GRAPHING EQUATIONS WITH GENERALIZED INTERVAL ARITHMETIC

by

Jeffrey Allen Tupper

A thesis submitted in conformity with the requirements
for the degree of Master of Science
Graduate Department of Computer Science
University of Toronto

Graphing Equations with Generalized Interval Arithmetic
Jeffrey Allen Tupper
Master of Science degree, 1996
Graduate Department of Computer Science
University of Toronto

Abstract

Floating-point is commonly used in numerical computations; this use has revealed its inherent inaccuracy and has spread uncertainty throughout the computer community. Interval techniques unite the precision provided by the modern computer with the accuracy accorded to traditional mathematics. As interval analysis matured, sophisticated optimization of interval techniques has occurred.

This thesis presents a framework which enables further optimization of interval routines, while shielding the numerical practitioner from the complexities that have recently surfaced in interval algorithms. This new framework is constructed by migrating variables from the problem domain into the interval arithmetic. Properties of functions within the problem domain may be tracked, so many common non-differentiable partial functions are handled naturally.

This new approach is briefly compared with the much earlier, independent approach offered by Eldon R. Hansen in 1975. The fundamental problem of reliably rendering graphs of implicit equations drives the comparison.

Acknowledgments

First, I acknowledge my parents: my father, who revealed the world of logic, and my mother, who revealed the world of life. That life has become ever more precious after my recent marriage: I thank my wife, Brenda, for all of her kind acts and continued support.

I was given a challenging standard by my supervisors, Eugene Fiume and Rudi Mathon: one cannot be given more.

Much of the clarity of this document is owed to my readers, namely: John Funge, Wayne Hayes, and Francois Pitt. The patience and diligence exhibited by each was exemplary. I thank Xiaoyuan for her warm words; I thank Mahdi for his true words. I express graditude to the others in the lab, for providing the needed distractions from writing.

Finally, I must thank Jim Little, as he introduced me to interval methods during my initial undergraduate year [2, pages 84–88].

Of course, I assume full responsibility for all errors and mistakes still present in this document. I hope that I have met the expectations of all who have helped me on my journey.

Contents

Notation

Number Systems

Interval Number Systems

Notation - Continued

Chapter 2

Letters

Functions

Intervals

Demotions

Miscellaneous

Notation - Continued

Chapter 3

δ_{ij}	Kronecker delta	50
D^-, D^+	lower and upper bound of set D	57
$f_{\geq 0}$	non-negative indicator function	83
$f_{>0}$	positive indicator function	87
$f_{\neq 0}$	non-zero indicator function	86
$f_{\neq k}$	non-integral indicator function	86
$g_{(x=\alpha,y)}, g_{(x,y=\alpha)}$	one-dimensional, axis-aligned projections of binary g	65
G	representative of function g	50
$\mathcal{L}_1$	$\mathcal{L}_1$ norm	72
L_G	Lagrange interpolating polynomial of G	50, 70, 98
ω	weight function for $\mathcal{L}_1$ norm	72
ψ_1^G	linear coefficient of linear L_G	50
ψ_2^G	quadratic coefficient of quadratic L_G	71
$\psi_1^{\downarrow}(G)$	G is monotonically decreasing	50
$\psi_2^{\downarrow}(G)$	G is concave down	71
$\psi_1^0(G)$	G is constant	50
$\psi_2^0(G)$	G is linear	71
$\psi_1^{\uparrow}(G)$	G is monotonically increasing	50
$\psi_2^{\uparrow}(G)$	G is concave up	71
$\psi_1^{\updownarrow}(G)$	G is monotonic	51
$\psi_2^{\updownarrow}(G)$	G is concave	71
$\Phi_{\mathbb{T}}$	description translation function	62
Φ_{ζ}	description translation function	82
$\Xi(g)$	sectioning of g	54
$\Xi^{\star}(g)$	preferred sectioning of g	54
$\Xi_1(g)$	sectioning of g into monotonic pieces	54
$\Xi_2(g)$	sectioning of g into concave pieces	73
$\Xi_{\pm}^{\mathbb{J}}(g)$	sectioning of $g^{\mathbb{J}}$ into periodic pieces	60
$\Xi_{\pm}^{\mathbb{M}}(g)$	sectioning of $g^{\mathbb{M}}$ into periodic pieces	81
$\Xi_{\mid}(g)$	sectioning of g into defined regions	61
$\Xi_{\Delta}(g)$	sectioning of g into continuous regions	62
$\xi_{\geq 0}, \zeta_{\geq 0}$	set of non-negative extended reals	83
$\xi_{>0}, \zeta_{>0}$	set of positive extended reals	87
$\xi_{\neq 0}, \zeta_{\neq 0}$	set of non-zero extended reals	86
$\xi_{\neq k}, \zeta_{\neq k}$	set of non-integral extended reals	86
$\mathrm{Z}_F(m,\xi)$	chooses description function, from F, for (m,ξ)	82
$\subseteq_k$	k-member subset	50

Notation - Continued

Chapter 4

$G[S]$	graph of S	103
$M(\boldsymbol{p})$	region $\boldsymbol{p}$ represents	104
$M^{\mathbb{Y}}(\boldsymbol{p})$	region $\boldsymbol{p}$ represents, expressed using $\mathbb{Y}^2$	108
$\boldsymbol{p}$	pixel	104
$\boldsymbol{P}$	pixel cluster	120
R	rendering	107
$R_{\square}$	rendering produced using pixel testing	109
$R_{\boxdot}$	rendering produced using subpixel sampling	112
$R_{\boxplus}$	rendering produced using exhaustive subpixel sampling	113
$R_{\triangle}$	rendering produced using continuity-based testing	114
$R_{f\square}$	rendering produced using floating-point sampling	117
$R[S]$	rendering of S	107
$R(\boldsymbol{p})$	value of $\boldsymbol{p}$ in R	105
S	graph specification	103
$S^{\mathbb{Y}}$	graph specification, using $\mathbb{Y}$	108

Interdependence Scheme

$$
\begin{tikzcd}
 & 1 \arrow[d] & & & \\
\mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C} & 2.1\text{–}2.4 \arrow[d] & & & \\
\mathbb{F}, \mathbb{R}^{\star} & 2.5\text{–}2.6 \arrow[d] \arrow[r] & 3.1 & 3.2 \arrow[d] \arrow[dr] & \\
\mathbb{I}, \mathbb{J} & 2.7\text{–}2.8 \arrow[d] \arrow[rr] & & 3.2.1\text{–}3.2.15 \arrow[dd] \arrow[dr] & 3.3 \arrow[d] \\
\mathbb{I}_f, \mathbb{J}_f & 2.9\text{–}2.10 \arrow[d] \arrow[rrr] & & & 3.3.1\text{–}3.3.15 \arrow[d] \\
\mathbb{Y}^{\between}, \mathbb{Y}^{|f}, \mathbb{Y}^{\hat{|}F} & 2.11 \arrow[d] \arrow[rr] & & 3.2.16\text{–}3.2.17 \arrow[d] \arrow[r] & 3.3.16\text{–}3.3.17 \arrow[d] \\
\mathbb{Y}^{\hat{|}\hat{\Delta}F} & 2.12 \arrow[d] \arrow[rr] & & 3.2.18\text{–}3.2.19 \arrow[d] \arrow[r] & 3.3.18\text{–}3.3.19 \arrow[d] \\
\mathbb{Y}^{\star} & 2.13 \arrow[d] \arrow[rr] & & 3.2.20\text{–}3.2.21 \arrow[d] \arrow[r] & 3.3.20\text{–}3.3.21 \arrow[d] \\
\mathbb{H} & 2.14\text{–}2.15 & & 3.2.22\text{–}3.2.28 & 3.2.22\text{–}3.2.30 \arrow[d] \\
\mathbb{D} & & & & 4 \arrow[d] \\
 & & & & 5
\end{tikzcd}
$$

The above is only a guide. In particular:

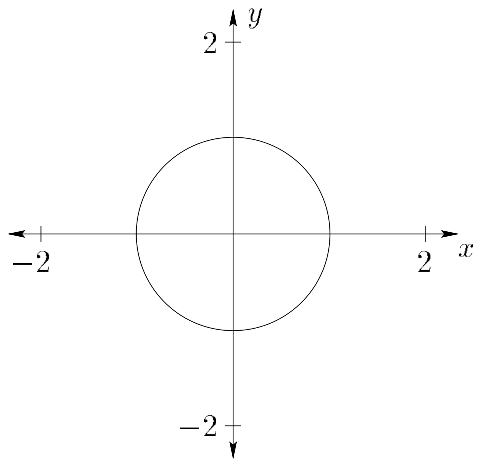
- reading of chapter 4 may commence once constant interval arithmetic is understood.

Chapter 1

Motivation

$$x^2 + y^2 = 1;$$

an unassuming equation, with a simple graph.



Graph of $x^2 + y^2 = 1$

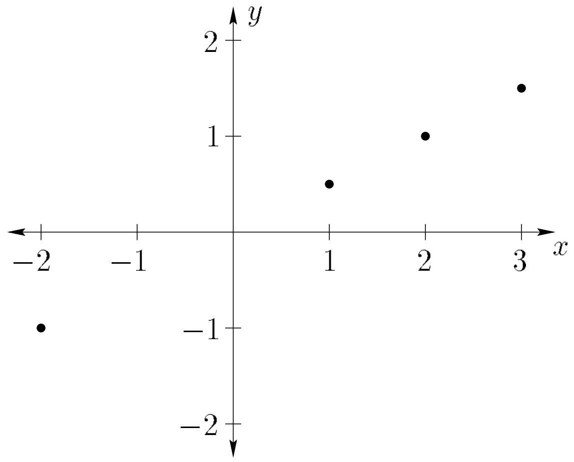
The graph and equation are a pair: the graph contains the points for which the equation is satisfied.

Early in school, as our teachers instill into us the logics of mathematics and geometry, we learn various sets of rules for producing graphs of equations. Given a function p we would first generate a table of the values of $p(x)$, for various values of x. An example table, for $p(x) = \frac{1}{2}x$, follows:

x	$p(x)$
-2	-1
1	$\frac{1}{2}$
2	1
3	$1\frac{1}{2}$

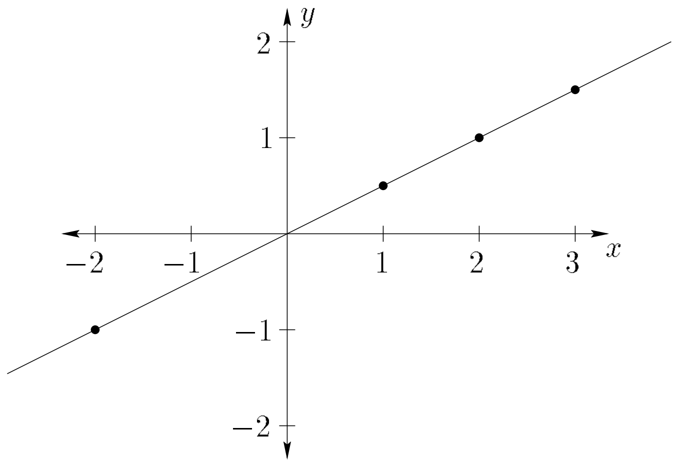
Table of $p(x)$

We would then painstakingly plot the points $(x, p(x))$, secure in the knowledge that these points satisfy the equation $y = p(x)$.



Subgraph of $y = p(x)$

We would then draw a line, connecting the points.



Graph of $y = p(x)$

Glad to be freed from the tedium of plotting points, few question the teacher as to why a line connecting the points may be drawn. During this lesson, the teacher with inquisitive students may well be unlucky, for there is no clear explanation as to why the connecting line may be drawn. It turns out that difficulties arise as this method is applied to general equations.

1.1 Sampling

There is the question as to how many times, and for which arguments, the function is computed. Consider the following equation:

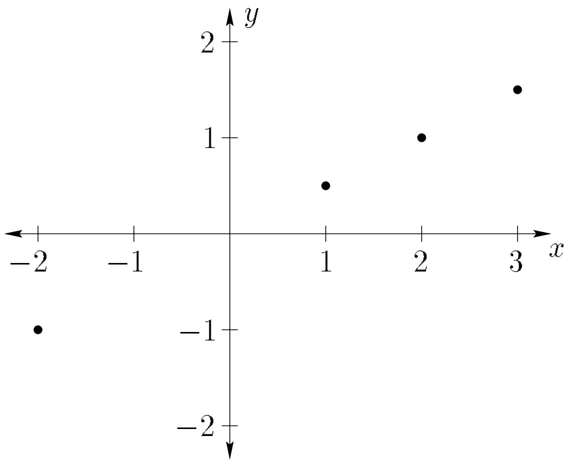
$$y = q(x), \quad q(x) = \tfrac{1}{2}x^4 - 2x^3 - \tfrac{1}{2}x^2 + 8\tfrac{1}{2}x - 6.$$

Computing $q(x)$ generates the sample $(x, q(x))$, of the graph of $y = q(x)$. Sampling, as we did before, generates the following table:

x	$q(x)$
-2	-1
1	$\frac{1}{2}$
2	1
3	$1\frac{1}{2}$

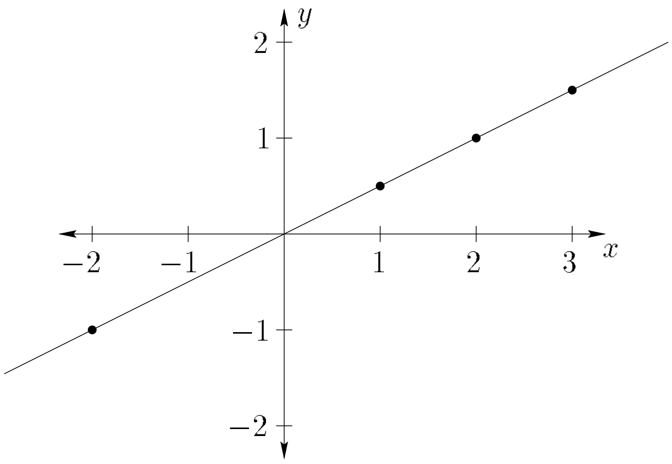
Table of $q(x)$

Surprisingly, the table matches our ealier one. It is not surprising that the plotted points match.



Subgraph of $y = q(x)$

Continuing our procedure by rote, the same graph is generated.



Candidate Graph of $y = q(x)$

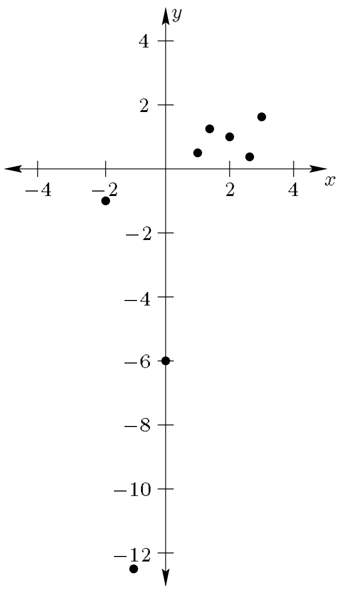
By adding more samples to our table, we see that the previous graph is incorrect.

x	$q(x)$
-2	-1
-1	$-12\frac{1}{2}$
0	-6
1	$\frac{1}{2}$

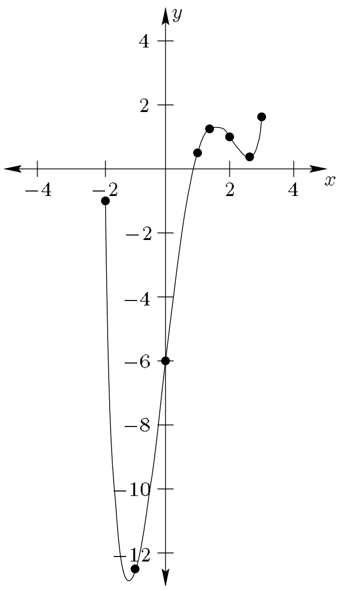
x	$q(x)$
$1\frac{1}{2}$	$1\frac{13}{32}$
2	1
$2\frac{1}{2}$	$\frac{13}{32}$
3	$1\frac{1}{2}$

Richer Table of $q(x)$

Using our richer table, we again plot the points which we know satisfy our equation.



Subgraph of $y = q(x)$



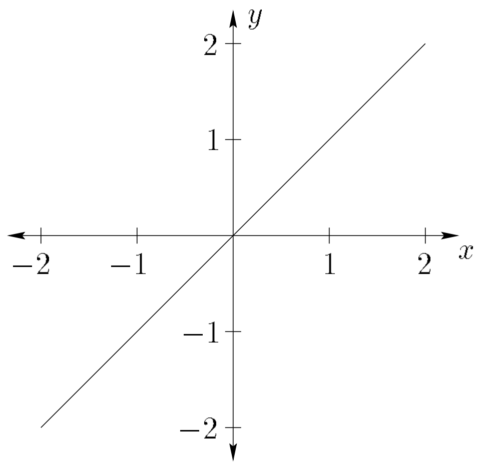
Graph of $y = q(x)$

We have lost confidence in the line which connects the points. Without warning, it has failed us. There is hope that we may be able to predict its failure for polynomials, or other classes of functions, but we aim to graph general equations.

Although calculating a large number of samples guarantees to consume vast resources, it does not guarantee that a more reliable graph is generated. Consider the equation

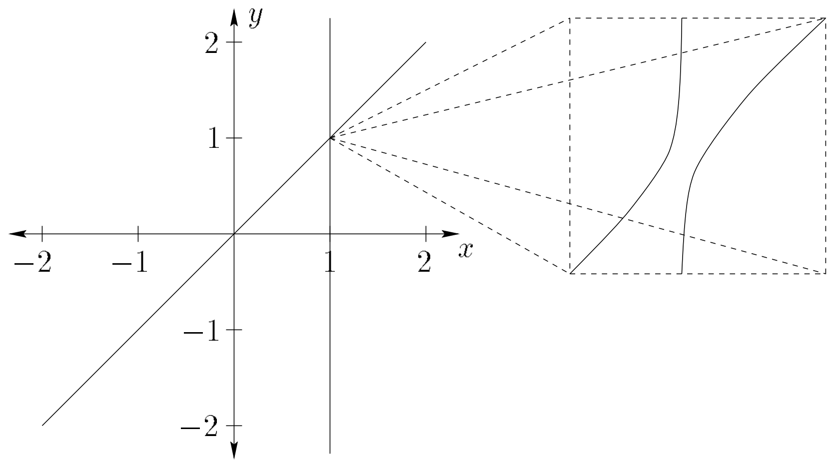
$$y = r(x), \quad r(x) = x\frac{x - 1 - 10^{-9}}{x - 1}.$$

Using over a million uniformly spaced samples of $(x, r(x))$, from $[-2, 2]$, results in the following graph:



Candidate Graph of $y = r(x)$

The actual graph follows, which may be reliably computed using a handful of samples.



Graph of $y = r(x)$

The actual graph is a very sharp hyperbola, and can not be generated by following our procedure, as the graph is composed of two disjoint curves.

1.2 Implicit Equations

An implicit equation, such as our motivating example

$$x^2 + y^2 = 1,$$

may not be expressed as a function g,

$$y = g(x),$$

since for $x = 0$, $y = -1$ and $y = 1$ both satisfy our equation. All hope is not lost, as our equation may be expressed as the union of two functions:

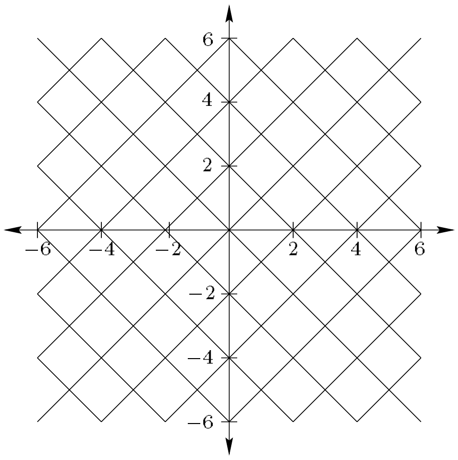
$$x^2 + y^2 = 1 \quad \Leftrightarrow \quad y = \sqrt{1 - x^2} \ \vee \ y = -\sqrt{1 - x^2}.$$

We may then graph each function separately, and then combine the two graphs into a single graph.

Consider the following equation:

$$\cos \pi x = \cos \pi y,$$

whose graph follows:



Graph of $\cos \pi x = \cos \pi y$

If this graph were to be separated into a collection of functions, an infinite number of functions would be needed, since each function may describe at most one point for each value of x. For any value of x, an infinite number of values of y satisfy the equation given. However, for any finite region of the plane, a finite number of functions suffice. Some equations, such as

$$\cos \frac{1}{x} = \cos \frac{1}{y},$$

require an infinite number of functions, even to graph finite regions of the plane, using the procedure just described.

1.3 Relations

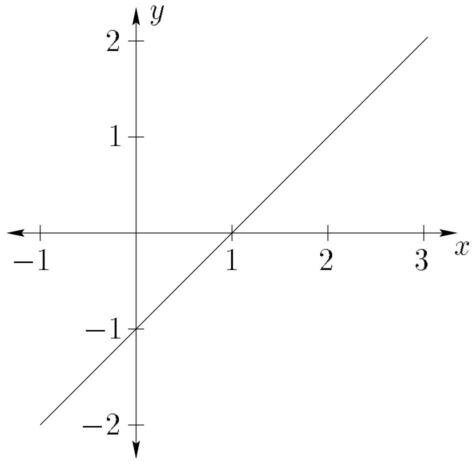
Let us turn our attention to a more difficult problem. Consider graphing a relation, such as

$$y \leq x - 1.$$

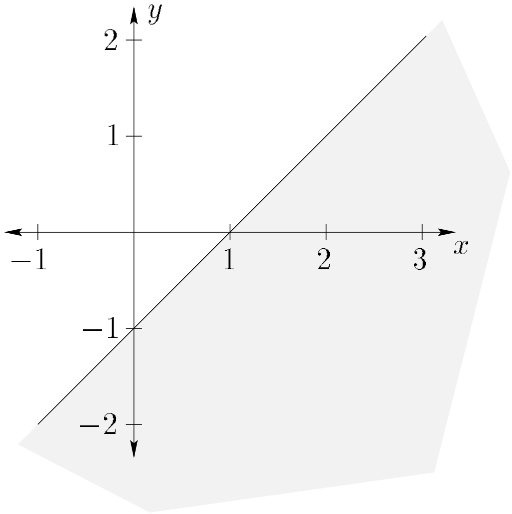
The procedure taught is to first graph the boundary; the boundary of $s(x, y) \leq t(x, y)$ is given by $s(x, y) = t(x, y)$. Our example's boundary is given by

$$y = x - 1.$$

We then shade in the appropriate side.



Graph of $y = x - 1$

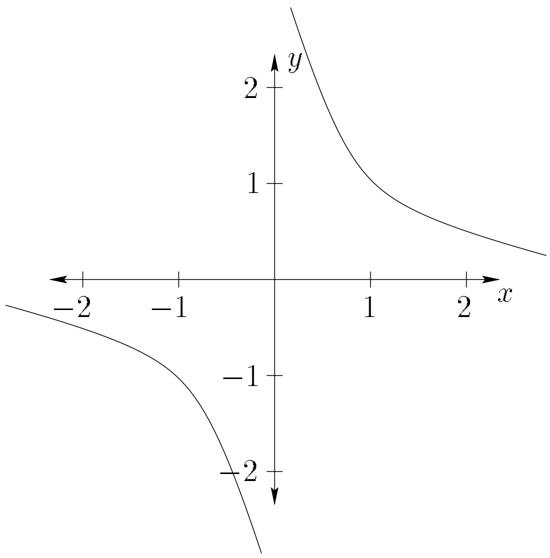


Graph of $y \leq x - 1$

It may seem that graphing relations is not much harder than graphing equations, given the simple approach outlined earlier. But consider the two relations

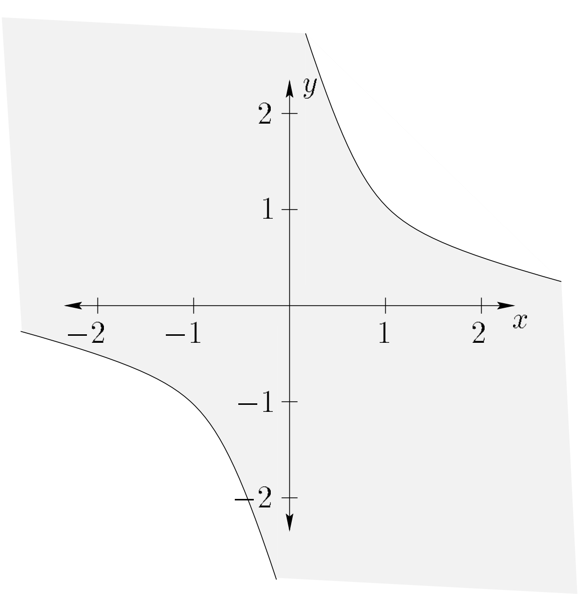
$$xy \leq 1 \quad \text{and} \quad y \leq \frac{1}{x}.$$

Both have the same boundary, which does not break the plane nicely into two "sides".

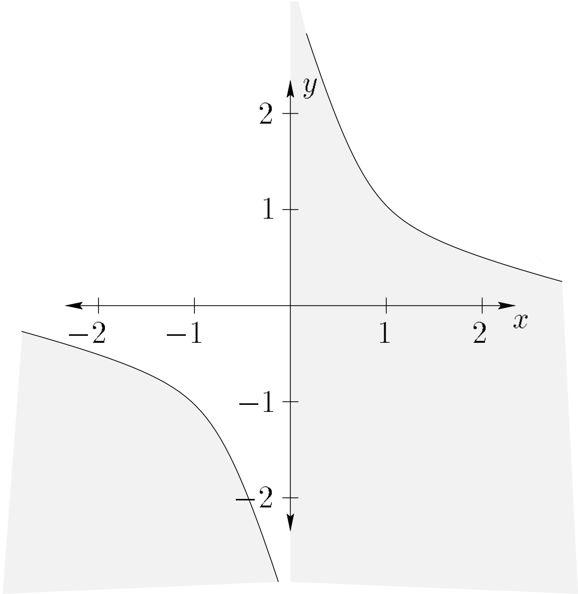


Graph of $xy = 1$, *or* $y = \frac{1}{x}$

The two relations have different graphs, one of which may be given by our side-testing procedure.



Graph of $xy \leq 1$

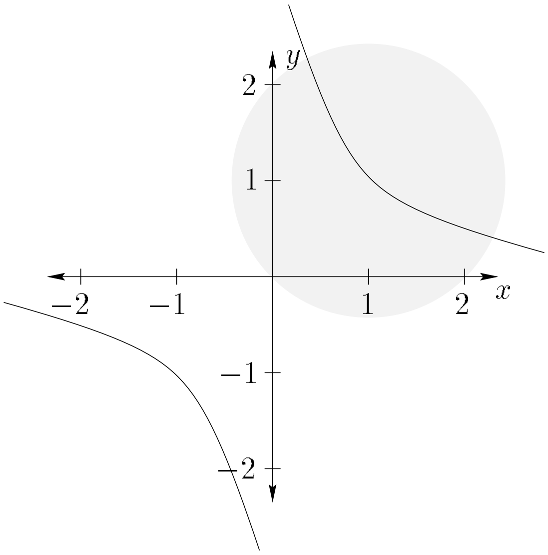


Graph of $y \leq \frac{1}{x}$

If this is not troubling enough, consider the relation

$$\frac{(xy-1)^2}{(x-1)^2+(y-1)^2-2} \leq 0;$$

which, again, has the same boundary as the two earlier relations $xy = 1$ and $y = \frac{1}{x}$. The graph of this new relation follows:



Graph of $\frac{(xy-1)^2}{(x-1)^2+(y-1)^2-2} \leq 0$

The graph contains all points that satisfy

$$xy = 1 \quad \text{or} \quad (x-1)^2 + (y-1)^2 < 2.$$

The only true relationship between the boundary of a graph and the actual graph is that of containment: the graph contains its boundary. In the case of strict inequality, the graph does not contain its boundary.

It appears that graphing relations is indeed more difficult than graphing equations. However, this is not the case; consider the following relation:

$$g(x, y) \geq 0.$$

This relation may be expressed as an equation, as follows:

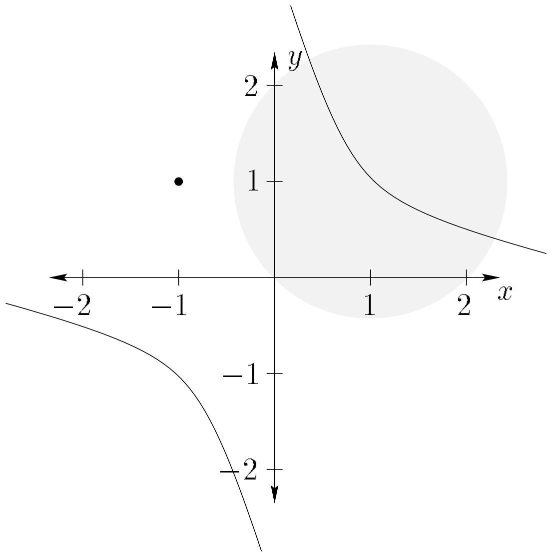
$$g(x, y) = |g(x, y)|.$$

We have not chosen a simpler problem: we have, however, illustrated some of its hidden difficulties. These difficulties are nicely illustrated in the next example.

A single graph may contain zero, one, and two-dimensional elements; consider the equation

$$h(x, y) = -|h(x, y)|, \quad h(x, y) = \frac{(xy-1)^2}{(x-1)^2 + (y-1)^2 - 2}\left((x+1)^2 + (y-1)^2\right),$$

whose graph follows:



Graph of $h(x, y) = -|h(x, y)|$

The graph contains all point which satisfy

$$xy = 1, \qquad (x-1)^2 + (y-1)^2 < 2, \qquad \text{or} \qquad (x+1)^2 + (y-1)^2 = 0.$$

It seems that the entire idea of generating graphs as collections of lines is fundamentally flawed, as a graph may contain two-dimensional elements. Representing two-dimensional elements with a collection of lines is inefficient at best, and simply unconscionable at worst.

1.4 Numerical Round-Off

Given that a basic algorithm has failed us, it is reasonable to do a survey of our basic tools. We are most interested in finding a graphing algorithm which we may program a modern computer to perform. One immediate concern is the representation such machines use for real numbers. The representation often used is analagous to scientific notation, keeping a fixed number of digits for any given quantity. This can lead to further difficulties.

Consider graphing the equation

$$y = n(x), \quad n(x) = (1000 + \sin x) - 999,$$

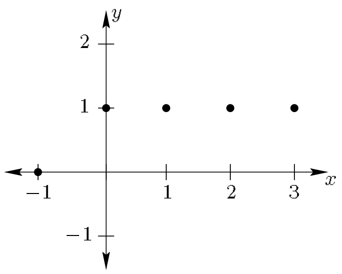
by sampling $n(x)$, limiting ourselves to three digits of precision. A transcription of the computations performed, while sampling $n(x)$ at $x = -1, 0, 1, 2,$ and 3 follows:

$$\begin{array}{ccccccccc}
n(-1) & \rightsquigarrow & (1000 + \sin(-1)) - 999 & \rightsquigarrow & (1000 + -0.841) - 999 & \rightsquigarrow & 999 - 999 & \rightsquigarrow & 0, \\
n(0) & \rightsquigarrow & (1000 + \sin(0)) - 999 & \rightsquigarrow & (1000 + 0) - 999 & \rightsquigarrow & 1000 - 999 & \rightsquigarrow & 1, \\
n(1) & \rightsquigarrow & (1000 + \sin(1)) - 999 & \rightsquigarrow & (1000 + 0.841) - 999 & \rightsquigarrow & 1000 - 999 & \rightsquigarrow & 1, \\
n(2) & \rightsquigarrow & (1000 + \sin(2)) - 999 & \rightsquigarrow & (1000 + 0.909) - 999 & \rightsquigarrow & 1000 - 999 & \rightsquigarrow & 1, \\
n(3) & \rightsquigarrow & (1000 + \sin(3)) - 999 & \rightsquigarrow & (1000 + 0.141) - 999 & \rightsquigarrow & 1000 - 999 & \rightsquigarrow & 1.
\end{array}$$

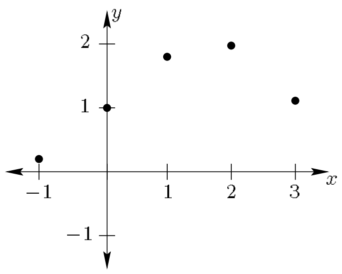
It is clear that for all x, our computations result in $n(x) \rightsquigarrow 1$ or $n(x) \rightsquigarrow 0$, due to numerical round-off. It is equally clear that

$$n(x) = (1000 + \sin x) - 999 = 1000 + \sin x - 999 = 1 + \sin x,$$

so that $n(x) = 1 + \sin x$.



Computed Subgraph of $y = n(x)$



Actual Subgraph of $y = n(x)$

Most calculations introduce some numerical round-off. With complicated equations, there will be long sequences of calculations, which allows numerical round-off to accumulate. For such equations, the generated graph may differ significantly from the actual graph.

1.5 Computability

Given that our problem seems quite difficult, let us focus on a simpler problem. Consider the equation

$$c(x, y) = 0,$$

with the restriction that $c(x, y)$ is an expression of fixed value. The graph of the equation would either be empty, if $c(x, y) \neq 0$, or the entire plane, if $c(x, y) = 0$. After a suitable formalization, this problem may be proven to be non-computable, unless a very restricted set of operators is allowed in the construction of c: for any fixed computer program, there will be equations which it cannot graph correctly.

1.6 Perseverance

Disregarding all of these difficulties, we shall carry on. It is clear that we will be able to find an algorithm that can correctly graph many common equations. In fact; for any finite set of equations, we know that there exists a program that will generate the correct graph for every equation in that set.

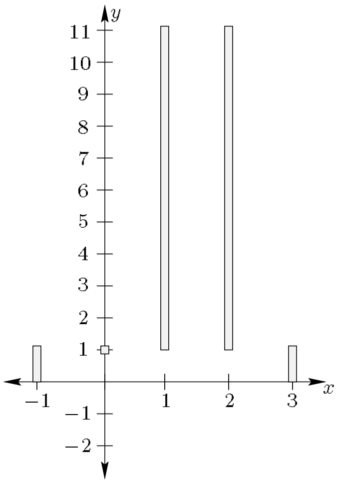
We start by defining a novel set of numbers, along with an arithmetic over these numbers, so that we may compute without worrying excessively about numerical round-off. This arithmetic is a generalization of interval arithmetic, so we will refer to it as "generalized interval arithmetic". With interval arithmetic, lower and upper bounds on computed result are kept. With our previous example, of graphing $y = n(x)$,

$$n(x) = (1000 + \sin x) - 999,$$

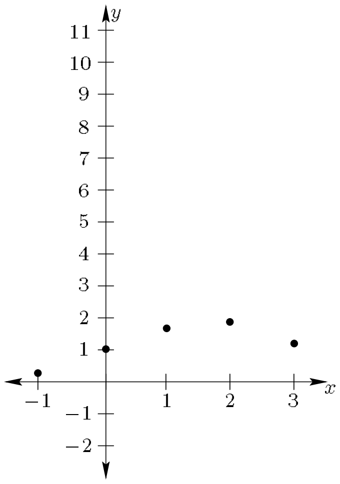
the sampling computation, for $x = 1$, would proceed as follows:

$$\begin{aligned} & n(\langle 1, 1\rangle) \\ \rightsquigarrow\ & (1000 + \sin(\langle 1, 1\rangle)) - 999 \\ \rightsquigarrow\ & (1000 + \langle 0.841, 0.842\rangle) - 999 \\ \rightsquigarrow\ & \langle 1000, 1010\rangle - 999 \\ \rightsquigarrow\ & \langle 1, 11\rangle; \end{aligned}$$

where in each $\langle a, b\rangle$ pair, the first element, a, is a lower bound, while the second element, b, is an upper bound. We have limited ourselves to three digits of precision, as before. Similarly computing samples for $x = -1, 0, 2, 3$ allows us to create a reliable subgraph of $y = n(x)$.



Computed Subgraph of $y = n(x)$



Actual Subgraph of $y = n(x)$

As our interval arithmetic was correctly carried out, we may place complete confidence in our produced subgraph: the actual samples for $x = -1, 0, 1, 2, 3$ lie within our computed samples.

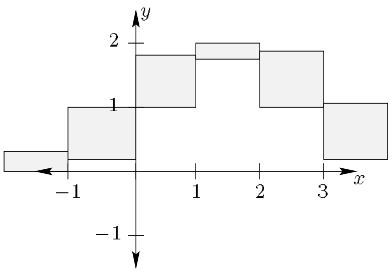
The true strength of interval arithmetic is revealed by sampling with intervals, rather than points. Consider graphing $y = m(x)$,

$$m(x) = 1 + \sin x,$$

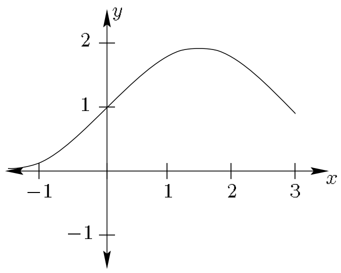
using interval arithmetic. We may sample the interval $\langle 0, 1\rangle$ by computing $m(\langle 0, 1\rangle)$, as follows:

$$\begin{array}{rl} & n(\langle 0,1\rangle) \\ \rightsquigarrow & 1+\sin(\langle 0,1\rangle)) \\ \rightsquigarrow & 1+\langle 0,0.842\rangle \\ \rightsquigarrow & \langle 1,1.85\rangle. \end{array}$$

Similarly computing samples for $x = \langle -2, -1\rangle, \langle -1, 0\rangle, \langle 1, 2\rangle, \langle 2, 3\rangle, \langle 3, 4\rangle$ allows us to create a reliable graph of $y = m(x)$.



Computed Graph of $y = m(x)$



Actual Graph of $y = m(x)$

Again, we have complete confidence in our computed graph: the true graph lies within our computed graph.

A detailed explanation of these techniques form the bulk of this thesis. The techniques are general and may be expanded to grapple other difficult problems.

1.7 Outline

We begin by formalizing interval arithmetic, after a brief formal review of some standard number systems. A variety of interval arithmetics are developed, which will allow us to cope with "badly behaved" equations. Much of the generalizations are novel, and developed by the author.

In the third chapter, a detailed exposition of the arithmetic of generalized intervals is presented. A general approach is taken, so that a similar set of rules may be followed when computing in any one of the myriad of interval arithmetics presented.

We will then precisely define what a graph is, to bring the mathematical idealization into the realm of Computer Science. This will allow for strong results, as we will then have a concrete, realizable goal. Results using several different interval arithmetics will be presented and briefly analysed.

Chapter 2

Numbers

This chapter is about numbers. A pertinent question to ask is: "What is a number?". Rather than answering such a philosophically contentious question I will answer a simpler question: "What are numbers used for?". Numbers are used to quantitatively describe things. For a description to be quantitative it must be possible to mechanistically compare and combine descriptions in a meaningful way.

In this chapter I will very briefly describe some common number systems before introducing a novel system of numbers. This chapter is intended to be pragmatic. Extraneous philosophical debate will be omitted. Readers are encouraged to consult [25, 67] for a deeper discussion of the nature of numbers.

A note, for the mathematically mature reader: the first four sections of this chapter define $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ along with some standard notation. Please begin reading with section 2.5, and use the initial sections for reference, as necessary.

2.1 Integers

The integer number system is a basic system of numbers. The set of all integers is denoted by $\mathbb{Z}$:

$$\mathbb{Z} = \{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}.$$

This number system is particularly simple and forms the basis for all of the other number systems presented here. Although the reader is assumed to be familiar with the integers, some semi-formal discussion follows, which serves to refresh the reader's memory and to illustrate common features of all number systems. The integers can be constructed from the natural numbers; purists construct the naturals using set theory [41, 17, 59].

Integers can be combined through addition and multiplication. Operators abstract the notion of combining numbers, by allowing for unary and 0-ary operators. The terms function and operator are interchangable. $\mathbb{X}$ denotes a set of numbers. An n-ary operator $\oplus$ maps an n-tuple of numbers to a single number. Formally stated,

$$\oplus : \mathbb{X}^n \mapsto \mathbb{X}.$$

Addition and multiplication are binary operators. An n-ary function $f : \mathbb{X}^n \mapsto \mathbb{X}$ may be represented as a set F, of $n+1$-tuples of numbers:

$$F = \{(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, y) : f(\boldsymbol{x}) = y\} \subseteq \mathbb{X}^{n+1}.$$

Boldface is used to indicate vectors.

A set of numbers $\mathbb{X}$ is closed under an n-ary operation $\oplus$ if

$$\forall[(x_1, x_2, \dots, x_n) \in \mathbb{X}^n] \quad \oplus(x_1, x_2, \dots, x_n) \in \mathbb{X}.$$

Integers are closed under addition and multiplication: the sum or product of any two integers is another integer.

Since binary operators are so prevalent several properties of binary operators will be relevant. A binary operator $\oplus$ is commutative if

$$\forall[(x, y) \in \mathbb{X}^2] \;\; (x \oplus y) = (y \oplus x),$$

it is associative if

$$\forall[(x, y, z) \in \mathbb{X}^3] \;\; ((x \oplus y) \oplus z) = (x \oplus (y \oplus z)),$$

it has identity $i \in \mathbb{X}$ if

$$\forall[x \in \mathbb{X}] \;\; (x \oplus i) = (i \oplus x) = x,$$

and it has $\oplus^{-1} : \mathbb{X} \mapsto \mathbb{X}$ as an inverse if

$$\forall[x \in \mathbb{X}] \;\; (x \oplus (\oplus^{-1} x)) = ((\oplus^{-1} x) \oplus x) = i,$$

where i is the identity for $\oplus$. A unary operator g has an inverse g^{-1} if

$$\forall[x \in \mathbb{X}] \;\; g^{-1}(g(x)) = x.$$

An n-ary function $g : \mathbb{X}^n \mapsto \mathbb{X}$ is total if

$$\forall[\boldsymbol{x} \in \mathbb{X}^n] \; g(\boldsymbol{x}) \neq \lambda,$$

where $g(\boldsymbol{x}) = \lambda$ states that g is undefined for agument $\boldsymbol{x}$. A function which is not total is a partial function. The function g is injective (invertible) if

$$\forall[\boldsymbol{x} \in \mathbb{X}^n] \; \forall[\boldsymbol{y} \in \mathbb{X}^n] \;\; [g(\boldsymbol{x}) = g(\boldsymbol{y})] \Rightarrow [\boldsymbol{x} = \boldsymbol{y}].$$

An inverse operator $\oplus^{-1}$ is a total inverse if it is a total operator, otherwise it is a partial inverse. A set of numbers $\mathbb{X}$ is closed under operator $\oplus : \mathbb{X}^n \mapsto \mathbb{X}$ if and only if $\oplus$ is total. The domain of a function g is written formally as $\mathrm{dom}(g)$.

$$\mathrm{dom}(g) \;\; \equiv_{\mathrm{def}} \;\; \{x \mid g(x) \neq \lambda\}.$$

An n-ary function $g : \mathbb{X}^n \mapsto \mathbb{X}$ may be restricted to a set $D \subseteq \mathbb{X}^n$, so that $g(\boldsymbol{x})$ is not defined for $\boldsymbol{x} \notin D$:

$$g|D \;\; \equiv_{\mathrm{def}} \;\; g \cap (D \times \mathbb{X}), \quad \text{so} \quad (g|D)(\boldsymbol{x}) = \begin{cases} g(\boldsymbol{x}) & \text{if } \boldsymbol{x} \in D, \\ \lambda & \text{if } \boldsymbol{x} \notin D. \end{cases}$$

Negation is the total inverse of addition. Subtraction is defined as the sum of a number with another number's additive inverse:

$$(x - y) \;\; \equiv_{\mathrm{def}} \;\; (x + (-y)).$$

A serious limitation of the integers is the lack of a total inverse of multiplication. Division is defined as the product of a number with another number's multiplicative inverse:

$$(x \div y) \ \equiv_{\text{def}} \ (x \times y^{-1}).$$

It follows that the integers are not closed under division.

Addition and multiplication over the integers jointly satisfy the following distributive law:

$$\forall[(x, y, z) \in \mathbb{Z}^3] \ \ (x \times (y + z)) = (x \times y) + (x \times z);$$

multiplication is said to distibute over addition. Addition and multiplication over the integers do not satisfy the following, alternative, distributive law:

$$\forall[(x, y, z) \in \mathbb{Z}^3] \ \ (x + (y \times z)) = (x + y) \times (x + z).$$

The first distributive law will be hereafter referred to as "the" distributive law.

Another nice property of the integers is that comparing any pair of integers will always result in exactly one of three orderings. Equivalently, every pair of distinct integers contains a larger member:

$$\forall[(x, y) \in \mathbb{Z}^2] \ \ (x \neq y) \Rightarrow ((x > y) \overline{\vee} (y > x)),$$

where $\overline{\vee}$ denotes exclusive or.

The comparison operator $\odot$ ($<$, $\leq$, $=$, $\geq$, or $>$) maps pairs of numbers to booleans:

$$\odot : \mathbb{X}^2 \mapsto \mathbb{B},$$

where $\mathbb{B} = \{\mathrm{F}, \mathrm{T}\}$, the set of booleans.

Common Practice

Almost all computers have hardware dedicated to performing very quick operations on integers. Many systems strictly limit the magnitude of the integers to guarantee certain limits on computational resource requirements, while some do not. Although the manipulations of integers by computers is a fascinating and vitally important research area we will envision integers as a basic data type with rudimentary operations.

2.2 Rational Numbers

The rational number system is an extension of the integer number system [42]. The set of all rationals is denoted by $\mathbb{Q}$:

$$\mathbb{Q} = \{\ldots, \tfrac{-2}{1}, \tfrac{-1}{2}, \tfrac{-1}{1}, \tfrac{1}{1}, \tfrac{1}{2}, \tfrac{2}{1}, \tfrac{1}{3}, \tfrac{2}{2}, \tfrac{3}{1}, \tfrac{1}{4}, \tfrac{2}{3}, \ldots\}.$$

Each rational number is a ratio of two integers: a numerator and a non-zero denominator. The rationals extend the integers since the integers are homomorphic to the rationals. An injective mapping $\psi : \mathbb{Z} \mapsto \mathbb{Q}$ is a homomorphism if all the properties of $\mathbb{Z}$ are preserved in $\psi(\mathbb{Z})$. The mapping $\phi_{\mathbb{Z}} : \mathbb{Z} \mapsto \mathbb{Q}$ (abbreviated as ϕ),

$$\phi(x) = \tfrac{x}{1},$$

is a homomorphism from $\mathbb{Z}$ to $\mathbb{Q}$. In order to verfiy this one must show that:

$$\forall[(x,y) \in \mathbb{Z}] \;\; \phi(x +^{\mathbb{Z}} y) = \phi(x) +^{\mathbb{Q}} \phi(y),$$

$$\forall[(x,y) \in \mathbb{Z}] \;\; \phi(x \times^{\mathbb{Z}} y) = \phi(x) \times^{\mathbb{Q}} \phi(y),$$

and

$$\forall[(x,y) \in \mathbb{Z}] \;\; (x >^{\mathbb{Z}} y) \Leftrightarrow (\phi(x) >^{\mathbb{Q}} \phi(y)).$$

When dealing with different number systems in close mutual proximity it will be useful to precisely specify operators as was done above. With a formal definition of rational addition and multiplication, showing that ϕ is a homomorphism is straightforward. Because of the simple homomorphism ϕ, the integers are commonly viewed as a subset of the rationals.

The nice properties of the integers extend to the rationals. The rationals are closed under addition and multiplication. Rational addition has a total inverse as did integer addition. Rational addition and multiplication are associative and commutative; together they obey the distributive law. Rational multiplication has an "almost" total inverse:

$$\left(\frac{x}{y}\right)^{-1} \;\; \equiv_{\text{def}} \;\; \frac{y}{x}.$$

The inverse is not total because zero is not allowed in the denominator of a rational. So the rationals are closed under division, except for division by zero. As with integers, comparisons between rational numbers result in one of three orderings. Most mathematicians do not view the lack of a total multiplicative inverse as a major failing of the rationals.

Consider the squaring operator defined by:

$$x^2 \;\; \equiv_{\text{def}} \;\; x \times x.$$

The inverse of the squaring operator is the square root operator, which satisfies the following:

$$\forall[x \in \mathbb{X}] \;\; (\sqrt{x})^2 = x.$$

There is no total square root operator over the rationals: consider $\sqrt{2}$ or $\sqrt{-1}$. This is one reason to extend the rationals. Many popular operators are not total over the rationals.

Common Practice

There are computer facilities, in the form of software libraries or hardware assist, for performing operations on rationals. These facilities store each rational as a pair of integers [51], or as a continued fraction [36]. Although arithmetic operations ($+$, $-$, $\times$, and $\div$) can be performed with these libraries, many other popular operations cannot be performed directly since many popular operations are not total over the rationals. It is also difficult to predict the computational resources required for a string of operations since the time required to perform an operation is dependent on the numbers involved. Limiting the computational requirements usually results in a system like floating point, which will be discussed shortly.

2.3 Real Numbers

Real numbers are a mathematical abstraction commonly used when modelling real-world phenomenon. Real numbers are an extension of the rational numbers [64]. The set of reals is denoted by $\mathbb{R}$.

$$\{\ldots, -2, -1.7, 0, \tfrac{1}{3}, \pi, \sqrt{3}, e^{1.4}, \sin(1), \ldots\} \subset \mathbb{R}.$$

Each real number can be specified by a converging infinite sequence of rational numbers [26]. The limit of the sequence is the value of the real number.

$$1 \stackrel{\mathbb{R}}{=} \langle \tfrac{1}{1}, \tfrac{1}{1}, \tfrac{1}{1}, \tfrac{1}{1}, \tfrac{1}{1}, \tfrac{1}{1}, \ldots \rangle \stackrel{\mathbb{R}}{=} \langle \tfrac{1}{2}, \tfrac{2}{3}, \tfrac{3}{4}, \tfrac{4}{5}, \tfrac{5}{6}, \tfrac{6}{7}, \ldots \rangle.$$

$$\sqrt{2} \stackrel{\mathbb{R}}{=} \langle \tfrac{1}{1}, \tfrac{14}{10}, \tfrac{141}{100}, \tfrac{1414}{1000}, \tfrac{14142}{10000}, \tfrac{141421}{100000}, \tfrac{1414213}{1000000}, \tfrac{14142136}{10000000}, \ldots \rangle.$$

The real numbers have a partial square root operator as did the rationals. Although the square root operator is defined for all non-negative real numbers, it is not defined for any negative real numbers. There is a natural homomorphism $\phi_{\mathbb{Q}}$ from the rationals to the reals, defined by:

$$\phi(\tfrac{a}{b}) = \langle \tfrac{a}{b}, \tfrac{a}{b}, \tfrac{a}{b}, \tfrac{a}{b}, \tfrac{a}{b}, \tfrac{a}{b}, \ldots \rangle.$$

The rationals are envisioned as being a subset of the reals because of this natural homomorphism. Addition and multiplication are associative and commutative over the reals and jointly satisfy the distributive law. Both operators have inverses, as they did with the rationals. The real number system is preferred over the rational system by mathematicians because many popular operations are closed over the reals. The set of real numbers not in $\phi_{\mathbb{Q}}(\mathbb{Q})$ are called the irrationals and are denoted by $\overline{\mathbb{Q}}$; it is these numbers that allow many common operators to be closed over the reals.

Common Practice

Modelling phenomenon with real numbers is overkill in most cases. Efficiently computing with real numbers directly is quite difficult. In some cases, operations involving real numbers are not computable [42, 60]. Many computational difficulties can be overcome by using a suitable representation for real numbers [69]. This will be discussed at this chapter's end. Even when numerical computation using reals is desirable, symbolic computation can sometimes be used instead.

2.4 Complex Numbers

The set of all complex numbers is denoted by $\mathbb{C}$. The square root operation is closed over $\mathbb{C}$ [8]. As was the case with real numbers, having a closed square root operation is only partly responsible for the importance of complex numbers.

Each complex number can be specified as a pair of real numbers:

$$\{\ldots, 1+i, \sqrt{2}+i\sqrt{3}, 0-2i, \ldots\} \subset \mathbb{C}.$$

The first element of each pair is the "real" part of a complex number, while the second element of each pair is the "imaginary" part of a complex number. The pairs can be written as above, with the imaginary part written as a real multiplied by i, $i = \sqrt{-1}$.

Since there is a simple homomorphism $\phi_{\mathbb{R}\mathbb{C}} : \mathbb{R} \mapsto \mathbb{C}$,

$$\phi(x) = x + 0i,$$

the real numbers are often viewed as a subset of the complex numbers [64].

The algorithms for computing with complex numbers are more intricate than those for real numbers. Even with the more intricate algorithms, this number system has many of the popular properties of the real number system. Addition and multiplication have inverses (partial for multiplication), are associative and commutative, and jointly satisfy the distributive law. Most common operators are closed over the complexes. However, there is no natural ordering relation for complex numbers.

The construction of the complex numbers from the real numbers can be viewed as an application of a general "doubling procedure", a procedure which creates number systems whose elements are represented as 2^n-tuples of real numbers. This same procedure can be used to construct the quaternion and Cayley number systems [33].

Common Practice

Complex numbers are very useful in modelling some phenomena. The same difficulties are encountered in computing directly with complex numbers as are encountered computing with real numbers. This is clear since the real numbers are homomorphic to the complex numbers; and conversely, the complex numbers are built from the real numbers in a remarkably simple way. All of the number systems built by application(s) of the doubling procedure can be emulated directly, using the real numbers as the base number system.

2.5 Floating Point

Floating point numbers are commonly used to approximate real numbers. Floating point facilities are common in computer hardware so most floating point operations can be performed very quickly on computers.

There are many different floating point number systems [5, 49, 50, 35], although they are all very similar. A floating point number can be written as:

$$a \times b^c,$$

where a,b, and c are all in a finite subdomain of the integers.

All of the numbers in a particular floating point number system can be specified with a single choice of b. The set of floating point numbers with $b = 2$ is denoted by $\mathbb{F}[2]$. $\mathbb{F}[2]$ is the system of choice for computer implementations since a and c are usually stored in binary.

Implementations usually represent a and c in a fixed number of bits. A common example is IEEE 754 [5] 64-bit double precision where a is stored in 53 bits (fifty-two bits for the magnititude, one for the sign) while c is stored in 11 bits (using biased binary representation). Such a system is compactly expressed as $\mathbb{F}[2, 53, -2^{10} + 2 \ldots 2^{10} - 1]$: two exponent values are reserved to indicate non-normalized numbers. The floating point operations described below are required in IEEE 754 compliant numerical libraries.

Formally, the system $\mathbb{F}[b, A, m \ldots M]$ includes all numbers which may be expressed as $(a \times b^c)$ and satisfy:

$$(-b^{A-1} < a < b^{A-1}) \wedge (m \leq c + (A-1) \leq M),$$

where a and c are integers. The subtraction present in the right conjunct shifts the "decimal place" so as to relate the exponent range with unity, rather than b^{A-1}.

Another view of the floating point numbers is to imagine the numbers of $\mathbb{F}[b, A, m \ldots M]$ as being described by A base b digits multiplied by b raised to an exponent between m and M:

$$d_0.d_1d_2d_3 \ldots d_{A-1} \times b^e \quad : \quad 0 \leq d_k < b, m \leq e \leq M.$$

Both describe the same system of numbers. The former description builds upon the preceding number systems while the latter gels with one's common experience of performing calculations. The relation between $m \ldots M$ and $\mathbb{F}[b, A, m \ldots M]$ is clearer; as are other important floating point concepts, such as the distinction between normalized numbers, where $d_0 \neq 0$, and denormalized numbers, where $d_0 = 0$.

Throughout this presentation the exact details of the underlying floating point system will not be important so $\mathbb{F}$ will be used to denote any particular floating point system. The exact format used to store floating point numbers does not concern us. The meticulous reader is encouraged to read one of [x,y,z] for details omitted in this brief exposé of floating point. We use $\mathbb{F}[10, 3, -9 \ldots 9]$ for numerical examples.

2.5.1 Infinity

There are two special numbers, ∞ and $-\infty$, which may not be expressed as above. Since numbers are stored in a fixed number of bits these "infinities" are very useful. Both of these numbers are members of all of our floating point number systems. Many properties of these special numbers are intuitive. For example:

$$\forall[x \in \mathbb{F}] \quad (x \neq -\infty) \Rightarrow (x +^{\mathbb{F}} \infty = \infty).$$

The floating point number ∞ can represent a real number too large to be described by a finite number of $\mathbb{F}$. Similar sign infinities are incomparable. Both $\infty < \infty$ and $\infty = \infty$ are false, since it is unknown which real number each ∞ represents.

2.5.2 NAN

Another number allowed for by computer implementations is λ. When a module implementing an operator is invoked with values for which the operator is not defined the module will return λ. For example:

$$\sqrt{-1} =^{\mathbb{F}} \lambda,$$

since floating point is an abstraction of real numbers rather than complex numbers. λ is referred to as a NAN (not-a-number) and is a member of $\mathbb{F}$.

The NAN is not crucial to our development of number systems since it is essentially a crutch to allow for detection of exceptional conditions after they occur. They do allow for compact computer routines. Interval arithmetic routines will detect upcoming exceptional conditions before they result in an application of an operator where it is not properly defined.

The NAN causes further erosion of the comparison operators. Any comparison involving a NAN is false; the NAN is an unordered number. Every pair (x, y) of floating point numbers is ordered in one of three ways unless x or y is λ, $x = y = \infty$, or $x = y = -\infty$.

2.5.3 Rounding

Floating point numbers approximate real numbers. Operations with floating point numbers approximate corresponding operations with real numbers. Consider the following addition operation:

$$1 \times 10^0 + 1 \times 10^3 = 1001 \times 10^0.$$

Both 1×10^0 and 1×10^3 are members of $\mathbb{F} = \mathbb{F}[10, 3, -9 \ldots 9]$; 1001×10^0 is not.

When the implied real result of a floating point operation is not a floating point number the result is rounded to a floating point number. The most common form of rounding is "rounding to nearest" where the result is rounded to the nearest floating point number. Using such rounding the previous example would result in:

$$1 \times 10^0 +^{\mathbb{F}} 1 \times 10^3 = 1 \times 10^3.$$

Another form of rounding is "upward rounding" where the result is rounded up to a larger floating point number. If the result is positive, it is rounded away from zero; if the result is negative, it is rounded towards zero. Using such rounding the previous example would result in:

$$1 \times 10^0 +^{\mathbb{F}+} 1 \times 10^3 = 101 \times 10^1.$$

Another form of rounding is "downward rounding" where the result is rounded down to a smaller floating point number. If the result is positive, it is rounded towards zero; if the result is negative, it is rounded away from zero. Using such rounding the previous example would result in:

$$1 \times 10^0 +^{\mathbb{F}-} 1 \times 10^3 = 1 \times 10^3.$$

Numerical libraries provide three forms of rounding: $\mathbb{F}=$, $\mathbb{F}+$, and $\mathbb{F}-$. The default mode of rounding is $\mathbb{F}=$. When an explicit rounding mode is not specified, as was done earlier, $\mathbb{F}=$ is assumed.

Although IEEE 754 requires that the algebraic operators $+$, $-$, $\times$, $\div$, and $\sqrt{x}$ are rounded to the nearest floating point number, other operators are not so favoured. The following example will illustrate what can happen with operators whose results are not guaranteed to be accurate to within one ULP (Unit in the Last Place). With a $\sin(x)$ implementation that is guaranteed to be accurate to within 40 ULPS the following may occur:

$$\begin{aligned}
\sin^{\mathbb{R}}(1) &\approx 0.841470984807896506652502321630298999622563 1 \\
&= 841.470984807896506652502321630298999622563 1 \times 10^{-3}; \\
\sin^{\mathbb{F}=}(1 \times 10^0) &= 843 \times 10^{-3}, \\
\sin^{\mathbb{F}+}(1 \times 10^0) &= 844 \times 10^{-3}, \\
\sin^{\mathbb{F}-}(1 \times 10^0) &= 810 \times 10^{-3}.
\end{aligned}$$

The actual value, $\sin^{\mathbb{R}}(1)$, is bracketed by $\sin^{\mathbb{F}-}(1 \times 10^0)$ and $\sin^{\mathbb{F}+}(1 \times 10^0)$. These brackets may be widely separated; with our example sine implementation they may differ by up to 80 ULPS. The result using "rounding to nearest" only guarantees that the true result will fall within the bracketed region.

Using real numbers directly in computations is currently infeasible. Floating point numbers are commonly used because of their computational advantages. Unfortunately, rounding causes the result returned to be inexact.

2.5.4 Algebraic Properties

Because of these inexact results, none of the three varieties of addition are associative, as shown below:

$$((1 \times 10^5 + -1 \times 10^5) + 1 \times 10^0)^{\mathbb{F}=} = 1 \times 10^0 \neq 0 \times 10^0 = (1 \times 10^5 + (-1 \times 10^5 + 1 \times 10^0))^{\mathbb{F}=},$$

$$((1 \times 10^5 + -1 \times 10^5) + 1 \times 10^0)^{\mathbb{F}+} = 1 \times 10^0 \neq 1 \times 10^3 = (1 \times 10^5 + (-1 \times 10^5 + 1 \times 10^0))^{\mathbb{F}+},$$

$$((1 \times 10^5 + -1 \times 10^5) + 1 \times 10^0)^{\mathbb{F}-} = 1 \times 10^0 \neq 1 \times 10^4 = (1 \times 10^5 + (-1 \times 10^5 + 1 \times 10^0))^{\mathbb{F}-}.$$

Similarly for multiplication: none of the three varieties are associative. Addition and multiplication do have the identities 0×10^0 and 1×10^0, respectively. All varieties of addition and multiplication are commutative over the floats, but the lack of associativity causes any non-trivial symbolic manipulation of an expression to affect the expression's value. Negation is a total inverse for all three addition operators. Since the base is a fixed integer none of the three multiplication operators have total inverses. None of the 3^5 possible distributive laws are obeyed. Algebraic properties of rounded computations are discussed in [37, 39, 38].

Common Practice

The floating point number system does not obey many nice formal rules [40]. Extensions and generalizations of IEEE 754 floating-point have been put forward [13, 6]. For many applications the use of floating point does not adversely affect the output, which has been envisioned as coming from computations using real numbers. With long streams of computations there is a worry that the floating point computation stream will radically diverge from the underlying real computation stream. In these cases, formal arguments involving particular implementations of the operators and particular sequences of computations must be made.

2.6 Extended Real Numbers

Since we will be discussing floating point numbers further, it will be useful to have an abstract model of floating point numbers. That model is the extended real number system, $\mathbb{R}^\star$:

$$\mathbb{R}^\star = \mathbb{R} \cup \{\infty, -\infty\}.$$

There is a natural homomorphism $\phi_{\mathbb{F}\mathbb{R}^\star} : \mathbb{F} \mapsto \mathbb{R}^\star$ from the floats to the extended reals. There is another natural homomorphism $\phi_{\mathbb{R}\mathbb{R}^\star} : \mathbb{R} \mapsto \mathbb{R}^\star$ from the reals to the extended reals. These homomorpisms allow for comparisons and operations to be applied between floats and reals by type promotion.

2.6.1 Hyperreal Numbers

Hyperreal numbers, denoted by ${}^\star\mathbb{R}$, are a similar extension of real numbers [16, 20, 63]. Hyperreals extend the reals by adding both infinite numbers, such as ∞, as well as infinitesimal numbers, such as Δ. Since Δ satisfies:

$$\forall [r \in \mathbb{R}] \;\; r > 0 \;\Rightarrow\; \Delta \in (0, r),$$

Δ is not a real number. Infinitesimals and infinities are very useful in presenting non-standard analysis which, for many, is more intuitive than standard real analysis. Some argument can be made for using $^{\star}\mathbb{R}$ as an idealized model of $\mathbb{F}$ since $\mathbb{F}$ contains two distinct numbers $+0$ and -0 which would correspond to Δ and $-\Delta$. IEEE 754 $\mathbb{F}$ does not, however, contain a third number 0 distinct from $+0$ and -0.

The hyperreals are an extension of the reals; they are constructed so that all statements which are provable over the reals are provable over the hyperreals, using a classical proof system. There is another, substantially different, approach to non-standard analysis [54]. With this "smooth non-standard analysis", all functions are infinitely differentiable.

2.6.2 Type Conversion

The three forms of floating point rounding are examples of type demotion. When a real number is rounded it is demoted into a relatively sparse system of numbers. Although the real number $\sqrt{2}$ and the floating point number 1×10^0 are incomparable, the two numbers may be compared by promotion via the natural homomorphisms, as follows:

$$\phi_{\mathbb{R}\mathbb{R}^\star}(\sqrt{2}) >^{\mathbb{R}^\star} \phi_{\mathbb{F}\mathbb{R}^\star}(1 \times 10^0).$$

There is a more compact syntax for describing type conversion. Any number may be converted to another type by attaching the target type. For example, $a^{\mathbb{R}^\star}$ specifies that the number a is converted to an extended real number. Type demotion can be more specific than promotion since there may be several ways to demote the number. Notice of promotion may be omitted as in the following example which adds three floating point numbers a, b, and c:

$$(a+b+c)^{\mathbb{R}^\star \to \mathbb{F}+} \equiv (a^{\mathbb{R}^\star} + b^{\mathbb{R}^\star} + c^{\mathbb{R}^\star})^{\mathbb{F}+}.$$

The inner promotions do not need to be mentioned since they can be inferred by the conversion from extended reals. The same addition using standard rounding could be specified as follows:

$$(a+b+c)^{\mathbb{F}=} \equiv (a +^{\mathbb{F}=} b +^{\mathbb{F}=} c).$$

Since floating point addition is not strictly associative the order of addition should be specified. Strict bounds on the allowable error of particular floating point operator implementations will not be used in this presentation, so lax expression specification can be tolerated.

2.6.3 Infinity Unveiled

A good intuition for the properties of ∞ may be formed by considering ∞ to be an ever growing, unbounded sequence of rationals, such as:

$$\infty =^{\mathbb{R}^\star} \{\tfrac{1}{3}, \tfrac{5}{4}, \tfrac{17}{2}, \tfrac{61}{5}, \tfrac{101}{7}, \ldots\} =^{\mathbb{R}^\star} \{\tfrac{1}{1}, \tfrac{3}{1}, \tfrac{2}{1}, \tfrac{4}{1}, \tfrac{3}{1}, \tfrac{5}{1}, \ldots\}.$$

Formally, $x = \{x_1, x_2, x_3, \ldots\} \in \mathbb{R}^\star$ is equal to ∞ if

$$\forall [l \in \mathbb{Q}]\ \exists j\ \forall [k > j]\ \ (x_k >^{\mathbb{Q}} l),$$

where j and k are integers. An intuition for $-\infty$ may be similarly formed.

$$\{x_1, x_2, x_3, \ldots\} =^{\mathbb{R}^\star} -\infty \quad \Leftrightarrow \quad \forall [l \in \mathbb{Q}]\ \exists j\ \forall [k > j]\ \ (x_k <^{\mathbb{Q}} l).$$

2.7 Interval Arithmetic

Although floating point computations are simple and efficient, rounding can cause a stream of floating point computations to quietly diverge from the envisioned stream of real computations. Interval arithmetic guarantees rigorous results yet is built from floating point arithmetic. Interval arithmetic will not prevent a series of computations from wandering but it will inform the user how much the computed result could deviate from the real result (the result using real numbers for the computations). The presentation given here differs somewhat from conventional introductions [4, 56, 57], due to the impending generalizations.

The set of intervals is denoted by $\mathbb{I}$. An interval is specified by two floating point numbers, a lower and upper bound.

$$\{\ldots, \langle 1\times 10^0, 2\times 10^0\rangle, \langle -1\times 10^0, 7\times 10^0\rangle, \langle -\infty, -2\times 10^3\rangle, \langle -\infty, \infty\rangle, \ldots\} \subset \mathbb{I}.$$

The interval $\langle a, b\rangle$ represents any particular real number between a and b. Rather than returning a single floating point number each operation will return a range of numbers which the real result is guaranteed to be in.

For example, π can be represented as the interval

$$\pi^{\mathbb{I}} = \langle 314\times 10^{-2}, 315\times 10^{-2}\rangle,$$

since

$$314\times 10^{-2} \leq^{\mathbb{R}^\star} \pi \leq^{\mathbb{R}^\star} 315\times 10^{-2}.$$

Operations involving $\pi^{\mathbb{I}}$ are not "aware" that $\pi^{\mathbb{I}}$ represents π. The operations only assume that $\pi^{\mathbb{I}}$ represents some fixed real number between 314×10^{-2} and 315×10^{-2}.

2.7.1 Syntax

The upper bound of interval i is denoted by i^+ while i^- denotes the lower bound:

$$i \equiv \langle i^-, i^+\rangle.$$

The width of an interval is the difference between the upper and lower bound, and is denoted by $i^{||}$ for the interval i:

$$i^{||} \equiv_{\text{def}} i^+ - i^-.$$

Every interval has non-negative width:

$$\forall[i\in\mathbb{I}]\ \ i^{||\mathbb{R}^\star} \geq 0.$$

Most intervals have positive width, but an interval could have zero width if it represents a particular real number which happens to coincide with a floating point number. A real number is contained in an interval if that interval can represent the real number:

$$\forall[x\in\mathbb{R}]\ \forall[\langle a,b\rangle\in\mathbb{I}]\ \ x\in\langle a,b\rangle \Leftrightarrow a\leq x\leq b.$$

The set of number within an interval i is denoted by $i^{\square}$:

$$i^{\square} = \{x : x\in i\}.$$

2.7.2 Order

The set of intervals do not have boolean comparison operators. The comparison operators for intervals are three valued logic operators. Operator $\odot$ compares intervals:

$$\odot : \mathbb{I}^2 \mapsto \mathbb{T}.$$

Three valued logic is denoted by $\mathbb{T}$.

$$\mathbb{T} \equiv_{\text{def}} \{\mathrm{F}, \mathrm{TF}, \mathrm{T}\}.$$

Consider the following examples:

$$\langle 1, 3\rangle <^{\mathbb{I}} \langle 4, 6\rangle = \mathrm{T},$$

$$\langle 1, 3\rangle >^{\mathbb{I}} \langle 4, 6\rangle = \mathrm{F},$$

$$\langle 1, 3\rangle <^{\mathbb{I}} \langle 2, 4\rangle = \mathrm{TF},$$

$$\langle 1, 3\rangle =^{\mathbb{I}} \langle 2, 4\rangle = \mathrm{TF},$$

$$\langle 3, 3\rangle =^{\mathbb{I}} \langle 3, 3\rangle = \mathrm{T}.$$

Three valued logic values can be demoted to boolean values in two ways; optimistically via $\mathbb{B}+ : \mathbb{T} \mapsto \mathbb{B}$, or pessimistically via $\mathbb{B}- : \mathbb{T} \mapsto \mathbb{B}$.

$$\mathbb{B}+ \equiv_{\text{def}} \{\mathrm{F} \mapsto \mathrm{F}, \mathrm{TF} \mapsto \mathrm{T}, \mathrm{T} \mapsto \mathrm{T}\}.$$

$$\mathbb{B}- \equiv_{\text{def}} \{\mathrm{F} \mapsto \mathrm{F}, \mathrm{TF} \mapsto \mathrm{F}, \mathrm{T} \mapsto \mathrm{T}\}.$$

2.7.3 Inclusion Property

There are guidelines to follow when implementing interval operators. It is crucial that the implementations follow the spirit of interval arithmetic: the intervals represent any fixed real number in their range, and that the operator's result represents every possible real result. The inclusion property formally states this.

Unary function $g^{\mathbb{I}}$ has the inclusion property if

$$\forall[i \in \mathbb{I}]\ \forall[x \in i]\ \ g^{\mathbb{R}}(x) \in g^{\mathbb{I}}(i).$$

A binary function $g^{\mathbb{I}}$ has the inclusion property if

$$\forall[(i, j) \in \mathbb{I}^2]\ \forall[(x, y) \in (i, j)]\ \ g^{\mathbb{R}}(x, y) \in g^{\mathbb{I}}(i, j).$$

A function which has the inclusion property can also be said to satisfy the inclusion property. Since intervals are essentially a computational tool, a function will often be identified with, or described by, an algorithm. The function $g^{\mathbb{I}}$ is said to model the underlying function $g^{\mathbb{R}}$.

In general, an n-ary function $g^{\mathbb{I}}$ satisfies the inclusion property if

$$\forall[\boldsymbol{i} \in \mathbb{I}^n]\ \forall[\boldsymbol{x} \in \boldsymbol{i}]\ \ g^{\mathbb{R}}(\boldsymbol{x}) \in g^{\mathbb{I}}(\boldsymbol{i}).$$

The inclusion property codifies validity. The implementation $g^{\mathbb{I}}$ of real function $g^{\mathbb{R}}$ is a valid implementation if $g^{\mathbb{I}}$ has the inclusion property.

2.7.4 Interval Extension

For any total function $g^{\mathbb{R}}$, an implementation $g^{\mathbb{I}}$ which always returns the universal interval $\langle -\infty, \infty \rangle$ is valid. Clearly implementation validity is not a sufficient indicator of quality. A good interval arithmetic implementation of a function returns small intervals.

The interval extension of a unary real function $g^{\mathbb{R}}$ is an interval function $g^{\mathbb{I}}$ defined by:

$$g^{\mathbb{I}}(i) = \langle l^{\mathbb{F}-}, u^{\mathbb{F}+} \rangle : \; l = \inf_{x \in i} g^{\mathbb{R}}(x), \; u = \sup_{x \in i} g^{\mathbb{R}}(x),$$

where l is an extended real number which bounds $g^{\mathbb{R}}(x)$ from below for all x in i; l is rounded down to determine the lower bound of $g^{\mathbb{I}}(i)$. l may have the value $-\infty$ if $g^{\mathbb{R}}(x)$ has no finite lower bound. u is similarly used to determine the upper bound. Although the interval extension is not a method to construct good interval operators, it can be used to show that a particular implementation returns optimal values.

The interval extension $g^{\mathbb{I}}$ of an n-ary function $g^{\mathbb{R}}$ is defined by:

$$g^{\mathbb{I}}(\boldsymbol{i}) = \langle l^{\mathbb{F}-}, u^{\mathbb{F}+} \rangle : \; l = \inf_{\boldsymbol{x} \in \boldsymbol{i}} g^{\mathbb{R}}(\boldsymbol{x}), \; u = \sup_{\boldsymbol{x} \in \boldsymbol{i}} g^{\mathbb{R}}(\boldsymbol{x}).$$

2.7.5 Algebraic Properties

Intervals generalize floating point numbers since there is an injective mapping $\phi_{\mathbb{F}\mathbb{I}} : \mathbb{F} \mapsto \mathbb{I}$ defined by:

$$\phi(x) = \langle x, x \rangle.$$

This mapping is not an isomorphism, although it allows one to identify the floating point numbers with $\phi_{\mathbb{F}\mathbb{I}}(\mathbb{F})$. The mapping $\phi_{\mathbb{F}\mathbb{I}} : \phi_{\mathbb{F}\mathbb{I}}(\mathbb{F}) \mapsto \mathbb{F}$ defined by:

$$\phi(\langle x, x \rangle) = x,$$

is an isomorphism between $\phi_{\mathbb{F}\mathbb{I}}(\mathbb{F})$, a subset of the intervals, and $\mathbb{F}$. A mapping $\phi_{\mathbb{X}_a\mathbb{X}_b} : \mathbb{X}_a \mapsto \mathbb{X}_b$ is an isomorphism if $\phi_{\mathbb{X}_a\mathbb{X}_b}$ is a homomorphism from $\mathbb{X}_a$ to $\mathbb{X}_b$, and ${\phi_{\mathbb{X}_a\mathbb{X}_b}}^{-1}$ is a homomorphism from $\mathbb{X}_b$ to $\mathbb{X}_a$.

Addition inherits the identity $\langle 0, 0 \rangle = \phi_{\mathbb{F}\mathbb{I}}(0)$ while multiplication inherits the identity $\langle 1, 1 \rangle = \phi_{\mathbb{F}\mathbb{I}}(1)$. Since intervals were constructed with mathematical rigor in mind, several nice properties are obeyed by intervals. Chief among these is the sub-distributive law:

$$\forall[(i, j, k) \in \mathbb{I}^3] \;\; (i \times (j + k)) \subseteq ((i \times j) + (i \times k)).$$

Although neither addition nor multiplication are associative, the operators preserve "associative trails". This property is expressed, for addition, as follows:

$$\forall[(i, j, k) \in \mathbb{I}^3] \; \forall[(a, b, c) \in (i, j, k)] \;\; ((a + b) + c) \in (i + (j + k)).$$

The property follows from the associativity of real addition and the inclusion property of interval addition. The above property can be extended by considering that interval addition is commutative. In general, a real computation result is guaranteed to be contained in the result of the associated interval computation because the interval inclusion property is transitive.

2.8 Real Interval Arithmetic

As extended numbers are useful when discussing floating point numbers, real intervals are a useful abstract model of floating point intervals. The set of real intervals is denoted by $\mathbb{J}$. Each real interval is specified by a lower and upper endpoint, both of which are extended real numbers.

$$\{\ldots, \langle -\infty, -\pi\rangle, \langle\sqrt{2}, \sqrt{3}+1\rangle, \langle 6,7\rangle, \langle -\infty, \infty\rangle, \langle 0,0\rangle, \ldots\} \subset \mathbb{J}.$$

The syntax for intervals is used for all forms of interval arithmetic, and will be used for abstract models of interval arithmetic as well. The ensuing development of interval arithmetic will flesh out the concepts introduced by floating point interval arithmetic.

Interval arithmetic is used to model computations with reals. Operators are defined over the reals and then modelled with interval operators. The interval inclusion property gives interval methods their rigor. An n-ary function $g^{\mathbb{J}}$ is a valid interval representation of the n-ary function $g^{\mathbb{R}}$ if $g^{\mathbb{J}}$ satisfies the interval inclusion property. The function $g^{\mathbb{J}}$ satisfies the inclusion property if

$$\forall[j \in \mathbb{J}]\; \forall[x \in j] \;\; g^{\mathbb{R}}(x) \in g^{\mathbb{J}}(j).$$

The judgement of model quality can again be guided by the interval extension $g^{\mathbb{J}} : \mathbb{J}^n \mapsto \mathbb{J}$ of a real function $g^{\mathbb{R}} : \mathbb{R}^n \mapsto \mathbb{R}$. The interval extension is defined as before:

$$g^{\mathbb{J}}(j) = \langle l, u\rangle : \;\; l = \inf_{x\in j} g^{\mathbb{R}}(x), \;\; u = \sup_{x \in j} g^{\mathbb{R}}(x).$$

The interval extension of a real function is the best possible model of that real function.

Real intervals behave much like the floating point intervals they abstract. The abstraction allows one to ignore the effects of rounding, which can simplify discussion and analysis.

2.9 Generalized Interval Arithmetic

Interval arithmetic can be generalized in several ways. Before further complicating the presentation I will unify floating point and real interval number systems.

2.9.1 Unification

The symbol $\mathcal{I}$ is a transformational operator which transforms number systems into interval number systems. Floating point interval arithmetic, $\mathbb{I}$, can be rewritten as $\mathcal{I}(\mathbb{F})$; while real interval arithmetic, $\mathbb{J}$, can be rewritten as $\mathcal{I}(\mathbb{R}^\star)$. As $\mathbb{X}$ denotes a number system, $\mathbb{Y}$ denotes an interval number system.

Interval arithmetic has been generalized through this simplification. Consider the number system $\mathcal{I}(\mathbb{Z}^\star)$ which denotes an interval system where the endpoints are extended integers:

$$\mathbb{Z}^\star = \mathbb{Z} \cup \{-\infty, \infty\}.$$

Infinities are useful in the underlying number system since intervals may need to describe arbitrarily distant numbers. Without them, some interval operators are forced to be only partially defined.

Consider the interval number system $\mathcal{I}(\mathbb{X})$; the previous example had $\mathbb{X} = \mathbb{Z}^\star$. The interval inclusion property for n-ary function g is clearly stated as:

$$\forall[\boldsymbol{i} \in (\mathcal{I}(\mathbb{X}))^n]\; \forall[\boldsymbol{x} \in \boldsymbol{i}] \;\; g(\boldsymbol{x}) \in g^{\mathcal{I}(\mathbb{X})}(\boldsymbol{i}).$$

The argument $\boldsymbol{x}$ is considered to vary over the domain of g. This property is equivalent to the inclusion property for both real and floating point intervals.

The interval extension of an n-ary function g is defined as:

$$g^{\mathcal{I}(\mathbb{X})}(\boldsymbol{i}) = \langle l^{\mathbb{X}-}, u^{\mathbb{X}+}\rangle : \ l = \inf_{\boldsymbol{x}\in\boldsymbol{i}} g(\boldsymbol{x}), \ u = \sup_{\boldsymbol{x}\in\boldsymbol{i}} g(\boldsymbol{x}).$$

The demotions $\mathbb{X}-$ and $\mathbb{X}+$ are used since the derived interval endpoints will need to be "rounded out" to ensure the endpoints are valid and of the correct type. The argument $\boldsymbol{x}$ is considered to vary over the domain of g. Demotions are not needed if the underlying number system is no poorer than the number system which the result of g belongs to, as was seen when g was a real valued function and the interval system was $\mathbb{J}$.

2.9.2 Three Valued Logic

Boolean logic can be thought of as a very simple number system, given our original framework; a boolean description is a quantitative description. Conjunction and disjunction are often thought of as multiplication and addition, respectively. Both are associative and commutative. Both distributive laws are obeyed:

$$a \wedge (b \vee c) = (a \wedge b) \vee (a \wedge c), \ a \vee (b \wedge c) = (a \vee b) \wedge (a \vee c).$$

T is an identity for conjunction while F is an identity for disjunction. Neither operator is invertible. The numbers can be ordered by agreeing that $\mathrm{F} <^{\mathbb{B}} \mathrm{T}$.

Three valued logic is isomorphic to $\mathcal{I}(\mathbb{B})$. The mapping $\phi_{\mathbb{T}} : \mathbb{T} \mapsto \mathcal{I}(\mathbb{B})$,

$$\phi = \{\mathrm{F} \mapsto \langle \mathrm{F}, \mathrm{F}\rangle, \mathrm{TF} \mapsto \langle \mathrm{F}, \mathrm{T}\rangle, \mathrm{T} \mapsto \langle \mathrm{T}, \mathrm{T}\rangle\},$$

is an isomophism between $\mathbb{T}$ and $\phi(\mathbb{T}) = \mathcal{I}(\mathbb{B})$. We let $a \sqsubseteq b$ denote that b is a valid description of a, where a and b are members of $\mathcal{I}(\mathbb{B}) = \mathbb{T}$:

$$a \sqsubseteq b \ \equiv_{\mathrm{def}} \ a \subseteq b, \qquad (a, b) \in \mathbb{T}^2.$$

This notation will clarify some later statements by reminding the reader that the arguments of $\sqsubseteq$ are members of $\mathbb{T}$.

All of the properties of three valued logic can be deduced from this, together with the properties of boolean logic. This is the spirit behind three valued logic: TF symbolizes a lack of knowledge. With further knowledge each TF can be reduced to either F or T.

2.9.3 Linear Intervals

The transformational operator $\mathcal{I}$ will be extended to further generalize interval arithmetic.

We envision numbers as a tool used to describe things. Many of the things described by numbers are parameterized. For example, we may be using a number to describe the mass of an iron ball. The iron ball can be parameterized by its radius. We will bring these parameters into our number system. This integration of parameters into numbers will give us hints as to how numbers depend on the parameters. These hints will enable the new interval routines to return much tighter results.

To start, we will consider a number system with a single parameter α. The parameter can vary from zero to one, and is a real number. Again, as a starting point we will consider only linear relationships between the parameter and the value's bounds. This new number system, using real numbers as the underlying number system, is denoted by $\mathcal{I}_{p+q\alpha}(\mathbb{R}^\star)$ or $\mathbb{J}_{p+q\alpha}$. The subscript, $p+q\alpha$, states that a linear relationship is used; the greek letter α signifies the parameter α of the linear function, while the latin letters p and q signify coefficients of the linear function.

A linear interval is described by a lower and upper endpoint, each of which is a linear function of α. The coefficients of the linear functions must be numbers of the underlying system $\mathbb{R}^\star$.

$$\{\ldots, \langle 1,2\rangle, \langle \alpha, 2+3\alpha\rangle, \langle -e^{e^2}, 1-6\alpha\rangle, \langle \sqrt{3}+\alpha, \sqrt{5}+\tfrac{1}{2}\alpha\rangle, \langle 1-\alpha, \infty\rangle, \ldots\} \in \mathbb{J}_{p+q\alpha}.$$

The semantics of the interval $\langle a+b\alpha, c+d\alpha\rangle \in \mathbb{J}_{p+q\alpha}$ is: when parameter α has value k, the interval represents a fixed real number between $a+bk$ and $c+dk$. This can be stated formally as:

$$x \in \langle a+b\alpha, c+d\alpha\rangle \quad\Leftrightarrow\quad a+b\alpha \leq^{\mathbb{R}^\star} x \leq^{\mathbb{R}^\star} c+d\alpha,$$

for real number x. Although the upper and lower endpoints of interval $x \in \mathbb{J}_{p+q\alpha}$ are linear functions of α, the real number represented by the interval may not be a linear function of the parameter α. The intervals cannot collapse. The interval between the lower and upper bound must be well-defined:

$$\forall[j \in \mathbb{J}_{p+q\alpha}]\ \forall[\alpha \in [0,1]]\ \ j^+(\alpha) \geq^{\mathbb{R}^\star} j^-(\alpha).$$

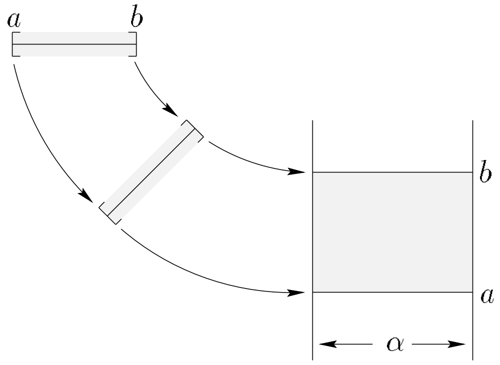
This follows from the original statement that intervals must have non-negative width:

$$j^{||\mathbb{R}^\star} \geq 0 \ \Rightarrow\ \forall[\alpha \in [0,1]]\ j^+(\alpha) - j^-\alpha \geq^{\mathbb{R}^\star} 0 \ \Rightarrow\ \forall[\alpha \in [0,1]]\ j^+(\alpha) \geq^{\mathbb{R}^\star} j^-(\alpha).$$

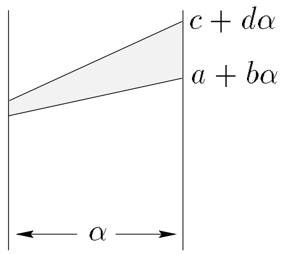
The upper and lower functions must both be well-defined over $[0,1]$:

$$\forall[j \in \mathbb{J}_{p+q\alpha}] \quad [0,1] \subseteq \mathrm{dom}(j^-) \ \wedge\ [0,1] \subseteq \mathrm{dom}(j^+).$$

A picture may soothe the intuition. Associate the interval $\langle a,b\rangle$ with the closed set $[a,b]$, of extended real numbers. The free variable α may be accommodated by introducing a new dimension. The interval $\langle a,b\rangle$ does not interact with this new dimension, although the interval $\langle a+b\alpha, c+d\alpha\rangle$ does. The earlier intervals may now be regarded as "constant intervals".



Constant Interval $\langle a,b\rangle$



Linear Interval $\langle a+b\alpha, c+d\alpha\rangle$

An example is appropriate. Consider the problem of determining the range of an arbitrary function $g : \mathbb{R} \mapsto \mathbb{R}$ over the domain $[0,1]$. The parameter α in this case is the argument to the function. The range may be computed by simply evaluating the function with x being a number representing the domain of interest, $[0,1]$.

Since α varies over $[0,1]$, the linear function $0 + 1\alpha$ represents the domain completely: every element of the domain $[0,1]$ is represented by $0+1\alpha$, for some value of $\alpha \in [0,1]$. It follows that the linear interval $\langle \alpha, \alpha \rangle$ represents the domain: every element of the domain is contained in the interval $\langle \alpha, \alpha\rangle$. The constant interval $\langle 0, 1\rangle$ represents the domain since every element of the domain is represented by an element of $\langle 0, 1\rangle$.

Consider the simple function $g(x) = x - x$. So, with $\mathbb{J}_{p+q\alpha}$, this proceeds as follows:

$$\text{domain} = \langle \alpha, \alpha \rangle,$$

$$\begin{aligned} g(\text{domain}) &\rightsquigarrow g(\langle \alpha, \alpha\rangle) \\ &\rightsquigarrow \langle \alpha, \alpha\rangle - \langle \alpha, \alpha\rangle \\ &\rightsquigarrow \langle 0, 0\rangle. \end{aligned}$$

The resulting bound for the range is $[0,0]$, which is the actual range. Using $\mathbb{J}$ this would proceed as follows:

$$\text{domain} = \langle 0, 1\rangle,$$

$$\begin{aligned} g(\text{domain}) &\rightsquigarrow g(\langle 0, 1\rangle) \\ &\rightsquigarrow \langle 0, 1\rangle - \langle 0, 1\rangle \\ &\rightsquigarrow \langle -1, 1\rangle. \end{aligned}$$

The resulting bound for the range is $[-1,1]$, which is valid, but definitely not optimal. To determine the range of g over the domain $[a,b]$, the domain would be represented by the linear interval $\langle a + (b-a)\alpha, a + (b-a)\alpha\rangle$, or the constant interval $\langle a, b\rangle$. The domain may also be represented by the linear interval $\langle b + (a-b)\alpha, b + (a-b)\alpha\rangle$. Any valid linear interval representative of the domain $[a,b]$ must contain either $\langle a + (b-a)\alpha, a+(b-a)\alpha\rangle$ or $\langle b + (a-b)\alpha, b + (a-b)\alpha\rangle$; unless it also represents a larger domain D, $[a,b] \subset D$.

The linear real interval number system may be denoted by $\mathbb{M}$. The notation follows from the denotation $\mathbb{L}$ for the linear floating point interval number system:

$$\mathbb{L} \equiv_{\text{def}} \mathbb{I}_{p+q\alpha} = \mathcal{I}_{p+q\alpha}(\mathbb{F}), \;\; \mathbb{M} \equiv_{\text{def}} \mathbb{J}_{p+q\alpha} = \mathcal{I}_{p+q\alpha}(\mathbb{R}^\star).$$

A linear interval model $g^{\mathbb{M}}$ of an n-ary function $g^{\mathbb{R}}$ satistfies the inclusion property if

$$\forall[\boldsymbol{m} \in \mathbb{M}^n]\; \forall[\alpha \in [0,1]]\; \forall[\boldsymbol{x} \in \boldsymbol{m}(\alpha)] \;\; g(\boldsymbol{x}) \in [g^{\mathbb{M}}(\boldsymbol{m})](\alpha).$$

Since $m \in \mathbb{M}$ is a function of α, $m(\alpha)$ is well defined as a closed real interval:

$$\langle a + b\alpha, c + d\alpha\rangle(k) = [a + bk, c + dk].$$

The linear interval extension of the n-ary function g is defined as follows:

$$g^{\mathbb{M}}(\boldsymbol{m}) = \langle l^{\mathbb{M}-}, u^{\mathbb{M}+}\rangle : \; l(\alpha) = \inf_{\boldsymbol{x}\in \boldsymbol{m}(\alpha)} g(\boldsymbol{x}), \;\; u(\alpha) = \sup_{\boldsymbol{x}\in\boldsymbol{m}(\alpha)} g(\boldsymbol{x}).$$

The lower and upper bounds l and u are functions of α. The lower bound $l(\alpha)$ bounds $f(\boldsymbol{x})$ from below; the range of $\boldsymbol{x}$ is $\boldsymbol{m} \in \mathbb{M}^n$, and is therefore a function of α. Although l is a function of α it is not guaranteed to be linear, or even continuous. The demotion from an arbitrary function to a linear function is significant. The demoted $l^{\mathbb{M}-}(\alpha)$ bounds $l(\alpha)$ from below while the demoted $u^{\mathbb{M}+}(\alpha)$ bounds $u(\alpha)$ from above. In both cases, α varies from zero to one.

The demotions $\mathbb{M}-$ and $\mathbb{M}+$ are significant in the definition of interval extension because there is no best demotion available. This is drastically different from the demotions $\mathbb{F}-$ and $\mathbb{F}+$ used in the definition of floating point interval extensions. When demoting an extended real to a float, there is a particular floating point number which is the best choice. When rounding down, the largest floating point number less than or equal to the extended real is chosen. The best choice when demoting an arbitrary extended real function l to a linear extended real function $l^{\mathbb{M}}$ depends on how $l^{\mathbb{M}}$ will be used. This significantly changes the character of the interval extension. It can no longer be used to show an interval model is optimal in general, although it may be used to show that an interval method is suboptimal.

2.9.4 Constant Intervals

The original interval arithmetic $\mathbb{I}$ can now be viewed as constant interval arithmetic, where the lower and upper function bounds are constants:

$$\mathbb{I} = \mathbb{I}_k = \mathcal{I}_k(\mathbb{F}),\ \ \mathbb{J} = \mathbb{J}_k = \mathcal{I}_k(\mathbb{R}^\star).$$

There is an injective mapping $\phi_{\mathbb{JM}} : \mathbb{J} \mapsto \mathbb{M}$,

$$\phi(\langle a, b\rangle) = \langle a + 0\alpha, b + 0\alpha\rangle.$$

With the operators defined via interval extension, the mapping $\phi_{\mathbb{JM}}$ is a homomorphism from $\mathbb{J}$ to $\mathbb{M}$. There is a similar mapping $\phi_{\mathbb{IL}}$ from $\mathbb{I}$ to $\mathbb{L}$.

2.9.5 Quadratic Intervals

Rather than using linear bounds for the intervals, quadratic bounds may be used. The quadratic real interval number system is denoted by $\mathbb{V}$:

$$\mathbb{U}\ \equiv_{\text{def}}\ \mathbb{I}_{p+q\alpha+r\alpha^2} = \mathcal{I}_{p+q\alpha+r\alpha^2}(\mathbb{F}),\ \ \mathbb{V}\ \equiv_{\text{def}}\ \mathbb{J}_{p+q\alpha+r\alpha^2} = \mathcal{I}_{p+q\alpha+r\alpha^2}(\mathbb{R}^\star).$$

Each interval u of $\mathbb{V}$ is specified by two quadratic functions, each of which is specified by three extended real numbers:

$$\forall[\boldsymbol{j}_1 + \boldsymbol{j}_2\alpha + \boldsymbol{j}_3\alpha^2\boldsymbol{k}_1 + \boldsymbol{k}_2\alpha + \boldsymbol{k}_3\alpha^2 \in \mathbb{V}]\ [(\boldsymbol{j}, \boldsymbol{k}) \in (\mathbb{R}^{\star 3}, \mathbb{R}^{\star 3})].$$

Since we require that both the lower and upper bound be well-defined functions, some possible descriptions are never valid. An example is the function $\infty - \infty\alpha$, which is not defined for $\alpha = \frac{1}{2} \in [0, 1]$. The methods used to implement interval operators will naturally avoid such descriptions.

Function demotion through $\mathbb{V}+$ and $\mathbb{V}-$ is more difficult than function demotion through $\mathbb{M}+$ and $\mathbb{M}-$. A later section will describe how function demotion is performed.

2.9.6 Multi-Dimensional Linear Intervals

A number may describe something with several parameters. Several parameters may be integrated into the number system. The simplest such system is $\mathbb{M}_2$, where the interval bounds are linear functions of α and β:

$$\mathbb{M}_2\ \equiv_{\text{def}}\ \mathbb{J}_{p+q\alpha+r\beta}.$$

Each parameter α and β may independently vary from zero to one.

In general, $\mathbb{M}_k$ is defined as a real linear interval number system with k parameters. Each parameter may vary from zero to one independently:

$$\mathbb{M}_k \equiv_{\text{def}} \mathbb{J}_{p+\Sigma q_i \boldsymbol{\alpha}_i}, \; \boldsymbol{\alpha} \in [0,1]^k.$$

The term linear interval was chosen over affine interval due to familiarity. Although the bounds are techincally affine functions, an interval system which used linear functions would not see much use. As will be seen when interval arithmetic application algorithms are discussed, there will often be a mapping from an "actual" parameter $\mathcal{A}_i$ to a system parameter $\boldsymbol{\alpha}_i$ to allow for more complex parameter domains. Forcing the upper and lower bounds to be zero when $\boldsymbol{\alpha}_i = 0$ would severely restrict these mappings, and the applicability of interval methods.

Consider our example problem, of determining the range of a function over a given domain. The linear interval chosen to represent the domain $[a, b]$ was $\langle a + (b-a)\alpha, a + (b-a)\alpha\rangle$. The upper and lower bounds are not always linear functions, since $a + (b-a)\alpha \neq 0$ for $\alpha = 0$.

2.9.7 Functional Intervals

Allowing functions with more descriptive power as an interval bounds is the obvious way to generalize interval arithmetic.

For any particular function $f : [0,1]^n \mapsto \mathbb{R}^\star$ there is an interval arithmetic number system $\mathcal{I}_f(\mathbb{F}) = \mathbb{I}_f$ with an abstract model $\mathcal{I}_f(\mathbb{R}^\star) = \mathbb{J}_f$. I will assume that the function has k "coefficients", and n "parameters". Each interval bound would then be specified by k extended real numbers, and would vary over an n-dimensional domain:

$$\forall[j \in \mathbb{J}_f] \;\; \exists[(\boldsymbol{a}, \boldsymbol{b}) \in (\mathbb{R}^{\star k}, \mathbb{R}^{\star k})] \;\; j = \langle f[\boldsymbol{a}], f[\boldsymbol{b}]\rangle;$$

$$f[\boldsymbol{a}] : [0,1]^n \mapsto \mathbb{R}^\star, \quad f[\boldsymbol{b}] : [0,1]^n \mapsto \mathbb{R}^\star.$$

The notation $f[\boldsymbol{a}]$ states that the k coefficients of f are filled in by the k elements of $\boldsymbol{a}$. An interval description is valid if the described interval does not collapse:

$$\forall[(\boldsymbol{a}, \boldsymbol{b}) \in (\mathbb{R}^{\star k}, \mathbb{R}^{\star k})] \;\; \langle f[\boldsymbol{a}], f[\boldsymbol{b}]\rangle^{||\mathbb{R}^\star} \geq 0 \;\Rightarrow\; \langle f[\boldsymbol{a}], f[\boldsymbol{b}]\rangle \in \mathbb{J}_f.$$

The width of an interval is interpreted as before:

$$\langle f[\boldsymbol{a}], f[\boldsymbol{b}]\rangle^{||\mathbb{R}^\star} \geq 0 \;\equiv\; \forall[\boldsymbol{\alpha} \in [0,1]^n] \;\; f[\boldsymbol{b}](\boldsymbol{\alpha}) \geq f[\boldsymbol{a}](\boldsymbol{\alpha}).$$

An interval model $g^{\mathbb{J}_f} : \mathbb{J}_f^m \mapsto \mathbb{J}_f$ of an m-ary function g is valid if $g^{\mathbb{J}_f}$ has the inclusion property. The model $g^{\mathbb{J}_f}$ has the inclusion property if

$$\forall[\boldsymbol{j} \in \mathbb{J}_f^m] \; \forall[\boldsymbol{\alpha} \in [0,1]^n] \; \forall[\boldsymbol{x} \in \boldsymbol{j}(\boldsymbol{\alpha})] \;\; g(\boldsymbol{x}) \in [g^{\mathbb{J}_f}(\boldsymbol{j})](\boldsymbol{\alpha}).$$

Containment is interpreted as before:

$$x \in \langle f[\boldsymbol{a}], f[\boldsymbol{b}]\rangle \;\Leftrightarrow\; f[\boldsymbol{a}](\boldsymbol{\alpha}) \leq^{\mathbb{R}^\star} x \leq^{\mathbb{R}^\star} f[\boldsymbol{b}](\boldsymbol{\alpha}).$$

The interval extension $g^{\mathbb{J}_f}$ of g is also defined as before:

$$g^{\mathbb{J}_f}(\boldsymbol{j}) = \langle l^{\mathbb{J}_f -}, u^{\mathbb{J}_f +}\rangle : \; l(\boldsymbol{\alpha}) = \inf_{\boldsymbol{x} \in \boldsymbol{j}(\boldsymbol{\alpha})} g(\boldsymbol{x}), \;\; u(\boldsymbol{\alpha}) = \sup_{\boldsymbol{x} \in \boldsymbol{j}(\boldsymbol{\alpha})} g(\boldsymbol{x}).$$

Different choices of f lead to differing complexity in the implementation of the operator models (such as $+^{\mathbb{J}_f}$, $\times^{\mathbb{J}_f}$, and $\div^{\mathbb{J}_f}$) and the demotion operators $\mathbb{J}_f+$ and $\mathbb{J}_f-$. The choice of f will affect how well the intervals can track the underlying stream of real computations as well as how useful the computed results will be.

2.9.8 Symbolic Intervals

The most general choice of f is to allow arbitrarily complex functions. Another way to view this is to have f as a universal function with an infinite number of coefficients:

$$f(\boldsymbol{\alpha}) = \boldsymbol{j}_1 + \sum_a \boldsymbol{k}_a \boldsymbol{\alpha}_a + \sum_a \boldsymbol{l}_a \frac{1}{\boldsymbol{\alpha}_a} + \sum_{a,b} \boldsymbol{m}_{a,b} \frac{\boldsymbol{\alpha}_a}{\boldsymbol{\alpha}_b} + \sum_{a,b,c} \boldsymbol{n}_{a,b,c} \frac{\boldsymbol{\alpha}_a \boldsymbol{\alpha}_b}{\boldsymbol{\alpha}_c} + \cdots \sum_a \boldsymbol{p}_a \sin(\boldsymbol{p'}_a \alpha_a + \boldsymbol{p''}_a) + \cdots$$

Of course, at any point in a computation only a finite number of coefficients are non-zero. A method of this form could completely avoid any difficult decisions by just "pushing" the computation into the interval symbolically, as shown in the example following.

Consider using such a system with our simple algorithm for determing a function's range over a given domain. An example problem instance is to determine the range of $g : \mathbb{R} \mapsto \mathbb{R}$,

$$g(x) = x \sin(x) + \tfrac{\sqrt{x}}{x+\pi},$$

over the domain $[0, 1]$. The algorithm would proceed as follows:

$$\text{domain} = \langle \alpha, \alpha \rangle,$$

$$\begin{aligned} g(\text{domain}) &\rightsquigarrow g(\langle \alpha, \alpha \rangle) \\ &\rightsquigarrow \langle \alpha, \alpha \rangle \times \sin(\langle \alpha, \alpha \rangle) + \tfrac{\sqrt{\langle \alpha, \alpha \rangle}}{\langle \alpha, \alpha \rangle + \langle \pi, \pi \rangle} \\ &\rightsquigarrow \langle \alpha \sin \alpha, \alpha \sin \alpha \rangle + \langle \tfrac{\sqrt{\alpha}}{\alpha+\pi}, \tfrac{\sqrt{\alpha}}{\alpha+\pi} \rangle \\ &\rightsquigarrow \langle \alpha \sin \alpha + \tfrac{\sqrt{\alpha}}{\alpha+\pi}, \alpha \sin \alpha + \tfrac{\sqrt{\alpha}}{\alpha+\pi} \rangle. \end{aligned}$$

Although the algorithm returned a description of the tightest bounds possible on the range of g, the results are obviously not of much use. We are no further along than when we started.

Symbolic computation is not a panacea. Although the operator models and demotion operators would be trivial to implement, interpreting the results becomes difficult. Symbolic simplification could be performed by the interval operators.

2.10 Generalized Floating Point Interval Arithmetic

Actual implementations of interval arithmetic use floating point numbers to describe interval bounds. I will only discuss $\mathbb{I}_f$ directly since a re-reading with an appropriate fixed choice of f will provide a discussion of $\mathbb{I}_k$, $\mathbb{L}_k$, or $\mathbb{U}$.

I will assume the bound description function f takes n parameters and has k floating point coefficients:

$$f[\boldsymbol{a}] : [0, 1]^n \mapsto \mathbb{R}^\star, \ \boldsymbol{a} \in \mathbb{F}^k.$$

As before, an interval description is valid if the described interval is non-collapsing. An interval i is a member of $\mathbb{I}_f$ if and only if a description of i is valid:

$$\forall[(\boldsymbol{a}, \boldsymbol{b}) \in (\mathbb{F}^k, \mathbb{F}^k)] \ \ \langle f[\boldsymbol{a}], f[\boldsymbol{b}] \rangle^{||\mathbb{R}^\star} \geq 0 \ \Leftrightarrow \ \langle f[\boldsymbol{a}], f[\boldsymbol{b}] \rangle \in \mathbb{I}_f,$$

$$\langle f[\boldsymbol{a}], f[\boldsymbol{b}] \rangle^{||\mathbb{R}^\star} \geq 0 \ \equiv \ \forall[\boldsymbol{\alpha} \in [0, 1]^n] \ f[\boldsymbol{b}](\boldsymbol{\alpha}) \geq f[\boldsymbol{a}](\boldsymbol{\alpha}).$$

An interval model $g^{\mathbb{I}_f} : \mathbb{I}_f^m \mapsto \mathbb{I}_f$ of an m-ary function g has the interval inclusion property if

$$\forall[\boldsymbol{j} \in \mathbb{I}_f^m]\ \forall[\boldsymbol{\alpha} \in [0,1]^n]\ \forall[\boldsymbol{x} \in \boldsymbol{j}(\boldsymbol{\alpha})]\ \ g(\boldsymbol{x}) \in [g^{\mathbb{I}_f}(\boldsymbol{j})](\boldsymbol{\alpha}),$$

$$x \in \langle f[\boldsymbol{a}], f[\boldsymbol{b}]\rangle \ \Leftrightarrow\ f[\boldsymbol{a}](\boldsymbol{\alpha}) \leq^{\mathbb{R}^\star} x \leq^{\mathbb{R}^\star} f[\boldsymbol{b}](\boldsymbol{\alpha}).$$

The interval extension $g^{\mathbb{I}_f}$ of g is derived from the real interval extension $g^{\mathbb{J}_f}$:

$$g^{\mathbb{I}_f}(\boldsymbol{j}) = \langle l^{\mathbb{J}_f - \to \mathbb{I}_f -}, u^{\mathbb{J}_f + \to \mathbb{I}_f +}\rangle :\ l(\boldsymbol{\alpha}) = \inf_{\boldsymbol{x} \in \boldsymbol{j}(\boldsymbol{\alpha})} g(\boldsymbol{x}),\ \ u(\boldsymbol{\alpha}) = \sup_{\boldsymbol{x} \in \boldsymbol{j}(\boldsymbol{\alpha})} g(\boldsymbol{x}).$$

Deriving the floating point extension from the real extension means that the really hard decision of how to bound an arbitrary function is made once. It also lends credence to the concept that $\mathbb{J}_f$ is an abstract model of $\mathbb{I}_f$. Deriving the bounds in this way will lead to suboptimal bounds since the demotion $\mathbb{J}_f-$ does not take into account the granularity of $\mathbb{I}_f$. The difference between the two stage demotion $([0,1]^n \mapsto \mathbb{R}^\star) \to \mathbb{J}_f- \to \mathbb{I}_f-$ and the direct demotion $([0,1]^n \mapsto \mathbb{R}^\star) \to \mathbb{I}_f-$ will only be on the order of machine precision, however.

There are two differences between $\mathbb{J}_f$ and $\mathbb{I}_f$. One is the two-stage demotion used in the interval extension. The other is the floating point evaluation of interval bounds.

A bound f is a function from $[0,1]^n$ to $\mathbb{R}^\star$. Implementations will have to evaluate f using floating point numbers. Formally, this is stated as a straight forward application of a demotion from $\mathbb{R}^\star$ to $\mathbb{F}$. A floating point number x is a member of interval $\langle f[\boldsymbol{a}], f[\boldsymbol{b}]\rangle$ for parameter value $\boldsymbol{\alpha}$ if

$$f[\boldsymbol{a}](\boldsymbol{\alpha}) \leq^{\mathbb{R}^\star} x \leq^{\mathbb{R}^\star} f[\boldsymbol{b}](\boldsymbol{\alpha}),$$

which promoted x to an extended real. This is not what an implementation would do. An implementation would round the bounds outward, so that a floating point number x is a member of $\langle f[\boldsymbol{a}], f[\boldsymbol{b}]\rangle$ for parameter value $\boldsymbol{\alpha}$ if

$$f^{\mathbb{F}-}[\boldsymbol{a}](\boldsymbol{\alpha}) \leq^{\mathbb{F}} x \leq^{\mathbb{F}} f^{\mathbb{F}+}[\boldsymbol{b}](\boldsymbol{\alpha}).$$

2.11 Interval Function Domains

The requirement that an interval number be non-collapsing will be relaxed. The number system $\mathbb{Y}^{\between}$ is the number system $\mathbb{Y}$ with the restriction, that intervals be non-collapsing, removed:

$$\forall[(\boldsymbol{a}, \boldsymbol{b}) \in (\mathbb{X}^k, \mathbb{X}^k)]\ \ \langle f[\boldsymbol{a}], f[\boldsymbol{b}]\rangle \in \mathcal{I}_f(\mathbb{X})^{\between}.$$

These number systems are not used directly. They are used in the construction of other number systems. Some research has shown that such number systems may be used, to simplify interval arithmetic proofs, and to align interval arithmetic with more established mathematics [34].

The number system $\mathcal{I}({\mathbb{Y}_a}^{\between} | f^{|}(\mathbb{Y}_b))$ extends the number system $\mathbb{Y}_a$ by allowing collapsing intervals. Each interval $j \in \mathcal{I}({\mathbb{Y}_a}^{\between} | f^{|}(\mathbb{Y}_b))$ can be described by two intervals, $v \in {\mathbb{Y}_a}^{\between}$ and $d \in \mathbb{Y}_b$:

$$\forall[j \in \mathcal{I}({\mathbb{Y}_a}^{\between} | f^{|}(\mathbb{Y}_b))]\ \exists[(v, d) \in ({\mathbb{Y}_a}^{\between}, \mathbb{Y}_b)]\ j = \langle v | f^{|}(d)\rangle.$$

Parameter value $\boldsymbol{\alpha}$ is in the domain of interval $\langle v | f^{|}(d)\rangle$ if the domain constraint is satisfied:

$$\boldsymbol{\alpha} \in \mathrm{dom}\langle v|f^{|}(d)\rangle \;\equiv_{\mathrm{def}}\; f^{|}(d(\boldsymbol{\alpha})).$$

The value of $\langle v|f^{|}(d)\rangle$ is given by v while the domain is given by d and $f^{|}$. At $\boldsymbol{\alpha}$ the interval is:

- not defined if $(\boldsymbol{\alpha} \in \mathrm{dom}\langle v|f^{|}(d)\rangle) = \mathrm{F}$,
- defined if $(\boldsymbol{\alpha} \in \mathrm{dom}\langle v|f^{|}(d)\rangle) = \mathrm{T}$, and
- potentially defined if $(\boldsymbol{\alpha} \in \mathrm{dom}\langle v|f^{|}(d)\rangle) = \mathbb{TF}$.

The number $x \in \mathbb{X}_a$ is contained in interval $\langle v|f^{|}(d)\rangle \in \mathcal{I}(\mathbb{Y}_a{}^{\between}|f^{|}(\mathbb{Y}_b))$, $\mathbb{Y}_a = \mathcal{I}(\mathbb{X}_a)$, for parameter value $\boldsymbol{\alpha}$ if the interval is (potentially) defined at $\boldsymbol{\alpha}$ and x lies within $v(\boldsymbol{\alpha})$:

$$[x \in \langle v|f^{|}(d)\rangle(\boldsymbol{\alpha})] \;\Leftrightarrow\; \left[(\boldsymbol{\alpha} \in \mathrm{dom}\langle v|f^{|}(d)\rangle)^{\mathbb{B}+} \wedge (x \in v(\boldsymbol{\alpha}))\right].$$

Since it is common to use the same number system as the basis for both the value and domain, there is an abbreviated syntax:

$$\mathcal{I}_f{}^{|f^{|}}(\mathbb{Y}) \;\equiv_{\mathrm{def}}\; \mathcal{I}_f(\mathbb{Y}|f^{|}) \;\equiv_{\mathrm{def}}\; \mathcal{I}(\mathcal{I}_f(\mathbb{Y})^{\between}|f^{|}(\mathcal{I}_f(\mathbb{Y}))).$$

2.11.1 Interval Inclusion

The number system $\mathbb{Y}_c = \mathcal{I}(\mathbb{Y}_a{}^{\between}|f^{|}(\mathbb{Y}_b))$ will be used in the definitions that follow. A model $g^{\mathbb{Y}_c}$ of the m-ary function g satisfies the inclusion property if, for every $\boldsymbol{j} \in \mathbb{Y}_c{}^m$:

$$\forall[\boldsymbol{\alpha} \in [0,1]^n]\; \forall[\boldsymbol{x} \in \boldsymbol{j}(\boldsymbol{\alpha})]\;\; g(\boldsymbol{x}) \in [g^{\mathbb{Y}_c}(\boldsymbol{j})](\boldsymbol{\alpha}) \wedge (g(\boldsymbol{x}) \neq \lambda \Leftrightarrow^{\mathbb{T}} \boldsymbol{\alpha} \in \mathrm{dom}[g^{\mathbb{Y}_c}(\boldsymbol{j})])^{\mathbb{B}+},$$

The statement can be factored into two parts. The first,

$$g(\boldsymbol{x}) \in [g^{\mathbb{Y}_c}(\boldsymbol{j})](\boldsymbol{\alpha}),$$

requires that the value returned is valid; while the second,

$$(g(\boldsymbol{x}) \neq \lambda \Leftrightarrow^{\mathbb{T}} \boldsymbol{\alpha} \in \mathrm{dom}[g^{\mathbb{Y}_c}(\boldsymbol{j})])^{\mathbb{B}+},$$

requires that the domain returned is valid.

2.11.2 Interval Extension

The interval extension $g^{\mathbb{Y}_c}$ of the m-ary g is also defined in two parts:

$$g^{\mathbb{Y}_c}(\boldsymbol{j}) = \langle v|f^{|}(d)\rangle;\; v = \langle v_l^{\mathbb{Y}_a{}^{\between}-}, v_u^{\mathbb{Y}_a{}^{\between}+}\rangle,\;\; d = \langle d_l^{f^{|}(\mathbb{Y}_b)-}, d_u^{f^{|}(\mathbb{Y}_b)+}\rangle.$$

The first part,

$$\langle v_l^{\mathbb{Y}_a{}^{\between}-}, v_u^{\mathbb{Y}_a{}^{\between}+}\rangle :\; v_l(\boldsymbol{\alpha}) = \inf_{\boldsymbol{x}\in\boldsymbol{j}(\boldsymbol{\alpha})} g(\boldsymbol{x}),\;\; v_u(\boldsymbol{\alpha}) = \sup_{\boldsymbol{x}\in\boldsymbol{j}(\boldsymbol{\alpha})} g(\boldsymbol{x}),$$

defines the value of $g^{\mathbb{Y}_c}(\boldsymbol{j})$. The demotions $\mathbb{Y}_a{}^{\between}-$ and $\mathbb{Y}_a{}^{\between}+$ gracefully handle undefined domains. When $v(\boldsymbol{\alpha})$ is undefined, $v^{\mathbb{Y}_a{}^{\between}-}(\boldsymbol{\alpha})$ and $v^{\mathbb{Y}_a{}^{\between}+}(\boldsymbol{\alpha})$ may take on any value. The second part,

$$\langle d_l^{f^{|}(\mathbb{Y}_b)-}, d_u^{f^{|}(\mathbb{Y}_b)+}\rangle \;:\; d_l(\boldsymbol{\alpha}) = \inf_{\boldsymbol{x}\in\boldsymbol{j}(\boldsymbol{\alpha})} (g(\boldsymbol{x}) \neq \lambda),\;\; d_u(\boldsymbol{\alpha}) = \sup_{\boldsymbol{x}\in\boldsymbol{j}(\boldsymbol{\alpha})} (g(\boldsymbol{x}) \neq \lambda),$$

defines the domain of $g^{\mathbb{Y}_c}(\boldsymbol{j})$.

The demotions $f^{|}(\mathbb{Y}_b)-$ and $f^{|}(\mathbb{Y}_b)+$ demote arbitrary functions, which map parameters to booleans (defined/undefined), to functions which are of the form permitted by $f^{|}(\mathbb{Y}_b)$. The mapping from parameters to booleans is done in two stages: first, the parameters are mapped to extended reals and then those extended reals are mapped to booleans, via $f^{|} : \mathbb{R}^{\star} \mapsto \mathbb{B}$. The downward demotion $f^{|}(\mathbb{Y}_b)-$ must preserve domain classifications:

$$\forall[\boldsymbol{\alpha} \in [0,1]^n] \quad [\boldsymbol{\alpha} \in \mathrm{dom}(\langle v|\langle d_l, d_u\rangle\rangle)] \subseteq [\boldsymbol{\alpha} \in \mathrm{dom}(\langle v|\langle d_l^{f^{|}(\mathbb{Y}_b)-}, d_u\rangle\rangle)];$$

any valid $\mathbb{Y}_b-$ demotion operator is a valid $f^{|}(\mathbb{Y}_b)-$ demotion operator. The upward demotion $f^{|}(\mathbb{Y}_b)+$ must also preserve domain classifications:

$$\forall[\boldsymbol{\alpha} \in [0,1]^n] \quad [\boldsymbol{\alpha} \in \mathrm{dom}(\langle v|\langle d_l, d_u\rangle\rangle)] \subseteq [\boldsymbol{\alpha} \in \mathrm{dom}(\langle v|\langle d_l, d_u^{f^{|}(\mathbb{Y}_b)+}\rangle\rangle)];$$

any valid $\mathbb{Y}_b+$ demotion operator is a valid $f^{|}(\mathbb{Y}_b)+$ demotion operator.

2.11.3 Domain Descriptions

We may employ several domain description functions within a single number system. The number system $\mathcal{I}_f(\mathbb{Y}|F^{|})$, with $F^{|} = \{f_i^{|}\}$, allows an interval to describe its domain with any particular member of $F^{|}$. Interval inclusion and extension are defined as before. This is possible since the definitions rely indirectly upon $f^{|}$, via the definition of domain membership. For $\mathcal{I}_f(\mathbb{Y}|f^{|})$,

$$\boldsymbol{\alpha} \in \mathrm{dom}\langle v|f^{|}(d)\rangle \ \equiv_{\mathrm{def}} \ f^{|}(d(\boldsymbol{\alpha}));$$

while for $\mathcal{I}_f(\mathbb{Y}|F^{|})$,

$$\boldsymbol{\alpha} \in \mathrm{dom}\langle v|f_i^{|}(d)\rangle \ \equiv_{\mathrm{def}} \ f_i^{|}(d(\boldsymbol{\alpha})),$$

with $f_i^{|} \in F^{|}$. The demotion operators $f^{|}(\mathbb{Y}_b)-$ and $f^{|}(\mathbb{Y}_b)+$ depend upon the function chosen to describe the interval whose domain description is being demoted. A valid $\mathbb{Y}_b-$ demotion operator is a valid $f^{|}(\mathbb{Y}_b)-$ demotion operator for any $f^{|} \in F^{|}$. Similarly for $\mathbb{Y}_b+$ and $f^{|}(\mathbb{Y}_b)+$.

2.11.4 Conjunctions

A variation of $\mathcal{I}_f(\mathbb{Y}|F^{|})$ is the number system $\mathcal{I}_f{}^{\hat{|}F^{|}}(\mathbb{Y})$:

$$\mathcal{I}_f{}^{\hat{|}F^{|}}(\mathbb{Y}) \ \equiv_{\mathrm{def}} \ \mathcal{I}_f(\mathbb{Y}\hat{|}F^{|}) \ \equiv_{\mathrm{def}} \ \mathcal{I}_f(\mathbb{Y}|\bigwedge F^{|}) \ \equiv_{\mathrm{def}} \ \mathcal{I}(\mathcal{I}_f(\mathbb{Y})^{\between}|\bigwedge F^{|}(\mathcal{I}_f(\mathbb{Y}))).$$

The number system $\mathcal{I}(\mathbb{Y}_a{}^{\between}|\bigwedge F^{|}(\mathbb{Y}_b))$ generalizes the number system $\mathcal{I}(\mathbb{Y}_a{}^{\between}|F^{|}(\mathbb{Y}_b))$ by allowing a set of constraints to describe an interval's domain. Each interval $j \in \mathcal{I}(\mathbb{Y}_a{}^{\between}|\bigwedge F^{|}(\mathbb{Y}_b))$ can be described by a value $v \in \mathbb{Y}_a{}^{\between}$ and a set of domain constraints $\{f_i^{|}(d_i)\}$, with $d_i \in \mathbb{Y}_b$ and $f_i^{|} \in F^{|}$.

Parameter value $\boldsymbol{\alpha}$ is in the domain of interval $\langle v|\{f_i^{|}(d_i)\}\rangle$ if

$$\boldsymbol{\alpha} \in \mathrm{dom}\langle v|\{f_i^{|}(d_i)\}\rangle \ \equiv_{\mathrm{def}} \ \bigwedge_i f_i^{|}(d_i(\boldsymbol{\alpha})).$$

The definitions of interval inclusion and interval extension are written as before, but with the new semantics behind parameter inclusion. The demotions $\bigwedge\mathbb{Y}_b+$ and $\bigwedge\mathbb{Y}_b-$ demote from $[0,1]^n \mapsto \mathbb{B}$ to $2^{[0,1]^n \mapsto \mathbb{R}^{\star} \overset{f^{|}}{\mapsto} \mathbb{B}}$. As before, $\bigwedge\mathbb{X}_b+$ and $\bigwedge\mathbb{Y}_b-$ must preserve domain classifications.

The implementation of models is simpler in $\mathcal{I}_f(\mathbb{Y}\hat{|}F^{|})$ than in $\mathcal{I}_f(\mathbb{Y}|F^{|})$. A description of a simple implementation of a model $g^{\mathbb{I}_f{}^{\hat{|}F^{|}}}$, of function $g : \mathbb{R}^m \mapsto \mathbb{R}$, follows.

- If g is total, an evaluation of $g^{\mathbb{I}_f{}^{\hat{\rceil}F|}}(\langle \boldsymbol{v}\hat{\rceil}\{\boldsymbol{f_j^|(d_j)}\}\rangle)$ would return an interval with a domain described by $\bigwedge_{i,j} f_{i,j}^|(d_{i,j})$.

- If g is partial, an evaluation of $g^{\mathbb{I}_f{}^{\hat{\rceil}F|}}(\langle \boldsymbol{v}\hat{\rceil}\{\boldsymbol{f_j^|(d_j)}\}\rangle)$ would return an interval with a domain described by $f_D^|(D) \wedge \bigwedge_{i,j} f_{i,j}^|(d_{i,j})$, where $f_D^|(D)$ is a domain constraint introduced by $g(\langle \boldsymbol{v}|\boldsymbol{f}^|(\boldsymbol{d})\rangle)$.

2.11.5 Simplicity

A simple interval arithmetic which handles partial functions gracefully is
$\mathcal{I}(\mathbb{Y}^{\between}|\mathbb{T}) \equiv \mathcal{I}(\mathbb{Y}^{\between}|\bigwedge \mathbb{T})$:

$$\mathcal{I}^{|\mathbb{T}}(\mathbb{Y}) \ \equiv_{\text{def}} \ \mathcal{I}(\mathbb{Y}^{\between}|\mathbb{T}).$$

This is not an utter abuse of notation, as $f^|(d) \in \mathbb{T}$ for any $d \in \mathbb{Y}$: every member of $f^|(\mathbb{Y})$ may be described as a member of $\mathbb{T}$. Furthermore, any non-trivial $f^|$ warrants the full descriptive power of $\mathbb{T}$. A trivial $f^|$ may be simulated by appropriately constructing $\mathcal{I}(\mathbb{Y}^{\between}|\mathbb{T})$ models.

Consider the square root operator, denoted here as g. It is a unary partial function, undefined for negative arguments. Consider using $\mathbb{I}_k$: $g^{\mathbb{I}_k}(j)$ is undefined if j contains only negative numbers. Consider the following examples:

$$g^{\mathbb{I}_k}\langle -2, -1\rangle \rightsquigarrow \lambda,$$

$$g^{\mathbb{I}_k{}^{|\mathbb{T}}}\langle\langle -2, -1\rangle|\mathrm{T}\rangle \rightsquigarrow \langle\langle -\infty, \infty\rangle|\mathrm{F}\rangle.$$

This illustrates one approach implementations may take when confronted with partial functions. Another approach is shown in the next section. Another pair of examples follow:

$$g^{\mathbb{I}_k}\langle -1, 1\rangle \rightsquigarrow \langle 0, 1\rangle,$$

$$g^{\mathbb{I}_k{}^{|\mathbb{T}}}\langle\langle -1, -1\rangle|\mathrm{T}\rangle \rightsquigarrow \langle\langle 0, 1\rangle|\mathrm{TF}\rangle.$$

2.12 Property Tracking

Properties may be tracked as interval computations are performed. This was done in section 2.11, where the domain was tracked. The framework introduced to track domains will be generalized to track various properties.

$\mathcal{P}$ denotes the property of interest. $\mathcal{P}$ is a function of the m-ary function g being checked, the point at which the function g is being checked at, and the results of checking the property for g's arguments.

$$\mathcal{P} : [\mathbb{X}^m \mapsto \mathbb{X}] \times \mathbb{X}^m \times \mathbb{B}^m \mapsto \mathbb{B}.$$

For a property to fit into this framework, the results of checking a property directly must be equivalent to checking a property recursively. Property $\mathcal{P}$ is recursively weakly checkable if

$$\mathcal{P}(g \circ \boldsymbol{h}, \boldsymbol{x}, \boldsymbol{b}) \Rightarrow \mathcal{P}(g, \boldsymbol{h}(\boldsymbol{x}), (\mathcal{P}(\boldsymbol{h}_i, \boldsymbol{x}, \boldsymbol{b}) : i = 1 \ldots m)).$$

Property $\mathcal{P}$ is recursively strongly checkable if

$$\mathcal{P}(g \circ \boldsymbol{h}, \boldsymbol{x}, \boldsymbol{b}) \Leftrightarrow \mathcal{P}(g, \boldsymbol{h}(\boldsymbol{x}), (\mathcal{P}(\boldsymbol{h}_i, \boldsymbol{x}, \boldsymbol{b}) : i = 1 \ldots m)).$$

The vector $\boldsymbol{x}$ is the union of all the arguments of the $\boldsymbol{h}_i$ functions. Each $\boldsymbol{h}_i$ can be considered a function of $\boldsymbol{x}$. If $\boldsymbol{h}_i$ is a function of $\boldsymbol{x}_{a_1}, \boldsymbol{x}_{a_2}, \ldots, \boldsymbol{x}_{a_k}$ then consider the above $\boldsymbol{h}_i$ to be $\boldsymbol{h}'_{\boldsymbol{i}}$, defined as:

$$\boldsymbol{h}'_{\boldsymbol{i}}(\boldsymbol{x}) = \boldsymbol{h}_i(\boldsymbol{x}_{a_1}, \boldsymbol{x}_{a_2}, \ldots, \boldsymbol{x}_{a_k}).$$

All of the $\boldsymbol{h}'_{\boldsymbol{i}}$ and $g \circ \boldsymbol{h}'$ are m-ary functions.

The value $\mathcal{P}(g, \boldsymbol{x})$ can be unambiguously built up recursively, using the syntactic definition of g, since each leaf node is a 0-ary function. For all discussed properties, $\mathcal{P}(g, \boldsymbol{x}) = \mathrm{T}$ if g is a non-empty 0-ary function. This can be verified by careful scrutiny of the formal definitions.

2.12.1 Properties

Some properties of interest are:

$$\begin{aligned}
\mathcal{P}_{|}(g, \boldsymbol{x}) &= g \text{ is defined at } \boldsymbol{x}, \\
\mathcal{P}_{\Delta}(g, \boldsymbol{x}) &= g \text{ is continuous at } \boldsymbol{x}, \text{ and} \\
\mathcal{P}_{k}(g, \boldsymbol{x}) &= g \text{ is locally constant at } \boldsymbol{x}.
\end{aligned}$$

Here are formal definitions of these properties:

$$\begin{aligned}
\mathcal{P}_{|}(g, \boldsymbol{x}) &\equiv_{\text{def}} g(\boldsymbol{x}) \neq \lambda, \\
\mathcal{P}_{\Delta}(g, \boldsymbol{x}) &\equiv_{\text{def}} \lim_{\boldsymbol{y} \to \boldsymbol{x}} g(\boldsymbol{y}) \overset{\mathbb{R}}{=} g(\boldsymbol{x}), \\
\mathcal{P}_{k}(g, \boldsymbol{x}) &\equiv_{\text{def}} \exists[\delta > 0]\ \forall \boldsymbol{y} \ \ (\|\boldsymbol{y} - \boldsymbol{x}\| < \delta) \Rightarrow (g(\boldsymbol{y}) = g(\boldsymbol{x})).
\end{aligned}$$

And here are checkable definitions, which correspond to the formal definitions given above:

$$\begin{aligned}
\mathcal{P}_{|}(g, \boldsymbol{x}, \boldsymbol{b}) &\equiv_{\text{def}} [g(\boldsymbol{x}) \neq \lambda] \wedge \bigwedge_i \boldsymbol{b}_i, \\
\mathcal{P}_{\Delta}(g, \boldsymbol{x}, \boldsymbol{b}) &\equiv_{\text{def}} [\lim_{\boldsymbol{y} \to \boldsymbol{x}} g(\boldsymbol{y}) \overset{\mathbb{R}}{=} g(\boldsymbol{x})] \wedge \bigwedge_i \boldsymbol{b}_i, \\
\mathcal{P}_{k}(g, \boldsymbol{x}, \boldsymbol{b}) &\equiv_{\text{def}} \exists[\delta > 0]\ \forall \boldsymbol{y} \ \ (\|\boldsymbol{y} - \boldsymbol{x}\| < \delta) \wedge \bigwedge_i (\boldsymbol{b}_i \Rightarrow \boldsymbol{y}_i = \boldsymbol{x}_i) \ \Rightarrow \ (g(\boldsymbol{y}) = g(\boldsymbol{x})).
\end{aligned}$$

With the above checkable definitions, $\mathcal{P}_{|}$ is strongly checkable, while $\mathcal{P}_{\Delta}$ and $\mathcal{P}_{k}$ are weakly checkable.

2.12.2 Interval Inclusion and Extension

The number system $\mathbb{Y}_b$, which uses $F(\mathbb{Y}_a)$ to track $\mathcal{P}$, will be used in the definitions that follow. The value of $\mathcal{P}$ for $j \in \mathbb{Y}_b$ is specified by prop$[j]$. For $\mathbb{Y}_b = \mathcal{I}_f{}^{|\mathbb{T}}(\mathbb{Y})$, $F(\mathbb{Y}_a) = \mathbb{T}$ and prop$[j]$ = dom$[j]$. When an interval has several properties, a specific property may be referenced by giving its label: prop$_{\Delta}[j]$ references the continuity property, for example.

A model $g^{\mathbb{Y}_b}$ of the m-ary function g satisfies the $\mathcal{P}$ inclusion property if for all appropriate $\boldsymbol{j}$, $\boldsymbol{\alpha}$, and $\boldsymbol{x}$:

$$\mathcal{P}(g, \boldsymbol{x}, (\text{prop}[\boldsymbol{j}_i](\boldsymbol{\alpha}) : i = 1 \ldots m)) \Leftrightarrow^{\mathbb{T}} \text{prop}[g^{\mathbb{Y}_b}(\boldsymbol{j})](\boldsymbol{\alpha}).$$

To fully satisfy the inclusion property, the model $g^{\mathbb{Y}_b}$ must also satisfy the value inclusion property. The value inclusion property is defined as before.

The interval $\mathcal{P}$ extension of the m-ary function g is defined as:

$$\mathrm{prop}[g^{\mathbb{Y}_b}(\boldsymbol{j})](\boldsymbol{\alpha}) = f(\langle p_l^{f(\mathbb{Y}\cdot a)-}, p_u^{f(\mathbb{Y}_a)+}\rangle(\boldsymbol{\alpha})),$$

where p_l and p_u are:

$$p_l(\boldsymbol{\alpha}) = \inf_{\boldsymbol{x}\in\boldsymbol{j}(\boldsymbol{\alpha})} p(\boldsymbol{x}),\ p_u(\boldsymbol{\alpha}) = \sup_{\boldsymbol{x}\in\boldsymbol{j}(\boldsymbol{\alpha})} p(\boldsymbol{x});\ p(\boldsymbol{x}) = \mathcal{P}(g, \boldsymbol{x}, (\mathrm{prop}[\boldsymbol{x}_i] : i = 1\ldots m)).$$

The interval value extension is defined as before. The interval extension $g^{\mathbb{X}_b}$ of g is both the interval value extension and the interval $\mathcal{P}$ extension.

2.12.3 Systems

For each property there is a compact syntax which states that the constructed number system tracks that property.

- The system $\mathcal{I}(\mathbb{Y}_a|F^|(\mathbb{Y}_b))$ enhances $\mathbb{Y}_a$ by tracking $\mathcal{P}_|$ with $F^|(\mathbb{Y}_b)$.
- The system $\mathcal{I}(\mathbb{Y}_a\Delta F^\Delta(\mathbb{Y}_b))$ enhances $\mathbb{Y}_a$ by tracking $\mathcal{P}_\Delta$ with $F^\Delta(\mathbb{Y}_b)$.
- The system $\mathcal{I}(\mathbb{Y}_a k F^k(\mathbb{Y}_b))$ enhances $\mathbb{Y}_a$ by tracking $\mathcal{P}_k$ with $F^k(\mathbb{Y}_b)$.

The description can be placed above the number system, as was done previously:

$$\mathcal{I}_f^{\mathrm{p}F^{\mathrm{p}}}(\mathbb{Y}) \ \equiv_{\mathrm{def}}\ \mathcal{I}(\mathcal{I}_f(\mathbb{Y})\mathrm{p}F^{\mathrm{p}}(\mathcal{I}_f(\mathbb{Y}))).$$

The exact number system used to track properties can also be specified:

$$\mathcal{I}_f^{\mathrm{p}F^{\mathrm{p}}(\mathbb{Y}_a)}(\mathbb{Y}_b) \ \equiv_{\mathrm{def}}\ \mathcal{I}(\mathcal{I}_f(\mathbb{Y}_b)\mathrm{p}F^{\mathrm{p}}(\mathbb{Y}_a)).$$

Conjuctive descriptions are also described as before:

$$\mathcal{I}_f^{\hat{\mathrm{p}}F^{\mathrm{p}}(\mathbb{Y}_a)}(\mathbb{Y}_b) \ \equiv_{\mathrm{def}}\ \mathcal{I}(\mathcal{I}_f(\mathbb{Y}_b)\mathrm{p}\bigwedge F^{\mathrm{p}}(\mathbb{Y}_a)).$$

Properties may be combined if they share the same underlying number system:

$$\mathcal{I}_f^{\mathrm{pq}F^{\mathrm{pq}}(\mathbb{Y})} \ \equiv_{\mathrm{def}}\ \mathcal{I}_f^{\mathrm{p}F^{\mathrm{pq}}(\mathbb{Y})\mathrm{q}F^{\mathrm{pq}}(\mathbb{Y})}.$$

A similar notation is used in describing intervals; for example, $\langle j\Delta d\rangle$ is an interval with value j, whose continuity is described by d.

In the next subsection, interval sets will be introduced. For any interval arithmetic $\mathbb{Y}$, there is an associated interval set arithmetic $\mathbb{Y}^\star$. $\mathbb{I}^{|\Delta\mathbb{T}\star} = \mathcal{I}^\star(\mathcal{I}(\mathbb{F})|\mathbb{T}\Delta\mathbb{T})$ and ${\mathbb{L}_k}^{\hat{|}\hat{\Delta}F\star} = \mathcal{I}^\star(\mathbb{L}_k|\bigwedge F(\mathbb{L}_k)\Delta\bigwedge F(\mathbb{L}_k))$ are likely implementations, both of which have been carried out by the author, for $k = 1, 2, 3$.

2.13 Interval Sets

A further extension of interval arithmetic allows better models of "bumpy" functions. A function is $\mathbb{I}_k$-bumpy if it is partial or discontinuous. In general, a function g is $\mathbb{Y}$-bumpy if a better model exists for g in $\mathbb{Y}^\star$ than in $\mathbb{Y}$. Bumpy functions are formally defined in section 2.13.2. This extension also allows for natural models of multi-functions; multi-functions are functions that may return multiple results, such as $\pm$.

Throughout this section, let $\mathbb{Y}$ denote $\mathbb{I}_f$, for a fixed choice of f. Each number in $\mathbb{Y}^\star$ is specified as a set of numbers from $\mathbb{Y}$:

$$\forall[j \in \mathbb{Y}^\star]\ j \in 2^{\mathbb{Y}}.$$

A real number x is contained in interval j if x is contained by any member of j:

$$x \in^{\mathbb{Y}^\star} j \ \Leftrightarrow\ \exists[j_i \in j]\ x \in^{\mathbb{Y}} j_i.$$

2.13.1 Interval Inclusion and Extension

A model $g^{\mathbb{Y}^\star}$ satisfies the inclusion property for function $g : \mathbb{R}^m \mapsto \mathbb{R}$ if

$$\forall[\boldsymbol{j} \in (\mathbb{Y}^\star)^m]\ \forall[\boldsymbol{\alpha} \in [0,1]^n]\ \forall[\boldsymbol{x} \in \boldsymbol{j}(\boldsymbol{\alpha})]\ \ g(\boldsymbol{x}) \in [g^{\mathbb{Y}^\star}(\boldsymbol{j})](\boldsymbol{\alpha}).$$

Note that for $\boldsymbol{x}$ to be contained in $\boldsymbol{j}$, $\boldsymbol{x}_i$ may be in any member of $\boldsymbol{j}_i$. This follows from the definition of vector containment.

The interval extension $g^{\mathbb{Y}^\star}$ of g can be defined as before:

$$g^{\mathbb{Y}^\star}(\boldsymbol{j}) = \langle l^{\mathbb{Y}^\star -}, u^{\mathbb{Y}^\star +}\rangle\ :\ l(\boldsymbol{\alpha}) = \inf_{\boldsymbol{x} \in \boldsymbol{j}(\boldsymbol{\alpha})} g(\boldsymbol{x})\ ,\ u(\boldsymbol{\alpha}) = \sup_{\boldsymbol{x} \in \boldsymbol{j}(\boldsymbol{\alpha})} g(\boldsymbol{x}).$$

This definition hides a great deal in the seemingly innocuous demotions $\mathbb{Y}^\star-$ and $\mathbb{Y}^\star+$. Consider the case $\mathbb{Y} = \mathbb{V}^{|f|}$. Perfect demotion operators $\mathbb{V}^{|f|_\star}-$ and $\mathbb{V}^{|f|_\star}+$ can be defined, as follows:

$$(\mathbb{V}^{|f|_\star}-)(g : [0,1]^n \mapsto \mathbb{R}^\star) = \{\ \langle\langle g(\boldsymbol{\beta}), g(\boldsymbol{\beta})\rangle \mid \sum_i (\boldsymbol{\alpha}_i - \boldsymbol{\beta}_i)^2 \le 0\rangle \mid \boldsymbol{\beta} \in [0,1]^n, g(\boldsymbol{\beta}) \ne \lambda\}.$$

The perfect demotion operators describe a demoted function as an infinite collection of points (denoted above as $\boldsymbol{\beta}$). Although the associated perfect demotion operators $\mathbb{U}^{|f|_\star}-$ and $\mathbb{U}^{|f|_\star}+$ would describe a function by a finite number of intervals, the size of the descriptions would be unmanageable.

Partial functions may be handled in a natural manner, as the following examples show, for $g(x) = \sqrt{x}$:

$$g^{\mathbb{I}^{|\mathbb{T}_\star}}\{\langle\langle -2, -1\rangle|\mathrm{T}\rangle\} \rightsquigarrow \{\},$$
$$g^{\mathbb{I}^{|\mathbb{T}_\star}}\{\langle\langle -1, 1\rangle|\mathrm{T}\rangle\} \rightsquigarrow \{\langle\langle 0, 1\rangle|\mathbb{TF}\rangle\}.$$

A good model $g^{\mathbb{I}^{|\mathbb{T}_\star}}$ of $g(x) = x^{-1}$ would behave as follows:

$$g^{\mathbb{I}^{|\mathbb{T}_\star}}\{\langle\langle -1, 1\rangle|\mathrm{T}\rangle\} \rightsquigarrow \{\langle\langle -\infty, -1\rangle|\mathbb{TF}\rangle, \langle\langle 1, \infty\rangle|\mathbb{TF}\rangle\}.$$

A poor model would behave as follows:

$$g^{\mathbb{I}^{|\mathbb{T}_\star}}\{\langle\langle -1, 1\rangle|\mathrm{T}\rangle\} \rightsquigarrow \{\langle\langle -\infty, \infty\rangle|\mathbb{TF}\rangle\},$$

Interval sets may be viewed as a simplification, or a generalization, of extended interval arithmetic, as defined in [32].

2.13.2 Bumpy Functions

Formally, g is $\mathbb{Y}$-bumpy if there is a model $g^{\mathbb{Y}^\star}$ such that

$$\exists[\boldsymbol{j} \in \mathbb{Y}^n]\ \forall g^{\mathbb{Y}}\quad g^{\mathbb{Y}^\star}(\{\boldsymbol{j}\}) \subset g^{\mathbb{Y}}(\boldsymbol{j}),$$

where the inclusion operator is strict. We assume that $g^{\mathbb{Y}^\star}$ returns a finite number of intervals. Interval inclusion is defined naturally:

$$j \subseteq k\ \equiv_{\mathrm{def}}\ \forall[x \in j]\ x \in k, \qquad j \subset k\ \equiv_{\mathrm{def}}\ (j \subseteq k) \wedge (k \not\subseteq j),$$

where j and k are intervals.

2.14 Variants

There is a more explicit syntax for specifying the number system the coefficients belong to. Rather than specifying a single number system which is used for all coefficients, a particular number system can be specified for each coefficient:

$$\mathcal{I}_{f[\boldsymbol{j}]}(\boldsymbol{j}^+ \in \mathbb{X}^n, \boldsymbol{j}^- \in \mathbb{X}^n) \equiv \mathcal{I}_{f[\boldsymbol{j}]}(\mathbb{X}).$$

Some systems will require that the upper and lower coefficients are identical:

$$\mathcal{I}_{f[\boldsymbol{j}]}(\boldsymbol{j}^\pm \in \mathbb{X}^n) \equiv \mathcal{I}_{f(\boldsymbol{j})}(\boldsymbol{j}^\pm \in \mathbb{X}^n), \text{ where}$$

$$a^\pm \in \mathbb{X} \quad \equiv_{\text{def}} \quad a^+ = a^- \in \mathbb{X},$$

Such a requirement will render the presented systems impotent. The system

$$\mathcal{I}_{p+q\alpha}(p^\pm \in \mathbb{I}, q^\pm \in \mathbb{I})$$

is not impotent; it is, in fact, similar to $\mathbb{L}$.

The free variables may be specified explicitly, as follows:

$$\mathcal{I}_{f(\boldsymbol{\alpha})}(\boldsymbol{\alpha} \in [0,1]^k) \equiv \mathcal{I}_{f(\boldsymbol{\alpha})}.$$

Hansen's generalized interval arithmetic [28] can be succinctly stated as:

$$\mathbb{H}_k \quad \equiv_{\text{def}} \quad \mathcal{I}_{p+\Sigma \boldsymbol{q}_i \boldsymbol{\alpha}_i}(p^\pm \in \mathbb{I}, \boldsymbol{q}_i^\pm \in \mathbb{I}^k, \boldsymbol{\alpha}_i \in \{-c_i, c_i\}).$$

2.15 Real Representations

To perform computations with the aid of digital computers we must build the reals out of a discrete system. Mathematicians have historically built up the reals with different approaches; this section details some of these approaches. Some of these approaches lead to mechanical algorithms which may be contrasted with the interval approach. Readers interested solely in the interval approach should proceed to the next chapter.

2.15.1 Dedekind Cuts

A real number $r \geq 0$ is represented by a cut $C \subseteq \mathbb{Q}^+$, $\mathbb{Q}^+ = \mathbb{Q} \cap [0, \infty)$. Every cut has the property that for all $q \in \mathbb{Q}^+$:

$$(c \in C) \wedge (q < c) \ \Rightarrow \ q \in C.$$

As presented, the cut $\mathbb{Q}^+$ represents ∞. Disallowing this special cut gives a representation for all non-negative real numbers. In general,

$$r = \text{lub} C {=^{\mathbb{R}}} C.$$

Most numbers have a representation that cannot be written out directly since the representation is an infinite set.

Operations on reals are inherited from the corresponding operations on rationals. For example, a binary operation on two real numbers, represented by cuts X and Y, is given by:

$$X \oplus Y \equiv_{\text{def}} \{x \oplus y : x \in X, y \in Y\}.$$

Difficulties are encountered when generalizing this to negative real numbers. If a cut is simply redefined to be a subset of $\mathbb{Q}$, then the product of two cuts is not a cut if the multiplicands correspond to negative numbers.

See [8, 64] for further details concerning this representation and associated methods.

2.15.2 Cauchy Sequences

A real number r is represented by a converging sequence $\{r_0, r_1, r_2 \ldots\}$ of rational numbers:

$$r = \lim_{k \to \infty} r_k =^{\mathbb{R}} \{r_0, r_1, r_2 \ldots\}.$$

Any particular real number has many different, but equivalent, representations.

Operations are again inherited from the corresponding operations on rationals:

$$x \oplus y \equiv_{\text{def}} \{x_0 \oplus y_0, x_1 \oplus y_1, x_2 \oplus y_2, \ldots\}.$$

It must be shown that the operation results are independent of the representation chosen for the operands:

$$(a =^{\mathbb{R}} x) \wedge (b =^{\mathbb{R}} y) \Rightarrow a \oplus b =^{\mathbb{R}} x \oplus y.$$

There is no guarantee as to the rate of convergence of the sequence $\{x_0, x_1, x_2, \ldots\}$ to the represented value x.

See [8, 42, 58] for further details concerning this representation and associated methods.

2.15.3 Decimal Expansions

A real number $r \in [0, 1]$ is represented by an infinite, base b, decimal expansion $0.d_1 d_2 d_3 \ldots$:

$$r = \sum_{k=1}^{\infty} d_k b^{-k} =^{\mathbb{R}} 0.d_1 d_2 d_3 \ldots \; : \; 0 \le d_k < b.$$

With a "floating" decimal place and a sign indicator, every real number has a representation. Some numbers, such as b^{-1}, may be expanded into a finite decimal expansion. An example with $b = 10$ is:

$$0.1 =^{\mathbb{R}} 0.100 \ldots =^{\mathbb{R}} 0.0999 \ldots,$$

which also shows that such numbers have two infinite forms. Only the second form, with the infinite tail of "0" digits, is to be used. With this convention, every real number has a unique representation.

Operations are defined as operations on the infinite sums corresponding to the real numbers' decimal expansions. Addition is a simple example:

$$x + y = \sum_{k=1}^{\infty} x_k b^{-k} + \sum_{k=1}^{\infty} y_k b^{-k} = \sum_{k=1}^{\infty} (x_k + y_k) b^{-k}.$$

After rearranging the expression so that one term corresponds to each digit "carrying" must take place to ensure $d_k \in [0, b-1]$.

This representation introduces difficulties and is not commonly used as a formal definition of real numbers. See [10] for further details concerning this representation and associated methods.

2.15.4 Continued Fractions

A positive real number r is represented by a, potentially infinite, continued fraction $[r_0/r_1/r_2/\cdots/r_m]$:

$$r = r_0 + \cfrac{1}{r_1 + \frac{1}{r_2+\frac{1}{r_3+\cdots}}} =^{\mathbb{R}} [r_0/r_1/r_2/\cdots/r_m] \ : \ \begin{array}{rcl} r_0 & \geq & 0, \\ r_k & \geq & 1 \text{ if } k \in [1, m), \\ r_m & \geq & 2 \text{ if } m \in [1, \infty). \end{array}$$

The restriction on the last term, r_m, enforces a unique representation for each real number. The continued fraction is finite if and only if r corresponds to a rational number. Euclid's algorithm is used to determine the sequence of terms for a given real number. There is a unique representation for each real number, as with decimal expansions, so:

$$x = y \ \Leftrightarrow \ [x_0/x_1/x_2/\cdots/x_m] = [y_0/y_1/y_2/\cdots/y_n] \ \equiv \ (n = m) \wedge \forall [k \in [0, m]] \ x_k = y_k.$$

Operations are also handled as per decimal expansions.

See [15, 46] for further details concerning this representation and associated methods.

2.15.5 Converging Intervals

A real number r is represented by a converging sequence of rational intervals:

$$r = \lim_{k\to\infty} l_k = \lim_{k\to\infty} u_k =^{\mathbb{R}} \{\langle l_0, u_0\rangle, \langle l_1, u_1\rangle, \langle l_2, u_2\rangle, \ldots\} \ : \ [l_{k+1}, u_{k+1}] \subseteq [l_k, u_k].$$

This representation is well suited to algorithmic manipulation. All common operations are well defined using interval arithmetic, as will be demonstrated in the next chapter.

A basic number, such as π, is provided as a computer program which produces consecutive terms of a representation of that basic number. Each term of the infinite sequence is produced after a finite number of operations is performed. Numbers can be combined by using interval arithmetic on the produced streams. It can be shown that the resulting stream also converges:

$$\{\langle l_0^x, u_0^x\rangle, \langle l_1^x, u_1^x\rangle, \ldots\} \oplus^{\mathbb{R}} \{\langle l_0^y, u_0^y\rangle, \langle l_1^y, u_1^y\rangle, \ldots\} \rightsquigarrow \{\langle l_0^{x+y}, u_0^{x+y}\rangle, \langle l_1^{x+y}, u_1^{x+y}\rangle, \ldots\}$$

$$\text{with } \langle l_{k+1}^{x+y}, u_{k+1}^{x+y}\rangle \subseteq \langle l_k^{x+y}, u_k^{x+y}\rangle, \text{ where } \langle l_k^{x+y}, u_k^{x+y}\rangle = \langle l_{k+d}^x, u_{k+d}^x\rangle \oplus^{\mathcal{I}(\mathbb{Q})} \langle l_{k+d}^y, u_{k+d}^y\rangle.$$

This assumes that the expression defines a real number. Some expressions, such as $1/0$, do not define a real number. Using $\mathcal{I}(\mathbb{Q})$ will cause delays to be introduced into the system. For example, a division will not produce output until the denominator does not contain zero. After this initial delay of d terms, one term is output for each set of input terms provided (one input term for each input stream). A system with this input-output relationship is termed an on-line arithmetic system. No delay will occur using $\mathcal{I}(\mathbb{Q}^\star)$, although the produced stream may begin with $\langle -\infty, \infty\rangle$ terms.

The value of any finite expression, built with the provided operators and basic numbers, can be determined to any reasonable accuracy:

$$\exists f \ \forall x \ \forall [\epsilon > 0] \ \max(|l_{f(x,\epsilon)}^x - x|, |u_{f(x,\epsilon)}^x - x|) < \epsilon.$$

The function $f(x, \epsilon)$ is computable, as it simply computes successive terms of the representation of x until $u_k^x - l_k^x < \epsilon$, and then returns k. This contrasts strongly with the previous representations. No algorithm, using a finite number of computable atomic operations, can compute:

- the first digit of a decimal expansion of r,
- the first term of a continued fraction representation of r,

as discussed in the literature [62, 42, 12]. Note that $x =^{\mathbb{R}} y$ does not imply $f(x, \epsilon) = f(y, \epsilon)$, where x and y are two different real number representations. The remaining representations are special forms of the general converging interval representation of real numbers.

2.15.6 Redundant Decimal Expansions

As with the conventional decimal expansion representation, each real number is represented by an infinite decimal sequence $0.d_1d_2d_3\ldots$. Digits may take on negative values as well as positive values:

$$r = \sum_{k=1}^{\infty} d_k b^{-k} =^{\mathbb{R}} 0.d_1d_2d_3\ldots \ : \ -b < d_k < b.$$

This representation is used in hardware [9, 21, 31, 68], partially due to the on-line property mentioned above. The on-line property for real arithmetic using redundant decimal expansions can be specified as: the kth digit of the result is produced before the $k+1+d$th digits of the inputs are used. Circuits have been designed and built with small d. The on-line property implies that parallel addition circuits for this representation can operate without the regular carry propagation delay required by conventional decimal representations.

2.15.7 Redundant Continued Fractions

As with the conventional continued fraction representation, each real number is represented by a continued fraction. Each term may now be positive or negative:

$$r = r_0 + \cfrac{1}{r_1 + \frac{1}{r_2 + \frac{1}{r_3 + \ddots}}} =^{\mathbb{R}} [r_0/r_1/r_2/\cdots/r_m] \ : \ \begin{array}{rcl} |r_k| & \geq & 2 \text{ if } k \in [1, m), \\ r_k r_{k+1} & > & 0 \text{ if } |r_k| = 2 \text{ and } k \in [1, m), \\ r_m & \neq & -2 \text{ if } m \in [1, \infty). \end{array}$$

Negative numbers can be represented immediately. An interval variant of Euclid's algorithm is used to determine the sequence of terms for a given real number.

Software designers have used this approach for real arithmetic [69].

2.15.8 Generalized Interval Arithmetic

With the converging interval representation, generalized interval arithmetic can be used for real arithmetic. Better convergence can be realized using generalized interval arithmetic rather than constant interval arithmetic. This will be argued later, but as a trivial example consider symbolic intervals. With symbolic intervals, convergence is immediate although unproductive.

Generalized interval arithmetic provides functional bounds for expressions. These functional bounds will be exploited by the geometric algorithms presented in later chapters. Conventional real number representations do not provide information as to the effect expression parameters have on evaluated expressions.

Chapter 3

Arithmetic

This chapter is about arithmetic. Arithmetic traditionally includes procedures for integer addition, subtraction, multiplication, and division. Here we will discuss procedures for interval operations. The discussion will not be of particular operations, such as addition or multiplication. Rather, properties of common operations, such as local concavity, will drive the discussion. This will lead toward a framework for implementing generalized interval models of real functions.

3.1 Floating Point

There is a vast body of literature concerning the implementation of floating point operators [22, 47, 29]; furthertmore, there is literature describing the implementation of correctly rounded floating-point arithmetics [35, 71, 53, 55]. Details concerning the implementation of the floating point system used are not relevant to the current discussion.

We assume that $\mathbb{F}$ satisfies IEEE standard 754. The IEEE 754 standard imposes strict requirements on the rounding of the algebraic operations $+$, $-$, $\times$, $\div$, and $\sqrt{x}$. An implementation must return the nearest floating point number to the actual real result, when an operation is carried out with the rounding mode set to "round to nearest". This implies that the result is accurate to within 1/2 ULP, unless underflow or overflow occurs. Moreover, the algebraic operations must correctly round the result when the current rounding mode is "round to $-\infty$" or "round to $+\infty$"; this only requires 1 ULP accuracy.

Other operators, such as sine, are not as favoured. The IEEE 754 standard does not require $\sin^{\mathbb{F}=}$ return the nearest floating point number to the actual real result. No claims are made concerning $\sin^{\mathbb{F}-}$ or $\sin^{\mathbb{F}+}$. Some systems assume the current rounding mode is "round to nearest" when sine is computed; using another rounding mode may adversely affect the trigonometric computation.

Some brief comments will illustrate how $g^{\mathbb{F}-}$ and $g^{\mathbb{F}+}$ may be constructed from $g^{\mathbb{F}}$, for various classes of functions.

3.1.1 Exact Functions

Consider the n-ary function $g : \mathbb{R}^{\star n} \mapsto \mathbb{R}^{\star}$. The function g is exact if:

$$\forall[\boldsymbol{x} \in \mathbb{F}^n] \;\; g(\boldsymbol{x}) \in \mathbb{F};$$

this is equivalent to $g|\mathbb{F}$ being closed over $\mathbb{F}$. If g is exact then $g^{\mathbb{R}\to\mathbb{F}=} = g^{\mathbb{R}\to\mathbb{F}-} = g^{\mathbb{R}\to\mathbb{F}+}$. The functions $|x|$, x, $-x$, $\min(x, y)$, $\max(x, y)$, $\lfloor x \rfloor$, and $\lceil x \rceil$ are all exact. Allow $g^{\mathbb{F}}$ to be an exact implementation of one of the preceding exact functions. An implementation of $g^{\mathbb{F}-}$ or $g^{\mathbb{F}+}$ may simply invoke $g^{\mathbb{F}}$.

3.1.2 Constant Functions

Consider the 0-ary function $g : \mathbb{R}^0 \mapsto \mathbb{R}$. The value of g may be evaulated, once, to a high precision so that $g^{\mathbb{F}-}$ and $g^{\mathbb{F}+}$ are precisely determined. An implementation may simply return the precalculated result. Examples include:

$$\pi \approx 3.14159265358979323846 26\ldots, \quad \pi^{\mathbb{F}-} = 314 \times 10^{-2}, \quad \pi^{\mathbb{F}+} = 315 \times 10^{-2};$$

and

$$e \approx 2.71828182845904523536 0287\ldots, \quad e^{\mathbb{F}-} = 271 \times 10^{-2}, \quad e^{\mathbb{F}+} = 272 \times 10^{-2}.$$

3.1.3 Provided Functions

Consider the n-ary function $g : \mathbb{R}^{\star n} \mapsto \mathbb{R}^{\star}$, with implementations of $g^{\mathbb{F}+}$ and $g^{\mathbb{F}-}$ provided. Consider the m-ary function $h : \mathbb{R}^{\star m} \mapsto \mathbb{R}^{\star}$, along with n m-ary functions $h_i : \mathbb{R}^{\star m} \mapsto \mathbb{R}$ which are exact over $\mathbb{F}$. If

$$\forall[\boldsymbol{x} \in \mathbb{R}^{\star m}] \;\; h(\boldsymbol{x}) = g(h_1(\boldsymbol{x}), h_2(\boldsymbol{x}), \ldots, h_n(\boldsymbol{x})),$$

then

$$\forall[\boldsymbol{x} \in \mathbb{F}^m] \;\; h^{\mathbb{F}-}(\boldsymbol{x}) = g^{\mathbb{F}-}(h_1(\boldsymbol{x}), h_2(\boldsymbol{x}), \ldots, h_n(\boldsymbol{x})),$$

and

$$\forall[\boldsymbol{x} \in \mathbb{F}^m] \;\; h^{\mathbb{F}+}(\boldsymbol{x}) = g^{\mathbb{F}+}(h_1(\boldsymbol{x}), h_2(\boldsymbol{x}), \ldots, h_n(\boldsymbol{x})).$$

An implementation of $h^{\mathbb{F}-}$ invokes $g^{\mathbb{F}-}$ with arguments given by exact implementations of $h_1, h_2, \ldots, h_n$; an implementation of $h^{\mathbb{F}+}$ similarly uses $g^{\mathbb{F}+}$. Examples include:

$$\begin{array}{llll} h(x) = x^{-1}, & g(x,y) = x \div y, & h_1(x) = 1, & h_2(x) = x; \\ h(x) = x^2, & g(x,y) = x \times y, & h_1(x) = x, & h_2(x) = x. \end{array}$$

The following examples demonstrate another method for computing common constant functions:

$$\begin{array}{lll} h = \pi, & g(x) = \arccos x, & h_1 = -1; \\ h = e, & g(x) = e^x, & h_1 = 1; \\ h = \sqrt{2}, & g(x) = \sqrt{x}, & h_1 = 2. \end{array}$$

This method is usually preferable to the high precision method outlined in the previous subsection, as it is more portable and easily implemented. This method is, of course, restricted to those functions and constants which may be directly constructed from the provided operators.

3.1.4 Accurate Functions

Consider the n-ary function $g : \mathbb{R}^{\star n} \mapsto \mathbb{R}^{\star}$, along with the functions $h_- : \mathbb{F}^n \mapsto \mathbb{F}$ and $h_+ : \mathbb{F}^n \mapsto \mathbb{F}$ which overestimate the error of $g^{\mathbb{F}}$, an accurate model of $g^{\mathbb{R}^\star}$. The function h_- overestimates the amount by which $g^{\mathbb{R}^\star}$ exceeds $g^{\mathbb{F}}$ while h_+ overestimates the amount by which $g^{\mathbb{F}}$ exceeds $g^{\mathbb{R}^\star}$:

$$\forall[\boldsymbol{x} \in \mathbb{F}^n] \;\; -h_-(\boldsymbol{x}) \le g^{\mathbb{F}}(\boldsymbol{x}) - g^{\mathbb{R}^\star}(\boldsymbol{x}) \le h_+(\boldsymbol{x}).$$

A simple implementation of $g^{\mathbb{F}-}$ would proceed as follows:

$$g^{\mathbb{F}-}(\boldsymbol{x}) \rightsquigarrow g^{\mathbb{F}}(\boldsymbol{x}) -^{\mathbb{F}-} h_-(\boldsymbol{x});$$

a simple implementation of $g^{\mathbb{F}+}$ would proceed as follows:

$$g^{\mathbb{F}+}(\boldsymbol{x}) \rightsquigarrow g^{\mathbb{F}}(\boldsymbol{x}) +^{\mathbb{F}+} h_+(\boldsymbol{x}).$$

In many cases, $h_-(\boldsymbol{x})$ and $h_+(\boldsymbol{x})$ would be computed concurrently with $g(\boldsymbol{x})$ and would use partial results computed during the computation of $g(\boldsymbol{x})$. If necessary, table lookup may be used to handle infinite arguments.

As an example, consider a model $g^{\mathbb{F}}$, of $g : \mathbb{R}^{\star n} \mapsto \mathbb{R}^{\star}$, which guarantees:

$$\forall [\boldsymbol{x} \in \mathbb{F}^n] \;\; |g^{\mathbb{F}}(\boldsymbol{x}) - g^{\mathbb{R}^\star}(\boldsymbol{x})| = \min_{y \in \mathbb{F}} |y - g^{\mathbb{R}^\star}(\boldsymbol{x})|,$$

which is equivalent to stating that no floating point number is closer to $g^{\mathbb{R}^\star}(\boldsymbol{x})$ than $g^{\mathbb{F}}(\boldsymbol{x})$. The functions

$$h_-(\boldsymbol{x}) = g^{\mathbb{F}}(\boldsymbol{x}) -^{\mathbb{F}+} \operatorname{pred}(g^{\mathbb{F}}(\boldsymbol{x})) \quad \text{and} \quad h_+(\boldsymbol{x}) = \operatorname{succ}(g^{\mathbb{F}}(\boldsymbol{x})) -^{\mathbb{F}+} g^{\mathbb{F}}(\boldsymbol{x})$$

correctly overestimate the error of $g^{\mathbb{F}}$, where $\operatorname{pred}(x)$ and $\operatorname{succ}(x)$ give the floating-point number immediately preceding and succeeding x, respectively:

$$\operatorname{pred}(x) = (x - \Delta)^{\mathbb{F}-}, \qquad \operatorname{succ}(x) = (x + \Delta)^{\mathbb{F}+}.$$

Using the preceding error overestimates, a simple implementation would proceed as follows:

$$g^{\mathbb{F}-}(\boldsymbol{x}) \rightsquigarrow \operatorname{pred}(g^{\mathbb{F}}(\boldsymbol{x})), \qquad g^{\mathbb{F}+}(\boldsymbol{x}) \rightsquigarrow \operatorname{succ}(g^{\mathbb{F}}(\boldsymbol{x})).$$

3.1.5 Argument Reduction

The accuracy of $g^{\mathbb{F}}$ may only be known over a restricted domain. Consider the sine function, which may be approximated by a finite polynomial:

$$\sin(x) \approx x - \frac{x^3}{3!} + \frac{x^5}{5!} \approx x - 167 \times 10^{-3} x^3 + 833 \times 10^{-5} x^5 = \sin^{\mathbb{F}}(x).$$

Although there are better ways of approximating the sine function, the McLaurin polynomial above will duly serve our purposes.

The $\sin^{\mathbb{F}}$ function is reasonably accurate for $x \in (-\pi, \pi)$ but it is not accurate for large angles. Large arguments are reduced by exploiting the trigonometric identity

$$\sin(x + 2\pi n) = \sin(x),$$

as follows:

$$\sin^{\mathbb{F}}(x) \rightsquigarrow \sin^{\mathbb{F}}(2\pi \operatorname{fract}(\tfrac{1}{2}\pi^{-1}x - \tfrac{1}{2}) - \pi).$$

The usual method for accurately computing the reduced argument employs high precision floating point arithmetic with an accurate representation of π^{-1}. Such an approach would use $\mathbb{F}[10, 19, -9 \ldots 9]$ to compute an accurate argument reduction when computing sine for $\mathbb{F}[10, 3, -9 \ldots 9]$. Some systems do not compute highly accurate reduced arguments hence it may be difficult to accurately estimate the error of $\sin^{\mathbb{F}}(x)$ for large x. Conventional argument reduction is discussed in [22, 47].

Another approach is to compute the argument reduction using interval arithmetic, and invoke the provided $\sin^{\mathbb{F}}$ function for small angles, where the error is known. First, the reduced angle is bounded:

$$\begin{array}{rrl} & b & \leftarrow & -\pi + 2\pi \text{ fract}^{\mathbb{I}}(\frac{1}{2}\pi^{-1}x - \frac{1}{2}) \\ \rightsquigarrow & b & \leftarrow & \langle -315\times 10^{-2}, -314\times 10^{-2}\rangle +^{\mathbb{I}} \langle 628\times 10^{-2}, 629\times 10^{-2}\rangle \times^{\mathbb{I}} \\ & & & \text{fract}^{\mathbb{I}}(\langle 159\times 10^{-3}, 160\times 10^{-3}\rangle \times^{\mathbb{I}} x -^{\mathbb{I}} \langle 500\times 10^{-3}, 500\times 10^{-3}\rangle). \end{array}$$

Then, a suitable $x' \in b$ is chosen and the provided $\sin^{\mathbb{F}}$ function is invoked with x' as an argument. Since $\sin^{\mathbb{F}}$ is an accurate function over b, the previous subsection on accurate functions details how $\sin^{\mathbb{F}-}(x')$ is constructed from $\sin^{\mathbb{F}}(x')$. An upper bound on $\sin(x)$ is similarly constructed from $\sin^{\mathbb{F}}(x')$ with $x' \in b$. Which $x' \in b$ is chosen depends on whether a lower or upper bound is desired. Sections 3.2.10 and 3.2.11 describes how x' is chosen.

3.1.6 Basic Methods

When little information is available on the provided functions, or the provided functions are unsatisfactory, rigorous upper and lower bounds may be computed by resorting to basic methods [22, 47, 29]. The methods employed will depend on the function to be computed, but usually a method for computing $g^{\mathbb{F}}$ can be adapted to compute $g^{\mathbb{F}-}$ or $g^{\mathbb{F}+}$.

With a thorough understanding of the method used to compute $g^{\mathbb{F}}$, an efficient, similar method may be used to compute $g^{\mathbb{F}-}$ or $g^{\mathbb{F}+}$. Let us consider the sine function:

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{\xi^7}{7!} \text{ with } \xi \in [0, x] \text{ for } x \in [0, \tfrac{1}{2}\pi).$$

Argument reduction may be used, as before, to reduce large angles. Since $\xi \in [0, x] \subseteq [0, \frac{1}{2}\pi)$,

$$x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{\pi^7}{2^7 7!} \leq \sin(x) \leq x - \frac{x^3}{3!} + \frac{x^5}{5!}.$$

Since $x \geq 0$ we may infer that x, x^3, and x^5 are all non-negative. We may now bound $\sin(x)$ using the provided floating point operations, as follows:

$$x^{\mathbb{F}-} -^{\mathbb{F}-} \left(\frac{x^3}{3!}\right)^{\mathbb{F}+} +^{\mathbb{F}-} \left(\frac{x^5}{5!}\right)^{\mathbb{F}-} -^{\mathbb{F}-} \left(\frac{\pi^7}{2^7 7!}\right)^{\mathbb{F}+} \leq \sin(x) \leq x^{\mathbb{F}+} -^{\mathbb{F}+} \left(\frac{x^3}{3!}\right)^{\mathbb{F}-} +^{\mathbb{F}+} \left(\frac{x^5}{5!}\right)^{\mathbb{F}+},$$

with evaluation proceeding from left to right in both cases. An implementation of $\sin^{\mathbb{F}-}$ and $\sin^{\mathbb{F}+}$ is now clear, for $x \in [0, \frac{1}{2}\pi)$. Using similar bounds on $\sin(x)$ for $x \in (-\frac{1}{2}\pi, 0]$, $\sin^{\mathbb{F}-}$ and $\sin^{\mathbb{F}+}$ may be evaluated for $x \in (-\frac{1}{2}\pi, \frac{1}{2}\pi)$. A similar implementation of the cosine function for $x \in (-\frac{1}{2}\pi, \frac{1}{2}\pi)$ along with appropriate argument reduction allows $\sin^{\mathbb{F}-}$, $\sin^{\mathbb{F}+}$, $\cos^{\mathbb{F}-}$, and $\cos^{\mathbb{F}+}$ to be evaluated for all finite floating-point numbers x. Infinite arguments may be handled by table lookup. The interval based argument reduction presented in the last subsection will correctly handle infinite arguments without table lookup.

Even with limited understanding of the method used to compute $g^{\mathbb{F}}$, both $g^{\mathbb{F}-}$ and $g^{\mathbb{F}+}$ may be implemented. Again, we consider the sine function:

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{\xi^7}{7!} \quad \text{with } \xi \in [0, x] \text{ for } x \in [0, \tfrac{1}{2}\pi).$$

Evaluating the approximation formula with an interval arithmetic provides strict bounds on $\sin(x)$:

$$\sin(x) \in \langle l, u\rangle = \langle x, x\rangle - \langle x, x\rangle^3 \div 3!^{\mathbb{I}} + \langle x, x\rangle^5 \div 5!^{\mathbb{I}} - \langle 0, x\rangle^7 \div 7!^{\mathbb{I}} \quad \text{for } x \in [0, \tfrac{1}{2}\pi),\ x \in \mathbb{F}.$$

An implementation of $\sin^{\mathbb{F}-}$ would compute the interval $\langle l, u\rangle$, as shown above, and return l. An implementation of $\sin^{\mathbb{F}+}$ would return u. The implementation may be extended, as the first was, to allow $\sin^{\mathbb{F}-}$ and $\sin^{\mathbb{F}+}$ to be computed for arbitrary arguments.

Although the two methods initially appear to be distinct, the first implementation of $\sin^{\mathbb{F}+}$ is simply a cleverly optimized version of the second implementation of $\sin^{\mathbb{F}+}$. Further optimization is possible. For example, one may precompute constants so that division operations may be replaced with multiplication operations:

$$\langle x, x\rangle \div 3!^{\mathbb{I}} \rightsquigarrow \langle x, x\rangle \times \langle (3!^{-1})^{\mathbb{R}\to\mathbb{F}-}, (3!^{-1})^{\mathbb{R}\to\mathbb{F}+}\rangle \rightsquigarrow \langle x, x\rangle \times \langle 166 \times 10^{-2}, 167 \times 10^{-2}\rangle,$$

which mildly reduces the accuracy. In general, interval methods allow one to build $g^{\mathbb{F}-}$ and $g^{\mathbb{F}+}$ from ${h_i}^{\mathbb{F}-}$ and ${h_i}^{\mathbb{F}+}$, using the method of computing $g^{\mathbb{F}}$ from ${h_i}^{\mathbb{F}}$ as a guide. Knowledge of $g^{\mathbb{R}}$ and ${h_i}^{\mathbb{R}}$ can help produce good implementations of $g^{\mathbb{F}-}$ and $g^{\mathbb{F}+}$.

3.2 Constant Interval Arithmetic

Let $\mathbb{Y}$ denote a constant interval number system, built from an underlying number system $\mathbb{X}$:

$$\mathbb{Y} = \mathcal{I}(\mathbb{X}).$$

Some candidates for $\mathbb{X}$, which will allow machine implementations of $\mathbb{Y}$, are the fixed-point, floating-point, fixed-slash, and floating-slash number systems [51]. The previous section discussed implementation details when $\mathbb{X} = \mathbb{F}$, although the comments made are relevant for the other possibilities of $\mathbb{X}$. We no longer consider details pertaining to the choice of $\mathbb{X}$.

A general methodology for constructing constant interval models of real functions will be presented in this section. We will assume that an order-preserving mapping $\phi_{\mathbb{X}\mathbb{R}^\star}$ exists:

$$\forall[(x, y) \in \mathbb{X}^2]\ \ x <^{\mathbb{X}} y \ \Rightarrow\ \phi_{\mathbb{X}\mathbb{R}^\star}(x) <^{\mathbb{R}^\star} \phi_{\mathbb{X}\mathbb{R}^\star}(y),$$

which allows us to focus on the case $\mathbb{X} = \mathbb{R}^\star$. We identify the number $x \in \mathbb{X}$ with the extended real number $\phi_{\mathbb{X}\mathbb{R}^\star}(x)$. This mapping need not be the obvious one. The construction of $g^{\mathbb{J}}$ from $g^{\mathbb{R}^\star} : \mathbb{R}^{\star n} \mapsto \mathbb{R}^\star$ will determine the construction of $g^{\mathbb{Y}}$ from $g^{\mathbb{X}-}, g^{\mathbb{X}+}$:

$$g^{\mathbb{J}}(\boldsymbol{j}) = \langle l_g(\boldsymbol{j}), u_g(\boldsymbol{j})\rangle \ \Rightarrow\ g^{\mathbb{Y}}(\boldsymbol{j}) = \langle l_g^{\mathbb{X}-}(\boldsymbol{j}), u_g^{\mathbb{X}+}(\boldsymbol{j})\rangle.$$

The endpoints $l_g(\boldsymbol{j})$ and $u_g(\boldsymbol{j})$ are procedures which evaluate g at point(s) $\boldsymbol{x} \in \boldsymbol{j}$ and return an endpoint based on those evaluations. In the first case, where $g^{\mathbb{J}}(\boldsymbol{j})$ is computed, $\boldsymbol{j} \in \mathbb{J}^n$ and $\boldsymbol{x} \in \mathbb{R}^{\star n}$; the evaluations of g are invocations of $g^{\mathbb{R}^\star}$. In the second case, where $g^{\mathbb{Y}}(\boldsymbol{j})$ is computed, $\boldsymbol{j} \in \mathbb{Y}^n$ and $\boldsymbol{x} \in \mathbb{X}^n$; the evaluations of g are invocations of $g^{\mathbb{X}-}$ or $g^{\mathbb{X}+}$. The same algorithm may be used for l_g (and u_g) in both cases.

Throughout this section we may treat members of $\mathbb{J}$ as constant functions, to ease the upcoming transition to linear interval arithmetic. Rather than describe the procedures l_g and u_g in a formal language, we will discuss evaluations of $g^{\mathbb{J}}(j)$ with examples. It is understood that much of the examination of g occurs while $g^{\mathbb{J}}$ is being implemented, rather than during execution. Of course, such examination is possible during execution, and may be useful for complicated functions; interval arithmetic may be used to help perform such examinations. Complicated functions may be handled without direct analysis; the interval inclusion property allows such functions to be treated as compositions of simpler functions.

Knowledge of basic vector calculus is assumed; see [48] for reference. See, for example, [19, 27] for other approaches to the implementation of constant interval arithmetic.

3.2.1 Constant Functions

Consider the constant, total function $g : \mathbb{R}^\star \mapsto \mathbb{R}^\star$. An optimal interval model $g^{\mathbb{J}}$ is obvious:

$$g^{\mathbb{J}}(j) \rightsquigarrow \langle g, g \rangle.$$

A simple theory, which defends this model, will be presented immediately. This theory will be subsequently extended in later sections, to defend other interval models.

3.2.2 Interpolating Polynomials

Given the set $G = \{(x_0, y_0), (x_1, y_1)\}$, consider the two functions $\varphi^G_{0,1} : \mathbb{R} \mapsto \mathbb{R}$ and $\varphi^G_{1,1} : \mathbb{R} \mapsto \mathbb{R}$, defined as follows:

$$\varphi^G_{0,1}(x) = \frac{x - x_1}{x_0 - x_1}, \quad \varphi^G_{1,1}(x) = \frac{x - x_0}{x_1 - x_0};$$

$\varphi^G_{i,d}$ is a d-degree polynomial with

$$\varphi^G_{i,d}(x_j) = \delta_{ij} \ \equiv_{\text{def}} \begin{cases} 0, & \text{if } i \neq j, \\ 1, & \text{if } i = j. \end{cases}$$

The above defines δ_{ij}, the Kronecker delta. The set G represents the function $g : \mathbb{R}^\star \mapsto \mathbb{R}^\star$ using two distinct elements of g:

$$G \subseteq_2 g, \quad \text{where } x \subseteq_k y \ \equiv_{\text{def}} \ (x \subseteq y) \wedge (|x| = k);$$

we here envision the unary function g as a set, as defined in section 2.1. From this, we may deduce that G is also a function, and that $x_0 \neq x_1$. It follows that the functions $\varphi^G_{0,1}$ and $\varphi^G_{1,1}$ are well defined, for our choice of G. Since $\varphi^G_{i,1}(x_j) = \delta_{ij}$, the function $L_G : \mathbb{R} \mapsto \mathbb{R}$,

$$L_G(x) = y_0 \varphi^G_{0,1} + y_1 \varphi^G_{1,1},$$

interpolates G:

$$L_G(x_i) = y_i.$$

L_G is the linear Lagrange interpolating polynomial of G.

L_G may be expressed in standard polynomial form:

$$L_G(x) = \psi^G_{1,1} x + \psi^G_{0,1}, \quad \psi^G_{1,1} = \frac{y_0}{x_0 - x_1} + \frac{y_1}{x_1 - x_0}, \ \psi^G_{0,1} = \frac{-x_1 y_0}{x_0 - x_1} + \frac{-x_0 y_1}{x_1 - x_0};$$

$\psi^G_{i,d}$ is the coefficient of x^i in $L_G(x)$, a d-degree polynomial. The leading coefficient, $\psi^G_{1,1}$, is of special interest, and may be denoted simply by ψ^G_1:

$$\psi^G_1 = \psi^G_{1,1}.$$

The set G, and the associated polynomial L_G, are:

- monotonically decreasing if $\psi_1^{\downarrow}(G)$,
- constant if $\psi_1^{0}(G)$, and
- monotonically increasing if $\psi_1^{\uparrow}(G)$;

where:

$$\psi_1^{\downarrow}(G) \ \equiv_{\text{def}} \ (\psi_1^G \leq 0), \quad \psi_1^{0}(G) \ \equiv_{\text{def}} \ (\psi_1^G = 0), \quad \psi_1^{\uparrow}(G) \ \equiv_{\text{def}} \ (\psi_1^G \geq 0).$$

Consider $G^\star$, a richer representation of g,

$$G^\star \subseteq_{\geq 2} g.$$

The representation $G^\star$ has one of the preceding properties if all two-member subsets of $G^\star$ have the same property:

$$\psi_1^{\chi}(G^\star) \ \equiv_{\text{def}} \ \forall[G \subseteq_2 G^\star] \ \psi_1^{\chi}(G), \qquad \text{where } \chi \in \mathbb{O} = \{\downarrow, 0, \uparrow\}.$$

All three properties are considered to be satisfied by sparse representations of g since

$$\forall[G \subseteq_{<2} g] \ \forall[\psi_1' \in \mathbb{R}] \ \exists[\psi_0' \in \mathbb{R}] \ \forall[(x_i, y_i) \in G] \ L_G'(x_i) = y_i,$$

where $L_G'(x) = \psi_1' x + \psi_0'$. For $G = g$, the usual definitions of constancy and monotonicity are equivalent to those given here. Let $\psi_1^{\updownarrow}(G^\star)$ state that $G^\star$ has one of the above properties:

$$\psi_1^{\updownarrow}(G^\star) \ \equiv_{\text{def}} \ \exists(\chi \in \mathbb{O}) \ \psi_1^{\chi}(G^\star).$$

For all representations $G \subseteq g$,

$$\psi_1^{0}(G) \ \Leftrightarrow \ \psi_1^{\downarrow}(G) \wedge \psi_1^{\uparrow}(G).$$

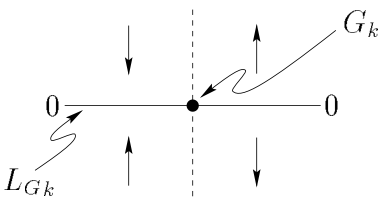
The Lagrange interpolating polynomial L_G for $G = \{(x_0, y_0)\} \subseteq_1 g$ is defined as follows:

$$L_G(x) = y_0.$$

Using the constant and linear interpolating polynomials we will construct constant bounds for many common functions.

3.2.3 ψ_1 Charts

Consider the following chart:



ψ_1 *Chart*

The ψ_1 chart is used to predict the sign of ψ_1^G, for $G = G_k \cup \{(x, y)\}$, given G_k. The chart divides $\mathbb{R}^2$ into six disjoint regions, as listed below.

- The forbidden region, indicated above by a dashed line. The point (x, y) may not reside in the forbidden region, since G is a function. The remaining five regions are each labelled with a member of $\mathbb{O}$.
- The zero region, indicated above by a solid line, and labelled with 0.
- Two up regions, each labelled with $\uparrow$.
- Two down regions, each labelled with $\downarrow$.

If the point (x, y) resides in a region labelled with $\chi \in \mathbb{O}$, then $\psi_1^\chi(G)$.

The rules for constructing a ψ_k chart are as follows:

1. The forbidden region, where $y \in \mathrm{dom}(G_k)$, is indicated by a dashed line.
2. The zero region, where $(x, y) \in L_{G_k}$, is labelled with 0. The zero region, along with the forbidden region divide the remainder of $\mathbb{R}^2$ into a checkerboard of regions.
3. The upper right region is an up region, labelled with $\uparrow$.
4. The remaining regions are up and down regions, labelled with $\uparrow$ and $\downarrow$ in checkerboard fashion, as shown above.

These rules work for any ψ_k chart, and are defended in section 3.4.2. There is a special case; when $(x, y) \in G_k$ the sign of ψ_1^G is arbitrary. This is forbidden with the above rules, since $(x, y) \in G_k \Rightarrow y \in \mathrm{dom}(G_k)$. These undercontrained cases are not important to us.

3.2.4 Constant Functions

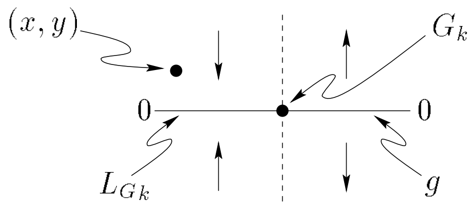
Consider the constant function $g : \mathbb{R} \mapsto \mathbb{R}$. Since g is constant, $\psi_1^0(g)$. Take any $G_k \subseteq_1 g$; a simple proof by contradiction, which follows, shows that L_{G_k} is an exact bound of g:

$$\forall[(x, y) \in g]\ L_{G_k}(x) = y.$$

Assume there is a point $(x, y) \in g$ such that $L_{G_k}(x) \neq y$. Let $G = G_k \cup \{(x, y)\}$, so $G \subseteq_2 g$:

$$\begin{aligned} L_{G_k}(x) \neq y \quad &\Rightarrow \quad (x, y) \notin G_k; \\ (x, y) \in g,\ (x, y) \notin G_k,\ G_k \subseteq_1 g \quad &\Rightarrow \quad G \subseteq_2 g. \end{aligned}$$

Furthermore, $G \subseteq g$ and $\psi_1^0(g)$ together imply that $\psi_1^0(G)$.



ψ_1 *Chart, with impossible* (x, y)

A quick review of the ψ_1 chart reveals this situation is impossible, since $\psi_1^0(G)$ implies that $(x, y) \in L_{G_k}$. The ψ_1 chart predicts the sign of ψ_1^G since $G = G_k \cup \{(x, y)\}$.

So, for any $G_k \subseteq_1 g$, $g \subseteq L_{G_k}$. It follows that L_{G_k} is a lower and upper bound for g, over $j \in \mathbb{J}$:

$$\forall[(x, y) \in g]\ L_{G_k}(x) = y \ \Rightarrow\ \forall[(x, y) \in g]\ L_{G_k}(x) \leq y \leq L_{G_k}(x).$$

The intuitive interval model originally given is now seen to be correct:

$$g^{\mathbb{J}}(j) \rightsquigarrow \langle g, g \rangle,$$

since it is equivalent to:

$$g^{\mathbb{J}}(j) \rightsquigarrow \langle L_{G_k}, L_{G_k} \rangle \text{ for any } G_k \subseteq_1 g,$$

assuming that g is total.

3.2.5 Optimality

Consider the interval model $g^{\mathbb{J}} : \mathbb{J} \mapsto \mathbb{J}$ of the function $g : \mathbb{R}^\star \mapsto \mathbb{R}^\star$. The function g has many interval models; we will now define when the model $g^{\mathbb{J}}$ is optimal.

A bound ${g_\star}^{\mathbb{J}+}$ is optimal, for interval arithmetic, if no better upper bound exists:

$$\text{optimal}^+({g_\star}^{\mathbb{J}+}, g) \ \equiv_{\text{def}}\ \forall g^{\mathbb{J}+}\ \ {g_\star}^{\mathbb{J}+} \leq g^{\mathbb{J}+}.$$

The model ${g_\star}^{\mathbb{J}}$ returns optimal upper bounds if the upper bound is optimal for all $j \in \mathbb{J}$:

$$\text{optimal}^+({g_\star}^{\mathbb{J}}, g) \ \equiv_{\text{def}}\ \forall[j \in \mathbb{J}]\ \text{optimal}^+({g_\star}^{\mathbb{J}}(j)^+, g(j^+)).$$

We may now prove that the interval extension of g is optimal. Consider the upper bound, for argument j:

$$g^{\mathbb{J}}(j)^+ = \sup_{x \in j} g(x),$$

from the definition of interval extension. The only way $g^{\mathbb{J}}(j)^+$ could fail to be optimal is for there to be a better bound of $g(j)$. This contradicts the definition of supremum; let the better bound be denoted as l,

$$\forall[x \in j]\ \ g(x) \leq l < g^{\mathbb{J}}(j)^+,$$

or, equivalently:

$$\forall[x \in j]\ \ g(x) \leq l < s, \quad s = \sup_{x \in j} g(x),$$

but:

$$s = \sup_{x \in j} g(x) \ \equiv\ \neg\exists[l \in \mathbb{R}^\star, l < s]\ \forall[x \in j]\ g(x) \leq l.$$

We now know that if g is differentiable over $j^\square$, then the upper bound given by $g^{\mathbb{J}}$ is obtained by $g(x_u)$ for some x_u in j:

$$\text{optimal}^+(g^{\mathbb{J}}(j), g) \ \Leftrightarrow\ \exists[x_u \in j]\ g(x_u) = g^{\mathbb{J}}(j)^+,$$

since j is closed. Lower bounds are handled similarly, and will be addressed in section 3.2.12.

Optimality can be defined without direct reference to the underlying function:

$$\text{optimal}^+(g^{\mathbb{J}}) \ \equiv_{\text{def}}\ \text{optimal}^+(g^{\mathbb{J}}, g_-), \quad g_-(x) = g^{\mathbb{J}}(\langle x, x \rangle)^-.$$

It is clear that $\text{optimal}^+(g^{\mathbb{J}}) \equiv \text{optimal}^+(g^{\mathbb{J}}, g|\text{dom}(g_-))$, since if $g(x) \neq g_-(x)$ then $g^{\mathbb{J}}(\langle x, x \rangle)$ is clearly not optimal, so $g^{\mathbb{J}}$ is not optimal. If $g^{\mathbb{J}}$ is valid, then $\text{dom}(g) \subseteq \text{dom}(g_-)$.

3.2.6 Piecewise Models

Any function $g : \mathbb{R}^\star \mapsto \mathbb{R}^\star$ may be cut into sections where each section fits into one monotonicity class:

$$\Xi_1(g) \;\equiv_{\text{def}}\; \{D : \psi_1^\ddagger(g|D), D \subseteq \mathbb{R}^\star\}.$$

A model of a function $g : \mathbb{R}^\star \mapsto \mathbb{R}^\star$ may be built up in pieces. To determine $g^{\mathbb{J}}(j)$, for $j \in \mathbb{J}$, a proper cover $C \subseteq \Xi_1(g)$ of j is found. The cover C is a set of sets. If C covers j, then every point in j is in a member of C:

$$C \text{ covers } j \;\equiv_{\text{def}}\; j^\square \subseteq \bigcup_{c \in C} c.$$

A cover is proper if it cannot be trivially shrunken:

$$C \text{ properly covers } j \;\equiv_{\text{def}}\; C \text{ covers } j \;\wedge\; \neg\exists[C' \subset C]\; C' \text{ covers } j.$$

Given a cover C of j it is trivial to construct a proper cover C' of j, simply by discarding members of C which do not overlap j. After a proper cover $C \subseteq \Xi_1(g)$ of j is found,

$$g^{\mathbb{J}}(j) \rightsquigarrow \bigcup_{\xi \in C} (g|\xi)^{\mathbb{J}}(j).$$

Since $\xi \in \Xi_1(g)$, $g|\xi$ is monotonic; hence $(g|\xi)^{\mathbb{J}}$ is simpler to evaluate than $g^{\mathbb{J}}$. The union of two intervals is an interval which includes the two given intervals:

$$j \cup^{\mathbb{J}} k \;\equiv_{\text{def}}\; \langle \min(j^-, k^-), \max(j^+, k^+) \rangle; \quad j \in \mathbb{J}, k \in \mathbb{J},$$

$$j \subseteq (j \cup k),\;\; k \subseteq (j \cup k).$$

Often, we form a set $\Xi_1^\star(g) \subseteq \Xi_1(g)$ from which proper covers of j may be easily formed, for any $j \in \mathbb{J}$. We will not mandate a particular choice of $\Xi_1^\star(g)$; there will be a natural choice for each g we consider. Using several $(g|\xi)^{\mathbb{J}}$, with $\xi \in \Xi_1^\star(g)$, we may then evaluate $g^{\mathbb{J}}(j)$ for any $j \in \mathbb{J}$ using the above strategy.

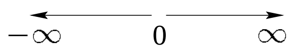
3.2.7 $\Xi_1^\star$ Charts

Some examples of $\Xi_1^\star(g)$ follow.

$$\begin{array}{rcccl}
\{[-\infty,\infty]\} & = & \Xi_1^\star(e^x) & \subseteq & \Xi_1(e^x), \\
\{[-\infty,0],[0,\infty]\} & = & \Xi_1^\star(x^2) & \subseteq & \Xi_1(x^2), \\
\{[-\infty,\infty]\} & = & \Xi_1^\star(\lfloor x \rfloor) & \subseteq & \Xi_1(\lfloor x \rfloor), \\
\{[-\infty,0],[0,\infty]\} & = & \Xi_1^\star(x^{-1}) & \subseteq & \Xi_1(x^{-1}), \\
\{\ldots,[-2\pi,-\pi],[-\pi,0],[0,\pi],[\pi,2\pi],\ldots\} & = & \Xi_1^\star(\cos(x)) & \subseteq & \Xi_1(\cos(x)).
\end{array}$$

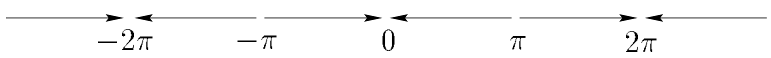
A $\Xi_1^\star(g)$ chart is used to visualize the sections the function g is cut into. Here are $\Xi_1^\star$ charts for the preceding examples:

$\Xi_1^\star(e^x)$ *Chart*

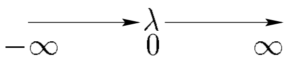


$\Xi_1^\star(x^2)$ *Chart*

$\Xi_1^\star(\lfloor x \rfloor)$ *Chart*



$\Xi_1^\star(\cos(x))$ *Chart*



$\Xi_1^\star(x^{-1})$ *Chart*

Determination of $\Xi_1^\star(g)$, for differentiable g, is aided by the relationship between $\frac{d}{dx}g$ and ψ_1^G: if $G \subseteq_2 g|[a,b]$ and $[a,b] \subseteq \mathrm{dom}(g)$, then

$$\exists[\xi \in [a,b]]\ \tfrac{d}{dx}g(\xi) = \psi_1^G.$$

As an example, consider $g|[0,\infty]$, $g(x) = x^2$; since

$$\tfrac{d}{dx}g = 2x,$$

for $\xi \in [0,\infty]$, which implies $\xi \geq 0$, the following holds:

$$\forall[\xi \in [0,\infty]]\ \tfrac{d}{dx}g(\xi) \geq 0.$$

From the aforementioned relationship between $\frac{d}{dx}g$ and ψ_1^G, it follows that $\psi_1^G \geq 0$, for any $G \subseteq_2 g|[0,\infty]$; so $\psi_1^\uparrow(g|[0,\infty])$.

3.2.8 Piecewise Constant Functions

We will determine $g^{\mathbb{J}}(j)$ for a constant function $g : \mathbb{R}^\star \mapsto \mathbb{R}^\star$. Piecewise constant functions are handled by considering $g|\xi$ for $\xi \in \Xi_1(g)$. The procedure is remarkably similar to the procedure for globally constant functions.

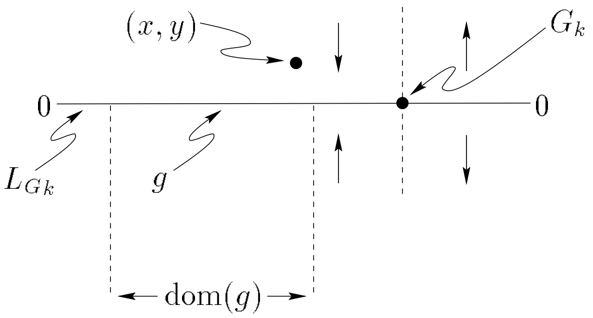
We have assumed that $\psi_1^0(g)$. Take any $G_k \subseteq_1 g$; a simple proof by contradiction, which follows, shows that L_{G_k} is an exact bound of g:

$$\forall[(x,y) \in g]\ L_{G_k}(x) \leq y \leq L_{G_k}.$$

Assume there is a point $(x,y) \in g$ such that $L_{G_k}(x) \neq y$. Let $G = G_k \cup \{(x,y)\}$, so $G \subseteq_2 g$:

$$\begin{aligned} L_{G_k}(x) \neq y \quad &\Rightarrow \quad (x,y) \notin G_k; \\ (x,y) \in g,\ (x,y) \notin G_k,\ G_k \subseteq_1 g \quad &\Rightarrow \quad G \subseteq_2 g. \end{aligned}$$

Furthermore, $G \subseteq g$ and $\psi_1^0(g)$ imply that $\psi_1^0(G)$.



ψ_1 Chart, with impossible (x,y)

A quick review of the ψ_1 chart reveals this situation is impossible, since $\psi_1^0(G)$ implies that $(x,y) \in L_{G_k}$. The ψ_1 chart predicts the sign of ψ_1^G since $G = G_k \cup \{(x,y)\}$.

3.2.9 Examples with Piecewise Constant Functions

Every function $g : \mathbb{R}^\star \mapsto \mathbb{R}^\star$ is a piecewise constant function, albeit with an infinite number of pieces. We are, however, concerned with functions which may be described using a finite number of pieces.

All globally constant functions are piecewise constant. We will consider such functions to be extended real functions, so that $[-\infty, \infty] \in \Xi_1(g)$. An example is $g(x) = \pi$:

$$\begin{array}{ll} & g^{\mathbb{J}}(\langle -1, 3\rangle) \\ \leadsto & {g_1}^{\mathbb{J}}(\langle -1, 3\rangle) \\ \leadsto & \langle g_1(x_1), g_1(x_1)\rangle, \quad x_1 \in \xi_1 \cap [-1, 3] \\ \leadsto & \langle \pi, \pi\rangle, \end{array}$$

with

$$C = \{\xi_1\} = \Xi_1^\star(g) \subseteq \Xi_1(g),\ \xi_1 = [-\infty, \infty],\ g_1 = g|\xi_1.$$

Consider the floor function, $g(x) = \lfloor x \rfloor : \mathbb{R} \mapsto \mathbb{R}$, which is a piecewise constant function. An example evaluation follows:

$$\begin{array}{lcccl} & g^{\mathbb{J}}(\langle 2.3, 3.1\rangle) & & & \\ \leadsto & {g_1}^{\mathbb{J}}(\langle 2.3, 3.1\rangle) & \cup & {g_2}^{\mathbb{J}}(\langle 2.3, 3.1\rangle) & \\ \leadsto & \langle g_1(x_1), g_1(x_1)\rangle & \cup & \langle g_2(x_2), g_2(x_2)\rangle, & x_i \in \xi_i \cap [2.3, 3.1] \\ \leadsto & \langle 2, 2\rangle & \cup & \langle 3, 3\rangle & \\ \leadsto & \langle 2, 3\rangle, & & & \end{array}$$

with

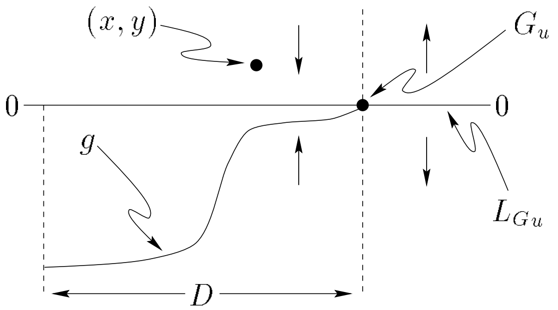
$$C = \{\xi_1, \xi_2\} \subseteq \Xi_1^\star(g) \subseteq \Xi_1(g),\ \xi_1 = [2, 3),\ \xi_2 = [3, 4),\ g_i = g|\xi_i.$$

3.2.10 Monotonically Increasing Functions

We will determine $g^{\mathbb{J}}(j)$ for any monotonically increasing function $g : \mathbb{R}^\star \mapsto \mathbb{R}^\star$. Since g is monotontically increasing, $\psi_1^\uparrow(g)$. We assume that $D = \mathrm{dom}(g) \subseteq j^\square$. We further assume that $D^+ \in D$, so we may take $G_u = \{(D^+, g(D^+))\}$. A simple proof by contradiction, which follows, shows that L_{G_u} is an upper bound for g:

$$\forall[(x, y) \in g]\ L_{G_u}(x) \geq y.$$

Assume that there is a point $(x, y) \in g$ such that $L_{G_u}(x) < y$. Let $G = G_u \cup \{(x, y)\}$, so $G \subseteq_2 g$. Furthermore, $G \subseteq g$ and $\psi_1^\uparrow(g)$ imply that $\psi_1^\uparrow(G)$.



ψ_1 Chart, with impossible (x, y)

A quick review of the ψ_1 chart reveals that this situation is impossible. There is no $(x, y) \in g$ such that $L_{G_u}(x) < y$ since $\psi_1^{\uparrow}(G)$, $G_u = \{(D^+, g(D^+))\}$, and $x \leq D^+$.

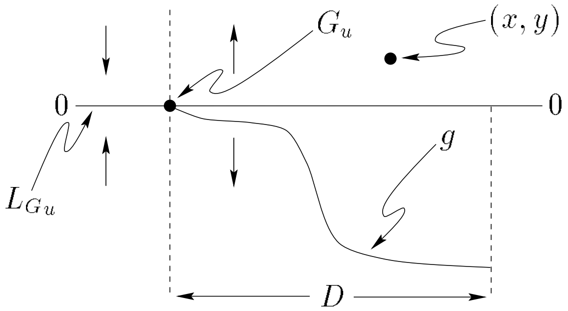
The two assumptions made do not overly restrict the applicability of the proof. If $D \not\subseteq j^{\square}$, consider $g|j^{\square}$ in place of g. If $D^+ \notin D$, consider $g' = g \cup \{(D^+, y)\}$ in place of g, such that g' is monotonically increasing. If $\lim_{x \to D^+} g(x)$ exists, it may be taken for y; otherwise, a trivial upper bound may be used.

3.2.11 Monotonically Decreasing Functions

We will determine $g^{\mathbb{J}}(j)$ for any monotonically decreasing function $g : \mathbb{R}^{\star} \mapsto \mathbb{R}^{\star}$. Since g is monotontically decreasing, $\psi_1^{\downarrow}(g)$. We assume that $D = \text{dom}(g) \subseteq j^{\square}$, and that $D^- \in D$, where $D^- = \inf D$; $D^+ = \sup D$, so that $D \subseteq [D^-, D^+]$. Take $G_u = \{(D^-, g(D^-))\}$; a simple proof by contradiction, which follows, shows that L_{G_u} is an upper bound for g:

$$\forall[(x, y) \in g]\ L_{G_u}(x) \geq y.$$

Assume that there is a point $(x, y) \in g$ such that $L_{G_u}(x) < y$. Let $G = G_u \cup \{(x, y)\}$, so $G \subseteq_2 g$. Furthermore, $G \subseteq g$ and $\psi_1^{\downarrow}(g)$ imply that $\psi_1^{\downarrow}(G)$.



ψ_1 *Chart, with impossible* (x, y)

A quick review of the ψ_1 chart reveals that this situation is impossible. There is no $(x, y) \in g$ such that $L_{G_u}(x) < y$ since $\psi_1^{\downarrow}(G)$, $G_u = \{(D^-, g(D^-))\}$, and $x \geq D^-$.

3.2.12 Lower Bounds

We have concentrated on upper bounds since lower bounds may be easily constructed using the rules given for upper bounds. This is achieved with the following identity:

$$g^{\mathbb{J}}(\boldsymbol{j})^- = -\left((-g)^{\mathbb{J}}(\boldsymbol{j})^+\right),$$

which follows directly from the definition of extremal bounds. Given that $(-g)^{\mathbb{J}}(\boldsymbol{j})^+$ is an upper bound of $(-g)(\boldsymbol{x}), \boldsymbol{x} \in \boldsymbol{j}$, it follows that $-\left((-g)^{\mathbb{J}}(\boldsymbol{j})^+\right)$ is a lower bound of $g(\boldsymbol{x}), \boldsymbol{x} \in \boldsymbol{j}$:

$$\begin{aligned}
&\forall[\boldsymbol{x} \in \boldsymbol{j}]\ (-g)^{\mathbb{J}}(\boldsymbol{j})^+ \geq (-g)(\boldsymbol{x}) \\
\Rightarrow\ &\forall[\boldsymbol{x} \in \boldsymbol{j}]\ -\left((-g)^{\mathbb{J}}(\boldsymbol{j})^+\right) \leq -(-g)(\boldsymbol{x}) \\
\Rightarrow\ &\forall[\boldsymbol{x} \in \boldsymbol{j}]\ -\left((-g)^{\mathbb{J}}(\boldsymbol{j})^+\right) \leq g(\boldsymbol{x}).
\end{aligned}$$

The proof is valid for all $g : \mathbb{R}^{\star n} \mapsto \mathbb{R}^\star$ since it only relies on properties of $\mathbb{R}^\star$.

So both lower and upper bounds for $g^{\mathbb{J}}(\boldsymbol{j})$ may be constructed from the upper bounds of $(-g)^{\mathbb{J}}(\boldsymbol{j})$ and $g^{\mathbb{J}}(\boldsymbol{j})$:

$$g^{\mathbb{J}}(\boldsymbol{j}) \rightsquigarrow \langle g^{\mathbb{J}}(\boldsymbol{j})^-, g^{\mathbb{J}}(\boldsymbol{j})^+\rangle \rightsquigarrow \langle -\left((-g)^{\mathbb{J}}(\boldsymbol{j})^+\right), g^{\mathbb{J}}(\boldsymbol{j})^+\rangle.$$

A similar process allows construction with lower bounds:

$$g^{\mathbb{J}}(\boldsymbol{j}) \rightsquigarrow \langle g^{\mathbb{J}}(\boldsymbol{j})^-, g^{\mathbb{J}}(\boldsymbol{j})^+\rangle \rightsquigarrow \langle g^{\mathbb{J}}(\boldsymbol{j})^-, -\left((-g)^{\mathbb{J}}(\boldsymbol{j})^-\right)\rangle.$$

We do assume that the number system underlying the interval number system has an exact negation operator which is total. Although the above construction could be taken literally, it is mainly a device to simplify exposition. In practice, upper and lower bounds are usually computed simultaneously by a single procedure.

3.2.13 Examples with Monotonic Functions

Consider the exponential function $g(x) = e^x : \mathbb{R}^\star \mapsto \mathbb{R}^\star$, which is a monotonically increasing function. An example evaluation follows:

$$\begin{aligned} & g^{\mathbb{J}}(\langle -3, 6\rangle) \\ \rightsquigarrow\ & {g_1}^{\mathbb{J}}(\langle -3, 6\rangle) \\ \rightsquigarrow\ & \langle g_1(-3), g_1(6)\rangle \\ \rightsquigarrow\ & \langle e^{-3}, e^6\rangle, \end{aligned}$$

with

$$C = \{\xi_1\} = \Xi_1^\star(g) \subseteq \Xi_1(g),\ \xi_1 = [-\infty, \infty],\ g_1 = g|\xi_1.$$

Similar functions include $\arctan(x), \operatorname{signum}(x), \lfloor x\rfloor, \lceil x\rceil, x^{2k+1}$ for all $k \in \mathbb{Z}$, $\sqrt[k]{x}$ for all $k \in \mathbb{Z}^+$, and k^x for all $k \in \mathbb{R}$, $k \geq 1$.

The floor function is both monotonically increasing and piecewise constant. Consider the evaluation of $g^{\mathbb{J}}(\langle 1.4, 7.2\rangle)$, $g(x) = \lfloor x\rfloor$, as a piecewise constant function:

$$\begin{array}{lcccl} & g^{\mathbb{J}}(\langle 1.4, 7.2\rangle) & & & \\ \rightsquigarrow & {g_1}^{\mathbb{J}}(\langle 1.4, 7.2\rangle) & \cup\ \cdots\ \cup & {g_7}^{\mathbb{J}}(\langle 1.4, 7.2\rangle) & \\ \rightsquigarrow & \langle g_1(x_1), g_1(x_1)\rangle & \cup\ \cdots\ \cup & \langle g_7(x_7), g_7(x_7)\rangle, & x_i \in \xi_i \cap [1.4, 7.2] \\ \rightsquigarrow & \langle 1, 1\rangle & \cup\ \cdots\ \cup & \langle 7, 7\rangle & \\ \rightsquigarrow & \langle 1, 7\rangle, & & & \end{array}$$

with

$$C = \{\xi_1, \ldots, \xi_7\} \subseteq \Xi_1^\star(g) \subseteq \Xi_1(g),\ \xi_i = [i, i+1),\ g_i = g|\xi_i;$$

and as a monotonically increasing function:

$$\begin{aligned} & g^{\mathbb{J}}(\langle 1.4, 7.2\rangle) \\ \rightsquigarrow\ & {g_1}^{\mathbb{J}}(\langle 1.4, 7.2\rangle) \\ \rightsquigarrow\ & \langle g_1(1.4), g_1(7.2)\rangle \\ \rightsquigarrow\ & \langle 1, 7\rangle, \end{aligned}$$

with

$$C = \{\xi_1\} = \Xi_1^\star(g) \subseteq \Xi_1(g),\ \xi_1 = [-\infty, \infty],\ g_1 = g|\xi_1.$$

The floor function should therefore be handled as a monotonically increasing function, for large arguments.

Consider the negation function $g(x) = -x : \mathbb{R}^\star \mapsto \mathbb{R}^\star$, which is a monotonically decreasing function. An example evaluation follows:

$$\begin{array}{rl} & g^{\mathbb{J}}(\langle 2, 4\rangle) \\ \rightsquigarrow & {g_1}^{\mathbb{J}}(\langle 2, 4\rangle) \\ \rightsquigarrow & \langle g_1(4), g_1(2)\rangle \\ \rightsquigarrow & \langle -4, -2\rangle, \end{array}$$

with

$$C = \{\xi_1\} = \Xi_1^\star(g) \subseteq \Xi_1(g),\ \xi_1 = [-\infty, \infty],\ g_1 = g|\xi_1.$$

The function k^x is similar, for $k \in \mathbb{R}$, $0 < k \leq 1$.

3.2.14 Examples with Piecewise Monotonic Functions

Consider the absolute value function $g(x) = |x| : \mathbb{R}^\star \mapsto \mathbb{R}^\star$, which is a piecewise monotonic function. An example evaluation follows:

$$\begin{array}{rccc} & g^{\mathbb{J}}(\langle -2.3, 3.1\rangle) & & \\ \rightsquigarrow & {g_1}^{\mathbb{J}}(\langle -2.3, 3.1\rangle) & \cup & {g_2}^{\mathbb{J}}(\langle -2.3, 3.1\rangle) \\ \rightsquigarrow & \langle g_1(0), g_1(-2.3)\rangle & \cup & \langle g_2(0), g_2(3.1)\rangle \\ \rightsquigarrow & \langle 0, 2.3\rangle & \cup & \langle 0, 3.1\rangle \\ \rightsquigarrow & \langle 0, 3.1\rangle, & & \end{array}$$

with

$$C = \{\xi_1, \xi_2\} \subseteq \Xi_1^\star(g) \subseteq \Xi_1(g),\ \xi_1 = [-\infty, 0],\ \xi_2 = [0, \infty],\ g_i = g|\xi_i.$$

For any $k \in \mathbb{Z}^+$, the function x^{2k} is similar.

Consider the reciprocation function $g^{\mathbb{R}^\star}(x) = x^{-1} : \mathbb{R}^\star \mapsto \mathbb{R}^\star$, which is a piecewise monotonically decreasing function. The function $g^{\mathbb{R}^\star}$ is an extension of the real function $g^{\mathbb{R}}$, $g^{\mathbb{R}}(x) = x^{-1}$:

$$g^{\mathbb{R}^\star}(x) = \begin{cases} g^{\mathbb{R}}(x) & \text{if } |x| \in \mathbb{R}^+, \\ \lambda & \text{if } |x| = 0, \\ 0 & \text{if } |x| = \infty. \end{cases}$$

Two example evaluations follow:

$$\begin{array}{rccc} & g^{\mathbb{J}}(\langle -2, 5\rangle) & & \\ \rightsquigarrow & {g_1}^{\mathbb{J}}(\langle -2, 5\rangle) & \cup & {g_2}^{\mathbb{J}}(\langle -2, 5\rangle) \\ \rightsquigarrow & \langle -\infty, g_1(-2)\rangle & \cup & \langle g_2(5), \infty\rangle \\ \rightsquigarrow & \langle -\infty, -0.5\rangle & \cup & \langle 0.2, \infty\rangle \\ \rightsquigarrow & \langle -\infty, \infty\rangle, & & \end{array}$$

with

$$C = \{\xi_1, \xi_2\} = \Xi_1^\star(g) \subseteq \Xi_1(g),\ \xi_1 = [-\infty, 0],\ \xi_2 = [0, \infty],\ g_i = g|\xi_i;$$

and:

$$\begin{array}{ll} & g^{\mathbb{J}}(\langle -2, 0\rangle) \\ \rightsquigarrow & {g_1}^{\mathbb{J}}(\langle -2, 0\rangle) \\ \rightsquigarrow & \langle -\infty, g_1(-2)\rangle \\ \rightsquigarrow & \langle -\infty, -0.5\rangle, \end{array}$$

with

$$C = \{\xi_1\} \subset \Xi_1^{\star}(g),\ g_1 = g|\xi_1.$$

The function x^{-k} is similar, for any $k \in \mathbb{Z}^{+}$.

3.2.15 Periodic Functions

Consider the piecewise monotonic function $g(x) = \sin(x)$. With thin arguments, the evaluation of $g^{\mathbb{J}}(j)$ proceeds as follows:

$$\begin{array}{lccc} & g^{\mathbb{J}}(\langle 1.5, 1.6\rangle) & & \\ \rightsquigarrow & {g_1}^{\mathbb{J}}(\langle 1.5, \frac{\pi}{2}\rangle) & \cup & {g_2}^{\mathbb{J}}(\langle \frac{\pi}{2}, 1.6\rangle) \\ \rightsquigarrow & \langle g_1(1.5), g_1(\frac{\pi}{2})\rangle & \cup & \langle g_2(1.6), g_2(\frac{\pi}{2})\rangle \\ \rightsquigarrow & \langle \sin(1.5), 1\rangle & \cup & \langle \sin(1.6), 1\rangle \\ \rightsquigarrow & \langle \sin(1.5), 1\rangle, & & \end{array}$$

with

$$C = \{\xi_1, \xi_2\} \subset \Xi_1^{\star}(g) \subset \Xi_1(g),\ \xi_1 = \left[\tfrac{-\pi}{2}, \tfrac{\pi}{2}\right],\ \xi_2 = \left[\tfrac{\pi}{2}, \tfrac{3\pi}{2}\right].$$

With thick arguments, the evaluation of $g^{\mathbb{J}}(j)$ is displeasing:

$$\begin{array}{lccccccc} & g^{\mathbb{J}}(\langle -0.2, 9.2\rangle) & & & & & & \\ \rightsquigarrow & {g_1}^{\mathbb{J}}(\langle -0.2, 9.2\rangle) & \cup & {g_2}^{\mathbb{J}}(\langle -0.2, 9.2\rangle) & \cup & {g_3}^{\mathbb{J}}(\langle -0.2, 9.2\rangle) & \cup & {g_4}^{\mathbb{J}}(\langle -0.2, 9.2\rangle) \\ \rightsquigarrow & \langle {g_1}^{\mathbb{J}}(-0.2), {g_1}^{\mathbb{J}}(\frac{\pi}{2})\rangle & \cup & \langle {g_2}^{\mathbb{J}}(\frac{\pi}{2}), {g_2}^{\mathbb{J}}(\frac{3\pi}{2})\rangle & \cup & \langle {g_3}^{\mathbb{J}}(\frac{3\pi}{2}), {g_3}^{\mathbb{J}}(\frac{5\pi}{2})\rangle & \cup & \langle {g_4}^{\mathbb{J}}(9.2), {g_4}^{\mathbb{J}}(\frac{5\pi}{2})\rangle \\ \rightsquigarrow & \langle \sin(-0.2), 1\rangle & \cup & \langle -1, 1\rangle & \cup & \langle -1, 1\rangle & \cup & \langle \sin(9.2), 1\rangle \\ \rightsquigarrow & \langle -1, 1\rangle & & & & & & \end{array}$$

with

$$C = \{\xi_1, \xi_2, \xi_3, \xi_4\} \subset \Xi_1^{\star}(g) \subset \Xi_1(g),\ \xi_i = \left[\tfrac{(2i-3)\pi}{2}, \tfrac{(2i-1)\pi}{2}\right],\ g_i = g|\xi_i;\ i \in \{1, 2, 3, 4\}.$$

Although the result returned is optimal, the amount of work performed to determine the result may be reduced.

We will cut the function $g : \mathbb{R}^{\star} \mapsto \mathbb{R}^{\star}$ into sections where each section attains the extreme values of g:

$$\Xi_{\pm}^{\mathbb{J}}(g) \ \equiv_{\text{def}} \ \{j : j \in \mathbb{J},\ \forall g^{\mathbb{J}} \ \ \langle g_-, g_+\rangle \subseteq g^{\mathbb{J}}(j)\},$$

$$g_- = \inf_{(x,y)\in g} y, \quad g_+ = \sup_{(x,y)\in g} y,$$

where

$$j \subseteq^{\mathbb{J}} k \ \equiv_{\text{def}} \ \forall[x \in^{\mathbb{J}} j]\ x \in^{\mathbb{J}} k.$$

When evaluating $g^{\mathbb{J}}(j)$, we may simply return $\langle g_-, g_+\rangle$ if any of the aforementioned sections lie within j:

$$\exists[j_{\pm} \in \Xi_{\pm}^{\mathbb{J}}(g),\ j_{\pm} \subseteq j]\ g^{\mathbb{J}}(j) \rightsquigarrow g^{\mathbb{J}}(\langle -\infty, \infty\rangle).$$

As with the previous sectioning scheme, there will often be a preferred sectioning, denoted by $\Xi^{\star\mathbb{J}}_{\pm}(g)$, which we will use to check containment.

With our example function g, $g(x) = \sin(x)$,

$$\{\langle x, x + 2\pi\rangle : x \in \mathbb{R}\} \subset \Xi^{\star\mathbb{J}}_{\pm}(g),$$

so the previous evaluation may be shortened. It may now proceed as follows:

$$\begin{array}{ll} & g^{\mathbb{J}}(\langle -0.2, 9.2\rangle) \\ \rightsquigarrow & g^{\mathbb{J}}(\langle -\infty, \infty\rangle) \qquad \text{since } \xi_{\pm} \subset [-0.2, 9.2] \\ \rightsquigarrow & \quad \langle -1, 1\rangle, \end{array}$$

with

$$\xi_{\pm} = [0, 2\pi] \in \Xi^{\star\mathbb{J}}_{\pm}(g).$$

This rejection test may be performed with a single subtraction, to find the width of the argument. Another quick rejection test is possible, by allowing another class of intervals into $\Xi^{\star\mathbb{J}}_{\pm}(g)$:

$$\{\langle k\pi, (k+1)\pi\rangle : k \in \mathbb{Z}\} \subset \Xi^{\star\mathbb{J}}_{\pm}(g).$$

3.2.16 Partial Functions

We have considered implementing a model $g^{\mathbb{J}}$, given $g^{\mathbb{R}}$. We now consider implementing $g^{\mathbb{J}^{|\mathbb{T}}}$. The property $\mathcal{P}_{|}$ is of interest; let $\Xi_{|}(g)$ denote the domain of g, defined in terms of $\mathcal{P}_{|}$:

$$\Xi_{|}(g) \;\equiv_{\text{def}}\; \{x : \mathcal{P}_{|}(g, x)\} \equiv \operatorname{dom}(G).$$

The function $\Phi_{\mathbb{T}}$,

$$\Phi_{\mathbb{T}} : \mathbb{J}^{|\mathbb{T}} \times 2^{\mathbb{R}^{\star}} \mapsto \mathbb{T},$$

when given an interval j and a set ξ of extended real numbers, produces a valid description of the relationship between j and ξ:

$$\Phi_{\mathbb{T}}(j, \xi) \rightsquigarrow d, \quad (j \in \xi) \sqsubseteq d.$$

The relationship between j and ξ is that of containment, formally defined as follows:

$$j \in \xi \;\equiv_{\text{def}}\; \bigcup_{x \in j} x \in \xi.$$

For the function $g : \mathbb{R}^{\star} \mapsto \mathbb{R}^{\star}$, an evaluation of the model $g^{\mathbb{J}^{|\mathbb{T}}}$ proceeds as follows:

$$g^{\mathbb{J}^{|\mathbb{T}}}(\langle v|d\rangle) \rightsquigarrow \langle v'|d'\rangle.$$

The resulting domain description d', $d' \in \mathbb{T}$, is determined using d, $\Xi_{|}$, and $\Phi_{\mathbb{T}}$:

$$d' = d \wedge \Phi_{\mathbb{T}}(\langle v|d\rangle, \Xi_{|}(g)).$$

The resulting value v', $v' \in \mathbb{J}$, depends on d'. If $d' \neq \mathrm{F}$, the resulting value is given by the methods outlined earlier:

$$d' \neq \mathrm{F} \;\Rightarrow\; v' = g^{\mathbb{J}}(v).$$

If $d' = \mathrm{F}$, the resulting value is arbitrary:

$$d' = \mathrm{F} \;\Rightarrow\; v' = \langle -\infty, \infty\rangle,$$

as $d' = \mathrm{F}$ implies that $g(x) = \lambda$ for all $x \in \langle v|d\rangle$.

3.2.17 Examples with a Partial Function

We now consider an example partial function, the square root function:

$$g : \mathbb{R}^{\star} \mapsto \mathbb{R}^{\star},\ g(x) = \sqrt{x}.$$

The function g is defined for non-negative extended real numbers:

$$\Xi_{|}(g) = [0, \infty],\quad \xi_{|} = \Xi_{|}(g).$$

The evaluation of $g^{\mathbb{J}|\mathbb{T}}(j)$, $j = \langle\langle -1, 1\rangle|\mathrm{T}\rangle$, proceeds as follows:

$$\begin{array}{llclcl} & g^{\mathbb{J}|\mathbb{T}}(\langle\langle -1,1\rangle|\mathrm{T}\rangle) & & & & \\ \rightsquigarrow & \langle \quad g^{\mathbb{J}}(\langle -1,1\rangle) & | & \mathrm{T} \wedge \Phi_{\mathbb{T}}(j, \xi_{|}) & \rangle & \\ \rightsquigarrow & \langle \quad g^{\mathbb{J}}(\langle -1,1\rangle) & | & \mathrm{T} \wedge \mathbb{F} & \rangle, & \text{since } j \in \xi_{|} = \{\mathrm{F}, \mathrm{T}\} \\ \rightsquigarrow & \langle \quad \langle 0, 1\rangle & | & \mathbb{F} & \rangle. & \end{array}$$

The evaluation of $g^{\mathbb{J}|\mathbb{T}}(j)$, $j = \langle\langle 1, 4\rangle|\mathbb{F}\rangle$, proceeds as follows:

$$\begin{array}{llclcl} & g^{\mathbb{J}|\mathbb{T}}(\langle\langle 1,4\rangle|\mathbb{F}\rangle) & & & & \\ \rightsquigarrow & \langle \quad g^{\mathbb{J}}(\langle 1,4\rangle) & | & \mathbb{F} \wedge \Phi_{\mathbb{T}}(j, \xi_{|}) & \rangle & \\ \rightsquigarrow & \langle \quad g^{\mathbb{J}}(\langle 1,4\rangle) & | & \mathbb{F} \wedge \mathrm{T} & \rangle, & \text{since } j \in \xi_{|} = \{\mathrm{T}\} \\ \rightsquigarrow & \langle \quad \langle 1, 2\rangle & | & \mathbb{F} & \rangle. & \end{array}$$

The evaluation of $g^{\mathbb{J}|\mathbb{T}}(j)$, $j = \langle\langle -3, -2\rangle|\mathrm{T}\rangle$, proceeds as follows:

$$\begin{array}{llclcl} & g^{\mathbb{J}|\mathbb{T}}(\langle\langle -3,-2\rangle|\mathrm{T}\rangle) & & & & \\ \rightsquigarrow & \langle \quad g^{\mathbb{J}}(\langle -3,-2\rangle) & | & \mathrm{T} \wedge \Phi_{\mathbb{T}}(j, \xi_{|}) & \rangle & \\ \rightsquigarrow & \langle \quad g^{\mathbb{J}}(\langle -3,-2\rangle) & | & \mathrm{T} \wedge \mathrm{F} & \rangle, & \text{since } j \in \xi_{|} = \{\mathrm{F}\} \\ \rightsquigarrow & \langle \quad \langle -\infty, \infty\rangle & | & \mathrm{F} & \rangle. & \end{array}$$

The evaluation of $g^{\mathbb{J}|\mathbb{T}}(j)$, $j = \langle\langle 1, 4\rangle|\mathrm{F}\rangle$, proceeds as follows:

$$\begin{array}{llclcl} & g^{\mathbb{J}|\mathbb{T}}(\langle\langle 1,4\rangle|\mathrm{F}\rangle) & & & & \\ \rightsquigarrow & \langle \quad g^{\mathbb{J}}(\langle 1,4\rangle) & | & \mathrm{F} \wedge \Phi_{\mathbb{T}}(j, \xi_{|}) & \rangle & \\ \rightsquigarrow & \langle \quad g^{\mathbb{J}}(\langle 1,4\rangle) & | & \mathrm{F} & \rangle, & \text{note } j \in \xi_{|} = \{\} \\ \rightsquigarrow & \langle \quad \langle -\infty, \infty\rangle & | & \mathrm{F} & \rangle. & \end{array}$$

3.2.18 Discontinuous Functions

We now consider implementing $g^{\mathbb{J}^{\Delta\mathbb{T}}}$. The property $\mathcal{P}_{\Delta}$ is of interest; consider $\Xi_{\Delta}(g)$, defined in terms of $\mathcal{P}_{\Delta}$:

$$\Xi_{\Delta}(g) \ \equiv_{\text{def}}\ \{x : \mathcal{P}_{\Delta}(g, x)\}.$$

For the function $g : \mathbb{R}^{\star} \mapsto \mathbb{R}^{\star}$, an evaluation of the model $g^{\mathbb{J}^{\Delta\mathbb{T}}}$ proceeds as follows:

$$g^{\mathbb{J}^{\Delta\mathbb{T}}}(\langle v\Delta d\rangle) \rightsquigarrow \langle v'\Delta d'\rangle.$$

The resulting continuity description d', $d' \in \mathbb{T}$, is determined using d, Ξ_{Δ}, and $\Phi_{\mathbb{T}}$:

$$d' = d \wedge \Phi_{\mathbb{T}}(\langle v\Delta d\rangle, \Xi_{\Delta}(g)).$$

The resulting value v', $v' \in \mathbb{J}$, is given by the methods outlined earlier.

3.2.19 Example with a Discontinuous Function

We now consider an example discontinuous function, the floor function:

$$g : \mathbb{R} \mapsto \mathbb{R},\ g(x) = \lfloor x \rfloor.$$

The function g is continuous for non-integral arguments:

$$\Xi_\Delta(g) = \{(k, k+1) : k \in \mathbb{Z}\},\ \ \xi_\Delta = \Xi_\Delta(g).$$

The evaluation of $g^{\mathbb{J}^{\Delta\mathbb{T}}}(j)$, $j = \langle\langle 0.2, 1.8\rangle \Delta \mathrm{T}\rangle$, proceeds as follows:

$$\begin{array}{llcll}
 & & g^{\mathbb{J}^{|\mathbb{T}}}(\langle\langle 0.2, 1.8\rangle\Delta \mathrm{T}\rangle) & & \\
\rightsquigarrow & \langle & g^{\mathbb{J}}(\langle 0.2, 1.8\rangle) & \mid & \mathrm{T} \wedge \phi_{\mathbb{T}}(j, \xi_\Delta) \quad \rangle \\
\rightsquigarrow & \langle & \langle 0, 1\rangle & \mid & \mathrm{T} \wedge \mathrm{T\!F} \quad \rangle, \quad \text{since } j \in \xi_\Delta = \{\mathrm{F}, \mathrm{T}\} \\
\rightsquigarrow & \langle & \langle 0, 1\rangle & \mid & \mathrm{T\!F} \quad \rangle.
\end{array}$$

3.2.20 Bumpy Functions

We now consider implementing $g^{\mathbb{J}^{|\mathbb{T}\star}}$ models. Each $\mathbb{J}^{|\mathbb{T}\star}$ interval is given by a set of $\mathbb{J}^{|\mathbb{T}}$ intervals. We previously defined the union of two $\mathbb{J}$ intervals. Extending that definition results in the following method for taking the union of two $\mathbb{J}^{|\mathbb{T}}$ intervals:

$$\langle j_v | j_d\rangle \cup^{\mathbb{J}^{|\mathbb{T}}} \langle k_v | k_d\rangle \ \equiv_{\text{def}}\ \langle j_v \cup^{\mathbb{J}} k_v | j_d \vee k_d\rangle; \quad \langle j_v | j_d\rangle \in \mathbb{J}^{|\mathbb{T}}, \langle k_v | k_d\rangle \in \mathbb{J}^{|\mathbb{T}}.$$

Another method, which uses $\mathbb{J}^{|\mathbb{T}\star}$ to describe the result, follows:

$$j \cup^{\mathbb{J}^{|\mathbb{T}} \to \mathbb{J}^{|\mathbb{T}\star}} k \ \equiv_{\text{def}}\ \{j, k\}; \quad j \in \mathbb{J}^{|\mathbb{T}}, k \in \mathbb{J}^{|\mathbb{T}}.$$

Since each $\mathbb{J}^{|\mathbb{T}\star}$ interval is a collection of $\mathbb{J}^{|\mathbb{T}}$ intervals, the union of two $\mathbb{J}^{|\mathbb{T}\star}$ intervals is simply the sum of the two collections:

$$j \cup^{\mathbb{J}^{|\mathbb{T}\star}} k = \{j_0, j_1, \ldots, j_m\} \cup^{\mathbb{J}^{|\mathbb{T}\star}} \{k_0, k_1, \ldots, k_n\} \ \equiv_{\text{def}}\ \{j_0, j_1, \ldots, j_m, k_0, k_1, \ldots, k_n\};$$

$j \in \mathbb{J}^{|\mathbb{T}\star}$, $k \in \mathbb{J}^{|\mathbb{T}\star}$. Good models of $\mathbb{J}$-bumpy functions may be built using the methods presented so far, using $\cup^{\mathbb{J}^{|\mathbb{T}} \to \mathbb{J}^{|\mathbb{T}\star}}$ rather than $\cup^{\mathbb{J}^{|\mathbb{T}}}$, where appropriate.

3.2.21 Examples with Bumpy Functions

We will evaluate a $\mathbb{J}^{|\mathbb{T}\star}$ model of the multiplicative inverse, for the interval $\langle -2, 5\rangle$: let $g(x) = x^{-1}$,

$$C = \{\xi_1, \xi_2\} = \Xi_1^\star(g) \subseteq \Xi_1(g),\ \xi_1 = [-\infty, 0],\ \xi_2 = [0, \infty],\ g_i = g|\xi_i.$$

The evaluation of $g^{\mathbb{J}^{|\mathbb{T}\star}}(j)$, $j = \{\langle\langle -2, 5\rangle | \mathrm{T}\rangle\}$, proceeds as follows:

$$\begin{array}{llll}
 & g^{\mathbb{J}^{|\mathbb{T}\star}}(\{\langle\langle -2, 5\rangle|\mathrm{T}\rangle\}) & & \\
\rightsquigarrow & g^{\mathbb{J}^{|\mathbb{T}}}(\langle\langle -2, 5\rangle|\mathrm{T}\rangle) & & \\
\rightsquigarrow & {g_1}^{\mathbb{J}^{|\mathbb{T}}}(\langle\langle -2, 5\rangle|\mathrm{T}\rangle) & \cup^{\mathbb{J}^{|\mathbb{T}} \to \mathbb{J}^{|\mathbb{T}\star}} & {g_2}^{\mathbb{J}^{|\mathbb{T}}}(\langle\langle -2, 5\rangle|\mathrm{T}\rangle) \\
\rightsquigarrow & \langle\langle -\infty, -0.5\rangle|\mathrm{T\!F}\rangle & \cup^{\mathbb{J}^{|\mathbb{T}} \to \mathbb{J}^{|\mathbb{T}\star}} & \langle\langle 0.2, \infty\rangle|\mathrm{T\!F}\rangle \\
\rightsquigarrow & \{\langle\langle -\infty, -0.5\rangle|\mathrm{T\!F}\rangle, \langle\langle 0.2, \infty\rangle|\mathrm{T\!F}\rangle\}. & &
\end{array}$$

In section 3.2.14 we showed that $g^{\mathbb{J}}(\langle -2,5\rangle) \rightsquigarrow \langle -\infty, \infty\rangle$.

We will evaluate a $\mathbb{J}^{|\mathbb{T}\star}$ model of the floor function, for the interval $\langle 2.3, 3.1\rangle$: let $g(x) = x^{-1}$,

$$C = \{\xi_1, \xi_2\} \subseteq \Xi_1^\star(g) \subseteq \Xi_1(g),\ \xi_1 = [2,3),\ \xi_2 = [3,4),\ g_i = g|\xi_i.$$

The evaluation of $g^{\mathbb{J}^{|\mathbb{T}\star}}(j)$, $j = \{\langle\langle 2.3, 3.1\rangle|\mathrm{T}\rangle\}$, proceeds as follows:

$$\begin{array}{lccc}
 & g^{\mathbb{J}^{|\mathbb{T}\star}}(\{\langle\langle 2.3,3.1\rangle|\mathrm{T}\rangle\}) & & \\
\rightsquigarrow & g^{\mathbb{J}^{|\mathbb{T}}}(\langle\langle 2.3,3.1\rangle|\mathrm{T}\rangle) & & \\
\rightsquigarrow & g_1{}^{\mathbb{J}^{|\mathbb{T}}}(\langle\langle 2.3,3.1\rangle|\mathrm{T}\rangle) & \cup^{\mathbb{J}^{|\mathbb{T}}\to\mathbb{J}^{|\mathbb{T}\star}} & g_2{}^{\mathbb{J}^{|\mathbb{T}}}(\langle\langle 2.3,3.1\rangle|\mathrm{T}\rangle) \\
\rightsquigarrow & \langle\langle 2,2\rangle|\mathrm{TF}\rangle & \cup^{\mathbb{J}^{|\mathbb{T}}\to\mathbb{J}^{|\mathbb{T}\star}} & \langle\langle 3,3\rangle|\mathrm{TF}\rangle \\
\rightsquigarrow & \{\langle\langle 2,2\rangle|\mathrm{TF}\rangle, \langle\langle 3,3\rangle|\mathrm{TF}\rangle\} & & .
\end{array}$$

In section 3.2.9 we showed that $g^{\mathbb{J}}(\langle 2.3, 3.1\rangle) \rightsquigarrow \langle 2,3\rangle$.

3.2.22 Common Binary Functions

Unary functions have been discussed fully, so we now turn our attention to binary functions. There are several ways of extending the methods presented to handle binary functions. The timely evaluation of arbitrary binary functions is difficult, so we will first list the binary functions that interest us. These functions are: $x+y$, $x-y$, $x\times y$, $x \div y$, x^y, $\min(x,y)$, $\max(x,y)$, and $\theta(x,y)$; we may later refer to these as the common binary functions. We consider the exponential function, x^y, for positive bases only: $\mathrm{dom}(x^y) = (0,\infty] \times [-\infty,\infty]$. The function $\theta(x,y)$ gives the angle from the origin to the point (x,y):

$$(x,y) = (\sqrt{x^2+y^2}\cos\alpha, \sqrt{x^2+y^2}\sin\alpha);\ \ \alpha = \theta(x,y) \in (-\pi,\pi].$$

There is a unique angle $\alpha \in (-\pi,\pi]$ satisfying the above equation, unless $\sqrt{x^2+y^2} = 0$. The function $\theta(x,y)$ is defined when there is a unique angle; $\mathrm{dom}(\theta) = \mathbb{R}^2 - \{(0,0)\}$.

The functions of interest may be rewritten, as follows:

$$\begin{aligned}
x - y &= x + (-y),\\
x \times y &= \tfrac{1}{2}[(x+y)^2 - (x^2+y^2)]\\
&= e^{\ln(x)+\ln(y)},\\
x \div y &= x \times y^{-1},\\
x^y &= e^{y\ln x},\\
\theta(x,y) &= \tan^{-1}(\tfrac{y}{x}) - \tfrac{1}{2}\pi[\mathrm{signum}(x)-1][\mathrm{signum}(y)],\\
\min(x,y) &= \tfrac{1}{2}(x+y-|x-y|),\\
\max(x,y) &= \tfrac{1}{2}(x+y+|x-y|).
\end{aligned}$$

The addition operator is the sole remaining binary operator. Interval evaluation using the above rules for the binary operators will not produce optimal results. However, a cursory implementation of an interval arithmetic may use the above rules. It is reasonable to rewrite division, since the result will be nearly optimal, with a simpler implemention. Rewriting may expose salient features to a symbolic optimizer.

Argument reduction may be performed. For example, we may consider multiplication for positive multiplicands only, by exploiting the following identity:

$$(x+a)(y+b) = xy + ay + bx + ab.$$

The interaction between the underlying number system and any argument reduction, or rewriting, should be carefully considered. Careless symbolic manipulation may produce a form which needlessly exacts horrendous round-off during evaluation, and subsequently cause sub-optimal intervals to be returned.

3.2.23 Binary Functions

Given the binary function $g : \mathbb{X}^2 \mapsto \mathbb{X}$, let $g_{(x,y=\alpha)}$ and $g_{(x=\alpha,y)}$ denote unary functions, for any $\alpha \in \mathbb{X}$. These functions are one-dimensional slices of g, for a constant x or constant y. The functions are defined as follows:

$$g_{(x,y=\alpha)}(x) = g(x,\alpha),$$
$$g_{(x=\alpha,y)}(x) = g(\alpha,x).$$

We now restrict our attention to grid functions. The function g is a grid function if it is defined over a grid:

$$\mathrm{grid}(g) \ \equiv_{\mathrm{def}}\ \mathrm{grid}(\mathrm{dom}(g));$$

where, for any $D \subseteq \mathbb{X}^2$:

$$\mathrm{grid}(D) \ \equiv_{\mathrm{def}}\ \exists[X \subseteq \mathbb{X}]\ \exists[Y \subseteq \mathbb{X}]\ D = X \times Y.$$

A grid function $g : \mathbb{R}^{\star 2} \mapsto \mathbb{R}^\star$ may be classified using the scheme set out for unary functions:

$$\psi_1^{\chi_1\chi_2}(g) \ \equiv_{\mathrm{def}}\ \forall[\alpha \in \mathbb{R}^\star]\ \psi_1^{\chi_1}(g_{(x,y=\alpha)}) \ \wedge\ \psi_1^{\chi_2}(g_{(x=\alpha,y)}).$$

A function $g : \mathbb{R}^{\star 2} \mapsto \mathbb{R}^\star$ fits into a class if it may be extended into a grid function which fits into that class:

$$\psi_1^{\chi_1\chi_2}(g) \ \text{ if }\ \exists[G \subseteq \mathbb{R}^{\star 3}]\ \ g \subseteq G\ \wedge\ \mathrm{grid}(G)\ \wedge\ \psi_1^{\chi_1\chi_2}(G).$$

With this classification scheme, the function g may be cut into sections where each section fits into a class:

$$\Xi_1(g) = \{D : \psi_1^{\chi_1\chi_2}(g|D), D \subseteq \mathbb{R}^{\star 2}\}.$$

As with unary functions, $\Xi_1^\star(g)$ denotes a preferred sectioning from which covers are formed.

An upper bound for $g^{\mathbb{J}}(x,y)$, where $\psi_1^{\ddagger\ddagger}(g)$, is determined by considering $g_{(x,y=\alpha)}{}^{\mathbb{J}}(x)$, for all $\alpha \in y$, and then $g_{(x=\beta,y)}{}^{\mathbb{J}}(y)$, for a particular $\beta \in x$. Since $\psi_1^{\ddagger\ddagger}(g)$, the same $\beta \in x$ produces an upper bound of $g_{(x,y=\alpha)}{}^{\mathbb{J}}(x)$. Exceptional functions, whether they are partial, discontinuous or bumpy, are handled as before. For g, where $\psi_1^{\ddagger\ddagger}(g|D)$:

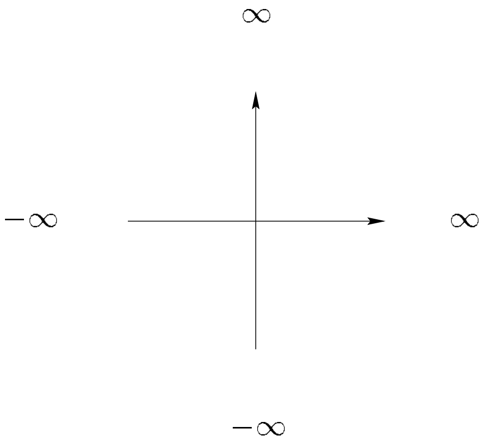
$$g^{\mathbb{J}}(j) = \langle L_{G_l}, L_{G_u}\rangle,\ \ G_l = \{(x_l,y_l)\},\ G_u = \{(x_u,y_u)\};$$

$$((x_l,y_l),(x_u,y_u)) = \begin{cases} ((j^+,k^+),(j^-,k^-)) & \text{if}\ \ \psi_1^{\downarrow\downarrow}(g|D), \\ ((j^+,k^-),(j^-,k^+)) & \text{if}\ \ \psi_1^{\downarrow\uparrow}(g|D), \\ ((j^-,k^+),(j^+,k^-)) & \text{if}\ \ \psi_1^{\uparrow\downarrow}(g|D), \\ ((j^-,k^-),(j^+,k^+)) & \text{if}\ \ \psi_1^{\uparrow\uparrow}(g|D); \end{cases}$$

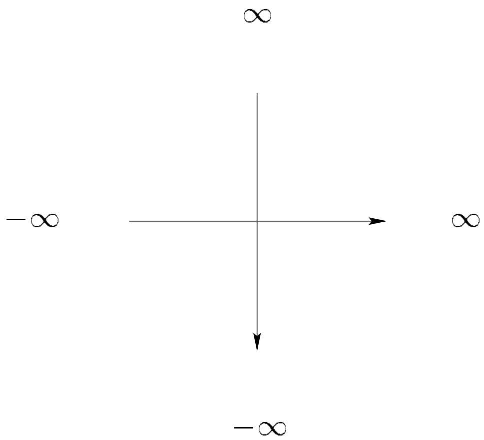
where $D = j^\square \times k^\square \subseteq \mathrm{dom}(g)$. We assume that $\{(x_l,y_l),(y_l,y_u)\} \subseteq g|D$; if not we may extend $g|D$, as was done with unary functions.

3.2.24 $\Xi_1^\star$ Charts

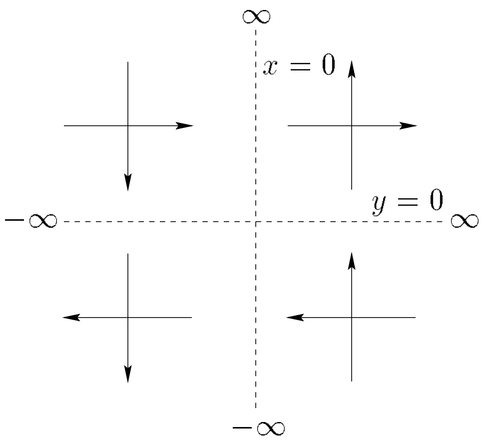
As with unary function, $\Xi_1^\star$ charts are used to graphically display the preferred sectioning of a function into monotonic pieces. Charts for some common binary functions follow.



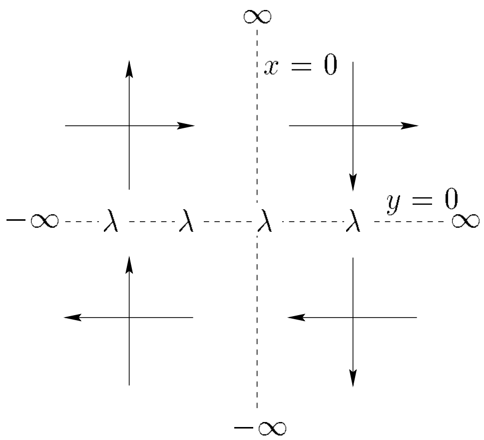
$\Xi_1^\star(x+y)$ *Chart*



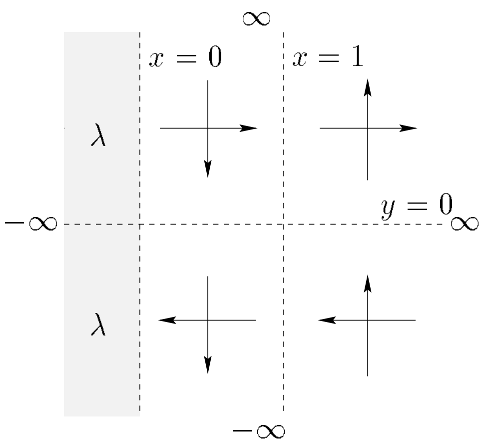
$\Xi_1^\star(x-y)$ *Chart*



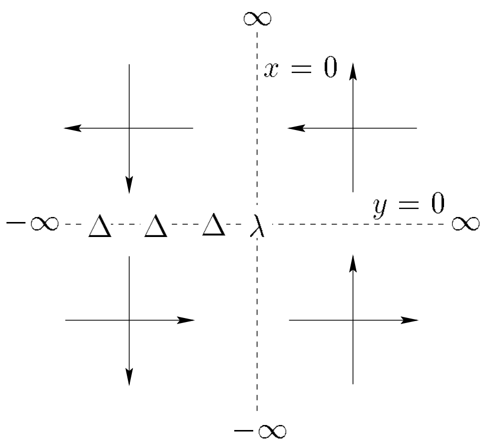
$\Xi_1^\star(xy)$ *Chart*



$\Xi_1^\star(x \div y)$ *Chart*



$\Xi_1^\star(x^y)$ *Chart*



$\Xi_1^\star(\theta(x,y))$ *Chart*

The $\Xi_1^\star(\min(x,y))$ and $\Xi_1^\star(\max(x,y))$ charts are both identical to the $\Xi_1^\star(x+y)$ chart. Regions where the function is not defined are labelled with λ; regions where the function is discontinuous are labelled with Δ. The horizontal and vertical arrows point in the direction of increasing g, for each component of g. For an interval box within a section the upper bound is given by $g(x,y)$, where (x,y) is the corner of the box that both arrows point towards. The lower bound is similarly given by $g(x,y)$, where (x,y) is the corner of the box that both arrows point away from. This is simply a graphical encoding of the rules given in the previous subsection.

Since

$$\tfrac{d}{dx}g_{(x,y=\alpha)}(\xi) = \tfrac{\partial}{\partial x}g(\xi,\alpha), \quad \tfrac{d}{dx}g_{(x=\alpha,y)}(\xi) = \tfrac{\partial}{\partial y}g(\alpha,\xi),$$

the relationship, between $\frac{d}{dx}g$ and ψ_1^G, used to aid the determination of $\Xi_1^\star(g)$, for unary g, may be used to aid the determination of $\Xi_1^\star(g)$, for binary g. As an example, consider $g|[0,\infty]^2$, $g(x,y)=xy$; since

$$\tfrac{\partial}{\partial x}g = y,\ \tfrac{\partial}{\partial y}g = x,$$

and $(\xi_x,\xi_y)\in[0,\infty]^2$, which implies $\xi_x\ge 0$ and $\xi_y\ge 0$, it follows that

$$\forall[(\xi_x,\xi_y)\in[0,\infty]^2]\ \tfrac{\partial}{\partial x}g(\xi_x,\xi_y)\ge 0,$$

$$\forall[(\xi_x,\xi_y)\in[0,\infty]^2]\ \tfrac{\partial}{\partial y}g(\xi_x,\xi_y)\ge 0.$$

From this, it follows that $\psi_1^{\uparrow\uparrow}(g|[0,\infty]^2)$.

3.2.25 Examples with a Binary Function

Consider the multiplication function, $g(x,y)=xy$,

$$\{\xi_{ij}\} = \Xi_1^\star(g),\ \xi_{ij} = R_i\times R_j,\ g_{ij} = g^{\mathbb{R}^\star}|\xi_{ij};$$

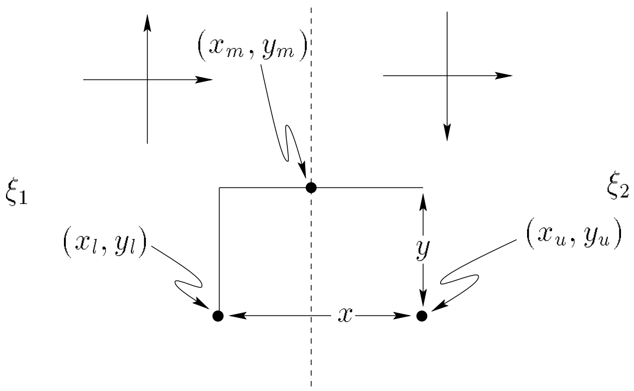
$$R_- = [-\infty,0], R_+ = [0,\infty];\ (i,j)\in\{-,+\}^2.$$

An example evaluation follows:

$$\begin{array}{rccc}
 & g^{\mathbb{J}}(\langle -2,3\rangle,\langle 5,7\rangle) & & \\
\rightsquigarrow & g_{-+}{}^{\mathbb{J}}(\langle -2,3\rangle,\langle 5,7\rangle) & \cup & g_{++}{}^{\mathbb{J}}(\langle -2,3\rangle,\langle 5,7\rangle) \\
\rightsquigarrow & \langle g_{-+}(-2,7), g_{-+}(0,5)\rangle & \cup & \langle g_{++}(0,5), g_{++}(3,7)\rangle \\
\rightsquigarrow & \langle -14,0\rangle & \cup & \langle 0,21\rangle \\
\rightsquigarrow & \langle -14,21\rangle. & &
\end{array}$$

For interval arguments that lie within one of the sections of $\Xi_1^\star(g)$, two applications of $g^{\mathbb{R}^\star}$ are used. For interval arguments which span two sections of $\Xi_1^\star(g)$, four applications of g are used, as was done in the above example evaluation. Since g is continuous, we may reduce the number of applications of $g^{\mathbb{R}^\star}$.

Consider evaluating $g(x, y)$, for the situation depicted below.



The argument, (x, y), is covered by two sections of $\Xi_1^{\star}(g)$:

$$C = \{\xi_1, \xi_2\},\ C \text{ covers } (x, y),\ g_i = g|\xi_i.$$

The sections share a common boundary, and points on the boundary are not strict local maxima or minima of g. This is true above, as both horizontal arrows point similarly, while the common boundary is vertical. With such a situation,

$$g^{\mathbb{J}}(x, y) = \langle l, m\rangle \cup \langle m, u\rangle;$$

$$l = g_1(x_l, y_l),\ m = g_1(x_m, y_m) = g_2(x_m, y_m),\ u = g_2(x_u, y_u).$$

It follows that $l \le m$, $m \le u$, so:

$$g^{\mathbb{J}}(x, y) = \langle l, u\rangle.$$

The bounds of $g^{\mathbb{J}}(x, y)$ may therefore be computed with two applications of g, rather than four. With the common binary functions, this situation always occurs when an argument spans two sections, unless the function is discontinuous at the common boundary. If the function is discontinuous at the common boundary, then the function is $\mathbb{J}$-bumpy. Similar reasoning may be used when the argument spans four sections, to reduce the number of applications of g from eight to four. Concern for efficiency is focussed on evaluation of $g^{\mathbb{J}}(\boldsymbol{x})$ when $\boldsymbol{x}^{\Box}$ is small.

3.2.26 Partial Binary Functions

Partial functions are handled as before. Let g denote a partial binary function. The domain of g may be defined in terms of $\mathcal{P}_|$:

$$\Xi_|(g)\ \equiv_{\text{def}}\ \{\boldsymbol{x} : \mathcal{P}_|(g, \boldsymbol{x})\} \equiv \text{dom}(g).$$

The function $\Phi_{\mathbb{T}}$,

$$\Phi_{\mathbb{T}} : \mathbb{J}^{|\mathbb{T}^2} \times 2^{\mathbb{R}^{\star 2}} \mapsto \mathbb{T},$$

again describes the relationship between $\boldsymbol{j}$ and $\boldsymbol{\xi}$:

$$\Phi_{\mathbb{T}}(\boldsymbol{j}, \boldsymbol{\xi}) \rightsquigarrow d,\quad (\boldsymbol{j} \in \boldsymbol{\xi}) \sqsubseteq d;$$

where $\boldsymbol{j} \in \mathbb{J}^{|\mathbb{T}^2}$ and $\boldsymbol{\xi} \in 2^{\mathbb{R}^{\star 2}}$. The relationship between $\boldsymbol{j}$ and $\boldsymbol{\xi}$ is that of containment, defined componentwise:

$$\boldsymbol{j} \in \boldsymbol{\xi} \ \equiv_{\text{def}} \ \forall i \ \boldsymbol{j}_i \in \boldsymbol{\xi}_i.$$

For the function $g : \mathbb{R}^{\star 2} \mapsto \mathbb{R}^{\star}$, an evaluation of the model $g^{\mathbb{J}|\mathbb{T}}$ proceeds as follows:

$$g^{\mathbb{J}|\mathbb{T}}(\langle j_v|j_d\rangle, \langle k_v|k_d\rangle) \rightsquigarrow \langle v'|d'\rangle.$$

The resulting domain description d', $d' \in \mathbb{T}$, is determined using j_d, k_d, $\Xi_|$, and $\Phi_{\mathbb{T}}$:

$$d' = j_d \wedge j_d \wedge \Phi_{\mathbb{T}}(\langle j_v|j_d\rangle, \langle k_v|k_d\rangle, \Xi_|(g)).$$

The resulting value v' depends on d', as the value depended on the domain for partial functions.

The methods for handling discontinuous and $\mathbb{J}$-bumpy functions are similarly extended.

3.2.27 Example with a Partial Binary Function

Consider the division function, $g(x, y) = x \div y$, which is both partial and $\mathbb{J}$-bumpy. For the division function,

$$\{\xi_{ij}\} = \Xi_1^{\star}(g), \ \xi_{ij} = R_i \times R_j, \ g_{ij} = g^{\mathbb{R}^{\star}}|\xi_{ij};$$

$$R_- = [-\infty, 0], R_+ = [0, \infty]; \ (i, j) \in \{-, +\}^2.$$

The evaluation of $g^{\mathbb{J}|\mathbb{T}\star}(x, y)$,

$$x = \{\langle\langle 30, 60\rangle|\mathrm{T}\rangle\}, \ y = \{\langle\langle -2, 3\rangle|\mathrm{T}\rangle\},$$

proceeds as follows:

$$\begin{array}{lccc}
 & g^{\mathbb{J}|\mathbb{T}\star}(\{\langle\langle 30, 60\rangle|\mathrm{T}\rangle, \langle\langle -2, 3\rangle|\mathrm{T}\rangle\}) & & \\
\rightsquigarrow & g^{\mathbb{J}|\mathbb{T}}(\langle\langle 30, 60\rangle|\mathrm{T}\rangle, \langle\langle -2, 3\rangle|\mathrm{T}\rangle\}) & & \\
\rightsquigarrow & g_{+-}{}^{\mathbb{J}|\mathbb{T}}(\langle\langle 30, 60\rangle|\mathrm{T}\rangle, \langle\langle -2, 3\rangle|\mathrm{T}\rangle) & \cup^{\mathbb{J}|\mathbb{T} \to \mathbb{J}|\mathbb{T}\star} & g_{++}{}^{\mathbb{J}|\mathbb{T}}(\langle\langle 30, 60\rangle|\mathrm{T}\rangle, \langle\langle -2, 3\rangle|\mathrm{T}\rangle) \\
\rightsquigarrow & \langle\langle -\infty, -15\rangle|\mathbb{TF}\rangle & \cup^{\mathbb{J}|\mathbb{T} \to \mathbb{J}|\mathbb{T}\star} & \langle\langle 10, \infty\rangle|\mathbb{TF}\rangle \\
\rightsquigarrow & \{\langle\langle -\infty, -15\rangle|\mathbb{TF}\rangle, \langle\langle 10, \infty\rangle|\mathbb{TF}\rangle\}. & &
\end{array}$$

The division function is defined unless the divisor is zero:

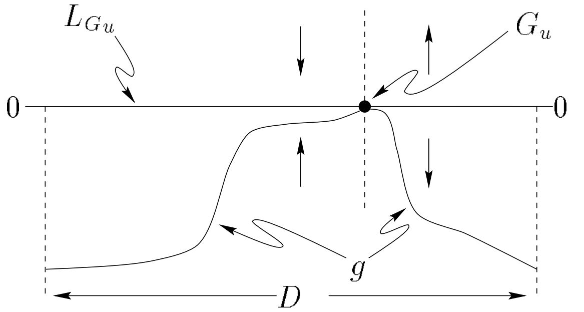
$$\Xi_|(g) = [-\infty, \infty] \times ([-\infty, 0) \cup (0, \infty]), \ \boldsymbol{\xi}_| = \Xi_|(g).$$

In the evaluation above, $g_{++}{}^{\mathbb{J}|\mathbb{T}}(\langle\langle 30, 60\rangle|\mathrm{T}\rangle, \langle\langle -2, 3\rangle|\mathrm{T}\rangle)$ is evaluated:

$$\begin{array}{llclc}
 & g_{++}{}^{\mathbb{J}|\mathbb{T}}(\langle\langle 30, 60\rangle|\mathrm{T}\rangle, \langle\langle -2, 3\rangle|\mathrm{T}\rangle) & & & \\
\rightsquigarrow & \langle \quad g_{++}{}^{\mathbb{J}}(\langle 30, 60\rangle, \langle -2, 3\rangle) & | & \mathrm{T} \wedge \mathrm{T} \wedge \Phi_{\mathbb{T}}(x, y, \boldsymbol{\xi}_|) & \rangle \\
\rightsquigarrow & \langle \quad g_{++}{}^{\mathbb{J}}(\langle 30, 60\rangle, \langle -2, 3\rangle) & | & \mathrm{T} \wedge \mathrm{T} \wedge \mathbb{TF} & \rangle \text{ since } (x, y) \in \boldsymbol{\xi}_| = \{\mathrm{F}, \mathrm{T}\} \\
\rightsquigarrow & \langle \quad \langle -\infty, -15\rangle & | & \mathbb{TF} & \rangle.
\end{array}$$

3.2.28 Monotonically Increasing, Decreasing Functions

There is another way to approach evaluating $g^{\mathbb{J}}$, when g is neither monotonically increasing nor monotonically decreasing. We consider g such that $\Xi_1^{\star}(g) = \{\xi_1, \xi_2\}$. Restricting our attention to continuous g, we may find an upper bound without splitting g into two parts. Consider the following chart, a ψ_1 chart where $\psi_1^{\uparrow}(g|\xi_1)$ and $\psi_1^{\downarrow}(g|\xi_2)$.



ψ_1 *Chart*

The bound may be seen to be correct by manually factoring the above chart into two separate charts, and then reasoning as before. A lower bound would be found for the above example with the procedures outlined earier.

3.3 Linear Interval Arithmetic

As with constant interval arithmetic, we will phrase our discussion in terms of our abstract model; which, for linear interval arithmetic, is $\mathbb{M}$. The procedures used to evaluate $g^{\mathbb{M}}$ may be used to evaluate $g^{\mathbb{Y}}$;

$$\mathbb{Y} = \mathcal{I}(\mathbb{X}),$$

$\mathbb{X}$ being the underlying number system. Section 3.3.30 details how this may be accomplished.

3.3.1 Interpolating Polynomials

Given the set $G = \{(x_0, y_0), (x_1, y_1), (x_2, y_2)\}$, consider the three functions $\varphi_{0,2}^G : \mathbb{R} \mapsto \mathbb{R}$, $\varphi_{1,2}^G : \mathbb{R} \mapsto \mathbb{R}$, and $\varphi_{2,2}^G : \mathbb{R} \mapsto \mathbb{R}$, defined as follows:

$$\varphi_{0,2}^G(x) = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)}, \quad \varphi_{1,2}^G(x) = \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)}, \quad \varphi_{2,2}^G(x) = \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)};$$

$\varphi_{i,d}^G$ is a d-degree polynomial with $\varphi_{i,d}^G(x_j) = \delta_{ij}$. Consider the function $g : \mathbb{R}^{\star} \mapsto \mathbb{R}^{\star}$, along with a representative G, $G \subseteq_3 g$. We may deduce that G is a function, and that:

$$(x_0 - x_1)(x_0 - x_2) \neq 0, \ (x_1 - x_0)(x_1 - x_2) \neq 0, \ (x_2 - x_0)(x_2 - x_1) \neq 0.$$

It follows that the functions $\varphi_{0,2}^G$, $\varphi_{1,2}^G$, and $\varphi_{2,2}^G$ are well defined, for our choice of G. Since $\varphi_{i,2}^G(x_j) = \delta_{ij}$, the function $L_G : \mathbb{R} \mapsto \mathbb{R}$,

$$L_G(x) = y_0\varphi_{0,1}^G + y_1\varphi_{1,1}^G + y_2\varphi_{0,2}^G,$$

interpolates G. L_G is the quadratic Lagrange interpolating polynomial of G.

L_G may be expressed in standard polynomial form:

$$L_G(x) = \psi^G_{2,2}x^2 + \psi^G_{1,2}x + \psi^G_{0,2},$$

with

$$\begin{aligned}
\psi^G_{2,2} &= \frac{y_0}{(x_0-x_1)(x_0-x_2)} + \frac{y_1}{(x_1-x_0)(x_1-x_2)} + \frac{y_2}{(x_2-x_0)(x_2-x_1)},\\
\psi^G_{1,2} &= \frac{-y_0(x_1+x_2)}{(x_0-x_1)(x_0-x_2)} + \frac{-y_1(x_0+x_2)}{(x_1-x_0)(x_1-x_2)} + \frac{-y_2(x_0+x_1)}{(x_2-x_0)(x_2-x_1)},\\
\psi^G_{0,2} &= \frac{y_0x_1x_2}{(x_0-x_1)(x_0-x_2)} + \frac{y_1x_0x_2}{(x_1-x_0)(x_1-x_2)} + \frac{y_2x_0x_1}{(x_2-x_0)(x_2-x_1)};
\end{aligned}$$

$\psi^G_{i,d}$ is the coefficient of x^i in $L_G(x)$, a d-degree polynomial. The leading coefficient, $\psi^G_{2,2}$, is of special interest, and may be denoted simply by ψ^G_2:

$$\psi^G_2 = \psi^G_{2,2}.$$

The set G, and the associated polynomial L_G, are:

- concave down if $\psi^{\downarrow}_2(G)$,
- linear if $\psi^0_2(G)$, and
- concave up if $\psi^{\uparrow}_2(G)$;

where:

$$\psi^{\downarrow}_2(G) \;\equiv_{\text{def}}\; (\psi^G_2 \le 0), \quad \psi^0_2(G) \;\equiv_{\text{def}}\; (\psi^G_2 = 0), \quad \psi^{\uparrow}_2(G) \;\equiv_{\text{def}}\; (\psi^G_2 \ge 0).$$

Consider $G^\star$, a richer representation of g; $G^\star \subseteq_{\ge 3} g$. The representation $G^\star$ has one of the preceding properties if all three-member subsets of $G^\star$ have the same property:

$$\psi^\chi_2(G^\star) \;\equiv_{\text{def}}\; \forall[G \subseteq_3 G^\star]\; \psi^\chi_2(G).$$

All three properties are considered to be satisfied by sparse representations of g since

$$\forall[G \subseteq_{<3} g]\; \forall[\psi'_2 \in \mathbb{R}]\; \exists[\psi_1 \in \mathbb{R}]\; \exists[\psi'_0 \in \mathbb{R}]\; \forall[(x_i, y_i) \in G]\; L'_G(x_i) = y_i,$$

where $L'_G(x) = \psi'_2x^2 + \psi'_1x + \psi'_0$. For $G = g$, the usual definitions of linearity and concavity are equivalent to those given here. Let $\psi^{\updownarrow}_2(G^\star)$ state that $G^\star$ has one of the above properties:

$$\psi^{\updownarrow}_2(G^\star) \;\equiv_{\text{def}}\; \exists(\chi \in \mathbb{O})\; \psi^\chi_2(G^\star).$$

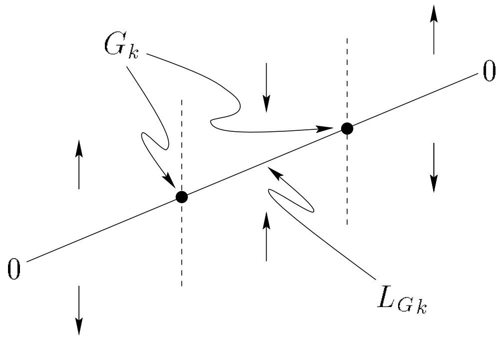
For all representations $G \subseteq g$,

$$\psi^0_2(G) \;\Leftrightarrow\; \psi^{\downarrow}_2(G) \wedge \psi^{\uparrow}_2(G).$$

Using the linear and quadratic interpolating polynomials we will construct linear bounds for many common functions.

3.3.2 ψ_2 Charts

Consider the following chart:

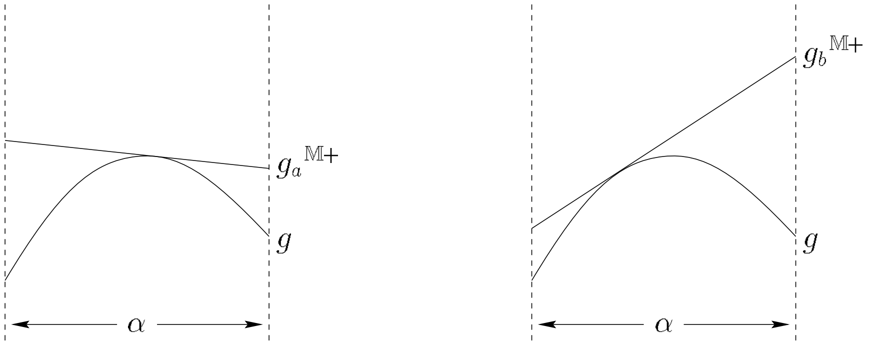


ψ_1 *Chart*

The ψ_2 chart is used to predict the sign of ψ_2^G, for $G = G_k \cup \{(x, y)\}$, given G_k. The chart divides $\mathbb{R}^2$ into nine disjoint regions.

3.3.3 Optimality

The notion of optimality is not as simple for linear interval arithmetic as it was for constant interval arithmetic. Consider the following function g, with two distinct bounds, ${g_a}^{\mathbb{M}-}$ and ${g_b}^{\mathbb{M}+}$:



Bound A *Bound B*

We now define a measure of bound goodness. Consider the $\mathcal{L}_1$ norm:

$$\mathcal{L}_1(g, g^{\mathbb{M}+}) \equiv_{\text{def}} \int_0^1 \omega(\alpha)(g^{\mathbb{M}+}(\alpha) - g(\alpha)),$$

where ω is a continuous positive function defined on $[0, 1]$. The $\mathcal{L}_1$ norm is always positive, since we consider only $g^{\mathbb{M}+}$ which are upper bounds of g. We consider the upper bound ${g_a}^{\mathbb{M}+}$ to be a better upper bound than ${g_b}^{\mathbb{M}+}$ if $\mathcal{L}_1(g, {g_a}^{\mathbb{M}+}) < \mathcal{L}_1(g, {g_b}^{\mathbb{M}+})$: a bound $g^{\mathbb{M}+}$ is good if $\mathcal{L}_1(g, g^{\mathbb{M}+})$ is small.

A bound ${g_\star}^{\mathbb{M}+}$ is optimal, for linear interval arithmetic, if no better linear interval upper bound exists:

$$\text{optimal}^+({g_\star}^{\mathbb{M}+}, g) \equiv_{\text{def}} \forall g^{\mathbb{M}+}\ \mathcal{L}_1(g, {g_\star}^{\mathbb{M}+}) \leq \mathcal{L}_1(g, g^{\mathbb{M}+}).$$

The model ${g_\star}^{\mathbb{M}}$ returns optimal upper bounds if the upper bound is optimal for all $m \in \mathbb{M}$:

$$\text{optimal}^+({g_\star}^{\mathbb{M}}, g) \equiv_{\text{def}} \forall [m \in \mathbb{M}]\ \text{optimal}^+(({g_\star}^{\mathbb{M}}(m))^+, g(m^+)).$$

As before, optimality can be defined without reference to the underlying function.

Also, arguing as before, we may show that an optimal model of g is an interval extension of g. This implies that for differentiable g, an optimal model produces bounds which touch g at two distinct points, allowing for infinitesimal separation between points. Infinitesimally separated points correspond to the upper bound matching both the value and the derivative at a point.

3.3.4 Piecewise Models

Any function $g : \mathbb{R}^\star \mapsto \mathbb{R}^\star$ may be cut into sections where each section fits into one class:

$$\Xi_2(g) \equiv_{\text{def}} \{D : \psi_2^\ddagger(g|D), D \subseteq \mathbb{R}^\star\}.$$

A model of a function $g : \mathbb{R}^\star \mapsto \mathbb{R}^\star$ may be built up in pieces. To determine $g^{\mathbb{M}}(m)$, for $m \in \mathbb{M}$, a proper cover $C \subseteq \Xi_2(g)$ of m is found. After a proper cover $C \subseteq \Xi_2(g)$ of m is found,

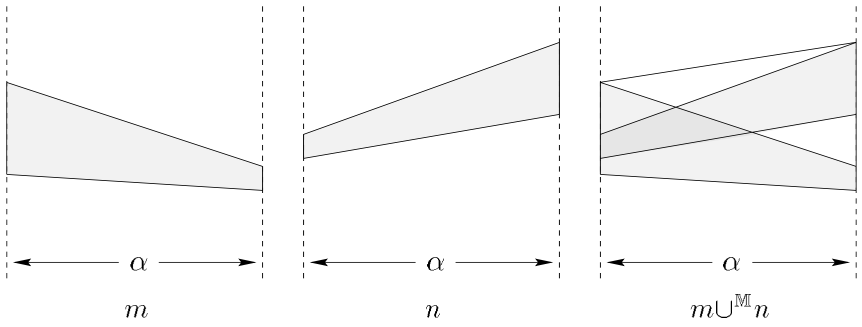
$$g^{\mathbb{M}}(m) \rightsquigarrow \bigcup_{\xi \in C} (g|\xi)^{\mathbb{M}}(m).$$

Since $\xi \in \Xi_2(g)$, $g|\xi$ is concave and is simpler to evaluate than g. The union of two linear intervals is a linear interval which includes the two given intervals:

$$m \cup^{\mathbb{M}} n \ \equiv_{\text{def}} \ \langle L_{G_l}, L_{G_u} \rangle, \quad m \subseteq (m \cup n), \quad n \subseteq (m \cup n);$$

$$\begin{aligned} G_l &= \{(0, \min(m(0), n(0))), (1, \min(m(1), n(1)))\}, \\ G_u &= \{(0, \max(m(0), n(0))), (1, \max(m(1), n(1)))\}, \end{aligned}$$

The following diagram displays the union of two linear intervals, m and n.



As before, covers are assembled from a preferred sectioning $\Xi_2^\star(g)$, $\Xi_2^\star(g) \subseteq \Xi_2(g)$.

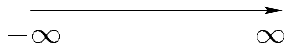
3.3.5 $\Xi_2^\star$ Charts

Some examples of $\Xi_2^\star(g)$ follow.

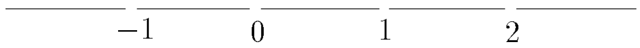
$$\begin{array}{rclcl} \{[-\infty, \infty]\} &=& \Xi_2^\star(e^x) &\subseteq& \Xi_2(e^x), \\ \{[-\infty, \infty]\} &=& \Xi_2^\star(x^2) &\subseteq& \Xi_2(x^2), \\ \{[-\infty, \infty]\} &=& \Xi_2^\star(|x|) &\subseteq& \Xi_2(|x|), \\ \{\ldots, [-2, -1), [-1, 0), [0, 1), [1, 2), \ldots\} &=& \Xi_2^\star(\lfloor x \rfloor) &\subseteq& \Xi_2(\lfloor x \rfloor), \\ \{[-\infty, 0], [0, \infty]\} &=& \Xi_2^\star(x^{-1}) &\subseteq& \Xi_2(x^{-1}), \\ \{\ldots, [-\frac{3}{2}\pi, -\frac{1}{2}\pi], [-\frac{1}{2}\pi, \frac{1}{2}\pi], [\frac{1}{2}\pi, \frac{3}{2}\pi], [\frac{3}{2}\pi, \frac{5}{2}\pi], \ldots\} &=& \Xi_2^\star(\cos(x)) &\subseteq& \Xi_2(\cos(x)). \end{array}$$

A $\Xi_2^\star(g)$ chart is used to visualize the sections the function g is cut into.

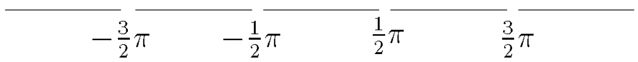
Here are $\Xi_2^\star$ charts for the preceding examples:



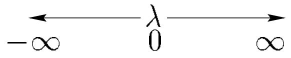
$\Xi_2^\star(e^x)$ *Chart*



$\Xi_2^\star(\lfloor x \rfloor)$ *Chart*



$\Xi_2^\star(\cos(x))$ *Chart*



$\Xi_2^\star(x^{-1})$ *Chart*

The $\Xi_2^\star(x^2)$ and $\Xi_2^\star(|x|)$ charts are both identical to the $\Xi_2^\star(e^x)$ chart.

Determination of $\Xi_2^\star(g)$, for twice differentiable g, is aided by the relationship between $\frac{d^2}{dx^2}g$ and ψ_2^G: if $G \subseteq_3 g|[a,b]$ and $[a,b] \subseteq \mathrm{dom}(g)$, then

$$\exists[\xi \in [a,b]]\ \tfrac{d^2}{dx^2}g(\xi) = \psi_2^G.$$

As an example, consider $g|[0,\infty]$, $g(x) = x^{-1}$; since

$$\tfrac{d^2}{dx^2}g = 2x^{-3},$$

for $\xi \in [0,\infty] \cap \mathrm{dom}(g) = (0,\infty]$, which implies $\xi \geq 0$, the following holds:

$$\forall[\xi \in [0,\infty] \cap \mathrm{dom}(g)]\ \tfrac{d^2}{dx^2}g(\xi) > 0.$$

From the aforementioned relationship between $\frac{d^2}{dx^2}g$ and ψ_2^G, it follows that $\psi_2^G > 0$, for any $G \subseteq_3 g|[0,\infty]$; so $\psi_2^\uparrow(g|[0,\infty])$.

3.3.6 Monotonic Sections

In the previous section, we designed $g^{\mathbb{J}}$ for a given function g, $g : \mathbb{R}^{\star n} \mapsto \mathbb{R}^\star$. We used $\Xi_1^\star(g)$ to limit our attention to monotonic sections of g. Monotonic sections will also help us design $g^{\mathbb{M}}$.

Consider a monotonically increasing function g, $g : \mathbb{R}^\star \mapsto \mathbb{R}^\star$. Let both

$$g^{\mathbb{R}^\star}(c + d\alpha) \leq u(\alpha),$$

and

$$l(\alpha) \leq g^{\mathbb{R}^\star}(a + b\alpha),$$

hold, for $\alpha \in [0,1]$. Since g is monotonically increasing,

$$a + b\alpha \leq x \leq c + d\alpha\ \Rightarrow g^{\mathbb{R}^\star}(a + b\alpha) \leq g^{\mathbb{R}^\star}(x) \leq g^{\mathbb{R}^\star}(c + d\alpha).$$

Combining this with the previous bounds results in the following:

$$l(\alpha) \leq g^{\mathbb{R}^\star}(a + b\alpha) \leq g^{\mathbb{R}^\star}(x) \leq g^{\mathbb{R}^\star}(c + d\alpha) \leq u(\alpha),$$

for $a + b\alpha \leq x \leq c + d\alpha$. This may be simplified to:

$$l(\alpha) \leq g^{\mathbb{R}^\star}(x) \leq u(\alpha),$$

for $x \in \langle a+b\alpha, c+d\alpha\rangle$. It is now established that $\langle l(\alpha), u(\alpha)\rangle$ bounds $g^{\mathbb{M}}(\langle a+b\alpha, c+d\alpha\rangle)$.

Consider a monotonically decreasing function g, $g : \mathbb{R}^\star \mapsto \mathbb{R}^\star$. Let both

$$g^{\mathbb{R}^\star}(a+b\alpha) \leq u(\alpha),$$

and

$$l(\alpha) \leq g^{\mathbb{R}^\star}(c+d\alpha),$$

hold, for $\alpha \in [0,1]$. Since g is monotonically decreasing,

$$a+b\alpha \leq x \leq c+d\alpha \;\Rightarrow g^{\mathbb{R}^\star}(c+d\alpha) \leq g^{\mathbb{R}^\star}(x) \leq g^{\mathbb{R}^\star}(a+b\alpha).$$

Reasoning as before forces us to conclude that $\langle l(\alpha), u(\alpha)\rangle$ bounds $g^{\mathbb{M}}(\langle a+b\alpha, c+d\alpha\rangle)$.

So, for a monotonic g, we may evaluate $g^{\mathbb{M}}(m)$ by proceeding as follows:

$$g^{\mathbb{M}}(m) \rightsquigarrow \begin{cases} \langle g(m^-)^{\mathbb{M}-}, g(m^+)^{\mathbb{M}+}\rangle & \text{if} \quad \psi_1^{\uparrow}(g), \\ \langle g(m^+)^{\mathbb{M}-}, g(m^-)^{\mathbb{M}+}\rangle & \text{if} \quad \psi_1^{\downarrow}(g). \end{cases}$$

Let g_- and g_+ denote two functions, from $\mathbb{R}^\star$ to $\mathbb{R}^\star$, defined as follows:

$$\begin{aligned} g_-(\alpha) &= g(m^-(\alpha)), \\ g_+(\alpha) &= g(m^+(\alpha)). \end{aligned}$$

We now focus on determining $g'^{\mathbb{M}-}$ and $g'^{\mathbb{M}+}$, for $g' = g_-$ and $g' = g_+$. As will be seen, this will give us a method for computing $g^{\mathbb{M}\between}$ rather than $g^{\mathbb{M}}$; appropriate demotions may be used to ensure the result is in $\mathbb{J}$, if necessary.

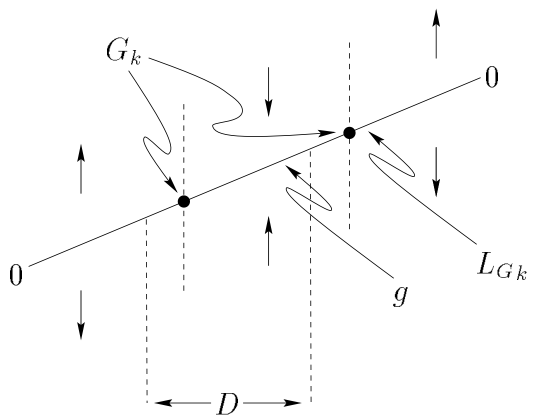
3.3.7 Linear Functions

We will determine $g^{\mathbb{M}-}$ and $g^{\mathbb{M}+}$ for a linear function $g : \mathbb{R}^\star \mapsto \mathbb{R}^\star$.

We have assumed that $\psi_2^0(g)$. Take any $G_k \subseteq_2 g$; a simple proof by contradiction, which follows, shows that L_{G_k} is an exact bound of g:

$$\forall[(x,y) \in g]\; L_{G_k}(x) \leq y \leq L_{G_k}(x).$$

Assume there is a point $(x,y) \in g$ such that $L_{G_k}(x) \neq y$. Let $G = G_k \cup \{(x,y)\}$, so $G \subseteq_3 g$. Furthermore, $G \subseteq g$ and $\psi_2^0(g)$ imply that $\psi_2^0(G)$.



ψ_2 *Chart*

A quick review of the ψ_2 chart reveals this situation is impossible, since $\psi_2^0(G)$ implies that $(x,y) \in L_{G_k}$. The ψ_2 chart predicts the sign of ψ_2^G since $G = G_k \cup \{(x,y)\}$.

3.3.8 Example with a Linear Function

Consider the negation function, $g(x) = -x : \mathbb{R}^\star \mapsto \mathbb{R}^\star$, which is a globally linear function. An example evaluation follows:

$$\begin{array}{lllll} & g^{\mathbb{M}}(\langle -3+\alpha, 7-2\alpha \rangle) & & & \\ \rightsquigarrow & {g_1^1}^{\mathbb{M}}(\langle -3+\alpha, 7-2\alpha \rangle) & & & \\ \rightsquigarrow & \langle \quad g_1^1(7-2\alpha)^{\mathbb{M}-} & , & g_1^1(-3+\alpha)^{\mathbb{M}+} & \rangle, \text{ since } \psi_1^{\downarrow}(g_1^1) \\ \rightsquigarrow & \langle \quad g_1^1(7-2\alpha) & , & g_1^1(-3+\alpha) & \rangle, \text{ since } \psi_2^0(g_1^1) \\ \rightsquigarrow & \langle \quad -7+2\alpha & , & 3-\alpha & \rangle, \end{array}$$

with

$$C = \{\xi_1^1\},\ \{\xi_1^1\} = \Xi_1^\star(g) \subseteq \Xi_1(g),\ \{\xi_1^1\} = \Xi_2^\star(g) \subseteq \Xi_2(g),$$
$$\xi_1^1 = [-\infty, \infty],\ g_1^1 = g|\xi_1^1.$$

3.3.9 Examples with a Piecewise Linear Function

Consider the floor function, $g(x) = \lfloor x \rfloor : \mathbb{R} \mapsto \mathbb{R}$, which is a piecewise constant function. We previously stated that

$$\Xi_2^\star(\lfloor x \rfloor) = \{\ldots, [-2,-1), [-1,0), [0,1), [1,2), \ldots\};$$

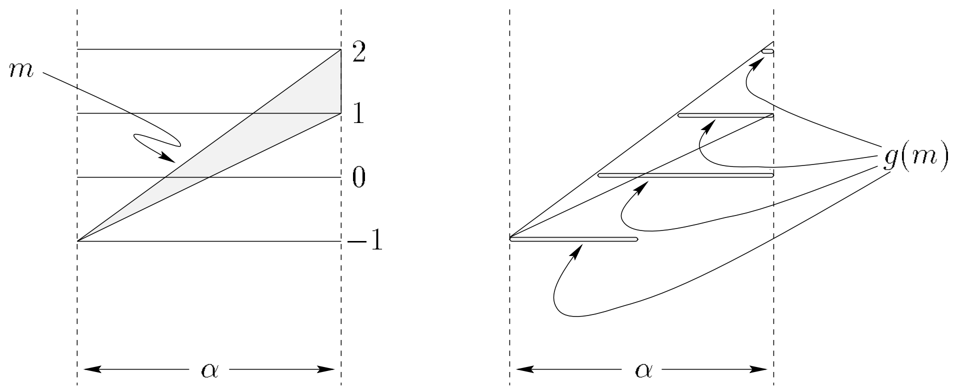
another possibility is to let

$$\Xi_2^\star(\lfloor x \rfloor) = \{\{\ldots, k-1, k, k+1, \ldots\} : k \in [0,1)\}.$$

With either sectioning, the function is seen to be piecewise linear, and is monotonically increasing:

$$\{\xi_1\} = \Xi_1^\star(g), \xi_1 = [-\infty, \infty].$$

Consider the following diagram:



m $\qquad$ $g(m)$

This type of diagram will be used throughout this section. The light grey region will often represent $g(m)$, as it does in the rightmost diagram:

$$x \in g(m)(\alpha) \text{ iff } (\alpha, x) \text{ is in the grey region.}$$

In the leftmost diagram, the region represents m; the two diagrams together illustrate how $g^{\mathbb{M}}(m)$ was determined.

The evaluation of $g^{\mathbb{M}}(m)$,

$$m = \langle -1+2\alpha, -1+3\alpha \rangle,$$

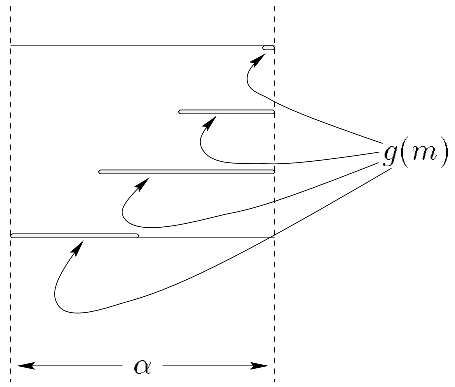
proceeds as follows, using the first sectioning:

$$\begin{array}{lccccccc} & g^{\mathbb{M}}(m) & & & & & & \\ \rightsquigarrow & g_1^{-1\mathbb{M}}(m) & \cup & g_1^{0\mathbb{M}}(m) & \cup & g_1^{1\mathbb{M}}(m) & \cup & g_1^{2\mathbb{M}}(m) \\ \rightsquigarrow & \langle -1,-1 \rangle & \cup & \langle 0,0 \rangle & \cup & \langle 1,1 \rangle & \cup & \langle 2,2 \rangle \\ \rightsquigarrow & \langle -1,2 \rangle, & & & & & & \end{array}$$

with

$$C = \{\xi_1^j\},\ \{\xi^j\} = \Xi_2^\star(g),\ \xi^j = [j, j+1),$$
$$\xi_1^j = \xi_1 \cap \xi^j,\ g_1^j = g|\xi_1^j;\ j \in \{-1,0,1,2\}.$$

Perusal of the following figure may ease the comprehension of the preceding evaluation. The grey region represents $g(m)$, while $g^{\mathbb{M}}(m)$ is displayed as the upper and lower solid lines.



$g^{\mathbb{M}}(m)$

The evaluation of $g^{\mathbb{M}}(m)$,

$$m = \langle -1+2\alpha, -1+3\alpha \rangle,$$

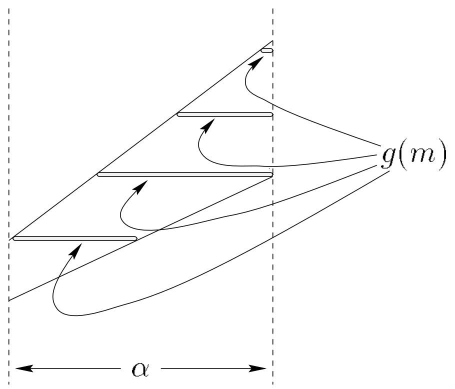
proceeds as follows, using the second sectioning:

$$\begin{array}{llccccl} & & g^{\mathbb{M}}(m) & & & & \\ \rightsquigarrow & \bigcup_{j\in[0,1)} & g_1^{j\mathbb{M}}(m) & & & & \\ \rightsquigarrow & \bigcup_{j\in[0,1)} & \langle & g_1^j(-1+2\alpha)^{\mathbb{M}-} & , & g_1^j(-1+3\alpha)^{\mathbb{M}+} & \rangle, \text{ since } \psi_1^{\uparrow}(g_1^j) \\ \rightsquigarrow & \bigcup_{j\in[0,1)} & \langle & g_1^j(-1+2\alpha) & , & g_1^j(-1+3\alpha) & \rangle, \text{ since } \psi_2^0(g_1^j) \\ \rightsquigarrow & \bigcup_{j\in[0,1)} & \langle & -1-j+2\alpha & , & -1-j+3\alpha & \rangle \\ \rightsquigarrow & & \langle & -2+2\alpha & , & -1+3\alpha & \rangle \end{array}$$

with

$$C = \{\xi_1^j\},\ \{\xi^j\} = \Xi_2^\star(g),\ \xi^j = \{\ldots, j-1, j, j+1, \ldots\},$$
$$\xi_1^j = \xi_1 \cap \xi^j,\ g_1^j = g|\xi_1^j;\ j \in [0,1).$$

The following figure graphically illustrates portions of the preceding evaluation.



$g^{\mathbb{M}}(m)$

Which of the two methods is used depends on the values of

$$(m^- - 1) - \lfloor m^- \rfloor \text{ and } m^+ - \lfloor m^+ \rfloor.$$

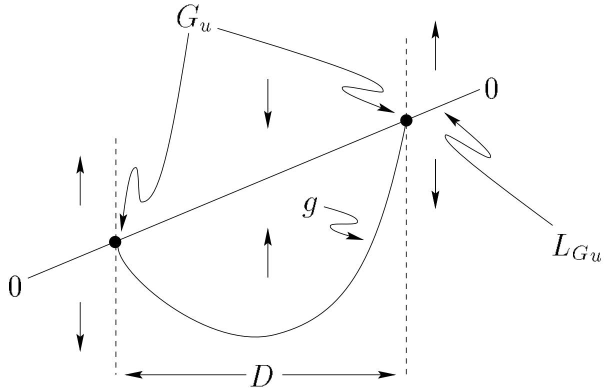
The same method need not be used for both bounds. With a reasonable choice of ω, an optimal bound is easily computed.

3.3.10 Concave Up Functions

We will determine $g^{\mathbb{M}+}$ for any concave up function $g : \mathbb{R}^\star \mapsto \mathbb{R}^\star$. Since g is concave up, $\psi_2^\uparrow(g)$. Let $D = \text{dom}(g)$; we assume that $\{D^-, D^+\} \subseteq D \subseteq [0,1]$, so we may take $G_u = \{(D^-, g(D^-)), (D^+, g(D^+))\}$. A simple proof by contradiction, which follows, shows that L_{G_u} is an upper bound for g:

$$\forall[(x,y) \in g]\ L_{G_u}(x) \geq y.$$

Assume that there is a point $(x,y) \in g$ such that $L_{G_u}(x) < y$. Let $G = G_u \cup \{(x,y)\}$, so $G \subseteq_3 g$. Furthermore, $G \subseteq g$ and $\psi_2^\uparrow(g)$ imply that $\psi_2^\uparrow(G)$.



ψ_2 *Chart*

A quick review of the ψ_2 chart reveals that this situation is impossible. There is no $(x,y) \in g$ such that $L_{G_u}(x) < y$ since $\psi_1^\uparrow(G)$, $G_u = \{(D^-, g(D^-)), (D^+, g(D^+))\}$, and $D^- \leq x \leq D^+$.

The assumptions made do not overly restrict the applicability of the proof.

- If $D \not\subseteq [0,1]$, consider $g|[0,1]$ in place of g.

- If $D^- \notin D$, consider $g' = g \cup \{(D^-, y)\}$ in place of g, such that g' is concave up. If $\lim_{x \to D^-} g(x)$ exists, it may be taken for y; otherwise, a trivial upper bound may be used.
- If $D^+ \notin D$, consider $g' = g \cup \{(D^+, y)\}$ in place of g, such that g' is concave up. If $\lim_{x \to D^+} g(x)$ exists, it may be taken for y; otherwise, a trivial upper bound may be used.

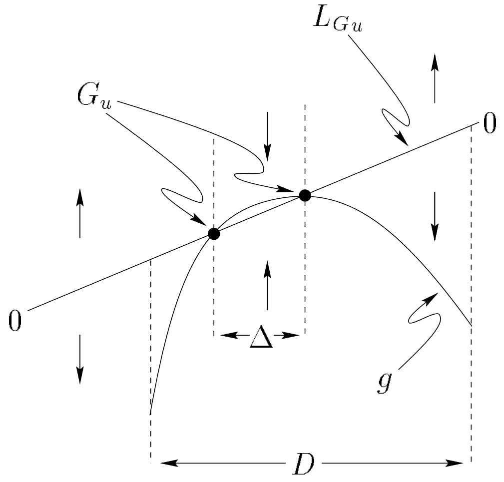
The bound is optimal, since L_{G_u} may not be lowered. Lowering L_{G_u} would lower $L_{G_u}(D^-)$ or $L_{G_u}(D^+)$, invalidating L_{G_u} as an upper bound.

3.3.11 Concave Down Functions

We will determine $g^{\mathbb{M}+}$ for any concave down function $g : \mathbb{R}^\star \mapsto \mathbb{R}^\star$. Since g is concave down, $\psi_2^\downarrow(g)$. Let $D = \mathrm{dom}(g)$; we assume that $\{x_1, x_1 + \Delta\} \subseteq D$, so we may take $G_u = \{(x_1, g(x_1)), (x_1 + \Delta, g(x_1 + \Delta))\}$. A simple proof by contradiction, which follows, shows that L_{G_u} is an upper bound for g, excepting $x \in (x_1, x_1 + \Delta)$:

$$\forall[(x, y) \in g] \quad L_{G_u}(x) \geq y \;\vee\; x \in (x_1, x_1 + \Delta).$$

Assume that there is a point $(x, y) \in g$ such that $L_{G_u}(x) < y$. Let $G = G_u \cup \{(x, y)\}$, so $G \subseteq_3 g$. Furthermore, $G \subseteq g$ and $\psi_2^\uparrow(g)$ imply that $\psi_2^\uparrow(G)$.



ψ_2 *Chart*

A quick review of the ψ_2 chart reveals that this situation is impossible. There is no $(x, y) \in g$, $x \notin (x_1, x_1 + \Delta)$ such that $L_{G_u}(x) < y$ since $\psi_1^\uparrow(G)$, $G_u = \{(x_1, g(x_1)), (x_1 + \Delta, g(x_1 + \Delta))\}$. We may take Δ to be infinitesimal for g differentiable at x_1; this corresponds to having L_{G_u} match, at x_1, both the value and the derivative of g. For discrete g we may take x_1 and $x_1 + \Delta$ to be neighbours. With such a choice of Δ, L_{G_u} is an upper bound for g since x may not be a member of $(x_1, x_1 + \Delta)$.

The assumptions made do not overly restrict the applicability of the proof.

For differentiable g and constant ω, the bound is optimal when $x = \frac{1}{2}(D^- + D^+)$; for other reasonable choices of ω, the optimal bound is similarly easy to determine. See section 3.4.3 for details.

3.3.12 Lower Bounds

Again, we concentrate on upper bounds, as lower bounds may be easily constructed using the rules given for upper bounds. The comments given in section 3.2.12 apply to linear bounds.

3.3.13 Example with a Concave Function

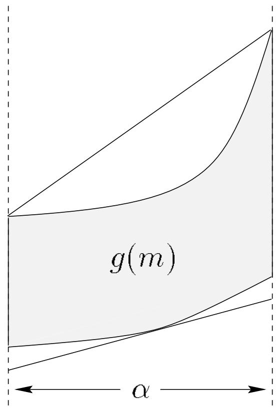
Consider the exponentiation function, $g(x) = e^x : \mathbb{R}^\star \mapsto \mathbb{R}^\star$, which is a globally concave up function. An example evaluation follows:

$$
\begin{array}{lllll}
 & g^{\mathbb{M}}(\langle -3+2\alpha, 11-4\alpha\rangle) & & & \\
\rightsquigarrow & {g_1^1}^{\mathbb{M}}(\langle -3+2\alpha, 11-4\alpha\rangle) & & & \\
\rightsquigarrow & \langle \quad g_1^1(-3+2\alpha)^{\mathbb{M}-} & , & g_1^1(11-4\alpha)^{\mathbb{M}+} & \rangle,\ \text{since } \psi_1^{\uparrow}(g_1^1) \\
\rightsquigarrow & \langle \quad g_l(p) - p\frac{d}{dx}g_l(p) + \frac{d}{dx}g_l(p)\alpha & , & g_u(0) + (g_u(1) - g_u(0))\alpha & \rangle,\ \text{since } \psi_2^{\uparrow}(g_1^1), \\
 & \quad g_l = g_1^1(-3+2x) & & g_u = g_1^1(11-4x) & \\
\rightsquigarrow & \langle \quad 2e^{-2}\alpha & , & e^7 + (e^3 - e^7)\alpha & \rangle,\ p = \frac{1}{2}
\end{array}
$$

with

$$C = \{\xi_1^1\},\ \{\xi_1^1\} = \Xi_1^\star(g) \subseteq \Xi_1(g),\ \{\xi_1^1\} = \Xi_2^\star(g) \subseteq \Xi_2(g),$$
$$\xi_1^1 = [-\infty, \infty],\ g_1^1 = g|\xi_1^1.$$

Portions of the preceding evaluation are graphically illustrated by the following figure:



$g^{\mathbb{M}}(m)$

3.3.14 Example with a Piecewise Concave Function

Consider the cubing function, $g(x) = x^3 : \mathbb{R}^\star \mapsto \mathbb{R}^\star$, which is a piecewise concave function. An example evaluation follows:

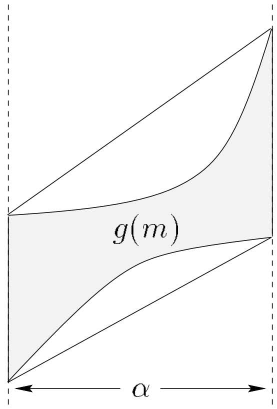
$$
\begin{array}{lllll}
 & g^{\mathbb{M}}(\langle -2+\alpha, 4-\alpha\rangle) & & & \\
\rightsquigarrow & \langle \quad g_1^1(-2+\alpha)^{\mathbb{M}-} & , & g_1^2(4-\alpha)^{\mathbb{M}+} & \rangle,\ \text{since } \psi_1^{\uparrow}(g_1^1 \cup g_1^2) \\
\rightsquigarrow & \langle \quad g_l(0) + (g_l(1) - g_l(0))\alpha & , & g_u(0) + (g_u(1) - g_u(0))\alpha & \rangle,\ \text{since } \psi_2^{\downarrow}(g_1^1),\ \psi_2^{\uparrow}(g_1^2), \\
 & \quad g_l = g_1^1(-2+x) & & g_u = g_1^2(4-x) & \\
\rightsquigarrow & \langle \quad -8+7\alpha & , & 64-37\alpha & \rangle,
\end{array}
$$

with

$$C = \{\xi_1^1, \xi_1^2\},\ \{\xi_1\} = \Xi_1^\star(g) \subseteq \Xi_1(g),\ \{\xi^j\} = \Xi_2^\star(g) \subseteq \Xi_2(g),$$

$$\xi_1 = [-\infty, \infty],\ \xi^1 = [-\infty, 0],\ \xi^2 = [0, \infty],$$
$$\xi_1^j = \xi_1 \cap \xi^j,\ g_1^j = g|\xi_1^j;\ j \in \{1, 2\}.$$

Examination of the following figure may aid the reader's comprehension of the preceding evaluation.



$g^{\mathbb{M}}(m)$

3.3.15 Periodic Functions

As with constant interval arithmetic, special care should be taken when evaluating periodic functions to avoid unnecessary computation.

We will again cut the function $g : \mathbb{R}^\star \mapsto \mathbb{R}^\star$ into sections where each section attains the extreme values of g:

$$\Xi_\pm^{\mathbb{M}}(g) \ \equiv_{\text{def}}\ \{m : m \in \mathbb{M},\ \forall g^{\mathbb{M}}\ \ \langle g_-, g_+\rangle \subseteq g^{\mathbb{M}}(m)\},$$
$$g_- = \inf_{(x,y)\in g} y, \quad g_+ = \sup_{(x,y)\in g} y,$$

where

$$m \subseteq^{\mathbb{M}} n \ \equiv_{\text{def}}\ \forall[\alpha \in [0,1]]\ \forall[x \in^{\mathbb{M}} m(\alpha)]\ x \in^{\mathbb{M}} n(\alpha).$$

When evaluating $g^{\mathbb{M}}(m)$, we may simply return $\langle g_-, g_+\rangle$ if any of the aforementioned sections lie within m:

$$\exists[m_\pm \in \Xi_\pm^{\mathbb{M}}(g),\ m_\pm \subseteq m]\ g^{\mathbb{M}}(m) \rightsquigarrow g^{\mathbb{M}}(\langle -\infty, \infty\rangle).$$

As with the previous sectioning scheme, there will often be a preferred sectioning, denoted by $\Xi_\pm^{\star\mathbb{M}}(g)$, which we will use to check containment.

The preferred sectioning $\Xi_\pm^{\star\mathbb{M}}$ of the sine function includes members from the preferred sectioning $\Xi_\pm^{\star\mathbb{J}}$ of the sine function:

$$\{\langle x, x + 2\pi\rangle : x \in \mathbb{R}\} \subset \Xi_\pm^{\star\mathbb{M}}(\sin),$$
$$\{\langle k\pi, (k+1)\pi\rangle : k \in \mathbb{Z}\} \subset \Xi_\pm^{\star\mathbb{M}}(\sin).$$

In general, all members of $\Xi_\pm^{\star\mathbb{J}}(g)$ may be transferred into $\Xi_\pm^{\star\mathbb{M}}(g)$, since $\Xi_\pm^{\mathbb{J}}(g)$ may be defined without reference to the underlying interval number system. We may add another set of intervals to $\Xi_\pm^{\star\mathbb{M}}(\sin)$:

$$\{\langle a + b\alpha, c + d\alpha\rangle : \max(|b|, |d|) \geq 4\pi\} \subset \Xi_\pm^{\star\mathbb{M}}(\sin).$$

This set is intrinsic to linear interval arithmetic: it need not transfer to another polynomial interval arithmetic.

3.3.16 Partial Functions

We have considered implementing a model $g^{\mathbb{M}}$, given $g^{\mathbb{R}}$. We now consider implementing $g^{\mathbb{M}|F^|}$. As before,

$$\Xi_|(g) \equiv_{\rm def} \{x : \mathcal{P}_|(g,x)\} \equiv \mathrm{dom}(G).$$

The function Φ_ζ,

$$\Phi_\zeta : \mathbb{M}^{|F^|} \times 2^{\mathbb{R}^\star} \mapsto \mathbb{M},$$

when given an interval m and a set ξ of extended real numbers, produces a valid description of the relationship between m and ξ, in terms of the provided set ζ, of extended real numbers:

$$\Phi_\zeta(m,\xi) \rightsquigarrow d, \quad \forall[\alpha \in [0,1]]\ (m(\alpha) \in \xi) \sqsubseteq (d(\alpha) \in \zeta).$$

The relationship between each interval and its associated set is that of containment, defined as before:

$$m(\alpha) \in \xi \equiv \bigcup_{x \in m(\alpha)} x \in \xi, \qquad d(\alpha) \in \zeta \equiv \bigcup_{x \in d(\alpha)} x \in \zeta.$$

The function Φ_ζ "translates" from ξ to ζ.

For the function $g : \mathbb{R}^\star \mapsto \mathbb{R}^\star$, an evaluation of the model $g^{\mathbb{M}|F^|}$ proceeds as follows:

$$g^{\mathbb{M}|F^|}(m) \rightsquigarrow m'; \quad m = \langle v|f(d)\rangle,\ m' = \langle v'|f'(d')\rangle.$$

The evaluation of $g^{\mathbb{M}|F^|}$ is analogous to the evaluation of $g^{\mathbb{J}|\mathbb{T}}$.

The resulting domain description $f'(d')$; $d' \in \mathbb{M}$, $f' \in F^|$; is determined using $f(d)$, ξ, ζ, and Φ_ζ:

$$f'(d') = f(d) \wedge f_i^|(\Phi_\zeta(m,\xi)).$$

The set ξ, given by $\Xi_|(g)$, corresponds to $\mathcal{P}_|(g,x)$:

$$\xi = \Xi_|(g) = \{x : \mathcal{P}_|(g,x)\}.$$

The set ζ, given indirectly by $\mathrm{Z}_{F^|}(m,\xi)$, similarly corresponds to $f_i^| \in F^|$:

$$\zeta = \{x : f_i^|(x)\},\ f_i^| = \mathrm{Z}_{F^|}(m,\xi);$$

the function $f_i^|$ is chosen, by $\mathrm{Z}_{F^|}$, to facilitate the impending computation of $\Phi_\zeta(m,\xi)$. The chosen $f_i^|$ is used to describe the domain of $g^{\mathbb{M}|F^|}(m)$.

The resulting value v', $v' \in \mathbb{M}^{|F^|}$, depends on $f'(d')$. If $f'(d') \neq \mathrm{F}$, the resulting value is given by the methods outlined earlier:

$$f'(d') \neq \mathrm{F} \ \Rightarrow\ v' = g^{\mathbb{M}}(v).$$

If $f'(d') = \mathrm{F}$, the resulting value is arbitrary:

$$f'(d') = \mathrm{F} \ \Rightarrow\ v' = \langle -\infty, \infty\rangle.$$

3.3.17 Examples with a Partial Function

We now consider an example partial function, the square root function:

$$g : \mathbb{R}^\star \mapsto \mathbb{R}^\star,\ g(x) = \sqrt{x}.$$

The function g is defined for non-negative extended real numbers:

$$\Xi_{|}(g) = [0, \infty],\quad \xi_{\geq 0} = \xi_{|} = \Xi_{|}(g).$$

We let the domain description set include $f_{\geq 0}$:

$$f_{\geq 0} : \mathbb{R}^\star \mapsto \mathbb{B},\quad f_{\geq 0}(x) = (x \geq 0),\quad f_{\geq 0} \in F^{|}.$$

This allows a trivial implementation of $\mathrm{Z}_{F^{|}}(m, \xi_{\geq 0})$:

$$\mathrm{Z}_{F^{|}}(m, \xi_{\geq 0}) = f_{\geq 0}.$$

Which, in turn, allows a trivial implementation of $\phi_{\zeta_{\geq 0}}(\langle v|d\rangle, \xi_{\geq 0})$:

$$\phi_{\zeta_{\geq 0}}(\langle v|d\rangle, \xi_{\geq 0}) = v,\quad \zeta_{\geq 0} = \{x : f_{\geq 0}(x)\};$$

if $\phi_{\zeta_{\geq 0}}$ may return a member of $\mathbb{M}^{\between}$, as is the case when implementing $\mathbb{M}^{\between|F^{|}}$ models. If $\phi_{\zeta_{\geq 0}}$ must return a member of $\mathbb{M}$, the result may simply be demoted:

$$f_{\geq 0}(\phi_{\zeta_{\geq 0}}(\langle v|d\rangle, \xi_{\geq 0})) = f_{\geq 0}(v)^{f_{\geq 0}(\mathbb{M}^{\between}) \to f_{\geq 0}(\mathbb{M})}.$$

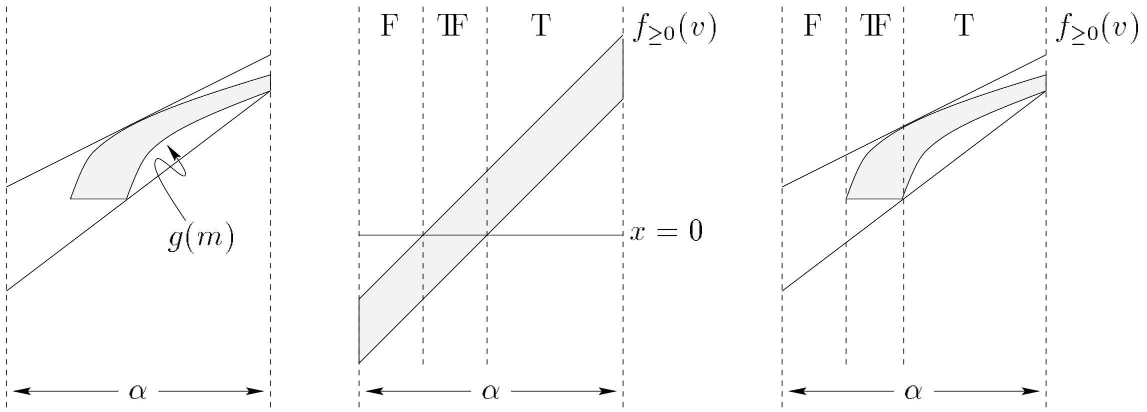
The evaluation of $g^{\mathbb{M}^{|F^{|}}}(m)$;

$$m = \langle v|d\rangle,\ v = \langle -2 + 4\alpha, -1 + 4\alpha\rangle,\ d = \mathrm{T};$$

proceeds as follows:

$$\begin{array}{lllll}
 & g^{\mathbb{M}^{|F^{|}}}(\langle\langle -2+4\alpha, -1+4\alpha\rangle|\mathrm{T}\rangle) & & & \\
\rightsquigarrow & \langle\ \ g^{\mathbb{M}^{\between}}(\langle -2+4\alpha, -1+4\alpha\rangle) & |\ \mathrm{T} \wedge f_{\geq 0}(\phi_{\zeta_{\geq 0}}(m, \xi_{\geq 0})) & \rangle, & \text{since } \mathrm{Z}_{F^{|}}(m, \xi_{\geq 0}) = f_{\geq 0} \\
\rightsquigarrow & \langle\ \ g^{\mathbb{M}^{\between}}(\langle -2+4\alpha, -1+4\alpha\rangle) & |\ \mathrm{T} \wedge f_{\geq 0}(v) & \rangle, & \text{since } \phi_{\zeta_{\geq 0}}(m, \xi_{\geq 0}) = v \in \mathbb{M} \\
\rightsquigarrow & \langle\ \langle -\sqrt{2} + 2\sqrt{2}\alpha, \frac{1}{2}\sqrt{2} + \sqrt{2}\alpha\rangle & |\ f_{\geq 0}(v) & \rangle. &
\end{array}$$

The following figures are presented to aid the reader in understanding the preceding evaluation.



$g^{\mathbb{M}}(v)$ $\qquad$ $f_{\geq 0}(v)$ $\qquad$ $g^{\mathbb{M}}(v)$ *with* $f_{\geq 0}(v)$

The evaluation of $g^{\mathbb{M}^{|F|}}(m)$;

$$m = \langle v | f_{\geq 0}(d)\rangle,\ v = \langle -1 + 2\alpha, -\tfrac{5}{6} + \tfrac{16}{9}\alpha\rangle,\ d = \langle 3 - 4\alpha, 3 - 4\alpha\rangle;$$

proceeds as follows:

$$\begin{array}{lllll}
 & g^{\mathbb{M}^{|F|}}(\langle\langle -1 + 2\alpha, -\tfrac{5}{6} + \tfrac{16}{9}\alpha\rangle | f_{\geq 0}(d)\rangle & & & \\
\rightsquigarrow & \langle\ g^{\mathbb{M}^{\between}}(\langle -1 + 2\alpha, -\tfrac{5}{6} + \tfrac{16}{9}\alpha\rangle) & | & f_{\geq 0}(d) \wedge f_{\geq 0}(\Phi_{\zeta_{\geq 0}}(m, \xi_{\geq 0})) & \rangle,\ \text{since } \mathrm{Z}_{F|}(m, \xi_{\geq 0}) = f_{\geq 0} \\
\rightsquigarrow & \langle\ g^{\mathbb{M}^{\between}}(\langle -1 + 2\alpha, -\tfrac{5}{6} + \tfrac{16}{9}\alpha\rangle) & | & f_{\geq 0}(d) \wedge f_{\geq 0}(v)^{f_{\geq 0}(\mathbb{M}^{\between}) \to f_{\geq 0}(\mathbb{M})} & \rangle \\
\rightsquigarrow & \langle\ \langle -1 + 2\alpha, -\tfrac{7}{6}\sqrt{2} + \tfrac{8}{3}\sqrt{2}\alpha\rangle & | & f_{\geq 0}(d) \wedge f_{\geq 0}(w) & \rangle \\
\rightsquigarrow & \langle\ \langle -1 + 2\alpha, -\tfrac{7}{6}\sqrt{2} + \tfrac{8}{3}\sqrt{2}\alpha\rangle & | & f_{\geq 0}(d') & \rangle,
\end{array}$$

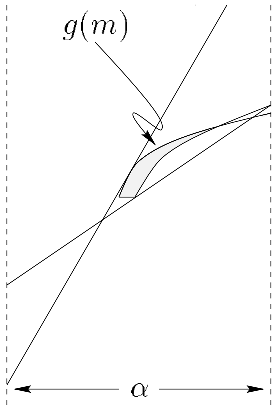
with

$$w = \langle -\tfrac{15}{32} + \alpha, -\tfrac{1}{2} + \alpha\rangle,$$

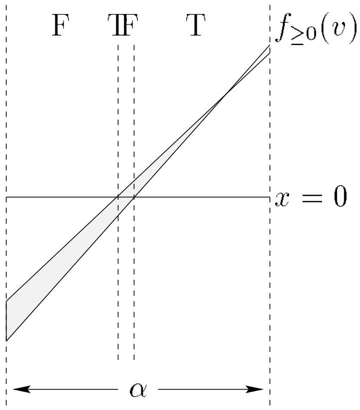
and

$$d' = d'_1 = \langle -1, \tfrac{3}{4} - \alpha\rangle,\ \text{or}\ d' = d'_2 = \langle -1, -\tfrac{15}{32} + \alpha\rangle.$$

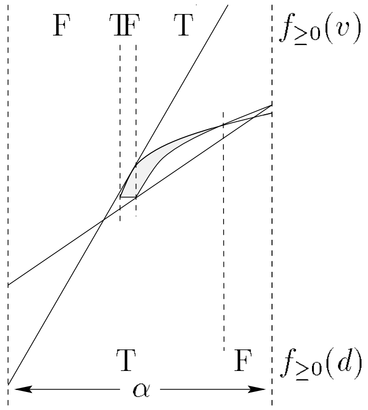
Perusal of the following figures may ease the comprehension of the preceding evaluation.



$g^{\mathbb{M}^{\between}}(v)$

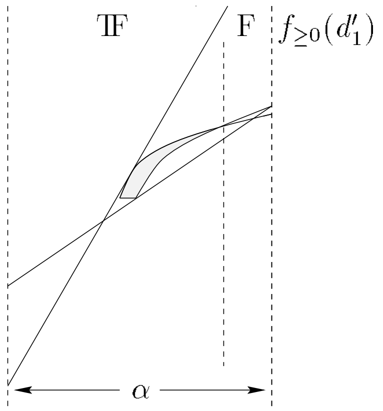


$f_{\geq 0}(v)$

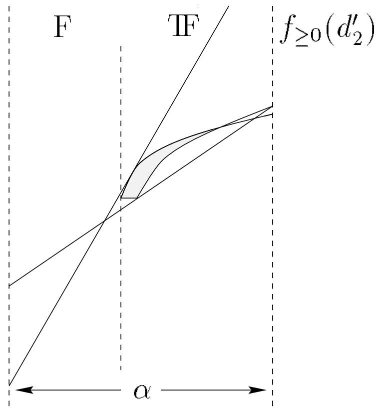


$g^{\mathbb{M}^{\between}}(v)$ *with* $f_{\geq 0}(v)$ *and* $f_{\geq 0}(d)$

Since the evaluation is of a $\mathbb{M}^{|F|}$ model, the domain constraint must be folded into a single constraint. An evaluation of a $\mathbb{M}^{\hat{|}F|}$ model may finish earlier, with the domain described by $f_{\geq 0}(d) \wedge f_{\geq 0}(w)$ rather than $f_{\geq 0}(d')$.



$g^{\mathbb{M}^{\between}}(v)$ *with* $f_{\geq 0}(d'_1)$

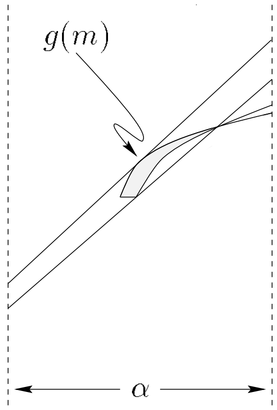


$g^{\mathbb{M}^{\between}}(v)$ *with* $f_{\geq 0}(d'_2)$

Note that a better bound is possible by taking $f_i^{|}(d)$ into account when determining v':

$$v' = (g|f_i^{|}(d))^{\mathbb{M}^{\between}}(v).$$

With such an approach, the bound appears as follows.



$(g|f_{\geq 0}(d))^{\mathbb{M}^{\between}}(v)$

3.3.18 Discontinuous Functions

We now consider implementing $g^{\mathbb{M}^{\Delta F^{\Delta}}}$. As before,

$$\Xi_{\Delta}(g) \;\equiv_{\text{def}}\; \{x : \mathcal{P}_{\Delta}(g, x)\}.$$

For the function $g : \mathbb{R}^{\star} \mapsto \mathbb{R}^{\star}$, an evaluation of the model $g^{\mathbb{M}^{\Delta F^{\Delta}}}$ proceeds as follows:

$$g^{\mathbb{M}^{\Delta F^{\Delta}}}(\langle v \Delta f(d) \rangle) \rightsquigarrow \langle v' \Delta f'(d') \rangle.$$

The evaluation of $g^{\mathbb{M}^{\Delta F^{\Delta}}}$ is analogous to the evaluation of $g^{\mathbb{J}^{\Delta\mathbb{T}}}$ or $g^{\mathbb{M}^{|F|}}$.

The resulting continuity description $f'(d')$ is determined using $f(d)$, ξ, ζ, and ϕ_{ζ}:

$$f'(d') = f(d) \wedge f_i^{|}(\phi_{\zeta}(m, \xi));$$

$$\xi = \Xi_{\Delta}(g) = \{x : \mathcal{P}_{\Delta}(g, x)\},\ \zeta = \{x : f_i^{\Delta}(x)\},\ f_i^{\Delta} = \mathrm{Z}_{F^{\Delta}}(m, \xi).$$

The resulting value v', for $v' \in \mathbb{M}$ or $v' \in \mathbb{M}^{\between}$, is given by the methods outlined earlier.

3.3.19 Examples with a Discontinuous Function

We now consider an example discontinuous function, the floor function:

$$g : \mathbb{R}^{\star} \mapsto \mathbb{R}^{\star},\ g(x) = \lfloor x \rfloor.$$

The function g is continuous for non-integral arguments:

$$\Xi_{\Delta}(g) = \{(k, k+1) : k \in \mathbb{Z}\},\quad \xi_{\neq k} = \xi_{\Delta} = \Xi_{\Delta}(g).$$

We let the continuity description set include $f_{\neq k}$:

$$f_{\neq k} : \mathbb{R}^{\star} \mapsto \mathbb{B}, \quad f_{\neq k}(x) = (x \notin \mathbb{Z}), \quad f_{\neq k} \in F^{\Delta}.$$

This allows a trivial implementation of $\mathrm{Z}_{F^{\Delta}}(m, \xi_{\neq k})$:

$$\mathrm{Z}_{F^{\Delta}}(m, \xi_{\neq k}) = f_{\neq k}.$$

Which, in turn, allows a trivial implementation of $\phi_{\zeta_{\neq k}}(\langle v|d\rangle, \xi_{\neq k})$:

$$\phi_{\zeta_{\neq k}}(\langle v|d\rangle, \xi_{\neq k}) = v, \quad \zeta_{\neq k} = \{x : f_{\neq k}(x)\};$$

since v is a member of $\mathbb{M}$. If a $\mathbb{M}^{\Delta F}$ model is evaluated, the resulting continuity description v may be demoted, as was done previously with domain descriptions.

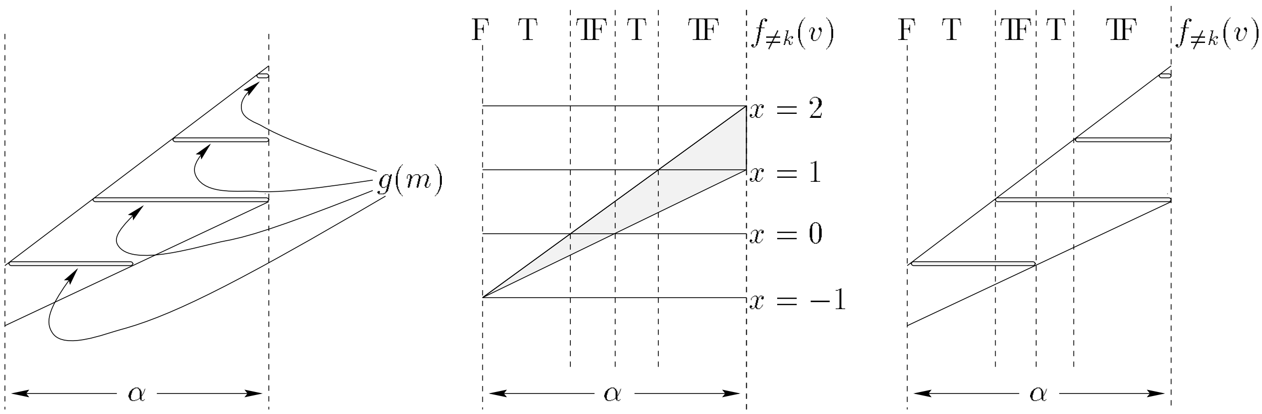
The evaluation of $g^{\mathbb{M}^{\Delta F^{\Delta}}}(m)$;

$$m = \langle v \Delta d\rangle, \; v = \langle -1 + 2\alpha, -1 + 3\alpha\rangle, \; d = \mathrm{T};$$

proceeds as follows:

$$\begin{array}{lllcll}
 & & g^{\mathbb{M}^{\Delta F^{\Delta}}}(\langle\langle -1+2\alpha, -1+3\alpha\rangle|\mathrm{T}\rangle) & & & \\
\rightsquigarrow & \langle & g^{\mathbb{M}}(\langle -1+2\alpha, -1+3\alpha\rangle) & | & \mathrm{T} \wedge f_{\neq k}(\phi_{\zeta_{\neq k}}(m, \xi_{\neq k})) & \rangle, \text{ since } \mathrm{Z}_{F^{\Delta}}(m, \xi_{\neq k}) = f_{\neq k} \\
\rightsquigarrow & \langle & g^{\mathbb{M}}(\langle -1+2\alpha, -1+3\alpha\rangle) & | & \mathrm{T} \wedge f_{\neq k}(v) & \rangle, \text{ since } \phi_{\zeta_{\neq k}}(m, \xi_{\neq k}) = v \\
\rightsquigarrow & \langle & \langle -2+2\alpha, -1+3\alpha\rangle & | & f_{\neq k}(v) & \rangle.
\end{array}$$

The following figures graphically illustrate portions of the preceding evaluation.



$g^{\mathbb{M}}(v)$ $\quad$ $f_{\neq k}(v)$ $\quad$ $g^{\mathbb{M}}(v)$ *with* $f_{\neq k}(v)$

Let F^{Δ} include $f_{\neq 0}$:

$$f_{\neq 0} : \mathbb{R}^{\star} \mapsto \mathbb{B}, \quad f_{\neq 0}(x) = (x \neq 0), \quad f_{\neq 0} \in F^{\Delta}.$$

We will now illustate, more fully, the role of Z and ϕ.

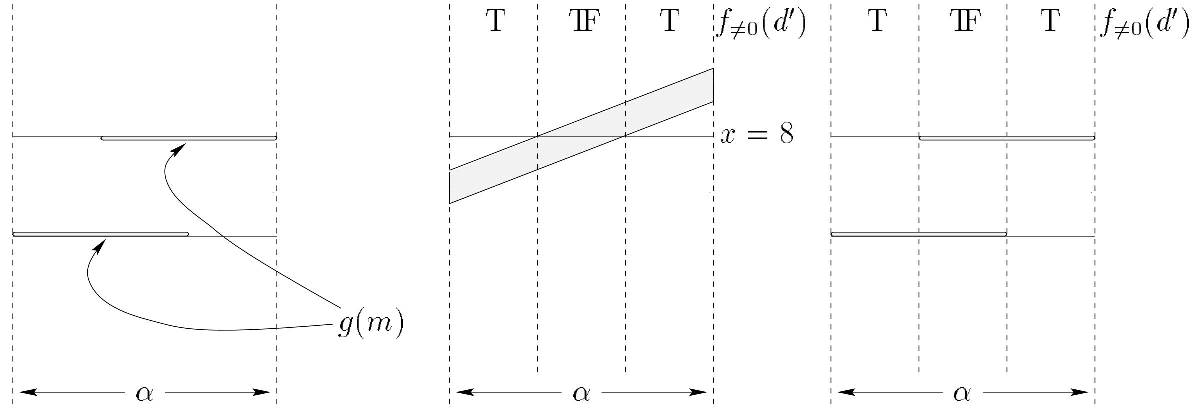
The evaluation of $g^{\mathbb{M}^{\Delta F^{\Delta}}}(m)$;

$$m = \langle v \Delta d\rangle, \; v = \langle 7\tfrac{1}{3} + \alpha, 7\tfrac{2}{3} + \alpha\rangle, \; d = \mathrm{T};$$

proceeds as follows:

$$
\begin{array}{lllcll}
 & & g^{\mathbb{M}^{\triangle F^{\triangle}}}(\langle\langle 7\frac{1}{3}+\alpha, 7\frac{2}{3}+\alpha\rangle|\mathrm{T}\rangle) & & & \\
\rightsquigarrow & \langle & g^{\mathbb{M}}(\langle 7\frac{1}{3}+\alpha, 7\frac{2}{3}+\alpha\rangle) & | & \mathrm{T}\wedge f_{\neq 0}(\Phi_{\zeta_{\neq 0}}(m,\xi_{\neq k})) & \rangle, \text{ since } \mathrm{Z}_{F^{\triangle}}(m,\xi_{\neq k}) = f_{\neq 0} \\
\rightsquigarrow & \langle & g^{\mathbb{M}}(\langle 7\frac{1}{3}+\alpha, 7\frac{2}{3}+\alpha\rangle) & | & \mathrm{T}\wedge f_{\neq 0}(v-8) & \rangle, \text{ since } \Phi_{\zeta_{\neq 0}}(m,\xi_{\neq k}) = d' = v-8 \\
\rightsquigarrow & \langle & \langle 7, 8\rangle & | & f_{\neq 0}(d') & \rangle.
\end{array}
$$

Portions of the preceding evaluation are depicted in the following figures.



$g^{\mathbb{M}}(v)$ $\qquad$ $f_{\neq 8}(v)$, *or* $f_{\neq 0}(d')$ $\qquad$ $g^{\mathbb{M}}(v)$ *with* $f_{\neq 0}(d')$

3.3.20 Bumpy Functions

We now consider implementing $g^{\mathbb{M}^{|F^|}\star}$ models. Each $\mathbb{M}^{|F^|}\star$ interval is given by a set of $\mathbb{M}^{|F^|}$ intervals. We previously defined the union of two $\mathbb{M}$ intervals. Extending that definition results in the following method for taking the union of two $\mathbb{M}^{|F^|}$ intervals:

$$\langle m_v|m_d\rangle \cup^{\mathbb{J}^{|\mathbb{T}}} \langle n_v|n_d\rangle \;\equiv_{\mathrm{def}}\; \langle m_v \cup^{\mathbb{M}} n_v | m_d \vee n_d\rangle; \quad \langle m_v|n_d\rangle \in \mathbb{M}^{|F^|}, \langle n_v|m_d\rangle \in \mathbb{M}^{|F^|}.$$

Another method, which uses $\mathbb{J}^{|F^|}\star$ to describe the result, follows:

$$m \cup^{\mathbb{M}^{|F^|} \to \mathbb{M}^{|F^|}\star} n \;\equiv_{\mathrm{def}}\; \{m, n\}; \quad m \in \mathbb{M}^{|F^|}, n \in \mathbb{M}^{|F^|}.$$

Since each $\mathbb{M}^{|F^|}\star$ interval is a collection of $\mathbb{M}^{|F^|}$ intervals, the union of two $\mathbb{M}^{|F^|}\star$ intervals is simply the sum of the two collections:

$$m \cup^{\mathbb{M}^{|F^|}\star} n = \{m_0, m_1, \ldots, m_j\} \cup^{\mathbb{J}^{|\mathbb{T}\star}} \{n_0, n_1, \ldots, n_k\} \;\equiv_{\mathrm{def}}\; \{m_0, m_1, \ldots, m_j, n_0, n_1, \ldots, n_k\}.$$

Good models of $\mathbb{M}$-bumpy functions may be built using the methods presented so far, using $\cup^{\mathbb{M}^{|F^|} \to \mathbb{M}^{|F^|}\star}$ rather than $\cup^{\mathbb{M}^{|F^|}}$, where appropriate.

3.3.21 Examples with Bumpy Functions

Let the domain description set include $f_{>0}$:

$$f_{>0} : \mathbb{R}^\star \mapsto \mathbb{B}, \quad f_{>0}(x) = (x > 0), \quad f_{>0} \in F^{|}.$$

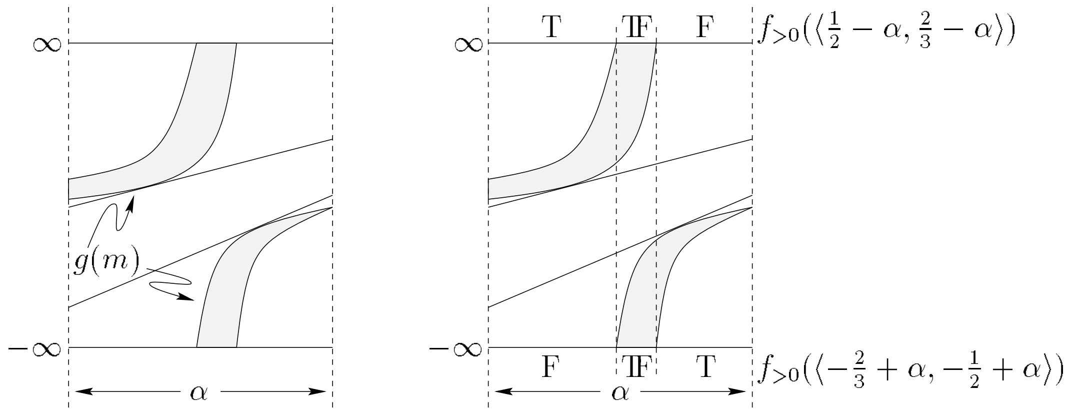
We will evaluate a $\mathbb{M}^{|F|_\star}$ model of the multiplicative inverse: let $g(x) = x^{-1}$. The evaluation of $g^{\mathbb{M}^{|F|_\star}}(m)$,

$$m = \{\langle\langle 1-2\alpha, 2-3\alpha\rangle|\mathrm{T}\rangle\},$$

proceeds as follows:

$$\begin{array}{lll}
 & g^{\mathbb{M}^{|F|_\star}}(\{\langle\langle 1-2\alpha, 2-3\alpha\rangle|\mathrm{T}\rangle\}) & \\
\rightsquigarrow & g^{\mathbb{M}^{|F|}}(\langle\langle 1-2\alpha, 2-3\alpha\rangle|\mathrm{T}\rangle) & \\
\rightsquigarrow & {g_1}^{\mathbb{M}^{|F|}}(\langle\langle 1-2\alpha, 2-3\alpha\rangle|\mathrm{T}\rangle) \quad \cup^{\mathbb{M}^{|F|}\to\mathbb{M}^{|F|_\star}} \quad {g_2}^{\mathbb{M}^{|F|}}(\langle\langle 1-2\alpha, 2-3\alpha\rangle|\mathrm{T}\rangle) & \\
\rightsquigarrow & \langle\langle -\infty, -8+8\alpha\rangle|f_{>0}(\langle -\frac{2}{3}+\alpha, -\frac{1}{2}+\alpha\rangle)\rangle \quad \cup^{\mathbb{M}^{|F|}\to\mathbb{M}^{|F|_\star}} \quad \langle\langle 3\alpha, \infty\rangle|f_{>0}(\langle \frac{1}{2}-\alpha, \frac{2}{3}-\alpha\rangle)\rangle & \\
\rightsquigarrow & \{\langle\langle -\infty, -8+8\alpha\rangle|f_{>0}(\langle -\frac{2}{3}+\alpha, -\frac{1}{2}+\alpha\rangle)\rangle, \langle\langle 3\alpha, \infty\rangle|f_{>0}(\langle \frac{1}{2}-\alpha, \frac{2}{3}-\alpha\rangle)\rangle\}. &
\end{array}$$

One's intuition of the preceding evaluation may be developed by careful scrutiny of the following figures.



$g^{\mathbb{M}}(v)$ $\qquad$ $g^{\mathbb{M}}(v)$ *with* $f_{>0}(v)$

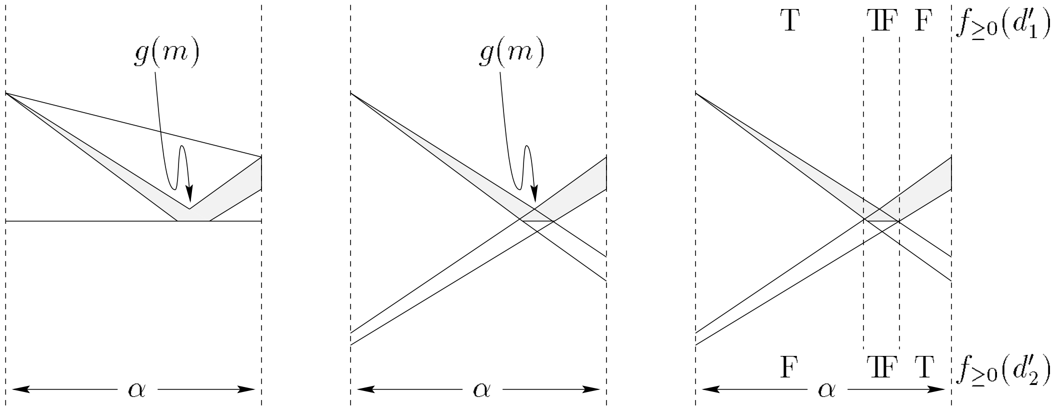
We will evaluate a $\mathbb{M}^{|F|_\star}$ model of absolute value: let $g(x) = |x|$. The evaluation of $g^{\mathbb{M}^{|F|_\star}}(m)$,

$$m = \{\langle\langle -4+5\alpha, -4+6\alpha\rangle|\mathrm{T}\rangle\},$$

proceeds as follows:

$$\begin{array}{ll}
 & g^{\mathbb{M}^{|F|_\star}}(\{\langle\langle m^-, m^+\rangle|\mathrm{T}\rangle\}) \\
\rightsquigarrow & g^{\mathbb{M}^{|F|}}(\langle\langle m^-, m^+\rangle|\mathrm{T}\rangle) \\
\rightsquigarrow & {g_1}^{\mathbb{M}^{|F|}}(\langle\langle m^-, m^+\rangle|\mathrm{T}\rangle) \quad \cup^{\mathbb{M}^{|F|}\to\mathbb{M}^{|F|_\star}} \quad {g_2}^{\mathbb{M}^{|F|}}(\langle\langle m^-, m^+\rangle|\mathrm{T}\rangle) \\
\rightsquigarrow & \langle\langle -m^+, -m^-\rangle|f_{\geq 0}(\langle -m^+, -m^-\rangle)\rangle \quad \cup^{\mathbb{M}^{|F|}\to\mathbb{M}^{|F|_\star}} \quad \langle\langle m^-, m^+\rangle|f_{\geq 0}(\langle m^-, m^+\rangle)\rangle \\
\rightsquigarrow & \{\langle\langle -m^+, -m^-\rangle|f_{\geq 0}(\langle -m^+, -m^-\rangle)\rangle, \langle\langle m^-, m^+\rangle|f_{\geq 0}(\langle m^-, m^+\rangle)\rangle\}.
\end{array}$$

Perusal of the following figures may cultivate the reader's appreciation of the preceding evaluation.



$g^{\mathbb{M}}(d)$ $g^{\mathbb{M}}(d)$ $g^{\mathbb{M}}(d)$ *with* $f_{>0}(v)$

3.3.22 Binary Functions: Two-Step Method

The approach taken for binary functions with constant interval arithmetic may be used with linear interval arithmetic. As before, we focus on binary grid functions. When presented with a binary function, we will cut it into sections where each section may be extended to a grid function which fits into a class.

As with unary functions, we will partition the domain based on monotonicity, so we may handle the upper and lower bounds independently. Concavity is also used when partitioning:

$$\Xi_2(g) = \{D : \psi_2^{\chi_1\chi_2}(g|D), D \subseteq \mathbb{R}^{\star 2}\};$$

$$\psi_2^{\chi_1\chi_2}(g) \;\text{ if }\; \exists[G \subseteq \mathbb{R}^{\star 3}]\ \forall[\alpha \in \mathbb{R}^\star]\ \ g \subseteq G \ \wedge\ \mathrm{grid}(G) \ \wedge\ \psi_2^{\chi_1}(G_{(x,\alpha)}) \ \wedge\ \psi_2^{\chi_2}(G_{(\alpha,y)}).$$

An upper bound will be determined for $g^{\mathbb{M}}(m)$ in two stages: first, a bilinear upper bound $h : [0,1]^2 \mapsto \mathbb{R}^\star$ of $g^{\mathbb{M}}(m)$ will be determined; then, a linear upper bound of h will be constructed; this upper bound will be an upper bound of g. A function $h : [0,1]^2 \mapsto \mathbb{R}^\star$ is bilinear if both $h_{(x,y=\alpha)}$ and $h_{(x=\alpha,y)}$ are linear.

An approximate upper bound $h^\star$ is constructed from g:

$$h^\star(x,y) = L_G;$$

where L_G is the bilinear Lagrange intepolating polynomial of the set G, a subgrid of g:

$$G = \{(x_i, y_j, g(x_i, y_j)) : (i,j) \in \{0,1\}^2\}.$$

Let $\mathrm{dom}(g) \subseteq X \times Y$. The location of the subgrid g is constrained by which concavity class g belongs to:

$$(x_0, x_1) = \begin{cases} (X^-, X^+) & \text{if } \psi_2^{\uparrow\updownarrow}, \\ (p, p+\Delta_p); \quad \{p, p+\Delta_p\} \in X & \text{if } \psi_2^{\downarrow\updownarrow}, \end{cases}$$

$$(y_0, y_1) = \begin{cases} (Y^-, Y^+) & \text{if } \psi_2^{\updownarrow\uparrow}, \\ (q, q+\Delta_q); \quad \{q, q+\Delta_q\} \in Y & \text{if } \psi_2^{\updownarrow\downarrow}. \end{cases}$$

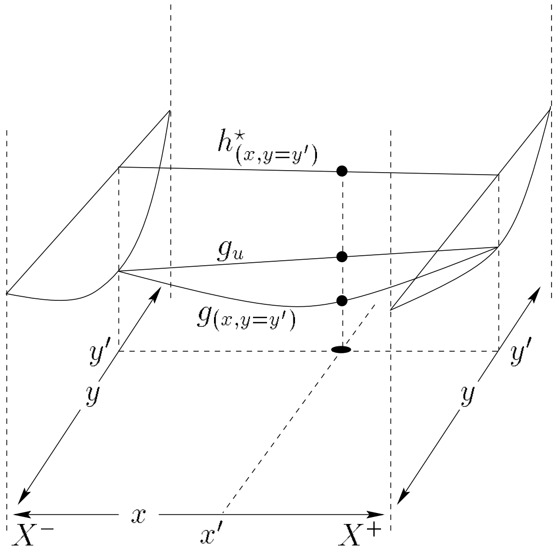
We assume that $\{(x_0, y_0), (x_1, y_1)\} \subseteq g$; if this is not the case we may extend g. Where g is concave up, we assume that G is finer than g:

$$\neg\exists[(x,y) \in \mathrm{dom}(g)]\ x \in (p, p+\Delta_p),$$

$$\neg\exists[(x,y) \in \mathrm{dom}(g)]\ y \in (q, q+\Delta_q).$$

For differentiable g, this corresponds to matching both the values and derivatives of g. When $\psi_2^{\downarrow\downarrow}(g)$, the mixed partial at (p,q) is matched.

When $\psi_2^{\uparrow\uparrow}(g)$, $h^\star$ is an upper bound. This is shown by the following proof by contradiction; the diagram given accompanies the proof.

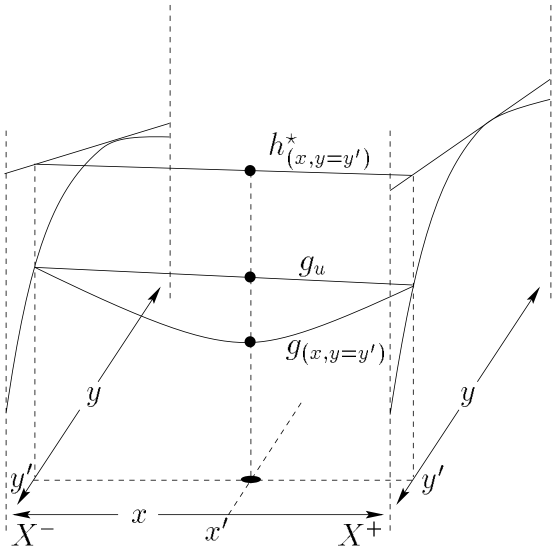


Assume that $h^\star$ fails, at (x',y'), to be an upper bound of g:

$$h^\star(x',y') < g(x',y').$$

Consider $g_{(x,y=y')}$, which has an upper bound g_u, given by L_G with $G = \{(X^-, y', g(X^-, y'), (X^+, y', g(X^+, y')\}$, since $\psi_1^{\downarrow}(g_{(x,y=y')})$. Since $h^\star_{(x=\alpha,y)}$ is an upper bound of $g_{(x=\alpha,y)}$ for $\alpha \in \{X^-, X^+\}$, $h^\star_{(x=\alpha,y)}(y')$ is an upper bound of $g_u(\alpha) = g_{(x=\alpha,y)}(y')$. It follows that $h^\star_{(x,y=\alpha)}$ is an upper bound of g_u, since both functions are linear; so $h^\star(x',y') \geq g(x',y')$, which contradicts our initial assumption.

When $\psi_2^{\uparrow\downarrow}(g)$, $h^\star$ is again an upper bound, and may be proven with a similar argument, which follows.



Assume that $h^\star$ fails, at (x', y'), to be an upper bound of g:

$$h^\star(x', y') < g(x', y').$$

Consider $g_{(x,y=y')}$, which has an upper bound g_u, given by L_G with $G = \{(X^-, y', g(X^-, y')), (X^+, y', g(X^+, y'))\}$, since $\psi_1^{\downarrow}(g_{(x,y=y')})$. Since $h^\star_{(x=\alpha,y)}$ is an upper bound of $g_{(x=\alpha,y)}$ for $\alpha \in \{X^-, X^+\}$, $h^\star_{(x=\alpha,y)}(y')$ is an upper bound of $g_u(\alpha) = g_{(x=\alpha,y)}(y')$. It follows that $h^\star_{(x,y=\alpha)}$ is an upper bound of g_u, since both functions are linear; so $h^\star(x', y') \geq g(x', y')$, which contradicts our initial assumption.

When $\psi_2^{\downarrow\uparrow}(g)$, $h^\star$ is again an upper bound, and may be proven with the preceding argument. Alternatively, one may consider $g'(x, y) = g(y, x)$, after ensuring that $h^\star_{g'}(x, y) = h^\star_g(y, x)$.

The proof does not hold for the last case, when $\psi_2^{\downarrow\downarrow}(g)$. In any of the other three cases, we may take $h = h^\star$; in this last case, we must further test g to determine an upper bound. We will not dwell on this since another method will be presented shortly.

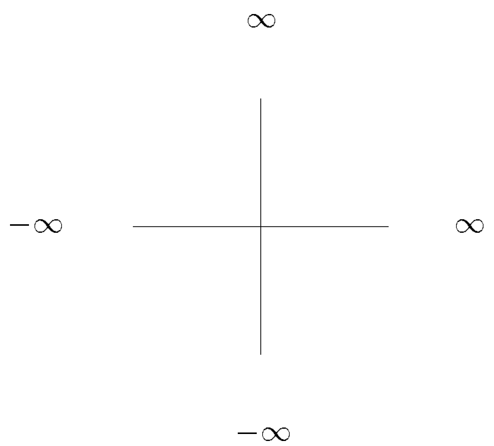
After h is determined, we construct a linear upper bound of $g(a+b\alpha, c+d\alpha)$ from $h(a+b\alpha, c+d\alpha)$. Given that

$$\begin{aligned} h(x, y) &= \psi + \psi_x x + \psi_y y + \psi_{xy} xy \\ \Rightarrow h(a + b\alpha, c + d\alpha) &= (\psi + \psi_x a + \psi_y b + \psi_{xy} bd) + (\psi_x b + \psi_y d + \psi_{xy}(ad + bc))\alpha + (\psi_{xy})\alpha^2, \end{aligned}$$

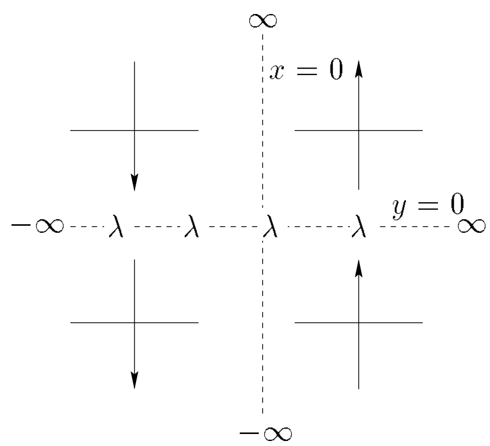
we may now treat h as a unary function of α. Previous subsections detail how an upper bound of a unary function may be found.

3.3.23 $\Xi_2^\star$ Charts

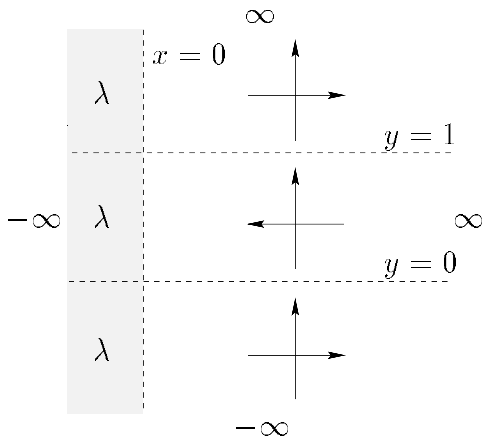
As was previously done, $\Xi_2^\star$ charts are used to graphically display the preferred sectioning of a function into concave pieces. Charts for some common binary functions follow.



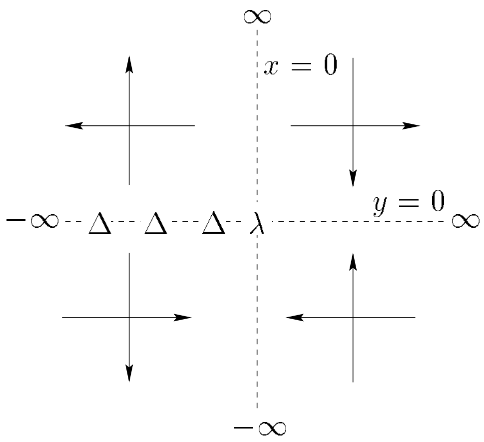
$\Xi_2^\star(x + y)$ *Chart*



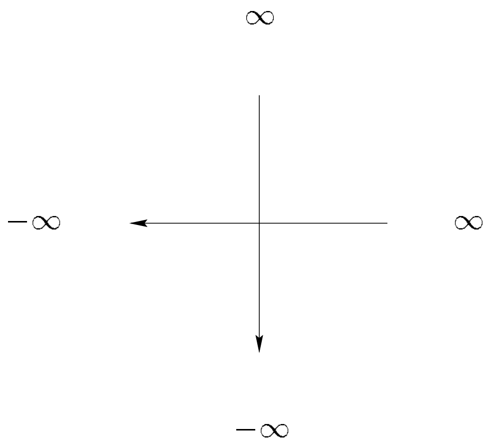
$\Xi_1^\star(x \div y)$ *Chart*



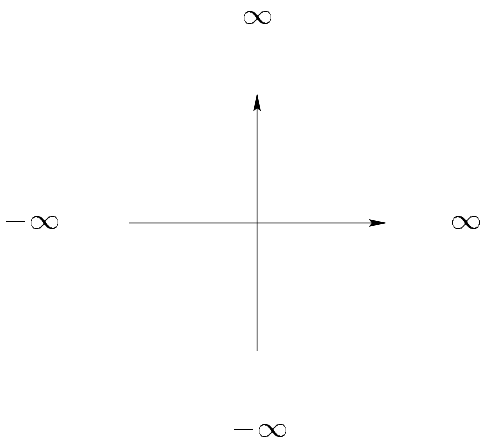
$\Xi_2^\star(x^y)$ *Chart*



$\Xi_2^\star(\theta(x, y))$ *Chart*



$\Xi_2^\star(\min(x, y))$ *Chart*



$\Xi_2^\star(\max(x, y))$ *Chart*

The $\Xi_2^\star(x - y)$ and $\Xi_2^\star(xy)$ charts are both identical to the $\Xi_2^\star(x + y)$ chart. The two-step method may handle charts which do not contain regions denoting that $\psi_2^{\uparrow\uparrow}(g|\xi_i)$ or $\psi_2^{\downarrow\downarrow}(g|\xi_i)$. The given charts reveal that all of the listed binary operators may be handled directly with the two-step method, with the exception of x^y. Since

$$\min(x, y) = -\max(-x, -y),$$

minimization and maximization may be handled, proceeding as follows:

$$\begin{aligned} \min^{\mathbb{M}}(x, y) &\rightsquigarrow \langle \min(x, y)^{\mathbb{M}-}, -(\max(-x, -y))^{\mathbb{M}+} \rangle, \\ \max^{\mathbb{M}}(x, y) &\rightsquigarrow \langle -(\min(-x, -y))^{\mathbb{M}-}, \max(x, y)^{\mathbb{M}+} \rangle. \end{aligned}$$

The above is only a formal justification of the obvious method,

$$\begin{aligned} \min^{\mathbb{M}}(x, y) &\rightsquigarrow \langle \min(x, y)^{\mathbb{M}-}, \min(x, y)^{\mathbb{M}+} \rangle, \\ \max^{\mathbb{M}}(x, y) &\rightsquigarrow \langle \max(x, y)^{\mathbb{M}-}, \max(x, y)^{\mathbb{M}+} \rangle. \end{aligned}$$

Addition, subtraction, and multiplication are particularly straightforward since each is a bilinear function. The bilinear bound is simply the function itself.

Since

$$\tfrac{d^2}{dx^2}g_{(x,y=\alpha)}(\xi) = \tfrac{\partial^2}{\partial x^2}g(\xi,\alpha), \quad \tfrac{d^2}{dx^2}g_{(x=\alpha,y)}(\xi) = \tfrac{\partial^2}{\partial y^2}g(\alpha,\xi),$$

the relationship, between $\frac{d^2}{dx^2}g$ and ψ_2^G, used to aid the determination of $\Xi_2^\star(g)$, for unary g, may be used to aid the determination of $\Xi_2^\star(g)$, for binary g. The partial derivatives for some common binary functions follow.

g	$\frac{\partial^2}{\partial x^2}g$	$\frac{\partial^2}{\partial y^2}g$
$x+y$	0	0
$x-y$	0	0
xy	0	0
$x \div y$	0	$2xy^{-3}$
x^y	$\frac{x^y}{x^2}(y^2-y)$	$x^y \ln^2(x)$
$\theta(x,y)$	$\frac{2xy}{(x^2+y^2)^2}$	$\frac{-2xy}{(x^2+y^2)^2}$

3.3.24 Examples with Binary Functions

Consider the subtraction function, $g(x,y) = x - y$,

$$\{\xi_1\} = \Xi_1^\star(g),\ \xi_1 = [-\infty,\infty]^2,\ g_1 = g^{\mathbb{R}^\star}|\xi_1;$$

The evaluation of $g^{\mathbb{M}}(m,n)$,

$$m = \langle -7+8\alpha, 3-\alpha\rangle, \quad n = \langle 3-2\alpha, 4-2\alpha\rangle,$$

proceeds as follows:

$$\begin{array}{llcccl}
 & g^{\mathbb{M}}(m,n) & & & & \\
 & {g_1}^{\mathbb{M}}(m,n) & & & & \\
\rightsquigarrow & \langle & g_1(m^-,n^+)^{\mathbb{M}-} & , & g_1(m^+,n^-)^{\mathbb{M}+} & \rangle, \text{ since } \psi_1^{\uparrow\downarrow}(g_1) \\
\rightsquigarrow & \langle & g_1(-7+8\alpha, 4-2\alpha)^{\mathbb{M}-} & , & g_1(3-\alpha, 3-2\alpha)^{\mathbb{M}+} & \rangle \\
\rightsquigarrow & \langle & (h_1(\alpha))^{\mathbb{M}-} & , & (h_2(\alpha))^{\mathbb{M}+} & \rangle \\
= & \langle & ((-7+8\alpha)-(4-2\alpha))^{\mathbb{M}-} & , & ((3-\alpha)-(3-2\alpha))^{\mathbb{M}+} & \rangle, \text{ since } g_1 \text{ is bilinear} \\
\rightsquigarrow & \langle & (10\alpha-11)^{\mathbb{M}-} & , & (-\alpha)^{\mathbb{M}+} & \rangle \\
\rightsquigarrow & \langle & -11+10\alpha & , & -\alpha & \rangle, \text{ since } \psi_1^0(h_1),\ \psi_1^0(h_2).
\end{array}$$

Consider the multiplication function, $g(x,y) = xy$,

$$\{\xi_{ij}\} = \Xi_1^\star(g),\ \xi_{ij} = R_i \times R_j,\ g_{ij} = g^{\mathbb{R}^\star}|\xi_{ij};$$

$$R_- = [-\infty, 0], R_+ = [0,\infty];\ (i,j) \in \{-,+\}^2.$$

The evaluation of $g^{\mathbb{M}}(m,n)$,

$$m = \langle 1+\alpha, 3-\alpha\rangle, \quad n = \langle 3-2\alpha, 4-2\alpha\rangle,$$

proceeds as follows:

$$\begin{array}{lcccl}
 & g^{\mathbb{M}}(m,n) & & & \\
 & {g_{++}}^{\mathbb{M}}(m,n) & & & \\
\rightsquigarrow & \langle \quad g_{++}(m^-,n^-)^{\mathbb{M}-} & , & g_{++}(m^+,n^+)^{\mathbb{M}+} & \rangle, \text{ since } \psi_1^{\uparrow\uparrow}(g_{++}) \\
\rightsquigarrow & \langle \quad g_{++}(1+\alpha,3-2\alpha)^{\mathbb{M}-} & , & g_{++}(3-\alpha,4-2\alpha)^{\mathbb{M}+} & \rangle \\
\rightsquigarrow & \langle \quad (h_1(\alpha))^{\mathbb{M}-} & , & (h_2(\alpha))^{\mathbb{M}+} & \rangle \\
= & \langle \quad ((1+\alpha)\times(3-2\alpha))^{\mathbb{M}-} & , & ((3-\alpha)\times(4-2\alpha))^{\mathbb{M}+} & \rangle, \text{ since } g_{++} \text{ is bilinear} \\
\rightsquigarrow & \langle \quad (-2\alpha^2+\alpha+3)^{\mathbb{M}-} & , & (2\alpha^2-10\alpha+12)^{\mathbb{M}+} & \rangle \\
\rightsquigarrow & \langle \quad 3-\alpha & , & 12-8\alpha & \rangle, \text{ since } \psi_1^{\downarrow}(h_1),\ \psi_1^{\uparrow}(h_2).
\end{array}$$

3.3.25 Binary Functions: One-Step Method

The framework presented results in the two-step method, which does not directly apply to the general exponentiation function x^y. With appropriate extensions, the two-step method may generate valid bounds for x^y but it will still generate bounds which are not optimal.

With the two-step method an upper bound h, of g, was found; h was then treated as a unary function so that our previous methods may be applied. We will remove the intermediate step, and consider g as a unary function of α:

$$g(\alpha) = g(p(\alpha), q(\alpha)).$$

The previous methods may now be applied.

The relevant derivatives appear in the table following. Positive multiplicative factors were removed from some table entries. Throughout the table $p = a + b\alpha$ while $q = c + d\alpha$.

g	$\frac{d}{d\alpha}g(p,q)$	$\frac{d^2}{d\alpha^2}g(p,q)$
$x+y$	$b+d$	0
$x-y$	$b-d$	0
xy	$bq+dp$	bd
$x\div y$	$bc-ad$	$d\frac{ad-bc}{q}$
x^y	$\frac{dp\ln p+bq}{p}$	$[dp\ln p][2bq+dp\ln p]$ $+b[bq(q-1)+2dp]$
$\theta(x,y)$	$dp-bq$	$-(bp+dq)(ad-bc)$

The one-step and two-step methods produce identical algorithms for addition, subtraction, multiplication, minimization and maximization. The sign of $\frac{d^2}{d\alpha^2}g$ is independent of the sign of bd, as expected. The sign of $\frac{d}{d\alpha}g$ reverses as the sign of bd reverses, as expected. With either the one-step or two-step method, efficiency may be improved by considering the $\Xi_1^\star(g)$ regions $m\times n$ overlap, as was done in section 3.2.25.

3.3.26 Examples with a Binary Function

Consider the division function, $g(x, y) = x \div y$,

$$\{\xi_{ij}\} = \Xi_1^\star(g),\ \xi_{ij} = R_i \times R_j,\ g_{ij} = g^{\mathbb{R}^\star}|\xi_{ij};$$

$$R_- = [-\infty, 0], R_+ = [0, \infty];\ (i, j) \in \{-, +\}^2.$$

The evaluation of $g^{\mathbb{M}}(m, n)$,

$$m = \langle \alpha, 5\alpha \rangle,\quad n = \langle 2\alpha, 1 + 3\alpha \rangle,$$

proceeds as follows:

$$\begin{array}{llcccl}
 & g^{\mathbb{M}}(m, n) & & & & \\
 & {g_{++}}^{\mathbb{M}}(m, n) & & & & \\
\rightsquigarrow & \langle & g_{++}(m^-, n^+)^{\mathbb{M}-} & , & g_{++}(m^+, n^-)^{\mathbb{M}+} & \rangle, \text{ since } \psi_1^{\uparrow\downarrow}(g_{++}) \\
\rightsquigarrow & \langle & g_{++}(\alpha, 1 + 3\alpha)^{\mathbb{M}-} & , & g_{++}(5\alpha, 2\alpha)^{\mathbb{M}+} & \rangle \\
\rightsquigarrow & \langle & (h_1(\alpha))^{\mathbb{M}-} & , & (h_2(\alpha))^{\mathbb{M}+} & \rangle \\
= & \langle & ((\alpha) \div (1 + 3\alpha))^{\mathbb{M}-} & , & ((5\alpha) \div (2\alpha))^{\mathbb{M}+} & \rangle \\
\rightsquigarrow & \langle & \frac{1}{4}\alpha & , & 2\frac{1}{2} & \rangle, \text{ since } \psi_1^{\downarrow}(h_1),\ \psi_1^{\uparrow}(h_2).
\end{array}$$

The sign of h_1 and h_2 are determined by the arguments and the relationship with the second derivative of g_{++}:

$$\frac{d^2}{d\alpha^2}h_1 = \frac{d^2}{d\alpha^2}g_{++}(m^-, n^+) = 3\frac{0 \times 3 - 1 \times 1}{q} \leq 0,\ q \in (0, 1],$$

$$\frac{d^2}{d\alpha^2}h_2 = \frac{d^2}{d\alpha^2}g_{++}(m^+, n^-) = 2\frac{0 \times 2 - 5 \times 0}{q} = 0,\ q \in (0, 1].$$

The corresponding evaluation with constant intervals produces a noticeably larger result:

$$\begin{array}{rcl}
g^{\mathbb{J}}(\langle 0, 5 \rangle, \langle 0, 4 \rangle) & \rightsquigarrow & \langle 0, \infty \rangle, \\
(g^{\mathbb{M}}(\langle \alpha, 5\alpha \rangle, \langle 2\alpha, 1 + 3\alpha \rangle))^{\mathbb{J}} \rightsquigarrow \langle \frac{1}{4}\alpha, 2\frac{1}{2} \rangle^{\mathbb{J}} & \rightsquigarrow & \langle \frac{1}{4}, 2\frac{1}{2} \rangle.
\end{array}$$

Similar behaviour may be observed near the origin with the general exponentiation function, x^y.

3.3.27 Partial Binary Functions

The methods for handling partial, discontinuous, and bumpy functions are extended in the obvious way. An example will suffice.

3.3.28 Examples with a Binary Partial Function

Consider the division function, $g(x, y) = x \div y$, which is both partial and $\mathbb{M}$-bumpy. For the division function,

$$\{\xi_{ij}\} = \Xi_1^\star(g),\ \xi_{ij} = R_i \times R_j,\ g_{ij} = g^{\mathbb{R}^\star}|\xi_{ij};$$

$$R_- = [-\infty, 0], R_+ = [0, \infty];\ (i, j) \in \{-, +\}^2.$$

The evaluation of $g^{\mathbb{J}^{|\mathbb{T}\star}}(\{\langle m|\mathrm{T}\rangle\}, \{\langle n|\mathrm{T}\rangle\})$,

$$\langle m|\mathrm{T}\rangle = \langle\langle 1 + 2\alpha, 1 + 3\alpha\rangle|\mathrm{T}\rangle,\ \langle n|\mathrm{T}\rangle = \langle\langle -1 + 3\alpha, -1 + 4\alpha\rangle|\mathrm{T}\rangle,$$

proceeds as follows:

$$
\begin{array}{lllll}
 & g^{\mathbb{M}|F^{|}\star}(\{\langle m|\mathrm{T}\rangle\},\{\langle n|\mathrm{T}\rangle\}) & & & \\
\rightsquigarrow & \quad g^{\mathbb{M}|F^{|}}(\langle m|\mathrm{T}\rangle,\langle n|\mathrm{T}\rangle) & & & \\
\rightsquigarrow & \quad {g_{+-}}^{\mathbb{M}|F^{|}}(\langle m|\mathrm{T}\rangle,\langle n|\mathrm{T}\rangle) & \cup^{\mathbb{M}|F^{|}\rightarrow\mathbb{M}|F^{|}\star} & & {g_{++}}^{\mathbb{M}|F^{|}}(\langle m|\mathrm{T}\rangle,\langle n|\mathrm{T}\rangle) \\
\rightsquigarrow & \langle\langle-\infty,1-24\alpha\rangle|f_{>0}(-n)\rangle & \cup^{\mathbb{M}|F^{|}\rightarrow\mathbb{M}|F^{|}\star} & & \langle\langle 3\frac{31}{36}-3\frac{1}{9}\alpha,\infty\rangle|f_{>0}(n)\rangle \\
\rightsquigarrow & \multicolumn{4}{l}{\{\langle\langle-\infty,1-24\alpha\rangle|f_{>0}(\langle 1-4\alpha,1-3\alpha\rangle)\rangle,\langle\langle 3\frac{31}{36}-3\frac{1}{9}\alpha,\infty\rangle|f_{>0}(\langle -1+3\alpha,-1+4\alpha\rangle)\rangle\}.}
\end{array}
$$

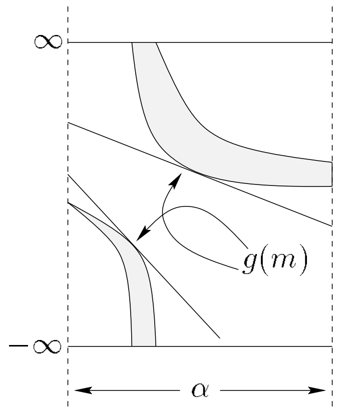
The division function is defined unless the divisor is zero:

$$\Xi_{|}(g)=[-\infty,\infty]\times([-\infty,0)\cup(0,\infty]),\ \boldsymbol{\xi}_{|}=\Xi_{|}(g).$$

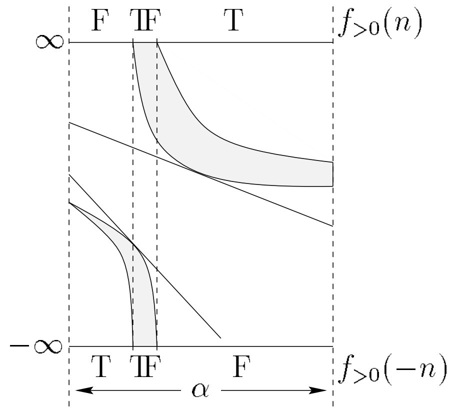
In the evaluation above, ${g_{++}}^{\mathbb{M}|F^{|}}(\langle m|\mathrm{T}\rangle,\langle n|\mathrm{T}\rangle)$ is evaluated:

$$
\begin{array}{llcll}
 & \multicolumn{4}{l}{{g_{++}}^{\mathbb{M}|F^{|}}(\langle m|\mathrm{T}\rangle,\langle n|\mathrm{T}\rangle)} \\
\rightsquigarrow & \langle\quad {g_{++}}^{\mathbb{M}|F^{|}}(m,n) & | & \mathrm{T}\wedge\mathrm{T}\wedge f_{>0}(\Phi_{\zeta_{>0}}(\langle m|\mathrm{T}\rangle,\langle n|\mathrm{T}\rangle,\boldsymbol{\xi}_{|})) & \rangle,\ \text{since } \mathrm{Z}_{F^{|}}(m,n,\boldsymbol{\xi}_{|})=f_{>0} \\
\rightsquigarrow & \langle\quad {g_{++}}^{\mathbb{M}|F^{|}}(m,n) & | & \mathrm{T}\wedge\mathrm{T}\wedge f_{>0}(n) & \rangle,\ \text{since } \Phi_{\zeta_{>0}}(m,n,\boldsymbol{\xi}_{|})=n \\
\rightsquigarrow & \langle\quad \langle 3\frac{31}{36}-3\frac{1}{9}\alpha,\infty\rangle & | & f_{>0}(\langle -1+3\alpha,-1+4\alpha\rangle) & \rangle.
\end{array}
$$

For general binary functions, it may be natural to introduce more than a single new constraint after a function application. Enhancing Z and Φ allows this; the common binary functions do not naturally introduce multiple constraints. The members of $F^{|}$ influence this decision.



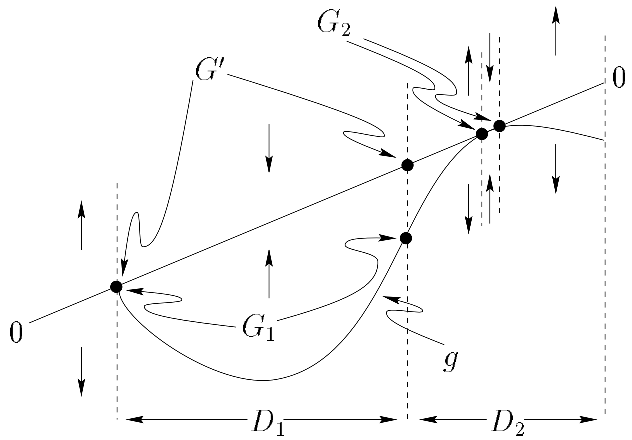
$g^{\mathbb{M}}(m,n)$



$g^{\mathbb{M}}(m,n)$ *with* $f_{>0}(-n)$, $f_{>0}(n)$

3.3.29 Concave Up, Down Functions

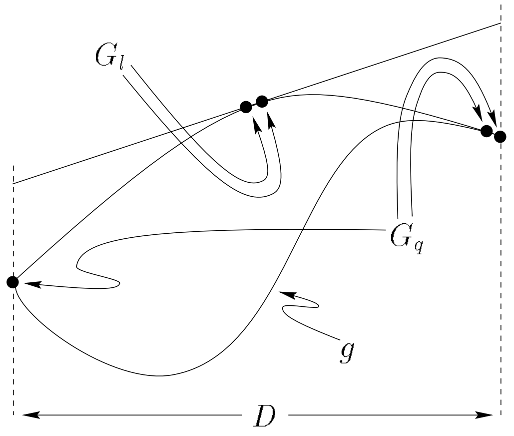
There is another way to approach evaluating $g^{\mathbb{M}}$, when g is neither concave up nor concave down. We consider g such that $\Xi_2^\star(g)=\{\xi_1,\xi_2\}$. Restricting our attention to continuous g, we may find an upper bound without splitting g into two parts. Consider the following chart, a ψ_2 chart where $\psi_2^{\uparrow}(g|\xi_1)$ and $\psi_2^{\downarrow}(g|\xi_2)$.



ψ_2 *Chart*

First, bounds of $g|\xi_1$ and $g|\xi_2$ are found; $(g|\xi_1)^{\mathbb{M}+}(j)^+ = L_{G_1}$, $(g|\xi_2)^{\mathbb{M}+}(j)^+ = L_{G_2}$. Let G' be formed by connecting the left endpoint of L_{G_1} with the left endpoint of L_{G_2}. If this line extends so that it overlaps L_{G_2}, then this may be taken as an upper bound of both sections. A lower bound would be found for the above example with the procedures outlined earier.

Another approach is to find a quadratic upper bound and then produce a linear upper bound of the quadratic upper bound. This may be done if $\psi_3^{\ddagger}$.



ψ_2, ψ_3 *Chart*

3.3.30 Floating Point

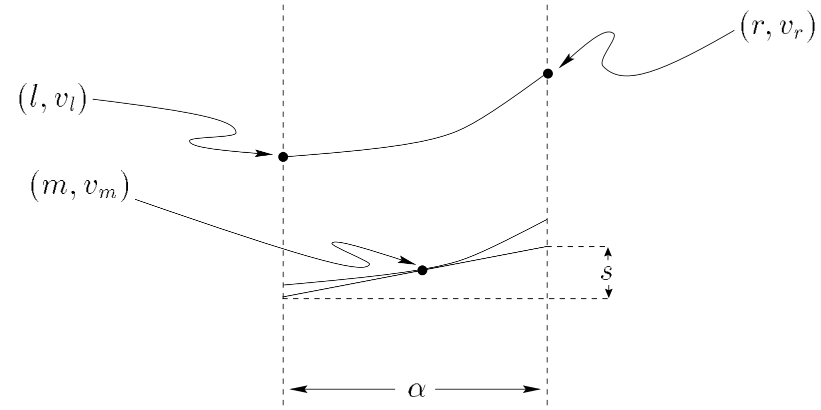
We will now demonstrate an evaluation of $g^{\mathbb{L}}$, for $g(x) = e^x$. The evaluation of $g^{\mathbb{M}}(\langle a + b\alpha, c + d\alpha\rangle$ proceeds as follows:

$$\begin{aligned} & g^{\mathbb{M}}(\langle a + b\alpha, c + d\alpha\rangle \\ \rightsquigarrow \quad & \langle (v_m - \tfrac{1}{2}s) + s\alpha, v_l + (v_r - v_l)\alpha\rangle, \end{aligned}$$

with

$$\begin{gathered} l = a, \qquad m = a + \tfrac{1}{2}b, \qquad r = a + b, \\ v_l = g^{\mathbb{R}^\star}(l), \quad v_m = g^{\mathbb{R}^\star}(m), \quad v_r = g^{\mathbb{R}^\star}(r), \\ s = bv_m. \end{gathered}$$

The following diagram illustrates the variables used.



The same procedure may be used when evaluating $g^{\mathbb{L}}$, after specifying which rounding mode is used for each operation. The evaluation of $g^{\mathbb{L}}(\langle a+b\alpha, c+d\alpha\rangle$ proceeds as follows:

$$\begin{array}{ll} & g^{\mathbb{M}}(\langle a+b\alpha, c+d\alpha\rangle \\ \rightsquigarrow & \langle (v_m -^{\mathbb{F}-} \tfrac{1}{2} \times^{\mathbb{F}+} s) + s\alpha, v_l + (v_r -^{\mathbb{F}+} v_l)\alpha\rangle, \end{array}$$

with

$$\begin{array}{ccc} l = a, & m = a +^{\mathbb{F}=} \tfrac{1}{2} \times^{\mathbb{F}=} b, & r = a +^{\mathbb{F}+} b, \\ v_l = g^{\mathbb{F}+}(l), & v_m = g^{\mathbb{F}=}(m), & v_r = g^{\mathbb{F}+}(r), \\ & s = b \times^{\mathbb{F}=} v_m, & \end{array}$$

assuming that $b \geq 0$. There is some freedom in the assignment of rounding modes; the above is intended as a guide to producing a valid model, not necessarily an optimal one. Even with a fixed assignment of rounding modes, the choice of rounding modes will influence the optimality of the model, as well as the execution cost.

3.4 Polynomial Interval Arithmetic

Many of the concepts introduced so far may be used to guide the implementation of more sophisticated interval arithmetics. In this section, the generalizations necessary for polynomial interval arithmetics are discussed.

3.4.1 Interpolating Polynomials

Lagrange interpolating polynomials are defined for arbitrary $G \subseteq_k g$ in many numerical texts. See, for example, [11].

3.4.2 ψ_k Charts

We now prove that the rules given in section 3.2.3, for constructing a ψ_k chart, are correct.

The forbidden region is clearly correct since L_G is a function; we simply decree that $(x, y) \notin G_k$ as our use of the ψ_k chart does not depend on how such points are treated. For (x, y) in the zero region, $L_G = L_{G_k}$ and $\deg(L_{G_k}) < \deg(L_G)$; so for any point in the zero region $\psi_k^G = 0$, which implies $\psi_k^0(G)$.

For the remaining regions, consider the polynomial

$$p(x) = \prod_{i=0}^{k-1}(x - x_i), \; G_k = \{(x_0, y_0), (x_1, y_1), \ldots, (x_{k-1}, y_{k-1})\}.$$

From the construction of $p(x)$ it is clear that

$$\forall[(x_i, y_i) \in G] \; mp(x_i) = 0,$$

for any $m \in \mathbb{R}$. Consider the polynomial

$$q(x) = L_{G_k}(x) + mp(x),$$

which interpolates G_k for any value of m. The k roots of the k degree polynomial p are $x_0, x_1, \ldots, x_{k-1}$. The polynomial p has no other roots since it is not identically zero. For large x, $p(x)$ is positive:

$$x > \max_{0 \le i < k} x_i \;\Rightarrow\; p(x) > 0.$$

Imagine $p(x)$ as x decreases; the sign of $p(x)$ will reverse each time x crosses a root of $p(x)$. This sign changing corresponds with the checkboard labelling of ψ_k.

Consider the point (x, y) which is y_Δ away from L_{G_k}:

$$y_\Delta = y - L_{G_k}(x).$$

If

$$m = \frac{y_\Delta}{p(x)},$$

then

$$\begin{aligned} q(x) &= L_{G_k}(x) + mp(x) \\ &= L_{G_k}(x) + \frac{y_\Delta}{p(x)}p(x) \\ &= L_{G_k}(x) + y_\Delta \\ &= L_{G_k}(x) + y - L_{G_k}(x) \\ &= y. \end{aligned}$$

Earlier we proved $q(x)$ interpolated G_k for any m. We have now shown $q(x)$ interpolates $G = G_k \cup \{(x, y)\}$. The leading coefficient of $q(x)$ is m. The sign of m relates to $y - L_{G_k}(x)$: the sign of m is positive if the region (x, y) resides in is labelled with $\uparrow$; the sign of m is negative if the region (x, y) resides in is labelled with $\downarrow$.

3.4.3 Optimality

Determining polynomial upper and lower bounds of general functions has been discussed in the literature [45, 14, 18, 72]. In the results cited, optimality is determined via the $\mathcal{L}_1$ norm, as done in section 3.3.3.

In [14], it is shown that if $g : \mathbb{R} \mapsto \mathbb{R}$ is bounded, and finite for at least $n + 1$ points, then there exists optimal lower and upper degree n bounds. It is also shown that if g is continuous on $[0, 1]$, and differentiable on $(0, 1)$, then the optimal bounds are unique. It is also established that for g with $\frac{d^n}{dx^n}g \geq 0$ or $\frac{d^n}{dx^n}g \leq 0$, the optimal bounds are found by interpolating g and $\frac{d}{dx}g$, as we have done. The optimal interpolation points are shown to be the nodes of a Gauss quadrature formula. With linear bounds, this corresponds to interpolating the value of g for $\alpha = 0$ and $\alpha = 1$, or interpolating the value and derivative of g for $\alpha = \frac{1}{2}$.

In [45], a collection of constrained approximation problems is brought together, with one-sided approximation treated as a special case of general constrained approximation problems. Linear programming is suggested as a method to determine bounds when the nth derivative crosses zero: [44] is cited. See [1, 43] for more recent work. Much of the current discussion is of spline aproximations, as in [43]. In all of the papers referenced, a detailed computational procedure must be followed to determine an approximate lower or upper bound.

In [72], another approach to proving upper and lower polynomial bounds optimal is taken. With this approach, it is shown that interpolating the value and/or derivative at the nodes of a Gauss quadrature formula constructs the optimal polynomial bound, provided that the bound does not interpolate the function elsewhere.

Most of these results generalize to non-polynomial bounds. Often, the bounds are taken from a Chebyshev system [18, 45, 1, 44]; the set of n degree polynomials form a Chebyshev system. In [72], bounds are taken as linear combinations of an arbitrary set of continuous functions. Characterizations of constrained approximation solutions has also been studied [30, 70].

3.4.4 Piecewise Models

As before, we may construct general polynomial interval operators by considering subdomains of the function to be implemented. As discussed earlier, optimal bounds are known when the $n+1$st derivative is bounded away from zero. When an interval straddles a zero of the $n+1$st derivative, the $n+k$th derivative will often be bounded away from zero. In such a case, an $n+k-1$ degree polynomial may be used to bound the function, which may then be demoted to find an n degree polynomial bound.

3.4.5 $\Xi_k^\star$ Charts

$\Xi_k^\star$ charts may be constructed for general k. The utility of such charts is somewhat limited, however. For constant bounds,

$$\frac{d}{dx}[g(x)] = \left[\frac{d}{dx}g\right](x),$$

so knowledge of $[\frac{d}{dx}g](x)$ is quite useful in bounding the derivative of $g(x)$; for linear bounds,

$$\frac{d^2}{dx^2}[g(ax+b)] = a^2\left[\frac{d^2}{dx^2}g\right](ax+b),$$

so knowledge of $[\frac{d^2}{dx^2}g](ax+b)$ is again quite useful in bounding the second derivative of $g(ax+b)$; for quadratic bounds,

$$\frac{d^3}{dx^3}[g(ax^2+bx+c)] = (2ax+b)^3\left[\frac{d^3}{dx^3}g\right](ax^2+bx+c) + a(2ax+b)\left[\frac{d^2}{dx^2}g\right](ax^2+bx+c),$$

so knowledge of $[\frac{d^3}{dx^3}g](ax^2+bx+c)$ does not allow us to bound the third derivative of $g(ax^2+bx+c)$. Additional knowledge of $[\frac{d^2}{dx^2}g](ax^2+bx+c)$ allows us to bound the third derivative of $g(ax^2+bx+c)$ in some cases. As the derivative analysis is not as simple as before, such a simple design aid is no longer sufficient. This does not preclude the construction of quadratic interval routines; it just shows that such construction cannot be guided by $\Xi_k^\star$ charts alone. A similar situation was encountered while constructing linear interval operators for binary functions.

Chapter 4

Graphs

We are now ready to approach our motivating problem. This chapter is about graphs, and graphing. A graph is a mathematical object; we must first take this ideal object and form a corresponding object that can be mechanically produced. We do not dwell on the issue; our discourse remains lofty, as we will continue to exploit our abstract models. The abstract models we exploit are, however, grounded in concepts that are realizable. Practical graphing concerns are addressed in [7].

4.1 Graphs

The graph of an equation may formally be defined as the set of points which satisfy that equation. Given a specification $S : \mathbb{R}^2 \mapsto \mathbb{B}$ the graph $G \subseteq \mathbb{R}^2$ of S is defined as follows:

$$G[S] \ \equiv_{\text{def}} \ \{\boldsymbol{x} : S(\boldsymbol{x})\}.$$

For example,

$$G[y = x^2] = \{(x, y) : y = x^2\}.$$

The formalism alone does not fully communicate the true nature of graphs: the visual nature of geometry must be reconciled with the dry, algebraic character of the formalism presented to fully capture the true nature of graphs. Mathematicians often refer to graphs as relations [65]. This is clear, as the formal definition of a graph is identical in graph theory [41] and geometry, but the two concepts are commonly thought of as distinct.

4.1.1 Rendering

As a graph is a visual object, we now preoccupy ourselves with rendering graphs. We will not attempt to fully render any graph; the influence we may exert on the world does not permit us to render either the intricate details or the vast stretches present in most graphs. We will content ourselves, for now, with discretely rendering small portions of graphs.

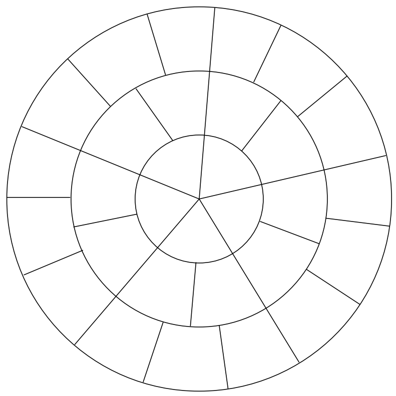
There is a wide variety of physical rendering devices which may be controlled by modern computers. Examples include: monitors, projection units, laser printers, ink-jet printers, dot-matrix printers, thermal printers, and plotters. We phrase our discussion in terms of an abstract rendering device, which approximates actual rendering devices. Our abstract rendering device outputs a perfectly rendered image R, which is a collection of pixels.

Usually, the pixels are rectangular in shape and form a $u \times v$ grid. Each pixel $\boldsymbol{p} \in R$ represents a region of the plane. The pixel $\boldsymbol{p}$ represents the region $M(\boldsymbol{p})$. If M is given by

$$M(\boldsymbol{p}) = \left[\frac{\boldsymbol{p}_u}{u}, \frac{\boldsymbol{p}_u + 1}{u}\right] \times \left[\frac{\boldsymbol{p}_v}{v}, \frac{\boldsymbol{p}_v + 1}{v}\right],$$

then the portion of the graph lying within the unit square is rendered, where $(\boldsymbol{p}_u, \boldsymbol{p}_v)$ is the coordinates of pixel $\boldsymbol{p}$. Other portions of the graph may be rendered by appropriately modifying M.

There are many other possibilities. When rendering with polar coordinates, it may be convenient to have pixels which are wedge-shaped and form concentric rings about the origin.



Wedge-Shaped Pixels

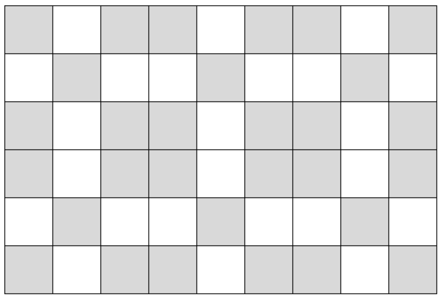
This type of pixel allows for a more natural mapping M, which will in turn enable better interval renderings. When rendering polar graphs, rectangular pixels may still be used, but with a more cumbersome mapping M.

Another possibility is for each pixel to be in the shape of a small cube. As with rectangular pixels, a simple axis-aligned affine mapping may be used, but would now allow portions of space to be presented. Three-dimensional rendering devices would allow the rendering to be presented directly; two-dimensional rendering devices may present projections and slices of the rendering.

In the remainder of this chapter, our illustrations will be of two-dimensional renderings using a grid of rectangular pixels. The accompanying descriptions are general, and may be applied to the other cases mentioned above. In higher (and lower) dimensions, references to $\mathbb{Y}^2$ and $\mathbb{R}^2$ may be replaced with references to $\mathbb{Y}^n$ and $\mathbb{R}^n$, respectively.

4.1.2 Batch Rendering

We assume each pixel can take on one of two colours, and that all 2^{uv} patterns may be realized.



9×6 *Grid of Pixels*

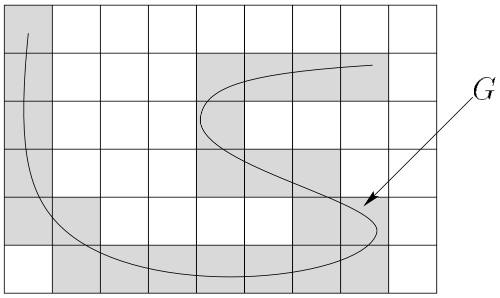
The rendering R represents the graph $G[S]$ if

$$\forall[\boldsymbol{p} \in R] \qquad \left(\sup_{\boldsymbol{x} \in M(\boldsymbol{p})} S(\boldsymbol{x})\right) \;=\; R(p),$$

where

$$R(\boldsymbol{p}) = \begin{cases} \mathrm{F} & \text{if } \boldsymbol{p} \text{ is white,} \\ \mathrm{T} & \text{if } \boldsymbol{p} \text{ is grey.} \end{cases}$$

For any given graph G, there is a unique rendering R which represents it.



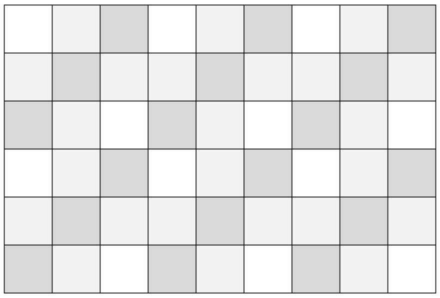
9×6 *Batch Rendering of G*

The information contained in the colour of $\boldsymbol{p} \in R$ is summarized in the following table:

$\boldsymbol{p}$	Information
White	$\forall[\boldsymbol{x} \in M(\boldsymbol{p})]\neg S(\boldsymbol{x})$
Gray	$\exists[\boldsymbol{x} \in M(\boldsymbol{p})] S(\boldsymbol{x})$

4.1.3 Progressive Rendering

We now assume that each pixel can take on one of three colours, and that all 3^{uv} patterns may be realized.



9×6 *Grid of Pixels*

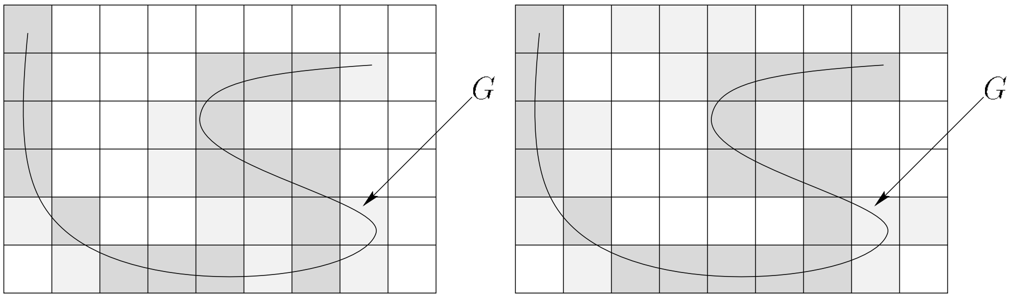
The rendering R represents the graph $G[S]$ if

$$\forall[\boldsymbol{p} \in R] \qquad \left(\sup_{\boldsymbol{x} \in M(\boldsymbol{p})} S(\boldsymbol{x})\right) \;\sqsubseteq\; R(p),$$

where

$$R(\boldsymbol{p}) = \begin{cases} \mathrm{F} & \text{if } \boldsymbol{p} \text{ is white,} \\ \mathbb{TF} & \text{if } \boldsymbol{p} \text{ is light grey,} \\ \mathrm{T} & \text{if } \boldsymbol{p} \text{ is dark grey.} \end{cases}$$

For any given graph G, there are many renderings which represent it.



9×6 *Progressive Renderings of* G

The information contained in the colour of $\boldsymbol{p} \in R$ is summarized in the following table:

$\boldsymbol{p}$	Information
White	$\forall[\boldsymbol{x} \in M(\boldsymbol{p})]\neg S(\boldsymbol{x})$
Light Gray	no information
Dark Gray	$\exists[\boldsymbol{x} \in M(\boldsymbol{p})]S(\boldsymbol{x})$

4.1.4 Syntax

A generous syntax is used to descibe relations to be graphed. This generosity does not burden graphing, as the generous syntax may be built up using another, more basic syntax. Here, we will show how some more luxurious elements may be built from a basic syntax.

General comparisons, namely $x < y, x \leq y, x = y, x \geq y, x > y$, may be emulated by exploiting the following identities:

$$\begin{aligned} G[g \odot h] &= G[g - h \odot 0], \\ G[g \leq 0] &= G[-g \geq 0], \\ G[g < 0] &= G[-g > 0], \\ G[g \geq 0] &= G[g = |g|], \\ G[g > 0] &= G[\mathrm{signum}(g) = 1]. \end{aligned}$$

Both g and h are arbitrary functions of x and y.

Conjunctions and disjunctions may be emulated by exploiting the following identities:

$$\begin{aligned} G[g = 0 \ \vee \ h = 0] &= G[gh = 0], \\ G[g = 0 \ \wedge \ h = 0] &= G[|g| + |h| = 0]. \end{aligned}$$

Equations which contain partial functions may be modified so that the equation itself is total. The following identities may be used for $\sqrt{x}$ and x^{-1}:

$$\begin{aligned} G[g(\sqrt{h}) = 0] &= G[g(\sqrt{|h|}) = 0 \ \wedge \ h \geq 0], \\ G[g(h^{-1}) = 0] &= G[g((h + |\mathrm{signum}(h)| - 1)^{-1}) = 0 \ \wedge \ |\mathrm{signum}(h)| = 1]; \end{aligned}$$

similar identities may be used for other partial functions.

Logical negation may now be emulated by exploiting the following identity:

$$G[g \neq 0] \quad = \quad G[|\mathrm{signum}(g)| = 1],$$

where g is a total function.

Some of the more exotic forms of syntax may be emulated by exploiting the following identities:

$$\begin{aligned} G[g \in \mathbb{Z}] &= G[g = \lfloor g \rfloor], \\ G[g \in [h, i)] &= G[h \leq g < i], \\ G[g \leq h < i] &= G[g \leq h \ \wedge \ h < i], \\ G[[g, h] \subseteq [i, j]] &= G[i \leq g \leq h \leq j]. \end{aligned}$$

Note that different emulation strategies may produce differing bounds when evaluated with an interval arithmetic. For example, exploiting the identity

$$G[g \leq h < i] \quad = \quad G[g \leq h \ \wedge \ h < i \ \wedge \ g < i],$$

rather than the identity given earlier, allows one to produce formulas which result in sharper bounds when evaluated using an interval arithmetic.

Although $\mathrm{signum}(x)$ is used above, it may be emulated by exploiting the following identity:

$$G[g(\mathrm{signum}(h)) = 0] \quad = \quad G[g(\lfloor 2\pi^{-1}\mathrm{Arctan} h \rfloor + \lceil 2\pi^{-1}\mathrm{Arctan} h \rceil) = 0].$$

Both $\lfloor x \rfloor$ and $\lceil x \rceil$ may be emulated as well. At some point, a set of basic operators must be defined. This emulation scheme may be foiled by providing a restrictive set of basic operators. Since it is not difficult to implement the more luxurious operators, we choose to provide them directly.

4.1.5 Notation

Given a specification $S : \mathbb{R}^2 \mapsto \mathbb{B}$, the graph $G \subseteq \mathbb{R}^2$ of S is defined as before:

$$G[S] = \{\boldsymbol{x} : S(\boldsymbol{x})\}.$$

A rendering R represents $G[S]$ if

$$\forall [\boldsymbol{p} \in R] \qquad \left(\sup_{\boldsymbol{x} \in M(\boldsymbol{p})} S(\boldsymbol{x}) \right) \ \sqsubseteq \ R(\boldsymbol{p}).$$

We say that R is a rendering of S if R represents $G[S]$; $R[S]$ denotes a rendering of S.

We now state some examples to clarify the definitions. Let R consist of a single pixel $\boldsymbol{p}$.

$M(\boldsymbol{p})$	$R[\sqrt{x+y} > \sqrt{2}](\boldsymbol{p})$
$[-2, -1]^2$	$\mathbb{TF}$ or F
$[-1, 1]^2$	$\mathbb{TF}$ or F
$[-1, 2]^2$	$\mathbb{TF}$ or T

4.2 Basic Rendering

We now turn our attention to rendering a graph. Our attention is focussed on a basic algorithm so that we may explore the obvious ramifications of rendering with interval arithmetic.

4.2.1 Constant Interval Arithmetic

Let $\mathbb{Y}$ denote a constant interval arithmetic, such as $\mathbb{J}$ or $\mathbb{J}^{|\mathbb{T}}$. We will not ensure that R represents $S : \mathbb{R}^2 \mapsto \mathbb{B}$ directly; we will instead work with $S^{\mathbb{Y}} : \mathbb{Y}^2 \mapsto \mathbb{T}$. The interval specification $S^{\mathbb{Y}}$ is computed by evaluating the specification S using the interval arithmetic $\mathbb{Y}$. The interval inclusion property assures us that

$$\forall[\boldsymbol{x} \in \boldsymbol{j}]\ S(\boldsymbol{x}) \subseteq S^{\mathbb{Y}}(\boldsymbol{j}).$$

Let $M^{\mathbb{Y}}(\boldsymbol{p})$ describe $M(\boldsymbol{p})$, using an element of $\mathbb{Y}^2$:

$$M(\boldsymbol{p}) \subseteq M^{\mathbb{Y}}(\boldsymbol{p}).$$

We may then determine $R(\boldsymbol{p})$ by considering

$$S^{\mathbb{Y}}(\boldsymbol{j}),$$

for $\boldsymbol{j} \subseteq M^{\mathbb{Y}}(\boldsymbol{p})$. The remaining sections detail how $R(\boldsymbol{p})$ may be determined.

As we only assume that $S^{\mathbb{Y}}$ may be computed, S may be partial: the domain of S must be taken into account. With interval arithmetics that track $\mathcal{P}_|$, such as $\mathbb{J}^{|\mathbb{T}}$, the domain of $S^{\mathbb{Y}}$ is bounded, and the domain of S may be accounted for. With interval arithmetics that do not track $\mathcal{P}_|$, such as $\mathbb{J}$, we have no information as to the domain of $S^{\mathbb{Y}}$. Two approaches may be taken with such arithmetics. This lack of information may be accounted for when performing the interval comparisons which occur while evaluating $S^{\mathbb{Y}}$, or after the evaluation of $S^{\mathbb{Y}}$ has completed. We call the former approach "early accounting", and the latter approach "deferred accounting". The latter approach is preferable, as it allows for better renderings. The two approaches are compared in section 4.2.3.

4.2.2 Sequential Rendering

A rendering R may be built up pixel by pixel. Each pixel is visited once. Throughout this section, simple graphs are presented as examples, to reduce clutter. The solutions presented, generally, handle more sophisticated problems well. Some of the interval bounds, and graphs, given may seem optimistic. Keep in mind that when each free variable, namely x and y, appears at most once within an evaluation, optimal bounds are produced, using constant interval arithmetic.

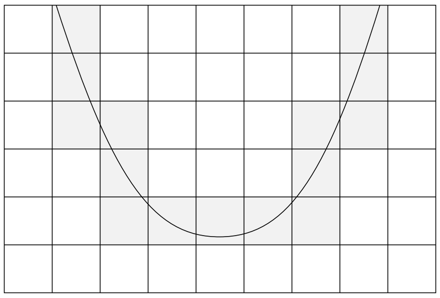
4.2.3 Pixel Testing

From the interval inclusion property we know that

$$\forall[\boldsymbol{p} \in R] \qquad \left(\sup_{\boldsymbol{x} \in M(\boldsymbol{p})} S(\boldsymbol{x}) \right) \sqsubseteq S^{\mathbb{Y}}(M^{\mathbb{Y}}(\boldsymbol{p})),$$

so setting $R(\boldsymbol{p})$ to $S^{\mathbb{Y}}(M^{\mathbb{Y}}(\boldsymbol{p}))$, for each pixel $\boldsymbol{p}$, will generate a rendering of $S^{\mathbb{Y}}$.

An example $\mathbb{J}$ rendering follows:

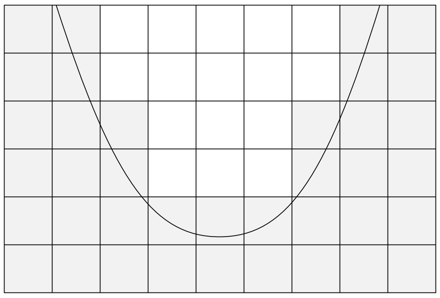


$R_{\Box}[(y = x^2)^{\mathbb{J}}]$

The solid line depicts the associated graph. $R_{\Box}$ denotes a rendering produced using pixel testing. For each pixel $\boldsymbol{p}$, $S^{\mathbb{J}}(M^{\mathbb{J}}(\boldsymbol{p}))$ is computed and $R(\boldsymbol{p})$ is then set accordingly. An example evaluation follows:

$$
\begin{array}{ll}
 & S^{\mathbb{J}}(M^{\mathbb{J}}(\boldsymbol{p})) \\
\leadsto & S^{\mathbb{J}}(\langle 1,2\rangle, \langle 2,3\rangle) \\
\leadsto & \langle 2,3\rangle = \langle 1,2\rangle^2 \\
\leadsto & \langle 2,3\rangle = \langle 1,4\rangle \\
\leadsto & \mathbb{TF}.
\end{array}
$$

Another example $\mathbb{J}$ rendering follows:



$R_{\Box}[(y < x^2)^{\mathbb{J}}]$

An example $\mathbb{J}$ evaluation of S, using $\mathbb{J}^{|\mathbb{T}}$ notation, follows:

$$
\begin{array}{ll}
 & S^{\mathbb{J}}(M^{\mathbb{J}}(\boldsymbol{p})) \\
\leadsto & S^{\mathbb{J}}(\langle\langle 1,2\rangle|\mathbb{TF}\rangle, \langle\langle -1,0\rangle|\mathbb{TF}\rangle) \\
\leadsto & \langle\langle -1,0\rangle|\mathbb{TF}\rangle < \langle\langle 1,2\rangle|\mathbb{TF}\rangle^2 \\
\leadsto & \langle\langle -1,0\rangle|\mathbb{TF}\rangle < \langle\langle 1,4\rangle|\mathbb{TF}\rangle \\
\leadsto & \langle\mathbb{TF}|\mathrm{T}\rangle.
\end{array}
$$

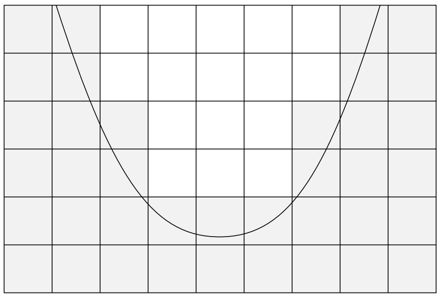
We have chosen to take into account the lack of knowledge of $\mathcal{P}_|$ when evaluating $<^{\mathbb{J}}$; an alternative evaluation of S, which defers this accounting, follows:

$$
\begin{array}{ll}
 & S^{\mathbb{J}}(M^{\mathbb{J}}(\boldsymbol{p})) \\
\leadsto & S^{\mathbb{J}}(\langle\langle 1,2\rangle|\mathbb{TF}\rangle, \langle\langle -1,0\rangle|\mathbb{TF}\rangle) \\
\leadsto & \langle\langle -1,0\rangle|\mathbb{TF}\rangle < \langle\langle 1,2\rangle|\mathbb{TF}\rangle^2 \\
\leadsto & \langle\langle -1,0\rangle|\mathbb{TF}\rangle < \langle\langle 1,4\rangle|\mathbb{TF}\rangle \\
\leadsto & \langle\mathrm{T}|\mathbb{TF}\rangle
\end{array}
$$

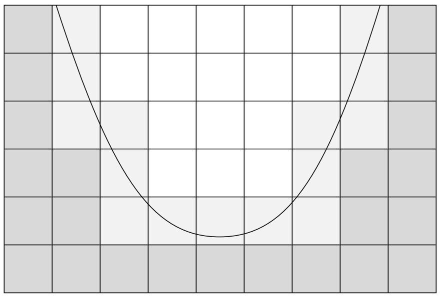
Although the evaluation result is T, we do not know $\mathcal{P}_{|}(S)$, so we must assume that $\mathcal{P}_{|}(S) = \mathbb{TF}$. We must therefore set $R(\boldsymbol{p})$ to $\mathbb{TF}$ using either accounting approach. The two approaches differ when rendering

$$y \ngeq x^2.$$

The two renderings follows:



Deferred Accounting



Early Accounting

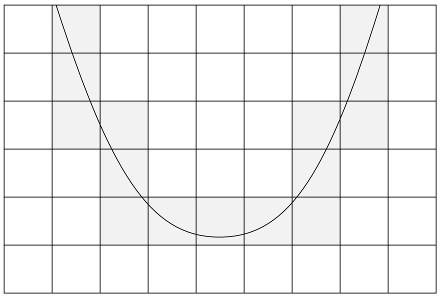
An example evaluation, using early accounting, follows:

$$\begin{array}{ll} & S^{\mathbb{J}}(M^{\mathbb{J}}(\boldsymbol{p})) \\ \rightsquigarrow & S^{\mathbb{J}}(\langle\langle 1,2\rangle|\mathbb{TF}\rangle, \langle\langle -1,0\rangle|\mathbb{TF}\rangle) \\ \rightsquigarrow & \langle\langle -1,0\rangle|\mathbb{TF}\rangle \ngeq \langle\langle 1,2\rangle|\mathbb{TF}\rangle^2 \\ \rightsquigarrow & \langle\langle -1,0\rangle|\mathbb{TF}\rangle \ngeq \langle\langle 1,4\rangle|\mathbb{TF}\rangle \\ \rightsquigarrow & \langle \mathrm{T}|\mathrm{T}\rangle; \end{array}$$

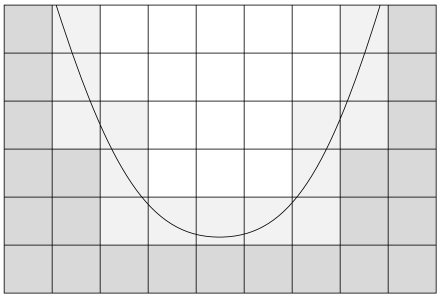
the corresponding evaluation using deferred accounting is as follows:

$$\begin{array}{ll} & S^{\mathbb{J}}(M^{\mathbb{J}}(\boldsymbol{p})) \\ \rightsquigarrow & S^{\mathbb{J}}(\langle\langle 1,2\rangle|\mathbb{TF}\rangle, \langle\langle -1,0\rangle|\mathbb{TF}\rangle) \\ \rightsquigarrow & \langle\langle -1,0\rangle|\mathbb{TF}\rangle \ngeq \langle\langle 1,2\rangle|\mathbb{TF}\rangle^2 \\ \rightsquigarrow & \langle\langle -1,0\rangle|\mathbb{TF}\rangle \ngeq \langle\langle 1,4\rangle|\mathbb{TF}\rangle \\ \rightsquigarrow & \langle \mathrm{T}|\mathbb{TF}\rangle. \end{array}$$

Two $\mathbb{J}^{|\mathbb{T}}$ renderings, corresponding to the preceding $\mathbb{J}$ renderings, follow:



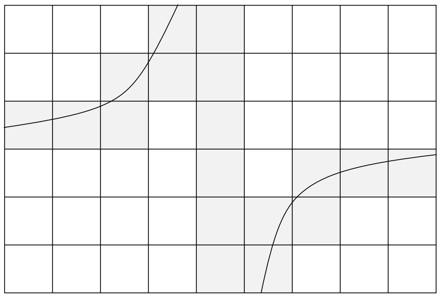
$R_{\Box}[(y = x^2)^{\mathbb{J}^{|\mathbb{T}}}]$



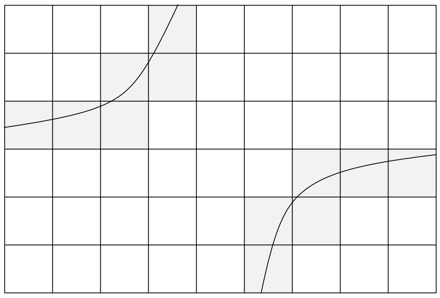
$R_{\Box}[(y < x^2)^{\mathbb{J}^{|\mathbb{T}}}]$

Superior inequality renderings are possible using $\mathbb{J}^{|\mathbb{T}}$ instead of $\mathbb{J}$.

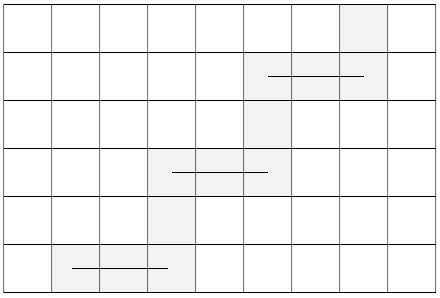
For discontinuous equations, $\mathbb{J}^{|\mathbb{T}\star}$ renderings are often superior to $\mathbb{J}^{|\mathbb{T}}$ renderings. Consider the following renderings:



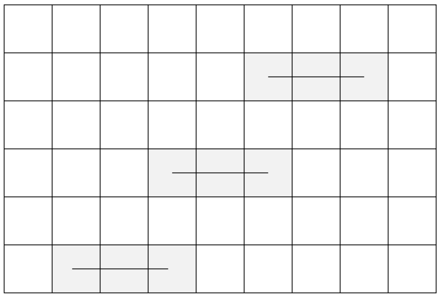
$R_{\square}[(y = x^{-1})^{\mathbb{J}^{|\mathbb{T}}}]$



$R_{\square}[(y = x^{-1})^{\mathbb{J}^{|\mathbb{T}\star}}]$



$R_{\square}[(y = \lfloor x \rfloor)^{\mathbb{J}^{|\mathbb{T}}}]$



$R_{\square}[(y = \lfloor x \rfloor)^{\mathbb{J}^{|\mathbb{T}\star}}]$

With discontinuous specifications, set-based interval arithmetics may sharply bound discontinuous pieces; without this ability to use several bounds, the discontinuous pieces must be bound with a single interval.

4.2.4 Subpixel Testing

After an uninformative pixel test, where

$$S^{\mathbb{Y}}(M^{\mathbb{Y}}(\boldsymbol{p})) \rightsquigarrow \mathbb{TF},$$

subpixel testing may be performed. If

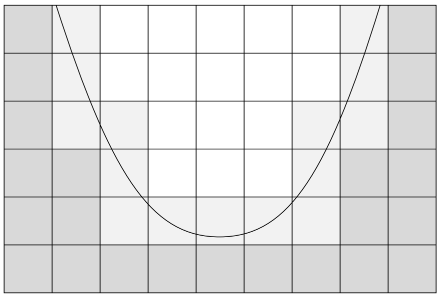
$$\boldsymbol{j} \in M^{\mathbb{Y}}(\boldsymbol{p}) \quad \text{and} \quad S^{\mathbb{Y}}(\boldsymbol{j}) \rightsquigarrow \mathrm{T},$$

then we set $R(\boldsymbol{p})$ to T, since

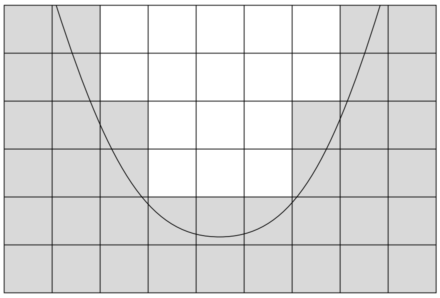
$$\boldsymbol{j} \in M^{\mathbb{Y}}(\boldsymbol{p}), \quad S^{\mathbb{Y}}(\boldsymbol{j}) \rightsquigarrow \mathrm{T} \;\Rightarrow\; \left(\sup_{\boldsymbol{x} \in M(\boldsymbol{p})} S(\boldsymbol{x}) \right) \;=\; \mathrm{T}.$$

The sample $\boldsymbol{j}$ is commonly chosen to be a corner of $M(\boldsymbol{p})$. We may refer to this method as subpixel sample testing, to distinguish it from the other forms of subpixel testing, which are presented later.

Compare the following two $\mathbb{J}^{|\mathbb{T}}$ renderings of $y < x^2$, produced using, and not using, subpixel testing:



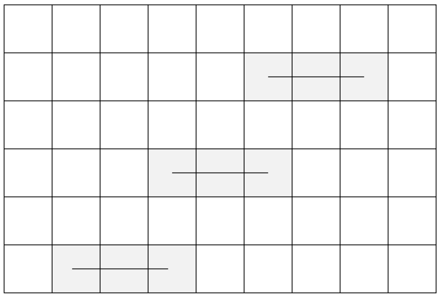
$R_{\Box}[(y < x^2)^{\mathbb{J}^{|\mathbb{T}}}]$



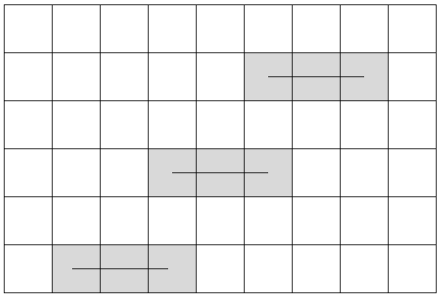
$R_{\boxdot}[(y < x^2)^{\mathbb{J}^{|\mathbb{T}}}]$

$R_{\boxdot}$ denotes a rendering produced using subpixel sample testing; all subpixel rendering methods work in conjunction with pixel testing. Of course, renderings produced using subpixel testing depend on which portion of the subpixel was chosen to test. The rendering depicted is optimistic; for every pixel on the boundary, a sample within the graph was chosen.

Consider rendering $y = \lfloor x \rfloor$; here are two $\mathbb{J}^{|\mathbb{T}\star}$ renderings, produced using, and not using, subpixel testing:



$R_{\Box}[(y = \lfloor x \rfloor)^{\mathbb{J}^{|\mathbb{T}\star}}]$



$R_{\boxdot}[(y = \lfloor x \rfloor)^{\mathbb{J}^{|\mathbb{T}\star}}]$

The rendering using subpixel testing was fortuitous, since the sample chosen from each dark grey pixel lay *within* the graph. Later methods will prove to be more reliable in rendering such graphs.

4.2.5 Exhaustive Subpixel Testing

After an uninformative pixel test, where

$$S^{\mathbb{Y}}(M^{\mathbb{Y}}(\boldsymbol{p})) \rightsquigarrow \mathbb{TF},$$

exhaustive subpixel testing may be performed. If

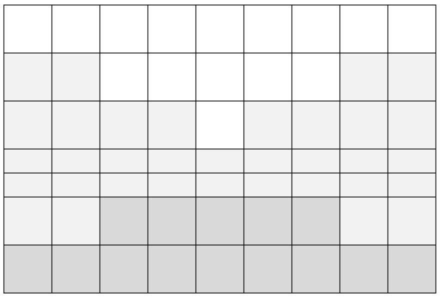
$$M^{\mathbb{Y}}(\boldsymbol{p}) \subseteq \bigcup_m \boldsymbol{j}_m \quad \text{and} \quad \forall m \;\; S^{\mathbb{Y}}(\boldsymbol{j}_m) \rightsquigarrow \mathrm{F},$$

then we set $R(\boldsymbol{p})$ to F, since

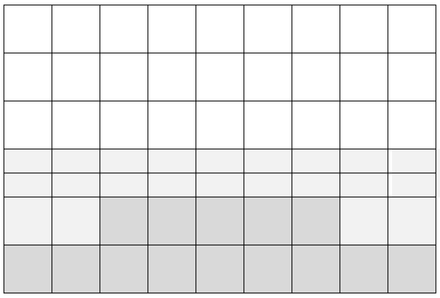
$$M^{\mathbb{Y}}(\boldsymbol{p}) \subseteq \bigcup_m \boldsymbol{j}_m, \;\; \forall m \;\; S^{\mathbb{Y}}(\boldsymbol{j}_m) \rightsquigarrow \mathrm{F} \;\Rightarrow\; \left(\sup_{\boldsymbol{x} \in M(\boldsymbol{p})} S(\boldsymbol{x}) \right) \;=\; \mathrm{F}.$$

Subpixel testing aims to prove a solution exists within a pixel; exhaustive subpixel testing aims to prove no solution exists within a pixel. Of course, if a solution is discovered during exhaustive subpixel testing, the test may abort prematurely and we may set $R(\boldsymbol{p})$ to T.

Consider rendering $y < x - x$; here are two $\mathbb{J}^{|\mathbb{T}}$ renderings, produced using, and not using, exhaustive subpixel testing:

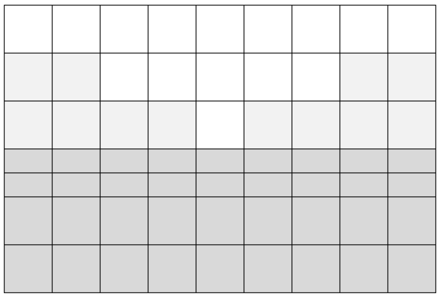


$R_{\square}[(y < x - x)^{\mathbb{J}^{|\mathbb{T}}}]$



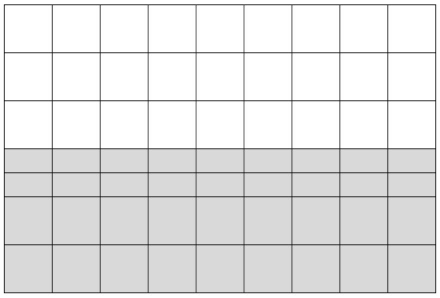
$R_{\boxplus}[(y < x - x)^{\mathbb{J}^{|\mathbb{T}}}]$

$R_{\boxplus}$ denotes a rendering produced using exhaustive subpixel testing. The uncertainty is caused by the form chosen for S: the rendering $R_{\square}[(y < 0)^{\mathbb{J}^{|\mathbb{T}}}]$ restricts the uncertainty to the border. A rendering produced using subpixel testing, but not exhaustive subpixel testing, follows:



$R_{\square}[(y < x - x)^{\mathbb{J}^{|\mathbb{T}\star}}]$

A rendering produced using both subpixel sample testing and exhaustive subpixel testing follows:



$R_{\square\boxplus}[(y < x - x)^{\mathbb{J}^{|\mathbb{T}\star}}]$

The two subpixel tests complement one another; after an uninformative pixel test both subpixel sample testing and exhaustive subpixel testing may be applied. If, after such testing, $R(\boldsymbol{p}) = \mathbb{TF}$, further subpixel testing may be applied.

4.2.6 Continuity-Based Testing

Subpixel sample testing rarely verifies one-dimensional elements of graphs. Consider a graph $G[g = 0]$, with $g : \mathbb{R}^2 \mapsto \mathbb{R}$. Consider $\boldsymbol{j} \in M^{\mathbb{Y}}(\boldsymbol{p})$, $\boldsymbol{k} \in M^{\mathbb{Y}}(\boldsymbol{p})$. If

$$g^{\mathbb{Y}}(\boldsymbol{j}) < 0 < g^{\mathbb{Y}}(\boldsymbol{k}) \quad \text{and} \quad \text{prop}_{\Delta}(S^{\mathbb{Y}}(\boldsymbol{j} \cup \boldsymbol{k})) = \mathrm{T},$$

then we set $R(\boldsymbol{p})$ to T, since

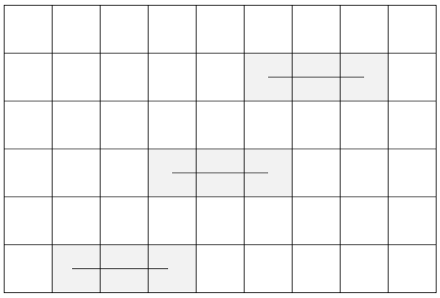
$$g^{\mathbb{Y}}(\boldsymbol{j}) < 0 < g^{\mathbb{Y}}(\boldsymbol{k}) \quad \text{and} \quad \text{prop}_{\Delta}(S^{\mathbb{Y}}(\boldsymbol{j} \cup \boldsymbol{k})) = \mathrm{T} \;\Rightarrow\; \exists[\boldsymbol{\xi} \in M^{\mathbb{Y}}(\boldsymbol{p})]\; g(\boldsymbol{\xi}) = 0,$$

so

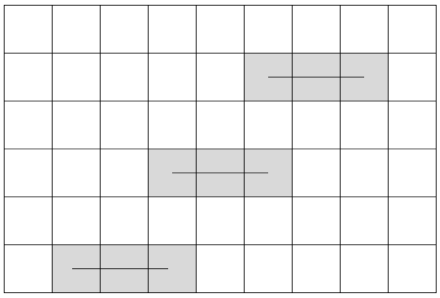
$$\left(\sup_{\boldsymbol{x} \in M(\boldsymbol{p})} S(\boldsymbol{x})\right) = \mathrm{T}.$$

The corners of $M(\boldsymbol{p})$ are common initial choices for $\boldsymbol{j}$ and $\boldsymbol{k}$. Continuity-based testing is often used in place of subpixel sample testing where $=$ occurs in a specification.

Compare the following two $\mathbb{J}|\mathbb{T}$ renderings of $y = \lfloor x \rfloor$, produced using, and not using, continuity-based testing:



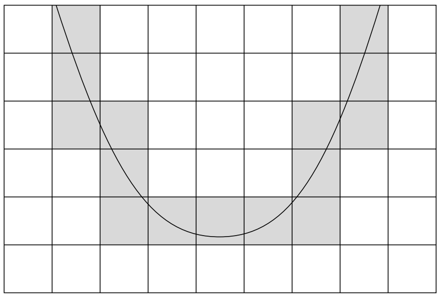
$R_{\square}[(y = \lfloor x \rfloor)^{\mathbb{J}|\mathbb{T}}]$



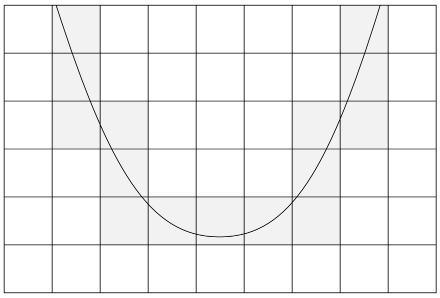
$R_{\Delta}[(y = \lfloor x \rfloor)^{\mathbb{J}|\mathbb{T}}]$

R_{Δ} denotes a rendering produced using continuity-based testing. With sample testing, a sample must be chosen which lies within a graph; with continuity-based testing, a pair of samples must be chosen such that the graph lies between the two samples. If each sample is a point within $M^{\mathbb{J}}(\boldsymbol{p})$, and the samples are chosen uniformly and independently, the outer pixels of each step have a 25% chance of being verified as T, while the inner pixel of each step has a 50% chance of being verified as T. With subpixel sample testing, each pixel has a 0% chance of being verified as T.

Consider the following two renderings, both of which employ continuity-based testing:



$R_{\Delta}[(y - x^2 = 0)^{\mathbb{J}|\Delta\mathbb{T}}]$



$R_{\Delta}[(y^2 - 2x^2y + x^4 = 0)^{\mathbb{J}|\Delta\mathbb{T}}]$

Continuity-based testing fares poorly with $(y^2 - 2x^2y + x^4 = 0)$, as $(y^2 - 2x^2y + x^4)$ is non-negative for all (x, y).

4.2.7 Linear Interval Arithmetic

Let $\mathbb{Y}$ denote a two-dimensional linear interval arithmetic, such as $\mathbb{M}_2$ or $\mathbb{M}_2{}^{|\mathbb{T}}$. We will not ensure that R represents $S : \mathbb{R}^2 \mapsto \mathbb{B}$ directly; we will instead work with $S^{\mathbb{Y}} : \mathbb{Y}^2 \mapsto F(\mathbb{M}_2)$. The interval specification $S^{\mathbb{Y}}$ is computed by evaluating the specification S using the interval arithmetic $\mathbb{Y}$. The interval inclusion property assures us that

$$\forall[\boldsymbol{x} \in \boldsymbol{m}]\ S(\boldsymbol{x}) \subseteq S^{\mathbb{Y}}(\boldsymbol{m}).$$

Let $M^{\mathbb{Y}}(\boldsymbol{p})$ describe $M(\boldsymbol{p})$, using an element of $\mathbb{Y}^2$:

$$M(\boldsymbol{p}) \subseteq M^{\mathbb{Y}}(\boldsymbol{p}).$$

We may then determine $R(\boldsymbol{p})$ by considering

$$S^{\mathbb{Y}}(\boldsymbol{m}),$$

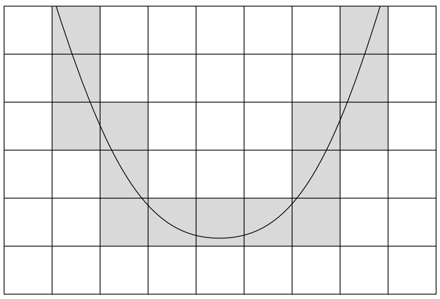
for $\boldsymbol{m} \subseteq M^{\mathbb{Y}}(\boldsymbol{p})$. The remaining sections detail how $R(\boldsymbol{p})$ may be determined. We account for the domain of S as before.

Other linear interval arithmetics may be used, such as $\mathbb{M}_1$ or $\mathbb{M}_3$, given an appropriate $M^{\mathbb{Y}}$. Better renderings may be obtained by taking S into account when choosing $\mathbb{Y}$ and $M^{\mathbb{Y}}$.

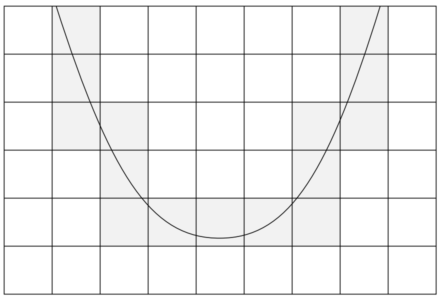
4.2.8 Sequential Rendering

As before, a rendering is built pixel by pixel. The four tests described in the previous section may be utilized with linear interval arithmetics. The tests simply utilize linear interval arithmetic in place of constant interval arithmetic.

For line-like renderings, subpixel testing is not always required when using a linear interval arithmetic. Consider the following two renderings:



$R_\square[(y-x^2=0)^{\mathbb{M}_2{}^{|\Delta\mathbb{T}}}]$



$R_\square[(y-x^2=0)^{\mathbb{J}}]$

An example evaluation follows, with $S = (g = 0)$ and $g = y - x^2$:

$$\begin{aligned}
& S^{\mathbb{M}_2{}^{|\Delta\mathbb{T}}}(M^{\mathbb{M}_2{}^{|\Delta\mathbb{T}}}(\boldsymbol{p})) \\
\rightsquigarrow\ & S^{\mathbb{M}_2{}^{|\Delta\mathbb{T}}}(\langle\langle 1+\alpha, 1+\alpha\rangle|\mathrm{T}\Delta\mathrm{T}\rangle, \langle\langle 2+\beta, 2+\beta\rangle|\mathrm{T}\Delta\mathrm{T}\rangle) \\
\rightsquigarrow\ & g^{\mathbb{M}_2{}^{|\Delta\mathbb{T}}}(\langle\langle 1+\alpha, 1+\alpha\rangle|\mathrm{T}\Delta\mathrm{T}\rangle, \langle\langle 2+\beta, 2+\beta\rangle|\mathrm{T}\Delta\mathrm{T}\rangle) = 0 \\
\rightsquigarrow\ & \langle\langle 2+\beta, 2+\beta\rangle|\mathrm{T}\Delta\mathrm{T}\rangle - \langle\langle 1+\alpha, 1+\alpha\rangle|\mathrm{T}\Delta\mathrm{T}\rangle^2 = 0 \\
\rightsquigarrow\ & \langle\langle 2+\beta, 2+\beta\rangle|\mathrm{T}\Delta\mathrm{T}\rangle - \langle\langle \tfrac{3}{4}+3\alpha, 1+3\alpha\rangle|\mathrm{T}\Delta\mathrm{T}\rangle = 0 \\
\rightsquigarrow\ & f_{=0}(\langle d(\alpha,\beta)|\mathrm{T}\Delta\mathrm{T}\rangle),\ \text{with } d(\alpha,\beta) = \langle 1-3\alpha+2\beta, 1\tfrac{1}{4}-3\alpha+2\beta\rangle.
\end{aligned}$$

From the evaluation we know that

$$d(0,0) = \langle 1, 1\tfrac{1}{4}\rangle, \quad d(1,0) = \langle -2, -1\tfrac{3}{4}\rangle,$$

and that g is continuous over $M(\boldsymbol{p})$; it follows that

$$\exists[\boldsymbol{\xi} \in M(\boldsymbol{p})] \;\; g(\boldsymbol{\xi}) = 0,$$

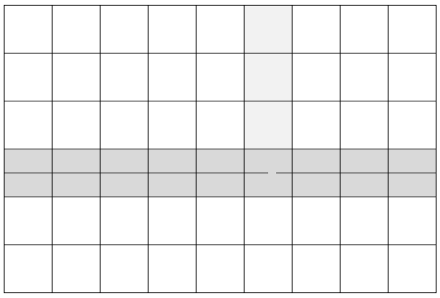
since

$$g(M^{\mathbb{M}_2|\Delta\mathbb{T}}(\boldsymbol{p}))(\alpha, \beta) \subseteq d(\alpha, \beta).$$

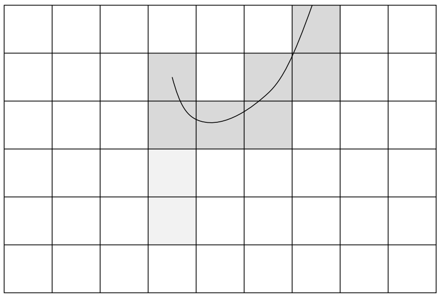
We may therefore set $R(\boldsymbol{p})$ to T.

Of course, subpixel evaluation is still needed to combat the interval over-estimation usually present in large specifications. Continuity information is usually needed when rendering specifications involving equality.

The following two renderings were produced using constant interval arithmetic and all of the subpixel tests described:



$R_{\square\boxplus\triangle}[(y = \frac{x}{x})^{\mathbb{J}|\Delta\mathbb{T}_\star}]$



$R_{\square\boxplus\triangle}[(y = x^x)^{\mathbb{J}|\Delta\mathbb{T}_\star}]$

The light grey pixels will not be resolved using any constant interval arithmetic. This is clear after noticing

$$\frac{\langle 0, \epsilon\rangle}{\langle 0, \epsilon\rangle} \rightsquigarrow \langle 0, \infty\rangle, \quad \frac{\langle -\epsilon, 0\rangle}{\langle -\epsilon, 0\rangle} \rightsquigarrow \langle 0, \infty\rangle,$$

and

$$\langle 0, \epsilon\rangle^{\langle 0, \epsilon\rangle} \rightsquigarrow \langle 0, 1\rangle,$$

for $\epsilon \in (0, 1]$.

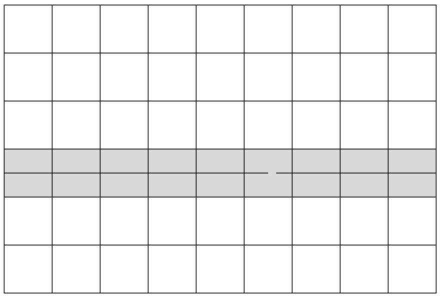
The light grey pixels may be resolved with linear interval arithmetic since operations may consider the dependence of the interval arguments upon the system parameters, namely x and y. For our preceding examples, note that

$$\frac{\langle \epsilon\alpha, \epsilon\alpha\rangle}{\langle \epsilon\alpha, \epsilon\alpha\rangle} \rightsquigarrow \langle 1, 1\rangle, \quad \frac{\langle -\epsilon + \epsilon\alpha, -\epsilon + \epsilon\alpha\rangle}{\langle -\epsilon + \epsilon\alpha, -\epsilon + \epsilon\alpha\rangle} \rightsquigarrow \langle 1, 1\rangle,$$

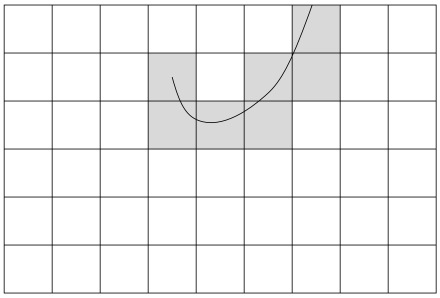
and

$$\langle \epsilon\alpha, \epsilon\alpha\rangle^{\langle \epsilon\alpha, \epsilon\alpha\rangle} \rightsquigarrow \langle -\tfrac{1}{2}(\tfrac{1}{2}\epsilon)^{\frac{1}{2}\epsilon}(\epsilon(1 + \ln\tfrac{1}{2}\epsilon) - 2) + (\tfrac{1}{2}\epsilon)^{\frac{1}{2}\epsilon}\epsilon(1 + \ln\tfrac{1}{2}\epsilon)\alpha, 1 + (\epsilon^\epsilon - 1)\alpha\rangle,$$

The following two renderings were rendered using linear interval arithmetic:



$R_{\square}[(y = \frac{x}{x})^{\mathbb{M}_2|\Delta\mathbb{T}}]$

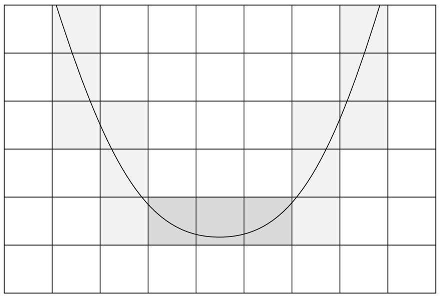


$R_{\square}[(y = x^x)^{\mathbb{M}_2|\Delta\mathbb{T}}]$

Of course, a symbolic optimizer may transform the equations to avoid evaluation difficulties when presented with the simple cases shown.

4.3 Optimization: Function Rendering

Rather than building up the rendering pixel by pixel, the rendering may be built up row by row, or column by column. Each row, or column, is visited once. Consider rendering $y = g(x)$, where $g : \mathbb{R} \mapsto \mathbb{R}$. An example rendering follows:



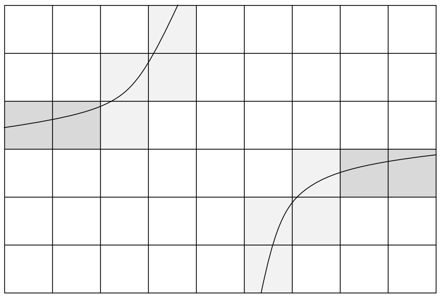
$R_{f\square}[(y = x^2)^{\mathbb{J}|\mathbb{T}}]$

$R_{f\square}$ denotes a rendering produced using column, or row, testing. The function g is evaluated for each column. An example evaluation follows:

$$
\begin{array}{ll}
 & g^{\mathbb{J}|\mathbb{T}}(M_x{}^{\mathbb{J}|\mathbb{T}}(\boldsymbol{r})) \\
\rightsquigarrow & g^{\mathbb{J}|\mathbb{T}}(\langle\langle\frac{1}{2}, 1\rangle|\mathrm{T}\rangle) \\
\rightsquigarrow & \langle\langle\frac{1}{2}, 1\rangle|\mathrm{T}\rangle^2 \\
\rightsquigarrow & \langle\langle\frac{1}{4}, 1\rangle|\mathrm{T}\rangle.
\end{array}
$$

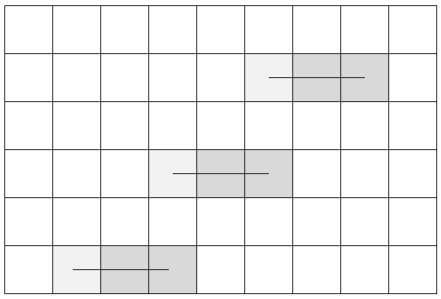
After $g^{\mathbb{J}|\mathbb{T}}(M_x{}^{\mathbb{J}|\mathbb{T}}(\boldsymbol{r}))$ is evaluated, pixels of R may be appropriately set. If $g^{\mathbb{J}|\mathbb{T}}(M_x{}^{\mathbb{J}|\mathbb{T}}(\boldsymbol{r}))$ spans a single pixel, that pixel may be set to T; pixels untouched by $g^{\mathbb{J}|\mathbb{T}}(M_x{}^{\mathbb{J}|\mathbb{T}}(\boldsymbol{r}))$ may be set to F.

Another example rendering follows:



$R_{f\square}[(y = x^{-1})^{\mathbb{J}|\mathbb{T}}]$

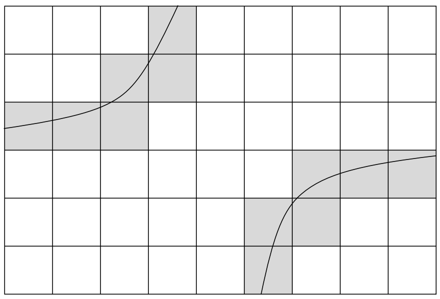
Finer tests may be performed, as with pixel-based testing. Using sample testing, g is evaluated for a portion of each partially undetermined column. An example rendering follows:



$R_{f\boxdot}[(y = \lfloor x \rfloor)^{\mathbb{J}|\mathbb{T}}]$ *or* $R_{f\square}[(y = \lfloor x \rfloor)^{\mathbb{M}|F|_\star}]$

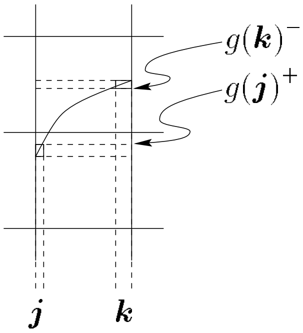
$R_{f\boxdot}$ denotes a rendering produced using sub-column, or sub-row, sample testing. The portions chosen for the above rendering lay on the left side of each pixel column. Another pass, choosing portions on the right side of each pixel column, would set the remaining undetermined pixels to T.

An example rendering, produced using sub-column continuity-based testing, follows:

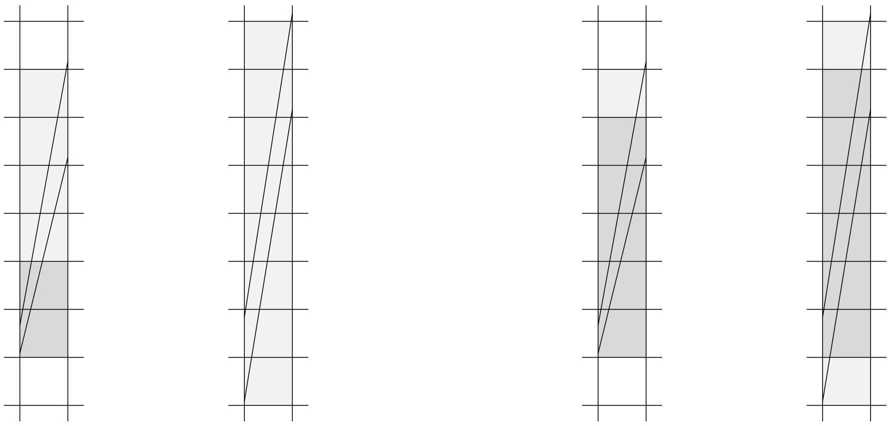


$R_{f\Delta}[(y = x^{-1})^{\mathbb{J}|\Delta\mathbb{T}}]$ *or* $R_{f\square}[(y = x^{-1})^{\mathbb{M}_2|\Delta\mathbb{T}}]$

Continuity-based testing allows column based testing to set long columns of pixels to T with a single test. With continuity-based testing, $g(\boldsymbol{j})$ and $g(\boldsymbol{k})$ are computed; if g is continuous over $\boldsymbol{j} \cup \boldsymbol{k}$ then g is known to smoothly pass from $g(\boldsymbol{j})^+$ to $g(\boldsymbol{k})^-$ if $g(\boldsymbol{k}) \geq g(\boldsymbol{j})$; g is known to smoothly pass from $g(\boldsymbol{k})^+$ to $g(\boldsymbol{j})^-$ if $g(\boldsymbol{j}) \geq g(\boldsymbol{k})$. The following diagram illustrates a portion of a single column from the preceding rendering:



Similar results may be obtained with linear interval arithmetic, without bothering with continuity-based testing. Continuity information is still quite important, however. The following diagram illustrates the information gained by a linear interval arithmetic evaluation of g, with and without continuity information:



$g(\boldsymbol{m})$ $g(\boldsymbol{n})$
Without Continuity Information

$g(\boldsymbol{m})$ $g(\boldsymbol{n})$
With Continuity Information

Without continuity information, pixels that completely enclose $g(\boldsymbol{m})(\alpha)$, for any α, may be set to T; determining such pixels is straight-forward. Such determination mimics the rules behind determining $\Xi^{\star\mathbb{M}}_{\pm}(\sin)$; see section 3.3.15. With continuity information, g must smoothly pass from $(g(\boldsymbol{m}))(0)^+$ to $(g(\boldsymbol{m}))(1)^-$ if $(g(\boldsymbol{m}))(1) \geq (g(\boldsymbol{m}))(0)$; g must smoothly pass from $(g(\boldsymbol{m}))(1)^+$ to $(g(\boldsymbol{m}))(0)^-$ if $(g(\boldsymbol{m}))(0) \geq (g(\boldsymbol{m}))(1)$.

Of course, row and column testing may be used on specifications containing inequalities. Logical combinations of equations and relations are easily accommodated, by passing row or column descriptions along with evaluation results when evaluating the upper levels of S.

4.4 Optimization: Super-Pixel Rendering

The algorithm presented so far is woefully inefficient for pedestrian graphs. Consider rendering

$$x^2 + y^2 = 1,$$

at a resolution of 1024×1024. Rendering $G[x^2 + y^2 = 1]$ with pixel-based testing would require more than one million interval evaluations of S. Efficiency may be improved considerably by using super-pixel testing: testing a group of pixels with a single interval evaluation of S.

Automated symbolic reasoning may deduce that

$$G[x^2 + y^2 = 1] = G[y = \sqrt{1 - x^2} \;\vee\; y = -\sqrt{1 - x^2}],$$

so that column-based testing may be used. Rendering $G[x^2 + y^2 = 1]$ with column-based testing would require more than one thousand interval evaluations of S. Efficiency may be improved considerably by using super-column testing: testing a group of pixels with a single interval evaluation of S.

With either method, partial information presented during rendering may be of limited utility. Super-pixel, and super-column, testing may present informative intermediate renderings.

4.4.1 Constant Interval Arithmetic

The super-pixel rendering algorithm maintains a list L, of pixel clusters. Initially, L consists of a single cluster of pixels; that cluster describes all of R. The rendering starts with all pixels set to $\mathbb{TF}$. On each iteration, a cluster $\boldsymbol{P}$ is removed from L; $S^{\mathbb{Y}}(M^{\mathbb{Y}}(\boldsymbol{P}))$ is then evaluated:

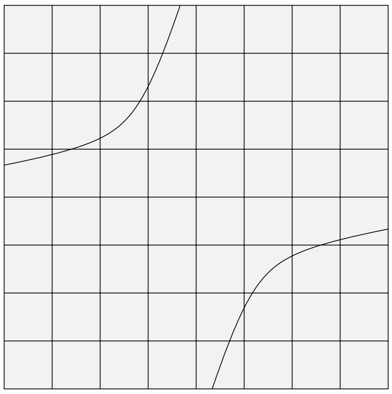
- If $S^{\mathbb{Y}}(M^{\mathbb{Y}}(\boldsymbol{P})) \rightsquigarrow \mathrm{F}$ then $R(\boldsymbol{p})$ is set to F for all $\boldsymbol{p} \in \boldsymbol{P}$.

- If $S^{\mathbb{Y}}(M^{\mathbb{Y}}(\boldsymbol{P})) \rightsquigarrow \mathrm{T}$ then $R(\boldsymbol{p})$ is set to T for all $\boldsymbol{p} \in \boldsymbol{P}$.

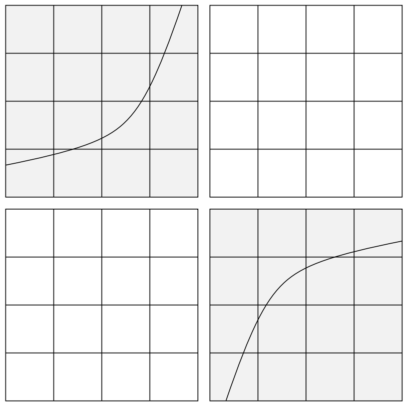
- If $S^{\mathbb{Y}}(M^{\mathbb{Y}}(\boldsymbol{P})) \rightsquigarrow \mathbb{TF}$ then $\boldsymbol{P}$ is cut into several subclusters $\boldsymbol{P}_1, \boldsymbol{P}_2, \ldots, \boldsymbol{P}_m$ with $\boldsymbol{P} \subseteq \cup_i \boldsymbol{P}_i$. All of the new clusters are added to L. The cuts are performed along pixel boundaries so that all pixels belong to at most one member of L.

Subdivision is not performed on clusters which describe single pixels; the pixel testing methods outlined earlier are performed on single pixel clusters.

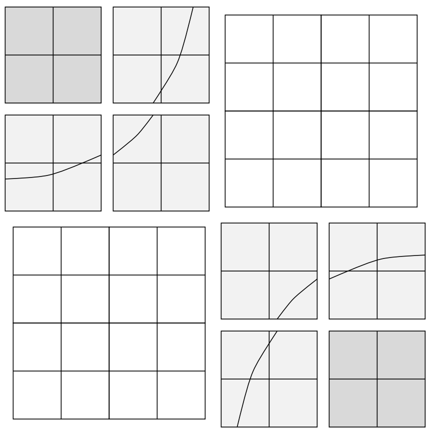
An example rendering, produced using super-pixel testing, follows:



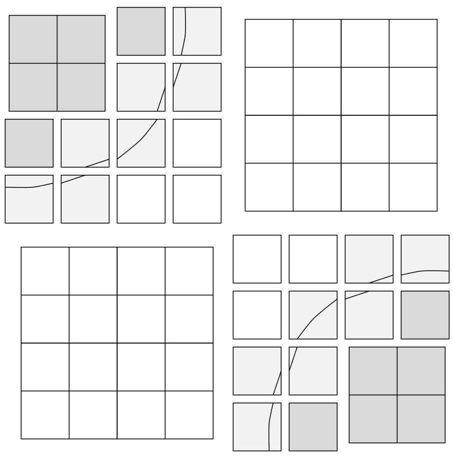
$R_{8\times 8}[(xy < -1)^{\mathbb{J}}]$



$R_{4\times 4}[(xy < -1)^{\mathbb{J}}]$



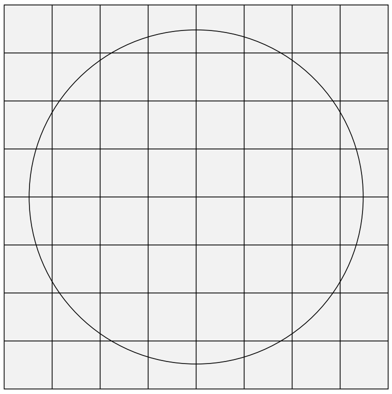
$R_{2\times 2}[(xy < -1)^{\mathbb{J}}]$



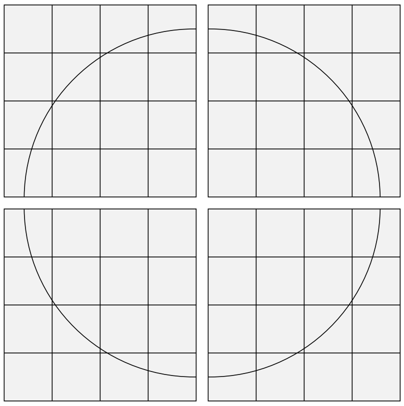
$R_{1\times 1}[(xy < -1)^{\mathbb{J}}]$

$R_{k\times k}$ denotes a super-pixel rendering, where L consists of $k \times k$ clusters of pixels. Section 4.5 will provide a motivation for keeping cluster cuts nicely aligned, as was done above. All intermediate renderings are renderings of $G[(xy < -1)^{\mathbb{J}}]$, and may be presented to the user.

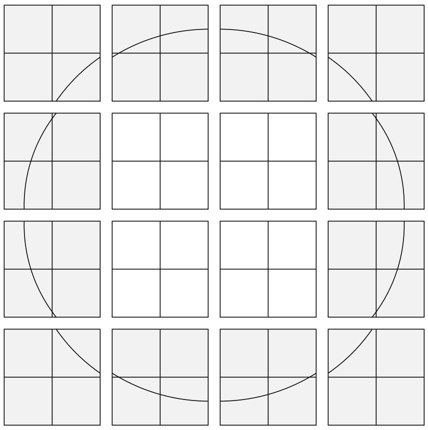
Another example super-pixel rendering follows:



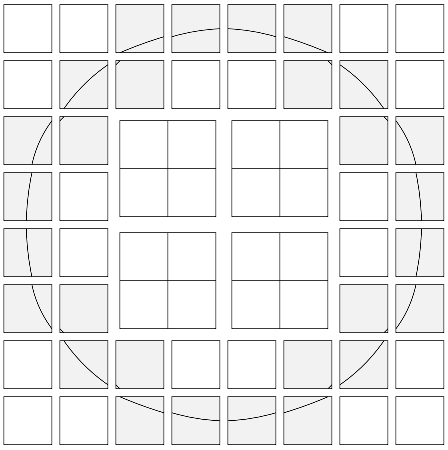
$R_{8\times8}[(x^2+y^2=1)^{\mathbb{J}}]$



$R_{4\times4}[(x^2+y^2=1)^{\mathbb{J}}]$



$R_{2\times2}[(x^2+y^2=1)^{\mathbb{J}}]$



$R_{1\times1}[(x^2+y^2=1)^{\mathbb{J}}]$

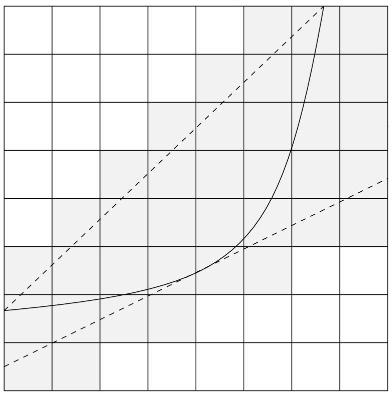
4.4.2 Linear Interval Arithmetic

A similar algorithm may be enacted, but with linear interval arithmetic used to evaluate $S^{\mathbb{Y}}$. The algorithm proceeds as before, unless

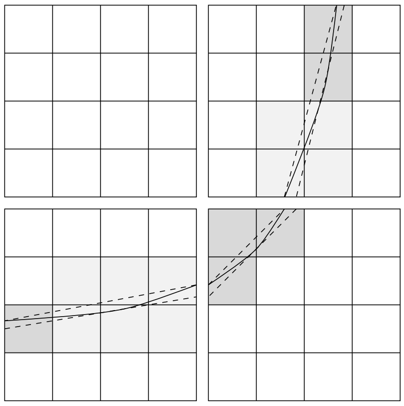
$$\left(S^{\mathbb{Y}}(M^{\mathbb{Y}}(\boldsymbol{p}))\right)^{\mathbb{T}} \rightsquigarrow \mathbb{TF},$$

in which case the relationship between $S^{\mathbb{Y}}(M^{\mathbb{Y}}(\boldsymbol{p}))$ and the system parameters x and y may allow for some pixels to be set to either T or F.

An example follows:



$R_{8\times 8}[(y = e^x)^{\mathbb{M}_2|\Delta\mathbb{T}}]$



$R_{4\times 4}[(y = e^x)^{\mathbb{M}_2|\Delta\mathbb{T}}]$

The dotted lines indicate the constraints determined by the linear interval evaluation of S. Pixels which lie outside of these constraints are set to F. A pixel $\boldsymbol{p}$ may be set to T if the evaluation of $S^{\mathbb{M}_2|\Delta\mathbb{T}}(M^{\mathbb{M}_2|\Delta\mathbb{T}}(\boldsymbol{P}))$ has shown that S is continuous over $\boldsymbol{p}$ and that $y - e^x$ attains both signs. This is seen visually when the constraints divide the pixel into three regions; the constraint region includes no corners. Such determination mimics the one-dimensional case, described in section 4.3.

Pixel assignment may be rapidly performed by using provided graphics primitives. When

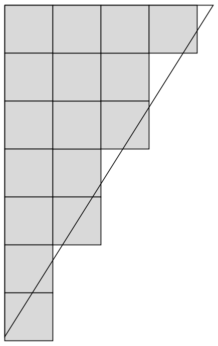
$$S^{\mathbb{Y}}(M^{\mathbb{Y}}(\boldsymbol{p})) \rightsquigarrow \mathrm{F} \quad \text{or} \quad S^{\mathbb{Y}}(M^{\mathbb{Y}}(\boldsymbol{p})) \rightsquigarrow \mathrm{T},$$

the appropriate rectangle is rendered; when

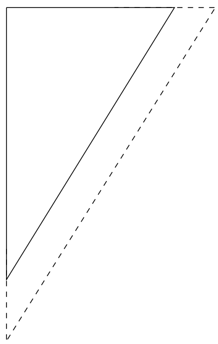
$$(S^{\mathbb{Y}}(M^{\mathbb{Y}}(\boldsymbol{p})))^{\mathbb{T}} \rightsquigarrow \mathbb{TF},$$

appropriate polygons are rendered. As demonstrated earlier, continuity information may also allow some pixels within the constraint region to be set when rendering equations. Usually, a white polygon on either side of the constraint region is rendered. With intimate knowledge of the provided graphics primitives, such rendering may be straightforward.

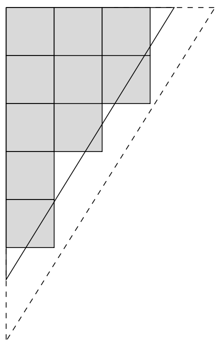
Slight perturbation of the polygon may ensure that pixels are not set incorrectly, as the following diagram suggests:



Original Polygon



Slight Perturbation



Shrunken Polygon

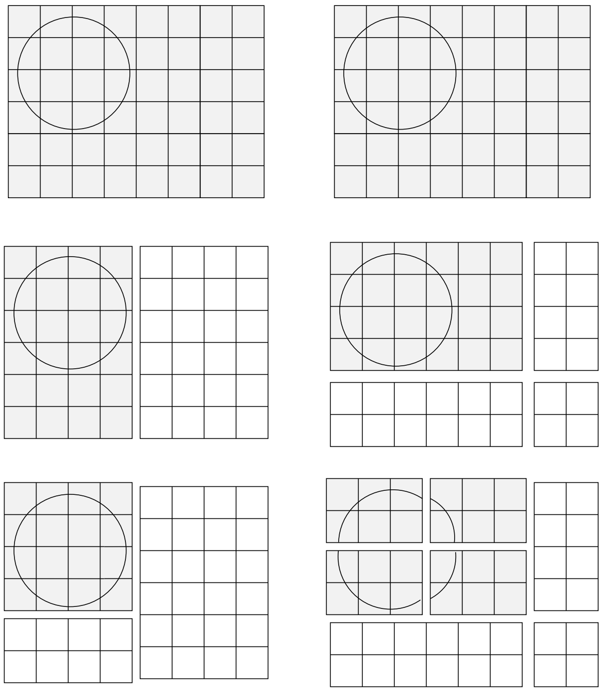
The affected pixels are shown in dark grey; unaffected pixels are not shown.

Precise, rapid pixel control is posssible; a rapid polygon rendering may be followed by manipulation of the pixels along the perimeter of the polygon. A precise rendering of the graph may be deferred until the clusters describe small collections of pixels; polygon perturbation may ensure all intervening renderings still represent G.

Employing sophisticated interval arithmetics requires sophisticated graphics primitives; $\mathbb{V}_2$ requires primitives which render conic sections. Unavailable graphics primitives may be implemented, but such implementation negates part of the advantage of using a more sophisticated interval arithmetic. When using sophisticated interval arithmetics, demotions may be used to reduce the variety of graphics primitives needed: $\mathbb{V}_2 \to \mathbb{M}_2$ allows $\mathbb{V}_2$ to be used with polygon-filling primitives; $\mathbb{M}_2 \to \mathbb{J}$ and $\mathbb{V}_2 \to \mathbb{J}$ allow $\mathbb{M}_2$ and $\mathbb{V}_2$ to be used with rectangle-filling primitives.

4.4.3 Cut Heuristics

The efficiency of the super-pixel method depends on the cuts performed. Compare the following two cutting patterns:



Cutting Pattern A *Cutting Pattern B*

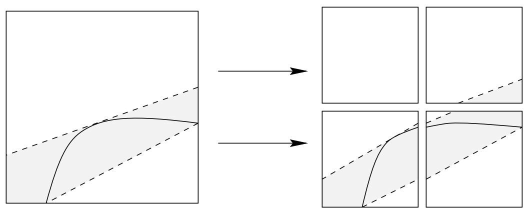
Cutting pattern A has produced more information than cutting pattern B, with fewer cuts and fewer interval evaluations. An optimal cutting pattern may be determined, but would take far more resources than rendering the graph with a simple cutting pattern. After rendering a graph with cutting pattern P, an improved cutting pattern P' may be deduced.

This improved cutting pattern P' may be useful when rendering a similar graph, or when rendering the same graph again. The pattern P' may require less storage than the rendering R, so it may be preferable to store the cutting pattern P' in place of the rendering R. Portions of R may be rendered on demand, using the stored cutting pattern. A sophisticated approach is to store large uniform stretches of R as a cutting pattern but store the intricate details of R directly, to speed later re-rendering. Regardless, an initial cutting pattern must be decided; cutting heuristics guide this decision.

The information already created while rendering is used by the cutting heuristics to determine the next cut. There is no optimal set of heuristics; a heuristic is designed by assuming the graph has certain properties. If the graph does have those properties, application of the heuristic will likely speed the rendering process; if not, application of the heuristic will likely hinder the rendering process.

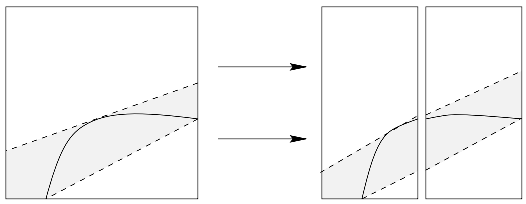
4.4.4 Examples of Cutting Heuristics

Many cutting heuristics are possible. The presented super-pixel renderings cut each pixel cluster into four equal pieces.



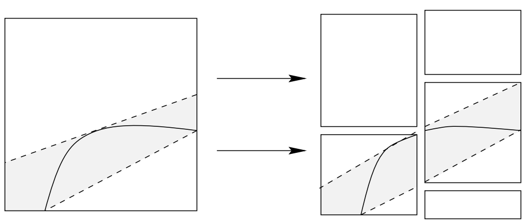
Four-Way Cut

Another possibility is to consider the consraint region, and to cut each pixel cluster into two equal pieces, cutting across the longer side of the region.

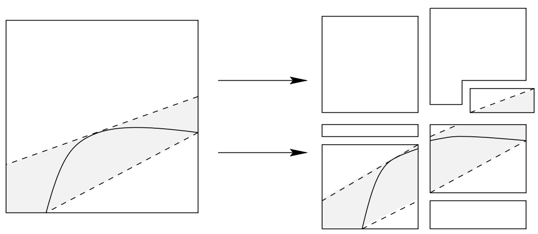


Two-Way Cut

After cutting is performed, the domain may be constricted to tightly bound the constrained region.



Constricting Two-Way Cut

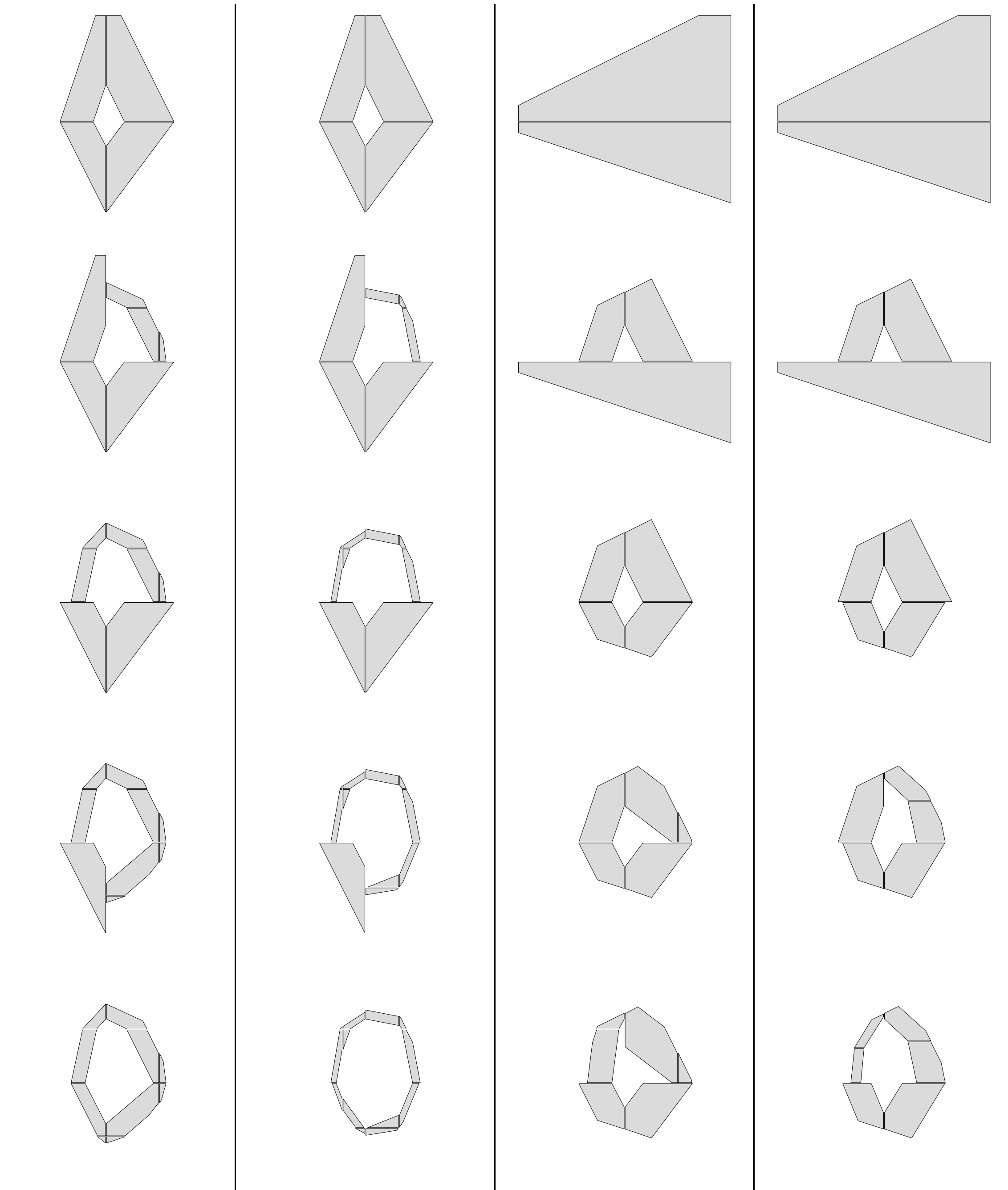


Constricting Four-Way Cut

The regions exluded by such constriction are already determined exactly. The following figure illustrates

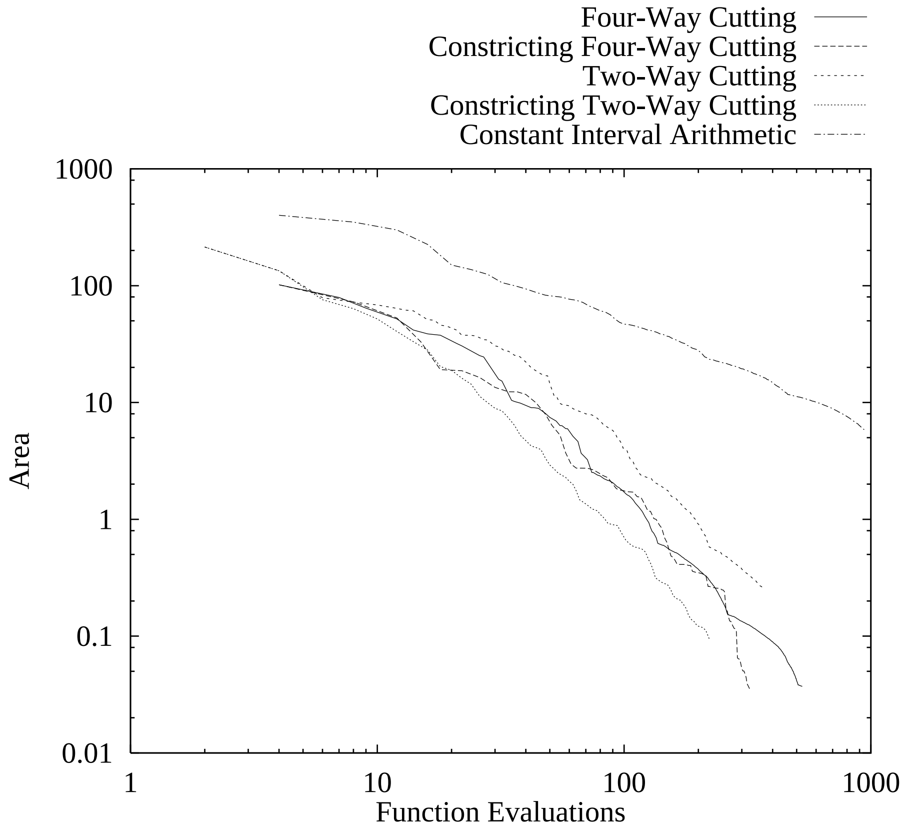
$$2(x-1)^2 + (y-1)^2 = 30,$$

being rendered over the region $[-10, 10] \times [-10, 10]$, with each column representing a different cutting heuristics. As each heuristic requires a different number of interval evaluations per stage, one should not compare the different techniques based on the following figure.



Four-Way Cutting | *Constricting Four-Way Cutting* | *Two-Way Cutting* | *Constricting Two-Way Cutting*

The following diagram illustrates the gains possible using the different cutting heuristics.



The diagram illustrates the area contained within the constrained regions while rendering

$$2(x-1)^2 + (y-1)^2 = 30,$$

over the region $[-10, 10] \times [-10, 10]$ using linear interval arithmetic with various cutting heuristics. Data from a rendering using constant interval arithmetic is also included, for reference.

4.5 Optimization: Caching

The specification S can be broken into several pieces, based upon its dependence on x and y:

$$G[S] = G[S'(\boldsymbol{S^1}, \boldsymbol{S^x}, \boldsymbol{S^y})].$$

For example,

$$G[x^2 + y^2 = x^4 + y^4]$$

may be transformed into

$$G[S'(\boldsymbol{S^x}, \boldsymbol{S^y})],$$

with

$$\boldsymbol{S^x}_0 = x^2, \quad \boldsymbol{S^x}_1 = x^4,$$

$$\boldsymbol{S^y}_0 = y^2, \quad \boldsymbol{S^y}_1 = y^4,$$

$$S' = (\boldsymbol{S^x}_0 + \boldsymbol{S^y}_0 = \boldsymbol{S^x}_1 + \boldsymbol{S^y}_1).$$

This is a natural extension of common sub-expression elimination, present in optimizing compilers [3]. Applying common sub-expression elimination in conjunction with symbolic rewriting, the example is transformed into

$$G[S'(\boldsymbol{S^x}, \boldsymbol{S^y})],$$

with

$$\boldsymbol{S^x}_0 = x^2, \quad \boldsymbol{S^x}_1 = (\boldsymbol{S^x}_0)^2,$$

$$\boldsymbol{S^y}_0 = y^2, \quad \boldsymbol{S^y}_1 = (\boldsymbol{S^y}_0)^2,$$

$$S' = (\boldsymbol{S^x}_0 + \boldsymbol{S^y}_0 = \boldsymbol{S^x}_1 + \boldsymbol{S^y}_1).$$

Such transformations are useful when evaluating S many times, which occurs during rendering.

Let $M(\boldsymbol{p}) = (M_x(\boldsymbol{p}), M_y(\boldsymbol{p}))$; after evaluating

$$S(M(\boldsymbol{p})) = S'(\boldsymbol{S^1}, \boldsymbol{S^x}(M_x(\boldsymbol{p})), \boldsymbol{S^y}(M_y(\boldsymbol{p}))),$$

the evaluation of

$$S(M(\boldsymbol{p}')) = S'(\boldsymbol{S^1}, \boldsymbol{S^x}(M_x(\boldsymbol{p}')), \boldsymbol{S^y}(M_y(\boldsymbol{p}')))$$

is more efficient if

$$M_x(\boldsymbol{p}) = M_x(\boldsymbol{p}') \quad \text{or} \quad M_y(\boldsymbol{p}) = M_y(\boldsymbol{p}');$$

some sub-expressions need not be re-evaluated.

With aligned cuts, it is likely that

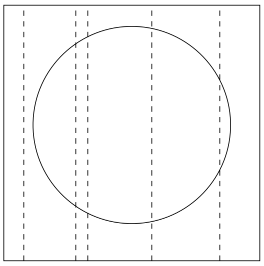
$$S(M(\boldsymbol{p})) = S'(\boldsymbol{S^1}, \boldsymbol{S^x}(M_x(\boldsymbol{p})), \boldsymbol{S^y}(M_y(\boldsymbol{p})))$$

has sub-expressions that have been evaluated before. For example, with cuts aligned along a 32×32 grid,

$$\boldsymbol{S^x}(M_x(\boldsymbol{p})) \quad \text{and} \quad \boldsymbol{S^y}(M_y(\boldsymbol{p}))$$

are each computed 32 times, instead of 1024 times, if sub-expressions are cached; these calculations assume that every grid cell contains one pixel cluster. Caching $\boldsymbol{S^1}$ takes minimal memory, and aids computation considerably; with our previous example, it would be computed once, instead of 1024 times, if $\boldsymbol{S^1}$ is cached.

A better estimate of cache utility may be made by considering the graph being rendered. For $\boldsymbol{S^x}(M_x(\boldsymbol{p}))$, consider vertical lines, as shown in the following figure:



$G[S]$, with Vertical Lines

In the example shown, most lines intersect G twice; it follows that $\boldsymbol{S^x}(M_x(\boldsymbol{p}))$ is usually computed twice, for each possible value of $M_x(\boldsymbol{p})$. Interval evaluation may "smear" the graph, so that $\boldsymbol{S^x}(M_x(\boldsymbol{p}))$ may be computed several times for each actual intersection.

4.6 Optimization: Removing Conditionals

The ideas behind caching may be extended; after the evaluation of

$$S^{\mathbb{Y}}(M^{\mathbb{Y}}(\boldsymbol{p}))$$

has been performed, the evaluation of

$$S^{\mathbb{Y}}(M^{\mathbb{Y}}(\boldsymbol{p}')), \quad \boldsymbol{p}' \subseteq \boldsymbol{p}$$

may be simplified. An example will sufficiently expose the ideas.

Consider rendering $G[y = x^2]$ with constant interval arithmetic. Eventually, x^2 is computed using interval arithmetic, by the following rule:

$$x^2 \rightsquigarrow \begin{cases} \langle {x^-}^2, {x^+}^2 \rangle & \text{if } x^- \geq 0, \\ \langle {x^+}^2, {x^-}^2 \rangle & \text{if } x^+ \leq 0, \\ \langle 0, {x^-}^2 \rangle & \text{if } 0 \in x \text{ and } -x^- \geq x^+, \\ \langle 0, {x^+}^2 \rangle & \text{if } 0 \in x \text{ and } -x^- \leq x^+. \end{cases}$$

Given that the evaluation of x^2 during the computation of $S^{\mathbb{J}}(M^{\mathbb{J}}(\boldsymbol{p}))$ falls into the first case, the evaluation of x^2 during the computation of $S^{\mathbb{J}}(M^{\mathbb{J}}(\boldsymbol{p}'))$, for $\boldsymbol{p}' \subseteq \boldsymbol{p}$, may drop immediately into the first case, without testing x. Similarly with the second case; the other two cases require that x be tested, to produce optimal bounds of x^2. The example given is simple; other operators perform many tests on their arguments before falling into one of many cases. The structure which holds cache information may also hold the additional information needed by the method alluded to in this section.

When rendering a small portion of a specific graph to a high resolution, it may be worthwhile to assemble a new operator to expose the conditionals. Consider rendering $G[y = x^2 + x^4]$ over $[5, 10] \times [-10, 10]$, using constant interval arithmetic. Each operator is evaluated by one of the following rules:

$$x^2 \rightsquigarrow \begin{cases} \langle {x^-}^2, {x^+}^2 \rangle & \text{if } x^- \geq 0, \\ \langle {x^+}^2, {x^-}^2 \rangle & \text{if } x^+ \leq 0, \\ \langle 0, {x^-}^2 \rangle & \text{if } 0 \in x \text{ and } -x^- \geq x^+, \\ \langle 0, {x^+}^2 \rangle & \text{if } 0 \in x \text{ and } -x^- \leq x^+; \end{cases}$$

$$x^4 \rightsquigarrow \begin{cases} \langle {x^-}^4, {x^+}^4 \rangle & \text{if } x^- \geq 0, \\ \langle {x^+}^4, {x^-}^4 \rangle & \text{if } x^+ \leq 0, \\ \langle 0, {x^-}^4 \rangle & \text{if } 0 \in x \text{ and } -x^- \geq x^+, \\ \langle 0, {x^+}^4 \rangle & \text{if } 0 \in x \text{ and } -x^- \leq x^+; \end{cases}$$

$$x + y \rightsquigarrow \langle x^- + y^-, x^+ + y^+ \rangle;$$

while the compound operator $x^2 + x^4$ is evaluated by the following rule:

$$x^2 + x^4 \rightsquigarrow \begin{cases} \langle {x^-}^2 + {x^-}^4, {x^+}^2 + {x^+}^4 \rangle & \text{if } x^- \geq 0, \\ \langle {x^+}^2 + {x^+}^4, {x^-}^2 + {x^-}^4 \rangle & \text{if } x^+ \leq 0, \\ \langle 0, {x^-}^2 + {x^-}^4 \rangle & \text{if } 0 \in x \text{ and } -x^- \geq x^+, \\ \langle 0, {x^+}^2 + {x^+}^4 \rangle & \text{if } 0 \in x \text{ and } -x^- \leq x^+; \end{cases}$$

The rule given was found automatically; the rule was determined by combining the cases of the basic operators. Over the area $[5, 10] \times [-10, 10]$, the rule may be simplified to the following:

$$x^2 + x^4 \rightsquigarrow \langle x^{-2} + x^{-4}, x^{+2} + x^{+4} \rangle.$$

The simplified rule was found automatically, by simply evaluating the operator over the domain of interest. Rounding control is accounted for, when evaluating with $\mathbb{I}$ or $\mathbb{L}$. The example rule would instead be the following:

$$x^2 + x^4 \rightsquigarrow \langle (x^{-2})^{\mathbb{F}-} +^{\mathbb{F}-} (x^{-4})^{\mathbb{F}-}, (x^{+2})^{\mathbb{F}+} +^{\mathbb{F}+} (x^{+4})^{\mathbb{F}+} \rangle.$$

which was again automatically deduced from the associated $\mathbb{I}$ rules for the appropriate basic operators. The larger rules, with larger chains of computation, let an optimizing compiler obtain better CPU utilization. Additionally, fewer rounding mode controls need to be issued with larger chains of computation. Rounding control is often overly expensive; on many systems changing the current rounding mode consumes more resources than a floating point operation, such as multiplication.

A more involved example is evaluating $g^{\mathbb{I}}(x)$,

$$g(x) = x \cos(x^2 - 1),$$

for $x \in [1, 2]$. After evaluating $g(\langle 1, 2 \rangle)$, it is known that evaluation of each basic operator falls into one case; combining these cases gives the following rule for evaluating $g^{\mathbb{I}}(x)$:

$$g^{\mathbb{I}}(x) \rightsquigarrow \langle x^- \times^{\mathbb{F}-} \cos^{\mathbb{F}-}((x^{+2})^{\mathbb{F}+} -^{\mathbb{F}+} 1), x^+ \times^{\mathbb{F}+} \cos^{\mathbb{F}+}((x^{-2})^{\mathbb{F}-} -^{\mathbb{F}-} 1) \rangle,$$

for any $x \in [1, 2]$. Given a specification S, the interval evaluation $S^{\mathbb{Y}}$ may be quite involved; the challenge is to expose the salient features of $S^{\mathbb{Y}}$ to a sophisticated automatic optimizer.

The derivative of g,

$$\tfrac{d}{dx} g = \cos(x^2 - 1) - 2x^2 \sin(x^2 - 1),$$

when evaluated over the interval $\langle 1.5, 2 \rangle$, is total, continuous, and negative:

$$\left(\tfrac{d}{dx} g\right)^{\mathbb{I}|\Delta\mathbb{T}} (\langle \langle 15 \times 10^{-1}, 2 \times 10^{0} \rangle | \mathrm{T}\Delta\mathrm{T} \rangle) \rightsquigarrow \langle \langle -899 \times 10^{-2}, -319 \times 10^{-3} \rangle | \mathrm{T}\Delta\mathrm{T} \rangle < 0.$$

This implies that g is monotonically decreasing over $[1.5, 2]$ and the preceding evaluation may be improved, by application of the following truth:

$$g^{\mathbb{I}}(\langle x^+, x^+ \rangle)^- \leq g(x^+) \leq g(x^-) \leq g^{\mathbb{I}}(\langle x^-, x^- \rangle)^+.$$

The improved rule is thus given by the following:

$$g^{\mathbb{I}}(x) \rightsquigarrow \langle x^+ \times^{\mathbb{F}-} \cos^{\mathbb{F}-}((x^{+2})^{\mathbb{F}+} -^{\mathbb{F}+} 1), x^- \times^{\mathbb{F}+} \cos^{\mathbb{F}+}((x^{-2})^{\mathbb{F}-} -^{\mathbb{F}-} 1) \rangle,$$

and is valid for any interval $x \subseteq [1.5, 2]$. Clearly, this rule may be found automatically, given the evaluation rules for the basic operators. Such determination may be naturally performed using an interval arithmetic with automatic differentiation.

4.7 Alternative Formalisms

Alternative formal definitions are possible. An interesting possibility is to view the specification as a "true rendering", and actual renderings as demoted forms of this true rendering. With this approach, the true rendering $R^\star$ is simply the specification,

$$R^\star : \mathbb{R}^2 \mapsto \mathbb{B}, \quad R^\star = S,$$

while an actual rendering R is a demotion of the true rendering,

$$R : \mathbb{R}^2 \mapsto \mathbb{T}, \quad \forall[\boldsymbol{x} \in \mathbb{R}^2]\ R^\star(\boldsymbol{x}) \sqsubseteq R(\boldsymbol{x}),$$

$$\forall[\boldsymbol{p} \in R]\ \forall[(\boldsymbol{x}, \boldsymbol{y}) \in \boldsymbol{p}^2]\ R(\boldsymbol{x}) = R(\boldsymbol{y}).$$

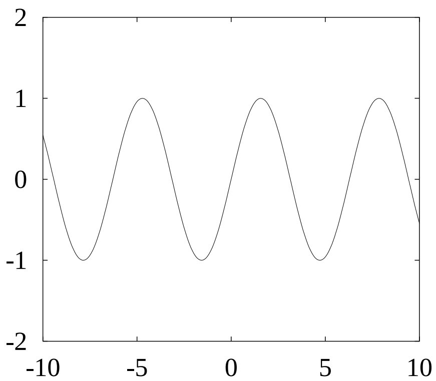
Other approaches are possible, see [23] for formal definitions of traditional rendering techniques.

4.8 Other Work

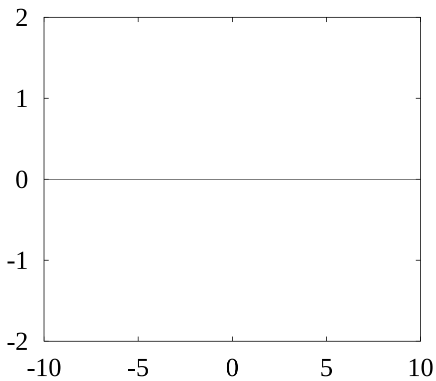
We will now compare the methods presented within this thesis to previous methods, reported by others. First, our approach to graphing will be compared to other approaches put forward. Afterwards, other interval arithmetic extensions will be presented, and compared to our interval arithmetic framework. Our motivating problem, two-dimensional equation rendering, provides a framework for such comparisons.

4.8.1 Sampling

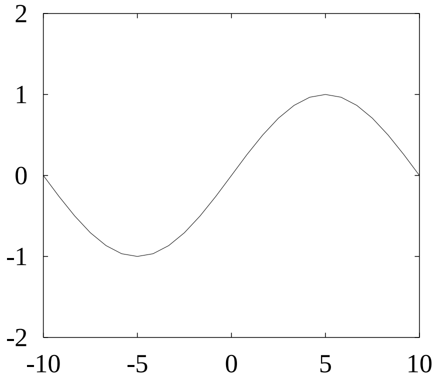
Traditional approaches to rendering functions sample the given function g at various places, by computing $(x, g^{\mathbb{F}}(x))$. After sampling, a rendering is produced by appropriately connecting the samples. We deam such approaches unsatisfactory, as the produced graph may mislead the viewer as to the nature of G. Some examples follow:



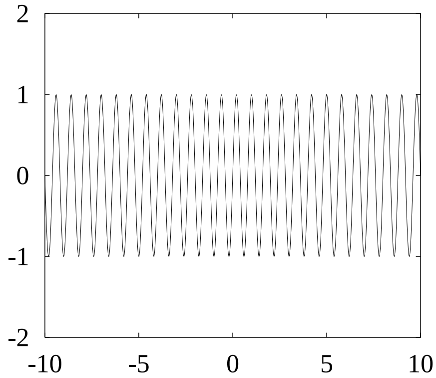
$R_{\mathbb{F}}[y = (9^9 - \sin x) - 9^9 + \sin x]$



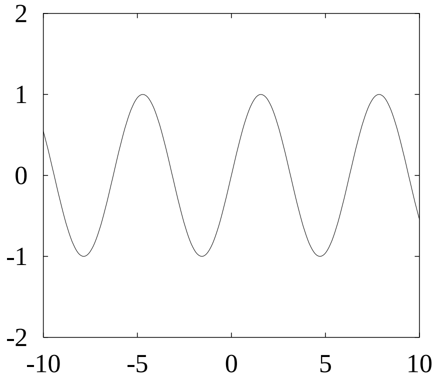
$R[y = (9^9 - \sin x) - 9^9 + \sin x]$



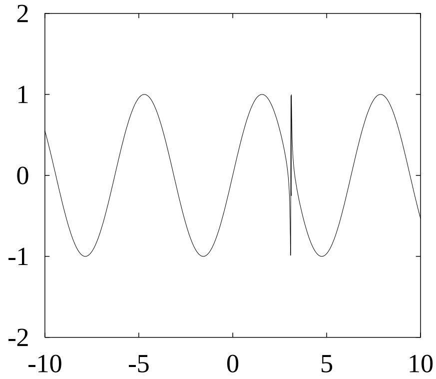
$R_{\mathbb{F}}[y = \sin 2.5\pi x]$



$R[y = \sin 2.5\pi x]$



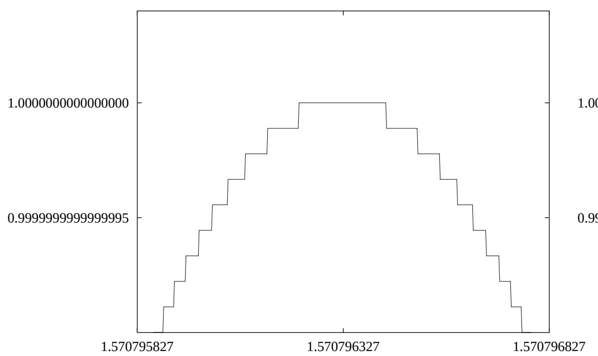
$R_{\mathbb{F}}[y = \sin x \frac{x-3.11}{x-3.1}]$



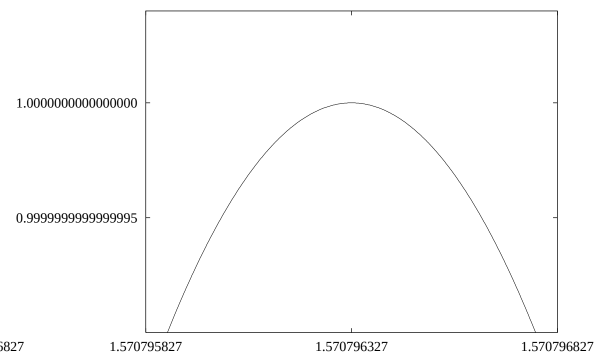
$R[y = \sin x \frac{x-3.11}{x-3.1}]$

$R_{\mathbb{F}}$ denotes a rendering produced using uniform floating-point sampling. The first graph was produced using 100 samples; the second and third were produced using 25 samples each. The third graph has a true rendering, at the given resolution, which may be somewhat misleading, although it would be quite natural for a user to "zoom" into the atypical region. Please see [24] for an extended discussion of sampling techniques in rendering.

These approaches typically break down when investigating minute details of the graph, as the following renderings illustrate:

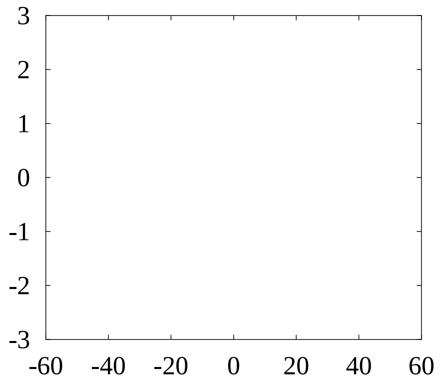


$R_{\mathbb{F}}[y = \sin x]$

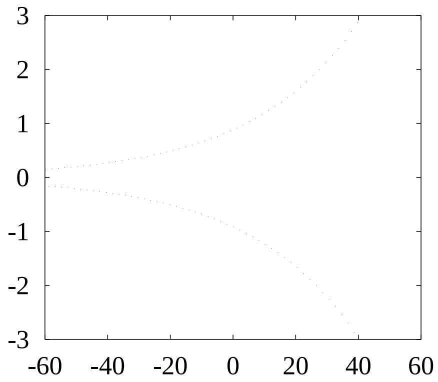


$R[y = \sin x]$

Such ineptitude at fine-scale detail may cause large-scale features to be poorly rendered, as a function may arbitrarily magnify regions. Of course, sampling methods fare poorly when interesting features of a graph occur between floating-point coordinates; consider the following renderings:



$R_{\mathbb{F}}[y = (-1.03)^{x-\pi}]$



$R[y = (-1.03)^{x-\pi}]$

For the preceding example, we consider the general exponentiation operator, which is defined for negative bases with integral exponents: for example, $(-1.1)^2 = 1.21$ and $(-1.1)^3 = -1.331$. All of the examples given in this section are rendered correctly using the appropriate interval techniques which were described in this chapter.

The renderings produced for the above examples finish with each pixel determined exactly, excepting that which displays fine-scale detail. With that function, a course approximation is generated, which contains the true curve. An erstwhile interval-based renderer may increase the precision of the underlying number system, as needed.

More sophisticated sampling algorithms may be similarly fooled, although with more convoluted examples. Without assumptions as to the shape of G, a finite number of floating-point samples gives no information, other than the samples actually computed. Samples computed using floating point rarely lie within G.

4.8.2 Line Tracing

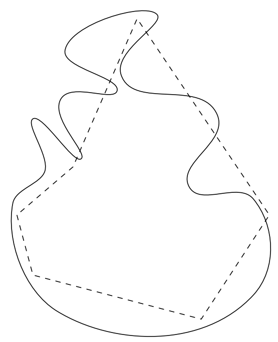
Interval arithmetic may form the basis for rendering algorithms other than those given here. Sophisticated methods are certainly possible. Care should be taken to preserve the strength of the underlying interval arithmetic, so that algorithm guarantees may be given.

In [66], an implicit curve approximation algorithm is given; it is argued, therein, that the given algorithm is more reliable than those based on sampling. This is clear, although the algorithm given assumes the following:

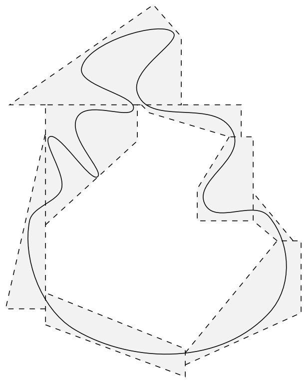
- G is a continuous, 1D manifold;
  - G has no isolated singularities,
  - G has no regions of dimension greater than one,
  - each disjoint curve segment of G is either closed, or has endpoints at the graphing boundary,
- G has no segments aligned with either axis.

The output of the algorithm is a collection of line segments, which approximate G.

Using a linear interval arithmetic in place of a constant interval arithmetic allows a similar algorithm to be constructed, which returns a collection of polygons which *include* G. With domain and continuity tracking in conjunction with automated derivative analysis, some of the assumptions may be lifted, as they may be verified as the algorithm proceeds.



Line Approximation



Polygon Approximation

Many algorithms which use interval arithmetic, do so peripherally, and would benefit from a re-engineering which allows the concepts of interval arithmetic to permeate the entire algorithm. A chief benefit of such re-engineering is that a clear, strong guarantee of program output may be given. With the curve approximation algorithm, one may output the polygons as a collection of (circular) lists, with the guarantee that the curve segment passes from polygon to polygon, in the order given. Strong guarantees may be given for the original algorithm, but they are intricate and somewhat unsatisifying. Strong guarantees may even be given for algorithms based on floating-point, but such guarantees are considerably more intricate, and consequently even less satisfying.

An adoption of linear interval arithmetic in place of constant interval arithmetic increases the efficiency of a method, and may be enacted using a minimal expenditure of effort. Portions which performed derivative analysis may be removed, as such analysis is done automatically, within the linear interval arithmetic library. The desired results may be obtained from the linear interval arithmetic library. A generalized interval arithmetic library may use demotions to shield an application from the interval arithmetic being used.

4.8.3 Extended Interval Arithmetic

Extended interval arithmetic may be used to render graphs. Extended interval arithmetic is similar to $\mathbb{J}^\star$, with the restriction that each interval set contain either one or two intervals. If the set contains two intervals, one interval must contain $-\infty$; the other interval must contain $+\infty$.

An algorithm employing extended interval arithmetic may readily graph specifications given using $+$, $-$, $\times$, and $\div$; the more general interval sets are needed if operations such as $\lfloor x \rfloor$ or $\pm x$ may occur in a specification.

4.8.4 Derivative-Based Methods

When working with interval methods, computed bounds on derivatives often supplement computed bounds on values. A simple example problem will illustrate the core technique: consider bounding the range of $g(x)$, for $x \in [0,1]$. In chapter two, we bounded the range of $g(x)$ by evaluating

$$g^{\mathbb{J}}(\langle 0,1\rangle)$$

or

$$g^{\mathbb{M}}(\langle \alpha,\alpha\rangle);$$

another approach is to evaluate

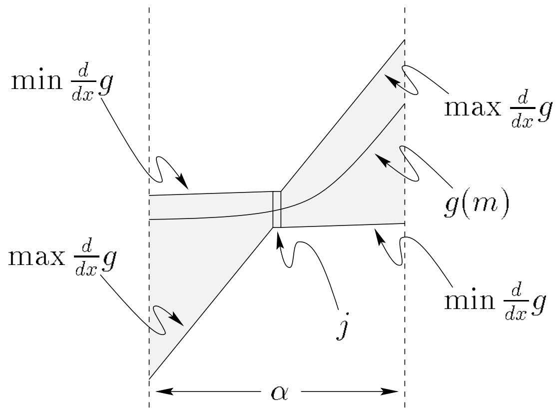
$$g^{\mathbb{J}}(j) + (\alpha - j)(\tfrac{d}{dx}g)^{\mathbb{J}}(\langle 0,1\rangle),$$

with $j \subseteq \langle 0,1\rangle$. Usually, j is taken to be the midpoint of the domain: $j = \langle \frac{1}{2},\frac{1}{2}\rangle$, so

$$g^{\mathbb{J}}(\langle \tfrac{1}{2},\tfrac{1}{2}\rangle) + (\alpha - \tfrac{1}{2})(\tfrac{d}{dx}g)^{\mathbb{J}}(\langle 0,1\rangle),$$

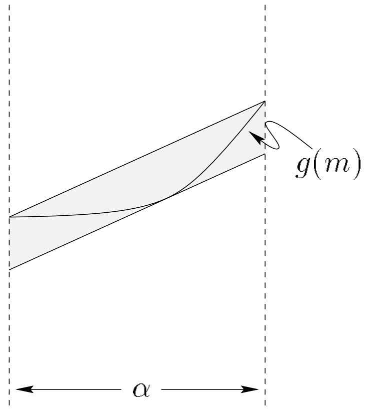
bounds $g(x)$ for $x \in [0,1]$. The midpoint is chosen, as it often produces the best bounds possible with this approach.

This approach, based on the first derivative, may be graphically depicted, as follows:



$$g^{\mathbb{J}}(j) + (\alpha - j)(\tfrac{d}{dx}g)^{\mathbb{J}}(\langle 0,1\rangle)$$

The linear interval approach may be similarly depicted, as follows:

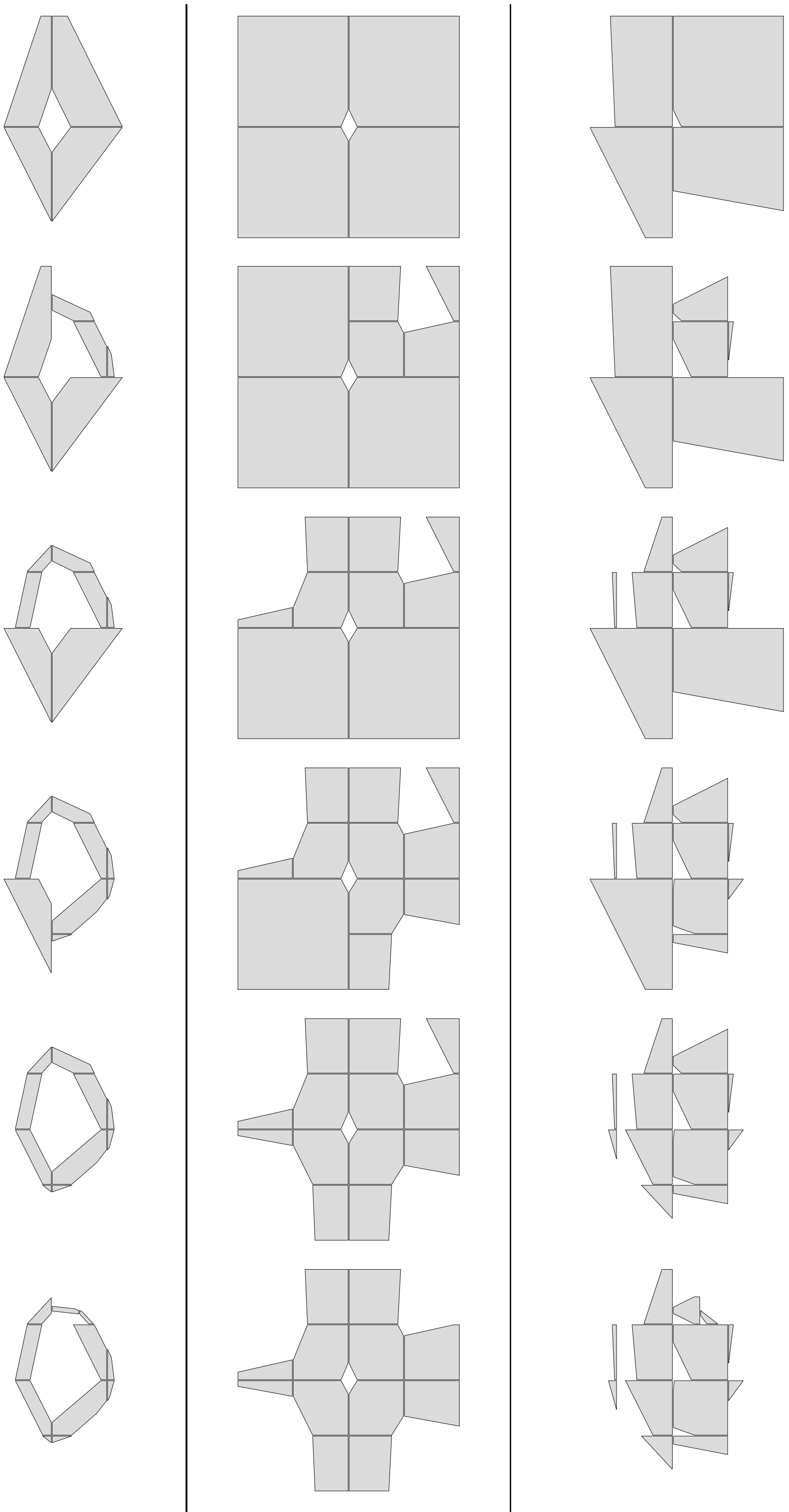


$$g^{\mathbb{M}}(\langle \alpha,\alpha\rangle)$$

The following diagram depicts renderings of

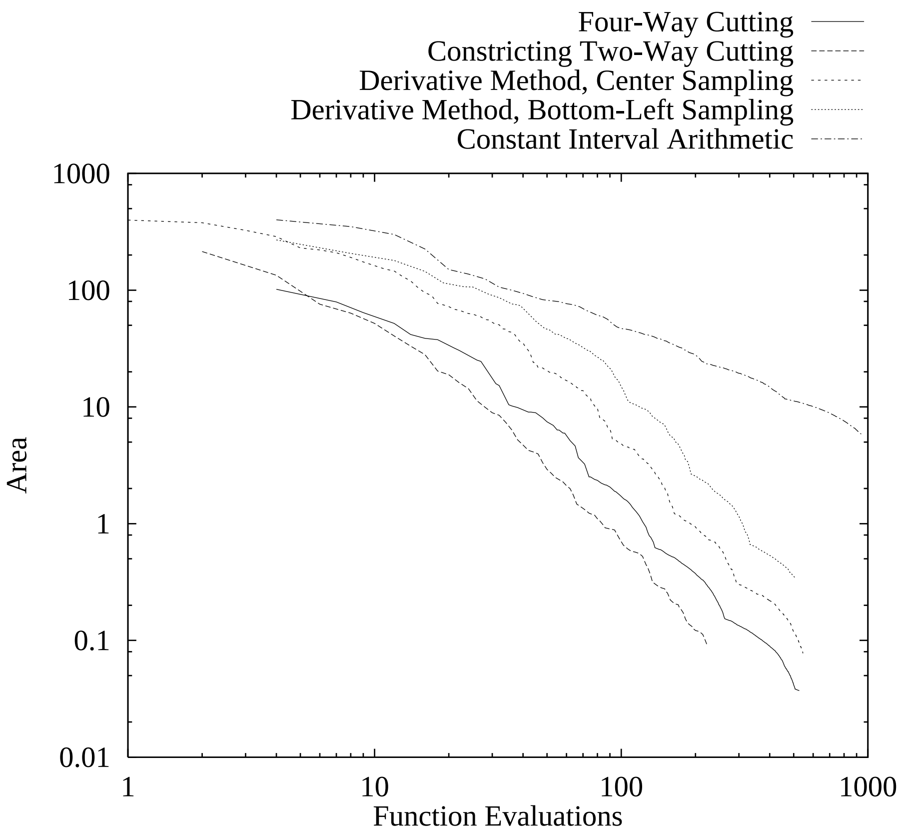
$$2(x-1)^2 + (y-1)^2 = 30,$$

over $[-10, 10] \times [-10, 10]$ using linear interval arithmetic, and two derivative-based methods. The linear interval arithmetic method uses four-way cutting, the simplest progressive method discussed. The two derivative-based methods both use the first derivative only, but differ in their placement of the sample $\boldsymbol{j}$. One places $\boldsymbol{j}$ at the center of the cluster; the other places $\boldsymbol{j}$ at the bottom left corner of the cluster. As each method uses a different number of interval evaluations per stage, the following diagram does not indicate the relative efficiencies of the different methods.



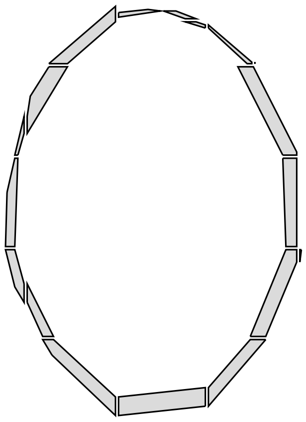
Four-Way Cutting | *Derivative-Based Method (Center Sampling)* | *Derivative-Based Method (Bottom-Left Sampling)*

Clearly, the linear interval arithmetic produces superior intervening renderings, as fewer spurious visual artifacts are present in the renderings produced using linear interval arithmetic. The following diagram illustrates the efficiency of the various methods when rendering the aforementioned equation:

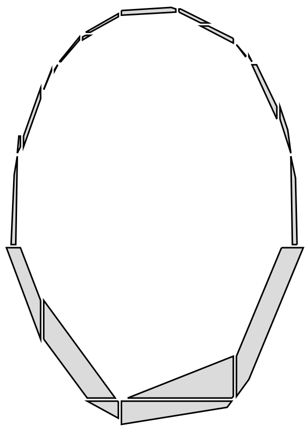


Of course, the derivative methods are not competitive when the underlying functions are not differentiable.

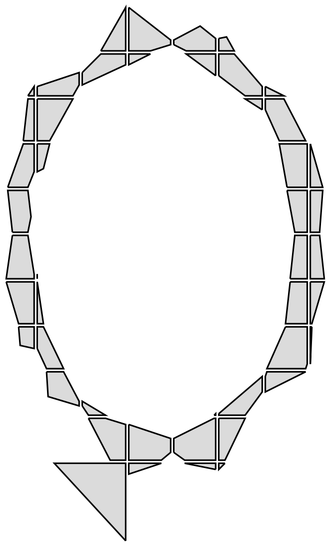
The following diagram illustrates the information gained, using each method, after 45 interval evaluations:



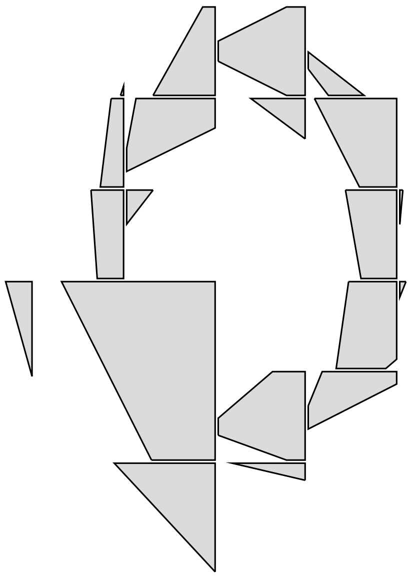
Four-Way Cutting
45 Interval Evaluations



Constricting Four-Way Cutting
44 Interval Evaluations



Derivative Method, Center Sampling
45 Interval Evaluations



Derivative Method, Bottom-Left Sampling
45 Interval Evaluations

An interval arithmetic similar to $\mathbb{M}$ may be implemented using derivative-based methods. $\mathbb{D}$ denotes such an arithmetic, which bounds both the value and first derivative of evaluated functions. Each element of $\mathbb{D}$ is given by a first component $v \in \mathbb{I}$, which bounds the value at $\alpha = 0$, and a second component $d \in \mathbb{I}$, which bounds the derivative for $\alpha \in [-1,1]$. Domains other than $[0,1]$ are possible; as are other sample locations. Further discussion of this approach can be found in the next subsection.

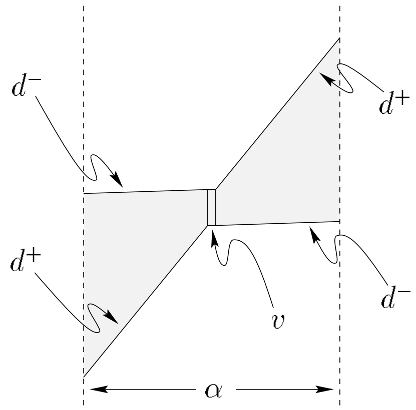
4.8.5 Hansen's Linear Interval Arithmetic

Linear interval arithmetic and Hansen's linear interval arithmetic share a common motivation. Hansen's linear interval arithmetic is, however, a closer relative of the derivative-based methods.

In $\mathbb{H}$, each interval is represented as a sampled value v, and a slope d:

$$\forall[(d,v) \in \mathbb{J}^2]\ \langle v + d\alpha\rangle \in \mathbb{H},\ \ \alpha \in [-c,c];$$

the bounds on α must also be stored. It seems reasonable to assume that c is fixed, as d may be adjusted to account for an arbitrary value of c. Intervals of $\mathbb{H}$ may be graphically depicted, as follows:



Interval $\langle v + d\alpha\rangle \in \mathbb{H}$

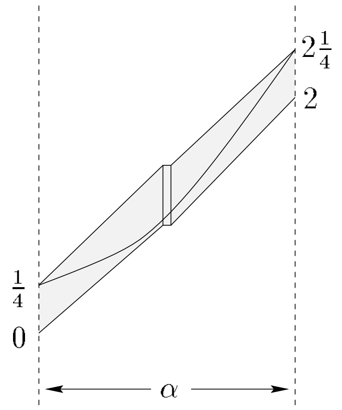
Bounds produced using Hansen's linear interval arithmetic are generally superior to bounds produced using a derivative-based method, as the relationship(s) between the free variable(s) and derivative(s) may be taken into account, as with linear interval arithmetic. An example evaluation is appropriate; let us bound the range of $g(x)$ for $x \in [0,1]$, by evaluating $g([0,1])$. Let $g(x) = (x+\frac{1}{2})^2$; with Hansen's linear interval arithmetic,

$$\begin{array}{ll} & g^{\mathbb{H}}([0,1]) \\ \rightsquigarrow & g^{\mathbb{H}}(\langle\langle\frac{1}{2},\frac{1}{2}\rangle + \langle 1,1\rangle\alpha\rangle), \quad \text{with } c = \frac{1}{2} \\ \rightsquigarrow & (\langle\langle\frac{1}{2},\frac{1}{2}\rangle + \langle 1,1\rangle\alpha\rangle + \langle\langle\frac{1}{2},\frac{1}{2}\rangle + \langle 0,0\rangle\alpha\rangle)^2 \\ \rightsquigarrow & (\langle\langle 1,1\rangle + \langle 1,1\rangle\alpha\rangle)^2 \\ \rightsquigarrow & \langle\langle 1,\frac{5}{4}\rangle + \langle 2,2\rangle\alpha\rangle; \end{array}$$

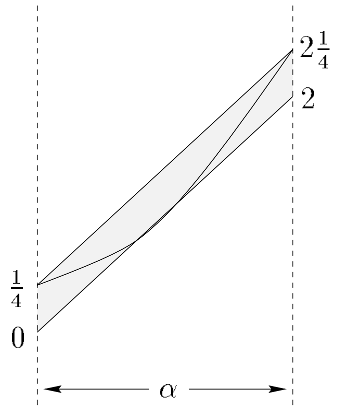
while with linear interval arithmetic,

$$\begin{array}{ll} & g^{\mathbb{M}}([0,1]) \\ \rightsquigarrow & g^{\mathbb{M}}(\langle\alpha,\alpha\rangle) \\ \rightsquigarrow & (\langle\alpha,\alpha\rangle + \langle\frac{1}{2},\frac{1}{2}\rangle)^2 \\ \rightsquigarrow & \langle\frac{1}{2}+\alpha,\frac{1}{2}+\alpha\rangle^2 \\ \rightsquigarrow & \langle 2\alpha,\frac{1}{4}+2\alpha\rangle. \end{array}$$

Evaluation rules for Hansen's linear interval arithmetic are given in [28], although sufficient rules are easily determined. A squaring operator is needed for the above result with $\mathbb{H}$; using a general multiplication operator produces a sub-optimal bound. The two bounds are seen to be identical, as the following diagrams illustrate:



$g^{\mathbb{H}}([0,1])$



$g^{\mathbb{M}}([0,1])$

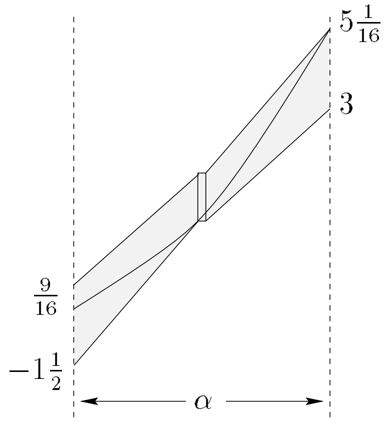
Although the diagram gives v a slight width, in actuality the two bounds are identical. We may extend the previous evaluations by bounding the range of $h(x) = (g(x))^2$ for $x \in [0,1]$. With Hansen's linear interval arithmetic,

$$\begin{array}{cc} & h^{\mathbb{H}}([0,1]) \\ \rightsquigarrow & (g^{\mathbb{H}}([0,1]))^2 \\ \rightsquigarrow & (\langle\langle 1, \frac{5}{4}\rangle + \langle 2,2\rangle\alpha\rangle)^2 \\ \rightsquigarrow & \langle\langle 1, 2\frac{9}{16}\rangle + \langle 4,5\rangle\alpha\rangle; \end{array}$$

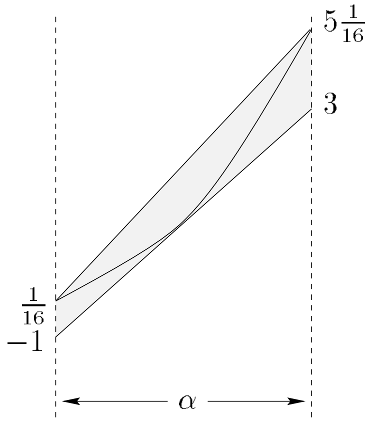
with linear interval arithmetic,

$$\begin{array}{cc} & h^{\mathbb{M}}([0,1]) \\ \rightsquigarrow & (g^{\mathbb{M}}([0,1]))^2 \\ \rightsquigarrow & (\langle 2\alpha, \frac{1}{4} + 2\alpha\rangle)^2 \\ \rightsquigarrow & \langle -1 + 4\alpha, \frac{1}{16} + 5\alpha\rangle. \end{array}$$

Linear interval arithmetic has produced a superior bound; illustrations of the two bounds follow:



$h^{\mathbb{H}}([0,1])$

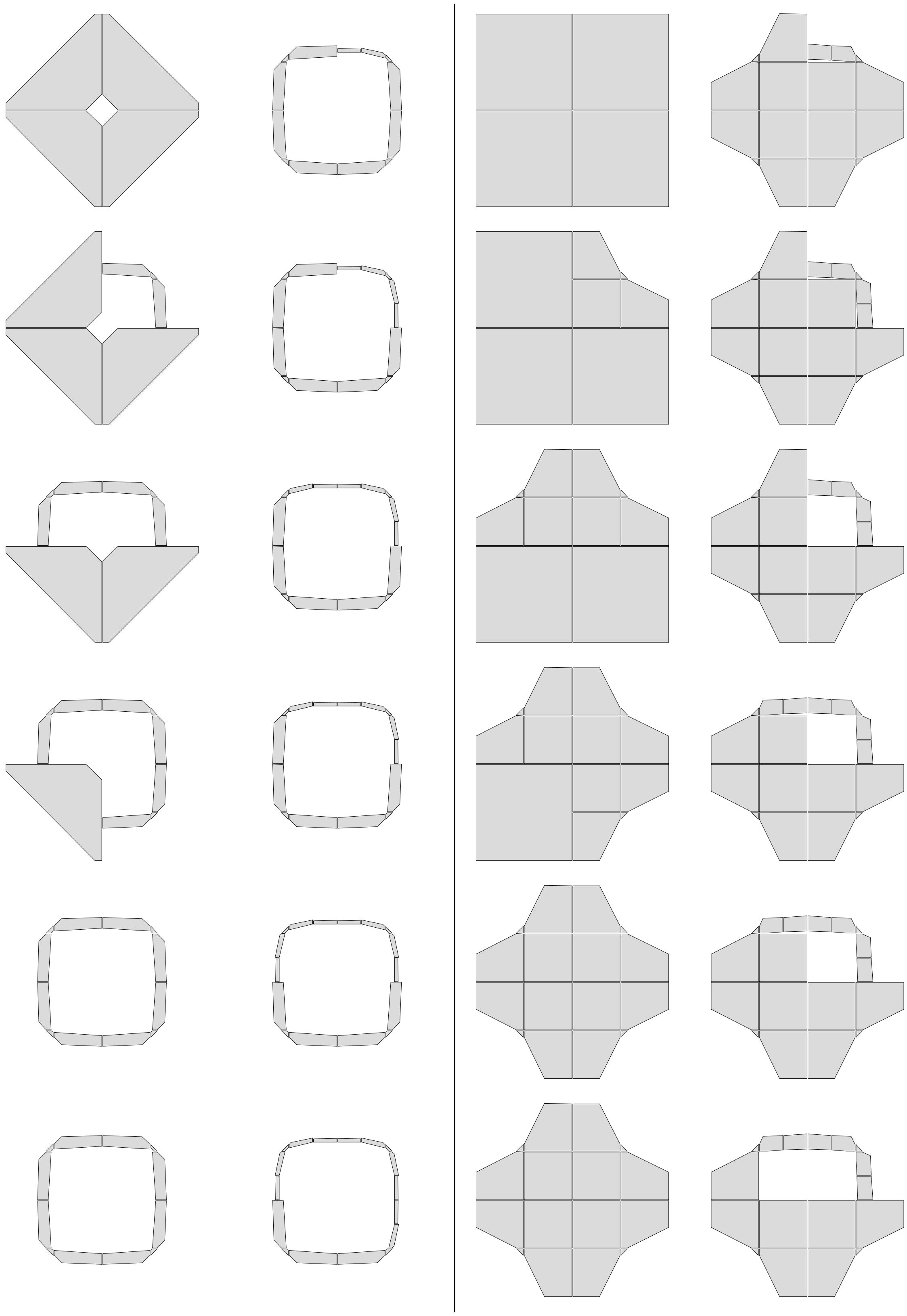


$h^{\mathbb{M}}([0,1])$

The following diagram depicts renderings of

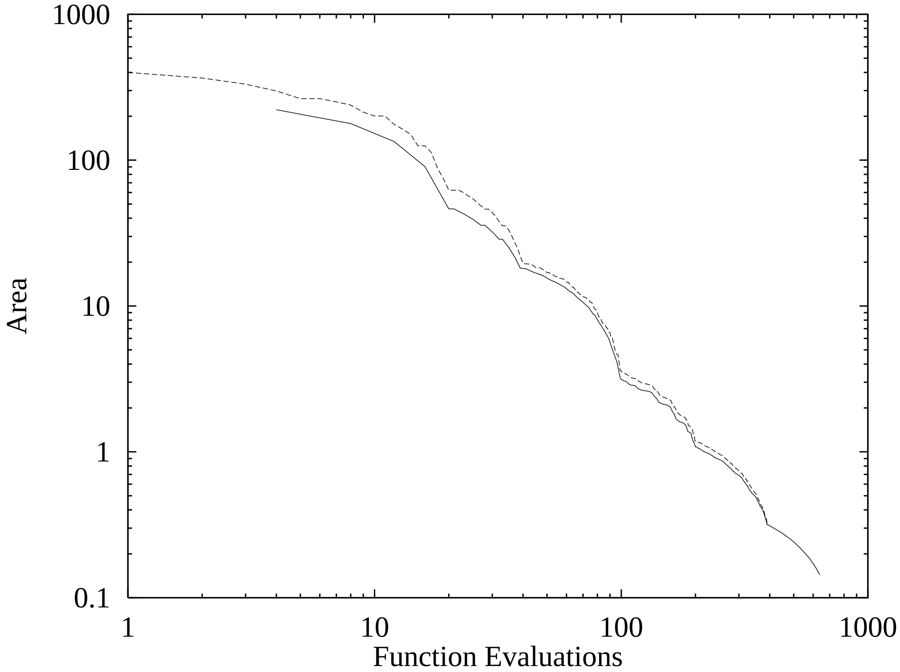
$$x^4 + y^4 = 1600,$$

over $[-10, 10] \times [-10, 10]$ using Hansen's two-dimensional linear interval arithmetic and two-dimensional linear interval arithmetic. As before, the two methods do not perform a similar amount of work per stage; the diagram does not indicate the relative efficiencies of the two methods.

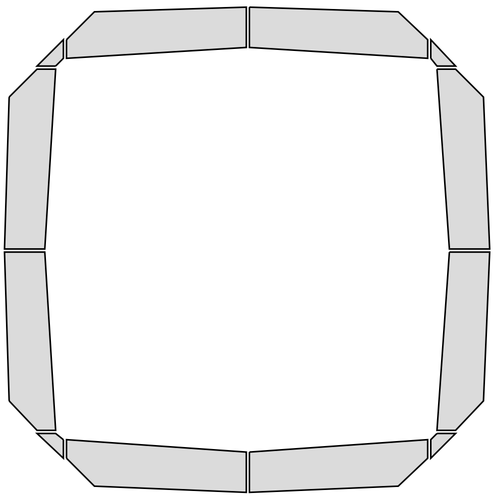


Four-Way Cutting Hansen's Four-Way Cutting

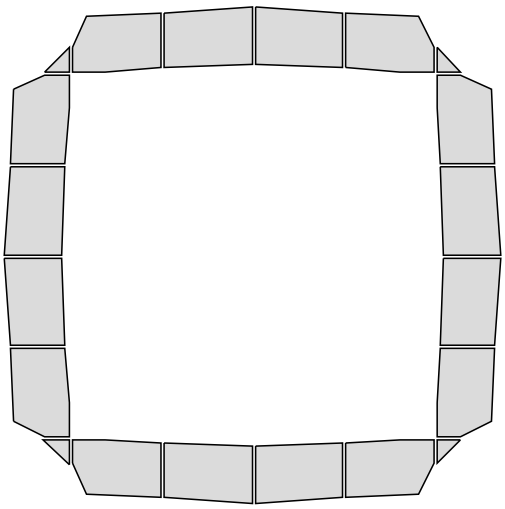
The following diagram depicts the efficiency of the methods when rendering the aforementioned equation:



The following diagrams illustrate the information gained, using each method, after 20 interval evaluations:

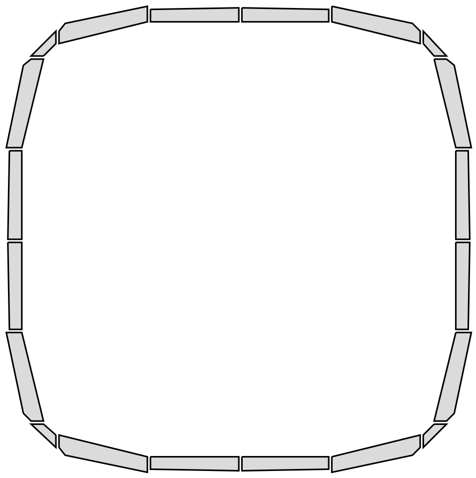


Four-Way Cutting with $\mathbb{L}_2$
20 Interval Evaluations

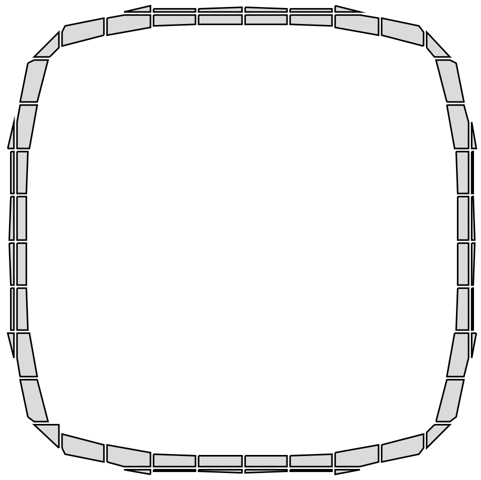


Four-Way Cutting with $\mathbb{H}_2$
20 Interval Evaluations

and after 40 interval evaluations:



Four-Way Cutting with $\mathbb{L}_2$
40 Interval Evaluations



Four-Way Cutting with $\mathbb{H}_2$
40 Interval Evaluations

Hansen's linear interval arithmetic $\mathbb{H}$ and our linear interval arithmetic $\mathbb{L}$ perform a similar amount of work computing bounds for any given operator. Consider the operator $g(x) = x^2$, which has the following evaluation rule:

$$g^{\mathbb{H}}(\langle a + b\alpha \rangle) \rightsquigarrow \langle (a^2 + \langle 0, c^2 \rangle b^2) + (2ab)\alpha \rangle.$$

Evaluation may be simplified by assuming $c = 1$, so α varies from -1 to 1. With that assumption, evaluation proceeds by the following rule:

$$g^{\mathbb{H}}(\langle a + b\alpha \rangle) \rightsquigarrow \langle (a^2 + \langle 0, 1 \rangle b^2) + (2ab)\alpha \rangle.$$

Evaluation efficiency may be improved considerably by expanding the interval operations, as was done with the evaluation of our interval operators. The evaluation breaks into a number of cases, depending on the relationship between a and zero, and b and zero. A portion of the evaluation rule, for our example g, follows:

$$g^{\mathbb{H}}(\langle a + b\alpha \rangle) \rightsquigarrow \begin{cases} \langle \langle {a^-}^2, {a^+}^2 + {b^+}^2 \rangle + \langle 2a^-b^-, 2a^+b^+ \rangle \alpha \rangle & \text{if } (a^- \geq 0) \ \wedge \ (b^- \geq 0), \\ \vdots & \vdots \\ \langle \langle {a^-}^2, {a^+}^2 + {b^+}^2 \rangle + & \text{if } (0 \in a) \ \wedge \ (a^+ \geq -a^-) \ \wedge \\ \quad \langle 2\min(a^-b^+, a^+b^-), 2\min(a^-b^-, a^+b^+) \rangle \alpha \rangle & \quad (0 \in b) \ \wedge \ (b^+ \geq -b^-). \end{cases}$$

The above portion is taken from an evaluation rule with 16 cases; a greater number of cases could have been used, in the aim of reducing the likelihood of computing a large number of floating-point operations. The evaluation rule for g using linear interval arithmetic follows:

$$g^{\mathbb{M}}(\langle a + b\alpha, c + d\alpha \rangle) \rightsquigarrow \begin{cases} \langle (m^2 - bm) + 2bm\alpha, c^2 + ((c+d)^2 - c^2)\alpha \rangle & \text{if } (A \geq 0), \\ \langle (n^2 - dn) + 2dn\alpha, a^2 + ((a+b)^2 - a^2)\alpha \rangle & \text{if } (C \leq 0), \\ \langle 0, A^2 + (C^2 - A^2)\alpha \rangle & \text{otherwise.} \end{cases}$$

where $m = a + \frac{1}{2}b$, $n = c + \frac{1}{2}d$, $A = \min(a, a + b)$, and $C = \max(c, c + d)$. With Hansen's method, most evaluations employ seven multiplications and one addition; with our method, most evaluations employ six multiplications and six additions. Our evaluation rule falls into fewer cases, and removal of conditionals occurs earlier. Changing the domain of α influences the efficiency of our method as well.

It is not completely clear why Hansen chose to evaluate $g(\langle a + b\alpha\rangle)$ using

$$g^{\mathbb{H}}(\langle a + b\alpha\rangle) \rightsquigarrow \langle (a^2 + \langle 0, c^2\rangle b^2) + (2ab)\alpha\rangle,$$

instead of

$$g^{\mathbb{H}}(\langle a + b\alpha\rangle) \rightsquigarrow \langle a^2 + b(2a + \langle -c, c\rangle)\alpha\rangle,$$

as

$$\langle a + b\alpha\rangle^2 = a^2 + 2ab\alpha + b^2\alpha^2.$$

Many evaluation rules are possible: what is needed is a methodology for choosing rules. With Hansen's intervals, there is a choice between minimizing the width of the resulting v or minimizing the width of the resulting d, where $g^{\mathbb{H}}(\langle a + b\alpha\rangle) \rightsquigarrow \langle v + d\alpha\rangle$. Both methods are computationally identical for linear operators. Hansen's methods outlined in [28] require more floating-point operations for division and multiplication.

With either approach, fewer floating-point operations may be used to compute bounds, if looser bounds are acceptable. Derivative methods are easily implemented and produce slightly larger bounds at a slightly higher evaluation cost. The efficiency graphs illustrate that each approach differs, in computational efficiency, by a constant factor.

With a modern language, such as C++, Hansen's intervals may be built using an underlying interval arithmetic class. Our methods may be modularly constructed by using an underlying linear bound class, which observes the current rounding mode.

It is unreasonable to expect a common optimizing compiler to produce code comparable to a direct implementation, as symbolic reasoning is employed when producing an efficient implementation. The author advocates the implementation of a program which employs sophisticated symbolic reasoning to automatically produce “direct” implementations, for any of a variety of interval arithmetics. Such a program would be given a description of an operator's properties, and produce a code fragment which implements the corresponding interval operator. Such routines may be folded into an optimizing compiler, but it is unclear what algorithms would benefit, other than interval arithmetic classes.

The chief advantage of our approach to generalized interval arithmetic is its mathematical simplicity, which allows for properties to be naturally tracked. This simplicity also provides for a superior handling of discontinuous functions, and multi-functions.

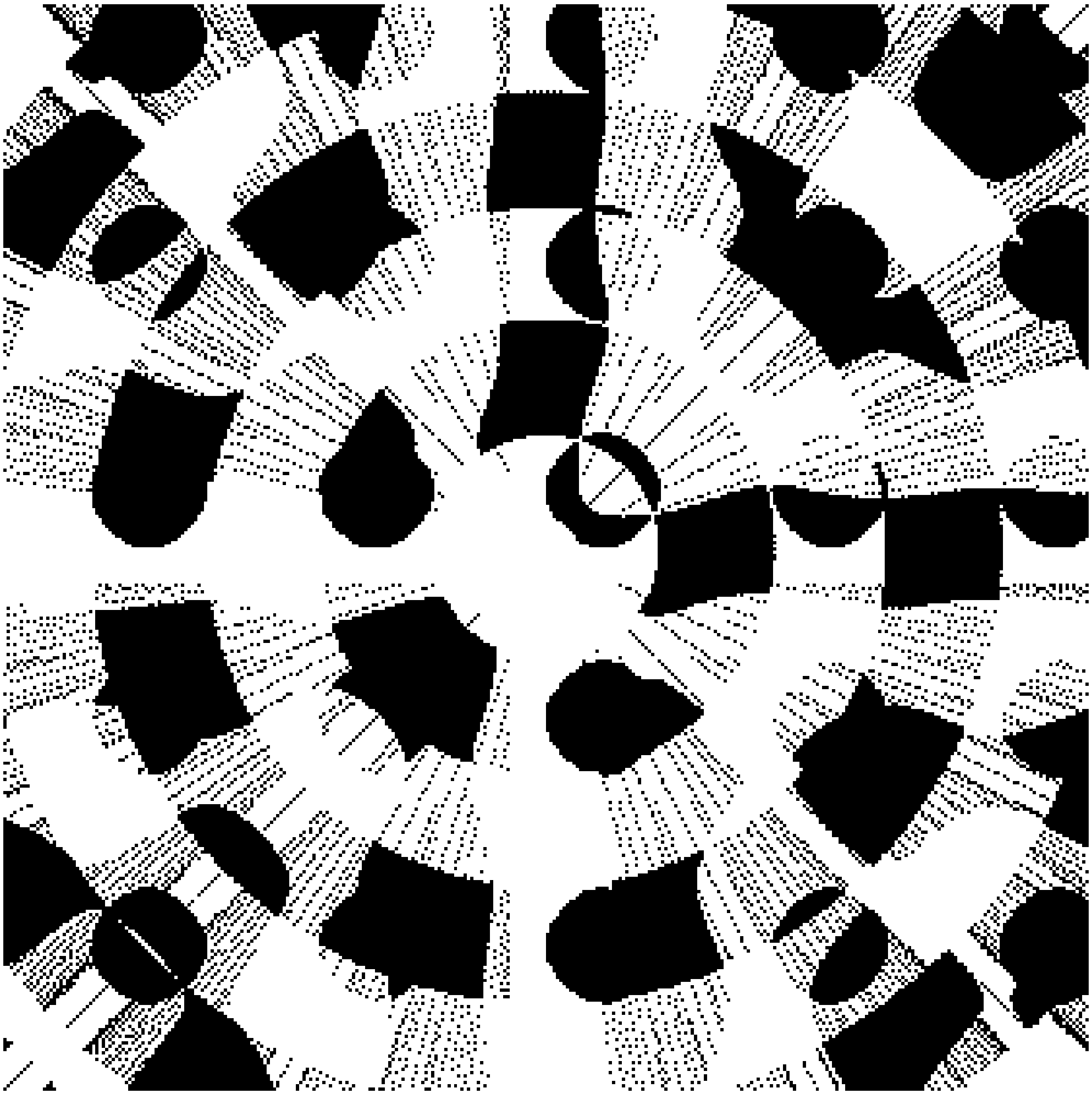
The chief advantage of Hansen's approach is that $\mathbb{H}$ is easily implemented, given an implementation of $\mathbb{I}$. Of course, $\mathbb{D}$ is implemented with even less work. With a naïve implementation, both $\mathbb{D}$ and $\mathbb{H}$ return sub-optimal bounds. Regardless, as j shrinks, the relative differences between $g^{\mathbb{D}}(j)$, $g^{\mathbb{H}}(j)$, and $g^{\mathbb{I}}(j)$ approach zero. With effort, better bounds may be returned, although this mitigates the cheif advantage of the two methods.

Our generalized interval arithmetic may be implemented using Hansen's methods, or using derivative-based methods. Temporary recourse to the methods outlined in this thesis is possible when considering non-differentiable operators.

Finally, it should be noted that with several minor changes to Hansen's fundamental definition of intervals, we may produce our fundamental definition of intervals. With $\mathbb{H}$, and $\mathbb{M}$, this proceeds as follows: for an interval $\langle a + b\alpha \rangle \in \mathbb{H}$, let α vary from 0 to 1, rather than from $-c$ to c; let b be a member of $\mathbb{J}^{\between}$, rather than a member of $\mathbb{J}$.

4.9 Example Renderings

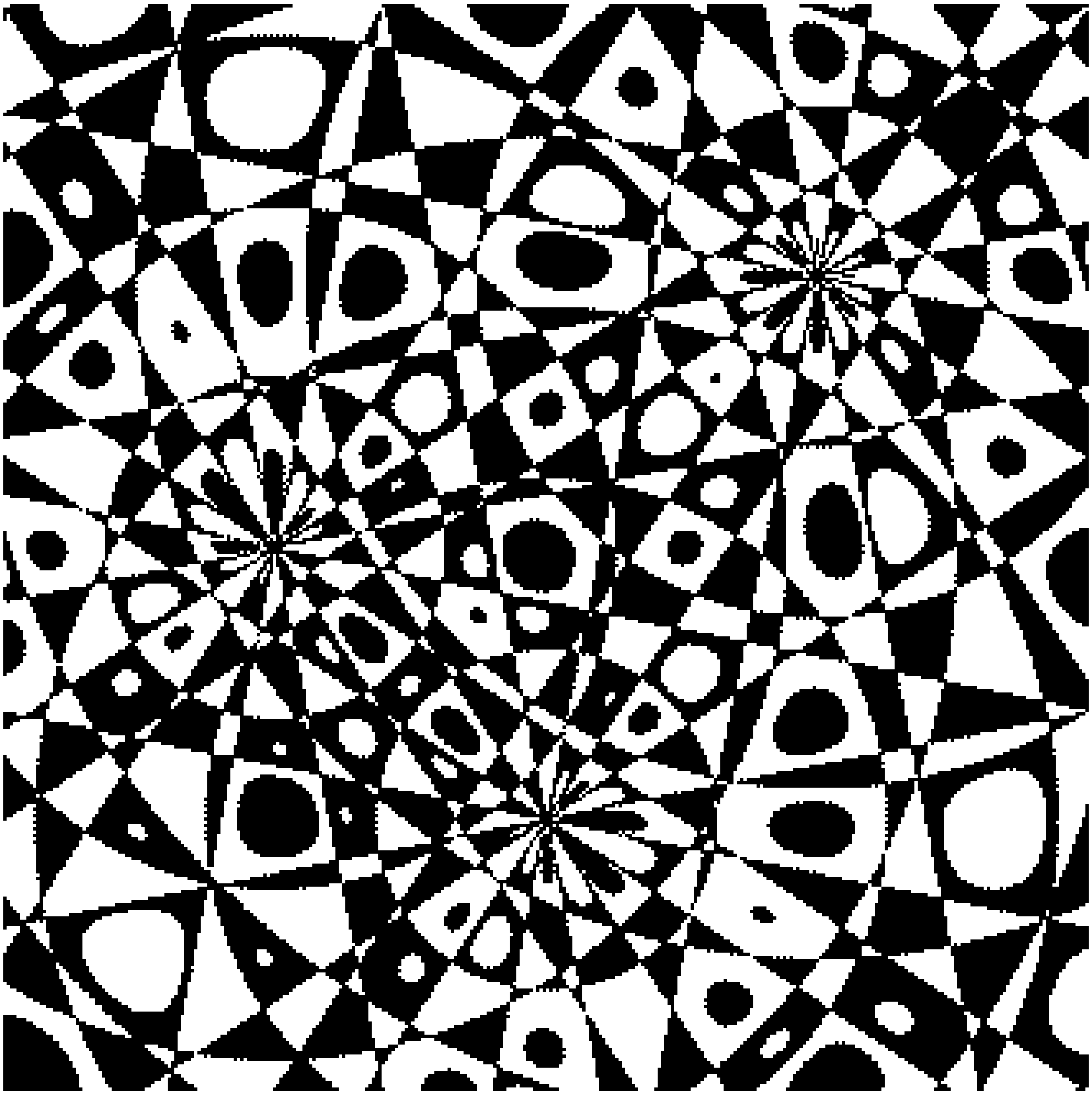
In this section, a few example renderings are illustrated. The presented renderings were produced using the methods described in this thesis.



The above rendering is of

$$\max_{\lfloor 2+\sin\sqrt{x^2+y^2}\rfloor}(\gcd(x,y), \sin x + \sin y, x\cos y + y\cos x) > 1,$$

at a resolution of 2048×2048. In the above specification, $\gcd(x, y)$ returns the greatest common real-valued divisor of x and y; $\max_k$ returns the value of the kth largest argument.



The above rendering is of

$$f(x+5,y)f(x-5,y-5)f(x,y+5) \in \{[\tfrac{-1}{10},0],[\tfrac{1}{5},\infty]\},$$

with

$$f(x,y) = \left(\sin\sqrt{x^2+y^2}\right)\left(\cos 8\mathrm{Arctan}\frac{y}{x}\right),$$

at a resolution of 3072×3072.

The above rendering is of

$$\tfrac{1}{4}\left[2\sin(x\sin y + y\sin x) + f(x-y) + f(x+y)\right] > \tfrac{1}{2}[\sin 160x + \sin 160y],$$

with

$$f(x) = \sqrt[3]{\sin 2.5\sqrt{2}x},$$

at a resolution of 3072×3072.

Chapter 5

Conclusion

5.1 Interval Techniques

Many distinct algorithms benefit from interval methods. Unfortunately, each algorithm typically re-implements code common to many such algorithms. A generalized interval arithmetic library rationalizes development by providing a framework which may contain code and concepts common to many algorithms. Fewer assumptions need be made by algorithms, as properties of the underlying functions may be tracked.

Essentially, generalized interval arithmetic bundles the lower layers of sophisticated interval arithmetic algorithms into a unified library. With this unification, sophisticated optimization within the common library is possible.

5.2 Graphing

A simple algorithm which graphs implicit equations has been presented. The algorithm progressively renders a graph: vast stretches are initially carved out; as the algorithm proceeds, intricate details of the graph are revealed. At all times, the rendering presents completely reliable information. This contrasts strongly with traditional sampling techniques, which evade the general problem and produce renderings which have no formal connection to the underlying graph. Computers bring unheralded speed and precision to mathematical tasks, such as graphing; this thesis demonstrates that such technology may produce accurate results, contrary to common practice.

5.3 Future Work

Generalized interval techniques may be explored further. Knowledge of theoretical results concerning the efficiency of generalized interval evaluation would be reassuring; complete knowledge is not possible, as the exact determination of the value of a function is, in general, not computable.

In practice, a system that mechanically produces generalized interval arithmetic libraries would be appreciated. With such a system, human effort is applied to a more abstract system, allowing for the deployment of a variety of efficient interval arithmetic libraries with a reduced likelihood of implementation error. Another implementation approach is, for example, to implement $\mathcal{I}_{p+q\alpha}(\mathbb{L})$ in place of $\mathbb{U}$; a thorough evaluation of this approach may be informative.

The interval arithmetic presented may be generalized further. Other data types and operations may be considered, as in [52]; probabilistic arithmetic is another possibility. Integrating automatic differentiation [61] into this framework is another possible pursuit.

An ongoing challenge is to integrate generalized interval arithmetic into a wide variety of applications. As with graphing, the free variables provided by a generalized interval arithmetic should be gainfully exploited by the benefitting application.

Graphing may be explored further. Algorithms for accurately rendering differential equations, integral equations, and iterated function systems may be explored. Higher dimensions are intriguing; the implementation of an interval-based renderer which models the interaction of light within a scene is a tempting challenge. Such algorithms would return reliable bounds on the colour assigned to a pixel; a quantized colour system would provide a natural stopping criteria.

Bibliography

[1] N. N. Abdelmalek. The discrete one-sided chebyshev approximation. *Inst. Maths. Applics.*, 18:361–370, 1976.

[2] Harold Abelson and Gerald Jay Sussman with Julie Sussman. *Structure and Interpretation of Computer Programs.* MIT Press, Cambridge, Massachusetts, 1985.

[3] Alfred V. Aho, Ravi Sethi, and Jeffrey D. Ullman. *Compilers: Principles, Techniques, and Tools.* Addison-Wesley Series in Computer Science. Addison-Wesley, Menlo Park, California, 1986.

[4] Götz Alefeld and Jürgen Herzberger. *Introduction to Interval Computations.* Academic Press, New York, 1983.

[5] American National Standards Institute / Institute of Electrical and Electronics Engineers, New York. *IEEE Standard for Binary Floating-Point Arithmetic*, 1985. ANSI/IEEE Standard 754-1985.

[6] American National Standards Institute / Institute of Electrical and Electronics Engineers, New York. *IEEE Standard for Radix-Independent Floating-Point Arithmetic*, 1985. ANSI/IEEE Standard 854-1987.

[7] Herbert Arkin and Raymond R. Colton. *Graphs: How to Make and Use Them.* Harper & Brothers Publishers, New York and London, 1936.

[8] B. Artmann. *The Concept of Number: from quaternions to monads and topological fields.* Ellis Horwood Limited, Market Cross House, Cooper Street, Chichester, West Sussex, PO19 1EB, England, 1988.

[9] Algirdas Avizienis. Signed-digit number representations for fast parallel arithmetic. *IEEE Transactions on Electronic Computers*, EC-10:389–400, 1961.

[10] F. A. Behrend. A contribution to the theory of magnitudes and the foundations of analysis. *Math. Zeitschrift*, 63:345–362, 1956.

[11] I. S. Berezin and N. P. Zhidkov. *Computing Methods*, volume I, chapter 2, pages 72–82. Addison-Wesley, 1965. Translated by O. M. Blunn.

[12] E. Bishop and D. Bridges. *Constructive Real Analysis.* Springer-Verlag, Berlin, 1985.

[13] G. Bohlender. *Computer Arithmetic and Self-Validating Numerical Methods*, volume 7 of *Notes and Reports in Mathematics in Science and Engineering*, chapter What Do We Need Beyond IEEE Arithmetic? Academic Press, New York, 1990.

[14] R. Bojanic and R. DeVore. On polynomials of best one sided approximation. *Enseignement Math.*, 12:139–164, 1966.

[15] Claude Brezinski. *History of Continued Fractions and Padé Approximants.* Number 12 in Springer Series in Computational Mathematics. Springer-Verlag, 1991.

[16] Nigel J. Cutland. No longer ghosts — the renaissance of infinitesimals. *Mathematical Perspectives: Four Recent Inaugural Lectures*, pages 43–74, 1990.

[17] Keith J. Devlin. *Constructibility.* Springer-Verlag, Berlin, 1984.

[18] Ronald DeVore. One-sided approximation of functions. *Journal of Approximation Theory*, 1:11–25, 1968.

[19] J. K. S. Dewar. Procedures for interval arithmetic. *Computing Journal*, 14:447–450, 1970.

[20] Herbert B. Enderton. *A Mathematical Introduction to Logic.* Academic Press, 24-28 Oval Road, London, NW1 7DX, 1972.

[21] Miloš D. Ercegovac and Thomas Lang. On-line arithmetic: A design methodology and applications. *VLSI Signal Processing III*, pages 252–263, 1988.

[22] C. T. Fike. *Computer Evaluation of Mathematical Functions.* Prentice-Hall Series in Automatic Computation. Prentice-Hall, Englewood Cliffs, New Jersey, 1968.

[23] Eugene L. Fiume. *The Mathematical Structure of Raster Graphics.* Academic Press, 1250 Sixth Avenue, San Diego, CA 92101, 1989.

[24] James Foley, Andries van Dam, Steven Feiner, and John Hughes. *Computer Graphics: Principles and Practice.* Addison-Wesley, Reading, Massachusetts, 2 edition, 1990.

[25] Dr. G. Frege. *The Foundations of Arithmetic.* Northwestern University Press, Evanston, Illinois, 1950. English Translation by J. L. Austin.

[26] Casper Goffman and George Pedrick. *Real Functions.* Prindle, Weber, and Schmidt, Boston, 1967.

[27] D. I. Good and R. L. London. Computer interval arithmetic: Definition and proof of correct implementation. *Journal of the Association for Computing Machinery*, 17:603–612, 1970.

[28] E. R. Hansen. A generalized interval arithmetic. In *Interval Mathematics: Proceedings of the International Symposium*, volume 29 of *Lecture Notes In Computer Science*, pages 7–18, Berlin, May 1975. Springer-Verlag.

[29] J. G. Hayes, editor. *Numerical Approximations to Functions and Data.* The Athlone Press, University of London, 2 Gower Street London, 1970. Based on a conference organized by the Institute of Mathematics and Its Applications, Canterbury, England, 1967.

[30] V. H. Hristov and K. G. Ivanov. Characterization of best approximations from below and above. *Colloquia Mathematica Societis János Bolyai*, 58:377–403, 1990.

[31] Mary Jane Irwin and Robert Michael Owens. Fully digit on-line networks. *IEEE Transactions on Computers*, C-32(4):402–406, April 1983.

[32] W. Kahan. A more complete interval arithmetic. Technical report, University of Toronto, 1968. Report.

[33] I.L. Kantor and A.S. Solodovnikov. *Hypercomplex Numbers*. Springer-Verlag, New York, 1989.

[34] E. Kaucher. Interval analysis in the extended interval space $\mathbb{IR}$. In *Fundamentals of Numerical Computation*, number 2 in Computing Supplementum, pages 33–49. Springer-Verlag, Wien, 1980.

[35] Donald E. Knuth. *The Art of Computer Programming: Seminumerical Algorithms*, volume 2 of *Computer Science and Information Processing*. Addison-Wesley, Reading, Massachusetts, 1969.

[36] Peter Kornerup and David W. Matula. Finite precision lexicographic continued fraction number systems. In *Proceedings of Seventh Symposium on Computer Arithmetic*, IEEE Symposium on Computer Arithmetic, pages 207–214, 1985.

[37] U. Kulisch. An axiomatic approach to rounded computation. *Numer. Math.*, 18:1–17, 1971.

[38] U. Kulisch. Implementation and formalization of floating-point arithmetics. In *Caratheodary Symposium*, Athen, 1973.

[39] U. Kulisch. On the concept of a screen. *Z. Angew. Math. Mech.*, 53:115–119, 1973.

[40] Ulrich Kulisch. *A New Approach to Scientific Computation*, chapter A New Arithmetic for Scientific Computation, pages 1–26. Academic Press, New York, 1983.

[41] K. Kuratowski and A. Mostowski. *Set Theory*. North-Holland, Amsterdam, 1968.

[42] B. A. Kushner. *Lectures on Constructive Mathematical Analysis*, volume 60. American Mathematical Society, Providence, Rhode Island, 1984.

[43] Burkhard Lenze. On constructive one-sided spline approximation. *Approximation Theory*, 6(2):383–386, 1989.

[44] James T. Lewis. Computation of best one-sided l_1 approximation. *Mathematics of Computation*, 24(111):529–536, July 1970.

[45] James T. Lewis. Approximation with convex constraints. *SIAM Review*, 15(1):193–217, January 1973.

[46] Lisa Lorentzen and Haakon Waadeland. *Continued Fractions with Applications*. Number 3 in Studies in Computational Mathematics. North-Holland, Amsterdam, 1992.

[47] L. A. Lyusternik, O. A. Chervonenkis, and A. R. Yanpol'skii. *Handbook for Computing Elementary Functions*, volume 76 of *International Series of Monographs in Pure and Applied Mathematics*. Pergamon Press, London, 1965.

[48] Jerrold E. Marsden and Anthony J. Tromba. *Vector Calculus*. W. H. Freeman and Company, New York, 3 edition, 1976.

[49] Shouichi Matsui and Masao Iri. An overflow/underflow-free floating-point representation of numbers. *Journal of Information Processing*, 4(3):123–133, 1981.

[50] David W. Matula. Towards an abstract mathematical theory of floating-point arithmetic. In *1969 Spring Joint Computer Conference*, AFIPS Proceedings, pages 765–772, Montvale, New Jersey, 1969. AFIPS Press.

[51] David W. Matula and Peter Kornerup. Finite precision rational arithmetic: Slash number systems. *IEEE Transactions on Computers*, C-34(1):3–18, January 1985.

[52] W. L. Miranker. *A New Approach to Scientific Computation*, chapter Ultra-Arithmetic: The Digital Computer in Function Space, pages 165–198. Academic Press, New York, 1983.

[53] W. L. Miranker and U. Kulisch. Computer arithmetic in theory and practice. Technical report, IBM Thomas K. Watson Research Center, Yorktown Heights, 1979. RC 7776 (33658), July 24, Mathematics.

[54] Ieke Moerdijk and Gonzalo E. Reyes. *Models for Smooth Infinitesimal Analysis*. Springer-Verlag, New York, 1991.

[55] R. E. Moore. *Elements of Scientific Computing*. Holt, New York, 1975.

[56] Ramon E. Moore. *Interval Analysis*. Prentice-Hall Series in Automatic Computation. Prentice-Hall, Englewood Cliffs, New Jersey, 1966.

[57] Ramon E. Moore. *Methods and Applications of Interval Analysis*. SIAM Studies in Applied Mathematics. SIAM, Philadelphia, 1979.

[58] George Pedrick. *A First Course in Analysis*. Springer-Verlag, New York, 1994.

[59] Michael D. Potter. *Sets, An Introduction*. Oxford University Press, New York, 1990.

[60] Marian Boykan Pour-El and Ian Richards. Computability and noncomputability in classical analysis. *Transactions of the American Mathematical Society*, 275(2):539–560, February 1983.

[61] L. B. Rall. *Computer Arithmetic and Self-Validating Numerical Methods*, chapter Differentiation Arithmetics. Notes and Reports in Mathematics in Science and Engineering. Academic Press, New York, 1990.

[62] H. G. Rice. Recursive real numbers. *Proceedings of the American Mathematical Society*, 5(5):784–791, 1954.

[63] Alain Robert. *Nonstandard Analysis*. John Wiley & Sons, New York, 1988.

[64] Walter Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, New York, 1953.

[65] Samuel Selby and Leonard Sweet. *Sets—Relations—Functions*. McGraw-Hill Book Company, New York, 1963.

[66] John M. Snyder. Interval analysis for computer graphics. *Computer Graphics*, 26(2):121–129, July 1992.

[67] Richard L. Tieszen. *Mathematical Intuition*. Kluwer Academic Publishers, P.O. Box 17, 3300 AA Dordrecht, The Netherlands, 1989.

[68] Kishor S. Trivedi and Miloš D. Ercegovac. On-line algorithms for division and multiplication. *IEEE Transactions on Computers*, C-26(7):681–687, July 1977.

[69] Jean E. Vuillemin. Exact real computer arithmetic with continued fractions. *IEEE Transactions on Computers*, 39(8):1087–1105, August 1990.

[70] Wolfgang Warth. Approximation with constraints in normed linear spaces. *Journal of Approximation Theory*, 21:303–312, 1977.

[71] J. H. Wilkinson. *Rounding Errors in Algebraic Processes*. Her Majesty's Stationary Office, London, 1968.

[72] A. Young and E. A. Kiountouzis. Best approximation in an asymmetrically weighted l_1 measure. *J. Inst. Maths. Applics.*, 24:379–394, 1979.