zero argument values for this operation. The values given for the operation $arc_F(x, y)$ for the four combinations of signed zeros for x and y are those given in [53]. The following table of values is given in [53] for the value of $arc_F(x, y)$ with both of the arguments zero:

Zero arguments		
x	y	$arc_F(x,y)$
0	0	0
$\mathbf{-0}$	0	π
$\mathbf{-0}$	$\mathbf{-0}$	$-\pi$
0	$\mathbf{-0}$	$\mathbf{-0}$

Note that the mathematical arc function is indeterminate (undefined) for (0,0), but these result are numerically more useful than giving an **invalid** notification for such arguments. LIA-2 therefore specifies results as above.

There is also a problem for argument values of $\mathbf{+\infty}$ or $\mathbf{-\infty}$ for this operation. The following table of values is given in [53] for the value of $arc_F(x, y)$ with at least one of the arguments infinite:

Infinite arguments		
x	y	$arc_F(x,y)$
$\mathbf{+\infty}$	$\geqslant 0$	0
$\mathbf{+\infty}$	$\mathbf{+\infty}$	$\pi/4$
finite	$\mathbf{+\infty}$	$\pi/2$
$\mathbf{-\infty}$	$\mathbf{+\infty}$	$3 \cdot \pi/4$
$\mathbf{-\infty}$	$\geqslant 0$	π
$\mathbf{-\infty}$	$\mathbf{-0}$	$-\pi$
$\mathbf{-\infty}$	< 0	$-\pi$
$\mathbf{-\infty}$	$\mathbf{-\infty}$	$-3 \cdot \pi/4$
finite	$\mathbf{-\infty}$	$-\pi/2$
$\mathbf{+\infty}$	$\mathbf{-\infty}$	$-\pi/4$
$\mathbf{+\infty}$	< 0	$\mathbf{-0}$
$\mathbf{+\infty}$	$\mathbf{-0}$	$\mathbf{-0}$

If one of x and y is infinite and the other is finite, the result tabulated is consistent with that obtained by a conventional limiting process. However, the results of $\pi/4$, $-\pi/4$, $3 \cdot \pi/4$, and $-3 \cdot \pi/4$ corresponding to infinite values for both x and y, are of questionable validity, since only the quadrant is known, not the angle within the quadrant. However, these results are numerically more useful than giving an **invalid** notification for such arguments. LIA-2 therefore specifies results as above.

B.5.3.10 Operations for trigonometrics given angular unit

At present only Ada specifies trigonometric operations with angular unit argument. LIA-2 has adopted angular unit argument operations in order to encourage uniformity among languages which might include such operations in the future. The angular units in T appear to be particularly important and have therefore been given a tighter error bound requirement. An implementation can of course have the same (tighter) error bound for all angular units.

The $big\_angle\_u_F$ parameter may be adjusted by bindings, or even by some compiler flag, or mode setting within a program. However, this method should only allow the value of this parameter to be set to a value greater than or equal to 1, so that at least arguments within the first two (plus and minus) cycles are allowed, and such that $ulp_F(big\_angle\_u_F) \leqslant 1/2000$, so that at least 2000 evenly distributed points within the 'last' cycle (farthest away from 0) are distinguishable. The latter gives a rather low accuracy at the far ends of the range, especially if p_F is comparatively large, so values this large for $big\_angle\_u_F$ are not recommendable unless the application is such that high accuracy trigonometric operations are not needed.

The $min\_angular\_unit_F$ parameter is specified for two reasons. Firstly, if the type F has no denormal values ($denorm_F$ = **false**), some angle values in F are not representable after normalisation if the angular unit has too small magnitude (this gives the firm limit above). Secondly, even if F has denormal values ($denorm_F$ = **true**), angular units with very small magnitude do not allow the representable angles to be particularly dense, not even if the angular value is within the first cycle. This does in itself not give rise to a particular limit value, but the limit value defined here is reasonable.

All of the argument angular unit trigonometric, and argument angular unit inverse trigonometric, approximation helper functions, including those for normalisation, angular unit conversion, and arc, are exempted from the monotonicity requirement for the angular unit argument.

If the angular unit argument, u, is such that $u/4 \in F$, the $tanu_F$ operation has poles at odd multiples of $u/4$. This is the case for degrees ($u = 360$). As for $tanu_F$, if the angular unit argument, u, is such that $u/4 \in F$ the $secu_F$ operation has poles at odd multiples of $u/4$.

The same comments hold for the $arcu_F$ operation as for arc_F operation, except that the discontinuity in the mathematical function is from $-u/2$ to $+u/2$.

B.5.3.11 Operations for angular-unit conversions

Conversion of an angular value x from angular unit a to angular unit b appears simple: compute $x \cdot b/a$. Basing a numerical conversion of angular values directly on the above mathematical equality (e.g. $div_F(mul_F(x,b),a)$) looses much absolute angular accuracy, however, especially for large angular values. Instead computing $arcu_F(b, cosu_F(a,x), sinu_F(a,x))$ gives a more accurate result. This might still not be within the accuracy required by LIA-2 for the angular unit conversion operations specified by LIA-2.

Note that all of the angular conversion operations return an angularly normalised result. This is in order to maintain high accuracy of the angle value being represented.

Angular conversion operations are commonly found on ‘scientific’ calculators and also in Java, though then only between degrees and radians.

B.5.4 Conversion operations

Clause 5.2 of LIA-1 covers conversions from an integer type to another integer type and to a floating point type, as well as between (LIA-1 conforming) floating point types of the same radix.

LIA-2 extends these conversions to cover conversions to and from non-LIA conforming datatypes, such as conversion to and from strings, and also extends the floating point conversion specifications to also handle conversions where the radices are different.

In ordinary string formats for numerals, the string “Hello world!” is an example of a signalling NaN.

LIA-2 does not specify any string formats, not even for the special values $\mathbf{-0}$, $\mathbf{+\infty}$, $\mathbf{-\infty}$, and quiet NaN, but possibilities for the special values include the strings used in the text of LIA-2, as well as strings like “+infinity” or “positiva oändligheten”, etc, and the strings used may depend on preference settings, as they may also for non-special values. E.g. one may use different notation for the decimal separator character (e.g., period, comma, Arabic comma, ...), use superscript digits for exponents, or use Arabic or traditional Thai digits. String formats for numerical values, and if and how they may depend on preference settings, is also an issue for bindings or programming language specifications, not for this part of LIA.

If the value converted is greater that those representable in the target, or less than those representable in the target, even after rounding, then an overflow will result. E.g., if the target is a character string of at most 3 digits, and the target radix is 10, then an integer source value of 1000 will result in an overflow. As for other operations, if the notification handling is by recording of indicators, a suitable continuation value must be used.

Most language standards contain (partial) format specifications for conversion to and from strings, usually for a decimal representation.

B.5.5 Numerals as operations in the programming language

B.5.5.1 Numerals for integer datatypes

Negative values (except $minint_I$ if $minint_I = -maxint_I - 1$) can be obtained by using the negation operation (neg_I).

Integer numerals in radix 10 are normally available in programming languages. Other radices may also be available for integer numerals, and the radix used may be part of determining the target integer datatype. E.g., radix 10 may be for signed integer datatypes, and radix 8 or 16 may be for unsigned integer datatypes.

Syntaxes for numerals for different integer datatypes need not be different, nor need they be the same. This Part does not further specify the format for integer numerals. That is an issue for bindings.

Overflow for integer numerals can be detected at “compile time”, and warned about.

B.5.5.2 Numerals for floating point datatypes

If the numerals used as operations in a program, and numerals read from other sources use the same radix, then “internal” numerals and “external” numerals (strings) denoting the same value in $\mathcal{R}$ and converted to the same target datatype should be converted to the same value.

Negative values (including negative 0, $\mathbf{-0}$) can be obtained by using the negation operation (neg_F).

Radices other than 10 may also be available for floating point numerals.

Integer numerals may also be floating point numerals, i.e. their syntaxes need not be different. Nor need syntaxes for numerals for different floating point datatypes be different, nor need they be the same. This Part does not specify the syntax for numerals. That is an issue for bindings or programming language specifications.

Overflow/underflow for floating point numerals can be detected at "compile time", and warned about.

B.6 Notification

An intermediate overflow on computing approximations to x^2 or y^2 during the calculation of $hypot_F(x, y) \approx \sqrt{x^2 + y^2})$ must not result in an overflow notification, unless the end result overflows. This is clear from the specification of the $hypot_F$ operation in this Part.

If a single argument operation op_F, for the corresponding mathematical function f, is such that $f(x)$ very closely approximates x, when $|x| \leqslant fminN_F$, then $op_F(x)$ returns x for $|x| \leqslant fminN_F$, and does not give a notification if there cannot be any denormalisation loss relative to $f(x)$. For details, see the individual operation specifications for $expm1_F$, $ln1p_F$, $sinh_F$, $arcsinh_F$, $tanh_F$, $arctanh_F$, sin_F, $arcsin_F$, tan_F, and $arctan_F$.

Operations specified in LIA-2 return **invalid**(**qNaN**) when passed a signaling **NaN** (**sNaN**) as an argument. Most operations specified in LIA-2 return **qNaN**, without any notification when passed a quiet **NaN** (**qNaN**) as an argument.

The different kinds of notifications occur under the following circumstances:

a) **invalid**: when an argument is not valid for the operation, and no value in F^* or any special value result makes mathematical sense.

b) **pole**: when the input operand corresponds to a pole of the mathematical function approximated by the operation.

c) **overflow**: when the (rounded) result is outside of the range of the result datatype.

d) **underflow**: when a sufficiently closely approximating result of the operation has a magnitude that is so small that it might not be sufficiently accurately represented in the result datatype.

e) **absolute_precision_underflow**: when the magnitude of the angle argument to a floating point trigonometric operation exceeds the maximum value of the argument for which the density of floating point values is deemed sufficient for the operation to make sense. See clause 5.3.8 and the associated discussion in this rationale (section B.5.3.8).

In order to avoid **absolute_precision_underflow** notifications, and to maintain a high accuracy, implementors are encouraged to provide, and programmers encouraged to use, the angle normalisation operations specified in 5.3.9.1 and 5.3.10.1.

The difference between the **pole** and **overflow** notifications is that the first corresponds to a true mathematical singularity, and the second corresponds to a well-defined mathematical result that happens to lie outside the range of F.

B.6.1 Continuation values

For handling of notifications, the method that does recording of indicators (LIA-1, clause 6.1.2) is preferred.

An implementation which supports recording of indicators must supply continuation values to be used when execution is continued following the occurrence of a notification. For systems which support signed zeros, infinities and **NaN**s, LIA-2 specifies how these values, as well as ordinary values, are used as continuation values. Other implementations which use recording of indicators must supply other suitable continuation values and document the values selected.

B.7 Relationship with language standards

An arithmetic expression might not be executed as written.

For example, if x is declared to be single precision (`SP`) floating point, and calculation is done in single precision, then the expression

`arcsin(x)`

might translate to

$arcsin_{SP}(\texttt{x})$

If the language in question did all computations in double precision (`DP`) floating point, the above expression might translate to

$arcsin_{DP}(cvt_{SP \rightarrow DP}(x))$

Alternatively, if `x` was declared to be an integer, and the expected result datatype is single precision float, the above expression might translate to

$cvt_{DP \rightarrow SP}(arcsin_{DP}(cvt_{I \rightarrow DP}(x)))$

The datatypes involved in implicit conversions need not be accessible to the programmer. For example, trigonometric operations may be evaluated in extended double precision, even though that datatype is not made available to programmers using a particular programming language. These extra datatypes should be made available, however, and the implicit conversions should be expressible as explicit conversions. At least in order to be able to show exactly which expression is going to be evaluated without having to look at the machine code.

B.8 Documentation requirements

Annex C
(informative)

Example bindings for specific languages

This annex describes how a computing system can simultaneously conform to a language standard (or publicly available specification) and to LIA-2. It contains suggestions for binding the "abstract" operations specified in LIA-2 to concrete language syntax. The format used for these example bindings in this annex is a short form version, suitable for the purposes of this annex. An actual binding is under no obligation to follow this format. An actual binding should, however, as in the bindings examples, give the LIA-2 operation name, or parameter name, bound to an identifier by the binding.

Portability of programs can be improved if two conforming LIA-2 systems using the same language agree in the manner with which they adhere to LIA-2. For instance, LIA-2 requires that the parameter $big\_angle\_r_F$ be provided (if any conforming radian trigonometric operations are provided), but if one system provides it by means of the identifier `BigAngle` and another by the identifier `MaxAngle`, portability is impaired. Clearly, it would be best if such names were defined in the relevant language standards or binding standards, but in the meantime, suggestions are given here to aid portability.

The following clauses are suggestions rather than requirements because the areas covered are the responsibility of the various language standards committees. Until binding standards are in place, implementors can promote "de facto" portability by following these suggestions on their own.

The languages covered in this annex are

Ada
Basic
C
C++
Fortran
Haskell
Java
Common Lisp
ISLisp
Modula-2
Pascal and Extended Pascal
PL/I
SML

This list is not exhaustive. Other languages and other computing devices (like 'scientific' calculators, 'web script' languages, and database 'query languages') are suitable for conformity to LIA-2.

In this annex, the parameters, operations, and exception behaviour of each language are examined to see how closely they fit the requirements of LIA-2. Where parameters, constants, or operations are not provided by the language, names and syntax are suggested. (Already provided syntax is marked with a $\star$.)

This annex describes only the language-level support for LIA-2. An implementation that wishes to conform must ensure that the underlying hardware and software is also configured to conform to LIA-2 requirements.

A complete binding for LIA-2 will include, or refer to, a binding for LIA-1. In turn, a complete binding for the LIA-1 may include, or refer to, a binding for IEC 60559. Such a joint LIA-2/LIA-1/IEC 60559 binding should be developed as a single binding standard. To avoid conflict with ongoing development, only the LIA-2 specific portions of such a binding are examplified in this annex.

C.1 General comments

Most language standards permit an implementation to provide, by some means, the parameters and operations required by LIA-2 that are not already part of the language. The method for accessing these additional parameters and operations depends on the implementation and language, and is not specified in LIA-2 nor examplified in this annex. It could include external subroutine libraries; new intrinsic functions supported by the compiler; constants and functions provided as global "macros"; and so on. The actual method of access through libraries, macros, etc. should of course be given in a real binding.

Most language standards do not constrain the accuracy of elementary numerical functions, or specify the subsequent behaviour after an arithmetic notification occurs.

In the event that there is a conflict between the requirements of the language standard and the requirements of LIA-2, the language binding standard should clearly identify the conflict and state its resolution of the conflict.

C.2 Ada

The programming language Ada is defined by ISO/IEC 8652:1995, *Information Technology – Programming Languages – Ada* [11], where the specifications for the operations for elementary functions are based on ISO/IEC 11430:1994 *Information technology – Programming languages – Generic package of elementary functions for Ada* [12].

An implementation should follow all the requirements of LIA-2 unless otherwise specified by this language binding.

The operations or parameters marked "†" are not part of the language and should be provided by an implementation that wishes to conform to LIA-2 for that operation or parameter. For each of the marked items a suggested identifier is provided.

The Ada datatype `Boolean` corresponds to the LIA datatype **Boolean**.

Every implementation of Ada has at least one integer datatype, and at least one floating point datatype. The notations *INT* and *FLT* are used to stand for the names of one of these datatypes in what follows.

The LIA-2 integer operations are listed below, along with the syntax used to invoke them:

$max_I(x, y)$	*INT*'Max(*x*, *y*)	⋆
$min_I(x, y)$	*INT*'Min(*x*, *y*)	⋆
$max\_seq_I(xs)$	Max(*xs*)	†
$min\_seq_I(xs)$	Min(*xs*)	†
$dim_I(x, y)$	Dim(*x*, *y*)	†
$power_I(x, y)$	*x* ** *y*	⋆
$shift2_I(x, y)$	Shift2(*x*, *y*)	†
$shift10_I(x, y)$	Shift10(*x*, *y*)	†
$sqrt_I(x)$	Sqrt(*x*)	†

$\mathit{divides}_I(x,y)$	`Divides(`x`, `y`)`	†
$\mathit{even}_I(x)$	x `mod 2 = 0`	⋆
$\mathit{odd}_I(x)$	x `mod 2 /= 0`	⋆
$\mathit{divf}_I(x,y)$	`Div(`x`, `y`)`	†
$\mathit{moda}_I(x,y)$	x `mod` y	⋆
$\mathit{group}_I(x,y)$	`Group(`x`, `y`)`	†
$\mathit{pad}_I(x,y)$	`Pad(`x`, `y`)`	†
$\mathit{quot}_I(x,y)$	`Quot(`x`, `y`)`	†
$\mathit{remr}_I(x,y)$	`Rem(`x`, `y`)`	†
$\mathit{gcd}_I(x,y)$	`Gcd(`x`, `y`)`	†
$\mathit{lcm}_I(x,y)$	`Lcm(`x`, `y`)`	†
$\mathit{gcd\_seq}_I(xs)$	`Gcd(`xs`)`	†
$\mathit{lcm\_seq}_I(xs)$	`Lcm(`xs`)`	†
$\mathit{add\_wrap}_I(x,y)$	`Add_Wrap(`x`, `y`)`	†
$\mathit{add\_ov}_I(x,y)$	`Add_Over(`x`, `y`)`	†
$\mathit{sub\_wrap}_I(x,y)$	`Sub_Wrap(`x`, `y`)`	†
$\mathit{sub\_ov}_I(x,y)$	`Sub_Over(`x`, `y`)`	†
$\mathit{mul\_wrap}_I(x,y)$	`Mul_Wrap(`x`, `y`)`	†
$\mathit{mul\_ov}_I(x,y)$	`Mul_Over(`x`, `y`)`	†

where x and y are expressions of type *INT* and where xs is an expression of type `array of` *INT*.

The LIA-2 basic floating point operations are listed below, along with the syntax used to invoke them:

$\mathit{max}_F(x,y)$	*FLT*`'Max(`x`, `y`)`	⋆
$\mathit{min}_F(x,y)$	*FLT*`'Min(`x`, `y`)`	⋆
$\mathit{mmax}_F(x,y)$	`MMax(`x`, `y`)`	†
$\mathit{mmin}_F(x,y)$	`MMin(`x`, `y`)`	†
$\mathit{max\_seq}_F(xs)$	`Max(`xs`)`	†
$\mathit{min\_seq}_F(xs)$	`Min(`xs`)`	†
$\mathit{mmax\_seq}_F(xs)$	`MMax(`xs`)`	†
$\mathit{mmin\_seq}_F(xs)$	`MMin(`xs`)`	†
$\mathit{dim}_F(x,y)$	`Dim(`x`, `y`)`	†
$\mathit{rounding}_F(x)$	*FLT*`'Unbiased_Rounding(`x`)`	⋆
$\mathit{floor}_F(x)$	*FLT*`'Floor(`x`)`	⋆
$\mathit{ceiling}_F(x)$	*FLT*`'Ceiling(`x`)`	⋆
$\mathit{rounding\_rest}_F(x)$	x `-` *FLT*`'Unbiased_Rounding(`x`)`	⋆
$\mathit{floor\_rest}_F(x)$	x `-` *FLT*`'Floor(`x`)`	⋆
$\mathit{ceiling\_rest}_F(x)$	x `-` *FLT*`'Ceiling(`x`)`	⋆
$\mathit{remr}_F(x,y)$	*FLT*`'Remainder(`x`, `y`)`	⋆
$\mathit{sqrt}_F(x)$	`Sqrt(`x`)`	⋆
$\mathit{rsqrt}_F(x)$	`RSqrt(`x`)`	†
$\mathit{add\_lo}_F(x,y)$	`Add_Low(`x`, `y`)`	†
$\mathit{sub\_lo}_F(x,y)$	`Sub_Low(`x`, `y`)`	†
$\mathit{mul\_lo}_F(x,y)$	`Mul_Low(`x`, `y`)`	†
$\mathit{div\_rest}_F(x,y)$	`Div_Rest(`x`, `y`)`	†
$\mathit{sqrt\_rest}_F(x)$	`Sqrt_Rest(`x`)`	†
$\mathit{mul}_{F \to F'}(x,y)$	`Prod(`x`, `y`)`	†

where x, y, and z are expressions of type *FLT*, and where xs is an expression of type `array of`

FLT.

The parameters for LIA-2 operations approximating real valued transcendental functions can be accessed by the following syntax:

$max\_err\_hypot_F$	`Err_Hypotenuse(`x`)`	†
$max\_err\_exp_F$	`Err_Exp(`x`)`	†
$max\_err\_power_F$	`Err_Power(`x`)`	†
$max\_err\_sinh_F$	`Err_Sinh(`x`)`	†
$max\_err\_tanh_F$	`Err_Tanh(`x`)`	†
$big\_angle\_r_F$	`Big_Radian_Angle(`x`)`	†
$max\_err\_sin_F$	`Err_Sin(`x`)`	†
$max\_err\_tan_F$	`Err_Tan(`x`)`	†
$min\_angular\_unit_F$	`Smallest_Angular_Unit(`x`)`	†
$big\_angle\_u_F$	`Big_Angle(`x`)`	†
$max\_err\_sinu_F(u)$	`Err_Sin_Cycle(`u`)`	†
$max\_err\_tanu_F(u)$	`Err_Tan_Cycle(`u`)`	†
$max\_err\_convert_F$	`Err_Convert(`x`)`	†
$max\_err\_convert_{F'}$	`Err_Convert_To_String`	†
$max\_err\_convert_{D'}$	`Err_Convert_To_String`	†

where x and u are expressions of type *FLT*. Several of the parameter functions are constant for each type (and library), the argument is then used only to differentiate among the floating point types.

The LIA-2 elementary floating point operations are listed below, along with the syntax used to invoke them:

$hypot_F(x, y)$	`Hypotenuse(`x`, `y`)`	†
$power_{FI}(b, z)$	b `**` z	⋆
$exp_F(x)$	`Exp(`x`)`	⋆
$expm1_F(x)$	`ExpM1(`x`)`	†
$exp2_F(x)$	`Exp2(`x`)`	†
$exp10_F(x)$	`Exp10(`x`)`	†
$power_F(b, y)$	b `**` y	⋆
$power1pm1_F(b, y)$	`Power1PM1(`b`, `y`)`	†
$ln_F(x)$	`Log(`x`)`	⋆
$ln1p_F(x)$	`Log1P(`x`)`	†
$log2_F(x)$	`Log2(`x`)`	†
$log10_F(x)$	`Log10(`x`)`	†
$logbase_F(b, x)$	`Log(`x`, `b`)` (note parameter order)	⋆
$logbase1p1p_F(b, x)$	`Log1P1P(`x`, `b`)`	†
$sinh_F(x)$	`SinH(`x`)`	⋆
$cosh_F(x)$	`CosH(`x`)`	⋆
$tanh_F(x)$	`TanH(`x`)`	⋆
$coth_F(x)$	`CotH(`x`)`	⋆
$sech_F(x)$	`SecH(`x`)`	†

$csch_F(x)$	CscH(x)	†
$arcsinh_F(x)$	ArcSinH(x)	⋆
$arccosh_F(x)$	ArcCosH(x)	⋆
$arctanh_F(x)$	ArcTanH(x)	⋆
$arccoth_F(x)$	ArcCotH(x)	⋆
$arcsech_F(x)$	ArcSecH(x)	†
$arccsch_F(x)$	ArcCscH(x)	†
$rad_F(x)$	Rad(x)	†
$axis\_rad_F(x)$	Rad(x, h, v) (note out parameters)	†
$sin_F(x)$	Sin(x)	⋆
$cos_F(x)$	Cos(x)	⋆
$tan_F(x)$	Tan(x)	⋆
$cot_F(x)$	Cot(x)	⋆
$sec_F(x)$	Sec(x)	†
$csc_F(x)$	Csc(x)	†
$cossin_F(x)$	CosSin(x, c, s) (note out parameters)	†
$arcsin_F(x)$	ArcSin(x)	⋆
$arccos_F(x)$	ArcCos(x)	⋆
$arctan_F(x)$	ArcTan(x)	⋆
$arccot_F(x)$	ArcCot(x)	⋆
$arcctg_F(x)$	ArcCtg(x)	†
$arcsec_F(x)$	ArcSec(x)	†
$arccsc_F(x)$	ArcCsc(x)	†
$arc_F(x,y)$	ArcTan(y, x) or ArcCot(x, y)	⋆
$cycle_F(u,x)$	Cycle(x, u) (note parameter order)	†
$axis\_cycle_F(u,x)$	Cycle(x, u, h, v)	†
$sinu_F(u,x)$	Sin(x, u) (note parameter order)	⋆
$cosu_F(u,x)$	Cos(x, u)	⋆
$tanu_F(u,x)$	Tan(x, u)	⋆
$cotu_F(u,x)$	Cot(x, u)	⋆
$secu_F(u,x)$	Sec(x, u)	†
$cscu_F(u,x)$	Csc(x, u)	†
$cossinu_F(u,x)$	CosSin(x, u, c, s)	†
$arcsinu_F(u,x)$	ArcSin(x, u)	⋆
$arccosu_F(u,x)$	ArcCos(x, u)	⋆
$arctanu_F(u,x)$	ArcTan(x, Cycle=>u)	⋆
$arccotu_F(u,x)$	ArcCot(x, Cycle=>u)	⋆
$arcctgu_F(u,x)$	ArcCtg(x, u)	†
$arcsecu_F(u,x)$	ArcSec(x, u)	†
$arccscu_F(u,x)$	ArcCsc(x, u)	†
$arcu_F(u,x,y)$	ArcTan(y, x, u) or ArcCot(x, y, u)	⋆
$rad\_to\_cycle_F(x,u)$	Rad_to_Cycle(x, u)	†
$cycle\_to\_rad_F(u,x)$	Cycle_to_Rad(u, x)	†

$\mathit{cycle\_to\_cycle}_F(u, x, v)$	`Cycle_to_Cycle(`*u*`, `*x*`, `*v*`)`	†

where b, x, y, u, and v are expressions of type *FLT*, and z is an expressions of type *INT*.

Arithmetic value conversions in Ada are always explicit and usually use the destination datatype name as the name of the conversion function, except when converting to/from strings.

$\mathit{convert}_{I\to I'}(x)$	*INT2*`(`*x*`)`	⋆
$\mathit{convert}_{I''\to I}(s)$	`Get(`*s*`,`*n*`,`*w*`)`	⋆
$\mathit{convert}_{I\to I''}(x)$	`Put(`*s*`,`*x*`,`*base*`?)`	⋆
$\mathit{convert}_{I''\to I}(f)$	`Get(`*f*`?,`*n*`,`*w*`?)`	⋆
$\mathit{convert}_{I\to I''}(x)$	`Put(`*h*`?,`*x*`,`*w*`?,`*base*`?)`	⋆
$\mathit{rounding}_{F\to I}(y)$	*INT*`(`*FLT*`'Unbiased_Rounding(`*y*`))`	⋆
$\mathit{floor}_{F\to I}(y)$	*INT*`(`*FLT*`'Floor(`*y*`))`	⋆
$\mathit{ceiling}_{F\to I}(y)$	*INT*`(`*FLT*`'Ceiling(`*y*`))`	⋆
$\mathit{convert}_{I\to F}(x)$	*FLT*`(`*x*`)`	⋆
$\mathit{convert}_{F\to F'}(y)$	*FLT2*`(`*y*`)`	⋆
$\mathit{convert}_{F''\to F}(s)$	`Get(`*s*`,`*n*`,`*w*`?)`	⋆
$\mathit{convert}_{F''\to F}(f)$	`Get(`*f*`?,`*n*`,`*w*`?)`	⋆
$\mathit{convert}_{F\to F''}(y)$	`Put(`*s*`,`*y*`,Aft=>`*a*`?,Exp=>`*e*`?)`	⋆
$\mathit{convert}_{F\to F''}(y)$	`Put(`*h*`?,`*y*`,Fore=>`*i*`?,Aft=>`*a*`?,Exp=>`*e*`?)`	⋆
$\mathit{convert}_{D\to F}(z)$	*FLT*`(`*z*`)`	⋆
$\mathit{convert}_{D'\to F}(s)$	`Get(`*s*`,`*n*`,`*w*`?)`	⋆
$\mathit{convert}_{D'\to F}(f)$	`Get(`*f*`?,`*n*`,`*w*`?)`	⋆
$\mathit{convert}_{F\to D}(y)$	*FXD*`(`*y*`)`	⋆
$\mathit{convert}_{F\to D'}(y)$	`Put(`*s*`,`*y*`,Aft=>`*a*`?,Exp=>0)`	⋆
$\mathit{convert}_{F\to D'}(y)$	`Put(`*h*`?,`*y*`,Fore=>`*i*`?,Aft=>`*a*`?,Exp=>0)`	⋆

where x is an expression of type *INT*, y is an expression of type *FLT*, and z is an expression of type *FXD*, where *FXD* is a fixed point type. *INT2* is the integer datatype that corresponds to I'. A ? above indicates that the parameter is optional. e is greater than 0.

Ada provides non-negative numerals for all its integer and floating point types. The default base is 10, but all bases from 2 to 16 can be used. There is no differentiation between the numerals for different floating point types, nor between numerals for different integer types, but integer numerals (without a point) cannot be used for floating point types, and 'real' numerals (with a point) cannot be used for integer types. Integer numerals can have an exponent part though. The details are not repeated in this example binding, see ISO/IEC 8652:1995, clause 2.4 Numeric Literals, clause 3.5.4 Integer Types, and clause 3.5.6 Real Types.

The Ada standard does not specify any numerals for infinities and NaNs. Suggestion:

$+\infty$	*FLT*`'Infinity`	†
qNaN	*FLT*`'NaN`	†
sNaN	*FLT*`'SigNaN`	†

as well as string formats for reading and writing these values as character strings.

Ada has a notion of 'exception' that implies a non-returnable, but catchable, change of control flow. Ada uses its exception mechanism as its default means of notification. Ada ignores **underflow** notifications since an Ada exception is inappropriate for an **underflow** notification. On **underflow** the continuation value (specified in LIA-2) is used directly without recording the **underflow** itself. Ada uses the exception `Constraint_Error` for **pole** and **overflow** notifications, and the exception `Numerics.Argument_Error` for **invalid** notifications. Since Ada exceptions are non-returnable changes of control flow, no continuation value is provided for these notifications.

An implementation that wishes to follow LIA-2 should provide recording of indicators as an alternative means of handling numeric notifications. Recording of indicators is the LIA-2 preferred means of handling numeric notifications.

C.3 BASIC

The programming language BASIC is defined by ANSI X3.113-1987 (R1998) [41], endorsed by ISO/IEC 10279:1991, *Information technology – Programming languages – Full BASIC* [17].

An implementation should follow all the requirements of LIA-2 unless otherwise specified by this language binding.

The operations or parameters marked "†" are not part of the language and should be provided by an implementation that wishes to conform to the LIA-2 for that operation. For each of the marked items a suggested identifier is provided.

BASIC has no user accessible datatype corresponding to the LIA datatype **Boolean**.

BASIC has one primitive computational data type, `numeric`. The model presented by the BASIC language is that of a real number with decimal radix and a specified (minimum) number of significant decimal digits. Numeric data is not declared directly, but any special characteristics are inferred from how they are used and from any `OPTIONS` that are in force.

The BASIC statement `OPTION ARITHMETIC NATIVE` ties the `numeric` type more closely to the underlying implementation. The precision and type of `NATIVE` numeric data is implementation dependent.

Since the BASIC numeric data type does not match the integer type required by the LIA-1, this binding example does not include any of the LIA-2 operations for integer data types.

The LIA-2 non-transcendental floating point operations are listed below, along with the syntax used to invoke them:

$max_F(x, y)$	MAX(x, y)	†
$min_F(x, y)$	MIN(x, y)	†
$mmax_F(x, y)$	MMAX(x, y)	†
$mmin_F(x, y)$	MMIN(x, y)	†
$max\_seq_F(xs)$	MAXS(xs)	†
$min\_seq_F(xs)$	MINS(xs)	†
$mmax\_seq_F(xs)$	MMAXS(xs)	†
$mmin\_seq_F(xs)$	MMINS(xs)	†
$dim_F(x, y)$	MONUS(x, y)	†
$rounding_F(x)$	ROUNDING(x)	†
$floor_F(x)$	FLOOR(x)	⋆
$ceiling_F(x)$	CEILING(x)	†
$rounding\_rest_F(x)$	x - ROUNDING(x)	†
$floor\_rest_F(x)$	x - FLOOR(x)	⋆
$ceiling\_rest_F(x)$	x - CEILING(x)	†
$remr_F(x, y)$	REMAINDER(x, y)	†
$sqrt_F(x)$	SQRT(x)	⋆
$rsqrt_F(x)$	RSQRT(x)	†
$add\_lo_F(x, y)$	ADD_LOW(x, y)	†

$\mathit{sub\_lo}_F(x, y)$	SUB_LOW(x, y)	†
$\mathit{mul\_lo}_F(x, y)$	MUL_LOW(x, y)	†
$\mathit{div\_rest}_F(x, y)$	DIV_REST(x, y)	†
$\mathit{sqrt\_rest}_F(x)$	SQRT_REST(x)	†

where x, y, and z are expressions of type `numeric`, and where xs is an expression of type array of `numeric`.

The LIA-2 parameters for operations approximating real valued transcendental functions can be accessed by the following syntax:

$\mathit{max\_err\_hypot}_F$	ERR_HYPOTENUSE	†
$\mathit{max\_err\_exp}_F$	ERR_EXP	†
$\mathit{max\_err\_power}_F$	ERR_POWER	†
$\mathit{max\_err\_sinh}_F$	ERR_SINH	†
$\mathit{max\_err\_tanh}_F$	ERR_TANH	†
$\mathit{big\_angle\_r}_F$	BIG_RADIAN_ANGLE	†
$\mathit{max\_err\_sin}_F$	ERR_SIN	†
$\mathit{max\_err\_tan}_F$	ERR_TAN	†
$\mathit{min\_angular\_unit}_F$	MIN_ANGLE_UNIT	†
$\mathit{big\_angle\_u}_F$	BIG_ANGLE	†
$\mathit{max\_err\_sinu}_F(u)$	ERR_SIN_CYCLE(u)	†
$\mathit{max\_err\_tanu}_F(u)$	ERR_TAN_CYCLE(u)	†
$\mathit{max\_err\_convert}_F$	ERR_CONVERT	†
$\mathit{max\_err\_convert}'_F$	ERR_CONVERT_TO_STRING	†
$\mathit{max\_err\_convert}_D$	ERR_CONVERT_TO_STRING	†

where u is an expression of type `numeric`.

The LIA-2 floating point operations are listed below, along with the syntax used to invoke them. BASIC has a degree mode and a radian mode for the trigonometric operations.

$\mathit{hypot}_F(x, y)$	HYPOT(x, y)	†
$\mathit{exp}_F(x)$	EXP(x)	⋆
$\mathit{expm1}_F(x)$	EXPM1(x)	†
$\mathit{exp2}_F(x)$	EXP2(x)	†
$\mathit{exp10}_F(x)$	EXP10(x)	†
$\mathit{power}_F(b, y)$	POWER(b, y)	†
$\mathit{power1pm1}_F(b, y)$	POWER1PM1(b, y)	†
$\mathit{ln}_F(x)$	LOG(x)	⋆
$\mathit{ln1p}_F(x)$	LOG1P(x)	†
$\mathit{log2}_F(x)$	LOG2(x)	⋆
$\mathit{log10}_F(x)$	LOG10(x)	⋆
$\mathit{logbase}_F(b, x)$	LOGBASE(b, x)	†
$\mathit{logbase1p1p}_F(b, x)$	LOGBASE1P1P(b, x)	†
$\mathit{sinh}_F(x)$	SINH(x)	⋆
$\mathit{cosh}_F(x)$	COSH(x)	⋆
$\mathit{tanh}_F(x)$	TANH(x)	⋆
$\mathit{coth}_F(x)$	COTH(x)	†

$sech_F(x)$	SECH(x)	†
$csch_F(x)$	CSCH(x)	†
$arcsinh_F(x)$	ARCSINH(x)	†
$arccosh_F(x)$	ARCCOSH(x)	†
$arctanh_F(x)$	ARCTANH(x)	†
$arccoth_F(x)$	ARCCOTH(x)	†
$arcsech_F(x)$	ARCSECH(x)	†
$arccsch_F(x)$	ARCCSCH(x)	†
$rad_F(x)$	NORMANGLE(x) (when in radian mode)	†
$sin_F(x)$	SIN(x) (when in radian mode)	⋆
$cos_F(x)$	COS(x) (when in radian mode)	⋆
$tan_F(x)$	TAN(x) (when in radian mode)	⋆
$cot_F(x)$	COT(x) (when in radian mode)	⋆
$sec_F(x)$	SEC(x) (when in radian mode)	†
$csc_F(x)$	CSC(x) (when in radian mode)	†
$arcsin_F(x)$	ARCSIN(x) (when in radian mode)	⋆
$arccos_F(x)$	ARCCOS(x) (when in radian mode)	⋆
$arctan_F(x)$	ARCTAN(x) (when in radian mode)	⋆
$arccot_F(x)$	ARCCOT(x) (when in radian mode)	⋆
$arcctg_F(x)$	ARCCTG(x) (when in radian mode)	†
$arcsec_F(x)$	ARCSEC(x) (when in radian mode)	†
$arccsc_F(x)$	ARCCSC(x) (when in radian mode)	†
$arc_F(x, y)$	ANGLE(x,y) (when in radian mode)	⋆
$cycle_F(u, x)$	NORMANGLEU(u,x)	†
$sinu_F(u, x)$	SINU(u,x)	†
$cosu_F(u, x)$	COSU(u,x)	†
$tanu_F(u, x)$	TANU(u,x)	†
$cotu_F(u, x)$	COTU(u,x)	†
$secu_F(u, x)$	SECU(u,x)	†
$cscu_F(u, x)$	CSCU(u,x)	†
$arcsinu_F(u, x)$	ARCSINU(u,x)	†
$arccosu_F(u, x)$	ARCCOSU(u,x)	†
$arctanu_F(u, x)$	ARCTANU(u,x)	†
$arccotu_F(u, x)$	ARCCOTU(u,x)	†
$arcctgu_F(u, x)$	ARCCTGU(u,x)	†
$arcsecu_F(u, x)$	ARCSECU(u,x)	†
$arccscu_F(u, x)$	ARCCSCU(u,x)	†
$arcu_F(u, x, y)$	ANGLEU(u,x,y)	†
$cycle_F(360, x)$	NORMANGLE(x) (when in degree mode)	†
$sinu_F(360, x)$	SIN(x) (when in degree mode)	⋆
$cosu_F(360, x)$	COS(x) (when in degree mode)	⋆
$tanu_F(360, x)$	TAN(x) (when in degree mode)	⋆

$cotu_F(360, x)$	`COT(x)` (when in degree mode)	⋆
$secu_F(360, x)$	`SEC(x)` (when in degree mode)	†
$cscu_F(360, x)$	`CSC(x)` (when in degree mode)	†

$arcsinu_F(360, x)$	`ARCSIN(x)` (when in degree mode)	⋆
$arccosu_F(360, x)$	`ARCCOS(x)` (when in degree mode)	⋆
$arctanu_F(360, x)$	`ARCTAN(x)` (when in degree mode)	⋆
$arccotu_F(360, x)$	`ARCCOT(x)` (when in degree mode)	⋆
$arcctgu_F(360, x)$	`ARCCTG(x)` (when in degree mode)	†
$arcsecu_F(360, x)$	`ARCSEC(x)` (when in degree mode)	†
$arccscu_F(360, x)$	`ARCCSC(x)` (when in degree mode)	†
$arcu_F(360, x, y)$	`ANGLE(x,y)` (when in degree mode)	⋆

$rad\_to\_cycle_F(x, u)$	`RAD_TO_CYCLE(x, u)`	†
$cycle\_to\_rad_F(u, x)$	`CYCLE_TO_RAD(u, x)`	†
$cycle\_to\_cycle_F(u, x, v)$	`CYCLE_TO_CYCLE(u, x, v)`	†

where b, x, y, u, and v are expressions of type `numeric`.

Arithmetic value conversions in BASIC are always tied to reading and writing text.

$convert_{F''\to F}(stdin)$	`READ x`	⋆
$convert_{F\to F''}(y)$	`PRINT y`	⋆
$convert_{D'\to F}(stdin)$	`READ x`	⋆

where x is a variable of type `numeric`, y is an expression of type `numeric`.

BASIC provides non-negative numerals for `numeric` in base 10.

BASIC does not specify any numerals for infinities and NaNs. Suggestion:

+∞	`INFINITY`	†
qNaN	`NAN`	†
sNaN	`SIGNAN`	†

as well as string formats for reading and writing these values as character strings.

BASIC has a notion of 'exception' that implies a non-returnable change of control flow. BASIC uses its exception mechanism as its default means of notification. BASIC ignores **underflow** notifications since a BASIC exception is inappropriate for an **underflow** notification. On **underflow** the continuation value (specified in LIA-2) is used directly without recording the **underflow** itself. BASIC uses the exception number `1003` (Numeric supplied function overflow) for **overflow**, the exception number `3001` (Division by zero) for **pole**, and the exception numbers `301x`(?) for **invalid**. Since BASIC exceptions are non-returnable changes of control flow, no continuation value is provided for these notifications.

An implementation that wishes to follow LIA-2 should provide recording of indicators as an alternative means of handling numeric notifications. Recording of indicators is the LIA-2 preferred means of handling numeric notifications.

C.4 C

The programming language C is defined by ISO/IEC 9899:1990, *Information technology – Programming languages – C* [18], currently under revision (C9x FDIS).

An implementation should follow all the requirements of LIA-2 unless otherwise specified by this language binding.

The operations or parameters marked "†" are not part of the language and should be provided by an implementation that wishes to conform to the LIA-2 for that operation. For each of the marked items a suggested identifier is provided.

The LIA-1 datatype **Boolean** is implemented by the C datatype `int` (1 = **true** and 0 = **false**). (Revised C will provide a `_Bool` datatype.)

Every implementation of C has integral datatypes `int`, `long int`, `unsigned int`, and `unsigned long int`. INT is used below to designate one of the integer datatypes.

C has three floating point datatypes: `float`, `double`, and `long double`. FLT is used below to designate one of the floating point datatypes.

The LIA-2 integer operations are listed below, along with the syntax used to invoke them:

$max_I(x,y)$	imax(x, y)	†
$min_I(x,y)$	imin(x, y)	†
$max\_seq_I(xs)$	imax_arr(xs,nr_of_items)	†*
$min\_seq_I(xs)$	imin_arr(xs,nr_of_items)	†*
$dim_I(x,y)$	idim(x, y)	†
$power_I(x,y)$	ipower(x, y)	†
$shift2_I(x,y)$	shift2(x, y)	†
$shift10_I(x,y)$	shift10(x, y)	†
$sqrt_I(x)$	isqrt(x)	†
$divides_I(x,y)$	does_divide(x, y)	†
$divides_I(x,y)$	x != 0 && y % x == 0	$\star$
$even_I(x)$	x % 2 == 0	$\star$
$odd_I(x)$	x % 2 != 0	$\star$
$divf_I(x,y)$	div(x, y)	†
$moda_I(x,y)$	mod(x, y)	†
$group_I(x,y)$	group(x, y)	†
$pad_I(x,y)$	pad(x, y)	†
$quot_I(x,y)$	quot(x, y)	†
$remr_I(x,y)$	iremainder(x, y)	†
$gcd_I(x,y)$	gcd(x, y)	†
$lcm_I(x,y)$	lcm(x, y)	†
$gcd\_seq_I(xs)$	gcd_arr(xs,nr_of_items)	†*
$lcm\_seq_I(xs)$	lcm_arr(xs,nr_of_items)	†*
$add\_wrap_I(x,y)$	add_wrap(x, y)	†
$add\_ov_I(x,y)$	add_over(x, y)	†
$sub\_wrap_I(x,y)$	sub_wrap(x, y)	†
$sub\_ov_I(x,y)$	sub_over(x, y)	†
$mul\_wrap_I(x,y)$	mul_wrap(x, y)	†
$mul\_ov_I(x,y)$	mul_over(x, y)	†

where x and y are expressions of the same integer type and where xs is an expression of type array of an integer type. (The operations marked with * needs one name per integer datatype.)

The LIA-2 non-transcendental floating point operations are listed below, along with the syntax (type generic macros) used to invoke them:

$min_F(x,y)$	nmin(x,y)	†
$max_F(x,y)$	nmax(x,y)	†
$mmin_F(x,y)$	fmin(x,y)	$\star$(C9x)

$\mathit{mmax}_F(x, y)$	`fmax(`x`,`y`)`	⋆(C9x)
$\mathit{min\_seq}_F(xs)$	`nmin_arr(`xs`,`$\mathit{nr\_of\_items}$`)`	†*
$\mathit{max\_seq}_F(xs)$	`nmax_arr(`xs`,`$\mathit{nr\_of\_items}$`)`	†*
$\mathit{mmin\_seq}_F(xs)$	`fmin_arr(`xs`,`$\mathit{nr\_of\_items}$`)`	†*
$\mathit{mmax\_seq}_F(xs)$	`fmax_arr(`xs`,`$\mathit{nr\_of\_items}$`)`	†*
$\mathit{dim}_F(x, y)$	`fdim(`x`,`y`)`	⋆(C9x)
$\mathit{rounding}_F(x)$	`nearbyint(`x`)` (when in round to nearest mode)	⋆(C9x)
$\mathit{floor}_F(x)$	`floor(`x`)`	⋆
$\mathit{ceiling}_F(x)$	`ceil(`x`)`	⋆
$\mathit{rounding\_rest}_F(x)$	x `- nearbyint(`x`)` (when in round to nearest mode)	⋆(C9x)
$\mathit{floor\_rest}_F(x)$	x `- floor(`x`)`	⋆
$\mathit{ceiling\_rest}_F(x)$	x `- ceil(`x`)`	⋆
$\mathit{remr}_F(x, y)$	`remainder(`x`,`y`)`	⋆(C9x)
$\mathit{sqrt}_F(x)$	`sqrt(`x`)`	⋆
$\mathit{rsqrt}_F(x)$	`rsqrt(`x`)`	†
$\mathit{add\_lo}_F(x, y)$	`add_low(`x`, `y`)`	†
$\mathit{sub\_lo}_F(x, y)$	`sub_low(`x`, `y`)`	†
$\mathit{mul\_lo}_F(x, y)$	`mul_low(`x`, `y`)`	†
$\mathit{div\_rest}_F(x, y)$	`div_rest(`x`, `y`)`	†
$\mathit{sqrt\_rest}_F(x)$	`sqrt_rest(`x`)`	†
$\mathit{mul}_{F \to F'}(x, y)$	`dprod(`x`, `y`)`	†

where x, y and z are expressions of the same floating point type, and where xs is an expression of type array of a floating point type.

The LIA-2 parameters for operations approximating real valued transcendental functions can be accessed by the following syntax (type generic macros):

$\mathit{max\_err\_hypot}_F$	`err_hypot(`x`)`	†
$\mathit{max\_err\_exp}_F$	`err_exp(`x`)`	†
$\mathit{max\_err\_power}_F$	`err_power(`x`)`	†
$\mathit{max\_err\_sinh}_F$	`err_sinh(`x`)`	†
$\mathit{max\_err\_tanh}_F$	`err_tanh(`x`)`	†
$\mathit{big\_angle\_r}_F$	`big_radian_angle(`x`)`	†
$\mathit{max\_err\_sin}_F$	`err_sin(`x`)`	†
$\mathit{max\_err\_tan}_F$	`err_tan(`x`)`	†
$\mathit{min\_angular\_unit}_F$	`smallest_angle_unit(`x`)`	†
$\mathit{big\_angle\_u}_F$	`big_angle(`x`)`	†
$\mathit{max\_err\_sinu}_F(u)$	`err_sin_cycle(`u`)`	†
$\mathit{max\_err\_tanu}_F(u)$	`err_tan_cycle(`u`)`	†
$\mathit{max\_err\_convert}_F$	`err_convert(`x`)`	†
$\mathit{max\_err\_convert}_{F'}$	`err_convert_to_string`	†
$\mathit{max\_err\_convert}_{D'}$	`err_convert_to_string`	†

where x and u are expressions of a floating point type. Several of the parameter functions are constant for each type (and library), the argument is then used only to differentiate among the floating point types.

C has a `pow` operation that does not conform to LIA-2, but may be specified in LIA-2 terms as:

$$\begin{array}{lll} pow_F(x,y) & = 1 & \text{if } x \text{ is a quiet NaN and } y = 0 \\ & = pow_F(x,0) & \text{if } y = \mathbf{-0} \\ & = power_{F\mathcal{Z}}(x,y) & \text{if } y \in F \cap \mathcal{Z} \\ & = power_F(x,y) & \text{otherwise} \end{array}$$

C has a `hypot` operation that does not conform to LIA-2, but may be specified in LIA-2 terms as:

$$\begin{array}{lll} hhypot_F(x,y) & = \mathbf{+\infty} & \text{if } x \text{ is a quiet NaN and } y \in \{\mathbf{-\infty}, \mathbf{+\infty}\} \\ & = \mathbf{+\infty} & \text{if } x \in \{\mathbf{-\infty}, \mathbf{+\infty}\} \text{ and } y \text{ is a quiet NaN} \\ & = hypot_F(x,y) & \text{otherwise} \end{array}$$

The LIA-2 elementary floating point operations are listed below, together with the non-LIA-2 pow_F and $hhypot_F$, along with the syntax (type generic macros) used to invoke them:

$hypot_F(x,y)$	`hypotenuse(x, y)`	†
$hhypot_F(x,y)$	`hypot(x, y)`	⋆ Not LIA-2!
$power_{FI}(b,z)$	`poweri(b, z)`	†
$exp_F(x)$	`exp(x)`	⋆
$expm1_F(x)$	`expm1(x)`	⋆(C9x)
$exp2_F(x)$	`exp2(x)`	⋆
$exp10_F(x)$	`exp10(x)`	†
$power_F(b,y)$	`power(b, y)`	†
$pow_F(b,y)$	`pow(b, y)`	⋆ Not LIA-2!
$power1pm1_F(b,y)$	`power1pm1(b, y)`	†
$ln_F(x)$	`log(x)`	⋆
$ln1p_F(x)$	`log1p(x)`	⋆(C9x)
$log2_F(x)$	`log2(x)`	⋆
$log10_F(x)$	`log10(x)`	⋆
$logbase_F(b,x)$	`logbase(b, x)`	†
$logbase1p1p_F(b,x)$	`logbase1p1p(b, x)`	†
$sinh_F(x)$	`sinh(x)`	⋆
$cosh_F(x)$	`cosh(x)`	⋆
$tanh_F(x)$	`tanh(x)`	⋆
$coth_F(x)$	`coth(x)`	†
$sech_F(x)$	`sech(x)`	†
$csch_F(x)$	`csch(x)`	†
$arcsinh_F(x)$	`asinh(x)`	⋆
$arccosh_F(x)$	`acosh(x)`	⋆
$arctanh_F(x)$	`atanh(x)`	⋆
$arccoth_F(x)$	`acoth(x)`	†
$arcsech_F(x)$	`asech(x)`	†
$arccsch_F(x)$	`acsch(x)`	†
$rad_F(x)$	`radian(x)`	†
$axis\_rad_F(x)$	`axis_rad(x, &h, &v)` (note out parameters)	†
$sin_F(x)$	`sin(x)`	⋆

$cos_F(x)$	`cos(x)`	⋆
$tan_F(x)$	`tan(x)`	⋆
$cot_F(x)$	`cot(x)`	†
$sec_F(x)$	`sec(x)`	†
$csc_F(x)$	`csc(x)`	†
$cossin_F(x)$	`cossin(x, &c, &s)`	†
$arcsin_F(x)$	`asin(x)`	⋆
$arccos_F(x)$	`acos(x)`	⋆
$arctan_F(x)$	`atan(x)`	⋆
$arccot_F(x)$	`acot(x)`	†
$arcctg_F(x)$	`actg(x)`	†
$arcsec_F(x)$	`asec(x)`	†
$arccsc_F(x)$	`acsc(x)`	†
$arc_F(x, y)$	`atan2(y, x)`	⋆
$cycle_F(u, x)$	`cycle(u, x)`	†
$axis\_cycle_F(u, x)$	`axis_cycle(u, x, &h, &v)`	†
$sinu_F(u, x)$	`sinu(u, x)`	†
$cosu_F(u, x)$	`cosu(u, x)`	†
$tanu_F(u, x)$	`tanu(u, x)`	†
$cotu_F(u, x)$	`cotu(u, x)`	†
$secu_F(u, x)$	`secu(u, x)`	†
$cscu_F(u, x)$	`cscu(u, x)`	†
$cossinu_F(u, x)$	`cossinu(u, x, &c, &s)`	†
$arcsinu_F(u, x)$	`asinu(u, x)`	†
$arccosu_F(u, x)$	`acosu(u, x)`	†
$arctanu_F(u, x)$	`atanu(u, x)`	†
$arccotu_F(u, x)$	`acotu(u, x)`	†
$arcctgu_F(u, x)$	`actgu(u, x)`	†
$arcsecu_F(u, x)$	`asecu(u, x)`	†
$arccscu_F(u, x)$	`acscu(u, x)`	†
$arcu_F(u, x, y)$	`atan2u(u, y, x)`	†
$rad\_to\_cycle_F(x, u)$	`radian_to_cycle(x, u)`	†
$cycle\_to\_rad_F(u, x)$	`cycle_to_radian(u, x)`	†
$cycle\_to\_cycle_F(u, x, v)$	`cycle_to_cycle(u, x, v)`	†

where b, x, y, u, and v are expressions of the same floating point type.

Arithmetic value conversions in C can be explicit or implicit. The explicit arithmetic value conversions are usually expressed as 'casts', except when converting to/from strings. The rules for when implicit conversions are applied is not repeated here, but work as if a cast had been applied.

$convert_{I\to I'}(x)$	(*INT2*)*x*	⋆
$convert_{I''\to I}(s)$	`sscanf(s,"%no",&r)`	⋆
$convert_{I''\to I}(s)$	`sscanf(s,"%nd",&r)`	⋆
$convert_{I''\to I}(s)$	`sscanf(s,"%nx",&r)`	⋆
$convert_{I''\to I}(f)$	`fscanf(f,"%no",&r)`	⋆
$convert_{I''\to I}(f)$	`fscanf(f,"%nd",&r)`	⋆

$convert_{I''\rightarrow I}(f)$	`fscanf(`f`,"%`n`x",&`r`)`	⋆
$convert_{I\rightarrow I''}(x)$	`sprintf(`s`,"%`n`o",`x`)`	⋆
$convert_{I\rightarrow I''}(x)$	`sprintf(`s`,"%`n`d",`x`)`	⋆
$convert_{I\rightarrow I''}(x)$	`sprintf(`s`,"%`n`x",`x`)`	⋆
$convert_{I\rightarrow I''}(x)$	`fprintf(`h`,"%`n`o",`x`)`	⋆
$convert_{I\rightarrow I''}(x)$	`fprintf(`h`,"%`n`d",`x`)`	⋆
$convert_{I\rightarrow I''}(x)$	`fprintf(`h`,"%`n`x",`x`)`	⋆
$rounding_{F\rightarrow I}(y)$	(*INT*)`nearbyint(`y`)` (when in round to nearest mode)	⋆
$floor_{F\rightarrow I}(y)$	(*INT*)`floor(`y`)`	⋆
$ceiling_{F\rightarrow I}(y)$	(*INT*)`ceil(`y`)`	⋆
$convert_{I\rightarrow F}(x)$	(*FLT*)x	⋆
$convert_{F\rightarrow F'}(y)$	(*FLT2*)y	⋆
$convert_{F''\rightarrow F}(s)$	`sscanf(`s`,"%e",&`r`)`	⋆
$convert_{F''\rightarrow F}(f)$	`fscanf(`f`,"%e",&`r`)`	⋆
$convert_{F\rightarrow F''}(y)$	`sprintf(`s`,"%.`d`e",`x`)`	⋆
$convert_{F\rightarrow F''}(y)$	`fprintf(`h`,"%.`d`e",`x`)`	⋆
$convert_{D'\rightarrow F}(s)$	`sscanf(`s`,"%f",&`r`)`	⋆
$convert_{D'\rightarrow F}(f)$	`fscanf(`f`,"%f",&`r`)`	⋆
$convert_{F\rightarrow D'}(y)$	`sprintf(`s`,"%.`d`f",`x`)`	⋆
$convert_{F\rightarrow D'}(y)$	`fprintf(`h`,"%.`d`f",`x`)`	⋆

where x is an expression of type *INT*, y is an expression of type *FLT*, and z is an expression of type *FXD*, where *FXD* is a fixed point type. *INT2* is the integer datatype that corresponds to I'. A ? above indicates that the parameter is optional. e is greater than 0.

C9x provides non-negative numerals for all its integer and floating point types. The default base is 10, but base 8 (for integers) and 16 (both integer and float) can be used too. Numerals for different integer types are distinguished by suffixes. Numerals for different floating point types are distinguished by suffix: `f` for `float`, no suffix for `double`, `l` for `long double`. Numerals for floating point types must have a '.' in them. The details are not repeated in this example binding, see ISO/IEC FDIS 9899, clause 6.4.4.1 Integer constants, and clause 6.4.4.2 Floating constants.

C9x specifies numerals (as macros) for infinities and NaNs for `float`:

+∞	`INFINITY`	⋆
qNaN	`NAN`	⋆
sNaN	`SIGNAN`	†

as well as string formats for reading and writing these values as character strings.

C9x has two ways of handling arithmetic errors. One, for backwards compatibility, is by assigning to `errno`. The other is by recording of indicators, the method preferred by LIA-2, which can be used for floating point errors. For C9x, the **absolute_precision_underflow** notification is ignored. The behaviour for notification upon integer operations initiating a notification is, however, not defined by C9x.

C.5 C++

The programming language C++ is defined by ISO/IEC 14882:1998, *Programming languages – C++* [19].

An implementation should follow all the requirements of LIA-2 unless otherwise specified by this language binding.

The operations or parameters marked "†" are not part of the language and should be provided by an implementation that wishes to conform to the LIA-2 for that operation. For each of the marked items a suggested identifier is provided.

This example binding recommends that all identifiers suggested here be defined in the namespace `std::math`.

The LIA-1 datatype **Boolean** is implemented by the C++ datatype `bool`.

Every implementation of C++ has integral datatypes `int`, `long int`, `unsigned int`, and `unsigned long int`. INT is used below to designate one of the integer datatypes.

C++ has three floating point datatypes: `float`, `double`, and `long double`. FLT is used below to designate one of the floating point datatypes.

The LIA-2 integer operations are listed below, along with the syntax used to invoke them:

$max_I(x,y)$	max(x, y)	⋆
$min_I(x,y)$	min(x, y)	⋆
$max\_seq_I(xs)$	xs.max()	⋆
$min\_seq_I(xs)$	xs.min()	⋆
$dim_I(x,y)$	dim(x, y)	†
$power_I(x,y)$	power(x, y)	†
$sqrt_I(x)$	sqrt(x)	†
$shift2_I(x,y)$	shift2(x, y)	†
$shift10_I(x,y)$	shift10(x, y)	†
$divides_I(x,y)$	does_divide(x, y)	†
$divides_I(x,y)$	y != 0 && y % x == 0	⋆
$even_I(x)$	x % 2 == 0	⋆
$odd_I(x)$	x % 2 != 0	⋆
$divf_I(x,y)$	div(x, y)	†
$moda_I(x,y)$	mod(x, y)	†
$group_I(x,y)$	group(x, y)	†
$pad_I(x,y)$	pad(x, y)	†
$quot_I(x,y)$	quot(x, y)	†
$remr_I(x,y)$	iremainder(x, y)	†
$gcd_I(x,y)$	gcd(x, y)	†
$lcm_I(x,y)$	lcm(x, y)	†
$gcd\_seq_I(xs)$	xs.gcd()	†
$lcm\_seq_I(xs)$	xs.lcm()	†
$add\_wrap_I(x,y)$	add_wrap(x, y)	†
$add\_ov_I(x,y)$	add_over(x, y)	†
$sub\_wrap_I(x,y)$	sub_wrap(x, y)	†
$sub\_ov_I(x,y)$	sub_over(x, y)	†
$mul\_wrap_I(x,y)$	mul_wrap(x, y)	†
$mul\_ov_I(x,y)$	mul_over(x, y)	†

where x and y are expressions of the same integer type and where xs is an expression of type valarray of an integer type.

The LIA-2 non-transcendental floating point operations are listed below, along with the syntax used to invoke them:

$\mathit{max}_F(x, y)$	`nmax(`x`, `y`)`	†
$\mathit{min}_F(x, y)$	`nmin(`x`, `y`)`	†
$\mathit{mmax}_F(x, y)$	`max(`x`, `y`)`	⋆(unclear)
$\mathit{mmin}_F(x, y)$	`min(`x`, `y`)`	⋆(unclear)
$\mathit{max\_seq}_F(xs)$	xs`.nmax()`	†
$\mathit{min\_seq}_F(xs)$	xs`.nmin()`	†
$\mathit{mmax\_seq}_F(xs)$	xs`.max()`	⋆(unclear)
$\mathit{mmin\_seq}_F(xs)$	xs`.min()`	⋆(unclear)

$\mathit{dim}_F(x, y)$	`dim(`x`, `y`)`	†
$\mathit{rounding}_F(x)$	`round(`x`)`	†
$\mathit{floor}_F(x)$	`floor(`x`)`	⋆
$\mathit{ceiling}_F(x)$	`ceil(`x`)`	⋆
$\mathit{rounding\_rest}_F(x)$	x` - round(`x`)`	†
$\mathit{floor\_rest}_F(x)$	x` - floor(`x`)`	⋆
$\mathit{ceiling\_rest}_F(x)$	x` - ceil(`x`)`	⋆

$\mathit{mul}_{F \to F'}(x, y)$	`dprod(`x`, `y`)`	†
$\mathit{remr}_F(x, y)$	`remainder(`x`, `y`)`	†
$\mathit{sqrt}_F(x)$	`sqrt(`x`)`	⋆
$\mathit{rsqrt}_F(x)$	`reciprocal_sqrt(`x`)`	†

$\mathit{add\_lo}_F(x, y)$	`add_low(`x`, `y`)`	†
$\mathit{sub\_lo}_F(x, y)$	`sub_low(`x`, `y`)`	†
$\mathit{mul\_lo}_F(x, y)$	`mul_low(`x`, `y`)`	†
$\mathit{div\_rest}_F(x, y)$	`div_rest(`x`, `y`)`	†
$\mathit{sqrt\_rest}_F(x)$	`sqrt_rest(`x`)`	†

where x, y and z are expressions of the same floating point type, and where xs is an expression of type valarray of a floating point type.

The parameters for operations approximating real valued transcendental functions can be accessed by the following syntax:

$\mathit{max\_err\_hypot}_F$	`err_hypotenuse<FLT>()`	†

$\mathit{max\_err\_exp}_F$	`err_exp<FLT>()`	†
$\mathit{max\_err\_power}_F$	`err_power<FLT>()`	†

$\mathit{max\_err\_sinh}_F$	`err_sinh<FLT>()`	†
$\mathit{max\_err\_tanh}_F$	`err_tanh<FLT>()`	†

$\mathit{big\_angle\_r}_F$	`big_radian_angle<FLT>()`	†
$\mathit{max\_err\_sin}_F$	`err_sin<FLT>()`	†
$\mathit{max\_err\_tan}_F$	`err_tan<FLT>()`	†

$\mathit{min\_angular\_unit}_F$	`smallest_angle_unit<FLT>()`	†
$\mathit{big\_angle\_u}_F$	`big_angle<FLT>()`	†
$\mathit{max\_err\_sinu}_F(u)$	`err_sin_cycle(`u`)`	†
$\mathit{max\_err\_tanu}_F(u)$	`err_tan_cycle(`u`)`	†

$\mathit{max\_err\_convert}_F$	`err_convert<FLT>()`	†
$\mathit{max\_err\_convert}_{F'}$	`err_convert_to_string()`	†

$\mathit{max\_err\_convert}_{D'}$	`err_convert_to_string()`	†

where u is an expression of a floating point type. Several of the parameter functions are constant for each type (and library).

The LIA-2 elementary floating point operations are listed below, along with the syntax (type generic macros) used to invoke them:

$\mathit{hypot}_F(x, y)$	`hypotenuse(x, y)`	†
$\mathit{power}_{FI}(b, z)$	`poweri(b, z)`	†
$\mathit{exp}_F(x)$	`exp(x)`	⋆
$\mathit{expm1}_F(x)$	`expm1(x)`	†
$\mathit{exp2}_F(x)$	`exp2(x)`	†
$\mathit{exp10}_F(x)$	`exp10(x)`	†
$\mathit{power}_F(b, y)$	`power(b, y)`	†
$\mathit{pow}_F(b, y)$	`pow(b, y)`	⋆ Not LIA-2! (See C.)
$\mathit{power1pm1}_F(b, y)$	`power1pm1(b, y)`	†
$\mathit{ln}_F(x)$	`log(x)`	⋆
$\mathit{ln1p}_F(x)$	`log1p(x)`	†
$\mathit{log2}_F(x)$	`log2(x)`	†
$\mathit{log10}_F(x)$	`log10(x)`	⋆
$\mathit{logbase}_F(b, x)$	`logbase(b, x)`	†
$\mathit{logbase1p1p}_F(b, x)$	`logbase1p1p(b, x)`	†
$\mathit{sinh}_F(x)$	`sinh(x)`	⋆
$\mathit{cosh}_F(x)$	`cosh(x)`	⋆
$\mathit{tanh}_F(x)$	`tanh(x)`	⋆
$\mathit{coth}_F(x)$	`coth(x)`	†
$\mathit{sech}_F(x)$	`sech(x)`	†
$\mathit{csch}_F(x)$	`csch(x)`	†
$\mathit{arcsinh}_F(x)$	`asinh(x)`	⋆
$\mathit{arccosh}_F(x)$	`acosh(x)`	⋆
$\mathit{arctanh}_F(x)$	`atanh(x)`	⋆
$\mathit{arccoth}_F(x)$	`acoth(x)`	†
$\mathit{arcsech}_F(x)$	`asech(x)`	†
$\mathit{arccsch}_F(x)$	`acsch(x)`	†
$\mathit{rad}_F(x)$	`rad(x)`	†
$\mathit{axis\_rad}_F(x)$	`axis_rad(x, &h, &v)` (note out parameters)	†
$\mathit{sin}_F(x)$	`sin(x)`	⋆
$\mathit{cos}_F(x)$	`cos(x)`	⋆
$\mathit{tan}_F(x)$	`tan(x)`	⋆
$\mathit{cot}_F(x)$	`cot(x)`	†
$\mathit{sec}_F(x)$	`sec(x)`	†
$\mathit{csc}_F(x)$	`csc(x)`	†
$\mathit{cossin}_F(x)$	`cossin(x, &c, &s)`	†
$\mathit{arcsin}_F(x)$	`asin(x)`	⋆
$\mathit{arccos}_F(x)$	`acos(x)`	⋆

$arctan_F(x)$	`atan(x)`	⋆
$arccot_F(x)$	`acot(x)`	†
$arcctg_F(x)$	`actg(x)`	†
$arcsec_F(x)$	`asec(x)`	†
$arccsc_F(x)$	`acsc(x)`	†
$arc_F(x, y)$	`atan2(y, x)`	⋆
$cycle_F(u, x)$	`cycle(u, x)`	†
$axis\_cycle_F(u, x)$	`axis_cycle(u, x, &h, &v)`	†
$sinu_F(u, x)$	`sinu(u, x)`	†
$cosu_F(u, x)$	`cosu(u, x)`	†
$tanu_F(u, x)$	`tanu(u, x)`	†
$cotu_F(u, x)$	`cotu(u, x)`	†
$secu_F(u, x)$	`secu(u, x)`	†
$cscu_F(u, x)$	`cscu(u, x)`	†
$cossinu_F(x)$	`cossinu(u, x, &c, &s)`	†
$arcsinu_F(u, x)$	`asinu(u, x)`	†
$arccosu_F(u, x)$	`acosu(u, x)`	†
$arctanu_F(u, x)$	`atanu(u, x)`	†
$arccotu_F(u, x)$	`acotu(u, x)`	†
$arcctgu_F(u, x)$	`actgu(u, x)`	†
$arcsecu_F(u, x)$	`asecu(u, x)`	†
$arccscu_F(u, x)$	`acscu(u, x)`	†
$arcu_F(u, x, y)$	`atan2u(u, y, x)`	†
$rad\_to\_cycle_F(x, u)$	`radian_to_cycle(x, u)`	†
$cycle\_to\_rad_F(u, x)$	`cycle_to_radian(u, x)`	†
$cycle\_to\_cycle_F(u, x, v)$	`cycle_to_cycle(u, x, v)`	†

where b, x, y, u, and v are expressions of type *FLT*, and z is an expressions of type *INT*

Arithmetic value conversions in C++ are can be explicit or implicit. The rules for when implicit conversions are applied is not repeated here. The explicit arithmetic value conversions are usually expressed as 'casts', except when converting to/from strings.

$convert_{I \to I'}(x)$	(*INT2*)x	⋆
$convert_{I'' \to I}(s)$	`sscanf(s,"%no",&r)`	⋆
$convert_{I'' \to I}(s)$	`sscanf(s,"%nd",&r)`	⋆
$convert_{I'' \to I}(s)$	`sscanf(s,"%nx",&r)`	⋆
$convert_{I'' \to I}(f)$	`fscanf(f,"%no",&r)`	⋆
$convert_{I'' \to I}(f)$	`fscanf(f,"%nd",&r)`	⋆
$convert_{I'' \to I}(f)$	`fscanf(f,"%nx",&r)`	⋆
$convert_{I \to I''}(x)$	`sprintf(s,"%no",x)`	⋆
$convert_{I \to I''}(x)$	`sprintf(s,"%nd",x)`	⋆
$convert_{I \to I''}(x)$	`sprintf(s,"%nx",x)`	⋆
$convert_{I \to I''}(x)$	`fprintf(h,"%no",x)`	⋆
$convert_{I \to I''}(x)$	`fprintf(h,"%nd",x)`	⋆
$convert_{I \to I''}(x)$	`fprintf(h,"%nx",x)`	⋆
$rounding_{F \to I}(y)$	(*INT*)`nearbyint(y)` (when in round to nearest mode)	⋆
$floor_{F \to I}(y)$	(*INT*)`floor(y)`	⋆

$ceiling_{F\to I}(y)$	`(INT)ceil(y)`	⋆
$convert_{I\to F}(x)$	`(FLT)x`	⋆
$convert_{F\to F'}(y)$	`(FLT2)y`	⋆
$convert_{F''\to F}(s)$	`sscanf(s,"%e",&r)`	⋆
$convert_{F''\to F}(f)$	`fscanf(f,"%e",&r)`	⋆
$convert_{F\to F''}(y)$	`sprintf(s,"%.de",x)`	⋆
$convert_{F\to F''}(y)$	`fprintf(h,"%.de",x)`	⋆
$convert_{D'\to F}(s)$	`sscanf(s,"%f",&r)`	⋆
$convert_{D'\to F}(f)$	`fscanf(f,"%f",&r)`	⋆
$convert_{F\to D'}(y)$	`sprintf(s,"%.df",x)`	⋆
$convert_{F\to D'}(y)$	`fprintf(h,"%.df",x)`	⋆

where x is an expression of type *INT*, y is an expression of type *FLT*, and z is an expression of type *FXD*, where *FXD* is a fixed point type. *INT2* is the integer datatype that corresponds to I'. A ? above indicates that the parameter is optional. e is greater than 0.

C++ provides non-negative numerals for all its integer and floating point types in base 10. Numerals for different integer types are distinguished by suffixes. Numerals for different floating point types are distinguished by suffix: `f` for `float`, no suffix for `double`, `l` for `long double`. Numerals for floating point types must have a '.' in them. The details are not repeated in this example binding, see ISO/IEC 14882, clause 2.9.1 Integer literals, and clause 2.9.4 Floating literals.

C++ does not specify numerals for infinities and NaNs. Suggestion:

+∞	`INFINITY`	†
qNaN	`NAN`	†
sNaN	`SIGNAN`	†

as well as string formats for reading and writing these values as character strings.

C++ has completely undefined behaviour on arithmetic notification. An implementation wishing to conform to LIA-2 should provide a means for recording of indicators, similar to C9x.

C.6 Fortran

The programming language Fortran is defined by ISO/IEC 1539-1:1997, *Information technology – Programming languages – Fortran – Part 1: Base language* [23].

An implementation should follow all the requirements of LIA-2 unless otherwise specified by this language binding.

The operations or parameters marked "†" are not part of the language and should be provided by an implementation that wishes to conform to the LIA-2 for that operation. For each of the marked items a suggested identifier is provided.

The Fortran datatype `LOGICAL` corresponds to the LIA datatype **Boolean**.

Every implementation of Fortran has one integer datatype, denoted as `INTEGER`, and two floating point data type denoted as `REAL` (single precision) and `DOUBLE PRECISION`.

An implementation is permitted to offer additional `INTEGER` types with a different range and additional `REAL` types with different precision or range, parameterised with the `KIND` parameter.

The LIA-2 integer operations are listed below, along with the syntax used to invoke them:

$\mathit{max}_I(x,y)$	MAX(x, y)	⋆
$\mathit{min}_I(x,y)$	MIN(x, y)	⋆
$\mathit{max\_seq}_I(xs)$	MAX($xs[1]$, $xs[2]$, ..., $xs[n]$) or MAXVAL(xs)	⋆
$\mathit{min\_seq}_I(xs)$	MIN($xs[1]$, $xs[2]$, ..., $xs[n]$) or MINVAL(xs)	⋆
$\mathit{dim}_I(x,y)$	DIM(x, y)	⋆
$\mathit{power}_I(x,y)$	x ** y	⋆
$\mathit{shift2}_I(x,y)$	SHIFT2(x, y)	†
$\mathit{shift10}_I(x,y)$	SHIFT10(x, y)	†
$\mathit{sqrt}_I(x)$	ISQRT(x)	†
$\mathit{divides}_I(x,y)$	DIVIDES(x, y)	†
$\mathit{even}_I(x)$	MODULO(x,2) == 0	⋆
$\mathit{odd}_I(x)$	MODULO(x,2) != 0	⋆
$\mathit{divf}_I(x,y)$	DIV(x, y)	†
$\mathit{moda}_I(x,y)$	MODULO(x, y)	⋆
$\mathit{group}_I(x,y)$	GROUP(x, y)	†
$\mathit{pad}_I(x,y)$	PAD(x, y)	†
$\mathit{quot}_I(x,y)$	QUOTIENT(x, y)	†
$\mathit{remr}_I(x,y)$	REMAINDER(x, y)	†
$\mathit{gcd}_I(x,y)$	GCD(x, y)	†
$\mathit{lcm}_I(x,y)$	LCM(x, y)	†
$\mathit{gcd\_seq}_I(xs)$	GCDVAL(xs)	†
$\mathit{lcm\_seq}_I(xs)$	LCMVAL(xs)	†
$\mathit{add\_wrap}_I(x,y)$	ADD_WRAP(x, y)	†
$\mathit{add\_ov}_I(x,y)$	ADD_OVER(x, y)	†
$\mathit{sub\_wrap}_I(x,y)$	SUB_WRAP(x, y)	†
$\mathit{sub\_ov}_I(x,y)$	SUB_OVER(x, y)	†
$\mathit{mul\_wrap}_I(x,y)$	MUL_WRAP(x, y)	†
$\mathit{mul\_ov}_I(x,y)$	MUL_OVER(x, y)	†

where x and y are expressions of type INTEGER and where xs is an expression of type array of INTEGER.

The additional non-transcendental floating point operations are listed below, along with the syntax used to invoke them:

$\mathit{max}_F(x,y)$	MAX(x, y)	⋆
$\mathit{min}_F(x,y)$	MIN(x, y)	⋆
$\mathit{mmax}_F(x,y)$	MMAX(x, y)	†
$\mathit{mmin}_F(x,y)$	MMIN(x, y)	†
$\mathit{max\_seq}_F(xs)$	MAX($xs[1]$, $xs[2]$, ..., $xs[n]$) or MAXVAL(xs)	⋆
$\mathit{min\_seq}_F(xs)$	MIN($xs[1]$, $xs[2]$, ..., $xs[n]$) or MINVAL(xs)	⋆
$\mathit{mmax\_seq}_F(xs)$	MMAX($xs[1]$, $xs[2]$, ..., $xs[n]$) or MMAXVAL(xs)	†
$\mathit{mmin\_seq}_F(xs)$	MMIN($xs[1]$, $xs[2]$, ..., $xs[n]$) or MMINVAL(xs)	†
$\mathit{dim}_F(x,y)$	DIM(x, y)	⋆
$\mathit{rounding}_F(x)$	IEEE_RINT(x) (if in round to nearest mode)	(⋆)
$\mathit{floor}_F(x)$	IEEE_RINT(x) (if in round towards $-\infty$ mode)	(⋆)
$\mathit{ceiling}_F(x)$	IEEE_RINT(x) (if in round towards $+\infty$ mode)	(⋆)
$\mathit{rounding\_rest}_F(x)$	x - IEEE_RINT(x) (if in round to nearest mode)	(⋆)

$floor\_rest_F(x)$	x `-` `IEEE_RINT(`x`)` (if in round towards $-\infty$ mode)	(⋆)
$ceiling\_rest_F(x)$	x `-` `IEEE_RINT(`x`)` (if in round towards $+\infty$ mode)	(⋆)
$remr_F(x,y)$	`IEEE_REM(`x`, `y`)`	(⋆)
$sqrt_F(x)$	`SQRT(`x`)`	⋆
$rsqrt_F(x)$	`RSQRT(`x`)`	†
$mul_{F \to F'}(x,y)$	`DPROD(`x`, `y`)`	⋆
$add\_lo_F(x,y)$	`ADD_LOW(`x`, `y`)`	†
$sub\_lo_F(x,y)$	`SUB_LOW(`x`, `y`)`	†
$mul\_lo_F(x,y)$	`MUL_LOW(`x`, `y`)`	†
$div\_rest_F(x,y)$	`DIV_REST(`x`, `y`)`	†
$sqrt\_rest_F(x)$	`SQRT_REST(`x`)`	†

where x, y and z are expressions of type *FLT*, and where xs is an expression of type array of *FLT*.

The LIA-2 parameters for operations approximating real valued transcendental functions can be accessed by the following syntax:

$max\_err\_hypot_F$	`ERR_HYPOTENUSE(`x`)`	†
$max\_err\_exp_F$	`ERR_EXP(`x`)`	†
$max\_err\_power_F$	`ERR_POWER(`x`)`	†
$max\_err\_sinh_F$	`ERR_SINH(`x`)`	†
$max\_err\_tanh_F$	`ERR_TANH(`x`)`	†
$big\_angle\_r_F$	`BIG_RADIAN_ANGLE(`x`)`	†
$max\_err\_sin_F$	`ERR_SIN(`x`)`	†
$max\_err\_tan_F$	`ERR_TAN(`x`)`	†
$min\_angular\_unit_F$	`MIN_ANGLE_UNIT(`x`)`	†
$big\_angle\_u_F$	`BIG_ANGLE(`x`)`	†
$max\_err\_sinu_F(u)$	`ERR_SIN_CYCLE(`u`)`	†
$max\_err\_tanu_F(u)$	`ERR_TAN_CYCLE(`u`)`	†
$max\_err\_convert_F$	`ERR_CONVERT(`x`)`	†
$max\_err\_convert_{F'}$	`ERR_CONVERT_TO_STRING`	†
$max\_err\_convert_{D'}$	`ERR_CONVERT_TO_STRING`	†

where b, x and u are expressions of type *FLT*. Several of the parameter functions are constant for each type (and library), the argument is then used only to differentiate among the floating point types.

The LIA-2 elementary floating point operations are listed below, along with the syntax used to invoke them:

$hypot_F(x,y)$	`HYPOT(`x`, `y`)`	†
$power_{FI}(b,z)$	b `**` z	⋆
$exp_F(x)$	`EXP(`x`)`	⋆
$expm1_F(x)$	`EXPM1(`x`)`	†
$exp2_F(x)$	`EXP2(`x`)`	†
$exp10_F(x)$	`EXP10(`x`)`	†
$power_F(b,y)$	b `**` y	⋆
$power1pm1_F(b,y)$	`POWER1PM1(`b`, `y`)`	†

$\mathit{ln}_F(x)$	LOG(x)	⋆
$\mathit{ln1p}_F(x)$	LOG1P(x)	†
$\mathit{log2}_F(x)$	LOG2(x)	†
$\mathit{log10}_F(x)$	LOG10(x)	⋆
$\mathit{logbase}_F(b, x)$	LOGBASE(b, x)	†
$\mathit{logbase1p1p}_F(b, x)$	LOGBASE1P1P(b, x)	†
$\mathit{sinh}_F(x)$	SINH(x)	⋆
$\mathit{cosh}_F(x)$	COSH(x)	⋆
$\mathit{tanh}_F(x)$	TANH(x)	⋆
$\mathit{coth}_F(x)$	COTH(x)	†
$\mathit{sech}_F(x)$	SECH(x)	†
$\mathit{csch}_F(x)$	CSCH(x)	†
$\mathit{arcsinh}_F(x)$	ASINH(x)	†
$\mathit{arccosh}_F(x)$	ACOSH(x)	†
$\mathit{arctanh}_F(x)$	ATANH(x)	†
$\mathit{arccoth}_F(x)$	ACOTH(x)	†
$\mathit{arcsech}_F(x)$	ASECH(x)	†
$\mathit{arccsch}_F(x)$	ACSCH(x)	†
$\mathit{rad}_F(x)$	RAD(x)	†
$\mathit{sin}_F(x)$	SIN(x)	⋆
$\mathit{cos}_F(x)$	COS(x)	⋆
$\mathit{tan}_F(x)$	TAN(x)	⋆
$\mathit{cot}_F(x)$	COT(x)	†
$\mathit{sec}_F(x)$	SEC(x)	†
$\mathit{csc}_F(x)$	CSC(x)	†
$\mathit{arcsin}_F(x)$	ASIN(x)	⋆
$\mathit{arccos}_F(x)$	ACOS(x)	⋆
$\mathit{arctan}_F(x)$	ATAN(x)	⋆
$\mathit{arccot}_F(x)$	ACOT(x)	†
$\mathit{arcctg}_F(x)$	ACTG(x)	†
$\mathit{arcsec}_F(x)$	ASEC(x)	†
$\mathit{arccsc}_F(x)$	ACSC(x)	†
$\mathit{arc}_F(x, y)$	ATAN2(y, x)	⋆
$\mathit{cycle}_F(u, x)$	CYCLE(u,x)	†
$\mathit{sinu}_F(u, x)$	SINU(u,x)	†
$\mathit{cosu}_F(u, x)$	COSU(u,x)	†
$\mathit{tanu}_F(u, x)$	TANU(u,x)	†
$\mathit{cotu}_F(u, x)$	COTU(u,x)	†
$\mathit{secu}_F(u, x)$	SECU(u,x)	†
$\mathit{cscu}_F(u, x)$	CSCU(u,x)	†
$\mathit{arcsinu}_F(u, x)$	ASINU(u,x)	†
$\mathit{arccosu}_F(u, x)$	ACOSU(u,x)	†

$arctanu_F(u,x)$	ATANU(u,x)	†
$arccotu_F(u,x)$	ACOTU(u,x)	†
$arcctgu_F(u,x)$	ACTGU(u,x)	†
$arcsecu_F(u,x)$	ASECU(u,x)	†
$arccscu_F(u,x)$	ACSCU(u,x)	†
$arcu_F(u,x,y)$	ATAN2U(u,y,x)	†
$cycle_F(360,x)$	DEGREES(x)	†
$sinu_F(360,x)$	SIND(x)	†
$cosu_F(360,x)$	COSD(x)	†
$tanu_F(360,x)$	TAND(x)	†
$cotu_F(360,x)$	COTD(x)	†
$secu_F(360,x)$	SECD(x)	†
$cscu_F(360,x)$	CSCD(x)	†
$arcsinu_F(360,x)$	ASIND(x)	†
$arccosu_F(360,x)$	ACOSD(x)	†
$arctanu_F(360,x)$	ATAND(x)	†
$arccotu_F(360,x)$	ACOTD(x)	†
$arcctgu_F(360,x)$	ACTGD(x)	†
$arcsecu_F(360,x)$	ASECD(x)	†
$arccscu_F(360,x)$	ACSCD(x)	†
$arcu_F(360,x,y)$	ATAN2D(y,x)	†
$rad\_to\_cycle_F(x,u)$	RAD_TO_CYCLE(x, u)	†
$cycle\_to\_rad_F(u,x)$	CYCLE_TO_RAD(u, x)	†
$cycle\_to\_cycle_F(u,x,v)$	CYCLE_TO_CYCLE(u, x, v)	†

where b, x, y, u, and v are expressions of type *FLT*, and z is an expressions of type *INT*

Arithmetic value conversions in Fortran are always explicit, and the conversion function is named like the target type, except when converting to/from strings.

$convert_{I \to I'}(x)$		INT(x,$kind$)	⋆
	lbla	FORMAT (Bn)	⋆(binary)
$convert_{I'' \to I}(f)$		READ (UNIT=f,FMT=*lbla*) r	⋆
$convert_{I \to I''}(x)$		WRITE (UNIT=h, FMT=*lbla*) x	⋆
	lblb	FORMAT (On)	⋆(octal)
$convert_{I'' \to I}(f)$		READ (UNIT=f,FMT=*lblb*) r	⋆
$convert_{I \to I''}(x)$		WRITE (UNIT=h, FMT=*lblb*) x	⋆
	lblc	FORMAT (In)	⋆(decimal)
$convert_{I'' \to I}(f)$		READ (UNIT=f,FMT=*lblc*) r	⋆
$convert_{I \to I''}(x)$		WRITE (UNIT=h, FMT=*lblc*) x	⋆
	lbld	FORMAT (Zn)	⋆(hexadecimal)
$convert_{I'' \to I}(f)$		READ (UNIT=f,FMT=*lbld*) r	⋆
$convert_{I \to I''}(x)$		WRITE (UNIT=h, FMT=*lbld*) x	⋆
$rounding_{F \to I}(y)$		ROUND(y,$kind$?)	†
$floor_{F \to I}(y)$		FLOOR(y,$kind$?)	⋆
$ceiling_{F \to I}(y)$		CEILING(y,$kind$?)	⋆

$convert_{I \rightarrow F}(x)$		REAL(x,*kind*) or sometimes DBLE(x)	⋆
$convert_{F \rightarrow F'}(y)$		REAL(y,*kind*) or sometimes DBLE(y)	⋆
	lble	FORMAT (Fw.d)	⋆
	lblf	FORMAT (Dw.d)	⋆
	lblg	FORMAT (Ew.d)	⋆
	lblh	FORMAT (Ew.dEe)	⋆
	lbli	FORMAT (ENw.d)	⋆
	lblj	FORMAT (ENw.dEe)	⋆
	lblk	FORMAT (ESw.d)	⋆
	lbll	FORMAT (ESw.dEe)	⋆
$convert_{F'' \rightarrow F}(f)$		READ (UNIT=f,FMT=*lblx*) t	⋆
$convert_{F \rightarrow F''}(y)$		WRITE (UNIT=h, FMT=*lblx*) y	⋆
$convert_{D' \rightarrow F}(f)$		READ (UNIT=f,FMT=*lblx*) t	⋆

where x is an expression of type *INT*, y is an expression of type *FLT*, and z is an expression of type *FXD*, where *FXD* is a fixed point type. *INT2* is the integer datatype that corresponds to I'.

Fortran provides base 10 non-negative numerals for all of its integer and floating point types. Numerals for floating point types must have a '.' in them. The KIND of the a numeral is indicated by a suffix. The details are not repeated in this example binding, see ISO/IEC 1539-1, clause 4.3.1.1 Integer type, and clause 4.3.1.2 Real type.

Fortran does not specify numerals for infinities and NaNs. Suggestion:

+∞	INFINITY	†
qNaN	NAN	†
sNaN	SIGNAN	†

as well as string formats for reading and writing these values as character strings.

Fortran provides recording of indicators for floating point arithmetic notifications, the LIA-2 preferred method. See ISO/IEC TR 15580:1998, *Information technology – Programming languages – Fortran – Floating-point exception handling* [24]. **absolute_precision_underflow** notifications are however ignored.

C.7 Haskell

The programming language Haskell is defined by *Report on the programming language Haskell 98* [66], together with *Standard libraries for the Haskell 98 programming panguage* [67].

An implementation should follow all the requirements of LIA-2 unless otherwise specified by this language binding.

The operations or parameters marked "†" are not part of the language and should be provided by an implementation that wishes to conform to the LIA-2 for that operation. For each of the marked items a suggested identifier is provided.

The Haskell datatype `Bool` corresponds to the LIA datatype **Boolean**.

Every implementation of Haskell has at least two integer datatypes `Integer`, which is unlimited, and `Int`, and at least two floating point datatypes, `Float`, and `Double`. The notation *INT* is used to stand for the name of one of the integer datatypes, and *FLT* is used to stand for the name of one of the floating point datatypes in what follows.

The LIA-2 integer operations are listed below, along with the syntax used to invoke them:

$max_I(x, y)$	max x y or x ‘max‘ y	⋆
$min_I(x, y)$	min x y or x ‘min‘ y	⋆
$max\_seq_I(xs)$	maximum xs	⋆
$min\_seq_I(xs)$	minimum xs	⋆
$dim_I(x, y)$	dim x y or x ‘dim‘ y	†
$power_I(x, y)$	x ^ y or (^) x y	⋆
$shift2_I(x, y)$	shift2 x y or x ‘shift2‘ y	†
$shift10_I(x, y)$	shift10 x y or x ‘shift10‘ y	†
$sqrt_I(x)$	isqrt x	†
$divides_I(x, y)$	divides x y or x ‘divides‘ y	†
$even_I(x)$	even x	⋆
$odd_I(x)$	odd x	⋆
$divf_I(x, y)$	div x y or x ‘div‘ y	⋆
$moda_I(x, y)$	mod x y or x ‘mod‘ y	⋆
$group_I(x, y)$	grp x y or x ‘grp‘ y	†
$pad_I(x, y)$	pad x y or x ‘pad‘ y	†
$quot_I(x, y)$	ratio x y or x ‘ratio‘ y	†
$remr_I(x, y)$	remainder x y or x ‘remainder‘ y	†
$gcd_I(x, y)$	gcd x y or x ‘gcd‘ y	⋆
$lcm_I(x, y)$	lcm x y or x ‘lcm‘ y	⋆
$gcd\_seq_I(xs)$	gcd_seq xs	†
$lcm\_seq_I(xs)$	lcm_seq xs	†
$add\_wrap_I(x, y)$	x +: y	†
$add\_ov_I(x, y)$	x +:+ y	†
$sub\_wrap_I(x, y)$	x -: y	†
$sub\_ov_I(x, y)$	x -:+ y	†
$mul\_wrap_I(x, y)$	x *: y	†
$mul\_ov_I(x, y)$	x *:+ y	†

where x and y are expressions of type INT and where xs is an expression of type [INT].

The LIA-2 non-transcendental floating point operations are listed below, along with the syntax used to invoke them:

$max_F(x, y)$	max x y or x ‘max‘ y	⋆
$min_F(x, y)$	min x y or x ‘min‘ y	⋆
$mmax_F(x, y)$	mmax x y or x ‘mmax‘ y	†
$mmin_F(x, y)$	mmin x y or x ‘mmin‘ y	†
$max\_seq_F(xs)$	maximum xs	⋆
$min\_seq_F(xs)$	minimum xs	⋆
$mmax\_seq_F(xs)$	mmaximum xs	†
$mmin\_seq_F(xs)$	mminimum xs	†
$dim_F(x, y)$	dim x y or x ‘dim‘ y	†
$rounding_F(x)$	fromInteger (round x)	⋆
$floor_F(x)$	fromInteger (floor x)	⋆
$ceiling_F(x)$	fromInteger (ceiling x)	⋆
$rounding\_rest_F(x)$	x - fromInteger (round x)	⋆
$floor\_rest_F(x)$	x - fromInteger (floor x)	⋆

$ceiling\_rest_F(x)$	`x - fromInteger (ceiling x)`	⋆
$remr_F(x, y)$	`remainder x y` or `x ‘remainder‘ y`	†
$sqrt_F(x)$	`sqrt x`	⋆
$rsqrt_F(x)$	`rsqrt x`	†
$add\_lo_F(x, y)$	`x +:- y`	†
$sub\_lo_F(x, y)$	`x -:- y`	†
$mul\_lo_F(x, y)$	`x *:- y`	†
$div\_rest_F(x, y)$	`x /:* y`	†
$sqrt\_rest_F(x)$	`sqrt_rest x`	†
$mul_{F \to F'}(x, y)$	`prod x y`	†

where x, y and z are expressions of type *FLT*, and where xs is an expression of type `[FLT]`.

The binding for the floor, round, and ceiling operations here take advantage of the unbounded `Integer` type in Haskell, and that `Integer` is the default integer type.

The LIA-2 parameters for operations approximating real valued transcendental functions can be accessed by the following syntax:

$max\_err\_hypot_F$	`err_hypotenuse x`	†
$max\_err\_exp_F$	`err_exp x`	†
$max\_err\_power_F$	`err_power x`	†
$max\_err\_sinh_F$	`err_sinh x`	†
$max\_err\_tanh_F$	`err_tanh x`	†
$big\_angle\_r_F$	`big_radian_angle x`	†
$max\_err\_sin_F$	`err_sin x`	†
$max\_err\_tan_F$	`err_tan x`	†
$min\_angular\_unit_F$	`min_angle_unit x`	†
$big\_angle\_u_F$	`big_angle x`	†
$max\_err\_sinu_F(u)$	`err_sin_cycle u`	†
$max\_err\_tanu_F(u)$	`err_tan_cycle u`	†
$max\_err\_convert_F$	`err_convert x`	†
$max\_err\_convert_{F'}$	`err_convert ""`	†
$max\_err\_convert_{D'}$	`err_convert ""`	†

where b, x and u are expressions of type *FLT*. Several of the parameter functions are constant for each type (and library), the argument is then used only to differentiate among the floating point types.

The LIA-2 elementary floating point operations are listed below, along with the syntax used to invoke them:

$hypot_F(x, y)$	`hypotenuse x y`	†
$power_{FI}(b, z)$	`b ^^ z` or `(^^) b z`	⋆
$exp_F(x)$	`exp x`	⋆
$expm1_F(x)$	`expM1 x`	†
$exp2_F(x)$	`exp2 x`	†
$exp10_F(x)$	`exp10 x`	†
$power_F(b, y)$	`b ** y` or `(**) b y`	⋆

$power1pm1_F(b, y)$	`power1PM1` b y or b `‘power1PM1‘` y	†
$ln_F(x)$	`log` x	⋆
$ln1p_F(x)$	`log1P` x	†
$log2_F(x)$	`log2` x	†
$log10_F(x)$	`log10` x	†
$logbase_F(b, x)$	`logBase` b x or b `‘logBase‘` x	⋆
$logbase1p1p_F(b, x)$	`logBase1P1P` b x	†
$sinh_F(x)$	`sinh` x	⋆
$cosh_F(x)$	`cosh` x	⋆
$tanh_F(x)$	`tanh` x	⋆
$coth_F(x)$	`coth` x	†
$sech_F(x)$	`sech` x	†
$csch_F(x)$	`csch` x	†
$arcsinh_F(x)$	`asinh` x	⋆
$arccosh_F(x)$	`acosh` x	⋆
$arctanh_F(x)$	`atanh` x	⋆
$arccoth_F(x)$	`acoth` x	†
$arcsech_F(x)$	`asech` x	†
$arccsch_F(x)$	`acsch` x	†
$axis\_rad_F(x)$	`axis_radians` x	†
$rad_F(x)$	`radians` x	†
$sin_F(x)$	`sin` x	⋆
$cos_F(x)$	`cos` x	⋆
$tan_F(x)$	`tan` x	⋆
$cot_F(x)$	`cot` x	†
$sec_F(x)$	`sec` x	†
$csc_F(x)$	`csc` x	†
$cossin_F(x)$	`cosSin` x	†
$arcsin_F(x)$	`asin` x	⋆
$arccos_F(x)$	`acos` x	⋆
$arctan_F(x)$	`atan` x	⋆
$arccot_F(x)$	`acot` x	†
$arcctg_F(x)$	`actg` x	†
$arcsec_F(x)$	`asec` x	†
$arccsc_F(x)$	`acsc` x	†
$arc_F(x, y)$	`atan2` y x	⋆
$cycle_F(u, x)$	`cycle` u x	†
$axis\_cycle_F(u, x)$	`axis_cycle` u x	†
$sinu_F(u, x)$	`sinU` u x	†
$cosu_F(u, x)$	`cosU` u x	†
$tanu_F(u, x)$	`tanU` u x	†
$cotu_F(u, x)$	`cotU` u x	†
$secu_F(u, x)$	`secU` u x	†

$cscu_F(u,x)$	`cscU` u x	†
$cossinu_F(x)$	`cosSinU` u x	†
$arcsinu_F(u,x)$	`asinU` u x	†
$arccosu_F(u,x)$	`acosU` u x	†
$arctanu_F(u,x)$	`atanU` u x	†
$arccotu_F(u,x)$	`acotU` u x	†
$arcctgu_F(u,x)$	`acotU` u x	†
$arcsecu_F(u,x)$	`asecU` u x	†
$arccscu_F(u,x)$	`acscU` u x	†
$arcu_F(u,x,y)$	`atan2U` u y x	†
$rad\_to\_cycle_F(x,u)$	`rad_to_cycle` x u	†
$cycle\_to\_rad_F(u,x)$	`cycle_to_rad` u x	†
$cycle\_to\_cycle_F(u,x,v)$	`cycle_to_cycle` u x v	†

where b, x, y, u, and v are expressions of type *FLT*, and z is an expressions of type *INT*.

Arithmetic value conversions in Haskell are always explicit. They are done with the overloaded `fromIntegral` and `fromFractional` operations.

$convert_{I \rightarrow I'}(x)$	`fromIntegral` x	⋆
$convert_{I'' \rightarrow I}(x)$	`read` s	⋆
$convert_{I \rightarrow I''}(x)$	`show` x	⋆
$rounding_{F \rightarrow I}(y)$	`round(`y`)`	⋆
$floor_{F \rightarrow I}(y)$	`floor(`y`)`	⋆
$ceiling_{F \rightarrow I}(y)$	`ceiling(`y`)`	⋆
$convert_{I \rightarrow F}(x)$	`fromIntegral` x	⋆
$convert_{F \rightarrow F'}(y)$	`fromFractional` y	⋆
$convert_{F'' \rightarrow F}(s)$	`read` s	.⋆
$convert_{F \rightarrow F''}(y)$	`show` y	...⋆
$convert_{D' \rightarrow F}(s)$	`read` s	.⋆
$convert_{F \rightarrow D'}(y)$	`show` y	...⋆

where x is an expression of type *INT*, y is an expression of type *FLT*, and z is an expression of type *FXD*, where *FXD* is a fixed point type.

Haskell provides non-negative numerals for all its integer and floating point types in base is 10. There is no differentiation between the numerals for different floating point types, nor between numerals for different integer types, and integer numerals can be used for floating point values. Integer numerals stand for a value in `Integer` (the unbounded integer type) and an implicit `fromInteger` operation is applied to it. Fractional numerals stand for a value in `Rationale` (the unbounded type of rational numbers) and an implicit `fromRational` operation is applied to it.

Haskell does not specify any numerals for infinities and NaNs. Suggestion:

$+\infty$	`infinity`	†
qNaN	`quietNaN`	†
sNaN	`sigallingNaN`	†

as well as string formats for reading and writing these values as character strings.

Haskell has the notion of `error`, which results in a change of ‘control flow’, which cannot be returned from, nor caught. An `error` results in the termination of the program. (There are suggestions to improve this.) **pole** for integer types and **invalid** (in general) are considered to be `error`. No notification results for **underflow**, and the continuation value (specified by LIA-2) is used directly, since recording of indicators is not available and `error` is inappropriate for **underflow**. The handling of integer **overflow** is implementation dependent. The handling of floating point **overflow** and **pole** should be to return a suitable infinity (specified by LIA-2), if possible, without any notification, since recording of indicators is not available.

C.8 Java

The programming language Java is defined by *The Java Language Specification* [65].

An implementation should follow all the requirements of LIA-2 unless otherwise specified by this language binding.

The operations or parameters marked "†" are not part of the language and should be provided by an implementation that wishes to conform to the LIA-2 for that operation. For each of the marked items a suggested identifier is provided. The LIA-2 operations that are provided in Java 2 (marked "⋆" below) are in the final class `java.lang.Math`.

The Java datatype `boolean` corresponds to the LIA datatype **Boolean**.

Every implementation of Java has the integral datatypes `int`, and `long`.

Java has two floating point datatypes, `float` and `double`, which must conform to IEC 60559.

The LIA-2 integer operations are listed below, along with the syntax used to invoke them:

$min_I(x, y)$	`min(x, y)`	⋆
$max_I(x, y)$	`max(x, y)`	⋆
$min\_seq_I(xs)$	`min_arr(xs)`	†
$max\_seq_I(xs)$	`max_arr(xs)`	†
$dim_I(x, y)$	`dim(x, y)`	†
$sqrt_I(x)$	`sqrt(x)`	†
$power_I(x, y)$	`power(x, y)`	†
$divides_I(x, y)$	`divides(x, y)`	†
$even_I(x)$	`x % 2 == 0`	⋆
$odd_I(x)$	`x % 2 != 0`	⋆
$divf_I(x, y)$	`div(x, y)`	†
$moda_I(x, y)$	`mod(x, y)`	†
$group_I(x, y)$	`group(x, y)`	†
$pad_I(x, y)$	`pad(x, y)`	†
$quot_I(x, y)$	`quot(x, y)`	†
$remr_I(x, y)$	`rem(x, y)`	†
$gcd_I(x, y)$	`gcd(x, y)`	†
$lcm_I(x, y)$	`lcm(x, y)`	†
$gcd\_seq_I(xs)$	`gcd_arr(xs)`	†
$lcm\_seq_I(xs)$	`lcm_arr(xs)`	†
$add\_wrap_I(x, y)$	`add_wrap(x, y)`	†
$add\_ov_I(x, y)$	`add_over(x, y)`	†
$sub\_wrap_I(x, y)$	`sub_wrap(x, y)`	†
$sub\_ov_I(x, y)$	`sub_over(x, y)`	†

$mul\_wrap_I(x, y)$	`mul_wrap(x, y)`	†
$mul\_ov_I(x, y)$	`mul_over(x, y)`	†

where x and y are expressions of type *INT* and where xs is an expression of type `array of` *INT*.

The LIA-2 non-transcendental floating point operations are listed below, along with the syntax used to invoke them:

$min_F(x, y)$	`min(x, y)`	⋆
$max_F(x, y)$	`max(x, y)`	⋆
$mmax_F(x, y)$	`mmax(x, y)`	†
$mmin_F(x, y)$	`mmin(x, y)`	†
$min\_seq_F(xs)$	`min_arr(xs)`	†
$max\_seq_F(xs)$	`max_arr(xs)`	†
$mmax\_seq_F(xs)$	`mmax(xs)`	†
$mmin\_seq_F(xs)$	`mmin(xs)`	†
$rounding_F(x)$	`rint(x)`	⋆
$floor_F(x)$	`floor(x)`	⋆
$ceiling_F(x)$	`ceil(x)`	⋆
$dim_F(x, y)$	`dim(x, y)`	†
$dprod_{F \to F'}(x, y)$	`dprod(x, y)`	†
$remr_F(x, y)$	`IEEEremainder(x, y)`	⋆ (only for `double`)
$sqrt_F(x)$	`sqrt(x)`	⋆
$rsqrt_F(x)$	`rsqrt(x)`	†
$add\_lo_F(x, y)$	`add_low(x, y)`	†
$sub\_lo_F(x, y)$	`sub_low(x, y)`	†
$mul\_lo_F(x, y)$	`mul_low(x, y)`	†
$div\_rest_F(x, y)$	`div_rest(x, y)`	†
$sqrt\_rest_F(x)$	`sqrt_rest(x)`	†

where x, y and z are expressions of type *FLT*, and where xs is an expression of type `array of` *FLT*.

The LIA-2 parameters for operations approximating real valued transcendental functions can be accessed by the following syntax:

$max\_err\_hypot_F$	`err_hypotenuse(x)`	†
$max\_err\_exp_F$	`err_exp(x)`	†
$max\_err\_power_F(b, x)$	`err_power(b, x)`	†
$max\_err\_sinh_F$	`err_sinh(x)`	†
$max\_err\_tanh_F$	`err_tanh(x)`	†
$big\_radian\_angle_F$	`big_radian_angle(x)`	†
$max\_err\_sin_F$	`err_sin(x)`	†
$max\_err\_tan_F$	`err_tan(x)`	†
$min\_angular\_unit_F$	`smallest_angular_unit(x)`	†
big_angle_F	`big_angle(x)`	†
$max\_err\_sinu_F(u)$	`err_sin_cycle(u)`	†
$max\_err\_tanu_F(u)$	`err_tan_cycle(u)`	†

$\mathit{max\_err\_convert}_F$	err_convert(x)	†
$\mathit{max\_err\_convert}_{F'}$	err_convert_to_string	†
$\mathit{max\_err\_convert}_{D'}$	err_convert_to_string	†

where b, x and u are expressions of type *FLT*. Several of the parameter functions are constant for each type (and library), the argument is then used only to differentiate among the floating point types.

The LIA-2 elementary floating point operations are listed below, along with the syntax used to invoke them. These are defined only for double not for float.

$\mathit{hypot}_F(x, y)$	hypotenuse(x, y)	†
$\mathit{power}_{FI}(b, z)$	poweri(b, z)	†
$\mathit{exp}_F(x)$	exp(x)	⋆
$\mathit{expm1}_F(x)$	expm1(x)	†
$\mathit{exp2}_F(x)$	exp2(x)	†
$\mathit{exp10}_F(x)$	exp10(x)	†
$\mathit{power}_F(b, y)$	power(b, y)	†
$\mathit{pow}_F(b, y)$	pow(b, y)	⋆ Not LIA-2!
$\mathit{power1pm1}_F(b, y)$	power1pm1(b, y)	†
$\mathit{ln}_F(x)$	log(x)	⋆
$\mathit{ln1p}_F(x)$	log1p(x)	†
$\mathit{log2}_F(x)$	log2(x)	†
$\mathit{log10}_F(x)$	log10(x)	†
$\mathit{logbase}_F(b, x)$	log(b, x)	†
$\mathit{logbase1p1p}_F(b, x)$	log1p1p(b, x)	†
$\mathit{sinh}_F(x)$	sinh(x)	†
$\mathit{cosh}_F(x)$	cosh(x)	†
$\mathit{tanh}_F(x)$	tanh(x)	†
$\mathit{coth}_F(x)$	coth(x)	†
$\mathit{sech}_F(x)$	sech(x)	†
$\mathit{csch}_F(x)$	csch(x)	†
$\mathit{arcsinh}_F(x)$	asinh(x)	†
$\mathit{arccosh}_F(x)$	acosh(x)	†
$\mathit{arctanh}_F(x)$	atanh(x)	†
$\mathit{arccoth}_F(x)$	acoth(x)	†
$\mathit{arcsech}_F(x)$	asech(x)	†
$\mathit{arccsch}_F(x)$	acsch(x)	†
$\mathit{rad}_F(x)$	radian(x)	†
$\mathit{axis\_rad}_F(x)$	axis_rad(x)	†
$\mathit{sin}_F(x)$	sin(x)	⋆
$\mathit{cos}_F(x)$	cos(x)	⋆
$\mathit{tan}_F(x)$	tan(x)	⋆
$\mathit{cot}_F(x)$	cot(x)	†
$\mathit{sec}_F(x)$	sec(x)	†
$\mathit{csc}_F(x)$	csc(x)	†

$arcsin_F(x)$	asin(x)	⋆
$arccos_F(x)$	acos(x)	⋆
$arctan_F(x)$	atan(x)	⋆
$arccot_F(x)$	acot(x)	†
$arcctg_F(x)$	actg(x)	†
$arcsec_F(x)$	asec(x)	†
$arccsc_F(x)$	acsc(x)	†
$arc_F(x,y)$	atan2(y, x)	⋆

$cycle_F(u,x)$	cycle(u, x)	†
$axis\_cycle_F(u,x)$	axis_cycle(u, x)	†

$sinu_F(u,x)$	sinu(u, x)	†
$cosu_F(u,x)$	cosu(u, x)	†
$tanu_F(u,x)$	tanu(u, x)	†
$cotu_F(u,x)$	cotu(u, x)	†
$secu_F(u,x)$	secu(u, x)	†
$cscu_F(u,x)$	cscu(u, x)	†

$arcsinu_F(u,x)$	asinu(u, x)	†
$arccosu_F(u,x)$	acosu(u, x)	†
$arctanu_F(u,x)$	atanu(u, x)	†
$arccotu_F(u,x)$	acotu(u, x)	†
$arcctgu_F(u,x)$	actgu(u, x)	†
$arcsecu_F(u,x)$	asecu(u, x)	†
$arccscu_F(u,x)$	acscu(u, x)	†
$arcu_F(u,x,y)$	atan2u(u, y, x)	†

$rad\_to\_cycle_F(x,u)$	radian_to_cycle(x, u)	†
$cycle\_to\_rad_F(u,x)$	cycle_to_radian(u, x)	†
$cycle\_to\_cycle_F(u,x,v)$	cycle_to_cycle(u, x, v)	†
$rad\_to\_cycle_F(x,360)$	toDegrees(x)	⋆
$cycle\_to\_rad_F(360,x)$	toRadians(x)	⋆

where b, x, y, u, and v are expressions of type *FLT*, and z is an expressions of type *INT*

Arithmetic value conversions in Java can be explicit or implicit. The rules for when implicit conversions are applied is not repeated here. The explicit arithmetic value conversions are usually expressed as 'casts', except when converting to/from strings.

$convert_{I\to I'}(x)$	(*INT2*)x	⋆
$convert_{I''\to I}(s)$	Integer.parseInt(s)	⋆
$convert_{I''\to I}(s)$	Integer.parseInt(s,*radix*)	⋆
$convert_{I''\to I}(s)$	Long.parseLong(s)	⋆
$convert_{I''\to I}(s)$	Long.parseLong(s,*radix*)	⋆
$convert_{I\to I''}(x)$	Integer.toString(x)	⋆
$convert_{I\to I''}(x)$	Integer.toString(x,*radix*)	⋆
$convert_{I\to I''}(x)$	Integer.toBinaryString(x)	⋆
$convert_{I\to I''}(x)$	Integer.toOctalString(x)	⋆
$convert_{I\to I''}(x)$	Integer.toHexString(x)	⋆
$convert_{I\to I''}(x)$	Long.toString(x)	⋆
$convert_{I\to I''}(x)$	Long.toString(x,*radix*)	⋆
$convert_{I\to I''}(x)$	Long.toBinaryString(x)	⋆

$convert_{I \rightarrow I''}(x)$	`Long.toOctalString(`x`)`	⋆
$convert_{I \rightarrow I''}(x)$	`Long.toHexString(`x`)`	⋆
$convert_{I \rightarrow I''}(x)$	`""+`x	⋆
$rounding_{F \rightarrow I}(y)$	(*INT*)`rint(`y`)`	⋆
$floor_{F \rightarrow I}(y)$	(*INT*)`floor(`y`)`	⋆
$ceiling_{F \rightarrow I}(y)$	(*INT*)`ceil(`y`)`	⋆
$convert_{I \rightarrow F}(x)$	(*FLT*)x	⋆
$convert_{F \rightarrow F'}(y)$	(*FLT2*)y	⋆
$convert_{F'' \rightarrow F}(s)$	`Float.parseFloat(`s`)`	⋆
$convert_{F'' \rightarrow F}(s)$	`Double.parseDouble(`s`)`	⋆
$convert_{F \rightarrow F''}(y)$	`Float.toString(`x`)`	⋆
$convert_{F \rightarrow F''}(y)$	`Double.toString(`x`)`	⋆
$convert_{D' \rightarrow F}(s)$	`Float.parseFloat(`s`)`	⋆
$convert_{D' \rightarrow F}(s)$	`Double.parseDouble(`s`)`	⋆

where x is an expression of type *INT*, y is an expression of type *FLT*, and z is an expression of type *FXD*, where *FXD* is a fixed point type. *INT2* is the integer datatype that corresponds to I'. A ? above indicates that the parameter is optional. e is greater than 0.

Java provides non-negative numerals for all its integer and floating point types. The default base is 10, but for integers base 8 and 16 can be used too. Numerals for different integer types are distinguished by suffixes. Numerals for different floating point types are distinguished by suffix: `f` for `float`, no suffix for `double`, `l` for `long double`. Numerals for floating point types must have a '.' in them. The details are not repeated in this example binding, see *The Java Language Specification*, clause 3.10.1 Integer literals, and clause 3.10.2 Floating-point literals.

Java specifies numerals for infinities and NaNs:

$+\infty$	`Float.POSITIVE_INFINITY`	⋆
$+\infty$	`Double.POSITIVE_INFINITY`	⋆
$-\infty$	`Float.NEGATIVE_INFINITY`	⋆
$-\infty$	`Double.NEGATIVE_INFINITY`	⋆
qNaN	`Float.NaN`	⋆
qNaN	`Double.NaN`	⋆
sNaN	`Float.SigNaN`	†
sNaN	`Double.SigNaN`	†

as well as string formats for writing these values as character strings. However, infinities and NaNs cannot be converted *from* string.

Java has a notion of 'exception' that implies a non-returnable, but catchable, change of control flow. Java uses its exception mechanism as its default means of notification. Java ignores **underflow** notifications since a Java exception is inappropriate for an **underflow** notification. On **underflow** the continuation value (specified in LIA-2) is used directly without recording the **underflow** itself. Java also ignores **pole** and **overflow** notifications for floating point operations, and the continuation value (specified in LIA-2) is used directly without recording the **pole** or **overflow** itself. Java uses the exception `java.lang.ArithmeticException` for **invalid** notifications and for **pole** notifications for integer operations. Java, however, ignores **pole** and **invalid** for `log` and `sqrt`. Java uses `java.lang.NumberFormatException` for **invalid** (and **pole**) notifications for operations that convert from string. Since Java exceptions are non-returnable changes of control flow, no continuation value is provided for these notifications.

An implementation that wishes to follow LIA-2 should provide recording of indicators as an alternative means of handling numeric notifications, including those that Java ignores when the numeric notification handling mechanism is by Java exceptions. Recording of indicators is the LIA-2 preferred means of handling numeric notifications.

C.9 Common Lisp

The programming language Common Lisp is defined by ANSI X3.226-1994, *Information Technology – Programming Language – Common Lisp* [43].

An implementation should follow all the requirements of LIA-2 unless otherwise specified by this language binding.

The operations or parameters marked "†" are not part of the language and should be provided by an implementation that wishes to conform to the LIA-2 for that operation. For each of the marked items a suggested identifier is provided.

Common Lisp does not have a single datatype that corresponds to the LIA-1 datatype **Boolean**. Rather, `NIL` corresponds to **false** and `T` corresponds to **true**.

Every implementation of Common Lisp has one unbounded integer datatype. Any mathematical integer is assumed to have a representation as a Common Lisp data object, subject only to total memory limitations.

Common Lisp has four floating point types: `short-float`, `single-float`, `double-float`, and `long-float`. Not all of these floating point types must be distinct.

The additional integer operations are listed below, along with the syntax used to invoke them:

$min_I(x, y)$	`(min x y)`	⋆
$max_I(x, y)$	`(max x y)`	⋆
$min\_seq_I(xs)$	`(min . xs)` or (min x_1 x_2 ... x_n)	⋆
$max\_seq_I(xs)$	`(max . xs)` or (max x_1 x_2 ... x_n)	⋆
$dim_I(x, y)$	`(dim x y)`	†
$sqrt_I(x)$	`(isqrt x)`	†
$power_I(x, y)$	`(expt x y)` (returns a rational on negative power)	⋆
$shift2_I(x, y)$	`(shift2 x y)`	†
$shift10_I(x, y)$	`(shift10 x y)`	†
$divides_I(x, y)$	`(dividesp x y)`	†
$even_I(x)$	`(evenp x)`	⋆
$odd_I(x)$	`(oddp x)`	⋆

(the `floor`, `ceiling`, and `round` can also accept floating point arguments)

`(multiple-value-bind (flr md) (floor x y))`

$divf_I(x, y)$	`flr` or `(floor x y)`	⋆
$moda_I(x, y)$	`md` or `(mod x y)`	⋆

`(multiple-value-bind (ceil pd) (ceiling x y))`

$group_I(x, y)$	`ceil` or `(ceiling x y)`	⋆
$pad_I(x, y)$	`(- pd)`	⋆

`(multiple-value-bind (rnd rm) (round x y))`

$quot_I(x, y)$	`rnd` or `(round x y)`	⋆
$remr_I(x, y)$	`rm`	⋆

$gcd_I(x,y)$	(gcd x y) (deviation: (gcd 0 0) is 0)	⋆
$lcm_I(x,y)$	(lcm x y)	⋆
$gcd\_seq_I(xs)$	(gcd . xs) or (gcd x_1 x_2 ... x_n)	⋆
$lcm\_seq_I(xs)$	(lcm . xs) or (lcm x_1 x_2 ... x_n)	⋆
$add\_wrap_I(x,y)$	(add-wrap x y)	†
$add\_ov_I(x,y)$	(add-over x y)	†
$sub\_wrap_I(x,y)$	(sub-wrap x y)	†
$sub\_ov_I(x,y)$	(sub-over x y)	†
$mul\_wrap_I(x,y)$	(mul-wrap x y)	†
$mul\_ov_I(x,y)$	(mul-over x y)	†

where x and y are expressions of type *INT* and where xs is an expression of type list of *INT*.

The LIA-2 non-transcendental floating point operations are listed below, along with the syntax used to invoke them:

$min_F(x,y)$	(min x y)	⋆
$max_F(x,y)$	(max x y)	⋆
$min\_seq_F(xs)$	(min . xs) or (min x_1 x_2 ... x_n)	⋆
$max\_seq_F(xs)$	(max . xs) or (max x_1 x_2 ... x_n)	⋆
	(multiple-value-bind (flr frem) (ffloor x))	
$floor_F(x)$	(ffloor x) or flr	⋆
$floor\_rest_F(x)$	frem	⋆
	(multiple-value-bind (rnd rrem) (fround x))	
$rounding_F(x)$	(fround x) or rnd	⋆
$rounding\_rest_F(x)$	rrem	⋆
	(multiple-value-bind (cln crem) (fceiling x))	
$ceiling_F(x)$	(fceiling x) or cln	⋆
$ceiling\_rest_F(x)$	crem	⋆
$dim_F(x,y)$	(dim x y)	†
	(multiple-value-bind (rqt remainder) (fround x y))	
$remr_F(x,y)$	remainder	⋆
$sqrt_F(x)$	(sqrt x) (returns a complex on negative arg.)	⋆
$rsqrt_F(x)$	(rsqrt x)	†
$dprod_{F \to F'}(x,y)$	(prod x y)	†
$add\_lo_F(x,y)$	(add-low x y)	†
$sub\_lo_F(x,y)$	(sub-low x y)	†
$mul\_lo_F(x,y)$	(mul-low x y)	†
$div\_rest_F(x,y)$	(div-rest x y)	†
$sqrt\_rest_F(x)$	(sqrt-rest x)	†

where x, y and z are data objects of the same floating point type, and where xs is a data objects that is a list of data objects of (the same, in this binding) floating point type. Note that Common Lisp allows mixed number types in many of its operations. This example binding does not explain that in detail.

The LIA-2 parameters for operations approximating real valued transcendental functions can be accessed by the following syntax:

$max\_err\_hypot_F$	(err-hypotenuse x)	†

$\mathit{max\_err\_exp}_F$	`(err-exp` x`)`	†
$\mathit{max\_err\_power}_F$	`(err-power` x`)`	†
$\mathit{max\_err\_sinh}_F$	`(err-sinh` x`)`	†
$\mathit{max\_err\_tanh}_F$	`(err-tanh` x`)`	†
$\mathit{big\_radian\_angle}_F$	`(big-radian-angle` x`)`	†
$\mathit{max\_err\_sin}_F$	`(err-sin` x`)`	†
$\mathit{max\_err\_tan}_F$	`(err-tan` x`)`	†
$\mathit{min\_angular\_unit}_F$	`(minimum-angular-unit` x`)`	†
$\mathit{big\_angle\_u}_F$	`(big-angle` x`)`	†
$\mathit{max\_err\_sinu}_F(u)$	`(err-sin-cycle` u`)`	†
$\mathit{max\_err\_tanu}_F(u)$	`(err-tan-cycle` u`)`	†
$\mathit{max\_err\_convert}_F$	`(err_convert` x`)`	†
$\mathit{max\_err\_convert}_F$	`err-convert-to-string`	†
$\mathit{max\_err\_convert}_D$	`err-convert-to-string`	†

where b, x and u are expressions of type *FLT*. Several of the parameter functions are constant for each type (and library), the argument is then used only to differentiate among the floating point types.

The LIA-2 elementary floating point operations are listed below, along with the syntax used to invoke them:

$\mathit{hypot}_F(x, y)$	`(hypotenuse` x y`)`	†
$\mathit{power}_{FI}(b, z)$	`(expt` b z`)`	⋆
$\mathit{exp}_F(x)$	`(exp` x`)`	⋆
$\mathit{exp2}_F(x)$	`(exp2` x`)`	†
$\mathit{exp10}_F(x)$	`(exp10` x`)`	†
$\mathit{expm1}_F(x)$	`(expm1` x`)`	†
$\mathit{power}_F(b, y)$	`(expt` b y`)` (deviation: `(expt 0.0 0.0)` is 1)	⋆
$\mathit{power1pm1}_F(b, y)$	`(expt1pm1` b y`)`	†
$\mathit{ln}_F(x)$	`(log` x`)` (returns a complex on negative arg.)	⋆
$\mathit{ln1p}_F(x)$	`(log1p` x`)`	†
$\mathit{log2}_F(x)$	`(log2` x`)`	†
$\mathit{log10}_F(x)$	`(log10` x`)`	†
$\mathit{logbase}_F(b, x)$	`(log` x b`)` (note parameter order)	⋆
$\mathit{logbase1p1p}_F(b, x)$	`(log1p` x b`)`	†
$\mathit{sinh}_F(x)$	`(sinh` x`)`	⋆
$\mathit{cosh}_F(x)$	`(cosh` x`)`	⋆
$\mathit{tanh}_F(x)$	`(tanh` x`)`	⋆
$\mathit{coth}_F(x)$	`(coth` x`)`	†
$\mathit{sech}_F(x)$	`(sech` x`)`	†
$\mathit{csch}_F(x)$	`(csch` x`)`	†
$\mathit{arcsinh}_F(x)$	`(asinh` x`)`	⋆
$\mathit{arccosh}_F(x)$	`(acosh` x`)` (returns a complex when $x < 1$)	⋆

$\mathit{arctanh}_F(x)$	(atanh x)	(returns a complex when $\lvert x\rvert > 1$)	$\star$
$\mathit{arccoth}_F(x)$	(acoth x)		†
$\mathit{arcsech}_F(x)$	(asech x)		†
$\mathit{arccsch}_F(x)$	(acsch x)		†
$\mathit{rad}_F(x)$	(radians x)		†
$\mathit{axis\_rad}_F(x)$	(axis_rad x)		†
$\mathit{sin}_F(x)$	(sin x)		$\star$
$\mathit{cos}_F(x)$	(cos x)		$\star$
$\mathit{tan}_F(x)$	(tan x)		$\star$
$\mathit{cot}_F(x)$	(cot x)		†
$\mathit{sec}_F(x)$	(sec x)		†
$\mathit{csc}_F(x)$	(csc x)		†
$\mathit{arcsin}_F(x)$	(asin x)	(returns a complex when $\lvert x\rvert > 1$)	$\star$
$\mathit{arccos}_F(x)$	(acos x)	(returns a complex when $\lvert x\rvert > 1$)	$\star$
$\mathit{arctan}_F(x)$	(atan x)		$\star$
$\mathit{arccot}_F(x)$	(acot x)		†
$\mathit{arcctg}_F(x)$	(actg x)		†
$\mathit{arcsec}_F(x)$	(asec x)		†
$\mathit{arccsc}_F(x)$	(acsc x)		†
$\mathit{arc}_F(x, y)$	(atan y x)		$\star$
$\mathit{cycle}_F(u, x)$	(cycle u x)		†
$\mathit{axis\_cycle}_F(u, x)$	(axis_cycle u x)		†
$\mathit{sinu}_F(u, x)$	(sinU u x)		†
$\mathit{cosu}_F(u, x)$	(cosU u x)		†
$\mathit{tanu}_F(u, x)$	(tanU u x)		†
$\mathit{cotu}_F(u, x)$	(cotU u x)		†
$\mathit{secu}_F(u, x)$	(secU u x)		†
$\mathit{cscu}_F(u, x)$	(cscU u x)		†
$\mathit{arcsinu}_F(u, x)$	(asinU u x		†
$\mathit{arccosu}_F(u, x)$	(acosU u x		†
$\mathit{arctanu}_F(u, x)$	(atanU u x		†
$\mathit{arccotu}_F(u, x)$	(acotU u x		†
$\mathit{arcctgu}_F(u, x)$	(actgU u x)		†
$\mathit{arcsecu}_F(u, x)$	(asecU u x)		†
$\mathit{arccscu}_F(u, x)$	(acscU u x)		†
$\mathit{arcu}_F(u, x, y)$	(atanU u y x)		†
$\mathit{rad\_to\_cycle}_F(x, u)$	(rad_to_cycle x u)		†
$\mathit{cycle\_to\_rad}_F(u, x)$	(cycle_to_rad u x)		†
$\mathit{cycle\_to\_cycle}_F(u, x, v)$	(cycle_to_cycle u x v)		†

where b, x, y, u, and v are expressions of type *FLT*, and z is an expressions of type *INT*.

Arithmetic value conversions in Common Lisp are can be explicit or implicit. The rules for when implicit conversions are done is implementation defined.

$\mathit{convert}_{I\to I''}(x)$	(format nil "~wB" x)	$\star$(binary)
$\mathit{convert}_{I\to I''}(x)$	(format nil "~wO" x)	$\star$(octal)
$\mathit{convert}_{I\to I''}(x)$	(format nil "~wD" x)	$\star$(decimal)

$convert_{I\rightarrow I''}(x)$	(format nil "~wX" x)	$\star$(hexadecimal)
$convert_{I\rightarrow I''}(x)$	(format nil "~r,wR" x)	$\star$(radix r)
$convert_{I\rightarrow I''}(x)$	(format nil "~@R" x)	$\star$(roman numeral)
$rounding_{F\rightarrow I}(y)$	(round y)	$\star$
$floor_{F\rightarrow I}(y)$	(floor y)	$\star$
$ceiling_{F\rightarrow I}(y)$	(ceiling y)	$\star$
$convert_{I\rightarrow F}(x)$	(float x $kind$)	$\star$
$convert_{F\rightarrow F'}(y)$	(float y $kind$)	$\star$
$convert_{F\rightarrow F''}(y)$	(format nil "~wF" y)	$\star$
$convert_{F\rightarrow F''}(y)$	(format nil "~w,e,k,cE" y)	$\star$
$convert_{F\rightarrow F''}(y)$	(format nil "~w,e,k,cG" y)	$\star$
$convert_{F\rightarrow D'}(y)$	(format nil "~r,w,0,#F" y)	$\star$

where x is an expression of type *INT*, y is an expression of type *FLT*, and z is an expression of type *FXD*, where *FXD* is a fixed point type. Convertion from string to numeric value is in Common Lisp done via a general read procedure, which reads Common Lisp 'S-expressions'.

Common Lisp provides non-negative numerals for all its integer and floating point types in base 10. There is no differentiation between the numerals for different floating point datatypes, nor between numerals for different integer types, and integer numerals can be used for floating point values.

Common Lisp does not specify numerals for infinities and NaNs. Suggestion:

+∞	infinity-*FLT*	†
qNaN	nan-*FLT*	†
sNaN	signan-*FLT*	†

as well as string formats for reading and writing these values as character strings.

Common Lisp has a notion of 'exception', but it is unclear if it is used for any of the arithmetic operations for overflow or pole. However, Common Lisp has no notion of compile time type checking, and an operation can return differently typed values for different arguments. When justifiable, Common Lisp arithmetic operations returns a complex floating point value rather than giving a notification, even if the argument(s) to the operation were not complex. For instance, `(sqrt -1)` (quietly) returns a representation of $0 + i$.

C.10 ISLisp

The programming language ISLisp is defined by ISO/IEC 13816:1997, *Information technology – Programming languages, their environments and system software interfaces – Programming language ISLISP* [25].

An implementation should follow all the requirements of LIA-2 unless otherwise specified by this language binding.

The operations or parameters marked "†" are not part of the language and should be provided by an implementation that wishes to conform to the LIA-2 for that operation. For each of the marked items a suggested identifier is provided.

ISLisp does not have a single datatype that corresponds to the LIA datatype **Boolean**. Rather, `NIL` corresponds to **false** and `T` corresponds to **true**.

Every implementation of ISLisp has one unbounded integer datatype. Any mathematical integer is assumed to have a representation as a ISLisp data object, subject only to total memory limitations.

ISLisp has one floating point type required to conform to IEC 60559.

The additional integer operations are listed below, along with the syntax used to invoke them:

$\mathit{min}_I(x, y)$	`(min x y)`	⋆
$\mathit{max}_I(x, y)$	`(max x y)`	⋆
$\mathit{min\_seq}_I(xs)$	`(min . xs)` or `(min x1 x2 ... xn)`	⋆
$\mathit{max\_seq}_I(xs)$	`(max . xs)` or `(max x1 x2 ... xn)`	⋆
$\mathit{dim}_I(x, y)$	`(dim x y)`	†
$\mathit{sqrt}_I(x)$	`(isqrt x)`	⋆
$\mathit{power}_I(x, y)$	`(expt x y)` (deviation: `(expt 0 0)` is 1)	⋆
$\mathit{shift2}_I(x, y)$	`(shift2 x y)`	†
$\mathit{shift10}_I(x, y)$	`(shift10 x y)`	†
$\mathit{divides}_I(x, y)$	`(dividesp x y)`	†
$\mathit{even}_I(x)$	`(evenp x)`	†
$\mathit{odd}_I(x)$	`(oddp x)`	†
$\mathit{divf}_I(x, y)$	`(div x y)`	⋆
$\mathit{moda}_I(x, y)$	`(mod x y)`	⋆
$\mathit{group}_I(x, y)$	`(group x y)`	†
$\mathit{pad}_I(x, y)$	`(pad x y)`	†
$\mathit{quot}_I(x, y)$	`(quot x y)`	†
$\mathit{remr}_I(x, y)$	`(remainder x y)`	†
$\mathit{gcd}_I(x, y)$	`(gcd x y)` (deviation: `(gcd 0 0)` is 0)	⋆
$\mathit{lcm}_I(x, y)$	`(lcm x y)`	⋆
$\mathit{gcd\_seq}_I(xs)$	`(gcds xs)`	†
$\mathit{lcm\_seq}_I(xs)$	`(lcms xs)`	†
$\mathit{add\_wrap}_I(x, y)$	`(add_wrap x y)`	†
$\mathit{add\_ov}_I(x, y)$	`(add_over x y)`	†
$\mathit{sub\_wrap}_I(x, y)$	`(sub_wrap x y)`	†
$\mathit{sub\_ov}_I(x, y)$	`(sub_over x y)`	†
$\mathit{mul\_wrap}_I(x, y)$	`(mul_wrap x y)`	†
$\mathit{mul\_ov}_I(x, y)$	`(mul_over x y)`	†

where x and y are expressions of type *INT* and where xs is an expression of type list of *INT*.

The LIA-2 non-transcendental floating point operations are listed below, along with the syntax used to invoke them:

$\mathit{min}_F(x, y)$	`(min x y)`	⋆
$\mathit{max}_F(x, y)$	`(max x y)`	⋆
$\mathit{mmin}_F(x, y)$	`(mmin x y)`	†
$\mathit{mmax}_F(x, y)$	`(mmax x y)`	†
$\mathit{min\_seq}_F(xs)$	`(min . xs)` or `(min x1 x2 ... xn)`	⋆
$\mathit{max\_seq}_F(xs)$	`(max . xs)` or `(max x1 x2 ... xn)`	⋆
$\mathit{mmin\_seq}_F(xs)$	`(mmin . xs)` or `(mmin x1 x2 ... xn)`	†
$\mathit{mmax\_seq}_F(xs)$	`(mmax . xs)` or `(mmax x1 x2 ... xn)`	†
$\mathit{floor}_F(x)$	`(float (floor x))`	⋆

$\mathit{rounding}_F(x)$	(float (round x))	⋆
$\mathit{ceiling}_F(x)$	(float (ceiling x))	⋆

$\mathit{dim}_F(x, y)$	(dim x y)	†
$\mathit{dprod}_{F \to F'}(x, y)$	(prod x y)	†
$\mathit{remr}_F(x, y)$	(remainder x y)	†
$\mathit{sqrt}_F(x)$	(sqrt x)	⋆
$\mathit{rsqrt}_F(x)$	(rsqrt x)	†

$\mathit{add\_lo}_F(x, y)$	(add_low x y)	†
$\mathit{sub\_lo}_F(x, y)$	(sub_low x y)	†
$\mathit{mul\_lo}_F(x, y)$	(mul_low x y)	†
$\mathit{div\_rest}_F(x, y)$	(div_rest x y)	†
$\mathit{sqrt\_rest}_F(x)$	(sqrt_rest x)	†

where x, y and z are data objects of the same floating point type, and where xs is an data objects that are lists of data objects of the same floating point type.

The LIA-2 parameters for operations approximating real valued transcendental functions can be accessed by the following syntax:

$\mathit{max\_err\_hypot}_F$	(err-hypotenuse x)	†

$\mathit{max\_err\_exp}_F$	(err-exp x)	†
$\mathit{max\_err\_power}_F$	(err-power x)	†

$\mathit{max\_err\_sinh}_F$	(err-sinh x)	†
$\mathit{max\_err\_tanh}_F$	(err-tanh x)	†

$\mathit{big\_radian\_angle}_F$	(big-radian-angle x)	†
$\mathit{max\_err\_sin}_F$	(err-sin x)	†
$\mathit{max\_err\_tan}_F$	(err-tan x)	†

$\mathit{min\_angular\_unit}_F$	(minimum-angular-unit x)	†
$\mathit{big\_angle}_F$	(big-angle x)	†
$\mathit{max\_err\_sinu}_F(u)$	(err-sin-cycle u)	†
$\mathit{max\_err\_tanu}_F(u)$	(err-tan-cycle u)	†

$\mathit{max\_err\_convert}_F$	err-convert-to-string	†
$\mathit{max\_err\_convert}_D$	err-convert-to-string	†

where b, x and u are expressions of type FLT. Several of the parameter functions are constant for each type (and library), the argument is then used only to differentiate among the floating point types.

The LIA-2 elementary floating point operations are listed below, along with the syntax used to invoke them:

$\mathit{hypot}_F(x, y)$	(hypotenuse x y)	†

$\mathit{power}_{FI}(b, z)$	(expt b z)	⋆
$\mathit{exp}_F(x)$	(exp x)	⋆
$\mathit{expm1}_F(x)$	(expm1 x)	†
$\mathit{exp2}_F(x)$	(exp2 x)	†
$\mathit{exp10}_F(x)$	(exp10 x)	†
$\mathit{power}_F(b, y)$	(expt b y)	⋆

$\mathit{power1pm1}_F(b, y)$	(expm1 b y)	†
$\mathit{ln}_F(x)$	(log x)	⋆
$\mathit{ln1p}_F(x)$	(log1p x)	†
$\mathit{log2}_F(x)$	(log2 x)	†
$\mathit{log10}_F(x)$	(log10 x)	†
$\mathit{logbase}_F(b, x)$	(logbase b x)	†
$\mathit{logbase1p1p}_F(b, x)$	(logbase1p b x)	†
$\mathit{sinh}_F(x)$	(sinh x)	⋆
$\mathit{cosh}_F(x)$	(cosh x)	⋆
$\mathit{tanh}_F(x)$	(tanh x)	⋆
$\mathit{coth}_F(x)$	(coth x)	†
$\mathit{sech}_F(x)$	(sech x)	†
$\mathit{csch}_F(x)$	(csch x)	†
$\mathit{arcsinh}_F(x)$	(asinh x)	†
$\mathit{arccosh}_F(x)$	(acosh x)	†
$\mathit{arctanh}_F(x)$	(atanh x)	⋆
$\mathit{arccoth}_F(x)$	(acoth x)	†
$\mathit{arcsech}_F(x)$	(asech x)	†
$\mathit{arccsch}_F(x)$	(acsch x)	†
$\mathit{axis\_rad}_F(x)$	(axis_rad x)	†
$\mathit{rad}_F(x)$	(radians x)	†
$\mathit{sin}_F(x)$	(sin x)	⋆
$\mathit{cos}_F(x)$	(cos x)	⋆
$\mathit{tan}_F(x)$	(tan x)	⋆
$\mathit{cot}_F(x)$	(cot x)	†
$\mathit{sec}_F(x)$	(sec x)	†
$\mathit{csc}_F(x)$	(csc x)	†
$\mathit{arcsin}_F(x)$	(asin x)	⋆
$\mathit{arccos}_F(x)$	(acos x)	⋆
$\mathit{arctan}_F(x)$	(atan x)	⋆
$\mathit{arccot}_F(x)$	(acot x)	†
$\mathit{arcctg}_F(x)$	(actg x)	†
$\mathit{arcsec}_F(x)$	(asec x)	†
$\mathit{arccsc}_F(x)$	(acsc x)	†
$\mathit{arc}_F(x, y)$	(atan2 y x)	⋆
$\mathit{axis\_cycle}_F(u, x)$	(axis_cycle u x)	†
$\mathit{cycle}_F(u, x)$	(cycle u x)	†
$\mathit{sinu}_F(u, x)$	(sinU u x)	†
$\mathit{cosu}_F(u, x)$	(cosU u x)	†
$\mathit{tanu}_F(u, x)$	(tanU u x)	†
$\mathit{cotu}_F(u, x)$	(cotU u x)	†
$\mathit{secu}_F(u, x)$	(secU u x)	†
$\mathit{cscu}_F(u, x)$	(cscU u x)	†

$arcsinu_F(u,x)$	`(asinU` u x	†
$arccosu_F(u,x)$	`(acosU` u x	†
$arctanu_F(u,x)$	`(atanU` u x	†
$arccotu_F(u,x)$	`(acotU` u x	†
$arcctgu_F(u,x)$	`(actgU` u x`)`	†
$arcsecu_F(u,x)$	`(asecU` u x`)`	†
$arccscu_F(u,x)$	`(acscU` u x`)`	†
$arcu_F(u,x,y)$	`(atan2U` u y x`)`	†

$rad\_to\_cycle_F(x,u)$	`(rad_to_cycle` x u`)`	†
$cycle\_to\_rad_F(u,x)$	`(cycle_to_rad` u x`)`	†
$cycle\_to\_cycle_F(u,x,v)$	`(cycle_to_cycle` u x v`)`	†

where b, x, y, u, and v are expressions of type *FLT*, and z is an expressions of type *INT*.

Arithmetic value conversions in ISLisp are can be explicit or implicit. The rules for when implicit conversions are done is implementation defined.

$convert_{I\to I''}(x)$	`(format` g `"~B"` x`)`	⋆(binary)
$convert_{I\to I''}(x)$	`(format` g `"~O"` x`)`	⋆(octal)
$convert_{I\to I''}(x)$	`(format` g `"~D"` x`)`	⋆(decimal)
$convert_{I\to I''}(x)$	`(format` g `"~X"` x`)`	⋆(hexadecimal)
$convert_{I\to I''}(x)$	`(format` g `"~`r`R"` x`)`	⋆(radix r)
$convert_{I\to I''}(x)$	`(format-integer` g x r`)`	⋆(radix r)

$rounding_{F\to I}(y)$	`(round` y`)`	⋆
$floor_{F\to I}(y)$	`(floor` y`)`	⋆
$ceiling_{F\to I}(y)$	`(ceiling` y`)`	⋆

$convert_{I\to F}(x)$	`(float` x $kind$`)`	⋆

$convert_{F\to F'}(y)$	`(float` y $kind$`)`	⋆
$convert_{F\to F''}(y)$	`(format` g `"~G"` y`)`	⋆
$convert_{F\to F''}(y)$	`(format-float` g y`)`	⋆

where x is an expression of type *INT*, y is an expression of type *FLT*, and z is an expression of type *FXD*, where *FXD* is a fixed point type. Conversion from string to numeric value is in ISLisp done via a general read procedure, which reads ISLisp 'S-expressions'.

ISLisp provides non-negative numerals for its integer and floating point types in base is 10.

ISLisp does not specify numerals for infinities and NaNs. Suggestion:

+∞	`infinity`	†
qNaN	`nan`	†
sNaN	`signan`	†

as well as string formats for reading and writing these values as character strings.

ISLisp has a notion of 'error' that implies a catchable, possibly returnable, change of control flow. ISLisp uses its exception mechanism as its default means of notification. ISLisp ignores **underflow** notifications. On **underflow** the continuation value (specified in LIA-2) is used directly without recording the **underflow** itself. ISLisp uses the error `domain-error` for **invalid** and some **pole** notifications, the error `arithmetic-error` for **overflow** notifications, and the error `division-by-zero` for other **pole** notifications.

An implementation that wishes to follow LIA-2 should provide recording of indicators as an alternative means of handling numeric notifications. Recording of indicators is the LIA-2 preferred means of handling numeric notifications.

C.11 Modula-2

The programming language Modula-2 is defined by ISO/IEC 10514-1:1996, *Information technology – Programming languages - Part 1: Modula-2, Base Language* [26]. An implementation should follow all the requirements of LIA-2 unless otherwise specified by this language binding.

The operations or parameters marked "†" are not part of the language and should be provided by an implementation that wishes to conform to the LIA-2 for that operation. For each of the marked items a suggested identifier is provided.

The Modula-2 datatype `Boolean` corresponds to the LIA datatype **Boolean**.

The additional integer operations are listed below, along with the syntax used to invoke them:

$\mathit{min}_I(x,y)$	imin(x, y)	†
$\mathit{max}_I(x,y)$	imax(x, y)	†
$\mathit{min\_seq}_I(xs)$	iminArr(xs)	†
$\mathit{max\_seq}_I(xs)$	imaxArr(xs)	†
$\mathit{dim}_I(x,y)$	idim(x, y)	†
$\mathit{sqrt}_I(x)$	isqrt(x)	†
$\mathit{power}_I(x,y)$	ipower(x, y)	†
$\mathit{divides}_I(x,y)$	divides(x, y)	†
$\mathit{even}_I(x)$	not odd(x)	⋆
$\mathit{odd}_I(x)$	odd(x)	⋆
$\mathit{divf}_I(x,y)$	div(x, y)	†
$\mathit{moda}_I(x,y)$	x mod y	⋆
$\mathit{group}_I(x,y)$	group(x, y)	†
$\mathit{pad}_I(x,y)$	pad(x, y)	†
$\mathit{quot}_I(x,y)$	ratio(x, y)	†
$\mathit{remr}_I(x,y)$	residue(x, y)	†
$\mathit{gcd}_I(x,y)$	gcd(x, y)	†
$\mathit{lcm}_I(x,y)$	lcm(x, y)	†
$\mathit{gcd\_seq}_I(xs)$	gcdarr(xs)	†
$\mathit{lcm\_seq}_I(xs)$	lcmarr(xs)	†
$\mathit{add\_wrap}_I(x,y)$	addwrap(x, y)	†
$\mathit{add\_ov}_I(x,y)$	addover(x, y)	†
$\mathit{sub\_wrap}_I(x,y)$	subwrap(x, y)	†
$\mathit{sub\_ov}_I(x,y)$	subover(x, y)	†
$\mathit{mul\_wrap}_I(x,y)$	mulwrap(x, y)	†
$\mathit{mul\_ov}_I(x,y)$	mulover(x, y)	†

where x and y are expressions of type *INT* and where xs is an expression of type `array [] of` *INT*.

The additional non-transcendental floating point operations are listed below, along with the syntax used to invoke them:

$\mathit{min}_F(x,y)$	min(x, y)	†
$\mathit{max}_F(x,y)$	max(x, y)	†

$mmin_F(x, y)$	`mmin(`x`, `y`)`	†
$mmax_F(x, y)$	`mmax(`x`, `y`)`	†
$min\_seq_F(xs)$	`minarr(`xs`)`	†
$max\_seq_F(xs)$	`maxarr(`xs`)`	†
$mmin\_seq_F(xs)$	`mminarr(`xs`)`	†
$mmax\_seq_F(xs)$	`mmaxarr(`xs`)`	†
$dim_F(x, y)$	`dim(`x`, `y`)`	†
$rounding_F(x)$	`rounding(`x`)`	†
$floor_F(x)$	`floor(`x`)`	†
$ceiling_F(x)$	`ceiling(`x`)`	†
$rounding\_rest_F(x)$	x `- rounding(`x`)`	†
$floor\_rest_F(x)$	x `- floor(`x`)`	†
$ceiling\_rest_F(x)$	x `- ceiling(`x`)`	†
$dprod_{F \to F'}(x, y)$	`prod(`x`, `y`)`	†
$remr_F(x, y)$	`remainder(`x`, `y`)`	†
$sqrt_F(x)$	`sqrt(`x`)`	⋆
$rsqrt_F(x)$	`rsqrt(`x`)`	†
$add\_lo_F(x, y)$	`addlow(`x`, `y`)`	†
$sub\_lo_F(x, y)$	`sublow(`x`, `y`)`	†
$mul\_lo_F(x, y)$	`mullow(`x`, `y`)`	†
$div\_rest_F(x, y)$	`divrest(`x`, `y`)`	†
$sqrt\_rest_F(x)$	`sqrtrest(`x`)`	†

where x, y and z are expressions of type *FLT*, and where xs is an expression of type `array [] of` *FLT*.

The LIA-2 parameters for operations approximating real valued transcendental functions can be accessed by the following syntax:

$max\_err\_hypot_F$	`err_hypotenuse(`x`)`	†
$max\_err\_exp_F$	`err_exp(`x`)`	†
$max\_err\_power_F$	`err_power(`x`)`	†
$max\_err\_sinh_F$	`err_sinh(`x`)`	†
$max\_err\_tanh_F$	`err_tanh(`x`)`	†
$big\_radian\_angle_F$	`big_radian_angle(`x`)`	†
$max\_err\_sin_F$	`err_sin(`x`)`	†
$max\_err\_tan_F$	`err_tan(`x`)`	†
$min\_angular\_unit_F$	`min_angle_unit(`x`)`	†
$big\_angle\_u_F$	`big_angle(`x`)`	†
$max\_err\_sinu_F(u)$	`err_sin_cycle(`u`)`	†
$max\_err\_tanu_F(u)$	`err_tan_cycle(`u`)`	†
$max\_err\_convert_F$	`err_convert(`x`)`	†
$max\_err\_convert_F$	`err_convert_to_string`	†
$max\_err\_convert_D$	`err_convert_to_string`	†

where x and u are expressions of type *FLT*. Several of the parameter functions are constant for each type (and library), the argument is then used only to differentiate among the floating point types.

The LIA-2 elementary floating point operations are listed below, along with the syntax used to invoke them:

$\mathit{hypot}_F(x, y)$	`hypotenuse(x, y)`	†
$\mathit{power}_{FI}(b, z)$	`powerI(b, z)`	†
$\mathit{exp}_F(x)$	`exp(x)`	⋆
$\mathit{expm1}_F(x)$	`expm1(x)`	†
$\mathit{exp2}_F(x)$	`exp2(x)`	†
$\mathit{exp10}_F(x)$	`exp10(x)`	†
$\mathit{power}_F(b, y)$	`power(b, y)`	⋆
$\mathit{power1pm1}_F(b, y)$	`power1PM1(b, y)`	†
$\mathit{ln}_F(x)$	`ln(x)`	⋆
$\mathit{ln1p}_F(x)$	`ln1P(x)`	†
$\mathit{log2}_F(x)$	`log2(x)`	†
$\mathit{log10}_F(x)$	`log10(x)`	†
$\mathit{logbase}_F(b, x)$	`log(x, b)`	†
$\mathit{logbase1p1p}_F(b, x)$	`log1P1P(x, b)`	†
$\mathit{sinh}_F(x)$	`sinh(x)`	†
$\mathit{cosh}_F(x)$	`cosh(x)`	†
$\mathit{tanh}_F(x)$	`tanh(x)`	†
$\mathit{coth}_F(x)$	`coth(x)`	†
$\mathit{sech}_F(x)$	`sech(x)`	†
$\mathit{csch}_F(x)$	`csch(x)`	†
$\mathit{arcsinh}_F(x)$	`arcsinh(x)`	†
$\mathit{arccosh}_F(x)$	`arccosh(x)`	†
$\mathit{arctanh}_F(x)$	`arctanh(x)`	†
$\mathit{arccoth}_F(x)$	`arccoth(x)`	†
$\mathit{arcsech}_F(x)$	`arcsech(x)`	†
$\mathit{arccsch}_F(x)$	`arccsch(x)`	†
$\mathit{rad}_F(x)$	`radian(x)`	†
$\mathit{axis\_rad}_F(x)$	`axis_rad(x)`	†
$\mathit{sin}_F(x)$	`sin(x)`	⋆
$\mathit{cos}_F(x)$	`cos(x)`	⋆
$\mathit{tan}_F(x)$	`tan(x)`	⋆
$\mathit{cot}_F(x)$	`cot(x)`	†
$\mathit{sec}_F(x)$	`sec(x)`	†
$\mathit{csc}_F(x)$	`csc(x)`	†
$\mathit{arcsin}_F(x)$	`arcsin(x)`	⋆
$\mathit{arccos}_F(x)$	`arccos(x)`	⋆
$\mathit{arctan}_F(x)$	`arctan(x)`	⋆
$\mathit{arccot}_F(x)$	`arccot(x)`	†
$\mathit{arcctg}_F(x)$	`arcctg(x)`	†
$\mathit{arcsec}_F(x)$	`arcsec(x)`	†
$\mathit{arccsc}_F(x)$	`arccsc(x)`	†

$arc_F(x, y)$	`angle(x, y)`	†
$cycle_F(u, x)$	`cycle(u, x)`	†
$axis\_cycle_F(u, x)$	`axis_cycle(u, x)`	†
$sinu_F(u, x)$	`sinu(u, x)`	†
$cosu_F(u, x)$	`cosu(u, x)`	†
$tanu_F(u, x)$	`tanu(u, x)`	†
$cotu_F(u, x)$	`cotu(u, x)`	†
$secu_F(u, x)$	`secu(u, x)`	†
$cscu_F(u, x)$	`cscu(u, x)`	†
$arcsinu_F(u, x)$	`arcsinu(u, x)`	†
$arccosu_F(u, x)$	`arccosu(u, x)`	†
$arctanu_F(u, x)$	`arctanu(u, x)`	†
$arccotu_F(u, x)$	`arccotu(u, x)`	†
$arcctgu_F(u, x)$	`arcctgu(u, x)`	†
$arcsecu_F(u, x)$	`arcsecu(u, x)`	†
$arccscu_F(u, x)$	`arccscu(u, x)`	†
$arcu_F(u, x, y)$	`angleu(u, x, y)`	†
$rad\_to\_cycle_F(x, u)$	`Radian_to_cycle(x, u)`	†
$cycle\_to\_rad_F(u, x)$	`Cycle_to_radian(u, x)`	†
$cycle\_to\_cycle_F(u, x, v)$	`Cycle_to_cycle(u, x, v)`	†

where b, x, y, u, and v are expressions of type *FLT*, and z is an expressions of type *INT*

Arithmetic value conversions in C are can be explicit or implicit. The rules for when implicit conversions are applied is not repeated here. The explicit arithmetic value conversions are usually expressed as 'casts', except when converting to/from strings.

$convert_{I\to I'}(x)$	*INT*(x)	⋆
$convert_{I''\to I'}(f)$	`ReadCard(f,r)`	⋆
$convert_{I''\to I}(f)$	`ReadInt(f,r)`	⋆
$convert_{I'\to I''}(x)$	`WriteCard(h,x)`	⋆
$convert_{I\to I''}(x)$	`WriteInt(h,x)`	⋆
$rounding_{F\to I}(y)$	`round(y)`	⋆
$floor_{F\to I}(y)$	`floor(y)`	⋆
$ceiling_{F\to I}(y)$	`ceiling(y)`	⋆
$convert_{I\to F}(x)$	*FLT*(x)	⋆
$convert_{F\to F'}(y)$	*FLT2*(y)	⋆
$convert_{F''\to F}(f)$	`ReadReal(f,z)`	⋆
$convert_{F\to F''}(y)$	`WriteFloat(f,y,a,w)`	⋆
$convert_{F\to F''}(y)$	`WriteEng(h,y,a,w)`	⋆
$convert_{F\to F''}(y)$	`WriteReal(h,y,a,w)`	⋆
$convert_{D'\to F}(f)$	`ReadReal(f,z)`	⋆
$convert_{F\to D'}(y)$	`WriteFixed(h,y,a,w)`	⋆

where x is an expression of type *INT*, y is an expression of type *FLT*, and z is an expression of type *FXD*, where *FXD* is a fixed point type. *INT2* is the integer datatype that corresponds to I'. A ? above indicates that the parameter is optional. e is greater than 0.

Modula-2 provides base 8, 10, and 16 non-negative numerals for all its integer types, and base 10 non-negative numerals for all its floating point types. Numerals for floating point types must have a '.' in them. The details are not repeated in this example binding, see ISO/IEC 10514-1, clause 6.8.7.1 Whole Number Literals, and clause 6.8.7.2 Real Literals.

Modula-2 does not specify numerals for infinities and NaNs. Suggestion:

$+\infty$	`INFINITY`	†
qNaN	`NAN`	†
sNaN	`SIGNAN`	†

as well as string formats for reading and writing these values as character strings.

Modula-2 has a notion of 'exception' that implies a non-returnable, but catchable, change of control flow. Modula-2 uses its exception mechanism as its default means of notification. Modula-2 ignores **underflow** notifications since an Modula-2 exception is inappropriate for an **underflow** notification. On **underflow** the continuation value (specified in LIA-2) is used directly without recording the **underflow** itself. Modula-2 uses the exceptions `WHOLE-ZERO-DIVISION`, `WHOLE-ZERO-REMAINDER`, `NEGATIVE-SQRT-ARG`, `NONPOSITIVE-LN-ARG`, `NONPOSITIVE-POWER-ARG`, `TAN-OVERFLOW` (for pole, not overflow?), `ARCSIN-ARG-MAGNITUDE`, and `ARCCOS-ARG-MAGNITUDE` for **pole** and **invalid** notifications. The exceptions `WHOLE-OVERFLOW` and `REAL-OVERFLOW` are used for **overflow** notifications. Since Modula-2 exceptions are non-returnable changes of control flow, no continuation value is provided for these notifications.

An implementation that wishes to follow LIA-2 should provide recording of indicators as an alternative means of handling numeric notifications. Recording of indicators is the LIA-2 preferred means of handling numeric notifications.

C.12 Pascal and Extended Pascal

The programming language Pascal is defined by ISO/IEC 7185:1990, *Information technology - Programming languages – Pascal* [28]. The programming language Extended Pascal is defined in ISO/IEC 10206:1991 *Information technology – Programming languages – Extended Pascal* [29].

An implementation should follow all the requirements of LIA-2 unless otherwise specified by this language binding.

The operations or parameters marked "†" are not part of the language and should be provided by an implementation that wishes to conform to the LIA-2 for that operation. For each of the marked items a suggested identifier is provided.

The Pascal datatype `Boolean` corresponds to the LIA datatype **Boolean**.

The additional integer operations are listed below, along with the syntax used to invoke them:

$min_I(x, y)$	`Imin(`x`, `y`)`	†
$max_I(x, y)$	`Imax(`x`, `y`)`	†
$min\_seq_I(xs)$	`IminArr(`xs`)`	†
$max\_seq_I(xs)$	`ImaxArr(`xs`)`	†
$dim_I(x, y)$	`Idim(`x`, `y`)`	†
$sqrt_I(x)$	`Isqrt(`x`)`	†
$power_I(x, y)$	x `pow` y	⋆(Extended Pascal)
$divides_I(x, y)$	`Divides(`x`, `y`)`	†
$even_I(x)$	`(not Odd(`x`))`	⋆

$odd_I(x)$	`Odd(x)`	⋆
$divf_I(x, y)$	`Divi(x, y)`	†
$moda_I(x, y)$	`Modulo(x, y)`	†
$group_I(x, y)$	`Group(x, y)`	†
$pad_I(x, y)$	`Pad(x, y)`	†
$quot_I(x, y)$	`Ratio(x, y)`	†
$remr_I(x, y)$	`Residue(x, y)`	†
$gcd_I(x, y)$	`Gcd(x, y)`	†
$lcm_I(x, y)$	`Lcm(x, y)`	†
$gcd\_seq_I(xs)$	`GcdArr(xs)`	†
$lcm\_seq_I(xs)$	`LcmArr(xs)`	†
$add\_wrap_I(x, y)$	`AddWrap(x, y)`	†
$add\_ov_I(x, y)$	`AddOver(x, y)`	†
$sub\_wrap_I(x, y)$	`SubWrap(x, y)`	†
$sub\_ov_I(x, y)$	`SubOver(x, y)`	†
$mul\_wrap_I(x, y)$	`MulWrap(x, y)`	†
$mul\_ov_I(x, y)$	`MulOver(x, y)`	†

where x and y are expressions of type *INT* and where xs is an expression of type `array of` *INT*.

The LIA-2 non-transcendental floating point operations are listed below, along with the syntax used to invoke them:

$min_F(x, y)$	`Min(x, y)`	†
$max_F(x, y)$	`Max(x, y)`	†
$mmin_F(x, y)$	`MMin(x, y)`	†
$mmax_F(x, y)$	`MMax(x, y)`	†
$min\_seq_F(xs)$	`MinArr(xs)`	†
$max\_seq_F(xs)$	`MaxArr(xs)`	†
$mmin\_seq_F(xs)$	`MMinarr(xs)`	†
$mmax\_seq_F(xs)$	`MMaxarr(xs)`	†
$rounding_F(x)$	`Rounding(x)`	†
$floor_F(x)$	`Floor(x)`	†
$ceiling_F(x)$	`Ceiling(x)`	†
$dim_F(x, y)$	`Dim(x, y)`	†
$dprod_{F \rightarrow F'}(x, y)$	`Prod(x, y)`	†
$remr_F(x, y)$	`Remainder(x, y)`	†
$sqrt_F(x)$	`Sqrt(x)`	⋆
$rsqrt_F(x)$	`Rsqrt(x)`	†
$add\_lo_F(x, y)$	`AddLow(x, y)`	†
$sub\_lo_F(x, y)$	`SubLow(x, y)`	†
$mul\_lo_F(x, y)$	`MulLow(x, y)`	†
$div\_rest_F(x, y)$	`DivRest(x, y)`	†
$sqrt\_rest_F(x)$	`SqrtRest(x)`	†

where x, y and z are expressions of type *FLT*, and where xs is an expression of type `array of` *FLT*.

The LIA-2 parameters for operations approximating real valued transcendental functions can be accessed by the following syntax:

$max\_err\_hypot_F$	`Err_hypotenuse(x)`	†

$max\_err\_exp_F$	`Err_exp(`x`)`	†
$max\_err\_power_F$	`Err_power(`x`)`	†
$max\_err\_sinh_F$	`Err_sinh(`x`)`	†
$max\_err\_tanh_F$	`Err_tanh(`x`)`	†
$big\_radian\_angle_F$	`Big_radian_angle(`x`)`	†
$max\_err\_sin_F$	`Err_sin(`x`)`	†
$max\_err\_tan_F$	`Err_tan(`x`)`	†
$min\_angular\_unit_F$	`Min_angle_unit(`x`)`	†
big_angle_F	`Big_angle(`x`)`	†
$max\_err\_sinu_F(u)$	`Err_sin_cycle(`u`)`	†
$max\_err\_tanu_F(u)$	`Err_tan_cycle(`u`)`	†
$max\_err\_convert_F$	`Err_convert(`x`)`	†
$max\_err\_convert_{F'}$	`Err_convert_to_string`	†
$max\_err\_convert_{D'}$	`Err_convert_to_string`	†

where x and u are expressions of type *FLT*. Several of the parameter functions are constant for each type (and library), the argument is then used only to differentiate among the floating point types.

The LIA-2 elementary floating point operations are listed below, along with the syntax used to invoke them:

$hypot_F(x, y)$	`Hypotenuse(`x`, `y`)`	†
$power_{FI}(b, z)$	b `pow` z	⋆(Extended Pascal)
$exp_F(x)$	`Exp(`x`)`	⋆
$expm1_F(x)$	`ExpM1(`x`)`	†
$exp2_F(x)$	`Exp2(`x`)`	†
$exp10_F(x)$	`Exp10(`x`)`	†
$power_F(b, y)$	b `**` y	⋆(Extended Pascal)
$power1pm1_F(b, y)$	`Power1PM1(`b`, `y`)`	†
$ln_F(x)$	`Ln(`x`)`	⋆
$ln1p_F(x)$	`Ln1P(`x`)`	†
$log2_F(x)$	`Log2(`x`)`	†
$log10_F(x)$	`Log10(`x`)`	†
$logbase_F(b, x)$	`Log(`x`, `b`)`	†
$logbase1p1p_F(b, x)$	`Log1P1P(`x`, `b`)`	†
$sinh_F(x)$	`Sinh(`x`)`	†
$cosh_F(x)$	`Cosh(`x`)`	†
$tanh_F(x)$	`Tanh(`x`)`	†
$coth_F(x)$	`Coth(`x`)`	†
$sech_F(x)$	`Sech(`x`)`	†
$csch_F(x)$	`Csch(`x`)`	†
$arcsinh_F(x)$	`Arcsinh(`x`)`	†
$arccosh_F(x)$	`Arccosh(`x`)`	†

$arctanh_F(x)$	Arctanh(x)	†
$arccoth_F(x)$	Arccoth(x)	†
$arcsech_F(x)$	Arcsech(x)	†
$arccsch_F(x)$	Arccsch(x)	†
$rad_F(x)$	Radian(x)	†
$axis\_rad_F(x)$	Axis_Radian(x, h, v)	†
$sin_F(x)$	Sin(x)	⋆
$cos_F(x)$	Cos(x)	⋆
$tan_F(x)$	Tan(x)	†
$cot_F(x)$	Cot(x)	†
$sec_F(x)$	Sec(x)	†
$csc_F(x)$	Csc(x)	†
$arcsin_F(x)$	Arcsin(x)	†
$arccos_F(x)$	Arccos(x)	†
$arctan_F(x)$	Arctan(x)	⋆
$arccot_F(x)$	Arccot(x)	†
$arcctg_F(x)$	Arccot(x)	†
$arcsec_F(x)$	Arcsec(x)	†
$arccsc_F(x)$	Arccsc(x)	†
$arc_F(x, y)$	Angle(x, y)	†
$cycle_F(u, x)$	Cycle(u, x)	†
$axis\_cycle_F(u, x)$	Axis_Cycle(u, x, h, v)	†
$sinu_F(u, x)$	SinU(u, x)	†
$cosu_F(u, x)$	CosU(u, x)	†
$tanu_F(u, x)$	TanU(u, x)	†
$cotu_F(u, x)$	CotU(u, x)	†
$secu_F(u, x)$	SecU(u, x)	†
$cscu_F(u, x)$	CscU(u, x)	†
$arcsinu_F(u, x)$	ArcsinU(u, x)	†
$arccosu_F(u, x)$	ArccosU(u, x)	†
$arctanu_F(u, x)$	ArctanU(u, x)	†
$arccotu_F(u, x)$	ArccotU(u, x)	†
$arcctgu_F(u, x)$	ArccotU(u, x)	†
$arcsecu_F(u, x)$	ArcsecU(u, x)	†
$arccscu_F(u, x)$	ArccscU(u, x)	†
$arcu_F(u, x, y)$	AngleU(u, x, y)	†
$rad\_to\_cycle_F(x, u)$	RadianToCycle(x, u)	†
$cycle\_to\_rad_F(u, x)$	CycleToRadian(u, x)	†
$cycle\_to\_cycle_F(u, x, v)$	CycleToCycle(u, x, v)	†

where b, x, y, u, and v are expressions of type FLT, and z is an expressions of type INT

Arithmetic value conversions in C are can be explicit or implicit. The rules for when implicit conversions are applied is not repeated here. The explicit arithmetic value conversions are usually expressed as ‘casts’, except when converting to/from strings.

$convert_{I''\rightarrow I}(f)$	read(f?, r)	⋆
$convert_{I\rightarrow I''}(x)$	write(h?, x:n?)	⋆
$rounding_{F\rightarrow I}(y)$	round(y)	⋆
$floor_{F\rightarrow I}(y)$	floor(y)	†
$ceiling_{F\rightarrow I}(y)$	ceiling(y)	†
$convert_{F''\rightarrow F}(f)$	read(f?,m)	⋆
$convert_{F\rightarrow F''}(y)$	write(h?,y:i)	⋆
$convert_{D'\rightarrow F}(f)$	read(f?,m)	⋆
$convert_{F\rightarrow D'}(y)$	write(h?,y:i:a)	⋆

where x is an expression of type *INT*, y is an expression of type *FLT*, and z is an expression of type *FXD*, where *FXD* is a fixed point type. *INT2* is the integer datatype that corresponds to I'. A ? above indicates that the parameter is optional. e is greater than 0.

Pascal provides base 10 non-negative numerals for its only integer type and only floating point type. Numerals for floating point types must have a '.' in them. The details are not repeated in this example binding, see ISO/IEC FDIS 9899, clause xxxxxx, and clause yyyyy.

Pascal does not specify numerals for infinities and NaNs. Suggestion:

$+\infty$	INFINITY	†
qNaN	NAN	†
sNaN	SIGNAN	†

as well as string formats for reading and writing these values as character strings.

Pascal has the notion of 'error', which results in a change of 'control flow', which cannot be returned from, nor caught. An 'error' results in the termination of the program. **pole** for integer types and **invalid** (in general) are considered to be `error`. No notification results for **underflow**, and the continuation value (specified by LIA-2) is used directly, since recording of indicators is not available and 'error' is inappropriate for **underflow**. The handling of integer **overflow** is implementation dependent. The handling of floating point **overflow** and **pole** should be to return a suitable infinity (specified by LIA-2), if possible, without any notification, since recording of indicators is not available.

An implementation that wishes to follow LIA-2 should provide recording of indicators as an alternative means of handling numeric notifications. Recording of indicators is the LIA-2 preferred means of handling numeric notifications.

C.13 PL/I

The programming language PL/I is defined by ANSI X3.53-l976 (R1998), *Programming languages – PL/I* [44], and endorsed by ISO 6160:1979, *Programming languages – PL/I* [30]. The programming language General Purpose PL/I is defined by ISO/IEC 6522:1992, *Information technology – Programming languages – PL/I general-purpose subset* [31], also: ANSI X3.74-1987 (R1998).

An implementation should follow all the requirements of LIA-2 unless otherwise specified by this language binding.

The operations or parameters marked "†" are not part of the language and should be provided by an implementation that wishes to conform to the LIA-2 for that operation. For each of the marked items a suggested identifier is provided.

The LIA datatype **Boolean** is implemented in the PL/I datatype `BIT(1)` (1 = **true** and 0 = **false**).

An implementation of PL/I provides at least one integer data type, and at least one floating point data type. The attribute `FIXED(`n`,0)` selects a signed integer datatype with at least n (decimal or binary) digits of storage. The attribute `FLOAT(`k`)` selects a floating point datatype with at least n (decimal or binary) digits of precision.

The LIA-2 integer operations are listed below, along with the syntax used to invoke them:

$min_I(x, y)$	min(x, y)	⋆
$max_I(x, y)$	max(x, y)	⋆
$min\_seq_I(xs)$	min(xs[1], xs[2], ..., xs[n])	⋆
$max\_seq_I(xs)$	max(xs[1], xs[2], ..., xs[n])	⋆
$dim_I(x, y)$	dim(x, y)	†
$sqrt_I(x)$	sqrt(x)	†
$power_I(x, y)$	x ** y	⋆
$divides_I(x, y)$	divides(x, y)	†
$even_I(x)$	mod(x) = 0	⋆
$odd_I(x)$	mod(x) ¬= 0	⋆
$divf_I(x, y)$	divi(x, y)	†
$moda_I(x, y)$	mod(x, y)	⋆
$group_I(x, y)$	group(x, y)	†
$pad_I(x, y)$	pad(x, y)	†
$quot_I(x, y)$	ratio(x, y)	†
$remr_I(x, y)$	residue(x, y)	†
$gcd_I(x, y)$	gcd(x, y)	†
$lcm_I(x, y)$	lcm(x, y)	†
$gcd\_seq_I(xs)$	gcd(xs)	†
$lcm\_seq_I(xs)$	lcm(xs)	†
$add\_wrap_I(x, y)$	add_wrap(x, y)	†
$add\_ov_I(x, y)$	add_over(x, y)	†
$sub\_wrap_I(x, y)$	sub_wrap(x, y)	†
$sub\_ov_I(x, y)$	sub_over(x, y)	†
$mul\_wrap_I(x, y)$	mul_wrap(x, y)	†
$mul\_ov_I(x, y)$	mul_over(x, y)	†

where x and y are expressions of type *INT* and where xs is an expression of type `array of` *INT*.

The LIA-2 non-transcendental floating point operations are listed below, along with the syntax used to invoke them:

$min_F(x, y)$	min(x, y)	⋆
$max_F(x, y)$	max(x, y)	⋆
$min\_seq_F(xs)$	min(xs[1], xs[2], ..., xs[n])	⋆
$max\_seq_F(xs)$	max(xs[1], xs[2], ..., xs[n])	⋆
$rounding_F(x)$	round(x)	⋆
$floor_F(x)$	floor(x)	†
$ceiling_F(x)$	ceil(x)	†
$dim_F(x, y)$	dim(x, y)	†
$dprod_{F \to F'}(x, y)$	prod(x, y)	†

$remr_F(x, y)$	remainder(x, y)	†
$sqrt_F(x)$	sqrt(x)	⋆
$rsqrt_F(x)$	rsqrt(x)	†
$add\_lo_F(x, y)$	add_low(x, y)	†
$sub\_lo_F(x, y)$	aub_low(x, y)	†
$mul\_lo_F(x, y)$	mul_low(x, y)	†
$div\_rest_F(x, y)$	div_rest(x, y)	†
$sqrt\_rest_F(x)$	sqrt_rest(x)	†

where x, y and z are expressions of type *FLT*, and where xs is an expression of type `array of` *FLT*.

The parameters for operations approximating real valued transcendental functions can be accessed by the following syntax:

$max\_err\_hypot_F$	err_hypotenuse(x)	†
$max\_err\_exp_F$	err_exp(x)	†
$max\_err\_power_F$	err_power(x)	†
$max\_err\_sinh_F$	err_sinh(x)	†
$max\_err\_tanh_F$	err_tanh(x)	†
$big\_radian\_angle_F$	big_radian_angle(x)	†
$max\_err\_sin_F$	err_sin(x)	†
$max\_err\_tan_F$	err_tan(x)	†
$min\_angular\_unit_F$	min_angle_unit(x)	†
big_angle_F	big_angle(x)	†
$max\_err\_sinu_F(u)$	err_sin_cycle(u)	†
$max\_err\_tanu_F(u)$	err_tan_cycle(u)	†
$max\_err\_convert_F$	err_convert_to_string	†
$max\_err\_convert_D$	err_convert_to_string	†

where x and u are expressions of type *FLT*. Several of the parameter functions are constant for each type (and library), the argument is then used only to differentiate among the floating point types.

The LIA-2 elementary floating point operations are listed below, along with the syntax used to invoke them:

$hypot_F(x, y)$	hypotenuse(x, y)	†
$power_{FI}(b, z)$	poweri(b, z)	†
$exp_F(x)$	exp(x)	⋆
$expm1_F(x)$	expm1(x)	†
$exp2_F(x)$	exp2(x)	†
$exp10_F(x)$	exp10(x)	†
$power_F(b, y)$	power(b, y)	†
$power1pm1_F(b, y)$	power1pm1(b, y)	†
$ln_F(x)$	log(x)	⋆
$ln1p_F(x)$	log1p(x)	†
$log2_F(x)$	log2(x)	⋆

$\mathit{log10}_F(x)$	`log10(`x`)`	⋆
$\mathit{logbase}_F(b, x)$	`log(`b`, `x`)`	†
$\mathit{logbase1p1p}_F(b, x)$	`log1p1p(`b`, `x`)`	†
$\mathit{sinh}_F(x)$	`sinh(`x`)`	⋆
$\mathit{cosh}_F(x)$	`cosh(`x`)`	⋆
$\mathit{tanh}_F(x)$	`tanh(`x`)`	⋆
$\mathit{coth}_F(x)$	`coth(`x`)`	†
$\mathit{sech}_F(x)$	`sech(`x`)`	†
$\mathit{csch}_F(x)$	`csch(`x`)`	†
$\mathit{arcsinh}_F(x)$	`arcsinh(`x`)`	⋆
$\mathit{arccosh}_F(x)$	`arccosh(`x`)`	⋆
$\mathit{arctanh}_F(x)$	`arctanh(`x`)`	⋆
$\mathit{arccoth}_F(x)$	`arccoth(`x`)`	†
$\mathit{arcsech}_F(x)$	`arcsech(`x`)`	†
$\mathit{arccsch}_F(x)$	`arccsch(`x`)`	†
$\mathit{rad}_F(x)$	`rad(`x`)`	†
$\mathit{axis\_rad}_F(x)$	`axis_rad(`x`)`	†
$\mathit{sin}_F(x)$	`sin(`x`)`	⋆
$\mathit{cos}_F(x)$	`cos(`x`)`	⋆
$\mathit{tan}_F(x)$	`tan(`x`)`	⋆
$\mathit{cot}_F(x)$	`cot(`x`)`	⋆
$\mathit{sec}_F(x)$	`sec(`x`)`	†
$\mathit{csc}_F(x)$	`csc(`x`)`	†
$\mathit{arcsin}_F(x)$	`arcsin(`x`)`	⋆
$\mathit{arccos}_F(x)$	`arccos(`x`)`	⋆
$\mathit{arctan}_F(x)$	`arctan(`x`)`	⋆
$\mathit{arccot}_F(x)$	`arccot(`x`)`	†
$\mathit{arcctg}_F(x)$	`arcctg(`x`)`	†
$\mathit{arcsec}_F(x)$	`arcsec(`x`)`	†
$\mathit{arccsc}_F(x)$	`arccsc(`x`)`	†
$\mathit{arc}_F(x, y)$	`arc(`x`, `y`)`	⋆
$\mathit{cycle}_F(u, x)$	`cycle(`u`,`x`)`	†
$\mathit{axis\_cycle}_F(u, x)$	`axis_cycle(`u`,`x`)`	†
$\mathit{sinu}_F(u, x)$	`sin(`u`,`x`)`	†
$\mathit{cosu}_F(u, x)$	`cos(`u`,`x`)`	†
$\mathit{tanu}_F(u, x)$	`tan(`u`,`x`)`	†
$\mathit{cotu}_F(u, x)$	`cot(`u`,`x`)`	†
$\mathit{secu}_F(u, x)$	`sec(`u`,`x`)`	†
$\mathit{cscu}_F(u, x)$	`csc(`u`,`x`)`	†
$\mathit{arcsinu}_F(u, x)$	`arcsin(`u`,`x`)`	†
$\mathit{arccosu}_F(u, x)$	`arccos(`u`,`x`)`	†
$\mathit{arctanu}_F(u, x)$	`arctan(`u`,`x`)`	⋆
$\mathit{arccotu}_F(u, x)$	`arccot(`u`,`x`)`	†

$\mathit{arcctgu}_F(u, x)$	arcctg(u,x)	⋆
$\mathit{arcsecu}_F(u, x)$	arcsec(u,x)	†
$\mathit{arccscu}_F(u, x)$	arccsc(u,x)	†
$\mathit{arcu}_F(u, x, y)$	arc(u,x,y)	†
$\mathit{sinu}_F(360, x)$	sind(x)	⋆
$\mathit{cosu}_F(360, x)$	cosd(x)	⋆
$\mathit{tanu}_F(360, x)$	tand(x)	⋆
$\mathit{cotu}_F(360, x)$	cotd(x)	⋆
$\mathit{secu}_F(360, x)$	secd(x)	†
$\mathit{cscu}_F(360, x)$	cscd(x)	†
$\mathit{arcsinu}_F(360, x)$	arcsind(x)	⋆
$\mathit{arccosu}_F(360, x)$	arccosd(x)	⋆
$\mathit{arctanu}_F(360, x)$	arctand(x)	⋆
$\mathit{arccotu}_F(360, x)$	arccotd(x)	⋆
$\mathit{arcctgu}_F(360, x)$	arcctgd(x)	⋆
$\mathit{arcsecu}_F(360, x)$	arcsecd(x)	†
$\mathit{arccscu}_F(360, x)$	arccscd(x)	†
$\mathit{arcu}_F(360, x, y)$	arcd(y,x)	†
$\mathit{rad\_to\_cycle}_F(x, u)$	rad_to_cycle(x, u)	†
$\mathit{cycle\_to\_rad}_F(u, x)$	cycle_to_rad(u, x)	†
$\mathit{cycle\_to\_cycle}_F(u, x, v)$	cycle_to_cycle(u, x, v)	†

where b, x, y, u, and v are expressions of type FLT, and z is an expressions of type INT.

Arithmetic value conversions in PL/I are can be explicit or implicit. The rules for when implicit conversions are applied is not repeated here. The explicit arithmetic value conversions are usually expressed as 'casts', except when converting to/from strings.

$\mathit{convert}_{I \rightarrow I'}(x)$	FIXED(x, p)	⋆
$\mathit{convert}_{I'' \rightarrow I}(f)$	GET FILE (f)? EDIT (r) (F(w));	⋆
$\mathit{convert}_{I \rightarrow I''}(x)$	PUT FILE (h)? EDIT (x) (F(w));	⋆
$\mathit{rounding}_{F \rightarrow I}(y)$	FIXED(ROUND(y, 0),p)	⋆
$\mathit{floor}_{F \rightarrow I}(y)$	FIXED(FLOOR(y), p)	⋆
$\mathit{ceiling}_{F \rightarrow I}(y)$	FIXED(CEIL(y), p)	⋆
$\mathit{convert}_{I \rightarrow F}(x)$	FLOAT(x, p)	⋆
$\mathit{convert}_{I \rightarrow F}(x)$	DECIMAL(x, p)	⋆
$\mathit{convert}_{I \rightarrow F}(x)$	BINARY(x, p)	⋆
$\mathit{convert}_{F \rightarrow F'}(y)$	FLOAT(y, p)	⋆
$\mathit{convert}_{F \rightarrow F'}(y)$	DECIMAL(y, p)	⋆
$\mathit{convert}_{F \rightarrow F'}(y)$	BINARY(y, p)	⋆
$\mathit{convert}_{F'' \rightarrow F}(f)$	GET FILE (f)? EDIT (t) (E(w,a));	⋆
$\mathit{convert}_{F \rightarrow F''}(y)$	PUT FILE (h)? EDIT (y) (E(w,a));	⋆
$\mathit{convert}_{D' \rightarrow F}(f)$	GET FILE (f)? EDIT (t) (F(w,a));	⋆
$\mathit{convert}_{F \rightarrow D'}(y)$	FIXED(y, p, a))	⋆
$\mathit{convert}_{F \rightarrow D'}(y)$	PUT FILE (h)? EDIT (y) (F(w,a));	⋆

where x is an expression of type *INT*, y is an expression of type *FLT*, and z is an expression of type *FXD*, where *FXD* is a fixed point type. *INT2* is the integer datatype that corresponds to I'. A ? above indicates that the parameter is optional. a is greater than 0.

PL/I provides base 10 non-negative numerals for all its integer and floating point types.

PL/I does not specify numerals for infinities and NaNs. Suggestion:

+∞	`INFINITY`	†
qNaN	`NAN`	†
sNaN	`SIGNAN`	†

as well as string formats for reading and writing these values as character strings.

PL/I has a notion of 'condition' that implies a non-returnable, but catchable (in an ON-unit), change of control flow. PL/I uses its condition mechanism as its default means of notification. PL/I uses the condition `UNDERFLOW` for **underflow** notifications. PL/I uses the condition `ZERODIVIDE` for **pole** notifications, and the conditions `FIXEDOVERFLOW`, `SIZE`, and `OVERFLOW` for **overflow** notifications, and the exception `UNDEFINED` (†) for **invalid** notifications. Since PL/I exceptions are non-returnable changes of control flow, no continuation value is provided for these notifications. This is inappropriate, especially for underflow, so UNDERFLOW notifications are ignored if there is no ON-clause for UNDERFLOW in the program.

An implementation that wishes to follow LIA-2 should provide recording of indicators as an alternative means of handling numeric notifications. Recording of indicators is the LIA-2 preferred means of handling numeric notifications.

C.14 SML

The programming language SML is defined by *The Definition of Standard ML (Revised)* [68].

An implementation should follow all the requirements of LIA-2 unless otherwise specified by this language binding.

The operations or parameters marked "†" are not part of the language and should be provided by an implementation that wishes to conform to the LIA-2 for that operation. For each of the marked items a suggested identifier is provided.

The SML datatype `Boolean` corresponds to the LIA datatype **Boolean**.

Every implementation of SML has at least one integer datatype, `int`, and at least one floating point datatype, `real`. The notation *INT* is used to stand for the name of one of the integer datatypes, and *FLT* is used to stand for the name of one of the floating point datatypes in what follows.

The LIA-2 integer operations are listed below, along with the syntax used to invoke them:

$min_I(x, y)$	x `min` y or `op min (`x`,`y`)`	⋆
$max_I(x, y)$	x `max` y or `op max (`x`,`y`)`	⋆
$min\_seq_I(xs)$	`minimum` xs	†
$max\_seq_I(xs)$	`maximum` xs	†
$dim_I(x, y)$	x `dim` y or `op dim (`x`,`y`)`	†
$sqrt_I(x)$	`isqrt` x	†
$power_I(x, y)$	x `pow` y or `op pow (`x`,`y`)`	†
$divides_I(x, y)$	`divides (`x`,`y`)`	†
$even_I(x)$	`even` x	†
$odd_I(x)$	`odd` x	†
$divf_I(x, y)$	x `div` y or `op div (`x`,`y`)`	⋆

$moda_I(x,y)$	x `mod` y or `op mod (`x`,`y`)`	⋆
$group_I(x,y)$	`group (`x`,`y`)`	†
$pad_I(x,y)$	`pad (`x`,`y`)`	†
$quot_I(x,y)$	`ratio (`x`,`y`)`	†
$remr_I(x,y)$	`residue (`x`,`y`)`	†
$gcd_I(x,y)$	`gcd (`x`,`y`)`	⋆
$lcm_I(x,y)$	`lcm (`x`,`y`)`	⋆
$gcd\_seq_I(xs)$	`gcd_seq` xs	†
$lcm\_seq_I(xs)$	`lcm_seq` xs	†

$add\_wrap_I(x,y)$	x `+:` y	†
$add\_ov_I(x,y)$	x `+:+` y	†
$sub\_wrap_I(x,y)$	x `-:` y	†
$sub\_ov_I(x,y)$	x `-:+` y	†
$mul\_wrap_I(x,y)$	x `*:` y	†
$mul\_ov_I(x,y)$	x `*:+` y	†

where x and y are expressions of type *INT* and where xs is an expression of type *INT* `list`.

The additional non-transcendental floating point operations are listed below, along with the syntax used to invoke them:

$min_F(x,y)$	x `min` y or `op min (`x`,`y`)`	⋆
$max_F(x,y)$	x `max` y or `op max (`x`,`y`)`	⋆
$mmin_F(x,y)$	x `mmin` y or `op mmin (`x`,`y`)`	†
$mmax_F(x,y)$	x `mmax` y or `op mmax (`x`,`y`)`	†
$min\_seq_F(xs)$	`minimum` xs	†
$max\_seq_F(xs)$	`maximum` xs	†
$mmin\_seq_F(xs)$	`mminimum` xs	†
$mmax\_seq_F(xs)$	`mmaximum` xs	†

$rounding_F(x)$	`realRound` x	†
$floor_F(x)$	`realFloor` x	⋆
$ceiling_F(x)$	`realCeil` x	⋆

$dim_F(x,y)$	`dim (`x`,`y`)`	†
$dprod_{F\to F'}(x,y)$	`prod (`x`,`y`)`	†
$remr_F(x,y)$	`remainder (`x`,`y`)`	†
$sqrt_F(x)$	`sqrt` x	⋆
$rsqrt_F(x)$	`rsqrt` x	†

$add\_lo_F(x,y)$	x `+:-` y	†
$sub\_lo_F(x,y)$	x `-:-` y	†
$mul\_lo_F(x,y)$	x `*:-` y	†
$div\_rest_F(x,y)$	x `/:*` y	†
$sqrt\_rest_F(x)$	`sqrt_rest` x	†

where x, y and z are expressions of type *FLT*, and where xs is an expression of type *FLT* `list`.

The binding for the floor, round, and ceiling operations here take advantage of the unlimited `Integer` type in SML, and that `Integer` is the default integer type.

The parameters for operations approximating real valued transcendental functions can be accessed by the following syntax:

$max\_err\_hypot_F$	`err_hypotenuse` x	†

$max\_err\_exp_F$	`err_exp` x	†
$max\_err\_power_F$	`err_power` x	†
$max\_err\_sinh_F$	`err_sinh` x	†
$max\_err\_tanh_F$	`err_tanh` x	†
$big\_angle\_r_F$	`big_radian_angle` x	†
$max\_err\_sin_F$	`err_sin` x	†
$max\_err\_tan_F$	`err_tan` x	†
$min\_angular\_unit_F$	`min_angular_unit` x	†
$big\_angle\_u_F$	`big_angle` x	†
$max\_err\_sinu_F(u)$	`err_sin_cycle` u	†
$max\_err\_tanu_F(u)$	`err_tan_cycle` u	†
$max\_err\_convert_F$	`err_convert(`x`)`	†
$max\_err\_convert_{F'}$	`err_convert_to_string`	†
$max\_err\_convert_{D'}$	`err_convert_to_string`	†

where x and u are expressions of type *FLT*. Several of the parameter functions are constant for each type (and library), the argument is then used only to differentiate among the floating point types.

The LIA-2 elementary floating point operations are listed below, along with the syntax used to invoke them:

$hypot_F(x, y)$	`hypotenuse (`x`,`y`)`	†
$power_{FI}(b, z)$	b `^^` z or `op ^^ (`x`,`y`)`	†
$exp_F(x)$	`exp` x	⋆
$expm1_F(x)$	`expM1` x	†
$exp2_F(x)$	`exp2` x	†
$exp10_F(x)$	`exp10` x	†
$power_F(b, y)$	b `**` y	†
$pow_F(b, y)$	b `pow` y or `op pow (`x`,`y`)`	⋆ Not LIA-2! (See C.)
$power1pm1_F(b, y)$	`power1PM1 (`b`,`y`)`	†
$ln_F(x)$	`ln` x	⋆
$ln1p_F(x)$	`ln1P` x	†
$log2_F(x)$	`log2` x	†
$log10_F(x)$	`log10` x	⋆
$logbase_F(b, x)$	`log_base (`b`,`x`)`	†
$logbase1p1p_F(b, x)$	`log_base1P1P (`b`,`x`)`	†
$sinh_F(x)$	`sinh` x	⋆
$cosh_F(x)$	`cosh` x	⋆
$tanh_F(x)$	`tanh` x	⋆
$coth_F(x)$	`coth` x	†
$sech_F(x)$	`sech` x	†
$csch_F(x)$	`csch` x	†
$arcsinh_F(x)$	`arcsinh` x	†

$\mathit{arccosh}_F(x)$	`arccosh` x	†
$\mathit{arctanh}_F(x)$	`arctanh` x	†
$\mathit{arccoth}_F(x)$	`arccoth` x	†
$\mathit{arcsech}_F(x)$	`arcsech` x	†
$\mathit{arccsch}_F(x)$	`arccsch` x	†
$\mathit{rad}_F(x)$	`radians` x	†
$\mathit{axis\_rad}_F(x)$	`axis_rad` x	†
$\mathit{sin}_F(x)$	`sin` x	⋆
$\mathit{cos}_F(x)$	`cos` x	⋆
$\mathit{tan}_F(x)$	`tan` x	⋆
$\mathit{cot}_F(x)$	`cot` x	†
$\mathit{sec}_F(x)$	`sec` x	†
$\mathit{csc}_F(x)$	`csc` x	†
$\mathit{arcsin}_F(x)$	`arcsin` x	⋆
$\mathit{arccos}_F(x)$	`arccos` x	⋆
$\mathit{arctan}_F(x)$	`arctan` x	⋆
$\mathit{arccot}_F(x)$	`arccot` x	†
$\mathit{arcctg}_F(x)$	`arcctg` x	†
$\mathit{arcsec}_F(x)$	`arcsec` x	†
$\mathit{arccsc}_F(x)$	`arccsc` x	†
$\mathit{arc}_F(x,y)$	`arctan2` (y,x)	⋆
$\mathit{cycle}_F(u,x)$	`cycle` (u,x)	†
$\mathit{axis\_cycle}_F(u,x)$	`axis_cycle` (u,x)	†
$\mathit{sinu}_F(u,x)$	`sinU` (u,x)	†
$\mathit{cosu}_F(u,x)$	`cosU` (u,x)	†
$\mathit{tanu}_F(u,x)$	`tanU` (u,x)	†
$\mathit{cotu}_F(u,x)$	`cotU` (u,x)	†
$\mathit{secu}_F(u,x)$	`secU` (u,x)	†
$\mathit{cscu}_F(u,x)$	`cscU` (u,x)	†
$\mathit{arcsinu}_F(u,x)$	`arcsinU` (u,x)	†
$\mathit{arccosu}_F(u,x)$	`arccosU` (u,x)	†
$\mathit{arctanu}_F(u,x)$	`arctanU` (u,x)	†
$\mathit{arccotu}_F(u,x)$	`arccotU` (u,x)	†
$\mathit{arcctgu}_F(u,x)$	`arcctgU` (u,x)	†
$\mathit{arcsecu}_F(u,x)$	`arcsecU` (u,x)	†
$\mathit{arccscu}_F(u,x)$	`arccscU` (u,x)	†
$\mathit{arcu}_F(u,x,y)$	`arctan2U` (u,y,x)	†
$\mathit{rad\_to\_cycle}_F(x,u)$	`rad_to_cycle` (x,u)	†
$\mathit{cycle\_to\_rad}_F(u,x)$	`cycle_to_rad` (u,x)	†
$\mathit{cycle\_to\_cycle}_F(u,x,v)$	`cycle_to_cycle` (u,x,v)	†

where b, x, y, u, and v are expressions of type *FLT*, and z is an expressions of type *INT*

Type conversions in SML are always explicit.

$\mathit{convert}_{I \to I'}(x)$	`fromLarge` x or `toLarge` x	⋆

$convert_{I'' \rightarrow I}(s)$	`fromString` s	⋆
$convert_{I'' \rightarrow I}(s)$	`scan` $radix$ $getc$ s	⋆
$convert_{I \rightarrow I''}(x)$	`toString` x	⋆
$rounding_{F \rightarrow I}(y)$	`round` y	⋆
$rounding_{F \rightarrow I}(y)$	`toInt IEEEReal.TO_NEAREST` y	⋆
$rounding_{F \rightarrow I}(y)$	`toLargeInt IEEEReal.TO_NEAREST` y	⋆
$floor_{F \rightarrow I}(y)$	`floor` y	⋆
$floor_{F \rightarrow I}(y)$	`toInt IEEEReal.TO_NEGINF` y	⋆
$floor_{F \rightarrow I}(y)$	`toLargeInt IEEEReal.TO_NEGINF` y	⋆
$ceiling_{F \rightarrow I}(y)$	`ceiling` y	⋆
$ceiling_{F \rightarrow I}(y)$	`toInt IEEEReal.TO_POSINF` y	⋆
$ceiling_{F \rightarrow I}(y)$	`toLargeInt IEEEReal.TO_POSINF` y	⋆
$convert_{I \rightarrow F}(x)$	`fromInt` x	⋆
$convert_{I \rightarrow F}(x)$	`fromLargeInt` x	⋆
$convert_{F \rightarrow F'}(y)$	`toLarge` y	⋆
$convert_{F \rightarrow F'}(y)$	`fromLarge IEEEReal.TO_NEAREST` y	⋆
$convert_{F'' \rightarrow F}(s)$	`fromString` s	⋆
$convert_{F'' \rightarrow F}(s)$	`fromDecimal` s	⋆
$convert_{F'' \rightarrow F}(s)$	`scan` $getc$ s	⋆
$convert_{F \rightarrow F''}(y)$	`fmt (SCI` a`)` y	⋆
$convert_{F \rightarrow F''}(y)$	`toDecimal` y	⋆
$convert_{D' \rightarrow F}(s)$	`fromString` s	⋆
$convert_{D' \rightarrow F}(s)$	`scan` $getc$ s	⋆
$convert_{F \rightarrow D'}(y)$	`fmt (FIX` a`)` y	⋆

where x is an expression of type *INT*, y is an expression of type *FLT*, and z is an expression of type *FXD*, where *FXD* is a fixed point type. *INT2* is the integer datatype that corresponds to I'.

SML provides non-negative base 10 numerals for all its integer and floating point types. There is no differentiation between the numerals for different floating point types, nor between numerals for different integer types, but integer numerals cannot be used for floating point values. The details are not repeated in this example binding, see *The Definition of Standard ML (Revised)* [68].

SML specifies numerals for infinities, but not NaNs:

+∞	`posInf`	⋆
−∞	`negInf`	⋆
qNaN	`NaN`	†
sNaN	`sigNaN`	†

An implementation wishing to conform to LIA-2 should also provide string formats for reading and writing these values as character strings.

SML has a notion of 'exception' that implies a non-returnable, but catchable, change of control flow. SML uses its exception mechanism as its default means of notification. SML ignores **underflow** notifications since an SML exception is inappropriate for an **underflow** notification. On **underflow** the continuation value (specified in LIA-2) is used directly without recording the **underflow** itself. SML uses the exception `Div` for **pole** notifications, the exception `Overflow` for **overflow** notifications, and the exception `Domain` for **invalid** notifications (except for `sin`, `cos`, or `tan` given an infinitary argument, where the **invalid** notification is ignored). Since SML exceptions are non-returnable changes of control flow, no continuation value is provided for these notifications.

An implementation that wishes to follow LIA-2 should provide recording of indicators as an alternative means of handling numeric notifications. Recording of indicators is the LIA-2 preferred means of handling numeric notifications.

Annex D
(informative)

Bibliography

This annex gives references to publications relevant to LIA-2.

EDITOR'S NOTE – The naming of the standards appears unsystematic, but is as taken from ISO's, IEC's and ANSI's websites.

International standards documents

[1] ISO/IEC JTC1 Directives, Part 3: *Drafting and presentation of International Standards*, 1989.

[2] IEC 60559:1989, *Binary floating-point arithmetic for microprocessor systems.* (Also: ANSI/IEEE Standard 754-1985, *IEEE Standard for Binary Floating-Point Arithmetic.*)

[3] ISO/IEC 10967-1:1994, *Information technology – Language independent arithmetic – Part 1: Integer and floating point arithmetic*, (LIA-1).

[4] ISO/IEC FCD4 10967-2, *Information technology – Language independent arithmetic – Part 2: Elementary numerical functions*, 1999, (LIA-2). (This document.)

[5] ISO/IEC 10967-3, *Information technology – Language independent arithmetic – Part 3: Complex floating point arithmetic and complex elementary numerical functions*, (LIA-3). (To be published.)

[6] ISO 6093:1985, *Information processing – Representation of numerical values in character strings for information interchange.*

[7] ISO/IEC TR 10176:1998, *Information technology – Guidelines for the preparation of programming language standards.*

[8] ISO/IEC TR 10182:1993, *Information technology – Programming languages, their environments and system software interfaces – Guidelines for language bindings.*

[9] ISO/IEC 13886:1996, *Information technology – Language-Independent Procedure Calling*, (LIPC).

[10] ISO/IEC 11404:1996, *Information technology – Programming languages, their environments and system software interfaces – Language-independent datatypes*, (LID).

[11] ISO/IEC 8652:1995, *Information technology – Programming languages – Ada.*

[12] ISO/IEC 11430:1994, *Information technology – Programming languages – Generic package of elementary functions for Ada.*

[13] ISO/IEC 13813:1998, *Information technology – Programming languages – Generic packages of real and complex type declarations and basic operations for Ada (including vector and matrix types).*

[14] ISO/IEC 13814:1998, *Information technology – Programming languages – Generic package of complex elementary functions for Ada.*

[15] ISO 8485:1989, *Programming languages – APL.*

[16] ISO/IEC DIS 13751, *Information technology – Programming languages, their environments and system software interfaces – Programming language APL, extended*, 1999.

[17] ISO/IEC 10279:1991, *Information technology – Programming languages – Full BASIC.* (Essentially an endorsement of ANSI X3.113-1987 (R1998) [41].)

[18] ISO/IEC 9899:1990, *Programming languages – C.* Currently under revision: ISO/IEC FDIS 9899, 1999.

[19] ISO/IEC 14882:1998, *Programming languages – C++.*

[20] ISO 1989:1985, *Programming languages – COBOL.* (Endorsement of ANSI X3.23-l985 (R1991) [42].) Currently under revision (1998).

[21] ISO/IEC 16262:1998, *Information technology - ECMAScript language specification.*

[22] ISO/IEC 15145:1997, *Information technology – Programming languages – FORTH.* (Also: ANSI X3.215-1994.)

[23] ISO/IEC 1539-1:1997, *Information technology – Programming languages – Fortran - Part 1: Base language.*

[24] ISO/IEC TR 15580:1998, *Information technology – Programming languages – Fortran – Floating-point exception handling.*

[25] ISO/IEC 13816:1997, *Information technology – Programming languages, their environments and system software interfaces – Programming language ISLISP.*

[26] ISO/IEC 10514-1:1996, *Information technology – Programming languages – Part 1: Modula-2, Base Language.*

[27] ISO/IEC 10514-2:1998, *Information technology – Programming languages – Part 2: Generics Modula-2.*

[28] ISO 7185:1990, *Information technology – Programming languages – Pascal.*

[29] ISO/IEC 10206:1991, *Information technology – Programming languages – Extended Pascal.*

[30] ISO 6160:1979, *Programming languages – PL/I.* (Endorsement of ANSI X3.53-l976 (R1998) [44].)

[31] ISO/IEC 6522:1992, *Information technology – Programming languages – PL/I general-purpose subset.* (Also: ANSI X3.74-1987 (R1998).)

[32] ISO/IEC 13211-1:1995, *Information technology – Programming languages – Prolog – Part 1: General core.*

[33] ISO/IEC 9075:1992, *Information technology – Database languages – SQL.*

[34] ISO/IEC 8824-1:1995, *Information technology – Abstract Syntax Notation One (ASN.1) – Part 1: Specification of basic notation.*

[35] ISO/IEC 9001:1994, *Quality systems – Model for quality assurance in design, development, production, installation and servicing.*

[36] ISO/IEC 9126:1991, *Information technology – Software product evaluation – Quality characteristics and guidelines for their use.*

[37] ISO/IEC 12119:1994, *Information technology – Software packages – Quality requirements and testing.*

[38] ISO/IEC 14598-1:1999, *Information technology – Software product evaluation – Part 1: General overview.*

National standards documents

[39] ANSI/IEEE Standard 754-1985, *IEEE Standard for Binary Floating-Point Arithmetic.*

[40] ANSI/IEEE Standard 854-1987, *IEEE Standard for Radix-Independent Floating-Point Arithmetic.*

[41] ANSI X3.113-1987 (R1998), *Information technology – Programming languages – Full BASIC.*

[42] ANSI X3.23-l985 (R1991), *Programming languages – COBOL.*

[43] ANSI X3.226-1994, *Information Technology – Programming Language – Common Lisp.*

[44] ANSI X3.53-l976 (R1998), *Programming languages – PL/I.*

[45] ANSI/IEEE 1178-1990, *IEEE Standard for the Scheme Programming Language.*

[46] ANSI/NCITS 319-1998, *Information Technology – Programming Languages – Smalltalk.*

Books, articles, and other documents

[47] J. S. Squire (ed), *Ada Letters*, vol. XI, No. 7, ACM Press (1991).

[48] M. Abramowitz and I. Stegun (eds), *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Tenth Printing, 1972, Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402.

[49] J. Du Croz and M. Pont, *The Development of a Floating-Point Validation Package*, *NAG Newsletter*, No. 3, 1984.

[50] J. W. Demmel and X. Li, *Faster Numerical Algorithms via Exception Handling*, 11th International Symposium on Computer Arithmetic, Winsor, Ontario, June 29 - July 2, 1993.

[51] D. Goldberg, *What Every Computer Scientist Should Know about Floating-Point Arithmetic.* ACM Computing Surveys, Vol. 23, No. 1, March 1991.

[52] J. R. Hauser, *Handling Floating-Point Exceptions in Numeric Programs.* ACM Transactions on Programming Languages and Systems, Vol. 18, No. 2, March 1986, Pages 139-174.

[53] W. Kahan, *Branch Cuts for Complex Elementary Functions, or Much Ado about Nothing's Sign Bit*, Chapter 7 in *The State of the Art in Numerical Analysis* ed. by M Powell and A Iserles (1987) Oxford.

[54] W. Kahan, *Lecture Notes on the Status of IEEE Standard 754 for Binary Floating-Point Arithmetic*, Panel Discussion of *Floating-Point Past, Present and Future*, May 23, 1995, in a series of San Francisco Bay Area Computer Historical Perspectives, sponsored by SUN Microsystems Inc.

[55] U. Kulisch and W. L. Miranker, *Computer Arithmetic in Theory and Practice*, Academic Press, 1981.

[56] U. Kulisch and W. L. Miranker (eds), *A New Approach to Scientific Computation*, Academic Press, 1983.

[57] D. C. Sorenson and P. T. P. Tang, *On the Orthogonality of Eigenvectors Computed by Divide-and-Conquer Techniques*, SIAM Journal of Numerical Analysis, Vol. 28, No. 6, p. 1760, algorithm 5.3.

[58] *Floating-Point C Extensions* in Technical Report Numerical C Extensions Committee X3J11, April 1995, SC22/WG14 N403, X3J11/95-004.

[59] David M. Gay, *Correctly Rounded Binary-Decimal and Decimal-Binary Conversions*, AT&T Bell Laboratories, Numerical Analysis Manuscript 90-10, November 1990.

[60] M. Payne and R. Hanek, *Radian Reduction for Trigonometric Functions*, SIGNUM Newsletter, Vol. 18, January 1983.

[61] M. Payne and R. Hanek, *Degree Reduction for Trigonometric Functions*, SIGNUM Newsletter, Vol. 18, April 1983.

[62] N. L. Schryer, *A Test of a Computer's Floating-Point Unit*, Computer Science Technical Report No. 89, AT&T Bell Laboratories, Murray Hill, NJ, 1981.

[63] G. Bohlender, W. Walter, P Kornerup, D. W. Matula, *Semantics for Exact Floating Point Operations*, IEEE Arithmetic 10, 1992.

[64] W. Walter et al., *Proposal for Accurate Floating-Point Vector Arithmetic*, Mathematics and Computers in Simulation, vol. 35, no. 4, pp. 375-382, IMACS, 1993.

[65] James Gosling, Bill Joy, Guy Steele, *The Java Language Specification.*

[66] Simon Peyton Jones et al., *Report on the programming language Haskell 98*, February 1999.

[67] Simon Peyton Jones et al., *Standard libraries for the Haskell 98 programming language*, February 1999.

[68] Robin Milner, Mads Tofte, Robert Harper, and David MacQueen, *The Definition of Standard ML (Revised)*, The MIT Press, 1997, ISBN: 0-262-63181-4.