Interval Methods for Bounding the Range of Polynomials and Solving Systems of Nonlinear Equations

Dissertation

zur Erlangung des akademischen Grades
"Doktor der technischen Wissenschaften"

Eingereicht von

Dipl.-Inf. Volker Stahl

September 1995

Erster Begutachter: o.Univ.-Prof. Dr. Dr. Bruno Buchberger
Zweiter Begutachter: o.Univ.-Prof. Dr. Jens Volkert

Angefertigt am Forschungsinstitut für Symbolisches Rechnen
Technisch-Naturwissenschaftliche Fakultät
Johannes Kepler Universität Linz

Eidesstattliche Erklärung

Ich versichere, daß ich die Dissertation selbständig verfaßt habe, andere als die angegebenen Quellen und Hilfsmittel nicht verwendet und mich auch sonst keiner unerlaubten Hilfe bedient habe.

Volker Stahl
Hagenberg, 15. September 1995

Acknowledgments

Heartfelt thanks go to my scientific adviser, Dr. Hoon Hong, for making me familiar with constraint solving and interval methods, for his patient support and his indefatigable interest in my thesis and last but not least for his friendship. He invested much more time and effort in me than one could expect from any adviser — also for raising research projects for my financial support. Thank you very much Hoon, I will never forget what you did for me.

I am deeply indebted to Prof. Dr. Dr. Bruno Buchberger for his selfless dedication to the Research Institute for Symbolic Computation, where this thesis was written. Thank you also for the excellent introduction to scientific working in the "Thinking, Speaking, Writing" courses.

I am grateful to Prof. Dr. Jens Volkert for serving on my thesis committee.

Several members of RISC contributed to this thesis in one way or another. In particular, I want to mention my colleagues in the STURM group — thanks for our close and fruitful teamwork and for the inspirations which I got from our discussions.

I want to thank my parents Siegfried and Irmgard Stahl who enabled me to study and do research in whatever area I decided. Together with my brother Axel they encouraged and motivated me in moments of frustration and gave me a home where I could take refuge any time.

A fellowship of the Deutsch–Österreichische Akademische Austauschdienst provided financial support during two years in Austria. I am grateful to Dr. Volker Strehl and to Dr. Jochen Pfalzgraf for their assistance in obtaining that support.

This research was done within the framework of the ACCLAIM project sponsored by the European Community Basic Research Action (ESPRIT 7195) and the Austrian Science Foundation (P9374-PHY).

Abstract

This thesis describes systematically known and new results on bounding the range of polynomials and solving systems of nonlinear equations using interval methods.

Range of Polynomials. Bounding the range of polynomials is an important sub–problem in many disciplines of scientific computing. The main original results of this thesis are as follows:

- *Centered Form.* We give a new, more elementary, proof of the known quadratic convergence of centered forms.
- *Horner Form.* For every polynomial f there exists an interval O_f such that for every interval X, whose interior is disjoint from O_f, the Horner evaluation of f on X is exact. The smallest such O_f is the hull of all roots of the intermediate polynomials arising during the Horner evaluation of f. During evaluation of f on X this non–overestimation condition can be decided without additional computation. We failed to find a necessary and sufficient non–overestimation condition but found some further sufficient conditions.

  If the input interval of the Horner form contains zero, then we bisect it at zero and evaluate both parts separately. As both parts have one endpoint zero, the total cost for both evaluations is comparable to the cost of the ordinary Horner form but gives usually tighter inclusions. If the dense Horner form evaluated on X overestimates, then any bisection of X gives an improvement. If X is centered, then bisection at the midpoint reduces the overestimation error of the dense Horner form at least by half. This observation is used to reduce the overestimation error of the dense Taylor form at least by half.
- *Mean Value Form.* The width of the mean value form is the same for any choice of a center between the optimal centers of the bicentered mean value form. This width is smaller than if the center is chosen differently. Hence, the midpoint is a width optimal center for the mean value form.
- *Bernstein Form.* For the Bernstein form we show that it is inclusion monotone and that it gives the exact range provided the input interval is "small enough". If the input interval contains zero, we bisect it at zero and evaluate the parts separately. This gives tighter inclusions and allows faster evaluation because each sub–interval has one endpoint zero and hence the Taylor coefficient computation is trivial.
- *Interpolation Form.* We present some accuracy improvements of interpolation forms by combining them with the concept of slopes and bicentered forms.
- *Nested Form.* For multivariate polynomials we define the nested form, which is motivated by eliminating common subexpressions and by exploiting the subdistributivity law. It is roughly as accurate as the Horner form but less expensive.
- *Experimental Comparison.* Finally, we compare various univariate and multivariate range computation methods experimentally in the context of Newton's method and a global optimization algorithm. Therefore, we give an implementation of extended interval arithmetic on top of the IEEE standard 754 and prove its correctness.

Solution of Systems of Equations. The second part of the thesis is devoted to solving systems of nonlinear equations. Our contribution is as follows:

- *Acceleration.* In order to accelerate the known algorithms we introduce a new operation called tightening. The idea is to evaluate a multivariate equation in all variables except for one on the given box and to solve the obtained univariate interval equation. Tightening can be applied directly to the given equations or to linearizations of them. When applied to linearizations, tightening is equivalent to the non–preconditioned Hansen–Sengupta operator with a different strategy for choosing equations and variables. We give a uniform and geometric framework for linearized tightening and the Hansen–Sengupta operator. Then we give a new condition for non–existence of solutions for linearized tightening. According to experimental results tightening usually leads to significant speedups.
- *Termination.* Certain interval methods fail to terminate if the search space is decomposed in such a way that a solution lies on the boundary of some sub–box. We solve this problem by slightly enlarging the given box $\boldsymbol{X}$ obtaining $\tilde{\boldsymbol{X}}$, testing whether the Hansen–Sengupta operator converges starting from $\tilde{\boldsymbol{X}}$ and if so, iterating until we obtain a box $\boldsymbol{Y}$ which is either in the interior of $\tilde{\boldsymbol{X}}$ or disjoint from $\boldsymbol{X}$. We prove correctness and termination of this algorithm under the assumption that the system has only finitely many simple solutions and the floating point number accuracy is high enough.
- *Application to Robotics.* We apply the techniques developed above to the robot inverse kinematics problem. Exploiting the inherent structure of this problem results in a significant efficiency improvement. Experimental results of a parallel implementation on a workstation network and on a super computer using PVM are reported.

Zusammenfassung

Inhalt dieser Arbeit ist eine systematische Darstellung bekannter und neuer Ergebnisse über die Eingrenzung des Wertebereichs von Polynomen und über die Lösung nichtlinearer Gleichungssysteme mit Intervallmethoden.

Wertebereich von Polynomen. Die Eingrenzung des Wertebereichs von Polynomen ist ein elementares Teilproblem in vielen Disziplinen wissenschaftlichen Rechnens. Die wichtigsten neuen Ergebnisse dieser Arbeit sind wie folgt:

- *Zentrische Form.* Wir geben einen neuen, elementareren Beweis der bekannten quadratischen Konvergenz zentrischer Formen.
- *Horner Form.* Für jedes Polynom f existiert ein Intervall O_f so daß für jedes X, dessen Inneres O_f nicht schneidet, das Horner Schema von f auf X exakt ist. Das kleinste solche O_f ist die Hülle der Nullstellen der Teilpolynome, die während der Hornerauswertung von f auftreten. Während der Auswertung von f auf X kann diese Nicht–Überabschätzungsbedingung ohne Zusatzaufwand getestet werden. Es ist uns nicht gelungen, eine notwendige und hinreichende Bedingung für Nicht–Überabschätzung zu formulieren, stattdessen fanden wir einige andere hinreichende Bedingungen.
  Falls das Argument Intervall der Horner Form Null enthält, spalten wir es bei Null und werten beide Teile separat aus. Da beide Teile einen Endpunkt Null haben, sind die Gesamtkosten für beide Auswertungen vergleichbar mit denen des gewöhnlichen Horner Schemas, aber die Eingrenzung ist im allgemeinen genauer. Falls die dichte Horner Form ausgewertet auf X eine Überabschätzung liefert, dann wird durch zerteilen von X immer eine Verbesserung erzielt. Falls X zentriert ist, dann wird durch zerteilen am Mittelpunkt der Überabschätzungsfehler der dichten Horner Form mindestens um die Hälfte reduziert. Diese Beobachtung wird ausgenutzt um den Überabschätzungsfehler der dichten Taylor Form um mindestens die Hälfte zu reduzieren.
- *Mittelwertsform.* Die Weite der Mittelwertsform ist die selbe für jede Wahl des Zentrums zwischen den optimalen Zentren der bizentrischen Mittelwertsform. Diese Weite ist kleiner als wenn ein anderes Zentrum gewählt wird. Daher ist der Mittelpunkt ein optimales Zentrum der Mittelwertsform.
- *Bernstein Form.* Für die Bernstein Form zeigen wir Inklusionsmonotonie und daß für "hinreichend kleine" Intervalle keine Überabschätzung eintritt. Wenn das Argument Intervall Null enthält, spalten wir es bei Null und werten die Teile separat aus. Dies führt zu engeren Eingrenzungen und erlaubt schnellere Auswertung, da jedes Teilintervall einen Endpunkt Null hat und daher die Berechnung der Taylor Koeffizienten trivial ist.
- *Interpolationsform.* Wir geben einige Verbesserungen der Genauigkeit von Interpolationsformen, indem wir sie mit dem Konzept von Steigungen und bizentrischen Formen kombinieren.
- *Geschachtelte Form.* Für multivariate Polynome definieren wir die geschachtelte Form, die durch die Eliminierung gemeinsamer Teilausdrücke und die Ausnutzung des Subdistributivitätsgesetzes motiviert ist. Sie ist etwa gleich genau wie die Horner Form, aber schneller.
- *Experimenteller Vergleich.* Wir vergleichen verschiedene univariate und multivariate Wertebereichsmethoden experimentell im Kontext des Newton Verfahrens und eines globalen Optimierungsalgorithmus. Dazu implementieren wir generalisierte Intervallarithmetik basierend auf dem IEEE Standard 754 und beweisen die Korrektheit.

Lösung von Gleichungssystemen. Der zweite Teil der Arbeit ist der Lösung nichtlinearer Gleichungssysteme gewidmet. Unser Beitrag ist wie folgt:

- *Effizienz.* Um die bekannten Algorithmen zu beschleunigen definieren wir eine neue Operation genannt Einengung. Die Idee ist, eine multivariate Gleichung in allen Variablen außer einer auf dem gegebenen Intervall auszuwerten und die so erhaltene Intervallgleichung zu lösen. Einengung kann entweder direkt auf die gegebenen Gleichungen angewandt werden oder auf Linearisierungen. Angewandt auf Linearisierungen ist Einengung gleichbedeutend mit dem unkonditionierten Hansen–Sengupta Operator mit einer anderen Strategie für die Auswahl von Gleichungen und Variablen. Wir geben eine einheitliche und geometrische Darstellung von linearisierter Einengung und Hansen–Sengupta Operator und zeigen ein neues Kriterium für die Nichtexistenz von Lösungen für linearisierte Einengung. Entsprechend unserer experimentellen Ergebnisse führt Einengung im allgemeinen zu deutlichen Beschleunigungen.
- *Terminierung.* Gewisse Intervallmethoden terminieren nicht wenn der Suchraum so zerteilt wird, daß eine Lösung am Rand eines Teilintervalls liegt. Wir lösen dieses Problem indem wir eine geringe Vergrößerung $\hat{\boldsymbol{X}}$ des gegebenen Intervalls $\boldsymbol{X}$ bestimmen, testen ob der Hansen–Sengupta Operator

ausgehend von $\tilde{\boldsymbol{X}}$ konvergiert und in diesem Fall so lang iterieren bis wir ein Interval $\boldsymbol{Y}$ erhalten, welches entweder im Innern von $\tilde{\boldsymbol{X}}$ liegt oder disjunkt von $\boldsymbol{X}$ ist. Wir beweisen Korrektheit und Terminierung dieses Algorithmus unter der Annahme, daß das Gleichungssystem nur endlich viele einfache Lösungen hat und die Gleitkommagenauigkeit ausreicht.

- *Anwendung auf Robotik.* Wir wenden die oben beschriebenen Techniken auf das inverse Roboter Kinematikproblem an. Durch Ausnutzung der inherenten Struktur dieses Problems wird eine deutliche Effizienzverbesserung erzielt. Experimentelle Ergebnisse einer Implementierung auf einem Workstation Netzwerk und einem Superrechner unter Verwendung von PVM werden berichtet.

Contents

Overview

This thesis describes systematically known and new results on bounding the range of polynomials and solving systems of nonlinear equations using interval methods. The known results are presented as described by the table of contents whereas the new results are scattered throughout the thesis. Thus, in this overview we emphasize the original contributions.

Chapter 1 contains an introduction to the main concepts of interval arithmetic. In particular we give a general definition of centered forms (Definition 1.3.25, page 14) and a new, more elementary, proof of their quadratic convergence (Theorem 1.3.27, page 15).

An implementation of extended interval arithmetic on top of the IEEE standard 754 for binary floating point numbers and the proof of its correctness is subject of **Chapter 2**.

In **Chapter 3** we discuss methods for bounding the range of univariate polynomials. The main original results are as follows:

- *Horner Form.* For every polynomial f there exists an interval O_f such that for every interval X which is disjoint from the interior of O_f, the Horner evaluation of f on X is exact (Theorem 3.1.12, page 43). The smallest such O_f is the hull of all roots of the intermediate polynomials arising during the Horner evaluation of f. Theorem 3.1.15, page 46). During evaluation of f on X this non–overestimation condition can be decided without additional computation. (Theorem 3.1.22, page 50). Further sufficient conditions for non–overestimation of the Horner form are given by Theorem 3.1.26, Corollary 3.1.30, Corollary 3.1.31 and Theorem 3.1.32.
  If the input interval of the Horner form does not contain zero then the evaluation can be accelerated by replacing some interval power computations by number power computations (Algorithm 3.1.33, page 58). This observation is also used to speed up the separate computation of upper or lower bound of the Horner form (Algorithm 3.1.35, page 59).
  If the input interval of the Horner form contains zero, then we bisect it at zero and evaluate both halves separately. As both halves have one endpoint zero, the total cost for the evaluations is comparable to the cost of the ordinary Horner form but gives usually tighter inclusions (Algorithm 3.1.38, page 61). If the dense Horner form evaluated on X overestimates, then any bisection of X gives an improvement (Theorem 3.1.43, page 64). If X is centered, then bisection at the midpoint reduces the overestimation error of the dense Horner form at least by half (Theorem 3.1.45, page 65). This observation is used to reduce the overestimation error of the dense Taylor form at least by half (Algorithm 3.3.12, page 88).
- *Mean Value Form.* Any choice of a center between the optimal centers of the bicentered mean value form results in the same width of the mean value form, which is smaller than if the center is chosen differently (Theorem 3.2.22, page 78). Hence, the midpoint is a width optimal center for the mean value form.
- *Bernstein Form.* For the Bernstein form we show that it is inclusion monotone (Theorem 3.4.17, page 98) and that it gives the exact range provided the input interval is "small enough" (Theorem 3.4.19, page 104). If the input interval contains zero, we bisect it at zero. This gives tighter inclusions and allows faster evaluation because each sub–interval has one endpoint zero and hence the Taylor coefficient computation is trivial (Algorithm 3.4.25, page 110).
- *Interpolation Form.* We present some accuracy improvements of interpolation forms by combining them with the concept of bicentered forms (Section 3.5.1) and slopes (Section 3.5.2).
- *Experimental Comparison.* Finally, we compare various range computation methods experimentally at some classes of random polynomials, in the context of Newton's method, and in the context of a global optimization algorithm.

Chapter 4 generalizes some results of Chapter 3 to the multivariate case. For efficiency reasons we restrict our considerations to methods which preserve sparsity of the given polynomial. In particular, we introduce the nested form, (Definition 4.1.11, page 140) which is motivated by eliminating common subexpressions and by exploiting the subdistributivity law. It is roughly as accurate as the Horner form but less expensive. Again, we compare multivariate range computation methods at some classes of random polynomials, in the context of a root finding algorithm, and in the context of a global optimization algorithm.

Chapter 5 and Chapter 6 are concerned with the solution of systems of nonlinear equations. In **Chapter 5** we review methods for testing existence, uniqueness and non–existence of solutions in a box and give elementary geometric proofs of the main theorems. We introduce a modification of the Hansen–Sengupta operator called linearized tightening and give a new non–existence property (Theorem 5.3.2, page 189). While linearized tightening is not as powerful as the Hansen–Sengupta operator for testing uniqueness and existence of solutions (Section 5.4.5), experimental results indicate that it still leads in many cases to an efficiency improvement (Table 5.11.1, page 238).
Certain algorithms for solving systems of nonlinear equations fail to terminate if the search space is decomposed in such a way that a solution lies on the boundary of some sub–box. We solve this problem by slightly enlarging the given box $\boldsymbol{X}$ obtaining $\tilde{\boldsymbol{X}}$, testing whether the Hansen–Sengupta operator converges starting from $\tilde{\boldsymbol{X}}$ and if so, iterating until we obtain a box $\boldsymbol{Y}$ which is either in the interior of $\tilde{\boldsymbol{X}}$ or disjoint from $\boldsymbol{X}$. (Algorithm 5.9.4, page 230). We prove correctness and termination of this algorithm under the assumption that the system has only finitely many simple solutions and the floating point number accuracy is high enough (Section 5.9). Finally, we review a method for finding all solutions of a system of polynomial equations when the search space is unbounded (Section 5.10).

In **Chapter 6** we introduce a new method called tightening for pruning the search space. Various experiments show that applying tightening in a Hansen–Sengupta operator based solver gives usually significant efficiency improvements (Table 6.5.1, page 245).

Finally, in **Chapter 7** we apply the methods developed above to the robot inverse kinematics problem. The computation is accelerated by exploiting the inherent structure of this problem (Table 7.4, page 253). Experimental results of a parallel implementation on a workstation network (Table 7.5.1, page 255) and on a super computer (Table 7.5.4, page 259) are reported.

Symbol Index

$M_f^{*(\mathrm{s})}$	Dense slope form of f	71
$\ddot{M}_f$	Univariate / multivariate bicentered mean value form of f	76 / 154
$M_f^{(\mathrm{H})}$	Mean value – Horner form of f	146
$M_f^{\widetilde{\rightarrow}}$	Successive mean value form	148
$M_f^{(\mathrm{s})\widetilde{\rightarrow}}$	Successive slope form	151
$\ddot{M}_f^{(\mathrm{H})}$	Mean value – Horner form of f	156
$\ddot{M}_f^{\widetilde{\rightarrow}}$	Successive bicentered mean value form of f	157
$\mathrm{mag}(X)$	Magnitude, $\mathrm{mag}(X) = \max\{\lvert x\rvert \mid x \in X\}$	6
$\mathrm{mig}(X)$	Mignitude, $\mathrm{mig}(X) = \min\{\lvert x\rvert \mid x \in X\}$	6
N_f	Nested form of f	140
O_f	Overestimation interval of f	43
$p_j^{(k)}$	The j-th Bernstein polynomial of order k	89
$p_j^{(k,X)}$	The j-th Bernstein polynomial of order k over X	92
p^*	Path, $p^* = \langle p_1, \ldots, p_m\rangle$, $p_\ell = (G_{i_\ell}, j_\ell)$	195
$\mathcal{P}$	Powerset	36
ψ	Interpretation function for floating point intervals, $\psi : \mathbb{IF}_\emptyset \to \overline{\mathbb{IR}}_\emptyset$	26
$\mathrm{q}(X, Y)$	Distance, $\mathrm{q}(X, Y) = \max\{\lvert\overline{X} - \overline{Y}\rvert, \lvert\underline{X} - \underline{Y}\rvert\}$	6
$\hat{\mathrm{q}}(\boldsymbol{X}, \boldsymbol{Y})$	Distance, $\hat{\mathrm{q}}(\boldsymbol{X}, \boldsymbol{Y}) = \max_i \mathrm{q}(X_i, Y_i)$	6
Q_f	Interval such that $Q_f \not\supseteq O_f$	46
$\tilde{\rho}, \hat{\rho}, \check{\rho}$	Round to nearest, round up, round down function, $\tilde{\rho}, \hat{\rho}, \check{\rho}\colon \mathbb{R} \to \mathbb{F}$	21
$\mathrm{rad}(X)$	Radius, $\mathrm{rad}(X) = 1/2(\overline{X} - \underline{X})$	6
$\mathbb{R}$	Set of real numbers	5
σ	Rounding function for intervals, $\sigma : \overline{\mathbb{IR}}_\emptyset \to \mathbb{IF}_\emptyset$	26
$\mathrm{smig}(X)$	Signed mignitude	15
T_f	Taylor form of f	83
T_f^*	Dense Taylor form of f	83
$\check{T}_f^*$	Dense Taylor form of f with bisection	87
tight	Linear tightening, $\mathrm{tight} : \mathbb{IR}^n \times (\mathbb{R}^n \to \mathbb{IR}) \times \{1, \ldots, n\} \to \mathbb{IR}_\emptyset^n$	193
$\mathrm{w}(X)$	Width, $\mathrm{w}(X) = \overline{X} - \underline{X}$	6
$\hat{\mathrm{w}}(\boldsymbol{X})$	Maximal width, $\hat{\mathrm{w}}(\boldsymbol{X}) = \max_i \mathrm{w}(X_i)$	6
$\mathrm{w}^+(\boldsymbol{X})$	Sum of widths, $\mathrm{w}^+(\boldsymbol{X}) = \sum_i \mathrm{w}(X_i)$	6
$\mathrm{w}^*(\boldsymbol{X})$	Volume, $\mathrm{w}^*(\boldsymbol{X}) = \prod_i \mathrm{w}(X_i)$	174
X	Interval	5
$\boldsymbol{X}$	Interval vector	5
$\underline{X}, \overline{X}$	Lower respectively upper endpoint of the interval X	5
$\lvert X\rvert$	Absolute value, $\lvert X\rvert = \{\lvert x\rvert \mid x \in X\}$	6
$\boldsymbol{X}^{\overline{(i)}}$	Upper i-face, $\boldsymbol{X}^{\overline{(i)}} = (X_1, \ldots, X_{i-1}, \overline{X}_i, X_{i+1}, \ldots, X_n)$	182
$\boldsymbol{X}^{\underline{(i)}}$	Lower i-face, $\boldsymbol{X}^{\underline{(i)}} = (X_1, \ldots, X_{i-1}, \overline{X}_i, X_{i+1}, \ldots, X_n)$	182
$\boldsymbol{X}^{\overline{\underline{(i)}}}$	i-face, $\boldsymbol{X}^{\overline{\underline{(i)}}} = \boldsymbol{X}^{\overline{(i)}} \cup \boldsymbol{X}^{\underline{(i)}}$	182
$[\ldots]$	Hull operator	5
$\top$	Floating point number for ∞	21
$\bot$	Floating point number for $-\infty$	21
$\natural$	Not a number	21
$\tilde{+}, \hat{+}, \check{+}$	Floating point addition rounded to nearest, up, down, $\tilde{+}, \hat{+}, \check{+}\colon \mathbb{F}_\natural \times \mathbb{F}_\natural \to \mathbb{F}_\natural$	23
$\tilde{-}, \hat{-}, \check{-}$	Floating point subtraction rounded to nearest, up, down, $\tilde{-}, \hat{-}, \check{-}\colon \mathbb{F}_\natural \times \mathbb{F}_\natural \to \mathbb{F}_\natural$	23
$\tilde{*}, \hat{*}, \check{*}$	Floating point multiplication rounded to nearest, up, down, $\tilde{*}, \hat{*}, \check{*}\colon \mathbb{F}_\natural \times \mathbb{F}_\natural \to \mathbb{F}_\natural$	23
$\tilde{/}, \hat{/}, \check{/}$	Floating point division rounded to nearest, up, down, $\tilde{/}, \hat{/}, \check{/}\colon \mathbb{F}_\natural \times \mathbb{F}_\natural \to \mathbb{F}_\natural$	23
$\hat{\cdot}, \check{\cdot}$	Modified floating point multiplication rounded up, down, $\hat{\cdot}, \check{\cdot}\colon \mathbb{F}_\natural \times \mathbb{F}_\natural \to \mathbb{F}_\natural$	27

Chapter 1

Interval Arithmetic

This chapter gives a brief introduction to the main concepts of interval arithmetic. For a more comprehensive survey we recommend the excellent monographs [Moore, 1966], [Moore, 1979], [Alefeld and Herzberger, 1983], [Ratschek and Rokne, 1984], [Ratschek and Rokne, 1988], [Neumaier, 1990], [Hansen, 1992] and [Hammer *et al.*, 1993].

In Section 1.1 we define interval arithmetic operations, and give some important algebraic properties. Further, we introduce notation which will be used throughout the thesis. A topology on the set of intervals is defined in Section 1.2. Having a topology at hand, important notions like continuity and convergence are defined. Some elementary properties of interval functions are subject of Section 1.3. Interval functions usually occur when the range of a real function over an interval has to be estimated. In this context we give a general definition of a centered form. Concrete instances of centered forms are presented in Chapter 3 and Chapter 4.

1.1 Foundations

By an interval we mean a set

$$[a, b] = \{x \mid a \leq x \leq b\}$$

where $a, b \in \mathbb{R}$, the set of real numbers. The set of intervals over $\mathbb{R}$ is denoted by $\mathbb{IR}$. Intervals are written in capital letters. The lower and upper endpoint of an interval X is denoted by $\underline{X}$ respectively $\overline{X}$.

In order to distinguish intervals from arbitrary sets, we use a calligraphic font for the latter. If $\mathcal{X}$ is a bounded set of real numbers, then we write $[\mathcal{X}]$ to denote the smallest (w.r.t. $\subseteq$) interval which contains all elements of $\mathcal{X}$. We call $[\mathcal{X}]$ the hull of $\mathcal{X}$. For ease of notation we define

$$[a, b] = [b, a]$$

for $a \geq b$.

By an n-dimensional interval vector we mean an ordered n-tuple of intervals $\boldsymbol{X} = (X_1, \ldots, X_n)$. The set of n-dimensional interval vectors is denoted by $\mathbb{IR}^n$. Tuples are written in bold face.

Every continuous function $f : \mathbb{R}^n \to \mathbb{R}$ can be extended to $\mathbb{IR}^n \to \mathbb{IR}$ by

$$f(\boldsymbol{X}) = \{f(\boldsymbol{x}) \mid \boldsymbol{x} \in \boldsymbol{X}\}.$$

This allows us to embed $\mathbb{R}$ into $\mathbb{IR}$ by the morphism

$$\mu(x) = [x, x]$$

which is defined for tuples component wise. Elements of $\mathbb{IR}$ which are not in $\mathbb{R}$ are called proper intervals, elements of $\mathbb{R}$ are called points.

For the extension of arithmetic functions to intervals it suffices to consider endpoints:

Theorem 1.1.1 (Interval Arithmetic) *For all $X, Y \in \mathbb{IR}$ it holds that*

$$\begin{aligned} X+Y &= [\underline{X}+\underline{Y}, \overline{X}+\overline{Y}] \\ X-Y &= [\underline{X}-\overline{Y}, \overline{X}-\underline{Y}] \\ XY &= [\{\underline{XY}, \underline{X}\overline{Y}, \overline{X}\underline{Y}, \overline{XY}\}] \\ X/Y &= \begin{cases} [\{\underline{XY}, \underline{X}\overline{Y}, \overline{X}\underline{Y}, \overline{XY}\}] & \text{if} \quad 0 \notin Y \\ \textit{undefined} & \text{else} \end{cases} \\ -X &= [-\overline{X}, -\underline{X}]. \ \square \end{aligned}$$

One easily checks the following properties of interval arithmetic operations.

Theorem 1.1.2 (Algebraic Properties) *For all $X, Y, Z \in \mathbb{IR}$ it holds that*

- *Associativity:* $(X+Y)+Z = X+(Y+Z)$, $(XY)Z = X(YZ)$
- *Commutativity:* $(X+Y) = (Y+X)$, $XY = YX$
- *Neutral Element:* $0+X = X$, $1X = X$. $\square$

However, proper intervals do not have additive or multiplicative inverses. Further, the distributivity law does not hold for intervals. Instead, we have the following weaker version.

Theorem 1.1.3 (Subdistributivity) *For all $X, Y, Z \in \mathbb{IR}$ it holds that*

$$\begin{aligned} X(Y+Z) &\subseteq XY + XZ \\ X(Y+Z) &= XY + YZ \text{ if } X \in \mathbb{R} \text{ or } \underline{Y}, \underline{Z} \geq 0 \text{ or } \overline{Y}, \overline{Z} \leq 0. \ \square \end{aligned}$$

Remark. As intervals do not have additive inverses, $\mathbb{IR}^n$ is not a vector space. Still, we call the elements of $\mathbb{IR}^n$ interval vectors. Sometimes we use the more intuitive term "box" or simply interval. $\square$

Some useful standard functions which are used frequently in this thesis are listed below:

$$\begin{array}{lrcl} \text{Midpoint:} & \text{mid}(X) & = & 1/2(\overline{X}+\underline{X}) \\ \text{Width:} & \text{w}(X) & = & \overline{X}-\underline{X} \\ \text{Radius:} & \text{rad}(X) & = & 1/2(\overline{X}-\underline{X}) \\ \text{Absolute Value:} & |X| & = & \{|x| \mid x \in X\} \\ \text{Mignitude:} & \text{mig}(X) & = & \min\{|x| \mid x \in X\} \\ \text{Magnitude:} & \text{mag}(X) & = & \max\{|x| \mid x \in X\} \\ \text{Interior:} & \text{int}(X) & = & X \backslash \{\underline{X}, \overline{X}\}. \end{array}$$

The functions extend to tuples component wise. For interval vectors we define

$$\begin{array}{lrcl} \text{Maximal Width:} & \hat{\text{w}}(\boldsymbol{X}) & = & \max_i \text{w}(X_i) \\ \text{Sum of Widths:} & \text{w}^+(\boldsymbol{X}) & = & \sum_i \text{w}(X_i). \end{array}$$

1.2 Topology on the Set of Intervals

In this section we define a metric topology on the set of intervals which induces important concepts like convergence, continuity, etc.

Definition 1.2.1 (Distance) *The distance between two intervals $X, Y \in \mathbb{IR}$ is defined as*

$$\text{q}(X,Y) = \max\{|\overline{X}-\overline{Y}|, |\underline{X}-\underline{Y}|\}.$$

For interval vectors $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{IR}^n$ we define

$$\hat{\text{q}}(\boldsymbol{X}, \boldsymbol{Y}) = \max_{i=1,\ldots,n} \text{q}(X_i, Y_i). \ \square$$

Theorem 1.2.2 *The distance function* $\hat{\mathrm{q}} : \mathbb{IR}^n \times \mathbb{IR}^n \to \mathbb{R}$ *is a metric on* $\mathbb{IR}^n$.

Proof. We have to show the following three properties:

- Positivity: For all $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{IR}^n$ it holds that

$$\hat{\mathrm{q}}(\boldsymbol{X}, \boldsymbol{Y}) \geq 0, \quad \text{and} \quad \hat{\mathrm{q}}(\boldsymbol{X}, \boldsymbol{Y}) = 0 \text{ iff } \boldsymbol{X} = \boldsymbol{Y}.$$

- Symmetry: For all $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{IR}^n$ it holds that

$$\hat{\mathrm{q}}(\boldsymbol{X}, \boldsymbol{Y}) = \hat{\mathrm{q}}(\boldsymbol{Y}, \boldsymbol{X}).$$

- Triangle Inequality: For all $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z} \in \mathbb{IR}^n$ it holds that

$$\hat{\mathrm{q}}(\boldsymbol{X}, \boldsymbol{Y}) \leq \hat{\mathrm{q}}(\boldsymbol{X}, \boldsymbol{Z}) + \hat{\mathrm{q}}(\boldsymbol{Y}, \boldsymbol{Z}).$$

Positivity and symmetry follow immediately from Definition 1.2.1, hence it remains to show the triangle inequality. Let $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z} \in \mathbb{IR}^n$ arbitrary but fixed, let $\hat{i}$ such that $\hat{\mathrm{q}}(\boldsymbol{X}, \boldsymbol{Y}) = \mathrm{q}(X_{\hat{i}}, Y_{\hat{i}})$ and let $X = X_{\hat{i}}, Y = Y_{\hat{i}}, Z = Z_{\hat{i}}$. As

$$\mathrm{q}(X, Z) + \mathrm{q}(Y, Z) \leq \hat{\mathrm{q}}(\boldsymbol{X}, \boldsymbol{Z}) + \hat{\mathrm{q}}(\boldsymbol{Y}, \boldsymbol{Z})$$

it suffices to show that

$$\mathrm{q}(X, Y) \leq \mathrm{q}(X, Z) + \mathrm{q}(Y, Z).$$

$$\begin{aligned} \mathrm{q}(X, Z) + \mathrm{q}(Y, Z) &= \max\{|\overline{X} - \overline{Z}|, |\underline{X} - \underline{Z}|\} + \max\{|\overline{Y} - \overline{Z}|, |\underline{Y} - \underline{Z}|\} \\ &\geq \max\{|\overline{X} - \overline{Z}| + |\overline{Y} - \overline{Z}|, |\underline{X} - \underline{Z}| + |\underline{Y} - \underline{Z}|\} \\ &\geq \max\{|\overline{X} - \overline{Y}|, |\underline{X} - \underline{Y}|\} \\ &= \mathrm{q}(X, Y). \ \square \end{aligned}$$

The metric $\hat{\mathrm{q}}$ induces a topology on $\mathbb{IR}^n$, which is given by its open sets in the following way:

Definition 1.2.3 (Open Subsets of $\mathbb{IR}^n$) *A set* $\mathcal{S} \subseteq \mathbb{IR}^n$ *is open in* $\mathbb{IR}^n$ *if for all* $\boldsymbol{X} \in \mathcal{S}$ *there exists a real number* $\varepsilon > 0$ *such that*

$$\{\boldsymbol{Y} \in \mathbb{IR}^n \mid \hat{\mathrm{q}}(\boldsymbol{X}, \boldsymbol{Y}) \leq \varepsilon\} \subseteq \mathcal{S}. \ \square$$

Theorem 1.2.4 *Let* $\mathcal{O}_n$ *be the set of open subsets of* $\mathbb{IR}^n$. *Then* $(\mathbb{IR}^n, \mathcal{O}_n)$ *is a topological space.*

Proof. See [Armstrong, 1983]. $\square$

In the sequel, when we write $\mathbb{IR}^n$, we always mean the topological space $(\mathbb{IR}^n, \mathcal{O}_n)$.

Definition 1.2.5 (Convergent Sequence) *A sequence* $\langle \boldsymbol{X}^{(k)} \in \mathbb{IR}^n, k = 1, 2, \ldots \rangle$ *is called convergent if there exists* $\boldsymbol{X} \in \mathbb{IR}^n$ *such that*

$$\lim_{k \to \infty} \hat{\mathrm{q}}(\boldsymbol{X}^{(k)}, \boldsymbol{X}) = 0.$$

In this case $\boldsymbol{X}$ *is called a limit of the sequence* $\langle \boldsymbol{X}^{(k)} \rangle$.

Theorem 1.2.6 *A sequence* $\langle \boldsymbol{X}^{(k)} \in \mathbb{IR}^n, k = 1, 2, \ldots \rangle$ *converges to* $\boldsymbol{X} \in \mathbb{IR}^n$ *iff*

$$\lim_{k \to \infty} \underline{X_i^{(k)}} = \underline{X_i} \quad \text{and} \quad \lim_{k \to \infty} \overline{X_i^{(k)}} = \overline{X_i}$$

for all $i = 1, \ldots, n$. $\square$

Proof.

"$\Rightarrow$" Assume $\langle \boldsymbol{X}^{(k)}, k = 1, 2, \ldots \rangle$ converges to $\boldsymbol{X}$ and let $i \in \{1, \ldots, n\}$ and $\varepsilon > 0$ arbitrary but fixed. Then there exists $k_\varepsilon \in \mathbb{N}$ such that

$$\mathrm{q}(X_i^{(k)}, X_i) < \varepsilon \quad \text{for all} \quad k \geq k_\varepsilon.$$

It follows that

$$|\overline{X_i^{(k)}} - \overline{X_i}| < \varepsilon, \quad |\underline{X_i^{(k)}} - \underline{X_i}| < \varepsilon$$

for all $k \geq k_\varepsilon$, thus

$$\lim_{k \to \infty} \overline{X_i^{(k)}} = \overline{X_i}, \quad \lim_{k \to \infty} \underline{X_i^{(k)}} = \underline{X_i}.$$

"$\Leftarrow$" Assume $\langle \overline{X_i^{(k)}} \rangle$ converges to $\overline{X_i}$ and $\langle \underline{X_i^{(k)}} \rangle$ converges to $\underline{X_i}$ for $i = 1, \ldots, n$ and let $\varepsilon > 0$ arbitrary but fixed. Then there exists $k_\varepsilon \in \mathbb{N}$ such that

$$|\overline{X_i^{(k)}} - \overline{X_i}| < \varepsilon, \quad |\underline{X_i^{(k)}} - \underline{X_i}| < \varepsilon$$

for all i and all $k \geq k_\varepsilon$. It follows that

$$\hat{\mathrm{q}}(\boldsymbol{X}^{(k)}, \boldsymbol{X}) < \varepsilon \quad \text{for all} \quad k \geq k_\varepsilon,$$

thus

$$\lim_{k \to \infty} \boldsymbol{X}^{(k)} = \boldsymbol{X}. \ \square$$

Definition 1.2.7 (Nested Sequence) *A sequence* $\langle \boldsymbol{X}^{(k)} \in \mathbb{IR}^n, k = 1, 2, \ldots \rangle$ *is called nested if*

$$\boldsymbol{X}^{(k+1)} \subseteq \boldsymbol{X}^{(k)} \quad \text{for all} \quad k \geq 1. \ \square$$

Theorem 1.2.8 *A nested sequence* $\langle \boldsymbol{X}^{(k)} \in \mathbb{IR}^n, k = 1, 2, \ldots \rangle$ *is convergent.* $\square$

Proof. Let $\langle \boldsymbol{X}^{(k)} \rangle$ be a nested sequence. According to Theorem 1.2.6 it suffices to show that

$$\lim_{k \to \infty} \overline{X_i^{(k)}} = \overline{X_i} \quad \text{and} \quad \lim_{k \to \infty} \underline{X_i^{(k)}} = \underline{X_i} \quad \text{for all} \quad i = 1, \ldots, n.$$

Let $i \in \{1, \ldots, n\}$ be arbitrary but fixed. Obviously $\langle X_i^{(k)} \rangle$ is a nested sequence. Hence $\langle \underline{X_i^{(k)}} \rangle$ is a monotone, nondecreasing sequence of real numbers, bounded above by $\overline{X_i^{(1)}}$ and so has a limit $a \in \mathbb{R}$. Similarly, $\langle \overline{X_i^{(k)}} \rangle$ is a monotone, non–increasing sequence of real numbers, bounded below by $\underline{X_i^{(1)}}$ and so has a limit $b \in \mathbb{R}$. As $\underline{X_i^{(k)}} \leq \overline{X_i^{(k)}}$ for all k, it holds that $a \leq b$, hence $\langle X_i^{(k)} \rangle$ converges to $[a, b]$. $\square$

1.3 Interval Functions

Interval functions are functions which map intervals to intervals. They arise naturally when one wants to enclose the range of a point function over an interval. Obviously, we are interested in interval functions, which bound the range as tight as possible and which are cheap to evaluate. The interval function, which returns the exact range is most accurate but usually very expensive. The interval function, which returns $[-\infty, \infty]$ for all arguments is cheap, but of little use.

First, we have to make more precise what is meant by "bounding the range tightly". A well known criterion is convergence: How fast does the the interval function value approximate the range, if the width of the argument interval approaches zero? Unfortunately the notion of convergence is only meaningful for "small" argument intervals. Only in this case, a higher convergence order means tighter inclusions.

It seems that no useful criterion was found so far to compare overestimation errors of interval functions in the non–asymptotic case systematically.

In Section 1.3.1 we review some important properties of interval functions in general: Having a topology at hand, continuity is already defined (Section 1.3.1.2). The other two properties, inclusion monotonicity (Section 1.3.1.1) and the Lipschitz property (Section 1.3.1.3) capture in what sense a "reduction" of the argument interval leads to a "reduction" of the function value.

In Section 1.3.2 we introduce the fundamental notion of an interval extension. An interval extension of a point function is an interval function which encloses the range of the point function over intervals. Section 1.3.2.2 is concerned with an important class of interval extensions called centered forms. We give a new elementary proof of their quadratic convergence.

1.3.1 Properties of Interval Functions

Throughout this thesis interval functions are denoted by capital letters. In the sequel let $F : \mathbb{IR}^n \to \mathbb{IR}$ be an interval function.

1.3.1.1 Inclusion Monotonicity

Definition 1.3.1 (Inclusion Monotonicity) *F is inclusion monotone if for all $\boldsymbol{X} \subseteq \boldsymbol{Y} \in \mathbb{IR}^n$ it holds that*

$$F(\boldsymbol{X}) \subseteq F(\boldsymbol{Y}). \quad \square$$

Theorem 1.3.2 *Constant functions and projections $\mathbb{IR}^n \to \mathbb{IR}$ are inclusion monotone.* $\square$

Theorem 1.3.3 *Interval addition, subtraction, multiplication and power by a natural constant are inclusion monotone.* $\square$

Proof. Follows immediately from the fact that $X \circ Y = \{x \circ y \mid x \in X, y \in Y\}$ for $\circ \in \{+, -, *\}$ and $X^d = \{x^d \mid x \in X\}$. $\square$

Remark. Mignitude, magnitude, midpoint and width function are *not* inclusion monotone. $\square$

Theorem 1.3.4 (Composition Preserves Inclusion Monotonicity) *Let $F : \mathbb{IR}^m \to \mathbb{IR}$ and $G_i : \mathbb{IR}^n \to \mathbb{IR}$, $i = 1, \ldots, m$ be inclusion monotone. Then $F(G_1, \ldots, G_m) : \mathbb{IR}^n \to \mathbb{IR}$ is inclusion monotone.* $\square$

Proof. Let F, G_i as in Theorem 1.3.4 and let $\boldsymbol{X} \subseteq \boldsymbol{Y} \in \mathbb{IR}^n$ arbitrary but fixed. Then

$$(G_1(\boldsymbol{X}), \ldots, G_m(\boldsymbol{X})) \subseteq (G_1(\boldsymbol{Y}), \ldots, G_m(\boldsymbol{Y}))$$

as all G_i are inclusion monotone and

$$F(G_1(\boldsymbol{X}), \ldots, G_m(\boldsymbol{X})) \subseteq F(G_1(\boldsymbol{Y}), \ldots, G_m(\boldsymbol{Y}))$$

as F is inclusion monotone. $\square$

Corollary 1.3.5 (Class of Inclusion Monotone Functions) *Let $\mathcal{F}^n$ be the smallest set of functions $\mathbb{IR}^n \to \mathbb{IR}$, which contains all constant functions, all projections, and is closed under addition, subtraction, multiplication and power by a natural constant. Then every $F \in \mathcal{F}^n$ is inclusion monotone.* $\square$

1.3.1.2 Continuity

Definition 1.3.6 (Continuous Function) *$F : \mathbb{IR}^n \to \mathbb{IR}$ is continuous at $\boldsymbol{X} \in \mathbb{IR}^n$ if*

$$\lim_{k \to \infty} F(\boldsymbol{X}^{(k)}) = F(\boldsymbol{X})$$

for every sequence $\langle \boldsymbol{X}^{(k)} \in \mathbb{IR}^n \mid k = 1, 2, \ldots \rangle$ *converging to* $\boldsymbol{X}$. F *is continuous if* F *is continuous at every* $\boldsymbol{X} \in \mathbb{IR}^n$. □

Theorem 1.3.7 *Constant functions and projections* $\mathbb{IR}^n \to \mathbb{IR}$ *are continuous.* □

Proof. Obvious. □

Theorem 1.3.8 *Mignitude, magnitude, midpoint and width are continuous functions.* □

Proof. Let $\langle X^{(k)} \in \mathbb{IR} \mid k = 1, \ldots, n \rangle$ be a sequence converging to X.

- Mignitude and Magnitude. Note that $\mathrm{q}(A, B) \geq |\mathrm{mig}(A) - \mathrm{mig}(B)|$ for all $A, B \in \mathbb{IR}$. Hence

$$\lim_{k \to \infty} |\mathrm{mig}(X^{(k)}) - \mathrm{mig}(X)| \leq \lim_{k \to \infty} \mathrm{q}(X^{(k)}, X) = 0,$$

  i.e. $\mathrm{mig}(X^{(k)})$ converges to $\mathrm{mig}(X)$. A similar argument holds for magnitude.

- Midpoint.

$$\begin{aligned} \lim_{k \to \infty} \mathrm{mid}(X^{(k)}) &= \lim_{k \to \infty} 1/2(\overline{X^{(k)}} + \underline{X^{(k)}}) \\ &= 1/2(\lim_{k \to \infty} \overline{X^{(k)}} + \lim_{k \to \infty} \underline{X^{(k)}}) \\ &= 1/2(\overline{X} + \underline{X}) \\ &= \mathrm{mid}(X). \end{aligned}$$

- Width.

$$\begin{aligned} \lim_{k \to \infty} \mathrm{w}(X^{(k)}) &= \lim_{k \to \infty} (\overline{X^{(k)}} - \underline{X^{(k)}}) \\ &= \lim_{k \to \infty} \overline{X^{(k)}} - \lim_{k \to \infty} \underline{X^{(k)}} \\ &= \overline{X} - \underline{X} \\ &= \mathrm{w}(X). \ \square \end{aligned}$$

Theorem 1.3.9 *Interval addition, subtraction, multiplication and power by a natural constant are continuous.* □

Proof. Let $\langle (X^{(k)}, Y^{(k)}) \in \mathbb{IR}^2 \mid k = 1, 2, \ldots \rangle$ be a sequence converging to (X, Y).

- Addition and Subtraction.

$$\begin{aligned} \lim_{k \to \infty} (X^{(k)} \pm Y^{(k)}) &= \lim_{k \to \infty} [\underline{X^{(k)}} \pm \underline{Y^{(k)}}, \overline{X^{(k)}} \pm \overline{Y^{(k)}}] \\ &= [\lim_{k \to \infty} (\underline{X^{(k)}} \pm \underline{Y^{(k)}}), \lim_{k \to \infty} (\overline{X^{(k)}} \pm \overline{Y^{(k)}})] \\ &= [\underline{X} \pm \underline{Y}, \overline{X} \pm \overline{Y}] \\ &= X \pm Y. \end{aligned}$$

- Multiplication.

$$\begin{aligned} \lim_{k \to \infty} (X^{(k)} Y^{(k)}) &= \lim_{k \to \infty} [\underline{X^{(k)}}\,\underline{Y^{(k)}}, \underline{X^{(k)}}\,\overline{Y^{(k)}}, \overline{X^{(k)}}\,\underline{Y^{(k)}}, \overline{X^{(k)}}\,\overline{Y^{(k)}}] \\ &= [\lim_{k \to \infty} \underline{X^{(k)}}\,\underline{Y^{(k)}}, \lim_{k \to \infty} \underline{X^{(k)}}\,\overline{Y^{(k)}}, \lim_{k \to \infty} \overline{X^{(k)}}\,\underline{Y^{(k)}}, \lim_{k \to \infty} \overline{X^{(k)}}\,\overline{Y^{(k)}}] \\ &= [\underline{X}\,\underline{Y}, \underline{X}\,\overline{Y}, \overline{X}\,\underline{Y}, \overline{X}\,\overline{Y}] \\ &= XY. \end{aligned}$$

- Power by $d > 0$, $d \in \mathbb{N}$.
  - Assume d is odd.

$$\begin{aligned}\lim_{k\to\infty}(X^{(k)^d}) &= \lim_{k\to\infty}[\underline{X^{(k)}}^d, \overline{X^{(k)}}^d] \\ &= [\lim_{k\to\infty}(\underline{X^{(k)}}^d), \lim_{k\to\infty}(\overline{X^{(k)}}^d)] \\ &= [\underline{X}^d, \overline{X}^d] \\ &= X^d.\end{aligned}$$

  - Assume d is even.

$$\begin{aligned}\lim_{k\to\infty}(X^{(k)^d}) &= \lim_{k\to\infty}[\mathrm{mig}(X^{(k)})^d, \mathrm{mag}(X^{(k)})^d] \\ &= [\lim_{k\to\infty}(\mathrm{mig}(X^{(k)})^d), \lim_{k\to\infty}(\mathrm{mag}(X^{(k)})^d)] \\ &= [\mathrm{mig}(X)^d, \mathrm{mag}(X)^d] \\ &= X^d. \ \square\end{aligned}$$

Theorem 1.3.10 (Composition Preserves Continuity) *Let $F : \mathbb{IR}^m \to \mathbb{IR}$ and $G_i : \mathbb{IR}^n \to \mathbb{IR}$, $i = 1, \ldots, m$ be continuous functions. Then $F(G_1, \ldots, G_m) : \mathbb{IR}^n \to \mathbb{IR}$ is continuous.* $\square$

Proof. See [Armstrong, 1983], Theorem 2.7. $\square$

Corollary 1.3.11 (Class of Continuous Functions) *Let $\mathcal{F}^n$ be the smallest set of functions $\mathbb{IR}^n \to \mathbb{IR}$, which contains all constant functions, all projections, and is closed under the mignitude, magnitude, midpoint, width function, addition, subtraction, multiplication and power by a natural constant. Then every $F \in \mathcal{F}^n$ is continuous.* $\square$

1.3.1.3 Lipschitz Property

Definition 1.3.12 (Lipschitz Constant, Lipschitz Function) *Let $\boldsymbol{A} \in \mathbb{IR}^n$. A real number $\lambda_{F,\boldsymbol{A}}$ is a Lipschitz constant of F in $\boldsymbol{A}$ if for all $\boldsymbol{X} \subseteq \boldsymbol{A}$ it holds that*

$$\mathrm{w}(F(\boldsymbol{X})) \leq \lambda_{F,\boldsymbol{A}}\hat{\mathrm{w}}(\boldsymbol{X}).$$

F is a Lipschitz function if there exists a Lipschitz constant of F in $\boldsymbol{A}$ for all $\boldsymbol{A} \in \mathbb{IR}^n$. $\square$

Theorem 1.3.13 *Constant functions and projections $\mathbb{IR}^n \to \mathbb{IR}$ are Lipschitz.* $\square$

Proof. Obvious. $\square$

Theorem 1.3.14 *Mignitude, magnitude, midpoint and width are Lipschitz functions.* $\square$

Proof. Obvious. $\square$

Theorem 1.3.15 *Interval addition, subtraction, multiplication and power by a natural constant are Lipschitz functions.* $\square$

Proof.

- Addition and Subtraction. Let $\boldsymbol{A} \in \mathbb{IR}^2$ and let (X, Y) range over subintervals of $\boldsymbol{A}$. Then

$$\begin{aligned}\mathrm{w}(X + Y) &= \mathrm{w}(X) + \mathrm{w}(Y) \\ &\leq 2\hat{\mathrm{w}}(X, Y).\end{aligned}$$

  Hence 2 is a Lipschitz constant of addition in $\boldsymbol{A}$.

- Multiplication. Let $\boldsymbol{A} \in \mathbb{IR}^2$ and let (X, Y) range over subintervals of $\boldsymbol{A}$. Then

$$\begin{aligned} \mathrm{w}(XY) &\leq \mathrm{mag}(X)\mathrm{w}(Y) + \mathrm{mag}(Y)\mathrm{w}(X), \text{see [Neumaier, 1990]} \\ &\leq \mathrm{mag}(A_1)\mathrm{w}(Y) + \mathrm{mag}(A_2)\mathrm{w}(X) \\ &\leq \max\{\mathrm{mag}(A_1), \mathrm{mag}(A_2)\}\big(\mathrm{w}(X) + \mathrm{w}(Y)\big) \\ &\leq 2\max\{\mathrm{mag}(A_1), \mathrm{mag}(A_2)\}\hat{\mathrm{w}}(X, Y). \end{aligned}$$

  Hence $2\max\{\mathrm{mag}(A_1), \mathrm{mag}(A_2)\}$ is a Lipschitz constant of multiplication in $\boldsymbol{A}$.

- Power. Let $A \in \mathbb{IR}$ and let X range over subintervals of A. By induction we show that for all $i > 0$ it holds that

$$\mathrm{w}(X^i) \leq i\,\mathrm{mag}(X)^{i-1}\mathrm{w}(X). \tag{1.3.1}$$

  Obviously 1.3.1 holds for $i = 1$. Assume 1.3.1 holds for some $i > 0$. Then

$$\begin{aligned} \mathrm{w}(X^{i+1}) &\leq \mathrm{w}(XX^i) \\ &\leq \mathrm{mag}(X)\mathrm{w}(X^i) + \mathrm{mag}(X^i)\mathrm{w}(X) \\ &\leq \mathrm{mag}(X)i\mathrm{mag}(X)^{i-1}\mathrm{w}(X) + \mathrm{mag}(X)^i\mathrm{w}(X) \\ &\leq i\,\mathrm{mag}(X)^i\mathrm{w}(X) + \mathrm{mag}(X)^i\mathrm{w}(X) \\ &= (i+1)\mathrm{mag}(X)^i\mathrm{w}(X). \end{aligned}$$

  Hence $\mathrm{w}(X^i) \leq i\,\mathrm{mag}(A)^{i-1}\mathrm{w}(X)$ and $i\,\mathrm{mag}(A)^{i-1}$ is a Lipschitz constant of the i-th power function in A. □

Theorem 1.3.16 (Preservation of Lipschitz Property under Composition) *Let $F : \mathbb{IR}^m \to \mathbb{IR}$ be a Lipschitz function and let $G_i : \mathbb{IR}^n \to \mathbb{IR}$, $i = 1, \ldots, m$ be Lipschitz functions such that each G_i is continuous or inclusion monotone. Then $F(G_1, \ldots, G_m) : \mathbb{IR}^n \to \mathbb{IR}$ is a Lipschitz function.* □

The proof of Theorem 1.3.16 needs some preparation. In the following let $\mathbb{I}\boldsymbol{A}$ be the set of subintervals of $\boldsymbol{A}$ with the subspace topology of $\mathbb{IR}^n$ defined on it.

Lemma 1.3.17 *$\mathbb{I}\boldsymbol{A}$ is compact for every $\boldsymbol{A} \in \mathbb{IR}^n$.* □

Proof. Let $\boldsymbol{A} \in \mathbb{IR}^n$ arbitrary but fixed, define

$$\mathcal{A} = \{(\boldsymbol{u}, \boldsymbol{v}) \in \mathbb{R}^{2n} \mid \underline{A}_i \leq u_i \leq v_i \leq \overline{A}_i, \;\; i = 1, \ldots, n\}$$

and let $h : \mathcal{A} \to \mathbb{I}\boldsymbol{A}$ be defined as

$$h(\boldsymbol{u}, \boldsymbol{v}) = [\boldsymbol{u}, \boldsymbol{v}].$$

$\mathcal{A}$ is a closed and bounded subset of $\mathbb{R}^{2n}$, hence it is compact ([Armstrong, 1983], Theorem 3.1). Obviously h is continuous and therefore $\mathbb{I}\boldsymbol{A}$ is compact ([Armstrong, 1983], Theorem 3.4). □

Lemma 1.3.18 *Let $\boldsymbol{A} \in \mathbb{IR}^n$ and assume $G : \mathbb{IR}^n \to \mathbb{IR}$ is continuous. Then*

$$\{\mathrm{mag}(G(\boldsymbol{X})) \mid \boldsymbol{X} \in \mathbb{I}A\}$$

is bounded. □

Proof. Follows from Lemma 1.3.17 and [Armstrong, 1983], Theorem 3.10. □

Proof of Theorem 1.3.16. Let F, G_i be as in Theorem 1.3.16. Let $\boldsymbol{A} \in \mathbb{IR}^n$ arbitrary but fixed. First, we show that there exists an interval vector $\boldsymbol{A}' \in \mathbb{IR}^n$ such that

$$A'_i \supseteq G_i(\boldsymbol{X}) \;\text{ for all }\; \boldsymbol{X} \in \mathbb{I}\boldsymbol{A}.$$

- If G_i is continuous then the existence of A'_i follows from Lemma 1.3.18.

- If G_i is inclusion monotone then we may choose $A'_i = G_i(\boldsymbol{A})$.

Let $\lambda_{F,A'}$ be a Lipschitz constant of F in $\boldsymbol{A}'$ and let $\lambda_{G_i,A}$ be Lipschitz constants of G_i in $\boldsymbol{A}$ for $i = 1, \ldots, m$ respectively. Then for every subinterval $\boldsymbol{X}$ of $\boldsymbol{A}$ it holds that

$$\begin{aligned} \mathrm{w}(F(G_1(\boldsymbol{X}), \ldots, G_m(\boldsymbol{X})) &\leq \lambda_{F,A'} \max\{\mathrm{w}(G_1(\boldsymbol{X})), \ldots, \mathrm{w}(G_m(\boldsymbol{X}))\} \\ &\leq \lambda_{F,A'} \max\{\lambda_{G_1,A}, \ldots, \lambda_{G_m,A}\} \hat{\mathrm{w}}(\boldsymbol{X}). \end{aligned}$$

Hence

$$\lambda_{F,A'} \max\{\lambda_{G_1,A}, \ldots, \lambda_{G_m,A}\}$$

is a Lipschitz constant of $F(G_1, \ldots, G_m)$ in $\boldsymbol{A}$. □

Corollary 1.3.19 *Let $\mathcal{F}^n$ be the smallest set of functions $\mathbb{IR}^n \to \mathbb{IR}$, which contains all constant functions, all projections, and is closed under the mignitude, magnitude, midpoint, width function, addition, subtraction, multiplication and power by a natural constant. Then every $F \in \mathcal{F}^n$ is a Lipschitz function.* □

1.3.2 Inclusion of the Range of Real Functions

1.3.2.1 Interval Extensions

In this section we are concerned with interval functions, which bound the range of real functions. In the following let $f : \mathbb{R}^n \to \mathbb{R}$ be a point function and let $F : \mathbb{IR}^n \to \mathbb{IR}$ be an interval function.

Notation. The range of f on $\boldsymbol{X}$ is denoted by

$$f(X) = \{f(\boldsymbol{x}) \mid \boldsymbol{x} \in \boldsymbol{X}\}. \quad \square$$

Theorem 1.3.20 *If f is continuous then $f(\boldsymbol{X}) \in \mathbb{IR}$ for all $\boldsymbol{X} \in \mathbb{IR}^n$.* □

Proof. Assume $f : \mathbb{R}^n \to \mathbb{R}$ is continuous and let $\boldsymbol{X} \in \mathbb{IR}^n$. As $\boldsymbol{X}$ is a compact and connected subset of $\mathbb{R}^n$ and f is continuous, it follows that $f(\boldsymbol{X})$ is a compact and connected subset of $\mathbb{R}$ ([Armstrong, 1983], Theorem 3.1, 3.4, 3.21). Hence $f(X) \in \mathbb{IR}$ ([Armstrong, 1983], Theorem 3.19). □

Definition 1.3.21 (Interval Extension) *F is an interval extension of f if*

$$\begin{aligned} f(\boldsymbol{x}) &= F(\boldsymbol{x}) \text{ for all } \boldsymbol{x} \in \mathbb{R}^n \text{ and} \\ f(\boldsymbol{X}) &\subseteq F(\boldsymbol{X}) \text{ for all } \boldsymbol{X} \in \mathbb{IR}^n. \quad \square \end{aligned}$$

If F is an interval extension of f then $F(\boldsymbol{X})$ bounds the range of f on $\boldsymbol{X}$. The following theorem gives a different characterization of interval extensions.

Theorem 1.3.22 *If F is inclusion monotone and*

$$f(\boldsymbol{x}) = F(\boldsymbol{x}) \text{ for all } \boldsymbol{x} \in \mathbb{R}^n,$$

then F is an interval extension of f. □

Proof. Let F be inclusion monotone and assume $f(\boldsymbol{x}) = F(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^n$. We have to show that $f(\boldsymbol{X}) \subseteq F(\boldsymbol{X})$ for all $\boldsymbol{X}$. Let $\boldsymbol{X} \in \mathbb{IR}^n$ and $y \in f(\boldsymbol{X})$ arbitrary but fixed. Then there exists $\boldsymbol{x} \in \boldsymbol{X}$ such that $y = f(\boldsymbol{x}) = F(\boldsymbol{x})$ and inclusion monotonicity of F implies $y \in F(\boldsymbol{X})$. Hence $f(\boldsymbol{X}) \subseteq F(\boldsymbol{X})$. □

Obviously we are interested in interval extensions of which enclose the precise range as tight as possible. A frequently used criterion for the accuracy of the overestimation is given by the following definition.

Definition 1.3.23 (Convergence) *An interval extension F of f converges to f with order r in $\boldsymbol{A} \subseteq \mathbb{IR}^n$ if there exists a real number $\lambda_{F,f,A}$ such that for all $\boldsymbol{X} \in \mathbb{I}\boldsymbol{A}$*

$$\mathrm{q}(F(\boldsymbol{X}), [f(\boldsymbol{X})]) \leq \lambda_{F,f,A} \hat{\mathrm{w}}(\boldsymbol{X})^r .$$

F converges to f with order r if F converges to f with order r in $\boldsymbol{A}$ for all $\boldsymbol{A} \in \mathbb{IR}^n$. □

Lipschitz interval extensions converge at least linearly.

Theorem 1.3.24 *A Lipschitz interval extension F of f converges linearly to f.* □

Proof. Assume F is a Lipschitz interval extension of f, let $\boldsymbol{A} \in \mathbb{IR}^n$ arbitrary but fixed and let $\lambda_{F,A}$ be a Lipschitz constant of F in $\boldsymbol{A}$. Then

$$\begin{aligned} \mathrm{q}(F(\boldsymbol{X}), [f(\boldsymbol{X})]) &\leq \mathrm{w}(F(\boldsymbol{X})) \\ &\leq \lambda_{F,A} \hat{\mathrm{w}}(\boldsymbol{X}) \end{aligned}$$

for all $\boldsymbol{X} \in \mathbb{I}\boldsymbol{A}$. □

1.3.2.2 Centered Forms

An important class of interval extensions are centered forms. Centered forms are quadratically convergent, hence they give tight inclusions if the with of the argument interval is small. The notion of a centered form is not defined consistently in the literature. The definition we use is similar to a very general one by [Ratschek and Rokne, 1984]. In Chapter 3 and Chapter 4 we study some concrete instances of centered forms for polynomials.

Centered forms are one of the main areas studied in the literature on interval arithmetic, see for example [Moore, 1966], [Hansen, 1968], [Hansen, 1969], [Chuba and Miller, 1972], [Miller, 1972], [Miller, 1973], [Alefeld and Herzberger, 1974], [Miller, 1975], [Ratschek, 1977], [Hansen, 1978b], [Ratschek, 1978], [Moore, 1979], [Caprani and Madsen, 1980], [Ratschek, 1980a], [Ratschek, 1980b], [Ratschek and Rokne, 1980b], [Ratschek and Rokne, 1980a], [Krawczyk, 1980a], [Krawczyk, 1980b], [Alefeld, 1981], [Alefeld and Rokne, 1981], [Ratschek and Schröder, 1981], [Rokne, 1981], [Krawczyk and Nickel, 1982], [Krawczyk, 1982], [Rokne and Wu, 1982], [Alefeld and Herzberger, 1983], [Krawczyk, 1983], [Rall, 1983], [Rokne and Wu, 1983], [Ratschek and Rokne, 1984], [Alefeld and Lohner, 1985], [Krawczyk and Neumaier, 1985], [Rokne, 1985], [Rokne, 1986], [Baumann, 1988], [Alefeld, 1990], [Neumaier, 1990].

Throughout this section let $f : \mathbb{R}^n \to \mathbb{R}$ be continuous.

Definition 1.3.25 (Centered Form)

- *Let $\boldsymbol{g} : \mathbb{R}^{2n} \to \mathbb{R}^n$ such that*

$$f(\boldsymbol{x}) = f(\boldsymbol{c}) + \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{c})(\boldsymbol{x} - \boldsymbol{c}) \quad \text{for all} \quad \boldsymbol{x}, \boldsymbol{c} \in \mathbb{R}^n.$$

- *Let $\boldsymbol{z} : \mathbb{IR}^n \to \mathbb{R}^n$ such that*

$$\boldsymbol{z}(\boldsymbol{X}) \in \boldsymbol{X} \quad \text{for all} \quad \boldsymbol{X} \in \mathbb{IR}^n.$$

- *Let $\boldsymbol{G} : \mathbb{IR}^n \to \mathbb{IR}^n$ such that*

$$\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{z}(\boldsymbol{X})) \in G(\boldsymbol{X}) \quad \text{for all} \quad \boldsymbol{X} \in \mathbb{IR}^n \quad \text{and} \quad \text{for all } \boldsymbol{x} \in \boldsymbol{X}$$

  and G_i is Lipschitz for $i = 1, \ldots, n$.

Then $F : \mathbb{IR}^n \to \mathbb{IR}$,

$$F(\boldsymbol{X}) = f(\boldsymbol{z}(\boldsymbol{X})) + \boldsymbol{G}(\boldsymbol{X})(\boldsymbol{X} - \boldsymbol{z}(\boldsymbol{X}))$$

is a centered form of f. □

In the sequel let $\boldsymbol{g}, \boldsymbol{z}, \boldsymbol{G}$ and F as in Definition 1.3.25. First, we show that centered forms are interval extensions.

Theorem 1.3.26 *F is an interval extension of f.* □

Proof.

- If $\boldsymbol{x} \in \mathbb{R}^n$ then $\boldsymbol{z}(\boldsymbol{x}) = \boldsymbol{x}$, hence $F(\boldsymbol{x}) = f(\boldsymbol{x})$.
- Let $\boldsymbol{X} \in \mathbb{IR}^n$ arbitrary but fixed and let $\boldsymbol{c} = \boldsymbol{z}(\boldsymbol{X})$. For all $\boldsymbol{x} \in \boldsymbol{X}$ it holds that

$$\begin{aligned} f(\boldsymbol{x}) &= f(\boldsymbol{c}) + \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{c})(\boldsymbol{x} - \boldsymbol{c}) \\ &\in f(\boldsymbol{c}) + \boldsymbol{G}(\boldsymbol{X})(\boldsymbol{x} - \boldsymbol{c}) \\ &\subseteq f(\boldsymbol{c}) + \boldsymbol{G}(\boldsymbol{X})(\boldsymbol{X} - \boldsymbol{c}) \\ &= F(\boldsymbol{X}), \end{aligned}$$

  hence $f(\boldsymbol{X}) \subseteq F(\boldsymbol{X})$.

Theorem 1.3.27 (Quadratic Convergence of Centered Forms.) *F converges quadratically to f.* □

The quadratic convergence of centered forms was first conjectured by [Moore, 1966] and for multivariate rational functions first proven by [Hansen, 1969]. A proof by induction on rational expressions is given by [Chuba and Miller, 1972] and [Miller, 1972]. Quadratic convergence is also proved in [Alefeld and Herzberger, 1974], [Alefeld and Herzberger, 1983]. To my knowledge, the most general proof is given by [Krawczyk and Nickel, 1982], [Ratschek and Rokne, 1984]. It uses topological properties and Miranda's Theorem [Miranda, 1940]. We give a new elementary proof. It requires some preparation, though, and is postponed until after Lemma 1.3.30.

Definition 1.3.28 *The signed mignitude of an interval $X \in \mathbb{IR}$ is defined as*

$$\operatorname{smig}(X) = \begin{cases} \underline{X} & \text{if } \underline{X} > 0 \\ \overline{X} & \text{if } \overline{X} < 0 \\ 0 & \text{if } 0 \in X. \end{cases}$$

For interval vectors $\boldsymbol{X} \in \mathbb{IR}^n$ we define

$$\operatorname{smig}(\boldsymbol{X}) = (\operatorname{smig}(X_1), \ldots, \operatorname{smig}(X_n)).$$ □

The following property of the smig function is trivial but will be useful later.

Lemma 1.3.29 *For all $X \in \mathbb{IR}$ it holds that*

$$\operatorname{smig}(X) + [-1, 1]\operatorname{w}(X) \supseteq X.$$ □ (1.3.2)

Proof. For all $x \in X$ it holds that $x + [-1, 1]\operatorname{w}(X) \supseteq X$. Further, $\operatorname{smig}(X) \in X$. □

Lemma 1.3.30 *For all $\boldsymbol{X} \in \mathbb{IR}^n$ it holds that*

$$f(\boldsymbol{z}(\boldsymbol{X})) + \operatorname{smig}(\boldsymbol{G}(\boldsymbol{X}))(\boldsymbol{X} - \boldsymbol{z}(\boldsymbol{X})) \subseteq f(\boldsymbol{X}).$$ □ (1.3.3)

Proof. Let $\boldsymbol{X} \in \mathbb{IR}^n$ arbitrary but fixed and let $\boldsymbol{c} = \boldsymbol{z}(\boldsymbol{X})$. We have to show that

$$\overline{f(\boldsymbol{c}) + \operatorname{smig}(\boldsymbol{G}(\boldsymbol{X}))(\boldsymbol{X} - \boldsymbol{c})} \leq \overline{f(\boldsymbol{X})} \quad (1.3.4)$$

$$\underline{f(\boldsymbol{c}) + \operatorname{smig}(\boldsymbol{G}(\boldsymbol{X}))(\boldsymbol{X} - \boldsymbol{c})} \geq \underline{f(\boldsymbol{X})}. \quad (1.3.5)$$

We prove (1.3.4), a similar argument shows (1.3.5). Let

$$\begin{aligned}\Im^+ &= \{i \mid \operatorname{smig}(G_i(\boldsymbol{X})) > 0, \quad i = 1, \ldots, n\}\\ \Im^- &= \{i \mid \operatorname{smig}(G_i(\boldsymbol{X})) < 0, \quad i = 1, \ldots, n\}\end{aligned}$$

and let $\hat{\boldsymbol{x}} \in \boldsymbol{X}$ be defined as

$$\hat{x}_i = \begin{cases} \overline{X}_i & \text{if } i \in \Im^+ \\ \underline{X}_i & \text{if } i \in \Im^- \\ c_i & \text{else.} \end{cases}$$

Then

$$\begin{aligned}\overline{f(\boldsymbol{c}) + \operatorname{smig}(\boldsymbol{G}(\boldsymbol{X}))(\boldsymbol{X} - \boldsymbol{c})} &= f(\boldsymbol{c}) + \sum_{i=1}^{n} \overline{\operatorname{smig}(G_i(\boldsymbol{X}))(X_i - c_i)}\\ &= f(\boldsymbol{c}) + \sum_{i \in \Im^+} \operatorname{smig}(G_i(\boldsymbol{X}))(\overline{X}_i - c_i) + \sum_{i \in \Im^-} \operatorname{smig}(G_i(\boldsymbol{X}))(\underline{X}_i - c_i)\\ &\le f(\boldsymbol{c}) + \sum_{i \in \Im^+} g_i(\hat{\boldsymbol{x}}, \boldsymbol{c})(\overline{X}_i - c_i) + \sum_{i \in \Im^-} g_i(\hat{\boldsymbol{x}}, \boldsymbol{c})(\underline{X}_i - c_i)\\ &= f(\boldsymbol{c}) + \sum_{i=1}^{n} g_i(\hat{\boldsymbol{x}}, \boldsymbol{c})(\hat{x}_i - c_i)\\ &= f(\boldsymbol{c}) + \boldsymbol{g}(\hat{\boldsymbol{x}}, \boldsymbol{c})(\hat{\boldsymbol{x}} - \boldsymbol{c})\\ &= f(\hat{\boldsymbol{x}})\\ &\le \overline{f(X)}. \ \square\end{aligned}$$

Proof of Theorem 1.3.27. Let $\boldsymbol{A} \in \mathbb{IR}^n$ and for $i = 1, \ldots, n$ let $\lambda_{G_i,A}$ be a Lipschitz constant of G_i in $\boldsymbol{A}$. Let $\boldsymbol{X} \in \mathbb{I}\boldsymbol{A}$ arbitrary but fixed and let $\boldsymbol{c} = \boldsymbol{z}(\boldsymbol{X})$. Then

$$\begin{aligned}F(\boldsymbol{X}) &= f(\boldsymbol{c}) + \sum_{i=1}^{n} G_i(\boldsymbol{X})(X_i - c_i)\\ &\overset{(1.3.2)}{=} f(\boldsymbol{c}) + \sum_{i=1}^{n} \big(\operatorname{smig}(G_i(\boldsymbol{X})) + [-1,1]\operatorname{w}(G_i(\boldsymbol{X}))\big)(X_i - c_i)\\ &\subseteq f(\boldsymbol{c}) + \sum_{i=1}^{n} \operatorname{smig}(G_i(\boldsymbol{X}))(X_i - c_i) + [-1,1]\operatorname{w}(G_i(\boldsymbol{X}))(X_i - c_i)\\ &= f(\boldsymbol{c}) + \sum_{i=1}^{n} \operatorname{smig}(G_i(\boldsymbol{X}))(X_i - c_i) + \sum_{i=1}^{n} [-1,1]\operatorname{w}(G_i(\boldsymbol{X}))(X_i - c_i)\\ &\overset{(1.3.3)}{\subseteq} f(\boldsymbol{X}) + \sum_{i=1}^{n} [-1,1]\operatorname{w}(G_i(\boldsymbol{X}))(X_i - c_i)\\ &\subseteq f(\boldsymbol{X}) + \sum_{i=1}^{n} [-1,1]\operatorname{w}(G_i(\boldsymbol{X}))\operatorname{w}(X_i)\\ &\subseteq f(\boldsymbol{X}) + [-1,1]\hat{\operatorname{w}}(\boldsymbol{X}) \sum_{i=1}^{n} \operatorname{w}(G_i(\boldsymbol{X}))\\ &\subseteq f(\boldsymbol{X}) + [-1,1]\hat{\operatorname{w}}(\boldsymbol{X})^2 \sum_{i=1}^{n} \lambda_{G_i,A}\\ &\subseteq f(\boldsymbol{X}) + [-1,1]\hat{\operatorname{w}}(\boldsymbol{X})^2 \lambda_{G,A},\end{aligned}$$

where

$$\lambda_{G,A} = \sum_{i=1}^{n} \lambda_{G_i,A}.$$

Hence,

$$\mathrm{q}(F(\boldsymbol{X}), f(\boldsymbol{X})) \leq \lambda_{G,A} \hat{\mathrm{w}}(\boldsymbol{X})^2$$

and F converges quadratically to f according to Definition 1.3.23. □

Chapter 2

Implementation of Interval Arithmetic

Interval arithmetic was introduced originally as a new approach to the fundamental problem of representing real numbers on a computer. As is well known, this problem cannot be solved because the set of real numbers is uncountable but the amount of memory of computers is only countable. (More precisely, it is finite but can be extended in principle without limit). So far, there have been two main approaches to deal with this situation:

- In many cases, it is sufficient to compute with countable subsets of the reals, e.g. integers, rational numbers or finite extensions of the rationals. Countable sets can be encoded, i.e. the elements of a countable set can be mapped injectively to finite length bitstrings. Thus, for every algorithm and input data there exists a finite memory request, such that the computation is possible if the requested memory is available. The disadvantage of this approach is efficency. The computing time of arithmetic operations depends on the length of the argument bitstrings and is therefore not constant.
- Another approach is to choose a finite subset of the reals and to represent them by fixed length bit strings. We call such a subset of the reals the set of represented numbers. Arithmetic operations on represented numbers do in general not yield a represented number. Further, the input numbers to some algorithm need not be represented. Hence, input numbers and intermediate results have to be rounded to a represented number before they can be processed further. This allows efficient computation because all (rounded) arithmetic operations can be performed in constant time. However, rounding introduces errors which accumulate during the computation, and in general it is highly non–trivial to estimate the size of the error in the output.

The main motivation for interval arithmetic is to overcome the correctness deficiency of the second approach. The idea is to enclose a real number by a real interval where the endpoints are represented numbers. Naturally, these intervals are called represented intervals and they are encoded simply by encoding the endpoints. As in the case of represented numbers, arithmetic operations on represented intervals do in general not yield represented intervals. Hence, intermediate results have to be approximated by represented intervals, and this approximation is always done by represented super intervals. Similarly, input values are enclosed by represented intervals. Thus, by a suitable modification of the predicates and the control structures, it can be guaranteed that the output of an algorithm which operates on represented intervals contains the correct result.

As each intermediate result is rounded outwards, the accuracy of the inclusion decreases during the computation, which leads sometimes to useless results. The simplest solution is to extend the set of represented numbers in order to reduce the amount of overestimation in each arithmetic operation. In many cases, however, one can obtain tighter results by rewriting the algorithm without extending the set of represented numbers. As the set of represented numbers is finite and hence bounded, it is not possible to enclose every real number by a represented interval. We will handle this problem by adding intervals

of the form $\{x \in \mathbb{R} \mid x \leq y\}$, $\{x \in \mathbb{R} \mid x \geq y\}$ and $\mathbb{R}$ to the set of represented intervals, for all represented numbers y.

Fortunately the definition of represented numbers is standardized by the IEEE standard on binary floating point numbers. The main features of this standard are reviewed in Section 2.1. A mathematical abstraction of the standard is subject of Section 2.2. In Section 2.3 we give an implementation of interval arithmetic on top of the standard and prove its correctness.

2.1 The IEEE Standard 754 for Binary Floating Point Arithmetic

Rounded arithmetic on finite subsets of the reals plays a central role in scientific computing. Much effort has been devoted to improve efficiency, and nowadays there is support by special hardware in almost every computer. A necessary precondition for building hardware was to standardize the encoding of represented numbers, which was achieved by the IEEE standard 754 [IEEE, 1985]. Below, we summarize briefly a subset of this standard, which will be used in the following sections for implementing machine interval arithmetic. The bitstrings by which the represented numbers are encoded are called "floating point numbers".

2.1.1 Floating Point Number Formats

The format of a floating point number describes its bitlength and its interpretation as a real number. The IEEE Standard 754 specifies 4 floating point number formats. In our implementations we are using mainly the double precision numbers, hence we restrict our considerations to this format. A double precision floating point number is a string of 64 bits, which consists of

- 1 bit for the sign $s \in \{-1, 1\}$
- 11 bits for the exponent $e \in \{-1023, \ldots, 1024\}$
- 52 bits for the fraction $f \in \{0, 2^{-52}, 2 \cdot 2^{-52}, 3 \cdot 2^{-52}, \ldots, (2^{52} - 1)2^{-52}\}$.

The value of a floating point number is either undefined or an element of $\mathbb{R} \cup \{-\infty, \infty\}$. Given values for s, e and f, the value v of a floating point number computes as follows:

- If $-1023 < e < 1024$, then $v = s(1 + f) * 2^e$.
- If $e = -1023$ then $v = sf * 2^{-1022}$. Non–zero floating point numbers with this property are called denormalized.
- If $e = 1024$ and $f \neq 0$ then v is undefined. Floating point numbers with this property are called Nan (not a number).
- If $e = 1024$ and $f = 0$ then $v = s\,\infty$. Floating point numbers with this property are called infinite.

In the first 2 cases, the floating point number is called ordinary. If no confusion can arise, we do not distinguish between a floating point number and its value.

2.1.2 Predicates

The predicates $=$, $\neq$, $<$, $\leq$, $>$, $\geq$ are defined on non–Nan floating point numbers according to their values. The IEEE standard 754 defines the result of the comparison operations also for the case when one or both operands are Nan, but we will not make use of this feature.

2.1.3 Direction Rounded Arithmetic

The arithmetic operations $+, -, *, /$ and the square root operation are defined on floating point numbers. The operations are performed as if they first produced a correct intermediate result which is then rounded to a neighboring floating point number. It is possible to decide whether the correct result is rounded up, down, to the nearest represented number, or towards zero. In case of rounding to nearest, there can be two floating point numbers which are equivalently near. In such a situation, the one with its least significant fraction bit zero is chosen.

If a real number, which is larger than the largest ordinary floating point number, is rounded up or to nearest, then the result is ∞. Rounding such a number to zero or down, results in the largest ordinary floating point number. Similarly, if a real number, which is is smaller than the smallest ordinary floating point number, is rounded down or to nearest, then the result is $-\infty$. Rounding such a number to zero or up results in the smallest ordinary floating point number.

Division of a non–zero ordinary floating point number by zero results in ∞ if the signs of the operands are equal, and $-\infty$ if they are different. (Note that there exist two floating point numbers with value zero but with different signs.)

Real arithmetic is extended to operands $\pm\infty$ in the usual way. The result of $\infty - \infty$, $-\infty + \infty$, $0 * \pm\infty$, $\pm\infty * 0$, $0/0$, $\pm\infty / \pm \infty$ or square root of a negative number is a Nan.

If at least one operand is a Nan, then the result is also a Nan.

2.1.4 Exception Handling

The IEEE standard 754 defines 5 exceptions which are signaled when detected during a floating point operation. For each type of exception there exists a status flag which is set on any occurrence of the corresponding exception. It is reset only at the user's request. The user may test and alter each flag individually. Further, the user can define a trap for each exception, which is taken unless the exception is masked. Again, the user has the possibility to mask or to unmask each exception individually. The 5 exceptions are as follows:

- Invalid Operation. The invalid operation exception is raised if an operand is invalid for the operation to be performed. In our applications these are $\infty - \infty$, $-\infty + \infty$, $0 * \pm\infty$, $\pm\infty * 0$, $0/0$, $\pm\infty / \pm \infty$ and square root of a negative floating point number.
- Division by Zero. The division by zero exception is raised during a division, where the numerator is an ordinary non–zero number and the denominator is zero.
- Overflow. The overflow exception is raised if the rounded result of an operation exceeds in magnitude what would have been the result if the exponent range were unbounded.
- Underflow. The precise definition of this exception depends on the implementation. Roughly, it is raised if the result of an operation is very small in magnitude and inaccurate because it is approximated by a denormalized number.
- Inexact. The inexact exception is raised if the rounded result of a floating point operation is not the same as the result of the corresponding real number operation.

2.2 Floating Point Number Arithmetic

Informally, the most important facts about floating point numbers for our purposes can be abstracted as follows:

- There are three special numbers, $+\infty$, $-\infty$ and an error symbol.

- Arithmetic is defined in the usual way, where the exact result is rounded in a specified direction after each operation. If the result is undefined, as e.g $\infty - \infty$ or $0 * \infty$, then the result is the error symbol.
- If round to nearest is specified and there are two floating point numbers which are equally near, then the one with its least significant fraction bit zero is delivered.
- In case of any kind of error, a signal is raised. The user can specify that this signal is ignored or that a trap handler is called.

In order to facilitate mathematical reasoning on floating point numbers and arithmetic, we define them more abstractly. The definitions are such that their implementation on a machine, which conforms to the IEEE standard 754, is straight forward. On the other hand, they capture only the most essential properties of floating point numbers which are needed for the implementation of floating point interval arithmetic. Thus, it is for example possible to implement floating point numbers with arbitrary bitlength and internal representation, which still conform to the definitions.

For those who are logically oriented, we remark that from now on we use the language of first order predicate calculus with equality and the usual axioms about the membership predicate $\in$ (some part of ZF set theory).

Definition 2.2.1 (Floating Point Numbers) *Let $\mathbb{F}_o$ be a finite set with odd cardinality. Let*

$$\mathbb{F} = \mathbb{F}_o \cup \{\top, \bot\},$$

where $\top \neq \bot$ and $\top, \bot \notin \mathbb{F}_o$. Finally, let

$$\mathbb{F}_\sharp = \mathbb{F} \cup \{\sharp\},$$

where $\sharp \notin \mathbb{F}$. □

$\mathbb{F}$ is called the set of floating point numbers, $\mathbb{F}_o$ is called the set of ordinary floating point numbers and $\mathbb{F}_\sharp$ is called the set of generalized floating point numbers. In the following we use a, b to denote floating point numbers and x, y to denote real numbers.

Definition 2.2.2 (Interpretation of Floating Point Numbers) *Let $\phi : \mathbb{F}_o \to \mathbb{R}$ such that*

- *$\phi(a) \neq \phi(b)$ if $a \neq b$ for all $a, b \in \mathbb{F}_o$*
- *$\phi(a) = 0$ for some $a \in \mathbb{F}_o$.*
- *for every $a \in \mathbb{F}_o$ there exists $b \in \mathbb{F}_o$ such that $-\phi(a) = \phi(b)$.* □

We do not interpret $\top$ and $\bot$ as ∞ and $-\infty$, because this would lead to confusion when these numbers occur as endpoint of intervals as described in the next section. For notational convenience we denote the uniquely defined element of $\mathbb{F}_o$, which is mapped to 0 under ϕ by 0.

Definition 2.2.3 (Predicates on Floating Point Numbers) *The predicates $<, >, \leq, \geq$ are defined on $\mathbb{F}$ as follows:*

$$\begin{array}{lll} a < b & \text{iff} & (a, b \in \mathbb{F}_o, \phi(a) < \phi(b)) \text{ or } (a \in \mathbb{F}_o, b = \top) \text{ or } (a = \bot, b \in \mathbb{F}_o) \text{ or } (a = \bot, b = \top) \\ a > b & \text{iff} & b < a \\ a \leq b & \text{iff} & a < b \text{ or } a = b \\ a \geq b & \text{iff} & a > b \text{ or } a = b. \end{array}$$ □

Note that $\leq, \geq$ are total orderings of $\mathbb{F}_o$, hence $\max \mathbb{F}_o$ and $\min \mathbb{F}_o$ are well defined.

Definition 2.2.4 (Rounding Functions) *The rounding functions $\breve{\rho}, \hat{\rho}, \check{\rho}: \mathbb{R} \to \mathbb{F}$ with intended meaning round to nearest, round up and round down respectively, are defined such that for all $x \in \mathbb{R}$*

- *if* $x > \phi(\max \mathbb{F}_o)$ *then*

$$\begin{aligned} \tilde{\tilde{\rho}}\,(x) &= \top \\ \check{\rho}(x) &= \max \mathbb{F}_o \\ \hat{\rho}(x) &= \top \end{aligned}$$

- *if* $x < \phi(\min \mathbb{F}_o)$ *then*

$$\begin{aligned} \tilde{\tilde{\rho}}\,(x) &= \bot \\ \check{\rho}(x) &= \bot \\ \hat{\rho}(x) &= \min \mathbb{F}_o \end{aligned}$$

- *otherwise,* $\tilde{\tilde{\rho}}\,(x), \hat{\rho}(x), \check{\rho}(x) \in \mathbb{F}_o$ *and for all* $a \in \mathbb{F}_o$

$$\begin{aligned} |x - \phi(\tilde{\tilde{\rho}}\,(x))| \leq |x - \phi(a)| \quad & \text{(nearest)} \\ \phi(\check{\rho}(x)) \leq x \text{ and } \big(\phi(a) > x \text{ or } a \leq \check{\rho}(x)\big) \quad & \text{(down, optimal)} \\ \phi(\hat{\rho}(x)) \geq x \text{ and } \big(\phi(a) < x \text{ or } a \geq \hat{\rho}(x)\big) \quad & \text{(up, optimal).} \ \square \end{aligned}$$

The following lemmas are immediate consequences of Definition 2.2.4.

Lemma 2.2.5 *For all* $x \in \mathbb{R}$ *it holds that*

$$\begin{aligned} &\text{if } \check{\rho}(x) \in \mathbb{F}_o \quad \text{then} \quad \phi(\check{\rho}(x)) \leq x \\ &\text{if } \hat{\rho}(x) \in \mathbb{F}_o \quad \text{then} \quad \phi(\hat{\rho}(x)) \geq x. \ \square \end{aligned}$$

Lemma 2.2.6 *For all* $x \in \mathbb{R}$ *it holds that*

$$\begin{aligned} &\text{if } x > 0 \quad \text{then} \quad \check{\rho}(x) \in \mathbb{F}_o \\ &\text{if } x < 0 \quad \text{then} \quad \hat{\rho}(x) \in \mathbb{F}_o. \ \square \end{aligned}$$

Lemma 2.2.7 *For all* $x \in \mathbb{R}$ *it holds that*

$$\begin{aligned} &\text{if } x < 0 \quad \text{then} \quad \check{\rho}(x) < 0 \\ &\text{if } x > 0 \quad \text{then} \quad \hat{\rho}(x) > 0. \ \square \end{aligned}$$

Lemma 2.2.8 *For all* $x, y \in \mathbb{R}$ *and for all* $\rho \in \{\tilde{\tilde{\rho}}, \hat{\rho}, \check{\rho}\}$ *it holds that*

$$\begin{aligned} &\text{if } x \leq y \quad \text{then} \quad \rho(x) \leq \rho(y) \\ &\text{if } x \geq y \quad \text{then} \quad \rho(x) \geq \rho(y). \ \square \end{aligned}$$

Remark. The inversion of Lemma 2.2.8 does *not* hold. $\square$

Lemma 2.2.9 *For all finite* $\mathcal{X} \subseteq \mathbb{R}$ *and for all* $\rho \in \{\tilde{\tilde{\rho}}, \hat{\rho}, \check{\rho}\}$ *it holds that*

$$\begin{aligned} \min\{\rho(x) \mid x \in \mathcal{X}\} &= \rho(\min \mathcal{X}) \\ \max\{\rho(x) \mid x \in \mathcal{X}\} &= \rho(\max \mathcal{X}). \ \square \end{aligned}$$

Lemma 2.2.10 *For all* $a \in \mathbb{F}_o$ *and for all* $\rho \in \{\tilde{\tilde{\rho}}, \hat{\rho}, \check{\rho}\}$ *it holds that*

$$\rho(\phi(a)) = a. \ \square$$

In particular, $\tilde{\tilde{\rho}}\,(0) = \hat{\rho}(0) = \check{\rho}(0) = 0$.

Definition 2.2.11 (Floating Point Number Arithmetic) *The direction rounded arithmetic operations* $\overset{\approx}{+}$, $\overset{\wedge}{+}$, $\overset{\vee}{+}$, $\overset{\approx}{-}$, $\overset{\wedge}{-}$, $\overset{\vee}{-}$, $\overset{\approx}{*}$, $\overset{\wedge}{*}$, $\overset{\vee}{*}$, $\overset{\approx}{/}$, $\overset{\wedge}{/}$, $\overset{\vee}{/}$, $\mathbb{F}_\sharp \times \mathbb{F}_\sharp \to \mathbb{F}_\sharp$ *and* $- : \mathbb{F}_\sharp \to \mathbb{F}_\sharp$ *are defined as follows:*

$$a \overset{\approx}{+} b = \begin{cases} \sharp & \text{if } a = \sharp \text{ or } b = \sharp \text{ or } (a = \top, b = \bot) \text{ or } (a = \bot, b = \top) \\ \top & \text{if } (a = \top, b \notin \{\bot, \sharp\}) \text{ or } (b = \top, a \notin \{\bot, \sharp\}) \\ \bot & \text{if } (a = \bot, b \notin \{\top, \sharp\}) \text{ or } (b = \bot, a \notin \{\top, \sharp\}) \\ \overset{\approx}{\rho}(\phi(a) + \phi(b)) & \text{if } a, b \in \mathbb{F}_o \end{cases}$$

$$a \overset{\wedge}{+} b = \begin{cases} a \overset{\approx}{+} b & \text{if } a \notin \mathbb{F}_o \text{ or } b \notin \mathbb{F}_o \\ \overset{\wedge}{\rho}(\phi(a) + \phi(b)) & \text{if } a, b \in \mathbb{F}_o \end{cases}$$

$$a \overset{\vee}{+} b = \begin{cases} a \overset{\approx}{+} b & \text{if } a \notin \mathbb{F}_o \text{ or } b \notin \mathbb{F}_o \\ \overset{\vee}{\rho}(\phi(a) + \phi(b)) & \text{if } a, b \in \mathbb{F}_o \end{cases}$$

$$a \overset{\approx}{-} b = \begin{cases} \sharp & \text{if } a = \sharp \text{ or } b = \sharp \text{ or } (a = \top, b = \top) \text{ or } (a = \bot, b = \bot) \\ \top & \text{if } (a = \top, b \notin \{\top, \sharp\}) \text{ or } (b = \bot, a \notin \{\bot, \sharp\}) \\ \bot & \text{if } (a = \bot, b \notin \{\bot, \sharp\}) \text{ or } (b = \top, a \notin \{\top, \sharp\}) \\ \overset{\approx}{\rho}(\phi(a) - \phi(b)) & \text{if } a, b \in \mathbb{F}_o \end{cases}$$

$$a \overset{\wedge}{-} b = \begin{cases} a \overset{\approx}{-} b & \text{if } a \notin \mathbb{F}_o \text{ or } b \notin \mathbb{F}_o \\ \overset{\wedge}{\rho}(\phi(a) - \phi(b)) & \text{if } a, b \in \mathbb{F}_o \end{cases}$$

$$a \overset{\vee}{-} b = \begin{cases} a \overset{\approx}{-} b & \text{if } a \notin \mathbb{F}_o \text{ or } b \notin \mathbb{F}_o \\ \overset{\vee}{\rho}(\phi(a) - \phi(b)) & \text{if } a, b \in \mathbb{F}_o \end{cases}$$

$$a \overset{\approx}{*} b = \begin{cases} \sharp & \text{if } a = \sharp \text{ or } b = \sharp \text{ or } (a \in \{\top, \bot\}, b = 0) \text{ or } (a = 0, b \in \{\top, \bot\}) \\ \top & \text{if } (a = \top, b > 0) \text{ or } (a = \bot, b < 0) \text{ or } (b = \top, a > 0) \text{ or } (b = \bot, a < 0) \\ \bot & \text{if } (a = \top, b < 0) \text{ or } (a = \bot, b > 0) \text{ or } (b = \top, a < 0) \text{ or } (b = \bot, a > 0) \\ \overset{\approx}{\rho}(\phi(a) * \phi(b)) & \text{if } a, b \in \mathbb{F}_o \end{cases}$$

$$a \overset{\wedge}{*} b = \begin{cases} a \overset{\approx}{*} b & \text{if } a \notin \mathbb{F}_o \text{ or } b \notin \mathbb{F}_o \\ \overset{\wedge}{\rho}(\phi(a) * \phi(b)) & \text{if } a, b \in \mathbb{F}_o \end{cases}$$

$$a \overset{\vee}{*} b = \begin{cases} a \overset{\approx}{*} b & \text{if } a \notin \mathbb{F}_o \text{ or } b \notin \mathbb{F}_o \\ \overset{\vee}{\rho}(\phi(a) * \phi(b)) & \text{if } a, b \in \mathbb{F}_o \end{cases}$$

$$a \overset{\approx}{/} b = \begin{cases} \sharp & \text{if } a = \sharp \text{ or } b = \sharp \text{ or } a = b = \top \text{ or } a = b = \bot \\ \top & \text{if } (a = \top, \top > b > 0) \text{ or } (a = \bot, \bot < b < 0) \\ \bot & \text{if } (a = \top, \bot < b < 0) \text{ or } (a = \bot, \top > b > 0) \\ \text{undefined} & \text{if } a \in \mathbb{F}, b = 0 \\ \overset{\approx}{\rho}(\phi(a)/\phi(b)) & \text{if } a, b \in \mathbb{F}_o,\ b \neq 0 \end{cases}$$

$$a \overset{\wedge}{/} b = \begin{cases} a \overset{\approx}{/} b & \text{if } a \notin \mathbb{F}_o \text{ or } b \notin \mathbb{F}_o \text{ or } b = 0 \\ \overset{\wedge}{\rho}(\phi(a)/\phi(b)) & \text{if } a, b \in \mathbb{F}_o, b \neq 0 \end{cases}$$

$$a \overset{\vee}{/} b = \begin{cases} a \overset{\approx}{/} b & \text{if } a \notin \mathbb{F}_o \text{ or } b \notin \mathbb{F}_o \text{ or } b = 0 \\ \overset{\vee}{\rho}(\phi(a)/\phi(b)) & \text{if } a, b \in \mathbb{F}_o, b \neq 0 \end{cases}$$

$$-a = \begin{cases} \sharp & \text{if } a = \sharp \\ \top & \text{if } a = \bot \\ \bot & \text{if } a = \top \\ \overset{\approx}{\rho}(-\phi(a)) & \text{if } a \in \mathbb{F}_o \end{cases}$$

Note that according to Definition 2.2.2 for all $a \in \mathbb{F}_o$ there exists $b \in \mathbb{F}_o$ such that $-\phi(a) = \phi(b)$, and by Lemma 2.2.10

$$\overset{\approx}{\rho}(-\phi(a)) = \overset{\vee}{\rho}(-\phi(a)) = \overset{\wedge}{\rho}(-\phi(a)). \tag{2.2.1}$$

2.2.1 Basic Algorithms on Floating Point Numbers

Algorithms for floating point number arithmetic are usually implemented in hardware and we do not discuss them here. On most machines algorithms for floating point approximations of elementary transcendental functions are available, but it is usually not specified whether the returned floating point number is greater or less than the exact result. In many cases, not even an upper bound for the approximation error is given, hence we cannot use these algorithms for implementing interval arithmetic correctly.

Apart from arithmetic functions we need in this thesis only exponentiation by a natural number. Usually, the exponent will not be large, hence the well–known binary method for exponentiation seems suitable, see e.g. [Knuth, 1981]. The following two algorithms compute upper respectively lower approximations.

Algorithm 2.2.12 ($\hat{\text{POW}}$) [Upper Approximation of Floating Point Number Power]

In: $a \in \mathbb{F}$,
$n \in \mathbb{N}_0$.

Out: $b \in \mathbb{F}$, $b = \hat{\rho}(1)$ if $n = 0$, $b = a$ if $n > 0$ and $a \notin \mathbb{F}_o$, otherwise $b = \top$ or ($b \in \mathbb{F}_o$ and $\phi(b) \geq \phi(a)^n$).

(1) [Case $a < 0$.]
if n is odd and $a < 0$ then return $-\check{\text{POW}}(-a, n)$.
$a \longleftarrow |a|$.

(2) [Initialize.]
$b \longleftarrow \hat{\rho}(1)$.
if $n = 0$ return b.

(3) [Iterate.]
if n is odd then
$b \longleftarrow b \mathbin{\hat{*}} a$
if $n = 1$ return b, else $n \longleftarrow n - 1$.
$a \longleftarrow a \mathbin{\hat{*}} a$.
$n \longleftarrow n/2$.
goto Step 3.

Algorithm 2.2.13 ($\check{\text{POW}}$) [Lower Approximation of Floating Point Number Power]

In: $a \in \mathbb{F}$,
$n \in \mathbb{N}_0$.

Out: $b \in \mathbb{F}$, $b = \check{\rho}(1)$ if $n = 0$, $b = a$ if $n > 0$ and $a \notin \mathbb{F}_o$, otherwise $b = \bot$ or ($b \in \mathbb{F}_o$ and $\phi(b) \leq \phi(a)^n$).

(1) [Case $a < 0$.]
if n is odd and $a < 0$ then return $-\hat{\text{POW}}(-a, n)$.
$a \longleftarrow |a|$.

(2) [Initialize.]
$b \longleftarrow \check{\rho}(1)$.
if $n = 0$ return b.

(3) [Iterate.]
if n is odd then
$b \longleftarrow b \mathbin{\check{*}} a$
if $n = 1$ return b, else $n \longleftarrow n - 1$.
$a \longleftarrow a \mathbin{\check{*}} a$.
$n \longleftarrow n/2$.
goto Step 3.

Remark. For the correctness of Algorithms 2.2.12 and 2.2.13 we have to assume that $\hat{\rho}(1) \neq \top$ and $\check{\rho}(1) \neq 0$. Otherwise $\hat{\mathrm{POW}}(0,n) = \sharp$ and $\check{\mathrm{POW}}(\top,n) = \sharp$ for $n > 0$. In the sequel we assume $\hat{\rho}(1) \neq \top$ and $\check{\rho}(1) \neq 0$. □

Theorem 2.2.14 (Complexity) *Algorithm 2.2.12 respectively 2.2.13 costs* $\lfloor \mathrm{ld}(n) \rfloor + \nu(n)$ *floating point number multiplications, where* $\nu(n)$ *is the number of ones in the binary representation of* n. □

Proof. See [Knuth, 1981]. □

2.3 Floating Point Interval Arithmetic

Floating point intervals are closed real intervals, where the endpoints are ordinary floating point numbers. Arithmetic is defined in the set theoretic sense where the result is rounded outwards. In order to handle overflow, we extend the set of floating point intervals by half–open intervals $\{x \mid x \geq y\}$ and $\{x \mid x \leq y\}$ for all floating point numbers y, and by the set $\mathbb{R}$ of all real numbers.

In this section we define floating point intervals and floating point interval arithmetic formally. The definitions are motivated by the following goals:

- Correctness. A floating point interval operation must yield an interval, which contains the result interval of the corresponding real interval operation.
- Completeness. Floating point interval arithmetic functions must be defined whenever the corresponding real interval arithmetic function is defined. In particular, the computation must not be interrupted because of overflow.
- Efficiency. An efficient implementation of the floating point interval arithmetic must be possible. In particular, the IEEE 754 floating point arithmetic presented in the previous section must be used because of its hardware support.

Before going into details of floating point interval arithmetic, we want to point out an important problem. A floating point interval is a pair of floating point numbers, where the first component is less than or equal to the second component. If overflow occurs during some operation on floating point intervals, we obtain an interval where one or both components are $\top$ or $\bot$. While floating point intervals, where both components are ordinary floating point numbers, are interpreted as elements of $\mathbb{IR}$ in the obvious way, it is not clear, how floating point intervals, where at least one component is $\top$ or $\bot$, should be interpreted. For correctness reasons they cannot be interpreted as elements of $\mathbb{IR}$, hence we have to extend the set $\mathbb{IR}$ in a suitable way. There are two possibilities:

- Closed subintervals of $\mathbb{R} \cup \{\infty, -\infty\}$. The floating point number interpretation ϕ is extended by $\phi(\top) = \infty$ and $\phi(\bot) = -\infty$. The floating point interval $(a,b) \in \mathbb{F} \times \mathbb{F}$ is then interpreted simply as $[\phi(a), \phi(b)]$. The disadvantage of this approach is that all arithmetic operations are partial functions. For example $[3,\infty] * [0,0]$ or $[\infty,\infty] - [\infty,3]$ are not defined. This means that even the evaluation of arithmetic terms over $\mathbb{IR}$, which do not contain division, might not be possible when floating point intervals are used and overflow occurs.
- Extend $\mathbb{IR}$ by the sets
$$\left.\begin{array}{l} \{x \in \mathbb{R} \mid x \geq y\} \\ \{x \in \mathbb{R} \mid x \leq y\} \end{array}\right\} \text{for all } y \in \mathbb{R}$$
and by the set of all real numbers $\mathbb{R}$. The floating point interval $(\bot, \top)$ is interpreted as $\mathbb{R}$. If a is an ordinary floating point number, then $(a, \top)$ is interpreted as $\{x \in \mathbb{R} \mid x \geq \phi(a)\}$ and $(\bot, a)$ is interpreted as $\{x \in \mathbb{R} \mid x \leq \phi(a)\}$. Note that $(\top, \top)$ and $(\bot, \bot)$ are not floating point intervals. This approach seems to be advantageous, because it avoids infinity arithmetic completely and floating point interval addition, subtraction and multiplication are total functions. A problem arises, because the floating point multiplication function as defined by the IEEE standard must

be modified slightly for the implementation of floating point interval multiplication. For example, the result of $(\bot, \top) * (0,0)$ must be $(0,0)$, but a straight forward implementation of floating point interval multiplication using the IEEE floating point number multiplication would yield $(\sharp, \sharp)$.

In our implementations we use the second approach, which is described more formally as follows.

Definition 2.3.1 (Extended Intervals) *The set of extended intervals is defined as*

$$\overline{\mathbb{IR}} = \mathbb{IR} \cup \{\{x \in \mathbb{R} \mid x \geq y\} \mid y \in \mathbb{R}\} \cup \{\{x \in \mathbb{R} \mid x \leq y\} \mid y \in \mathbb{R}\} \cup \{\mathbb{R}\}$$

Further,

$$\begin{aligned} \mathbb{IR}_\emptyset &= \mathbb{IR} \cup \emptyset \\ \overline{\mathbb{IR}}_\emptyset &= \overline{\mathbb{IR}} \cup \emptyset. \ \square \end{aligned}$$

Throughout this section, elements of $\overline{\mathbb{IR}}$ are simply called intervals. An interval $X \in \overline{\mathbb{IR}}$ is called left bounded, if there exists $y \in X$ such that $x \geq y$ for all $x \in X$. In this case we define $\underline{X} = y$. Otherwise, X is called left unbounded. Similarly, X is called right bounded if there exists $y \in X$ such that $x \leq y$ for all $x \in X$. In this case we define $\overline{X} = y$. Otherwise, X is called right unbounded.

The interval arithmetic operations are extended from $\mathbb{IR}$ to $\overline{\mathbb{IR}}$ in the straight forward way.

Definition 2.3.2 (Extended Interval Arithmetic) *Let* $X, Y \in \overline{\mathbb{IR}}$.

$$\begin{aligned} X + Y &= \{x + y \mid x \in X, y \in Y\} \\ X - Y &= \{x - y \mid x \in X, y \in Y\} \\ X * Y &= \{x * y \mid x \in X, y \in Y\} \\ X/Y &= \{x/y \mid x \in X, y \in Y\} \text{ if } 0 \notin Y \\ -X &= \{-x \mid x \in X\}. \ \square \end{aligned}$$

Definition 2.3.3 (Floating Point Intervals) *The set of floating point intervals is defined as*

$$\mathbb{IF} = \{(a,b) \mid a, b \in \mathbb{F}_o, a \leq b\} \cup \{(\bot, a) \mid a \in \mathbb{F}_o\} \cup \{(a, \top) \mid a \in \mathbb{F}_o\} \cup \{(\bot, \top)\}$$

Further,

$$\mathbb{IF}_\emptyset = \mathbb{IF} \cup \{(\sharp, \sharp)\}. \ \square$$

Note that $(\bot, \bot), (\top, \top) \notin \mathbb{IF}$. The first and the second component of a floating point interval $A \in \mathbb{IF}$ is denoted by $\underline{A}$ respectively $\overline{A}$. In the following we will use A, B to denote floating point intervals and X, Y to denote real intervals.

Definition 2.3.4 (Interpretation of Floating Point Intervals) *Let* $\psi : \mathbb{IF}_\emptyset \to \overline{\mathbb{IR}}_\emptyset$ *such that*

$$\begin{aligned} \psi(a,b) &= [\phi(a), \phi(b)] && \text{if } a, b \in \mathbb{F}_o, \\ \psi(\bot, b) &= \{x \in \mathbb{R} \mid x \leq \phi(b)\} && \text{if } b \in \mathbb{F}_o, \\ \psi(a, \top) &= \{x \in \mathbb{R} \mid x \geq \phi(a)\} && \text{if } a \in \mathbb{F}_o, \\ \psi(\bot, \top) &= \mathbb{R} \\ \psi(\sharp, \sharp) &= \emptyset. \ \square \end{aligned}$$

Definition 2.3.5 (Rounding Function for Intervals) *Let* $\sigma : \overline{\mathbb{IR}}_\emptyset \to \mathbb{IF}_\emptyset$ *such that*

$$\sigma(X) = \begin{cases} (\check{\rho}(\underline{X}), \hat{\rho}(\overline{X})) & \text{if } X \in \mathbb{IR} \\ (\check{\rho}(y), \top) & \text{if } X = \{x \in \mathbb{R} \mid x \geq y\} \quad \text{for some } y \in \mathbb{R} \\ (\bot, \hat{\rho}(y)) & \text{if } X = \{x \in \mathbb{R} \mid x \leq y\} \quad \text{for some } y \in \mathbb{R} \\ (\bot, \top) & \text{if } X = \mathbb{R} \\ (\sharp, \sharp) & \text{if } X = \emptyset \ . \ \square \end{cases}$$

Definition 2.3.6 (Predicates on Floating Point Intervals) *Let* $\sim \in \{\subset, \subseteq, \supset, \supseteq\}$. *Then for all* $A, B \in \mathbb{IF}_\emptyset$

$$A \sim B \text{ iff } \psi(A) \sim \psi(B).$$

For all $a \in \mathbb{F}$, $A \in \mathbb{IF}_\emptyset$

$$a \in A \text{ iff } \begin{cases} \phi(a) \in \psi(A) & \text{if } a \in \mathbb{F}_o, \\ a = \underline{A} \text{ or } a = \overline{A} & \text{else. } \square \end{cases}$$

Lemma 2.3.7

- *For all* $X \in \mathbb{IR}_\emptyset$ *it holds that*
$$\psi(\sigma(X)) \supseteq X.$$
- *For all* $A \in \mathbb{IF}_\emptyset$ *it holds that*
$$\sigma(\psi(A)) = A. \ \square$$

Proof. Follows from Definition 2.3.4, Definition 2.3.5 and Lemma 2.2.5. $\square$

Before defining floating point interval arithmetic, we modify the floating point multiplication in order to facilitate the definition of the interval multiplication. If the IEEE standard would define floating point multiplication in this modified way, then interval arithmetic could be implemented more efficiently. However, the modified multiplication is not suitable for implementing infinity arithmetic, and it seems that the IEEE committee considered this aspect as more important.

Definition 2.3.8 (Modified Floating Point Multiplication) *The modified direction rounded floating point multiplications* $\overset{\wedge}{\cdot}, \overset{\vee}{\cdot} : \mathbb{F}_\sharp \times \mathbb{F}_\sharp \to \mathbb{F}_\sharp$ *are defined as*

$$a \overset{\wedge}{\cdot} b = \begin{cases} 0 & \text{if } (a = 0, b \in \{\top, \bot\}) \text{ or } (b = 0, a \in \{\top, \bot\}) \\ a \overset{\wedge}{*} b & \text{else} \end{cases}$$

$$a \overset{\vee}{\cdot} b = \begin{cases} 0 & \text{if } (a = 0, b \in \{\top, \bot\}) \text{ or } (b = 0, a \in \{\top, \bot\}) \\ a \overset{\vee}{*} b & \text{else. } \square \end{cases}$$

Definition 2.3.9 (Floating Point Interval Arithmetic) *The arithmetic operations* $+, -, *, / : \mathbb{IF} \times \mathbb{IF} \to \mathbb{IF}$ *and* $- : \mathbb{IF} \to \mathbb{IF}$ *are defined as follows:*

$$\begin{aligned} A + B &= (\underline{A} \overset{\vee}{+} \underline{B}, \overline{A} \overset{\wedge}{+} \overline{B}) \\ A - B &= (\underline{A} \overset{\vee}{-} \overline{B}, \overline{A} \overset{\wedge}{-} \underline{B}) \\ A * B &= (\min\{\underline{A} \overset{\vee}{\cdot} \underline{B}, \underline{A} \overset{\vee}{\cdot} \overline{B}, \overline{A} \overset{\vee}{\cdot} \underline{B}, \overline{A} \overset{\vee}{\cdot} \overline{B}\}, \max\{\underline{A} \overset{\wedge}{\cdot} \underline{B}, \underline{A} \overset{\wedge}{\cdot} \overline{B}, \overline{A} \overset{\wedge}{\cdot} \underline{B}, \overline{A} \overset{\wedge}{\cdot} \overline{B}\}) \\ A/B &= (\min\{\underline{A} \overset{\vee}{/} \underline{B}, \underline{A} \overset{\vee}{/} \overline{B}, \overline{A} \overset{\vee}{/} \underline{B}, \overline{A} \overset{\vee}{/} \overline{B}\}, \max\{\underline{A} \overset{\wedge}{/} \underline{B}, \underline{A} \overset{\wedge}{/} \overline{B}, \overline{A} \overset{\wedge}{/} \underline{B}, \overline{A} \overset{\wedge}{/} \overline{B}\}) \text{ if } 0 \notin B \\ -A &= (-\overline{A}, -\underline{A}). \ \square \end{aligned}$$

The following theorem states that the floating point interval arithmetic operations are performed as if they first produced a correct intermediate result which is then rounded outwards to a floating point interval.

Theorem 2.3.10 *For all* $A, B \in \mathbb{IF}$ *it holds that*

$$A + B = \sigma(\psi(A) + \psi(B)) \tag{2.3.1}$$
$$A - B = \sigma(\psi(A) - \psi(B)) \tag{2.3.2}$$
$$A * B = \sigma(\psi(A) * \psi(B)) \tag{2.3.3}$$
$$A/B = \sigma(\psi(A)/\psi(B)) \text{ if } 0 \notin B \tag{2.3.4}$$
$$-A = \sigma(-\psi(A)). \ \square \tag{2.3.5}$$

Proof. Let $A, B \in \mathbb{IF}$ arbitrary but fixed. We show that the left endpoints of the floating point intervals in (2.3.1) – (2.3.5) coincide. The proof for the right endpoints is equivalent.

- Proof of (2.3.1).
  - Assume $\underline{A} \neq \bot$, $\underline{B} \neq \bot$. Then $\psi(A), \psi(B)$ and $\psi(A) + \psi(B)$ are left bounded.

$$\begin{array}{rcl} \underline{A+B} & \overset{\text{Definition 2.3.9}}{=} & \underline{A} \overset{\vee}{+} \underline{B} \\ & \overset{\text{Definition 2.2.11}}{=} & \overset{\vee}{\rho}(\phi(\underline{A}) + \phi(\underline{B})) \\ & \overset{\text{Definition 2.3.4}}{=} & \overset{\vee}{\rho}(\underline{\psi(A)} + \underline{\psi(B)}) \\ & \overset{\text{Theorem 1.1.1}}{=} & \overset{\vee}{\rho}(\underline{\psi(A) + \psi(B)}) \\ & \overset{\text{Definition 2.3.5}}{=} & \underline{\sigma(\psi(A) + \psi(B))}. \end{array}$$

  - Assume $\underline{A} = \bot$ or $\underline{B} = \bot$. Then

$$\underline{A+B} = \bot,$$

    $\psi(A) + \psi(B)$ is left unbounded, and

$$\underline{\sigma(\psi(A) + \psi(B))} = \bot.$$

- Proof of (2.3.2). Analogous to the proof of (2.3.1).
- Proof of (2.3.3).
  - Assume the condition

$$(\underline{A} = \bot, \overline{B} > 0) \text{ or } (\overline{A} = \top, \underline{B} < 0) \text{ or } (\underline{B} = \bot, \overline{A} > 0) \text{ or } (\overline{B} = \top, \underline{A} < 0) \tag{2.3.6}$$

    does not hold. First, we eliminate some trivial cases. If $A = (\bot, \top)$ then $B = (0,0)$, $\psi(A) = \mathbb{R}$, $\psi(B) = 0$. Similarly, if $B = (\bot, \top)$ then $A = (0,0)$, $\psi(B) = \mathbb{R}$, $\psi(A) = 0$. In both cases

$$\underline{A * B} = 0 = \underline{\sigma(\psi(A) * \psi(B))}.$$

    In the sequel assume $A \neq (\bot, \top)$, $B \neq (\bot, \top)$. Next, we show that

$$\underline{A * B} = \min\{a \overset{\vee}{\cdot} b \mid a \in \{\overline{A}, \underline{A}\} \cap \mathbb{F}_o, b \in \{\overline{B}, \underline{B}\} \cap \mathbb{F}_o\} \tag{2.3.7}$$

    According to Definition 2.3.9 there exist $a \in \{\overline{A}, \underline{A}\}$, $b \in \{\overline{B}, \underline{B}\}$ such that $a \overset{\vee}{\cdot} b = \underline{A * B}$. If $a, b \in \mathbb{F}_o$ then (2.3.7) follows from Definition 2.3.9. The other cases are as follows:
    - If $a = \bot$ then $b \leq 0$, $\overline{A} \in \mathbb{F}_o$ and $\overline{A} \overset{\vee}{\cdot} b \leq a \overset{\vee}{\cdot} b$.
    - If $a = \top$ then $b \geq 0$, $\underline{A} \in \mathbb{F}_o$ and $\underline{A} \overset{\vee}{\cdot} b \leq a \overset{\vee}{\cdot} b$.
    - If $b = \bot$ then $a \leq 0$, $\overline{B} \in \mathbb{F}_o$ and $a \overset{\vee}{\cdot} \overline{B} \leq a \overset{\vee}{\cdot} b$.
    - If $b = \top$ then $a \geq 0$, $\underline{B} \in \mathbb{F}_o$ and $a \overset{\vee}{\cdot} \underline{B} \leq a \overset{\vee}{\cdot} b$.

    Finally, we show that $\psi(A) * \psi(B)$ is left bounded and

$$\underline{\psi(A) * \psi(B)} = \min\{\phi(a) * \phi(b) \mid a \in \{\overline{A}, \underline{A}\} \cap \mathbb{F}_o, b \in \{\overline{B}, \underline{B}\} \cap \mathbb{F}_o\} \tag{2.3.8}$$

    - Assume $\overline{A}, \underline{A}, \overline{B}, \underline{B} \in \mathbb{F}_o$. Then $\psi(A), \psi(B)$ are right and left bounded and

$$\begin{array}{rcl} \underline{\psi(A) * \psi(B)} & \overset{\text{Theorem 1.1.1}}{=} & \min\{x * y \mid x \in \{\underline{\psi(A)}, \overline{\psi(A)}\}, y \in \{\underline{\psi(B)}, \overline{\psi(B)}\}\} \\ & \overset{\text{Definition 2.3.4}}{=} & \min\{x * y \mid x \in \{\phi(\underline{A}), \phi(\overline{A})\}, y \in \{\phi(\underline{B}), \phi(\overline{B})\}\} \\ & = & \min\{\phi(a) * \phi(b) \mid a \in \{\underline{A}, \overline{A}\}, b \in \{\underline{B}, \overline{B}\}\}. \end{array}$$

    - Assume $\underline{A} = \bot$. Then $\overline{B} \leq 0$, hence $\overline{B} \in \mathbb{F}_o$. Further, if $\overline{A} > 0$ then $\underline{B} \in \mathbb{F}_o$.

$$\begin{array}{rcl} \underline{\psi(A) * \psi(B)} & = & \begin{cases} \phi(\overline{A}) * \phi(\overline{B}) & \text{if } \overline{A} \leq 0 \\ \phi(\overline{A}) * \phi(\underline{B}) & \text{else} \end{cases} \\ & = & \min\{\phi(a) * \phi(b) \mid a \in \{\overline{A}, \underline{A}\} \cap \mathbb{F}_o, b \in \{\overline{B}, \underline{B}\} \cap \mathbb{F}_o\}. \end{array}$$

· Assume $\overline{A} = \top$. Then $\underline{B} \geq 0$, hence $\underline{B} \in \mathbb{F}_o$. Further, if $\underline{A} < 0$ then $\overline{B} \in \mathbb{F}_o$.

$$\begin{aligned}\underline{\psi(A) * \psi(B)} &= \begin{cases} \phi(\underline{A}) * \phi(\underline{B}) & \text{if } \underline{A} \geq 0 \\ \phi(\underline{A}) * \phi(\overline{B}) & \text{else} \end{cases} \\ &= \min\{\phi(a) * \phi(b) \mid a \in \{\overline{A}, \underline{A}\} \cap \mathbb{F}_o, b \in \{\overline{B}, \underline{B}\} \cap \mathbb{F}_o\}.\end{aligned}$$

· The cases $\underline{B} = \bot$ and $\overline{B} = \top$ can be reduced to the cases $\underline{A} = \bot$ respectively $\overline{A} = \top$ by using commutativity of real and interval multiplication.

Thus,

$$\begin{aligned}\underline{A * B} &\overset{(2.3.7)}{=} \min\{a \overset{\vee}{*} b \mid a \in \{\overline{A}, \underline{A}\} \cap \mathbb{F}_o, b \in \{\overline{B}, \underline{B}\} \cap \mathbb{F}_o\} \\ &\overset{\text{(Definition 2.2.11)}}{=} \min\{\overset{\vee}{\rho}(\phi(a) * \phi(b)) \mid a \in \{\overline{A}, \underline{A}\} \cap \mathbb{F}_o, b \in \{\overline{B}, \underline{B}\} \cap \mathbb{F}_o\} \\ &\overset{\text{(Lemma 2.2.9)}}{=} \overset{\vee}{\rho}(\min\{\phi(a) * \phi(b) \mid a \in \{\overline{A}, \underline{A}\} \cap \mathbb{F}_o, b \in \{\overline{B}, \underline{B}\} \cap \mathbb{F}_o\}) \\ &\overset{(2.3.8)}{=} \overset{\vee}{\rho}(\underline{\psi(A) * \psi(B)}) \\ &\overset{\text{(Definition 2.3.5)}}{=} \underline{\sigma(\psi(A) * \psi(B))}.\end{aligned}$$

– Assume (2.3.6) holds. Then

$$\underline{A * B} = \bot,$$

$\psi(A) * \psi(B)$ is left unbounded, and hence

$$\underline{\sigma(\psi(A) * \psi(B))} = \bot.$$

- Proof of (2.3.4). Analogous to the proof of (2.3.3).
- Proof of (2.3.5).
  – Assume $\overline{A} \in \mathbb{F}_o$.

$$\begin{aligned}\underline{-A} &\overset{\text{Definition 2.3.9}}{=} -\overline{A} \\ &\overset{\text{Definition 2.2.11}}{=} \overset{\approx}{\rho}(-\phi(\overline{A})) \\ &\overset{(2.2.1)}{=} \overset{\vee}{\rho}(-\phi(\overline{A})) \\ &\overset{\text{Definition 2.3.4}}{=} \overset{\vee}{\rho}(-\overline{\psi(A)}) \\ &\overset{\text{Theorem 1.1.1}}{=} \overset{\vee}{\rho}(\underline{-\psi(A)}) \\ &\overset{\text{Definition 2.3.5}}{=} \underline{\sigma(-\psi(A))}.\end{aligned}$$

  – Assume $\overline{A} \notin \mathbb{F}_o$. Then $\overline{A} = \top$ and

$$\underline{-A} = \bot.$$

  Further, $\psi(A)$ is right unbounded, hence $-\psi(A)$ is left unbounded and

$$\underline{\sigma(-\psi(A))} = \bot. \ \square$$

In the following, for $\boldsymbol{A} \in \mathbb{IF}^n$ we define

$$\sigma(\boldsymbol{A}) = (\sigma(A_1), \ldots, \sigma(A_n))$$

and for $\boldsymbol{X} \in \overline{\mathbb{IR}}^n$ we define

$$\psi(\boldsymbol{X}) = (\psi(X_1), \ldots, \psi(X_n)).$$

Theorem 2.3.10 motivates the following Definition:

Definition 2.3.11 (Rounding) *A function* $\mathtt{F} : \mathbb{IF}^n \to \mathbb{IF}$ *is called rounding of a function* $F : \overline{\mathbb{IR}}^n \to \overline{\mathbb{IR}}$ *if for all* $\boldsymbol{A} \in \mathbb{IF}^n$

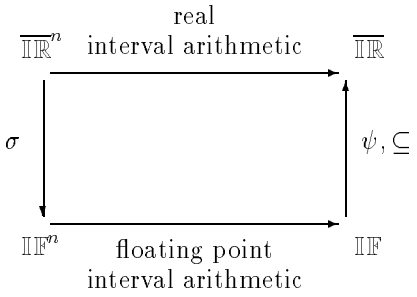
$$\mathtt{F}(\boldsymbol{A}) \supseteq \sigma(F(\psi(\boldsymbol{A}))).$$

$\mathtt{F}$ *is the exact rounding of* F *if for all* $\boldsymbol{A} \in \mathbb{IF}^n$

$$\mathtt{F}(\boldsymbol{A}) = \sigma(F(\psi(\boldsymbol{A}))). \ \square$$

Theorem 2.3.10 states that the floating point interval arithmetic functions are the exact roundings of the corresponding real interval arithmetic functions.

Now we are ready to prove correctness of floating point interval arithmetic in the sense that the evaluation of an arithmetic expression using floating point interval arithmetic yields an inclusion of the result obtained by using real interval arithmetic, provided the results are defined. This is described by the following diagram:



Theorem 2.3.12 *Let $F : \overline{\mathbb{IR}}^n \to \overline{\mathbb{IR}}$ be inclusion monotone and let $\mathtt{F} : \mathbb{IF}^n \to \mathbb{IF}$ be a rounding of F. Then for all $\boldsymbol{X} \in \overline{\mathbb{IR}}^n$ it holds that*

$$F(\boldsymbol{X}) \subseteq \psi(\mathtt{F}(\sigma(\boldsymbol{X}))). \ \square$$

Proof. Let F and $\mathtt{F}$ as in Theorem 2.3.12 and let $\boldsymbol{X} \in \overline{\mathbb{IR}}^n$ arbitrary but fixed. Note that

$$\begin{array}{lcl} \mathtt{F}(\sigma(\boldsymbol{X})) & \overset{\text{Definition 2.3.11}}{\supseteq} & \sigma(F(\psi(\sigma(\boldsymbol{X})))) \\ & \overset{\text{Lemma 2.3.7}}{\supseteq} & \sigma(F(\boldsymbol{X})). \end{array}$$

Hence,

$$\begin{array}{lcl} \psi(\mathtt{F}(\sigma(\boldsymbol{X}))) & \supseteq & \psi(\sigma(F(\boldsymbol{X}))) \\ & \overset{\text{Lemma 2.3.7}}{\supseteq} & F(\boldsymbol{X}). \ \square \end{array}$$

Corollary 2.3.13 *For all $X, Y \in \overline{\mathbb{IR}}$ it holds that*

$$\begin{array}{rcl} X + Y & \subseteq & \psi(\sigma(X) + \sigma(Y)) \\ X - Y & \subseteq & \psi(\sigma(X) - \sigma(Y)) \\ X * Y & \subseteq & \psi(\sigma(X) * \sigma(Y)) \\ X/Y & \subseteq & \psi(\sigma(X)/\sigma(Y)) \text{ if } 0 \notin \sigma(Y) \\ -X & \subseteq & \psi(-\sigma(X)). \ \square \end{array}$$

Proof. Follows from Theorem 2.3.10 and Theorem 2.3.12. $\square$

It remains to show that the composition of roundings is a rounding of the composition of functions.

Theorem 2.3.14 (Composition of Roundings) *Let $F : \overline{\mathbb{IR}}^m \to \overline{\mathbb{IR}}$, $G_i : \overline{\mathbb{IR}}^n \to \overline{\mathbb{IR}}$ be inclusion monotone and let $\mathtt{F} : \mathbb{IF}^m \to \mathbb{IF}$, $\mathtt{G}_i : \mathbb{IF}^n \to \mathbb{IF}$ be roundings of F, G_i for $i = 1, \ldots, m$ respectively. Then $\mathtt{F}(\mathtt{G}_1, \ldots, \mathtt{G}_m)$ is a rounding of $F(G_1, \ldots, G_m)$.* $\square$

Proof. Let $F, G_i, \mathtt{F}, \mathtt{G}_i$ as in Theorem 2.3.14 and let $\boldsymbol{A} \in \mathbb{IF}^n$ arbitrary but fixed. Then

$$\begin{array}{lcl} \mathtt{F}(\mathtt{G}_1, \ldots, \mathtt{G}_m)(\boldsymbol{A}) & \overset{\text{Definition 2.3.11}}{\supseteq} & \sigma(F(\psi(\mathtt{G}_1, \ldots, \mathtt{G}_m)))(\boldsymbol{A}) \\ & \overset{\text{Lemma 2.3.7}}{=} & \sigma(F(\psi(\mathtt{G}_1, \ldots, \mathtt{G}_n)))(\sigma(\psi(\boldsymbol{A}))) \\ & \overset{\text{Theorem 2.3.12}}{\supseteq} & \sigma(F(G_1, \ldots, G_m))(\psi(\boldsymbol{A})). \ \square \end{array}$$

Remark. Theorem 2.3.12 and Theorem 2.3.14 show that inclusion monotonicity is a very important property for the implementation of interval functions. □

Finally, we show that floating point interval arithmetic is inclusion monotone. We begin by proving that exact roundings of inclusion monotone functions are inclusion monotone. This implies inclusion monotonicity of the basic interval arithmetic functions. Next, we show that inclusion monotonicity is preserved under composition.

Theorem 2.3.15 *Let $F : \overline{\mathbb{IR}}^n \to \overline{\mathbb{IR}}$ be inclusion monotone and let $\mathtt{F} : \mathbb{IF}^n \to \mathbb{IF}$ be the exact rounding of F. Then $\mathtt{F}$ is inclusion monotone.* □

Proof. Let F and $\mathtt{F}$ as in Theorem 2.3.15 and let $\boldsymbol{A}, \boldsymbol{A}_\diamond \in \mathbb{IF}^n$ arbitrary but fixed such that $\boldsymbol{A}_\diamond \subseteq \boldsymbol{A}$. Then

$$\begin{array}{rcl} \mathtt{F}(\boldsymbol{A}_\diamond) & \overset{\text{Definition 2.3.11}}{=} & \sigma(F(\psi(\boldsymbol{A}_\diamond))) \\ & \overset{\text{Definition 2.3.6}}{\subseteq} & \sigma(F(\psi(\boldsymbol{A}))) \\ & \overset{\text{Definition 2.3.11}}{=} & \mathtt{F}(\boldsymbol{A}). \ \square \end{array}$$

Note that Theorem 2.3.15 can not be generalized to non–exact roundings.

Corollary 2.3.16 (Inclusion Monotonicity) *Let $A, B, A_\diamond, B_\diamond \in \mathbb{IF}$ such that $A_\diamond \subseteq A$, $B_\diamond \subseteq B$. Then*

$$\begin{array}{rcl} A_\diamond + B_\diamond & \subseteq & A + B \\ A_\diamond - B_\diamond & \subseteq & A - B \\ A_\diamond * B_\diamond & \subseteq & A * B \\ A_\diamond / B_\diamond & \subseteq & A/B \text{ if } 0 \notin B \\ -A_\diamond & \subseteq & -A. \ \square \end{array}$$

Proof. Follows from Theorem 2.3.10 and Theorem 2.3.15. □

Theorem 2.3.17 (Composition Preserves Inclusion Monotonicity) *Let $\mathtt{F} : \mathbb{IF}^m \to \mathbb{IR}$, $\mathtt{G}_i : \mathbb{IF}^n \to \mathbb{IF}$ be inclusion monotone for $i = 1, \ldots, m$. Then $\mathtt{F}(\mathtt{G}_1, \ldots, \mathtt{G}_m)$ is inclusion monotone.* □

Proof. Let $\mathtt{F}$, $\mathtt{G}_i$, $i = 1, \ldots, m$ as in Theorem 2.3.17. Let $\boldsymbol{A}_\diamond, \boldsymbol{A} \in \mathbb{IF}^n$ arbitrary but fixed such that $\boldsymbol{A}_\diamond \subseteq \boldsymbol{A}$. Then

$$\begin{array}{rcl} \mathtt{F}(\mathtt{G}_1, \ldots, \mathtt{G}_m)(\boldsymbol{A}_\diamond) & = & \mathtt{F}(\mathtt{G}_1(\boldsymbol{A}_\diamond), \ldots, \mathtt{G}_m(\boldsymbol{A}_\diamond)) \\ & \subseteq & \mathtt{F}(\mathtt{G}_1(\boldsymbol{A}), \ldots, \mathtt{G}_m(\boldsymbol{A})) \\ & = & \mathtt{F}(\mathtt{G}_1, \ldots, \mathtt{G}_m)(\boldsymbol{A}). \ \square \end{array}$$

2.3.1 Basic Algorithms on Floating Point Intervals

In this section we give algorithms for floating point interval arithmetic and some other operations, which will be used later.

Algorithm 2.3.18 (IADD) [Floating Point Interval Addition]

In: $A, B \in \mathbb{IF}$.

Out: $\mathrm{IADD}(A, B) \in \mathbb{IF}$, $\mathrm{IADD}(A, B) = A + B$.

(1) return $(\underline{A} \overset{\vee}{+} \underline{B}, \overline{A} \overset{\wedge}{+} \overline{B})$.

Theorem 2.3.19 (Complexity) *Algorithm 2.3.18 costs 2 floating point number additions.* □

Algorithm 2.3.20 (ISUB) [Floating Point Interval Subtraction]

In: $A, B \in \mathbb{IF}$.

Out: $\mathrm{ISUB}(A, B) \in \mathbb{IF}$, $\mathrm{ISUB}(A, B) = A - B$.

(1) return $(\underline{A} \overset{\vee}{-} \overline{B}, \overline{A} \overset{\wedge}{-} \underline{B})$.

Theorem 2.3.21 (Complexity) *Algorithm 2.3.20 costs 2 floating point number subtractions.* □

Algorithm 2.3.22 (INEG) [Floating Point Interval Negative]

In: $A \in \mathbb{IF}$.

Out: $\mathrm{INEG}(A) \in \mathbb{IF}$, $\mathrm{INEG}(A) = -A$.

(1) return $(-\overline{A}, -\underline{A})$.

Theorem 2.3.23 (Complexity) *Algorithm 2.3.22 costs 2 floating point number sign inversions.* □

Algorithm 2.3.24 for multiplying intervals is relatively complicated. This is mainly due to the fact, that the IEEE standard 754 does not provide the modified floating point multiplication functions $\overset{\wedge}{\cdot}$, $\overset{\vee}{\cdot}$. In order to minimize the number of floating point comparisons during interval multiplication, we do not call algorithms for $\overset{\wedge}{\cdot}$, $\overset{\vee}{\cdot}$, but use information on the sign of floating point number operands whenever possible. In any case, we have to avoid the multiplication of 0 and $\perp$ or $\top$ by $\overset{\wedge}{*}$ or $\overset{\vee}{*}$, because this would result in $\sharp$. According to Definition 2.3.8, the desired result in this case is 0.

Algorithm 2.3.24 (IMUL) [Floating Point Interval Multiplication]

In: $A, B \in \mathbb{IF}$.

Out: $\mathrm{IMUL}(A, B) \in \mathbb{IF}$, $\mathrm{IMUL}(A, B) = A * B$.

(1) [Case $\underline{A} > 0$.]
if $\underline{A} \leq 0$ goto Step 2.

(1.1) [Case $\underline{A} > 0$, $\underline{B} > 0$.]
if $\underline{B} > 0$ return $(\underline{A} \overset{\vee}{*} \underline{B}, \overline{A} \overset{\wedge}{*} \overline{B})$.

(1.2) [Case $\underline{A} > 0$, $\overline{B} < 0$.]
if $\overline{B} < 0$ return $(\overline{A} \overset{\vee}{*} \underline{B}, \overline{A} \overset{\wedge}{*} \overline{B})$.

(1.3) [Case $\underline{A} > 0$, $0 \in B$.]
if $\underline{B} = 0$ then $\underline{C} \longleftarrow 0$, else $\underline{C} \longleftarrow \overline{A} \overset{\vee}{*} \underline{B}$.
if $\overline{B} = 0$ then $\overline{C} \longleftarrow 0$, else $\overline{C} \longleftarrow \overline{A} \overset{\wedge}{*} \overline{B}$.
return C.

(2) [Case $\overline{A} < 0$.]
if $\overline{A} \geq 0$ goto Step 3.

(2.1) [Case $\overline{A} < 0$, $\underline{B} > 0$.]
if $\underline{B} > 0$ return $(\underline{A} \overset{\vee}{*} \overline{B}, \overline{A} \overset{\wedge}{*} \underline{B})$.

(2.2) [Case $\overline{A} < 0$, $\overline{B} < 0$.]
if $\overline{B} < 0$ return $(\overline{A} \overset{\vee}{*} \overline{B}, \underline{A} \overset{\wedge}{*} \underline{B})$.

(2.3) [Case $\overline{A} < 0$, $0 \in B$.]
if $\overline{B} = 0$ then $\underline{C} \longleftarrow 0$, else $\underline{C} \longleftarrow \underline{A} \overset{\vee}{*} \overline{B}$.
if $\underline{B} = 0$ then $\overline{C} \longleftarrow 0$, else $\overline{C} \longleftarrow \underline{A} \overset{\wedge}{*} \underline{B}$.
return C.

(3) [Case $0 \in A$.]

(3.1) [Case $0 \in A$, $\underline{B} > 0$.]
if $\underline{B} \leq 0$ goto Step 3.2.
if $\underline{A} = 0$ then $\underline{C} \longleftarrow 0$, else $\underline{C} \longleftarrow \underline{A} \overset{\vee}{*} \overline{B}$.
if $\overline{A} = 0$ then $\overline{C} \longleftarrow 0$, else $\overline{C} \longleftarrow \overline{A} \overset{\wedge}{*} \overline{B}$.
return C.

(3.2) [Case $0 \in A$, $\overline{B} < 0$.]
if $\overline{B} \geq 0$ goto Step 3.3.
if $\overline{A} = 0$ then $\underline{C} \longleftarrow 0$, else $\underline{C} \longleftarrow \overline{A} \overset{\vee}{*} \underline{B}$.
if $\underline{A} = 0$ then $\overline{C} \longleftarrow 0$, else $\overline{C} \longleftarrow \underline{A} \overset{\wedge}{*} \underline{B}$.
return C.

(3.3) [Case $0 \in A$, $0 \in B$.]

(3.3.1) [Case $\overline{A} = 0$, $0 \in B$.]
if $\overline{A} \neq 0$ goto Step 3.3.2.
if $\underline{A} = 0$ return $(0, 0)$.
if $\overline{B} = 0$ then $\underline{C} \longleftarrow 0$, else $\underline{C} \longleftarrow \underline{A} \overset{\vee}{*} \overline{B}$.
if $\underline{B} = 0$ then $\overline{C} \longleftarrow 0$, else $\overline{C} \longleftarrow \underline{A} \overset{\wedge}{*} \underline{B}$.
return C.

(3.3.2) [Case $\underline{A} = 0$, $\overline{A} \neq 0$, $0 \in B$.]
if $\underline{A} \neq 0$ goto Step 3.3.3.
if $\underline{B} = 0$ then $\underline{C} \longleftarrow 0$, else $\underline{C} \longleftarrow \overline{A} \overset{\vee}{*} \underline{B}$.
if $\overline{B} = 0$ then $\overline{C} \longleftarrow 0$, else $\overline{C} \longleftarrow \overline{A} \overset{\wedge}{*} \overline{B}$.
return C.

(3.3.3) [Case $0 \in A$, $\underline{A} \neq 0$, $\overline{A} \neq 0$, $\overline{B} = 0$.]
if $\overline{B} \neq 0$ goto Step 3.3.4.
if $\underline{B} = 0$ return $(0, 0)$, else return $(\overline{A} \overset{\vee}{*} \underline{B}, \underline{A} \overset{\wedge}{*} \underline{B})$.

(3.3.4) [Case $0 \in A$, $\underline{A} \neq 0$, $\overline{A} \neq 0$, $\underline{B} = 0$, $\overline{B} \neq 0$.]
if $\underline{B} = 0$ return $(\underline{A} \overset{\vee}{*} \overline{B}, \overline{A} \overset{\wedge}{*} \overline{B})$.

(3.3.5) [Case $0 \in A, 0 \in B, 0 \notin \{\overline{A}, \underline{A}, \overline{B}, \underline{B}\}$.]
$\underline{C} \longleftarrow \min\{\overline{A} \overset{\vee}{*} \underline{B}, \underline{A} \overset{\vee}{*} \overline{B}\}$.
$\overline{C} \longleftarrow \max\{\overline{A} \overset{\wedge}{*} \overline{B}, \underline{A} \overset{\wedge}{*} \underline{B}\}$.
return C.

Theorem 2.3.25 (Complexity) *If $0 \in \mathrm{int}(A)$ and $0 \in \mathrm{int}(B)$ then Algorithm 2.3.24 costs 4 floating point number multiplications. Otherwise, Algorithm 2.3.24 costs 2 floating point number multiplications.* □

Algorithm 2.3.26 (IDIV**)** [Floating Point Interval Division]

In: $A, B \in \mathbb{IF}$, $0 \notin B$.

Out: $\mathrm{IDIV}(A, B) \in \mathbb{IF}$, $\mathrm{IDIV}(A, B) = A/B$.

(1) [Distinguish cases $\underline{B} > 0$ and $\overline{B} < 0$.]
if $\underline{B} > 0$ goto Step 2, else goto Step 3.

(2) [Case $\underline{B} > 0$.]
if $\overline{A} < 0$ return $(\underline{A} \overset{\vee}{/} \underline{B}, \overline{A} \overset{\wedge}{/} \overline{B})$.

if $\underline{A} > 0$ return $(\underline{A} \overset{\vee}{/} \overline{B}, \overline{A} \overset{\wedge}{/} \underline{B})$.
return $(\underline{A} \overset{\vee}{/} \underline{B}, \overline{A} \overset{\wedge}{/} \underline{B})$.

(3) [Case $\overline{B} < 0$.]
if $\overline{A} < 0$ return $(\overline{A} \overset{\vee}{/} \underline{B}, \underline{A} \overset{\wedge}{/} \overline{B})$.
if $\underline{A} > 0$ return $(\overline{A} \overset{\vee}{/} \overline{B}, \underline{A} \overset{\wedge}{/} \underline{B})$.
return $(\overline{A} \overset{\vee}{/} \overline{B}, \underline{A} \overset{\wedge}{/} \overline{B})$.

Theorem 2.3.27 (Complexity) *Algorithm 2.3.26 costs 2 floating point number divisions.* □

The following algorithm gives the exact rounding of the interval square function.

Algorithm 2.3.28 (ISQR) [Floating Point Interval Square]

In: $A \in \mathbb{IF}$.

Out: $\mathrm{ISQR}(A) \in \mathbb{IF}$, $\mathrm{ISQR}(A) = \sigma(\psi(A)^2)$.

(1) [Distinguish cases $0 \in A$ and $0 \notin A$.]
if $\underline{A} \leq 0$ and $\overline{A} \geq 0$ goto Step 2, else goto Step 3.

(2) [Case $0 \in A$.]
if $-\underline{A} > \overline{A}$ return $(0, \underline{A} \overset{\wedge}{*} \underline{A})$ else return $(0, \overline{A} \overset{\wedge}{*} \overline{A})$.

(3) [Case $0 \notin A$.]
if $\overline{A} > 0$ return $(\underline{A} \overset{\vee}{*} \underline{A}, \overline{A} \overset{\wedge}{*} \overline{A})$ else return $(\overline{A} \overset{\vee}{*} \overline{A}, \underline{A} \overset{\wedge}{*} \underline{A})$.

Theorem 2.3.29 (Complexity) *Algorithm 2.3.28 costs 1 floating point number multiplication if* $0 \in A$ *and 2 floating point multiplications else.* □

Algorithm 2.3.30 (IPOW) [Floating Point Interval Power]

In: $A \in \mathbb{IF}$, $n \in \mathbb{N}$.

Out: $\mathrm{IPOW}(A, n) \in \mathbb{IF}$, $\psi(\mathrm{IPOW}(A, n)) \supseteq \psi(A)^n$.

(1) [Distinguish cases n even and n odd.]
if n is even goto Step 2, else goto Step 5.

(2) [Distinguish cases $0 \in A$ and $0 \notin A$.]
if $\underline{A} \leq 0$ and $\overline{A} \geq 0$ goto Step 3, else goto Step 4.

(3) [Case n even, $0 \in A$.]
if $-\underline{A} > \overline{A}$ return $(0, \hat{\mathrm{POW}}(\underline{A}, n))$ else return $(0, \hat{\mathrm{POW}}(\overline{A}, n))$.

(4) [Case n even, $0 \notin A$.]
if $\overline{A} > 0$ return $(\check{\mathrm{POW}}(\underline{A}, n), \hat{\mathrm{POW}}(\overline{A}, n))$ else return $(\check{\mathrm{POW}}(\overline{A}, n), \hat{\mathrm{POW}}(\underline{A}, n))$.

(5) [Case n odd.]
return $(\check{\mathrm{POW}}(\underline{A}, n), \hat{\mathrm{POW}}(\overline{A}, n))$.

Theorem 2.3.31 (Complexity) *Algorithm 2.3.28 costs 1 floating point number power computation if* n *is even and* $0 \in A$ *and 2 floating point number power computations else.* □

The following observation will be useful later on.

Theorem 2.3.32 *If* $0 \notin \mathrm{int}(A)$ *then* $0 \notin \mathrm{int}(\mathrm{IPOW}(A, n))$ *for all* $n \in \mathbb{N}$. □

Algorithm 2.3.33 computes a floating point approximation of the midpoint of an interval.

Algorithm 2.3.33 (MID) [Midpoint of a Floating Point Interval]

In: $A \in \mathbb{IF}$.

Out: $\mathrm{MID}(A) \in \mathbb{F}_o$, $\mathrm{MID}(A) \in A$.

(1) [Case $A = (\bot, \top)$.]
if $\underline{A} = -\overline{A}$ return 0.

(2) [Case $\underline{A} = \bot$ or $\overline{A} = \top$.]
if $\underline{A} = \bot$ return $\min \mathbb{F}_o$.
if $\overline{A} = \top$ return $\max \mathbb{F}_o$.

(3) return $\underline{A} \stackrel{\approx}{+} \stackrel{\approx}{\rho}(0.5) \stackrel{\approx}{*} (\overline{A} \stackrel{\vee}{-} \underline{A})$.

Theorem 2.3.34 (Correctness) *Algorithm 2.3.33 (*MID*) is correct.* □

Proof. Let $A \in \mathbb{IF}$ and let $b = \mathrm{MID}(A)$. If $\underline{A} = \bot$ or $\overline{A} = \top$ then obviously $b \in \mathbb{F}_o$ and $b \in A$. Assume $\underline{A}, \overline{A} \in \mathbb{F}_o$ and let $a = \overline{A} \stackrel{\vee}{-} \underline{A}$. According to Definition 2.2.11 $a \in \mathbb{F}_o$ and

$$0 \leq \phi(a) \leq \phi(\overline{A}) - \phi(\underline{A}).$$

Let $a' = \stackrel{\approx}{\rho}(0.5) \stackrel{\vee}{*} a$. As $\phi(0) = 0$ it follows from Definition 2.2.4 that $0 \leq \phi(\stackrel{\approx}{\rho}(0.5)) \leq 1$ and therefore

$$0 \leq \phi(a') \leq \phi(a) \leq \phi(\overline{A}) - \phi(\underline{A}).$$

Hence,

$$\phi(\underline{A}) \leq \phi(\underline{A}) + \phi(a') \leq \phi(\overline{A}).$$

From Lemma 2.2.8 it follows that

$$\stackrel{\approx}{\rho}(\phi(\underline{A})) \leq \stackrel{\approx}{\rho}(\phi(\underline{A}) + \phi(a')) \leq \stackrel{\approx}{\rho}(\phi(\overline{A})),$$

and according to Lemma 2.2.10 and Definition 2.2.11

$$\underline{A} \leq \underline{A} \stackrel{\approx}{+} a' \leq \overline{A}.$$

As $b = \underline{A} \stackrel{\approx}{+} a'$ it follows that $b \in \mathbb{F}_o$ and $b \in A$. □

Theorem 2.3.35 (Complexity) *Algorithm 2.3.33 (*MID*) costs 1 number multiplication and 2 number additions.* □

Algorithm 2.3.36 computes a floating point approximation of the width of an interval. The computed value is greater or equal the exact width of the input interval.

Algorithm 2.3.36 (WIDTH) [Width of a Floating Point Interval]

In: $A \in \mathbb{IF}$.

Out: $\mathrm{WIDTH}(A) \in \mathbb{F}$,

$$\mathrm{WIDTH}(A) = \begin{cases} \top & \text{if } \psi(A) \text{ is right or left unbounded} \\ \hat{\rho}(\mathrm{w}(\psi(A))) & \text{else.} \end{cases}$$

(1) return $\overline{A} \stackrel{\wedge}{-} \underline{A}$.

Theorem 2.3.37 *Algorithm 2.3.36 (*WIDTH*) costs 1 floating point addition.* □

Next, we give an algorithm for the exact rounding of the generalized and hull division of extended intervals.

Definition 2.3.38 (Generalized and Hull Division) *The function* $\mathrm{gdiv} : \overline{\mathbb{IR}}_\emptyset \times \overline{\mathbb{IR}}_\emptyset \times \overline{\mathbb{IR}}_\emptyset \to \mathcal{P}(\mathbb{R})$ *is defined as*

$$\mathrm{gdiv}(N, D, X) = \{x \in X \mid n = dx \text{ for some } n \in N, d \in D\}.$$

The function $\mathrm{hdiv} : \overline{\mathbb{IR}}_\emptyset \times \overline{\mathbb{IR}}_\emptyset \times \overline{\mathbb{IR}}_\emptyset \to \overline{\mathbb{IR}}_\emptyset$ *is defined as*

$$\mathrm{hdiv}(N, D, X) = [\mathrm{gdiv}(N, D, X)]. \ \square$$

In order to simplify notation, we define $\overline{X} = \infty$ if X is right unbounded and $\underline{X} = -\infty$ if X is left unbounded. Arithmetic on $\mathbb{R} \cup \{-\infty, \infty\}$ is defined in the usual way.

Theorem 2.3.39 (Generalized Division) *For all* $N, D, X \in \overline{\mathbb{IR}}$ *it holds that*

$$\mathrm{gdiv}(N, D, X) = \begin{cases} N/D \cap X & \text{if } 0 \notin D \\ X & \text{if } 0 \in N, 0 \in D \\ \emptyset & \text{if } 0 \notin N, D = 0 \\ \{x \in X \mid x \geq \underline{N}/\overline{D}\} & \text{if } \underline{N} > 0, \underline{D} = 0, \overline{D} > 0 \\ \{x \in X \mid x \leq \underline{N}/\underline{D}\} & \text{if } \underline{N} > 0, \underline{D} < 0, \overline{D} = 0 \\ \{x \in X \mid x \leq \underline{N}/\underline{D}\} \cup \{x \in X \mid x \geq \underline{N}/\overline{D}\}) & \text{if } \underline{N} > 0, 0 \in \mathrm{int}(D) \\ \{x \in X \mid x \leq \overline{N}/\overline{D}\} & \text{if } \overline{N} < 0, \underline{D} = 0, \overline{D} > 0 \\ \{x \in X \mid x \geq \overline{N}/\underline{D}\} & \text{if } \overline{N} < 0, \underline{D} < 0, \overline{D} = 0 \\ \{x \in X \mid x \leq \overline{N}/\overline{D}\} \cup \{x \in X \mid x \geq \overline{N}/\underline{D}\}) & \text{if } \overline{N} < 0, 0 \in \mathrm{int}(D). \ \square \end{cases}$$

Thus, for all $N, D, X \in \overline{\mathbb{IR}}_\emptyset$ there exist $Y_1, Y_2 \in \overline{\mathbb{IR}}_\emptyset$ such that

$$\mathrm{gdiv}(N, D, X) = Y_1 \cup Y_2.$$

Algorithm 2.3.40 (GDIV) [Generalized Division]

In: $N, D, X \in \mathbb{IF}_\emptyset$.

Out: $A_1, A_2 \in \mathbb{IF}_\emptyset$, $A_1 = \sigma(Y_1)$, $A_2 = \sigma(Y_2)$ for some $Y_1, Y_2 \in \overline{\mathbb{IR}}_\emptyset$ such that $\mathrm{gdiv}(\psi(N), \psi(D), \psi(X)) = Y_1 \cup Y_2$.

(1) [Case $N = \emptyset$ or $D = \emptyset$ or $X = \emptyset$.]
if $N = \emptyset$ or $D = \emptyset$ or $X = \emptyset$ return $(\sharp, \sharp), (\sharp, \sharp)$.

(2) [Case $0 \notin D$.]
if $\underline{D} > 0$ or $\overline{D} < 0$ return $N/D \cap X, (\sharp, \sharp)$.

(3) [Case $0 \in N$, $0 \in D$.]
if $0 \in N$ return $X, (\sharp, \sharp)$.

(4) [Case $0 \notin N$, $D = (0, 0)$.]
if $D = (0, 0)$ return $(\sharp, \sharp), (\sharp, \sharp)$.

(5) [Distinguish cases $\underline{N} > 0$ and $\overline{N} < 0$.]
if $\underline{N} > 0$ goto Step 5, else goto Step 6.

(6) [Case $\underline{N} > 0$, $0 \in D$, $D \neq (0, 0)$.]
if $\underline{D} = 0$ return $(\underline{N} \mathbin{\check{/}} \overline{D}, \top) \cap X, (\sharp, \sharp)$.
if $\overline{D} = 0$ return $(\bot, \underline{N} \mathbin{\hat{/}} \underline{D}) \cap X, (\sharp, \sharp)$.
return $(\bot, \underline{N} \mathbin{\hat{/}} \underline{D}) \cap X, (\underline{N} \mathbin{\check{/}} \overline{D}, \top) \cap X$.

(7) [Case $\overline{N} < 0$, $0 \in D$, $D \neq (0, 0)$.]
if $\underline{D} = 0$ return $(\bot, \overline{N} \mathbin{\hat{/}} \overline{D}) \cap X, (\sharp, \sharp)$.

if $\overline{D} = 0$ return $(\overline{N} \mathbin{\check{/}} \underline{D}, \top) \cap X, (\sharp, \sharp)$.
return $(\bot, \overline{N} \mathbin{\hat{/}} \overline{D}) \cap X, (\overline{N} \mathbin{\check{/}} \underline{D}, \top) \cap X$.

Algorithm 2.3.41 (HDIV) [Hull Division]

In: $N, D, X \in \mathbb{IF}_\emptyset$.

Out: $\mathrm{HDIV}(N, D, X) \in \mathbb{IF}_\emptyset$, $\mathrm{HDIV}(N, D, X) = \sigma(\mathrm{hdiv}(\psi(N), \psi(D), \psi(X)))$.

(1) [Generalized Division.]
$A_1, A_2 \longleftarrow \mathrm{GDIV}(N, D, X)$.

(2) [Hull]
if $A_1 = (\sharp, \sharp)$ return A_2.
if $A_2 = (\sharp, \sharp)$ return A_1.
return $(\underline{A}_1, \overline{A}_2)$.

Theorem 2.3.42 (Complexity) *Algorithm 2.3.40 and Algorithm 2.3.41 cost 2 floating point number divisions.* □

Algorithms in the following sections often require arithmetic operations, where one operand is a real number and the other operand is a real interval. These operations are defined in the real case by the embedding of real numbers into the set of real intervals. However, in the floating point case such an embedding is not possible, hence we have to define an explicit type conversion function which has *not* the properties of a homomorphism.

Definition 2.3.43 (Floating Point Number to Interval Conversion) *The function* $\gamma : \mathbb{F} \to \mathbb{IF}$ *is defined as*

$$\gamma(a) = \begin{cases} (a, a) & \text{if } a \in \mathbb{F}_o \\ (\bot, \min \mathbb{F}_o) & \text{if } a = \bot \\ (\max \mathbb{F}_o, \top) & \text{if } a = \top. \end{cases} \quad \square$$

From Definition 2.3.4 it follows that for all $a \in \mathbb{F}_o$

$$\phi(a) = \psi(\gamma(a)).$$

Further, by Definition 2.3.9 for all $a, b \in \mathbb{F}_o$, $\circ \in \{+, -, *, /\}$

$$(a \mathbin{\check{\circ}} b, a \mathbin{\hat{\circ}} b) = \gamma(a) \circ \gamma(b)$$

and for all $a \in \mathbb{F}$

$$\gamma(-a) = -\gamma(a).$$

The definition of γ for $a = \top$ and $a = \bot$ avoids sometimes the necessity of treating special cases. For example

$$(\underline{\gamma(\underline{A})}, \overline{\gamma(\overline{A})}) = A$$

for all $A \in \mathbb{IF}$. An algorithm for computing γ is straight forward:

Algorithm 2.3.44 (CONVERT) [Floating Point Number to Interval Conversion]

In: $a \in \mathbb{F}$.

Out: $\mathrm{CONVERT}(a) \in \mathbb{IF}$, $\mathrm{CONVERT}(a) = \gamma(a)$.

(1) if $a = \bot$ return $(\bot, \min \mathbb{F}_o)$.
if $a = \top$ return $(\max \mathbb{F}_o, \top)$.
return (a, a).

Definition 2.3.45 *Let* $\circ \in \{+,-,*,/\}$. *Then* $\circ : \mathbb{F} \times \mathbb{IF} \to \mathbb{IF}$ *is defined as*

$$a \circ B = \gamma(a) \circ B.$$

Similarly, $\circ : \mathbb{IF} \times \mathbb{F} \to \mathbb{IF}$ *is defined as*

$$A \circ b = A \circ \gamma(b). \ \square$$

Algorithms for arithmetic operations on a floating point number and a floating point interval, which do not explicitly call CONVERT are straight forward. Usually, they require less control structures and less floating point number comparisons, but the same number of arithmetic floating point number operations as algorithms which first convert the floating point number argument to an interval. As an example we give an Algorithm which computes the product of a floating point number and a floating point interval.

Algorithm 2.3.46 (NIMUL) [Floating Point Number / Interval Multiplication]

In: $a \in \mathbb{F}$, $B \in \mathbb{IF}$.

Out: $\mathrm{NIMUL}(a,B) \in \mathbb{IF}$, $\mathrm{NIMUL}(a,B) = \mathrm{IMUL}(\gamma(a),B)$.

(1) [Case $a = \top$.]
if $a = \top$
if $\underline{\underline{B}} < 0$ then $\underline{\underline{C}} \longleftarrow \bot$, else $\underline{\underline{C}} \longleftarrow \max \mathbb{F}_o \overset{\vee}{*} \underline{\underline{B}}$.
if $\overline{\overline{B}} > 0$ then $\overline{\overline{C}} \longleftarrow \top$, else $\overline{\overline{C}} \longleftarrow \max \mathbb{F}_o \overset{\wedge}{*} \overline{\overline{B}}$.
return C.

(2) [Case $a = \bot$.]
if $a = \bot$
if $\underline{B} < 0$ then $\overline{C} \longleftarrow \top$, else $\overline{C} \longleftarrow \min \mathbb{F}_o \overset{\wedge}{*} \underline{B}$.
if $\overline{B} > 0$ then $\underline{C} \longleftarrow \bot$, else $\underline{C} \longleftarrow \min \mathbb{F}_o \overset{\vee}{*} \overline{B}$.
return C.

(3) [Case $a \in \mathbb{F}_o$.]
if $a > 0$ return $(\underline{B} \overset{\vee}{*} a, \overline{B} \overset{\wedge}{*} a)$.
if $a < 0$ return $(\overline{B} \overset{\vee}{*} a, \underline{B} \overset{\wedge}{*} a)$.
return (0,0).

Theorem 2.3.47 (Complexity) *For all* $a \in \mathbb{F}_o$, $B \in \mathbb{IF}$, $\underline{B} \neq 0$, $\overline{B} \neq 0$ *Algorithm 2.3.46 with input* a, B *executes as many floating point number multiplications as Algorithm 2.3.24 with input* $\gamma(a), B$. $\square$

In the sequel, we omit $\phi, \psi, \tilde{\tilde{\rho}}, \hat{\rho}, \check{\rho}, \sigma, \gamma$ if no confusion can arise. Further, we write $+,-,-,*,/$ instead of IADD, ISUB, INEG, IMUL, IDIV respectively and use the mathematical notation for exponentiation instead of IPOW. Instead of CONVERT(a) we will write $[a]$ or simply a.

Chapter 3

Inclusion of the Range of Univariate Polynomials

The overestimation of the range of functions is a fundamental problem in interval mathematics. We will not treat the problem in its full generality, rather we restrict ourselves to polynomials. Apart from linear functions, for which range computation is trivial, polynomials are the "simplest" subclass of elementary transcendental functions. We can therefore expect more efficient methods and more refined results for polynomials than for arbitrary functions. The practical importance of polynomials is out of question, therefore it seems worthwhile to study them separately. In this chapter we consider univariate polynomials, the multivariate case is treated in the next chapter.

In the following let $f : \mathbb{R} \to \mathbb{R}$,

$$f(x) = a_n x^{d_n} + a_{n-1} x^{d_{n-1}} + \ldots + a_1 x^{d_1}, \quad a_i \neq 0, \quad d_{i+1} > d_i$$

be a polynomial and let

$$\begin{aligned} \Delta_1 &= d_1 \\ \Delta_i &= d_i - d_{i-1}, \quad i = 2, \ldots, n. \end{aligned}$$

3.1 Horner Form

There are many ways to rewrite an arithmetic expression and it is well known, that expressions, which are equivalent as functions over the reals are usually different as functions over intervals. A standard arrangement of expressions which contain only addition and multiplication is the Horner form – it is numerically stable and allows efficient evaluation. Concerning the interval case, Horner form is optimal with respect to subdistributivity. Among all methods described in this chapter, Horner form is cheapest. On the other hand, there are many examples, where Horner form gives the exact range, where "more sophisticated" methods overestimate.

Section 3.1.1 contains a new criterion to decide whether Horner form returns the exact range or a proper overestimation. An algorithmic test of the criterion is straight forward. Improvements if the argument interval does not contain 0 are discussed in Section 3.1.2. Sometimes it is sufficient to compute only an upper or lower bound of the range. A new algorithm for computing the endpoints of the Horner form separately is given in Section 3.1.2.2. Bisection of the input interval in connection with Horner form is studied in Section 3.1.3. We give some original theorems for estimating the reduction of the overestimation error through bisection and devise an efficient algorithm for the special case when the input interval is bisected at 0.

Definition 3.1.1 (Horner Form) *The Horner form $H_f : \mathbb{IR} \to \mathbb{IR}$ of f is defined as*

$$H_f(X) = \Big(\big(\ldots (a_n X^{\Delta_n} + a_{n-1}) X^{\Delta_{n-1}} + \ldots + a_2 \big) X^{\Delta_2} + a_1 \Big) X^{\Delta_1}. \ \square$$

If iterated multiplication is used instead of the interval power function in the definition of the Horner form, we obtain the dense Horner form.

Definition 3.1.2 (Dense Horner Form) *The dense Horner form $H_f^* : \mathbb{IR} \to \mathbb{IR}$ of f is defined as*

$$H_f^*(X) = \Big(\big(\ldots (a_n \underbrace{X \cdots X}_{\Delta_n \text{times}} + a_{n-1}) \underbrace{X \cdots X}_{\Delta_{n-1}\text{times}} + \ldots + a_2 \big) \underbrace{X \cdots X}_{\Delta_2 \text{times}} + a_1 \Big) \underbrace{X \cdots X}_{\Delta_1 \text{times}}. \ \square$$

Theorem 3.1.3 *For all $X \in \mathbb{IR}$ it holds that*

$$H_f(X) \subseteq H_f^*(X).$$

If f is dense or $0 \notin \operatorname{int}(X)$ then

$$H_f(X) = H_f^*(X). \ \square$$

Proof. Let

$$\begin{array}{rclcrcl} P_n(X) & = & a_n X^{\Delta_n}, & & P_i(X) & = & (P_{i+1}(X) + a_i) X^{\Delta_i} \\ P_n^*(X) & = & a_n \underbrace{X \cdots X}_{\Delta_n \text{times}}, & & P_i^*(X) & = & (P_{i+1}^*(X) + a_i) \underbrace{X \cdots X}_{\Delta_i \text{times}} \end{array}$$

for $i = 1, \ldots, n-1$. Note that $P_1(X) = H_f(X)$ and $P_1^*(X) = H_f^*(X)$. By induction it follows that $P_i(X) \subseteq P_i^*(X)$ for all i. If f is dense then $\Delta_i = 1$, for all $i > 1$ and $\Delta_1 = 0$, hence $P_i(X) = P_i^*(X)$ for all i. If $0 \notin \operatorname{int}(X)$ then X^{Δ_i} is equal to Δ_i times the product of X with itself, and $P_i(X) = P_i^*(X)$ for all i. $\square$

Theorem 3.1.4 (Inclusion Monotonicity) *H_f and H_f^* are inclusion monotone interval extensions of f.* $\square$

Proof. Obviously $H_f(X) = H_f^*(X) = f(X)$ if $X \in \mathbb{R}$. The inclusion monotonicity of H_f and H_f^* follows from Corollary 1.3.5 and Theorem 1.3.22 shows that H_f, H_f^* are interval extensions of f. $\square$

Theorem 3.1.5 (Convergence) *H_f and H_f^* converge linearly to f.* $\square$

Proof. H_f and H_f^* are interval extensions of f (Theorem 3.1.4) and Lipschitz (Corollary 1.3.19). Hence H_f and H_f^* converge linearly to f (Theorem 1.3.24). $\square$

The following theorem simplifies some considerations later on.

Theorem 3.1.6 *For all $c \in \mathbb{R}$, $X \in \mathbb{IR}$ the following holds.*

(i) *If $g(x) = f(x) + c$ then*

$$H_g(X) = H_f(X) + c.$$

(ii) *If $g(x) = cf(x)$ then*

$$H_g(X) = cH_f(X).$$

(iii) *If $g(x) = f(cx)$ then*

$$H_g(X) = H_f(cX). \ \square$$

Proof. Let $c \in \mathbb{R}$ and $X \in \mathbb{IR}$ arbitrary but fixed.

(i) Let $g(x) = f(x) + c$. Then

$$\begin{array}{rcl} H_g(X) & = & \Big(\big(\ldots (a_n X^{\Delta_n} + a_{n-1}) X^{\Delta_{n-1}} + \ldots + a_2 \big) X^{\Delta_2} + a_1 \Big) X^{\Delta_1} + c \\ & = & H_f(X). \end{array}$$

(ii) Let $g(x) = c\,f(x)$. Then

$$\begin{aligned} H_g(X) &= \Big(\big(\ldots(ca_nX^{\Delta_n} + ca_{n-1})X^{\Delta_{n-1}} + \ldots + ca_2\big)X^{\Delta_2} + ca_1\Big)X^{\Delta_1} \\ &= \Big(\big(\ldots c(a_nX^{\Delta_n} + a_{n-1})X^{\Delta_{n-1}} + \ldots + ca_2\big)X^{\Delta_2} + ca_1\Big)X^{\Delta_1} \\ &\ldots \\ &= c\Big(\big(\ldots(a_nX^{\Delta_n} + a_{n-1})X^{\Delta_{n-1}} + \ldots + a_2\big)X^{\Delta_2} + a_1\Big)X^{\Delta_1} \\ &= cH_f(X). \end{aligned}$$

(iii) Let $g(x) = f(cx)$. Then

$$\begin{aligned} H_g(X) &= \Big(\big(\ldots(a_nc^{d_n}X^{\Delta_n} + a_{n-1}c^{d_{n-1}})X^{\Delta_{n-1}} + \ldots + a_2c^{d_2}\big)X^{\Delta_2} + a_1c^{d_1}\Big)X^{\Delta_1} \\ &= \Big(\big(\ldots(a_n(cX)^{\Delta_n} + a_{n-1})c^{d_{n-1}}X^{\Delta_{n-1}} + \ldots + a_2c^{d_2}\big)X^{\Delta_2} + a_1c^{d_1}\Big)X^{\Delta_1} \\ &\ldots \\ &= \Big(\big(\ldots(a_n(cX)^{\Delta_n} + a_{n-1})(cX)^{\Delta_{n-1}} + \ldots + a_2\big)(cX)^{\Delta_2} + a_1\Big)(cX)^{\Delta_1} \\ &= H_f(cX). \ \square \end{aligned}$$

Remark. In general it is *not* true that $g(x) = f(x + c)$ implies $H_g(X) = H_f(X + c)$. For example, let $f(x) = x^2 - x$, $X = [0, 2]$ and $c = -1$. Then $g(x) = x^2 - 3x + 2$, $H_g(X) = [-4, 2]$ but $H_f(X + c) = [-2, 2]$. This observation can be exploited to reduce the overestimation error of the Horner form, see Section 3.3. $\square$

From Definition 3.1.1 we obtain the following algorithm for evaluating the Horner form.

Algorithm 3.1.7 (HF) [Horner Form]

In: $f(x) = a_nx^{d_n} + a_{n-1}x^{d_{n-1}} + \ldots + a_1x^{d_1} \in \mathbb{F}[x]$,
$X \in \mathbb{IF}$.

Out: $\mathrm{HF}(f, X) \in \mathbb{IF}$, $\mathrm{HF}(f, X) \supseteq H_f(X)$.

(1) [Initialize.]
$P_n \longleftarrow a_n$.

(2) [Accumulate.]
for $i = n - 1, \ldots, 1$
$P_i \longleftarrow P_{i+1}X^{d_{i+1}-d_i} + a_i$.

(3) [Last power.]
$P_0 \longleftarrow P_1X^{d_1}$.

(4) [Return.]
return P_0.

Remark. The specification of Algorithm 3.1.7 (HF) requires some explanation. It is clear how a polynomial $f(x) \in \mathbb{F}[x]$ is interpreted as a polynomial in $\mathbb{R}[x]$ if all coefficients are ordinary floating point numbers. As we allow $\bot$ and $\top$ as coefficients, the output must be specified more precisely as

$$\mathrm{HF}(f, X) \in \mathbb{IF}, \quad \mathrm{HF}(f, X) \supseteq \bigcup_{\tilde{f} \in \mathcal{F}} H_{\tilde{f}}(X),$$

where

$$\mathcal{F} = \left\{ \sum_{i=1}^{n} \tilde{a}_i x^{d_i} \;\middle|\; \begin{array}{ll} \tilde{a}_i = a_i & \text{if } a_i \in \mathbb{F}_o \\ \tilde{a}_i \geq \max \mathbb{F}_o & \text{if } a_i = \top \\ \tilde{a}_i \leq \min \mathbb{F}_o & \text{if } a_i = \bot \end{array} \right\}.$$

In order to simplify notation we write in the specification of subsequent range computation algorithms as in Algorithm 3.1.7. $\square$

Theorem 3.1.8 (Complexity) *Algorithm 3.1.7 (*HF*) costs*

$$\begin{array}{rl} n & \textit{interval power computations,} \\ n & \textit{interval multiplications and} \\ 2n-2 & \textit{number additions.} \ \square \end{array}$$

If $0 \notin \mathrm{int}(X)$ then every interval multiplication in Algorithm 3.1.7 (HF) costs only 2 number multiplications:

Theorem 3.1.9 (Complexity) *If $0 \notin \mathrm{int}(X)$ then Algorithm 3.1.7 (*HF*) costs*

$$\begin{array}{rl} n & \textit{interval power computations,} \\ 2n & \textit{number multiplications and} \\ 2n-2 & \textit{number additions.} \ \square \end{array}$$

Proof. Follows from Theorem 2.3.32 and Theorem 2.3.25. □

3.1.1 Non–Overestimation of Horner Form

Experimentally it was observed that there are many cases where the Horner form is exact, i.e. $H_f(X) = f(X)$.

- For example if $f(x) = x^2 - x$ and $X = [1, 2]$ we obtain $H_f(X) = [0, 2] = f(X)$. More generally, the Horner form of $f(x)$ evaluated on X does not lead to overestimation for all intervals where $\underline{X} \geq 1$ or $\overline{X} \leq 0$.
- Another example is $f(x) = x^3 - 2x$ evaluated on $[-2, -1.5]$. Again, we can generalize: Horner evaluation of $f(x)$ on X does not lead to overestimation if $\underline{X} \geq \sqrt{2}$ or $\overline{X} \leq -\sqrt{2}$.

In this section we present a new criterion to decide whether Horner Form computes the exact range or a proper overestimation. If $\mathrm{mig}(X)$ is sufficiently large, i.e. if

$$\underline{X} \geq \overline{O}_f \quad \text{or} \quad \overline{X} \leq \underline{O}_f$$

for some bounds

$$\overline{O}_f \geq 0 \geq \underline{O}_f$$

(Theorem 3.1.12), then Horner form is exact. The computation of the bounds $\overline{O}_f, \underline{O}_f$ is expensive. However, it can be tested very cheaply whether a given interval is outside the bounds, without actually knowing them (Algorithm 3.1.23).

It should be noted that this simple condition for non–overestimation is sufficient but *not* necessary. Counterexamples have been found, which could be generalized to whole classes of polynomials, where Horner form gives the range although X is not outside O_f (Section 3.1.1.4). Examples of two rather obvious classes are as follows:

- The Horner form of $f(x) = x^4 + x^2$ never overestimates.
- The Horner form of $x^3 + 2x$ gives the range for all argument intervals, which have midpoint zero.

A complete characterization of all cases where Horner form is exact is still unknown. The results of this section are original. Before going into details we eliminate some trivial cases from the discussion.

Theorem 3.1.10 *For all $c \in \mathbb{R}$, $X \in \mathbb{IR}$, if $g(x) = f(x) + c$ or $g(x) = cf(x)$, $c \neq 0$, then*

$$H_f(X) = f(X) \text{ iff } H_g(X) = g(X). \ \square$$

Proof. Theorem 3.1.10 is an immediate consequence of Theorem 3.1.6. Let $c \in \mathbb{R}$ and $X \in \mathbb{IR}$ arbitrary but fixed.

- Let $g(x) = f(x) + c$. Then $H_f(X) = f(X)$ iff $H_f(X) + c = f(X) + c$ iff $H_g(X) = g(X)$.
- Let $g(x) = cf(x)$, $c \neq 0$. Then $H_f(X) = f(X)$ iff $cH_f(X) = cf(X)$ iff $H_g(X) = g(X)$. □

Thus, we can make the following assumptions.

- Whether Horner form gives the range does not depend on an additive constant. Hence we can assume that f has no constant monomial, i.e. $d_1 > 0$.
- It is also independent of a constant factor whether Horner form overestimates. Therefore we may assume $a_n > 0$.
- Finally, if f is a monomial, then Horner form never overestimates. Thus, we assume $n > 1$.

Summarizing, throughout this section let

$$\begin{array}{rcl} f(x) & = & a_n x^{d_n} + a_{n-1} x^{d_{n-1}} + \ldots + a_1 x^{d_1} \\ a_i & \in & \mathbb{R}\backslash\{0\}, \quad i = 1, \ldots, n \\ a_n & > & 0 \\ d_1 & > & 0 \\ n & > & 1. \end{array} \tag{3.1.1}$$

3.1.1.1 A Sufficient Condition for Non–Overestimation

In this section we show that if the distance of X and 0 is sufficiently large, then Horner form does not overestimate. The key of the proof is that in this case certain "sub–polynomials" of f are monotone in X. Note however, that monotonicity of f in X is not sufficient for non–overestimation, for example $f(x) = x^2 - x$ is monotone in $X = [0.5, 1]$ but $H_f(X) = [-0.5, 0]$ whereas $f(X) = [-0.25, 0]$.

Let $p_i : \mathbb{R} \to \mathbb{R}$, $i = 0, \ldots, n$ be defined as

$$\begin{array}{rcl} p_n(x) & = & a_n \\ p_{n-1}(x) & = & p_n(x)\ x^{\Delta_n} + a_{n-1} \\ p_{n-2}(x) & = & p_{n-1}(x)\ x^{\Delta_{n-1}} + a_{n-2} \\ & \vdots & \\ p_1(x) & = & p_2(x)\ x^{\Delta_2} + a_1 \\ p_0(x) & = & p_1(x)\ x^{\Delta_1}. \end{array} \tag{3.1.2}$$

Note that $p_0(x) = f(x)$.

Definition 3.1.11 (Overestimation Interval) *The overestimation interval $O_f \in \mathbb{IR}$ of f is defined as*

$$O_f = [\{x \in \mathbb{R} \mid p_i(x) = 0 \text{ for some } n \geq i \geq 0\}]. \ \square$$

In other words, O_f is the smallest interval which contains all roots of the p_i. Note that $p_0(0) = 0$, hence $0 \in O_f$ and

$$\begin{array}{rcl} \overline{O_f} & \geq & 0, \\ \underline{O_f} & \leq & 0. \end{array}$$

Now we are ready for the main theorem of this section.

Theorem 3.1.12 (Non–Overestimation of Horner Form) *For all $X \in \mathbb{IR}$, if*

$$\text{int}(X) \cap O_f = \emptyset$$

then

$$H_f(X) = f(X). \ \square$$

For the proof of Theorem 3.1.12 we need some monotonicity properties of the polynomials p_i, which are subject of Lemma 3.1.13 and Lemma 3.1.14.

In the sequel let

$$\begin{array}{rcl} s_n & = & 1 \\ s_{i-1} & = & (-1)^{\Delta_i} s_i, \quad i = n, \ldots, 1. \end{array}$$

For the inductive proofs below we define $a_0 = 0$.

Lemma 3.1.13 *For $i = 0, \ldots, n$ it holds that*

$$\left.\begin{array}{rcl} p_i'(x) & \geq & 0 \\ p_i(x) & \geq & 0 \end{array}\right\} \quad \text{for all } x \geq \overline{O}_f \tag{3.1.3}$$

$$\left.\begin{array}{rcl} s_i p_i'(x) & \leq & 0 \\ s_i p_i(x) & \geq & 0 \end{array}\right\} \quad \text{for all } x \leq \underline{O}_f. \ \square \tag{3.1.4}$$

Proof.

- We show (3.1.3) by proving inductively for $i = n, \ldots, 0$

$$\left.\begin{array}{rcl} p_i(u) & \geq & p_i(v) \\ p_i(u) & \geq & 0 \end{array}\right\} \quad \text{for all } \ u \geq v \geq \overline{O}_f. \tag{3.1.5}$$

  As p_n is a constant, (3.1.5) holds for $i = n$. Assume (3.1.5) holds for some $i = \hat{i} > 0$.

$$\begin{array}{rcl} p_{\hat{i}-1}(u) & = & p_{\hat{i}}(u) u^{\Delta_{\hat{i}}} + a_{\hat{i}-1} \\ & \geq & p_{\hat{i}}(v) v^{\Delta_{\hat{i}}} + a_{\hat{i}-1} \\ & = & p_{\hat{i}-1}(v). \end{array}$$

  Assume $p_{\hat{i}-1}(u) < 0$. As $p_{\hat{i}-1}(x)$ is non–constant and increasing for $x \geq \overline{O}_f$, there exists $\hat{x} > u \geq \overline{O}_f$ such that $p_{\hat{i}-1}(\hat{x}) = 0$, which contradicts the definition of O_f.

- We show (3.1.4) by proving inductively for $i = n, \ldots, 0$

$$\left.\begin{array}{rcl} s_i p_i(u) & \geq & s_i p_i(v) \\ s_i p_i(u) & \geq & 0 \end{array}\right\} \quad \text{for all } \ u \leq v \leq \underline{O}_f. \tag{3.1.6}$$

  As p_n is a constant, (3.1.6) holds for $i = n$. Assume (3.1.5) holds for some $i = \hat{i} > 0$.

$$\begin{array}{rcl} s_{\hat{i}-1} p_{\hat{i}-1}(u) & = & s_{\hat{i}-1} p_{\hat{i}}(u) u^{\Delta_{\hat{i}}} + s_{\hat{i}-1} a_{\hat{i}-1} \\ & = & s_{\hat{i}} p_{\hat{i}}(u) (-u)^{\Delta_{\hat{i}}} + s_{\hat{i}-1} a_{\hat{i}-1} \\ & \geq & s_{\hat{i}} p_{\hat{i}}(v) (-v)^{\Delta_{\hat{i}}} + s_{\hat{i}-1} a_{\hat{i}-1} \\ & = & s_{\hat{i}-1} p_{\hat{i}}(v) v^{\Delta_{\hat{i}}} + s_{\hat{i}-1} a_{\hat{i}-1} \\ & = & s_{\hat{i}-1} p_{\hat{i}-1}(v) \end{array}$$

  Assume $s_{\hat{i}-1} p_{\hat{i}-1}(u) < 0$. As $s_{\hat{i}-1} p_{\hat{i}-1}(x)$ is non–constant and decreasing for $x \leq \underline{O}_f$, there exists $\hat{x} < u \leq \underline{O}_f$ such that $s_{\hat{i}-1} p_{\hat{i}-1}(\hat{x}) = 0$, which contradicts the definition of O_f. $\square$

We define multivariate polynomials $\tilde{p}_i : \mathbb{R}^{n-i} \to \mathbb{R}$, $i = 0, \ldots, n$ and $\tilde{f} : \mathbb{R}^n \to \mathbb{R}$ as

$$\begin{array}{rcl} \tilde{p}_n() & = & a_n \\ \tilde{p}_{n-1}(x_n) & = & \tilde{p}_n() \; x_n^{\Delta_n} + a_{n-1} \\ \tilde{p}_{n-2}(x_{n-1}, x_n) & = & \tilde{p}_{n-1}(x_n) \; x_{n-1}^{\Delta_{n-1}} + a_{n-2} \\ & \vdots & \\ \tilde{p}_1(x_2, \ldots, x_n) & = & \tilde{p}_2(x_3, \ldots, x_n) \; x_2^{\Delta_2} + a_1 \\ \tilde{p}_0(x_1, \ldots, x_n) & = & \tilde{p}_1(x_2, \ldots, x_n) \; x_1^{\Delta_1} = \tilde{f}(x_1, \ldots, x_n) \end{array} \tag{3.1.7}$$

Note that

$$\{\tilde{f}(x_1, \ldots, x_n) \mid x_1 \in X, \ldots, x_n \in X\} = H_f(X).$$

Lemma 3.1.14

$$\left.\begin{array}{rcl}\tilde{f}(\boldsymbol{u}) & \geq & \tilde{f}(\boldsymbol{v}) \\ \tilde{f}(\boldsymbol{u}) & \geq & 0\end{array}\right\} \quad \text{for all } \boldsymbol{u},\boldsymbol{v}\in\mathbb{R}^n, u_1\geq v_1\geq \overline{O}_f,\ldots,u_n\geq v_n\geq \overline{O}_f \tag{3.1.8}$$

$$\left.\begin{array}{rcl}s_0\tilde{f}(\boldsymbol{u}) & \geq & s_0\tilde{f}(\boldsymbol{v}) \\ s_0\tilde{f}(\boldsymbol{u}) & \geq & 0\end{array}\right\} \quad \text{for all } \boldsymbol{u},\boldsymbol{v}\in\mathbb{R}^n, u_1\leq v_1\leq \underline{O}_f,\ldots,u_n\leq v_n\leq \underline{O}_f \ \square \tag{3.1.9}$$

Proof. For $\boldsymbol{u},\boldsymbol{v}\in\mathbb{R}^n$, $i=0,\ldots,n$ let

$$\begin{array}{rcl}\boldsymbol{u}_i & = & (u_{i+1},u_{i+2},\ldots,u_n)\\ \boldsymbol{v}_i & = & (v_{i+1},v_{i+2},\ldots,v_n).\end{array}$$

- We show (3.1.8) by proving inductively for $i=n,\ldots,0$

$$\left.\begin{array}{rcl}\tilde{p}_i(\boldsymbol{u}_i) & \geq & \tilde{p}_i(\boldsymbol{v}_i) \\ \tilde{p}_i(\boldsymbol{u}_i) & \geq & 0\end{array}\right\} \text{ for all } \boldsymbol{u},\boldsymbol{v}\in\mathbb{R}^n, u_{i+1}\geq v_{i+1}\geq \overline{O}_f,\ldots,u_n\geq v_n\geq \overline{O}_f. \tag{3.1.10}$$

  As $\tilde{p}_n$ is a constant, (3.1.10) holds for $i=n$. Assume (3.1.10) holds for some $i=\hat{i}>0$.

$$\begin{array}{rcl}\tilde{p}_{\hat{i}-1}(\boldsymbol{u}_{\hat{i}-1}) & = & \tilde{p}_{\hat{i}}(\boldsymbol{u}_{\hat{i}})u_{\hat{i}}^{\Delta_{\hat{i}}}+a_{\hat{i}-1}\\ & \geq & \tilde{p}_{\hat{i}}(\boldsymbol{v}_{\hat{i}})v_{\hat{i}}^{\Delta_{\hat{i}}}+a_{\hat{i}-1}\\ & = & \tilde{p}_{\hat{i}-1}(\boldsymbol{v}_{\hat{i}-1}).\\ \tilde{p}_{\hat{i}-1}(\boldsymbol{u}_{\hat{i}-1}) & = & \tilde{p}_{\hat{i}}(\boldsymbol{u}_{\hat{i}})u_{\hat{i}}^{\Delta_{\hat{i}}}+a_{\hat{i}-1}\\ & \geq & \tilde{p}_{\hat{i}}(\overline{O}_f,\ldots,\overline{O}_f)\overline{O}_f{}^{\Delta_{\hat{i}}}+a_{\hat{i}-1}\\ & = & p_{\hat{i}}(\overline{O}_f)\overline{O}_f{}^{\Delta_{\hat{i}}}+a_{\hat{i}-1}\\ & = & p_{\hat{i}-1}(\overline{O}_f)\\ & \geq & 0.\end{array}$$

- We show (3.1.9) by proving inductively for $i=n,\ldots,0$

$$\left.\begin{array}{rcl}s_i\tilde{p}_i(\boldsymbol{u}_i) & \geq & s_i\tilde{p}_i(\boldsymbol{v}_i) \\ s_i\tilde{p}_i(\boldsymbol{u}_i) & \geq & 0\end{array}\right\} \text{ for all } \boldsymbol{u},\boldsymbol{v}\in\mathbb{R}^n, u_{i+1}\leq v_{i+1}\leq \overline{O}_f,\ldots,u_n\leq v_n\leq \underline{O}_f. \tag{3.1.11}$$

  As $\tilde{p}_n$ is a constant, (3.1.11) holds for $i=n$. Assume (3.1.11) holds for some $i=\hat{i}>0$.

$$\begin{array}{rcl}s_{\hat{i}-1}\tilde{p}_{\hat{i}-1}(\boldsymbol{u}_{\hat{i}-1}) & = & s_{\hat{i}-1}\tilde{p}_{\hat{i}}(\boldsymbol{u}_{\hat{i}})u_{\hat{i}}^{\Delta_{\hat{i}}}+s_{\hat{i}-1}a_{\hat{i}-1}\\ & = & s_{\hat{i}}\tilde{p}_{\hat{i}}(\boldsymbol{u}_{\hat{i}})(-u_{\hat{i}})^{\Delta_{\hat{i}}}+s_{\hat{i}-1}a_{\hat{i}-1}\\ & \geq & s_{\hat{i}}\tilde{p}_{\hat{i}}(\boldsymbol{v}_{\hat{i}})(-v_{\hat{i}})^{\Delta_{\hat{i}}}+s_{\hat{i}-1}a_{\hat{i}-1}\\ & = & s_{\hat{i}-1}\tilde{p}_{\hat{i}}(\boldsymbol{v}_{\hat{i}})v_{\hat{i}}^{\Delta_{\hat{i}}}+s_{\hat{i}-1}a_{\hat{i}-1}\\ & = & s_{\hat{i}-1}\tilde{p}_{\hat{i}-1}(\boldsymbol{v}_{\hat{i}-1}).\\ s_{\hat{i}-1}\tilde{p}_{\hat{i}-1}(\boldsymbol{u}_{\hat{i}-1}) & = & s_{\hat{i}-1}\tilde{p}_{\hat{i}}(\boldsymbol{u}_{\hat{i}})u_{\hat{i}}^{\Delta_{\hat{i}}}+s_{\hat{i}-1}a_{\hat{i}-1}\\ & = & s_{\hat{i}}\tilde{p}_{\hat{i}}(\boldsymbol{u}_{\hat{i}})(-u_{\hat{i}})^{\Delta_{\hat{i}}}+s_{\hat{i}-1}a_{\hat{i}-1}\\ & \geq & s_{\hat{i}}\tilde{p}_{\hat{i}}(\underline{O}_f,\ldots,\underline{O}_f)(-\underline{O}_f)^{\Delta_{\hat{i}}}+s_{\hat{i}-1}a_{\hat{i}-1}\\ & = & s_{\hat{i}}p_{\hat{i}}(\underline{O}_f)(-\underline{O}_f)^{\Delta_{\hat{i}}}+s_{\hat{i}-1}a_{\hat{i}-1}\\ & = & s_{\hat{i}-1}p_{\hat{i}-1}(\underline{O}_f)\\ & \geq & 0. \ \square\end{array}$$

Proof of Theorem 3.1.12. Let $X\in\mathbb{IR}$ such that $\mathrm{int}(X)\cap O_f=\emptyset$. Then either $\overline{X}\leq\underline{O}_f$ or $\underline{X}\geq\overline{O}_f$ and according to Lemma 3.1.13 f is monotone in X, i.e.

$$f(X)=[f(\underline{X}),f(\overline{X})].$$

By the definition of $\tilde{f}$ it holds that

$$\begin{aligned} f(\underline{X}) &= \tilde{f}(\underline{X}, \ldots, \underline{X}) \\ f(\overline{X}) &= \tilde{f}(\overline{X}, \ldots, \overline{X}). \end{aligned}$$

Finally, it follows from Lemma 3.1.14 that

$$[\tilde{f}(\underline{X}, \ldots, \underline{X}), \tilde{f}(\overline{X}, \ldots, \overline{X})] = \{\tilde{f}(x_1, \ldots, x_n) \mid x_1 \in X, \ldots, x_n \in X\} = H_f(X).$$

Hence

$$f(X) = H_f(X). \square$$

3.1.1.2 Optimality of the Non–Overestimation Condition

In the previous section we proved the sufficient condition $\mathrm{int}(X) \cap O_f = \emptyset$ for $H_f(X) = f(X)$. This condition is not necessary in general. For example let $f(x)$ such that $a_i > 0$ and Δ_i is even for all i. In this case $O_f = \{0\}$, but $H_f(X) = f(X)$ for all X in $\mathbb{IR}$, even if $\mathrm{int}(X) \cap O_f \neq \emptyset$. However, in this section we show that O_f is minimal in the sense that there exists no interval Q_f such that $Q_f \not\supseteq O_f$ and $\mathrm{int}(X) \cap Q_f = \emptyset$ still implies $H_f(X) = f(X)$. Throughout this section let $f(x)$ as in (3.1.1).

Theorem 3.1.15 (Optimality of the Overestimation Interval) *Assume $a_i < 0$ or Δ_i is odd for at least one $i \in \{1, \ldots, n\}$. Let $Q_f \in \mathbb{IR}$ such that $Q_f \not\supseteq O_f$. Then there exists $X \in \mathbb{IR}$ such that*

$$\mathrm{int}(X) \cap Q_f = \emptyset \quad \text{and} \quad H_f(X) \supset f(X). \ \square \tag{3.1.12}$$

For the proof of Theorem 3.1.15 we distinguish the case $O_f = \{0\}$, which is treated in Lemma 3.1.16 and Lemma 3.1.17, and the case $O_f \neq \{0\}$, which is treated in Lemma 3.1.18, Lemma 3.1.19 and Lemma 3.1.20. In the sequel let $Q_f \in \mathbb{IR}$ such that $Q_f \not\supseteq O_f$.

Lemma 3.1.16 *If $O_f = \{0\}$ then*

(i) *$a_i > 0$ for $i = 1, \ldots, n$ and*

(ii) *Δ_i is even for $i = 2, \ldots, n$. $\square$*

Proof. Assume $O_f = \{0\}$.

(i) Let $n \geq i > 0$ arbitrary but fixed. From the definition of O_f it follows that $p_i(x) \neq 0$ for $x \neq 0$. Further, $p_i(0) = a_i \neq 0$, hence $p_i(x) \neq 0$ for all $x \in \mathbb{R}$. As the leading coefficient of p_i is $a_n > 0$, it holds that $p_i(x) > 0$ for sufficiently large x. Hence $p_i(x) > 0$ for all x, and $a_i = p_i(0) > 0$.

(ii) Assume Δ_i is odd for some $i > 1$ and let i be the largest index such that Δ_i is odd. Then the degree of

$$p_i(x) = \big(\ldots (a_n x^{\Delta_n} + a_{n-1}) x^{\Delta_{n-1}} + \ldots + a_{i+1} \big) x^{\Delta_{i+1}} + a_i$$

is even and the degree of

$$p_{i-1}(x) = p_i(x) x^{\Delta_i} + a_{i-1}$$

is odd. Hence $p_{i-1}(x) < 0$ if x is small enough, which is a contradiction to $p_i(x) > 0$ for all x and $i > 0$. $\square$

Thus, in order to prove Theorem 3.1.15 for the case $O_f = \{0\}$ it remains to show the following lemma:

Lemma 3.1.17 *Assume $O_f = \{0\}$ and Δ_1 is odd. Then there exists $X \in \mathbb{IR}$ such that*

$$\mathrm{int}(X) \cap Q_f = \emptyset \quad \text{and} \quad H_f(X) \supset f(X). \ \square$$

Proof. Assume $O_f = \{0\}$ and Δ_1 is odd. From Lemma 3.1.16 it follows that $a_i > 0$ for $i = 1, \ldots, n$ and Δ_i is even for $i = 2, \ldots, n$. Hence the degree of f is odd and from Lemma 3.1.13 it follows that $f(x)$ is monotonically increasing. Thus

$$\underline{f(X)} = f(\underline{X})$$

for all $X \in \mathbb{IR}$. As $0 \notin Q_f$ there exists $X \in \mathbb{IR}$ such that

$$\begin{aligned} \mathrm{int}(X) \cap Q_f &= \emptyset \\ 0 &\in \mathrm{int}(X) \\ \overline{X} &> -\underline{X}. \end{aligned}$$

The monomials of $p_1(x)$ have even power and positive coefficient, hence

$$p_1(\overline{X}) > p_1(\underline{X}) > 0.$$

As $\underline{X}^{\Delta_1} < 0$, it holds that

$$\begin{aligned} f(\underline{X}) &= p_1(\underline{X})\underline{X}^{\Delta_1} \\ &> p_1(\overline{X})\underline{X}^{\Delta_1} \\ &= \tilde{f}(\underline{X}, \overline{X}, \overline{X}, \ldots, \overline{X}). \end{aligned}$$

Hence

$$\underline{H_f(X)} < \underline{f(X)},$$

which implies $H_f(X) \supset f(X)$. □

The following lemmas are a preparation for the proof of Theorem 3.1.15 for the case $O_f \neq \{0\}$.

Lemma 3.1.18

(i) *If $x \geq 0$ and $p_i(x) \geq 0$ for all $i = 1, \ldots, n$ then $x \geq \overline{O}_f$.*

(ii) *If $x \leq 0$ and $s_i p_i(x) \geq 0$ for all $i = 1, \ldots, n$ then $x \leq \underline{O}_f$.* □

Proof.

(i) Assume $x \geq 0$ and $p_i(x) \geq 0$ for all $i = 1, \ldots, n$. According to the definition of O_f we have to show that $p_i(y) \neq 0$ for all i and for all $y > x$. Hence, let $y > x$ arbitrary but fixed. Obviously $p_n(y) = a_n > 0$. By induction we show for $i = n-1, \ldots, 0$

$$p_i(y) > p_i(x). \tag{3.1.13}$$

As $p_{n-1}(y) = a_n y^{\Delta_n} + a_{n-1} > a_n x^{\Delta_n} + a_{n-1} = p_{n-1}(x)$, (3.1.13) holds for $i = n-1$. Assume (3.1.13) holds for some $i = \hat{i} > 0$.

$$\begin{aligned} p_{\hat{i}-1}(y) &= p_{\hat{i}}(y) y^{\Delta_{\hat{i}}} + a_{\hat{i}-1} \\ &> p_{\hat{i}}(x) x^{\Delta_{\hat{i}}} + a_{\hat{i}-1} \\ &= p_{\hat{i}-1}(x). \end{aligned}$$

Hence, $p_i(y) \neq 0$ for $i = 0, \ldots, n$.

(ii) Assume $x \leq 0$ and $s_i p_i(x) \geq 0$ for all $i = 1, \ldots, n$. According to the definition of O_f we have to show that $p_i(y) \neq 0$ for all i and for all $y < x$. Hence, let $y < x$ arbitrary but fixed. Obviously $p_n(y) = a_n > 0$. By induction we show for $i = n-1, \ldots, 0$

$$s_i p_i(y) > s_i p_i(x). \tag{3.1.14}$$

As $s_{n-1} p_{n-1}(y) = a_n (-y)^{\Delta_n} + s_{n-1} a_{n-1} > a_n (-x)^{\Delta_n} + s_{n-1} a_{n-1} = s_{n-1} p_{n-1}(x)$, (3.1.14) holds for $i = n-1$. Assume (3.1.14) holds for some $i = \hat{i} > 0$.

$$\begin{aligned} s_{\hat{i}-1} p_{\hat{i}-1}(y) &= s_{\hat{i}} p_{\hat{i}}(y) (-y)^{\Delta_{\hat{i}}} + s_{\hat{i}-1} a_{\hat{i}-1} \\ &> s_{\hat{i}} p_{\hat{i}}(x) (-x)^{\Delta_{\hat{i}}} + s_{\hat{i}-1} a_{\hat{i}-1} \\ &= s_{\hat{i}-1} p_{\hat{i}-1}(x). \end{aligned}$$

Hence, $p_i(y) \neq 0$ for $i = 0, \ldots, n$. □

In the sequel let

$$\tilde{f}_i(x_1,\ldots,x_n) = \frac{\partial}{\partial x_i}\tilde{f}(x_1,\ldots,x_n), \quad i=1,\ldots,n$$

be the partial derivatives of $\tilde{f}$. Note that

$$\begin{aligned}\tilde{f}_i(x_1,\ldots,x_n) &= x_1^{\Delta_1}x_2^{\Delta_2}\ldots x_{i-1}^{\Delta_{i-1}}\frac{\partial}{\partial x_i}\tilde{p}_{i-1}(x_i,\ldots,x_n)\\ &= x_1^{\Delta_1}x_2^{\Delta_2}\ldots x_{i-1}^{\Delta_{i-1}}\Delta_i x_i^{\Delta_i-1}\tilde{p}_i(x_{i+1},\ldots,x_n),\end{aligned}$$

where $\tilde{p}_i$, $\tilde{f}$ is as in (3.1.7).

Lemma 3.1.19 *Let $X \in \mathbb{IR}\backslash\mathbb{R}$ and let $x \in X$. If there exist $i,j \in \{1,\ldots,n\}$ such that*

$$\tilde{f}_i(x,\ldots,x)\tilde{f}_j(x,\ldots,x) < 0$$

then $\tilde{f}(x,\ldots,x)$ is not an extremum of $\tilde{f}$ in X. □

Proof. Assume $\tilde{f}(x,\ldots,x)$ is an extremum of $\tilde{f}$ in X, $\underline{X} \neq \overline{X}$ and $\tilde{f}_i(x,\ldots,x) > 0$, $\tilde{f}_j(x,\ldots,x) < 0$ for some $i,j \in \{1,\ldots,n\}$.

- As the i-th and the j-th partial derivative of $\tilde{f}$ are not vanishing at $(x,\ldots,x)$ it follows that $\tilde{f}(x,\ldots,x)$ is a boundary extremum, hence $x \notin \mathrm{int}(X)$.
- Assume $x = \overline{X}$. Then $\tilde{f}(x,\ldots,x)$ cannot be a maximum of $\tilde{f}$ in X because $\tilde{f}_j(x,\ldots,x) < 0$. Further $\tilde{f}(x,\ldots,x)$ cannot be a minimum of $\tilde{f}$ in X because $\tilde{f}_i(x,\ldots,x) > 0$. Hence $x \neq \overline{X}$.
- Assume $x = \underline{X}$. Then $\tilde{f}(x,\ldots,x)$ cannot be a maximum of $\tilde{f}$ in X because $\tilde{f}_i(x,\ldots,x) > 0$. Further $\tilde{f}(x,\ldots,x)$ cannot be a minimum of $\tilde{f}$ in X because $\tilde{f}_j(x,\ldots,x) < 0$. Hence $x \neq \overline{X}$. □

Lemma 3.1.20 *Assume $O_f \neq \{0\}$. Then there exists $X \in \mathbb{IR}\backslash\mathbb{R}$ such that*

$$\mathrm{int}(X) \cap Q_f = \emptyset$$

and for all $x \in X$ there exists $i \in \{1,\ldots,n\}$ such that

$$\tilde{f}_i(x,\ldots,x)\tilde{f}_n(x,\ldots,x) < 0.$$ □

Proof. Assume $O_f \neq \{0\}$. Then $\underline{O_f} \neq \overline{O_f}$ and there exists an interval $X \in \mathbb{IR}\backslash\mathbb{R}$ such that

$$\begin{aligned} X &\subseteq \mathrm{int}(O_f),\\ 0 &\notin X \text{ and}\\ \mathrm{int}(X) \cap Q_f &= \emptyset.\end{aligned}$$

We distinguish the cases $\underline{X} > 0$ and $\overline{X} < 0$.

- Assume $\underline{X} > 0$. As $\overline{X} < \overline{O_f}$ and $\underline{X} > 0$ it follows from Lemma 3.1.18 that for all $x \in X$ there exists $i \in \{1,\ldots,n\}$ such that $p_i(x) < 0$. Hence,

$$\begin{aligned}\tilde{f}_i(x,\ldots,x) &= x^{\Delta_1}x^{\Delta_2}\ldots x^{\Delta_{i-1}}\Delta_i x^{\Delta_i-1}\tilde{p}_i(x,\ldots,x)\\ &= \Delta_i x^{d_i-1}p_i(x)\\ &< 0.\\ \tilde{f}_n(x,\ldots,x) &= x^{\Delta_1}x^{\Delta_2}\ldots x^{\Delta_{n-1}}\Delta_n x^{\Delta_n-1}\tilde{p}_n()\\ &= \Delta_n x^{d_n-1}a_n\\ &> 0.\end{aligned}$$

- Assume $\overline{X} < 0$. As $\underline{X} > \underline{Q_f}$ and $\overline{X} < 0$ it follows from Lemma 3.1.18 that for all $x \in X$ there exists $i \in \{1, \ldots, n\}$ such that $s_i p_i(x) < 0$. Note that

$$\begin{aligned} s_0 &= (-1)^{d_n} \\ s_i &= (-1)^{d_n - d_i}, \quad i = 1, \ldots, n. \end{aligned}$$

Hence,

$$\begin{aligned} \tilde{f}_i(x, \ldots, x) &= x^{\Delta_1} x^{\Delta_2} \ldots x^{\Delta_{i-1}} \Delta_i x^{\Delta_i - 1} \tilde{p}_i(x, \ldots, x) \\ &= \Delta_i x^{d_i - 1} p_i(x) \\ &= \Delta_i (-1)^{d_n - d_i} x^{d_i - 1} s_i p_i(x) \\ &= \Delta_i (-1)^{d_n - 1} (-x)^{d_i - 1} s_i p_i(x) \\ &\quad \begin{cases} < 0 & \text{if } d_n \text{ is odd} \\ > 0 & \text{else.} \end{cases} \end{aligned}$$

$$\begin{aligned} \tilde{f}_n(x, \ldots, x) &= x^{\Delta_1} x^{\Delta_2} \ldots x^{\Delta_{n-1}} \Delta_n x^{\Delta_n - 1} \tilde{p}_n() \\ &= \Delta_n x^{d_n - 1} a_n \\ &\quad \begin{cases} < 0 & \text{if } d_n \text{ is even} \\ > 0 & \text{else.} \end{cases} \end{aligned}$$

□

The proof of Theorem 3.1.15 follows now easily from Lemma 3.1.16 – 3.1.20.

Proof of Theorem 3.1.15. Let f and Q_f as in Theorem 3.1.15.

- Assume $O_f = \{0\}$. From Lemma 3.1.16 it follows that $a_i > 0$ for $i = 1, \ldots, n$ and Δ_i is even for $i = 2, \ldots, n$. Hence, Δ_1 must be odd and $H_f(X) \supset f(X)$ follows from Lemma 3.1.17.
- Assume $O_f \neq \{0\}$. According to Lemma 3.1.20 there exists $X \in \mathbb{IR} \backslash \mathbb{R}$ such that $\text{int}(X) \cap Q_f = \emptyset$ and for all $x \in X$ there exists $i \in \{1, \ldots, n\}$ such that

$$\tilde{f}_i(x, \ldots, x) \tilde{f}_n(x, \ldots, x) < 0.$$

According to Lemma 3.1.19, for all $x \in X$, $\tilde{f}(x, \ldots, x)$ is not an extremum of $\tilde{f}$ in X. Thus, there exist $u_1, \ldots, u_n, v_1, \ldots, v_n \in X$ such that

$$\begin{aligned} \tilde{f}(u_1, \ldots, u_n) &< \underline{f(X)} \\ \tilde{f}(v_1, \ldots, v_n) &> \overline{f(X)}, \end{aligned}$$

hence

$$H_f(X) \supset f(X). \; \square$$

3.1.1.3 Algorithmic Test of the Non–Overestimation Condition

If $\text{int}(X) \cap O_f = \emptyset$, then Horner form gives the range of f on X without overestimation. As the computation of O_f is expensive, we are interested in methods to decide whether $\text{int}(X)$ and O_f are disjoint, without actually computing O_f. The following theorem shows how this can be done efficiently. Throughout this section let f as in (3.1.1).

Theorem 3.1.21 *Let $X \in \mathbb{IR} \backslash \mathbb{R}$. Then*

$$\text{int}(X) \cap O_f = \emptyset$$

if and only if

$$\begin{aligned} \underline{X} &\geq 0 \quad \text{and} \\ p_i(\underline{X}) &\geq 0 \quad \text{for } i = 1, \ldots, n \end{aligned}$$

or

$$\begin{aligned} \overline{X} &\leq 0 \quad \text{and} \\ s_i p_i(\overline{X}) &\geq 0 \quad \text{for } i = 1, \ldots, n. \; \square \end{aligned}$$

Proof. Let $X \in \mathbb{IR}$, $\underline{X} \neq \overline{X}$.

"$\Rightarrow$" Follows from Lemma 3.1.13. If $\mathrm{int}(X) \cap O_f = \emptyset$ then either $\underline{X} \geq \overline{O}_f$ or $\overline{X} \leq \underline{O}_f$.

- If $\underline{X} \geq \overline{O}_f$, then $\underline{X} \geq 0$ and $p_i(\underline{X}) \geq 0$ for all $i = 1, \ldots, n$.
- If $\overline{X} \leq \underline{O}_f$, then $\overline{X} \leq 0$ and $s_i p_i(\overline{X}) \geq 0$ for all $i = 1, \ldots, n$.

"$\Leftarrow$" Follows from Lemma 3.1.18.

- If $\underline{X} \geq 0$ and $p_i(\underline{X}) \geq 0$ for $i = 1, \ldots, n$, then $\underline{X} \geq \overline{O}_f$, hence $\mathrm{int}(X) \cap O_f = \emptyset$.
- If $\overline{X} \leq 0$ and $s_i p_i(\overline{X}) \geq 0$ for $i = 1, \ldots, n$, then $\overline{X} \leq \underline{O}_f$, hence $\mathrm{int}(X) \cap O_f = \emptyset$. □

If Theorem 3.1.21 is used for testing disjointness of $\mathrm{int}(X)$ and O_f one has to evaluate the p_i on the endpoints of X. The next theorem shows how this computation can be avoided by using intermediate results of the Horner evaluation of f on X instead. In the sequel let

$$H_{p_i} : \mathbb{IR} \to \mathbb{IR}$$

be the Horner form of p_i and let O_{p_i} be the overestimation interval of p_i, $i = 1, \ldots, n$.

Theorem 3.1.22 *Let $X \in \mathbb{IR}\backslash\mathbb{R}$. Then*

$$\mathrm{int}(X) \cap O_f = \emptyset$$

if and only if

$$\begin{array}{rcll} \underline{X} & \geq & 0 & \text{and} \\ \underline{H_{p_i}(X)} & \geq & 0 & \text{for } i = 1, \ldots, n \end{array}$$

or

$$\begin{array}{rcll} \overline{X} & \leq & 0 & \text{and} \\ \underline{s_i H_{p_i}(X)} & \geq & 0 & \text{for } i = 1, \ldots, n. \ \square \end{array}$$

Proof. Let $X \in \mathbb{IR}\backslash\mathbb{R}$.

"$\Rightarrow$" Assume $\mathrm{int}(X) \cap O_f = \emptyset$. Then either $\underline{X} \geq \overline{O}_f$ or $\overline{X} \leq \underline{O}_f$. As $O_f \supseteq O_{p_i}$, it holds that $H_{p_i}(X) = p_i(X)$ for $i = 1, \ldots, n$.

- If $\underline{X} \geq \overline{O}_f$ then $\underline{X} \geq 0$ and $\underline{p_i(X)} = p_i(\underline{X}) \geq 0$ by Lemma 3.1.13, hence $\underline{H_{p_i}(X)} \geq 0$.
- If $\overline{X} \leq \underline{O}_f$ then $\overline{X} \leq 0$ and $\underline{s_i p_i(X)} = s_i p_i(\overline{X}) \geq 0$ by Lemma 3.1.13, hence $\underline{s_i H_{p_i}(X)} \geq 0$.

"$\Leftarrow$" Follows from Theorem 3.1.21 and $H_{p_i}(X) \supseteq p_i(X)$. □

Algorithm 3.1.23 is a modification of Algorithm 3.1.7 which, in addition to the computation of $H_f(X)$, decides whether $\mathrm{int}(X) \cap O_f = \emptyset$. The decision does not cost any additional arithmetic operations.

Algorithm 3.1.23 (HFOT) [Horner Form with Overestimation Test]

In: $f(x) = a_n x^{d_n} + a_{n-1} x^{d_{n-1}} + \ldots + a_1 x^{d_1} \in \mathbb{F}[x]$,
$X \in \mathbb{IF}$.

Out: $\mathrm{HF}(f, X)$,
$t \in \{\texttt{true}, \texttt{false}\}$, if $\mathrm{int}(X) \cap O_f \neq \emptyset$ then $t = \texttt{false}$.

(1) [Initialize.]
$P_n \longleftarrow a_n$.

(2) [Accumulate.]
for $i = n-1, \ldots, 1$
$P_i \longleftarrow P_{i+1} X^{d_{i+1} - d_i} + a_i$.

(3) [Last power.]
$P_0 \longleftarrow P_1 X^{d_1}$.

(4) [Trivial cases for disjointness decision.]
if $0 \in \mathrm{int}(X)$ then $t \longleftarrow$ `false`, goto 8.
if $\underline{X} = \overline{X}$ then $t \longleftarrow$ `true`, goto 8.

(5) [If $d_1 = 0$ then ignore a_1.]
if $d_1 = 0$ then $l \longleftarrow 2$, else $l \longleftarrow 1$.

(6) [Signs.]
$s_n = \mathrm{sign}(a_n)$.
if $\underline{X} \geq 0$
for $i = n-1, \ldots, l$ do $s_i \longleftarrow s_{i+1}$.
if $\overline{X} \leq 0$
for $i = n-1, \ldots, l$ do $s_i \longleftarrow (-1)^{d_{i+1}-d_i} s_{i+1}$.

(7) [Decide disjointness of $\mathrm{int}(X)$ and O_f.]
$t \longleftarrow$ `true`.
for $i = l, \ldots, n-1$
if $\underline{s_i P_i} < 0$ then $t \longleftarrow$ `false`.

(8) [Return.]
return P_0, t.

Remark. If exact arithmetic is used in Algorithm 3.1.23 then $t =$ `true` if and only if $\mathrm{int}(X) \cap O_f = \emptyset$. □

If $Q_f \supseteq O_f$, then a sufficient condition for non–overestimation of the Horner form is $\mathrm{int}(X) \cap Q_f = \emptyset$. Thus, it is sometimes useful to know an overestimations of O_f. Overestimations of O_f can be obtained cheaply by the use of root bound theorems.

Theorem 3.1.24 *Let*

$$f(x) = \sum_{i=0}^{d_n} a_i^* x^i = \sum_{i=1}^{n} a_i x^{d_i}.$$

If $f(r) = 0$ for some $r \in \mathbb{R}$ then

$$\begin{aligned} |r| &\leq \max\{|a_0^*/a_{d_n}^*|, 1 + |a_1^*/a_{d_n}^*|, \ldots, 1 + |a_{d_n-1}^*/a_{d_n}^*|\} \\ |r| &\leq \max\{1, \sum_{i=0}^{d_n-1} |a_i^*/a_{d_n}^*|\} \\ |r| &\leq 2\max\{|a_{d_n-1}^*/a_{d_n}^*|, |a_{d_n-2}^*/a_{d_n}^*|^{1/2}, |a_{d_n-3}^*/a_{d_n}^*|^{1/3}, \ldots, |a_0^*/a_{d_n}^*|^{1/d_n}\}. \ \square \end{aligned}$$

Proof. See for example [Householder, 1970]. □

If $b \geq |r|$ for all real roots r of f, then b is called a root bound of f. If b is a common root bound of $p_0, \ldots, p_n$ then $[-b, b] \supseteq O_f$.

Corollary 3.1.25 (Bounds for Overestimation Interval) *If $b \in \mathbb{R}$ satisfies one of the conditions*

$$\begin{aligned} b &\geq \max\{1 + |a_1/a_n|, 1 + |a_2/a_n|, \ldots, 1 + |a_{n-1}/a_n|\} \\ b &\geq \max\{1, \sum_{i=1}^{n-1} |a_i/a_n|\} \\ b &\geq 2\max\{|a_{n-1}/a_n|^{1/(d_n-d_{n-1})}, |a_{n-2}/a_n|^{1/(d_n-d_{n-2})}, \ldots, |a_1/a_n|^{1/(d_n-d_1)}\} \end{aligned}$$

then

$$[-b, b] \supseteq O_f. \ \square$$

Proof. If b satisfies one of the conditions of Corollary 3.1.25, then b is a common root bound of $p_0, \ldots, p_n$. □

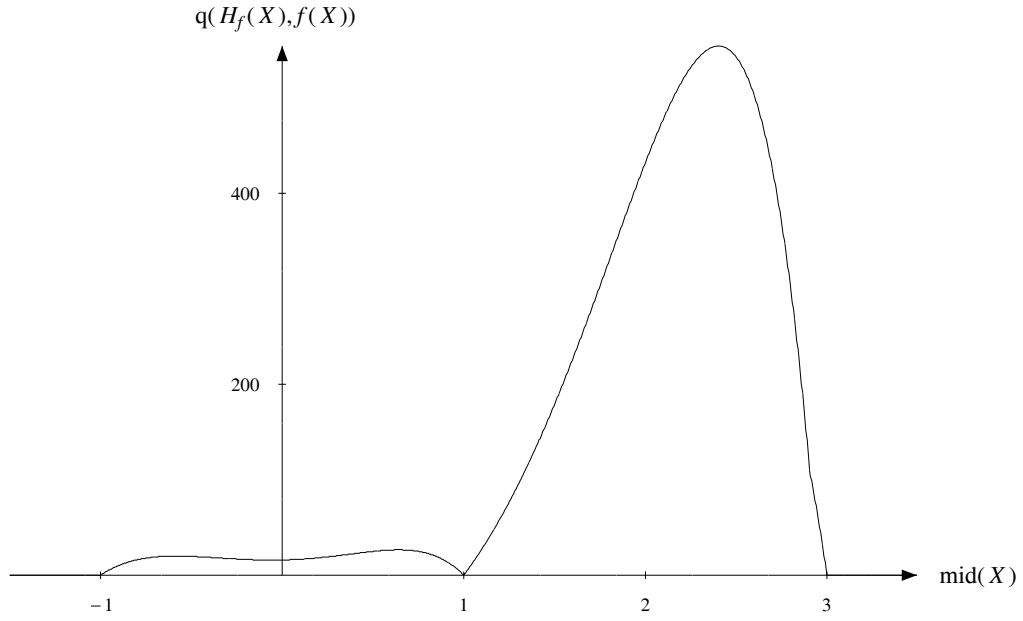


Figure 3.1.1: $q(H_f(X), f(X))$ in dependence of $\mathrm{mid}(X)$, where $f(x) = x^7 - 8x^4 + 64x$ and $w(X) \equiv 2$.

3.1.1.4 Further Cases where Horner Form is Exact

In Section 3.1.1.1 we proved the sufficient condition $\mathrm{int}(X) \cap O_f = \emptyset$ for $H_f(X) = f(X)$. In the following we study cases where $H_f(X) = f(X)$ but $\mathrm{int}(X) \cap O_f \neq \emptyset$. A complete characterization of all cases when Horner form is exact is still unknown. The results in this section are original.

Example. Consider the polynomial $f(x) = x^7 - 8x^4 + 64x$ and let $X = [0, 2]$. We obtain $O_f = [0, 2]$, hence $\mathrm{int}(X) \cap O_f \neq \emptyset$, i.e. the non–overestimation condition of Theorem 3.1.12 is not satisfied. Still,

$$H_f(X) = [0, 128] = f(X).$$

Figure 3.1.1 shows $q(H_f(X), f(X))$ for $f(x) = x^7 - 8x^4 + 64x$ in dependence of $\mathrm{mid}(X)$, where $w(X) \equiv 2$. According to Theorem 3.1.12, $H_f(X) = f(X)$ if $\mathrm{mid}(X) \geq 3$ or $\mathrm{mid}(X) \leq -1$. As one can see, the non–overestimation for $\mathrm{mid}(X) = 1$ is a "special case". The following Theorem gives a generalization.

Theorem 3.1.26 *If there exists $c \in \mathbb{R}$ and $\Delta \in \mathbb{N}$ such that*

$$a_i = ca_{i+1}, \quad d_i = d_{i+1} - \Delta, \quad i = 1, \ldots, n-1,$$

$d_1 > 0$, n is odd, and

$$X = [0, \sqrt[\Delta]{|c|}] \quad \text{or} \quad X = [-\sqrt[\Delta]{|c|}, 0],$$

then

$$H_f(X) = f(X). \ \square$$

For the example $f(x) = x^7 - 8x^4 + 64x$ above, we obtain $c = -8$, $\Delta = 3$, $d_0 = 1$ and by Theorem 3.1.26, $H_f(X) = f(X)$ for $X = [0, 2]$ and for $X = [-2, 0]$.

Proof. Let $f(x)$ as in Theorem 3.1.26. According to Theorem 3.1.10 we may assume $a_n = 1$. Then

$$f(x) = \left(\ldots ((x^\Delta + c)x^\Delta + c^2)x^\Delta + \ldots + c^{n-1} \right) x^\Delta x^{d_0}$$

where $d_0 = d_1 - \Delta$. Let

$$\begin{aligned} g_1(x) &= x^\Delta \\ g_{i+1}(x) &= (g_i(x) + c^i)x^\Delta, \quad i = 1, \ldots, n-1. \end{aligned}$$

and let $H_{g_i}(X) : \mathbb{IR} \to \mathbb{IR}$ be the Horner form of $g_i(x)$, $i = 1, \ldots, n$. Then

$$\begin{aligned} f(x) &= g_n(x)x^{d_0} \\ H_f(X) &\subseteq H_{g_n}(X)X^{d_0}, \end{aligned}$$

and it suffices to show

$$H_{g_n}(X)X^{d_0} \subseteq f(X) \text{ for } X = [0, \sqrt[\triangle]{-c}] \text{ and } X = [0, -\sqrt[\triangle]{-c}].$$

We distinguish 8 cases:

- $c \geq 0$, $X = [0, \sqrt[\triangle]{c}]$, Δ even.
  $O_f = \{0\}$, hence $\text{int}(X) \cap O_f = \emptyset$ and $H_f(X) = f(X)$ by Theorem 3.1.12.
- $c \geq 0$, $X = [0, \sqrt[\triangle]{c}]$, Δ odd.
  $\overline{O}_f = 0$, hence $\text{int}(X) \cap O_f = \emptyset$ and $H_f(X) = f(X)$ by Theorem 3.1.12.
- $c \geq 0$, $X = [-\sqrt[\triangle]{c}, 0]$, Δ even.
  $O_f = \{0\}$, hence $\text{int}(X) \cap O_f = \emptyset$ and $H_f(X) = f(X)$ by Theorem 3.1.12.
- $c \geq 0$, $X = [-\sqrt[\triangle]{c}, 0]$, Δ odd.
  By induction it follows for $i = 1, \ldots, n$

  $$\begin{aligned} H_{g_i}(X) &= [-c^i, 0] \\ g_i(\underline{X}) &= \begin{cases} -c^i, & i \text{ odd} \\ 0, & i \text{ even} \end{cases} \\ g_i(\overline{X}) &= 0. \end{aligned}$$

  Hence, as n is odd

  $$\begin{aligned} H_{g_n}(X)X^{d_0} &= [-c^n, 0][\underline{X}^{d_0}, 0] \\ &= [g_n(\underline{X}), 0][\underline{X}^{d_0}, 0] \\ &= [g_n(\underline{X})\underline{X}^{d_0}, 0] \\ &= [f(\underline{X}), 0] \\ &\subseteq f(X). \end{aligned}$$

- $c < 0$, $X = [0, \sqrt[\triangle]{-c}]$, Δ even.
  By induction it follows for $i = 1, \ldots, n$

  $$\begin{aligned} H_{g_i}(X) &= [0, -c^i] \\ g_i(\underline{X}) &= 0 \\ g_i(\overline{X}) &= \begin{cases} -c^i, & i \text{ odd} \\ 0, & i \text{ even.} \end{cases} \end{aligned}$$

  Hence, as n is odd

  $$\begin{aligned} H_{g_n}(X)X^{d_0} &= [0, -c^n][0, \overline{X}^{d_0}] \\ &= [0, g_n(\overline{X})][0, \overline{X}^{d_0}] \\ &= [0, g_n(\overline{X})\overline{X}^{d_0}] \\ &= [0, f(\overline{X})] \\ &\subseteq f(X). \end{aligned}$$

- $c < 0$, $X = [0, \sqrt[\triangle]{-c}]$, Δ odd.
  This case is equivalent to the previous case.

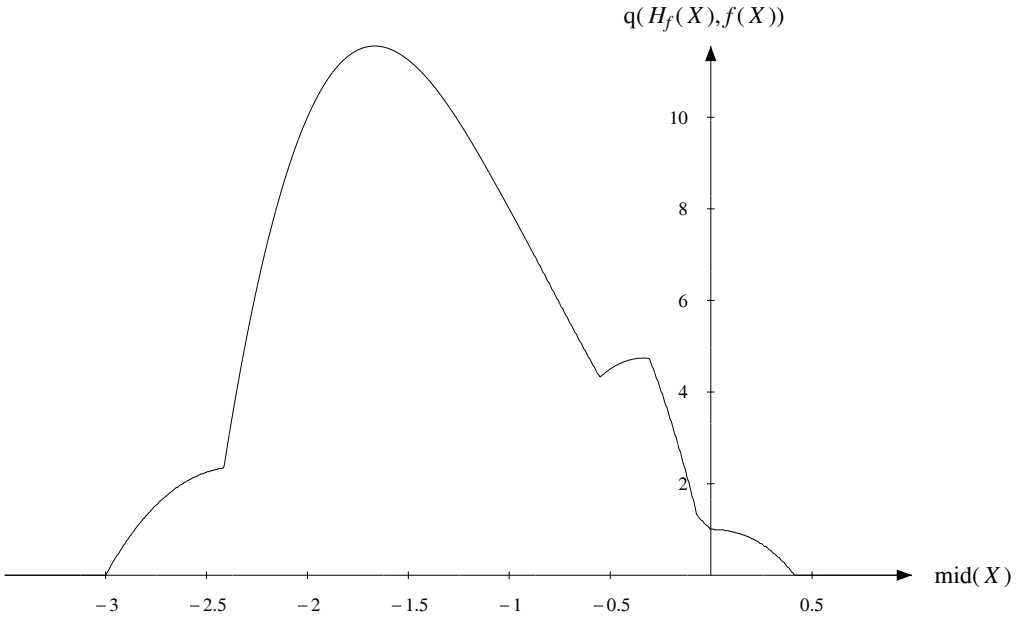


Figure 3.1.2: $\mathrm{q}(H_f(X), f(X))$ in dependence of $\mathrm{mid}(X)$, where $f(x) = x^4 + 2x^3 + 2x^2$ and $\mathrm{w}(X) \equiv 2$.

- $c < 0$, $X = [-\sqrt[\Delta]{-c}, 0]$, Δ even.
  By induction it follows for $i = 1, \ldots, n$

$$\begin{aligned} H_{g_i}(X) &= [0, -c^i] \\ g_i(\underline{X}) &= \begin{cases} -c^i, & i \text{ odd} \\ 0, & i \text{ even} \end{cases} \\ g_i(\overline{X}) &= 0. \end{aligned}$$

  Hence, as n is odd

$$\begin{aligned} H_{g_n}(X)X^{d_0} &= [0, -c^n][\underline{X}^{d_0}, 0] \\ &= [0, g_n(\underline{X})][\underline{X}^{d_0}, 0] \\ &= [g_n(\underline{X})\underline{X}^{d_0}, 0] \\ &= [f(\underline{X}), 0] \\ &\subseteq f(X). \end{aligned}$$

- $c < 0$, $X = [-\sqrt[\Delta]{-c}, 0]$, Δ odd.
  $\underline{O}_f = 0$, hence $\mathrm{int}(X) \cap O_f = \emptyset$ and $H_f(X) = f(X)$ by Theorem 3.1.12. □

Let us consider another example where Horner form is exact.

Example. Consider the polynomial $f(x) = x^4 + 2x^3 + 2x^2$ and let $X = [-0.5, 1.5]$. We obtain $O_f = [-2, 0]$, hence $\mathrm{int}(X) \cap O_f \neq \emptyset$, i.e. the non–overestimation condition of Theorem 3.1.12 is not satisfied. Still,

$$H_f(X) = [0, 16.3125] = f(X).$$

Figure 3.1.2 shows $\mathrm{q}(H_f(X), f(X))$ for $f(x) = x^4 + 2x^3 + 2x^2$ in dependence of $\mathrm{mid}(X)$, where $\mathrm{w}(X) \equiv 2$. According to Theorem 3.1.12, $H_f(X) = f(X)$ if $\mathrm{mid}(X) \geq 1$ or $\mathrm{mid}(X) \leq -3$. The observation that $H_f(X) = f(X)$ already for $\mathrm{mid}(X) \geq \sqrt{2} - 1 \approx 0.41$ is generalized by the following Theorem. □

In the sequel let $p_i, \tilde{p}_i, i = 0, \ldots, n$ as in (3.1.2) respectively (3.1.7), and let $H_{p_1}(X)$ be the Horner form of $p_1(x)$.

Theorem 3.1.27 *If $a_i > 0$ for $i = 1, \ldots, n$, $d_1 > 0$ is even, $\mid(X) \geq 0$ and $\underline{H_{p_1}(X)} \geq 0$, then*

$$H_f(X) = f(X). \ \square$$

For the example $f(x) = x^4 + 2x^3 + 2x^2$, $X = [-0.5, 1.5]$ above, we obtain

$$H_{p_1}(X) = (X + 2)X + 2 = [0.25, 7.25],$$

hence $H_f(X) = f(X)$ by Theorem 3.1.27. For the proof of Theorem 3.1.27 we need the following two lemmas.

Lemma 3.1.28 *If $a_i > 0$ for $i = 1, \ldots, n$, then $\overline{O}_f = 0$.* $\square$

Proof. Assume $a_i > 0$ for $i = 1, \ldots, n$. Then all coefficients of $p_i(x)$, are positive and $p_i(x)$ has no positive roots for $i = 0, \ldots, n$. Hence $\overline{O}_f = 0$. $\square$

Lemma 3.1.29 *If $a_i > 0$ for $i = 1, \ldots, n$ and $\mid(X) \geq 0$ then*

$$\overline{H_{p_1}(X)} = p_1(\overline{X}). \ \square$$

Proof. Assume $a_i > 0$ for $i = 1, \ldots, n$ and $\mid(X) \geq 0$. By induction we show for $i = n, \ldots, 1$

$$0 \leq |\tilde{p}_i(x_{i+1}, \ldots, x_n)| \leq \tilde{p}_i(\overline{X}, \ldots, \overline{X}) \ \text{ for all } \ x_{i+1}, \ldots, x_n \in X. \tag{3.1.15}$$

As $a_n > 0$, (3.1.15) holds for $i = n$. Assume (3.1.15) holds for some $i = \hat{i} > 1$ and let $x_{\hat{i}}, \ldots, x_n \in X$ arbitrary but fixed.

$$\begin{aligned}
|\tilde{p}_{\hat{i}-1}(x_{\hat{i}}, \ldots, x_n)| &= |\tilde{p}_{\hat{i}}(x_{\hat{i}+1}, \ldots, x_n)x_{\hat{i}}^{\Delta_{\hat{i}}} + a_{\hat{i}}| \\
&\leq |\tilde{p}_{\hat{i}}(x_{\hat{i}+1}, \ldots, x_n)||x_{\hat{i}}^{\Delta_{\hat{i}}}| + a_{\hat{i}} \\
&\leq |\tilde{p}_{\hat{i}}(\overline{X}, \ldots, \overline{X})|\overline{X}^{\Delta_{\hat{i}}} + a_{\hat{i}} \\
&= \tilde{p}_{\hat{i}}(\overline{X}, \ldots, \overline{X})\overline{X}^{\Delta_{\hat{i}}} + a_{\hat{i}} \\
&= \tilde{p}_{\hat{i}-1}(\overline{X}, \ldots, \overline{X}).
\end{aligned}$$

Hence

$$\begin{aligned}
\overline{H_{p_1}(X)} &= \max_{x_i \in X} \tilde{p}_1(x_1, \ldots, x_n) \\
&\leq \max_{x_i \in X} |\tilde{p}_1(x_1, \ldots, x_n)| \\
&\leq \tilde{p}_1(\overline{X}, \ldots, \overline{X}) \\
&= p_1(\overline{X}).
\end{aligned}$$

As $H_{p_1}(X) \supseteq p_1(X)$ it follows that $\overline{H_{p_1}(X)} = p_1(\overline{X})$. $\square$

Proof of Theorem 3.1.27. Assume $a_i > 0$ for $i = 1, \ldots, n$, $d_1 > 0$ is even, $\mid(X) \geq 0$ and $\underline{H_{p_1}(X)} \geq 0$. We distinguish the cases $0 \notin X$ and $0 \in X$.

- Assume $0 \notin X$. As $\mid(X) \geq 0$ it holds that $\underline{X} > 0$. From Lemma 3.1.28 it follows that $\overline{O}_f = 0$, hence $X \cap O_f = \emptyset$ and $H_f(X) = f(X)$ by Theorem 3.1.12.
- Assume $0 \in X$. As $\underline{H_{p_1}(X)} \geq 0$, $\overline{H_{p_1}(X)} = p_1(\overline{X})$ by Lemma 3.1.29 and $X^{d_1} = [0, \overline{X}^{d_1}]$, it holds that

$$\begin{aligned}
H_f(X) &= H_{p_1}(X)X^{d_1} \\
&= [\underline{H_{p_1}(X)}, p_1(\overline{X})][0, \overline{X}^{d_1}] \\
&= [0, p_1(\overline{X})\overline{X}^{d_1}] \\
&= [f(0), f(\overline{X})] \\
&\subseteq f(X). \ \square
\end{aligned}$$

Corollary 3.1.30

(i) *If* $a_i < 0$ *for* $i = 1, \ldots, n$, $d_1 > 0$ *is even,* $\mathrm{mid}(X) \geq 0$ *and* $\overline{H_{p_1}(X)} \leq 0$, *then* $H_f(X) = f(X)$.

(ii) *If* $a_i > 0$ *for all* i *where* d_i *is even,* $a_i < 0$ *for all* i *where* d_i *is odd,* $d_1 > 0$ *is even,* $\mathrm{mid}(X) \leq 0$ *and* $\underline{H_{p_1}(X)} \geq 0$, *then* $H_f(X) = f(X)$.

(iii) *If* $a_i < 0$ *for all* i *where* d_i *is even,* $a_i > 0$ *for all* i *where* d_i *is odd,* $d_1 > 0$ *is even,* $\mathrm{mid}(X) \leq 0$ *and* $\overline{H_{p_1}(X)} \leq 0$, *then* $H_f(X) = f(X)$. □

Proof. Follows from Theorem 3.1.27 applied to $-f(x)$, $f(-x)$, respectively $-f(-x)$. □

Corollary 3.1.31 *If*

$$\mathrm{sign}(a_n) = \mathrm{sign}(a_{n-1}) = \ldots = \mathrm{sign}(a_1)$$

and d_i *is even for* $i = 1, \ldots, n$ *then for all* $X \in \mathbb{IR}$

$$H_f(X) = f(X). \ \square$$

Proof. Let $f(x)$ as in Corollary 3.1.31. Without loss of generality assume

$$\mathrm{sign}(a_n) = \mathrm{sign}(a_{n-1}) = \ldots = \mathrm{sign}(a_1) = 1, \ \ d_1 > 0.$$

Let $X \in \mathbb{IR}$ arbitrary but fixed. If $\mathrm{mid}(X) \geq 0$ we apply Theorem 3.1.27, if $\mathrm{mid}(X) \leq 0$ we apply Corollary 3.1.30. In both cases we have to show that

$$\underline{H_{p_1}(X)} \geq 0.$$

As all coefficients of $\tilde{p}_1$ are positive and each x_i has even power in $\tilde{p}_1(x_1, \ldots, x_n)$, it holds that

$$\tilde{p}_1(x_1, \ldots, x_n) \geq 0 \ \ \text{for all} \ \ x_i \in \mathbb{R},$$

hence $\underline{H_{p_1}(X)} \geq 0$. □

Another example where Horner form is exact is as follows:

Example. Consider the polynomial $f(x) = x^7 + 2x^5 + 4x^3 + x$ and let $X = [-1, 1]$. We obtain $O_f = \{0\}$, hence $\mathrm{int}(X) \cap O_f \neq \emptyset$, i.e. the non–overestimation condition of Theorem 3.1.12 is not satisfied. Still,

$$H_f(X) = [-8, 8] = f(X).$$

Figure 3.1.3 shows $\mathrm{q}(H_f(X), f(X))$ for $f(x) = x^7 + 2x^5 + 4x^3 + x$ in dependence of $\mathrm{mid}(X)$, where $\mathrm{w}(X) \equiv 2$. According to Theorem 3.1.12, $H_f(X) = f(X)$ if $\mathrm{mid}(X) \geq 1$ or $\mathrm{mid}(X) \leq -1$. The observation that $H_f(X) = f(X)$ for $\mathrm{mid}(X) = 0$, i.e. $X = [-1, 1]$ is generalized by the following Theorem.

Theorem 3.1.32 *If*

$$\mathrm{sign}(a_n) = \mathrm{sign}(a_{n-1}) = \ldots = \mathrm{sign}(a_1),$$

d_i *is odd for* $i = 1, \ldots, n$ *and* $\mathrm{mid}(X) = 0$, *then*

$$H_f(X) = f(X). \ \square$$

Proof. Let $f(x)$ as in Theorem 3.1.32 and $\mathrm{mid}(X) = 0$. Note that Δ_i is even for $i = 2, \ldots, n$ and $\Delta_1 = d_1$ is odd. Further, $f(\overline{X}) = -f(\underline{X})$. Without loss of generality assume

$$\mathrm{sign}(a_n) = \mathrm{sign}(a_{n-1}) = \ldots = \mathrm{sign}(a_1) = 1.$$

By induction we show for $i = n, \ldots, 0$ that

$$0 \leq |\tilde{p}_i(x_{i+1}, \ldots, x_n)| \leq \tilde{p}_i(\overline{X}, \ldots, \overline{X}) \ \ \text{for all} \ \ x_{i+1}, \ldots, x_n \in X. \tag{3.1.16}$$

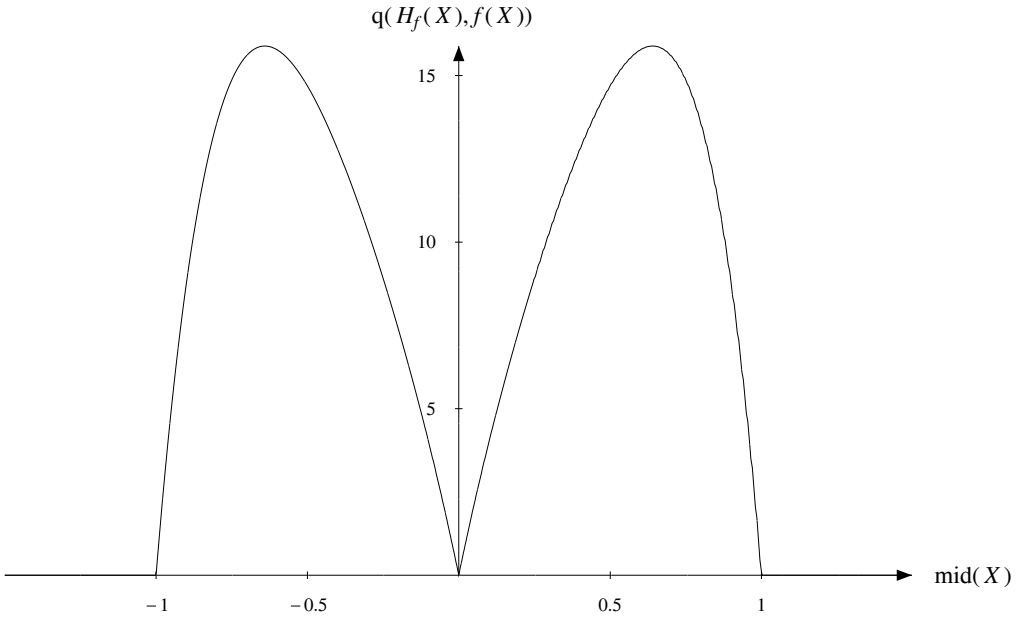


Figure 3.1.3: $q(H_f(X), f(X))$ in dependence of $\mathrm{mid}(X)$, where $f(x) = x^7 + 2x^5 + 4x^3 + x$ and $w(X) \equiv 2$.

As $a_n > 0$, (3.1.16) holds for $i = n$. Assume (3.1.16) holds for some $i = \hat{i} > 0$.

$$\begin{aligned} |\tilde{p}_{\hat{i}-1}(x_{\hat{i}}, \ldots, x_n)| &= |\tilde{p}_{\hat{i}}(x_{\hat{i}+1}, \ldots, x_n) x_{\hat{i}}^{\Delta_{\hat{i}}} + a_{\hat{i}}| \\ &\leq |\tilde{p}_{\hat{i}}(x_{\hat{i}+1}, \ldots, x_n)||x_{\hat{i}}^{\Delta_{\hat{i}}}| + a_{\hat{i}} \\ &\leq \tilde{p}_{\hat{i}}(\overline{X}, \ldots, \overline{X})\overline{X}^{\Delta_{\hat{i}}} + a_{\hat{i}} \\ &= \tilde{p}_{\hat{i}-1}(\overline{X}, \ldots, \overline{X}). \end{aligned}$$

Hence,

$$\begin{aligned} H_f(X) &= \{\tilde{f}(x_1, \ldots, x_n) \mid x_1, \ldots, x_n \in X\} \\ &\subseteq [-f(\overline{X}), f(\overline{X})] \\ &= [f(\underline{X}), f(\overline{X})] \\ &\subseteq f(X). \ \square \end{aligned}$$

Remark. The theorems in this section do not give a complete characterization of all cases where $H_f(X) = f(X)$. For example, let

$$\begin{aligned} f(x) &= x^7 - 2x^4 + 4x, \\ X &= [-1, 1] + (r^2 - 1)/r, \quad \text{where} \quad r = \sqrt[3]{\tfrac{1}{2}(1 + \sqrt{5})}. \end{aligned}$$

Then $H_f(X) = f(X)$, but none of the criteria presented in this section applies. $\square$

3.1.2 Improvements if the Input Interval does not Contain Zero

In this section we consider two improvements of the Horner form if the input interval does not contain zero in its interior. Section 3.1.2.1 contains an algorithm for evaluating the Horner form which is faster than Algorithm 3.1.7 (HF) if $0 \notin \mathrm{int}(X)$. Sometimes it is sufficient to compute only the upper or the lower bound of the Horner form. Obviously this can be done by evaluating the Horner form and selecting the desired endpoint. In Section 3.1.2.2 we show a more efficient method for the case $0 \notin \mathrm{int}(X)$.

3.1.2.1 Efficiency Improvement

If $0 \notin \mathrm{int}(X)$ then some interval power computations of Algorithm 3.1.7 (HF) can be replaced by number power computations. More precisely, if $0 \in P_{i+1}$ and $\underline{X} \geq 0$, then

$$P_{i+1}X^{d_{i+1}-d_i} = P_{i+1}\overline{X^{d_{i+1}-d_i}} = P_{i+1}\overline{X}^{d_{i+1}-d_i}.$$

The case $\overline{X} \leq 0$ is treated analogously. The case distinction $0 \in P_{i+1}$ does not introduce additional overhead because it is necessary anyways during the interval multiplication $P_{i+1}X^{d_{i+1}-d_i}$. Thus, we obtain the following algorithm.

Algorithm 3.1.33 (HFS) [Horner Form, Special Case for $0 \notin \mathrm{int}(X)$]

In: $f(x) = a_n x^{d_n} + a_{n-1}x^{d_{n-1}} + \ldots + a_1 x^{d_1} \in \mathbb{F}[x]$,
$X \in \mathbb{IF}$.

Out: $\mathrm{HF}(f, X)$.

(1) [Reduce to case $\underline{X} \geq 0$.]
if $0 \in \mathrm{int}(X)$ then return $\mathrm{HF}(f, X)$.
if $\overline{X} < 0$ then $f(x) \longleftarrow f(-x), X \longleftarrow -X$.

(2) [Initialize.]
$P_n \longleftarrow a_n$.

(3) [Accumulate.]
for $i = n-1, \ldots, 1$
  if $0 \in P_{i+1}$ then
    $P_i \longleftarrow P_{i+1}\ \hat{\mathrm{POW}}\ (\overline{X}, d_{i+1} - d_i) + a_i$
  else
    $P_i \longleftarrow P_{i+1}X^{d_{i+1}-d_i} + a_i$.

(4) [Last power.]
if $0 \in P_1$ then
  $P_0 \longleftarrow P_1\ \hat{\mathrm{POW}}\ (\overline{X}, d_1)$.
else
  $P_0 \longleftarrow P_1 X^{d_1}$.

(5) [Return.]
return P_0.

3.1.2.2 Separate Computation of Upper and Lower Bound

There are many algorithms, in particular algorithms for optimization, which require the computation of an upper (or lower) bound of $f(X)$. If u is an upper bound of $-f$ in X then $-u$ is a lower bound of f in X, hence it suffices to consider the computation of upper bounds. Obviously an upper bound of f in X can be obtained by evaluating $H_f(X)$ and selecting the upper bound of the result. In this section we present a more efficient method if $0 \notin \mathrm{int}(X)$. The results in this section are original.

From now on assume $0 \notin \mathrm{int}(X)$. It suffices to consider the case $\underline{X} \geq 0$. If $\overline{X} \leq 0$, then we form the polynomial $g(x) = f(-x)$, compute $\overline{H_g(-X)}$, and according to Theorem 3.1.6, $\overline{H_g(-X)} = \overline{H_f(X)}$. In the following let $\tilde{p}_i$ and $\tilde{f}$ as in (3.1.7).

Theorem 3.1.34 *Assume $\underline{X} \geq 0$ and define $\hat{\boldsymbol{x}} \in \mathbb{R}^n$ recursively as*

$$\hat{x}_i = \begin{cases} \overline{X} & \text{if } \tilde{p}_i(\hat{x}_{i+1}, \ldots, \hat{x}_n) \geq 0 \\ \underline{X} & \text{else.} \end{cases}$$

Then

$$\tilde{f}(\hat{\boldsymbol{x}}) = \overline{H_f(X)}. \ \square$$

Proof. Let X and $\hat{\boldsymbol{x}}$ as in Theorem 3.1.34. By induction we show for $i = n, \ldots, 0$

$$\tilde{p}_i(\hat{x}_{i+1}, \ldots, \hat{x}_n) \geq \tilde{p}_i(x_{i+1}, \ldots, x_n) \text{ for all } x_{i+1}, \ldots, x_n \in X. \tag{3.1.17}$$

Obviously (3.1.17) holds for $i = n$. Assume (3.1.17) holds for some $i = \hat{i} > 0$ and let $x_{\hat{i}}, \ldots, x_n \in X$ arbitrary but fixed. Then

$$\begin{aligned}
\tilde{p}_{\hat{i}-1}(x_{\hat{i}}, \ldots, x_n) &= \tilde{p}_{\hat{i}}(x_{\hat{i}+1}, \ldots, x_n) x_{\hat{i}}^{\Delta_{\hat{i}}} + a_{\hat{i}-1} \\
&\leq \tilde{p}_{\hat{i}}(\hat{x}_{\hat{i}+1}, \ldots, \hat{x}_n) x_{\hat{i}}^{\Delta_{\hat{i}}} + a_{\hat{i}-1} \\
&\leq \begin{cases} \tilde{p}_{\hat{i}}(\hat{x}_{\hat{i}+1}, \ldots, \hat{x}_n) \overline{X}^{\Delta_{\hat{i}}} + a_{\hat{i}-1} & \text{if } \tilde{p}_{\hat{i}}(\hat{x}_{\hat{i}+1}, \ldots, \hat{x}_n) \geq 0 \\ \tilde{p}_{\hat{i}}(\hat{x}_{\hat{i}+1}, \ldots, \hat{x}_n) \underline{X}^{\Delta_{\hat{i}}} + a_{\hat{i}-1} & \text{else} \end{cases} \\
&= \tilde{p}_{\hat{i}}(\hat{x}_{\hat{i}+1}, \ldots, \hat{x}_n) \hat{x}_{\hat{i}}^{\Delta_{\hat{i}}} + a_{\hat{i}-1} \\
&= \tilde{p}_{\hat{i}-1}(\hat{x}_{\hat{i}}, \ldots, \hat{x}_n). \ \square
\end{aligned}$$

Based on Theorem 3.1.34 we give algorithms for computing $\overline{H_f(X)}$ and $\underline{H_f(X)}$. In order to prevent wrong results in case of overflow we have to check the invalid operation flag explicitly.

Algorithm 3.1.35 (HFUB) [Upper Bound of Horner Form]

In: $f(x) = a_n x^{d_n} + a_{n-1} x^{d_{n-1}} + \ldots + a_1 x^{d_1} \in \mathbb{F}[x]$,
$X \in \mathbb{IF}$.

Out: $\mathrm{HFUB}(f, X) \in \mathbb{F}$, $\mathrm{HFUB}(f, X) \geq \overline{H_f(X)}$.

(1) [Reduce to case $\underline{X} \geq 0$.]
if $0 \in \mathrm{int}(X)$ return $\overline{\mathrm{HF}(f, X)}$.
if $\overline{X} < 0$ then $f(x) \longleftarrow f(-x)$, $X \longleftarrow -X$.

(2) [Initialize.]
$\tilde{p}_n \longleftarrow a_n$.
clear invalid operation flag.

(3) [Accumulate.]
for $i = n-1, \ldots, 1$
if $\tilde{p}_{i+1} \geq 0$
then $\tilde{p}_i \longleftarrow \tilde{p}_{i+1} \mathbin{\hat{*}} \hat{\mathrm{POW}}(\overline{X}, d_{i+1} - d_i) \mathbin{\hat{+}} a_i$.
else $\tilde{p}_i \longleftarrow \tilde{p}_{i+1} \mathbin{\hat{*}} \check{\mathrm{POW}}(\underline{X}, d_{i+1} - d_i) \mathbin{\hat{+}} a_i$.

(4) [Last power.]
if $\tilde{p}_1 \geq 0$
then $\tilde{p}_0 \longleftarrow \tilde{p}_1 \mathbin{\hat{*}} \hat{\mathrm{POW}}(\overline{X}, d_1)$
else $\tilde{p}_0 \longleftarrow \tilde{p}_1 \mathbin{\hat{*}} \check{\mathrm{POW}}(\underline{X}, d_1)$.

(5) [Check invalid operation and return.]
if invalid operation flag is raised, then $\tilde{p}_0 \longleftarrow \top$.
return $\tilde{p}_0$.

Algorithm 3.1.36 (HFLB) [Lower Bound of Horner Form]

In: $f(x) = a_n x^{d_n} + a_{n-1} x^{d_{n-1}} + \ldots + a_1 x^{d_1} \in \mathbb{F}[x]$,
$X \in \mathbb{IF}$.

Out: $\mathrm{HFLB}(f, X) \in \mathbb{F}$, $\mathrm{HFLB}(f, X) \leq \underline{H_f(X)}$.

(1) [Reduce to upper bound.]
return $-\mathrm{HFUB}(-f, X)$.

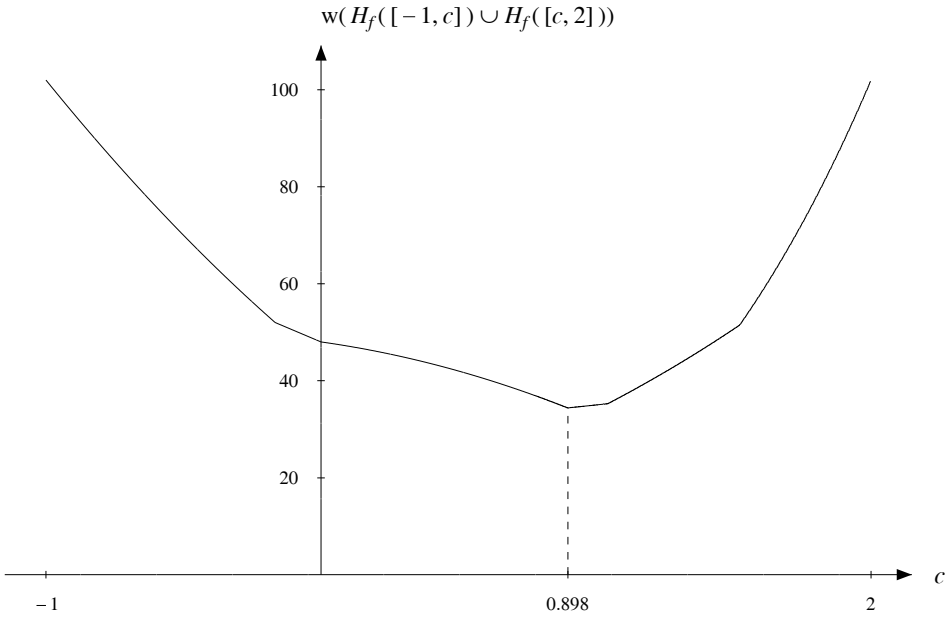


Figure 3.1.4: $\mathrm{w}(H_f([\underline{X}, c]) \cup H_f([c, \overline{X}]))$ in dependence of c for $f(x) = x^5 + x^4 - 3x^3 - 5x^2 + 7x$ and $X = [-1, 2]$. The optimal bisection point is $\hat{c} \approx 0.9$.

Theorem 3.1.37 (Complexity)

(i) *If* $0 \in \mathrm{int}(X)$ *then Algorithm 3.1.35 and 3.1.36 cost*

n	*interval power computations,*
n	*interval multiplications and*
$n-1$	*interval additions.*

(ii) *If* $0 \notin \mathrm{int}(X)$ *then Algorithm 3.1.35 and 3.1.36 cost*

n	*number power computations,*
n	*number multiplications and*
$n-1$	*number additions.* □

3.1.3 Bisection of the Input Interval

The overestimation error of the Horner form can usually be reduced by dividing the input interval X at a point $c \in \mathrm{int}(X)$ into two subintervals $X_1 = [\underline{X}, c]$, $X_2 = [c, \overline{X}]$ and evaluating the Horner form on both intervals separately. From the inclusion monotonicity of the Horner form it follows that

$$H_f(X) \supseteq H_f(X_1) \cup H_f(X_2) \supseteq f(X).$$

It seems to be non–trivial to compute an optimal bisection point $\hat{c}$ for a given polynomial f and $X \in \mathbb{IR}$, for example in the sense that

$$\mathrm{w}\big(H_f([\underline{X}, \hat{c}]) \cup H_f([\hat{c}, \overline{X}])\big) \leq \mathrm{w}\big(H_f([\underline{X}, c]) \cup H_f([c, \overline{X}])\big) \quad \text{for all} \quad c \in X. \tag{3.1.18}$$

Figure 3.1.4 shows $\mathrm{w}\big(H_f([\underline{X}, c]) \cup H_f([c, \overline{X}])\big)$ in dependence of c for $f(x) = x^5 + x^4 - 3x^3 - 5x^2 + 7x$ and $X = [-1, 2]$.

As the computation of an optimal or at least a "good" bisection point seems too expensive just for the purpose of reducing the overestimation error, we consider two special cases.

- *Bisection at zero.* In general, the computation of $H_f(X_1) \cup H_f(X_2)$ is twice as expensive as the computation of $H_f(X)$. However, as shown in Section 3.1.3.1, the costs are comparable if $c = 0$. This means that if $0 \in \mathrm{int}(X)$, then the computation of $H_f([\underline{X}, 0]) \cup H_f([0, \overline{X}])$ is roughly as expensive as the computation of $H_f(X)$, but gives usually better inclusions of the range.
- *Bisection at the midpoint.* The midpoint is usually not an optimal bisection point in the sense of (3.1.18). Yet, it is a "standard choice" in many cases and in Section 3.1.3.2 we give some bounds for the reduction of the overestimation error through bisection at the midpoint.

The results presented in this section are new. They will be used later on in Section 3.3 for the improvement of the Taylor form.

3.1.3.1 Bisection at Zero

In the sequel let $\check{H}_f : \mathbb{IR} \to \mathbb{IR}$ be defined as

$$\check{H}_f(X) = \begin{cases} H_f([\underline{X}, 0]) \cup H_f([0, \overline{X}]) & \text{if } 0 \in \mathrm{int}(X) \\ H_f(X) & \text{else.} \end{cases}$$

We give an algorithm for evaluating $\check{H}_f(X)$, which costs less than twice as many arithmetic floating point operations as Algorithm 3.1.7 (HF). The efficiency improvement is due to the fact, that an endpoint of $[\underline{X}, 0]$ and $[0, \overline{X}]$ is zero, and multiplications by zero can be saved.

Algorithm 3.1.38 (HFBZ) [Horner Form with Bisection at Zero]

In: $f(x) = a_n x^{d_n} + a_{n-1} x^{d_{n-1}} + \ldots + a_1 x^{d_1} \in \mathbb{F}[x]$,
$X \in \mathbb{IF}$.

Out: $\mathrm{HFBZ}(f, X) \in \mathbb{IF}$, $\mathrm{HFBZ}(f, X) \supseteq \check{H}_f(X)$.

(1) [Case $0 \notin \mathrm{int}(X)$.]
If $0 \notin \mathrm{int}(X)$ then return $\mathrm{HFS}(X)$.

(2) [Bisect and evaluate.]
$Y_1 \longleftarrow \mathrm{HFLZ}(f(-x), [0, -\underline{X}])$.
$Y_2 \longleftarrow \mathrm{HFLZ}(f(x), [0, \overline{X}])$.

(3) [Join.]
return $Y_1 \cup Y_2$.

Algorithm 3.1.39 (HFLZ) [Horner Form for $\underline{X} = 0$]

In: $f(x) = a_n x^{d_n} + a_{n-1} x^{d_{n-1}} + \ldots + a_1 x^{d_1} \in \mathbb{F}[x]$,
$X \in \mathbb{IF}$, $\underline{X} = 0$.

Out: $\mathrm{HFLZ}(f, X) \in \mathbb{IF}$, $\mathrm{HFLZ}(f, X) \supseteq H_f(X)$.

(1) [Initialize.]
$P_n \longleftarrow a_n$.
clear invalid operation flag.

(2) [Accumulate.]
for $i = n-1, \ldots, 1$
$p \longleftarrow \hat{\mathrm{POW}}(\overline{X}, d_{i+1} - d_i)$.
if $\underline{P_{i+1}} < 0$ then $\underline{P_i} \longleftarrow \underline{P_{i+1}} \mathbin{\check{*}} p \mathbin{\check{+}} a_i$, else $\underline{P_i} \longleftarrow a_i$.
if $\overline{P_{i+1}} > 0$ then $\overline{P_i} \longleftarrow \overline{P_{i+1}} \mathbin{\hat{*}} p \mathbin{\hat{+}} a_i$, else $\overline{P_i} \longleftarrow a_i$.

Monomials	3	6	9	12	15	18	21
Flops for $H_f(X)$	25.59	40.86	52.65	63.19	73.79	85.87	102.14
Flops for $\check{H}_f(X)$	31.15	47.15	58.50	68.25	77.57	87.00	96.72
$\mathrm{q}(H_f(X), f(X))/\mathrm{w}(X)$	0.05	0.15	0.24	0.37	0.49	0.62	0.75
$\mathrm{q}(\check{H}_f(X), f(X))/\mathrm{w}(X)$	0.03	0.09	0.12	0.16	0.19	0.22	0.25

Table 3.1.1: Comparison of Horner form $H_f(X)$, computed by Algorithm 3.1.7 (HF), and Horner form with bisection at zero $\check{H}_f(X)$, computed by Algorithm 3.1.38 (HFBZ), for random polynomials of degree 20 and $0 \in X \subseteq [-1, 1]$.

(3) [Last power.]
if $d_1 > 0$
$p \longleftarrow \hat{\mathrm{POW}}\ (\overline{X}, d_1)$
if $\underline{P_1} < 0$ then $\underline{P_0} \longleftarrow \underline{P_1} \mathbin{\check{*}} p$, else $\underline{P_0} \longleftarrow 0$
if $\overline{P_1} > 0$ then $\overline{P_0} \longleftarrow \overline{P_1} \mathbin{\hat{*}} p$, else $\overline{P_0} \longleftarrow 0$

(4) [Check invalid operation and return.]
if invalid operation flag is raised, then $P_0 \longleftarrow [\bot, \top]$
return P_0.

Theorem 3.1.40 (Complexity) *Algorithm 3.1.39* (HFLZ) *costs*

n *number power computations,*
$2n$ *number multiplications and*
$2n-2$ *number additions.* □

Theorem 3.1.41 (Complexity)

(i) *If* $0 \notin \mathrm{int}(X)$ *then Algorithm 3.1.38* (HFBZ) *costs*

n *interval power computations,*
$2n$ *number multiplications and*
$2n-2$ *number additions.*

(ii) *If* $0 \in \mathrm{int}(X)$ *then Algorithm 3.1.38* (HFBZ) *costs*

$2n$ *number power computations,*
$4n$ *number multiplications and*
$4n-4$ *number additions.* □

Theorem 3.1.41 does not justify the claim that the costs for computing $H_f(X)$ and $\check{H}_f(X)$ are comparable. The reason is, that the operations were counted for the worst case. For the average case, an experimental comparison seems to be more appropriate.

Table 3.1.1 shows the average cost and overestimation error of Algorithm 3.1.7 (HF) and Algorithm 3.1.38 (HFBZ) for 10^4 random polynomials f of degree 20 with different numbers of monomials and random intervals X. The polynomial coefficients are randomly chosen from $[-1, 1]$, $\underline{X}$ from $[-1, 0]$ and $\overline{X}$ from $[0, 1]$. All random numbers are uniformly distributed. The cost is the total number of arithmetic floating point instructions, including those which were executed during interval operations. The quantity $\mathrm{q}(H_f(X), f(X))/\mathrm{w}(X)$, respectively $\mathrm{q}(\check{H}_f(X), f(X))/\mathrm{w}(X)$ was chosen to measure the overestimation error.

Finally, one should mention that there are cases where $H_f(X) \neq f(X)$, $0 \in \mathrm{int}(X)$, but $\check{H}_f(X) = H_f(X)$, i.e. bisection at zero does not reduce overestimation. This can happen, even if 0 is the midpoint of X, as is shown in Figure 3.1.5. However, in the next section we show that this is not possible if f is dense.

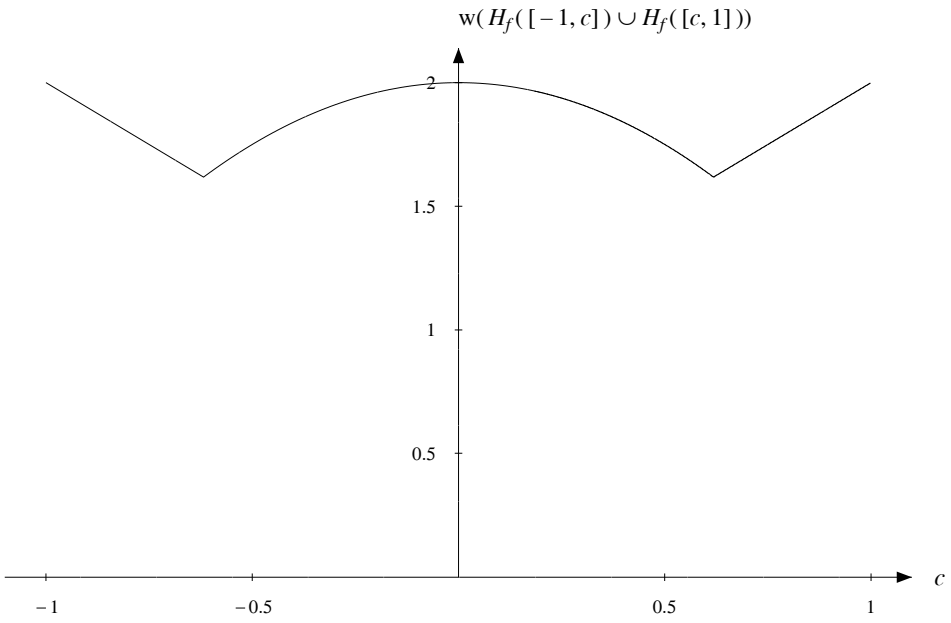


Figure 3.1.5: $\mathrm{w}(H_f([\underline{X}, c]) \cup H_f([c, \overline{X}]))$ in dependence of c for $f(x) = x^3 - x$ and $X = [-1, 1]$. Bisection at $c = 0$ does not reduce the overestimation error.

3.1.3.2 Bisection at the Midpoint for the Dense Horner Form

In this section we study the reduction of the overestimation error of the dense Horner form H_f^*, when the input interval X is bisected at its midpoint. In particular, if $\mathrm{mid}(X) = 0$, then the overestimation error is reduced at least by half. First, we show that if $H_f^*(X) \neq f(X)$, then bisection leads always to a reduction of the overestimation error (Corollary 3.1.44), i.e.

$$H_f^*([\underline{X}, c]) \cup H_f^*([c, \overline{X}]) \subset H_f^*(X) \quad \text{for all} \quad c \in \mathrm{int}(X).$$

Throughout this section let $X \in \mathbb{IR}$,

$$f(x) = \sum_{i=0}^{d_n} a_i^* x^i$$

and let

$$\begin{aligned} \tilde{f}(x_1, \ldots, x_{d_n}) \;=\; & \sum_{i=0}^{d_n} a_i^* \prod_{j=1}^{i} x_j \;= \\ & a_{d_n}^* \quad x_1 x_2 \cdots x_{d_n - 1} x_{d_n} \\ + \; & a_{d_n - 1}^* \quad x_1 x_2 \cdots x_{d_n - 1} \\ & \vdots \qquad\quad \vdots \\ + \; & a_2^* \quad x_1 x_2 \\ + \; & a_1^* \quad x_1 \\ + \; & a_0^*. \end{aligned}$$

Note that

$$H_f^*(X) = \{\tilde{f}(x_1, \ldots, x_{d_n}) \mid x_i \in X \quad \text{for all} \quad i\}.$$

Let

$$\tilde{f}_i(x_1, \ldots, x_n) = \frac{\partial}{\partial x_i} \tilde{f}(x_1, \ldots, x_n), \quad i = 1, \ldots, n$$

be the partial derivatives of $\tilde{f}$. As $\tilde{f}$ is linear in each variable, $\tilde{f}_i$ does not depend on x_i.

The following lemma states that $\tilde{f}$ achieves its maximum and minimum in the $n-$cube $(X,\ldots,X)$ at a corner of $(X,\ldots,X)$.

Lemma 3.1.42 *There exist* $\hat{x}_1,\ldots,\hat{x}_{d_n},\check{x}_1,\ldots,\check{x}_{d_n}\in\{\underline{X},\overline{X}\}$ *such that*

$$\begin{aligned}\max_{x_1,\ldots,x_{d_n}\in X}\tilde{f}(x_1,\ldots,x_{d_n}) &= \tilde{f}(\hat{x}_1,\ldots,\hat{x}_{d_n})\\ \min_{x_1,\ldots,x_{d_n}\in X}\tilde{f}(x_1,\ldots,x_{d_n}) &= \tilde{f}(\check{x}_1,\ldots,\check{x}_{d_n}).\ \square\end{aligned}$$

Proof. Follows immediately from the linearity of $\tilde{f}$ in each variable. $\square$

If $H_f^*(X)\neq f(X)$ then bisection at any point in the interior of X reduces the overestimation error.

Theorem 3.1.43 *For all* $c\in\mathrm{int}(X)$ *it holds that*

(i) *If* $\overline{f(X)}<\overline{H_f^*(X)}$ *then*

$$\overline{H_f^*([\underline{X},c])\cup H_f^*([c,\overline{X}])}<\overline{H_f^*(X)}.$$

(ii) *If* $\underline{f(X)}>\underline{H_f^*(X)}$ *then*

$$\underline{H_f^*([\underline{X},c])\cup H_f^*([c,\overline{X}])}>\underline{H_f^*(X)}.\ \square$$

Proof. As (ii) follows from (i) if f is replaced by $-f$, it suffices to prove (i). Assume

$$\overline{H_f^*([\underline{X},c])\cup H_f^*([c,\overline{X}])}=\overline{H_f^*(X)}$$

and let

$$Y=\begin{cases}[\underline{X},c] & \text{if } \overline{H_f^*([\underline{X},c])}\geq\overline{H_f^*([c,\overline{X}])}\\ [c,\overline{X}] & \text{else.}\end{cases}$$

Then

$$\overline{H_f^*(Y)}=\overline{H_f^*(X)}.$$

Note that

$$\begin{aligned}\overline{H_f^*(Y)} &= \max_{x_1,\ldots,x_{d_n}\in Y}\tilde{f}(x_1,\ldots,x_{d_n})\\ \overline{H_f^*(X)} &= \max_{x_1,\ldots,x_{d_n}\in X}\tilde{f}(x_1,\ldots,x_{d_n}).\end{aligned}$$

According to Lemma 3.1.42 there exist $y_1,\ldots,y_{d_n}\in\{\underline{Y},\overline{Y}\}$ such that

$$\begin{aligned}\tilde{f}(y_1,\ldots,y_{d_n}) &= \max_{x_1,\ldots,x_{d_n}\in Y}\tilde{f}(x_1,\ldots,x_{d_n})\\ &= \max_{x_1,\ldots,x_{d_n}\in X}\tilde{f}(x_1,\ldots,x_{d_n}).\end{aligned}$$

From the linearity of $\tilde{f}$ it follows that if $y_i=c$ for some i, then

$$\tilde{f}(y_1,\ldots,y_{i-1},\underline{X},y_{i+1},\ldots,y_n)=\tilde{f}(y_1,\ldots,y_n)=\tilde{f}(y_1,\ldots,y_{i-1},\overline{X},y_{i+1},\ldots,y_n).\tag{3.1.19}$$

Repeated application of (3.1.19) yields

$$\tilde{f}(y_1,\ldots,y_{d_n})=\begin{cases}\tilde{f}(\underline{X},\ldots,\underline{X}) & \text{if } Y=[\underline{X},c]\\ \tilde{f}(\overline{X},\ldots,\overline{X}) & \text{if } Y=[c,\overline{X}]\end{cases}\leq\overline{f(X)}.$$

Hence,

$$\overline{f(X)}=\tilde{f}(y_1,\ldots,y_{d_n})=H_f^*(X).\ \square$$

Corollary 3.1.44 *If* $f(X) \subset H_f^*(X)$ *then*

$$H_f^*([\underline{X}, c]) \cup H_f^*([c, \overline{X}]) \subset H_f^*(X)$$

for all $c \in \mathrm{int}(X)$. □

In the following we consider the special case $\mathrm{mid}(X) = 0$, i.e. the input interval is centered. Here, bisection at the midpoint reduces the overestimation error of the dense Horner form at least by half. This result is used later on for the reduction of the overestimation error of the Taylor form, see Section 3.3.

Theorem 3.1.45 (Bisection if Midpoint is Zero) *Assume* $\mathrm{mid}(X) = 0$ *and let*

$$C = H_f^*([\underline{X}, 0]) \cup H_f^*([0, \overline{X}]).$$

Then

$$|\overline{f(X)} - \overline{C}| \leq 1/2|\overline{f(X)} - \overline{H_f^*(X)}| \tag{3.1.20}$$
$$|\underline{f(X)} - \underline{C}| \leq 1/2|\underline{f(X)} - \underline{H_f^*(X)}|. \ \square \tag{3.1.21}$$

Proof. Let X and C as in Theorem 3.1.45. As (3.1.21) follows from (3.1.20) if f is replaced by $-f$, it suffices to show (3.1.20). From

$$\overline{H_f^*(X)} \geq \overline{C} \geq \overline{f(X)}$$

it follows that (3.1.20) is equivalent to

$$\overline{H_f^*(X)} - \overline{C} \geq \overline{C} - \overline{f(X)}.$$

Further,

$$\overline{C} = \max\{\overline{H_f^*([0, \overline{X}])}, \overline{H_f^*([\underline{X}, 0])}\},$$

hence it suffices to show

$$\overline{H_f^*(X)} - \overline{H_f^*([0, \overline{X}]} \geq \overline{H_f^*([0, \overline{X}])} - \overline{f(X)} \tag{3.1.22}$$
$$\overline{H_f^*(X)} - \overline{H_f^*([\underline{X}, 0]} \geq \overline{H_f^*([\underline{X}, 0])} - \overline{f(X)}. \tag{3.1.23}$$

The proofs of (3.1.22) and (3.1.23) are analogous, but for the sake of completeness we give both of them explicitly.

- Proof of (3.1.22). According to Lemma 3.1.42 there exist $\hat{x}_1, \ldots, \hat{x}_{d_n} \in \{0, \overline{X}\}$ such that

$$\max_{x_1, \ldots, x_{d_n} \in [0, \overline{X}]} \tilde{f}(x_1, \ldots, x_{d_n}) = \tilde{f}(\hat{x}_1, \ldots, \hat{x}_{d_n}).$$

Let k be the largest index such that $\hat{x}_i = \overline{X}$ for all $i \leq k$. If no such k exists, let $k = 0$. Then

$$\begin{aligned} \overline{H_f^*([0, \overline{X}])} &= \tilde{f}(\overline{X}, \ldots, \overline{X}, 0, \hat{x}_{k+2}, \ldots, \hat{x}_n) \\ &= a_k \overline{X}^k + a_{k-1} \overline{X}^{k-1} + \ldots + a_1 \overline{X} + a_0 \\ &\leq |a_k \overline{X}^k| + |a_{k-1} \overline{X}^{k-1}| + \ldots + |a_1 \overline{X}| + a_0. \end{aligned}$$

As $\mathrm{mid}(X) = 0$, it holds that

$$\overline{H_f^*(X)} = |a_{d_n} \overline{X}^{d_n}| + |a_{d_n - 1} \overline{X}^{d_n - 1}| + \ldots + |a_1 \overline{X}| + a_0.$$

Thus,

$$\begin{aligned} \overline{H_f^*(X)} - \overline{H_f^*([0, \overline{X}])} &\geq |a_{d_n}| \overline{X}^{d_n} + |a_{d_n - 1}| \overline{X}^{d_n - 1} + \ldots + |a_{k+1}| \overline{X}^{k+1} \\ &\geq -(a_{d_n} \overline{X}^{d_n} + a_{d_n - 1} \overline{X}^{d_n - 1} + \ldots + a_{k+1} \overline{X}^{k+1}) \\ &= \overline{H_f^*([0, \overline{X}])} - f(\overline{X}) \\ &\geq \overline{H_f^*([0, \overline{X}])} - \overline{f(X)}. \end{aligned}$$

- Proof of (3.1.23). According to Lemma 3.1.42 there exist $\hat{x}_1, \ldots, \hat{x}_{d_n} \in \{\underline{X}, 0\}$ such that

$$\max_{x_1,\ldots,x_{d_n} \in [\underline{X},0]} \tilde{f}(x_1, \ldots, x_{d_n}) = \tilde{f}(\hat{x}_1, \ldots, \hat{x}_{d_n}).$$

  Let k be the largest index such that $\hat{x}_i = \underline{X}$ for all $i \leq k$. If no such k exists, let $k = 0$. Then

$$\begin{aligned} \overline{H_f^*([\underline{X}, 0])} &= \tilde{f}(\underline{X}, \ldots, \underline{X}, 0, \hat{x}_{k+2}, \ldots, \hat{x}_n) \\ &= a_k \underline{X}^k + a_{k-1}\underline{X}^{k-1} + \ldots + a_1 \underline{X} + a_0 \\ &\leq |a_k \underline{X}^k| + |a_{k-1}\underline{X}^{k-1}| + \ldots + |a_1 \underline{X}| + a_0. \end{aligned}$$

  As $\mathrm{mid}(X) = 0$, it holds that

$$\overline{H_f^*(X)} = |a_{d_n}\underline{X}^{d_n}| + |a_{d_n-1}\underline{X}^{d_n-1}| + \ldots + |a_1\underline{X}| + a_0.$$

  Thus,

$$\begin{aligned} \overline{H_f^*(X)} - \overline{H_f^*([\underline{X}, 0])} &\geq |a_{d_n}\underline{X}^{d_n}| + |a_{d_n-1}\underline{X}^{d_n-1}| + \ldots + |a_{k+1}\underline{X}^{k+1}| \\ &\geq -(a_{d_n}\underline{X}^{d_n} + a_{d_n-1}\underline{X}^{d_n-1} + \ldots + a_{k+1}\underline{X}^{k+1}) \\ &= \overline{H_f^*([\underline{X}, 0])} - f(\underline{X}) \\ &\geq \overline{H_f^*([\underline{X}, 0])} - \overline{f(X)}. \ \square \end{aligned}$$

Another special case, where bisection at the midpoint reduces the overestimation error of the dense Horner form at least by half, are parabolas.

Theorem 3.1.46 (Bisection for Parabolas) *Assume $d_n = 2$, let $c = \mathrm{mid}(X)$ and*

$$C = H_f^*([\underline{X}, c]) \cup H_f^*([c, \overline{X}]).$$

Then

$$|\overline{f(X)} - \overline{C}| \leq 1/2|\overline{f(X)} - \overline{H_f^*(X)}| \tag{3.1.24}$$

$$|\underline{f(X)} - \underline{C}| \leq 1/2|\underline{f(X)} - \underline{H_f^*(X)}|. \ \square \tag{3.1.25}$$

Proof. Let f, c and C as in Theorem 3.1.46. As (3.1.25) follows from (3.1.24) if f is replaced by $-f$, it suffices to show (3.1.24). From

$$\overline{H_f^*(X)} \geq \overline{C} \geq \overline{f(X)}$$

it follows that (3.1.24) is equivalent to

$$\overline{C} - \overline{f(X)} \leq \overline{H_f^*(X)} - \overline{C}.$$

Let

$$\mathcal{Y} = ([\underline{X}, c], [\underline{X}, c]) \cup ([c, \overline{X}], [c, \overline{X}]).$$

Note that $\overline{C} \geq \tilde{f}(y_1, y_2)$ for all $(y_1, y_2) \in \mathcal{Y}$, $\overline{f(X)} \geq \tilde{f}(x, x)$ for all $x \in X$, and $\overline{H_f^*(X)} \geq \tilde{f}(x_1, x_2)$ for all $x_1, x_2 \in X$. Hence, it suffices to show that for all $(y_1, y_2) \in \mathcal{Y}$ there exist $x, x_1, x_2 \in X$ such that

$$\tilde{f}(y_1, y_2) - \tilde{f}(x, x) \leq \tilde{f}(x_1, x_2) - \tilde{f}(y_1, y_2).$$

In fact, we show that for all $(y_1, y_2) \in \mathcal{Y}$ there exist $x, x_1, x_2 \in X$ such that

$$\tilde{f}(y_1, y_2) - \tilde{f}(x, x) = \tilde{f}(x_1, x_2) - \tilde{f}(y_1, y_2).$$

Let $(y_1, y_2) \in \mathcal{Y}$ arbitrary but fixed.

- If $y_1 \geq y_2 \geq c$ or $y_1 \leq y_2 \leq c$ then let $x = y_1$, $x_1 = y_1$ and $x_2 = 2y_2 - y_1$. Note that $x_2 \in X$. As $\tilde{f}$ is linear in its second argument, it holds that

$$\begin{aligned} \tilde{f}(y_1, y_2) - \tilde{f}(x, x) &= \tilde{f}(y_1, y_2) - \tilde{f}(y_1, y_1) \\ &= \tilde{f}(y_1, 2y_2 - y_1) - \tilde{f}(y_1, y_2) \\ &= \tilde{f}(x_1, x_2) - \tilde{f}(y_1, y_2). \end{aligned}$$

- If $y_2 \geq y_1 \geq c$ or $y_2 \leq y_1 \leq c$ then let $x = y_2$, $x_1 = 2y_1 - y_2$ and $x_2 = y_2$. Note that $x_1 \in X$. As $\tilde{f}$ is linear in its first argument, it holds that

$$\begin{aligned} \tilde{f}(y_1, y_2) - \tilde{f}(x, x) &= \tilde{f}(y_1, y_2) - \tilde{f}(y_2, y_2) \\ &= \tilde{f}(2y_1 - y_2, y_2) - \tilde{f}(y_1, y_2) \\ &= \tilde{f}(x_1, x_2) - \tilde{f}(y_1, y_2). \ \square \end{aligned}$$

Remark. In general it is not the case that bisection of the input interval at its midpoint reduces the overestimation error of the dense Horner form at least by half. For example, let $f(x) = -3x^3 - x^2 + 9x$, $X = [0, 1]$ and $c = \mathrm{mid}(X)$. Then

$$\begin{aligned} \overline{f(X)} - \overline{H_f^*(X)} &\approx 3.9 \\ \overline{f(X)} - \overline{H_f^*([\underline{X}, c]) \cup H_f^*([c, \overline{X}])} &\approx 2.6, \end{aligned}$$

i.e. the error of the upper bound was reduced by less than half.

According to extensive experimental results, the reduction of the overestimation error of the dense Horner form, which is obtained through a bisection of the input interval at its midpoint, can be estimated optimally as follows:

Conjecture 3.1.47 (Bisection at Midpoint in General) *Let $X \in \mathbb{IR}$ such that $\underline{X} \neq 0$, $\overline{X} \neq 0$, $\underline{X} \neq \overline{X}$, let $c = \mathrm{mid}(X)$ and let*

$$\begin{aligned} r &= \begin{cases} \overline{X}/\underline{X} & \text{if } c \geq 0 \\ \underline{X}/\overline{X} & \text{else} \end{cases} \\ t &= \frac{r+1}{2r}. \end{aligned}$$

Further

$$\begin{aligned} b_{\max} &= \begin{cases} 2/t^{d_n - 2} & \text{if } 0 \notin X \\ (r-1)/(rt^{d_n - 1} - 1) & \text{else} \end{cases} \\ b_{\min} &= \begin{cases} 1 + t & \text{if } d_n = 2 \\ 1 + 1/r^{d_n - 2} & \text{if } d_n > 2 \ \text{ and } \ 0 \notin X \\ 1 & \text{else} \end{cases} \end{aligned}$$

and

$$C = H_f^*([\underline{X}, c]) \cup H_f^*([c, \overline{X}]).$$

Then

$$\begin{aligned} b_{\min}(\overline{H_f^*(X)} - \overline{C}) \ \leq \ \overline{H_f^*(X)} - \overline{f(X)} \ &\leq \ b_{\max}(\overline{H_f^*(X)} - \overline{C}) \\ b_{\min}(\underline{C} - \underline{H_f^*(X)}) \ \leq \ \underline{f(X)} - \underline{H_f^*(X)} \ &\leq \ b_{\max}(\underline{C} - \underline{H_f^*(X)}). \ \square \end{aligned}$$

3.1.4 Horner Form for Interval Polynomials

Several interval extensions of polynomials which are considered in subsequent sections require the overestimation of the range of a polynomial with interval coefficients. Therefore, we extend the Horner form in a straight forward way to interval polynomials. In the following let $F : \mathbb{R} \to \mathbb{IR}$,

$$F(x) = A_n x^{d_n} + A_{n-1} x^{d_{n-1}} + \ldots + A_1 x^{d_1} \in \mathbb{IR}[x]$$

be an interval polynomial. F can be considered as a set of real polynomials, i.e.

$$F(x) = \left\{ a_n x^{d_n} + a_{n-1} x^{d_{n-1}} + \ldots + a_1 x^{d_1} \;\middle|\; a_n \in A_n, a_{n-1} \in A_{n-1}, \ldots, a_1 \in A_1 \right\}.$$

Definition 3.1.48 *The Horner form $H_F : \mathbb{IR} \to \mathbb{IR}$ of F is defined as*

$$H_F(X) = \Big(\big(\ldots (A_n X^{\Delta_n} + A_{n-1}) X^{\Delta_{n-1}} + \ldots + A_2 \big) X^{\Delta_2} + A_1 \Big) X^{\Delta_1}. \ \square$$

The following theorem shows that splitting F into F_1, F_2 by splitting a coefficient and evaluating the Horner form of F_1 and F_2 separately does not lead to a better inclusion than evaluating the Horner form of F directly. More precisely, if $F = F_1 \cup F_2$ then

$$H_F(X) = H_{F_1}(X) \cup H_{F_2}(X).$$

Theorem 3.1.49

$$H_F(X) = \{H_f(X) \mid f \in F\}. \ \square$$

Proof.

$$\begin{aligned} H_F(X) &= \Big(\big(\ldots(A_n X^{\Delta_n} + A_{n-1})X^{\Delta_{n-1}} + \ldots + A_2\big)X^{\Delta_2} + A_1\Big)X^{\Delta_1} \\ &= \Big\{\Big(\big(\ldots(a_n X^{\Delta_n} + a_{n-1})X^{\Delta_{n-1}} + \ldots + a_2\big)X^{\Delta_2} + a_1\Big)X^{\Delta_1} \mid a_i \in A_i, i = 1, \ldots, n\Big\} \\ &= \{H_f(X) \mid f \in F\}. \ \square \end{aligned}$$

Algorithm 3.1.50 (HFI) for evaluating $H_F(X)$ differs from Algorithm 3.1.7 (HF) only in that the the coefficients of the input polynomial are floating point intervals instead of floating point numbers.

Algorithm 3.1.50 (HFI) [Horner Form of Interval Polynomial]

In: $F(x) = A_n x^{d_n} + A_{n-1}x^{d_{n-1}} + \ldots + A_1 x^{d_1} \in \mathbb{IF}[x]$,
$X \in \mathbb{IF}$.

Out: $\mathrm{HFI}(F, X) \in \mathbb{IF}, \mathrm{HFI}(F, X) \supseteq H_F(X)$.

(1) [Initialize.]
$P_n \longleftarrow A_n$.

(2) [Accumulate.]
for $i = n-1, \ldots, 1$
$P_i \longleftarrow P_{i+1} X^{d_{i+1}-d_i} + A_i$.

(3) [Last power.]
$P_0 \longleftarrow P_1 X^{d_1}$.

(4) [Return.]
return P_0.

Theorem 3.1.51 (Complexity) *Algorithm 3.1.50 (*HFI*) costs*

n	*interval power computations,*
n	*interval multiplications and*
$2n-2$	*number additions.* $\square$

Theorem 3.1.52 (Complexity) *If* $0 \notin \mathrm{int}(X)$ *then Algorithm 3.1.50 (*HFI*) costs*

n	*interval power computations,*
$2n$	*number multiplications and*
$2n-2$	*number additions.* $\square$

3.2 Mean Value Form

The mean value form is a quadratically convergent interval extension and gives therefore tighter inclusions than the Horner form if the width of the argument interval is sufficiently small. The convergence of the

mean value form is discussed e.g. in [Alefeld and Herzberger, 1974], [Skelboe, 1974], [Caprani and Madsen, 1980], and [Alefeld and Herzberger, 1983]. The inclusion monotonicity of the mean value form is proved in [Caprani and Madsen, 1980].

The mean value form can be derived easily from the mean value Theorem: For all $x, c \in X$ there exists $\xi \in X$ such that

$$f(x) = f(c) + f'(\xi)(x - c).$$

Hence, for all $x, c \in X$

$$f(x) \in f(c) + f'(X)(x - c)$$

and for all $c \in X$

$$f(X) \subseteq f(c) + f'(X)(X - c).$$

Thus, every interval extension F' of f' and any choice of $c \in X$, give rise to an interval extension $F : \mathbb{IR} \to \mathbb{IR}$ of f:

$$F(X) = f(c) + F'(X)(X - c) \supseteq f(X). \tag{3.2.1}$$

The mean value form is a special case of (3.2.1), where F' is the Horner form of f' and $c = \mathrm{mid}(X)$. The choice $c = \mathrm{mid}(X)$ is justified later on by Corollary 3.2.23.

Definition 3.2.1 (Mean Value Form) *The mean value form $M_f : \mathbb{IR} \to \mathbb{IR}$ and the dense mean value form $M_f^* : \mathbb{IR} \to \mathbb{IR}$ are defined as*

$$\begin{aligned} M_f(X) &= f(\mathrm{mid}(X)) + H_{f'}(X)(X - \mathrm{mid}(X)) \\ M_f^*(X) &= f(\mathrm{mid}(X)) + H_{f'}^*(X)(X - \mathrm{mid}(X)). \ \square \end{aligned}$$

Theorem 3.2.2 *M_f and M_f^* are interval extensions of f.* $\square$

Proof. Follows from (3.2.1). $\square$

The following theorem is taken from [Caprani and Madsen, 1980].

Theorem 3.2.3 (Inclusion Monotonicity) *M_f and M_f^* are inclusion monotone.* $\square$

Proof. Let $X_\diamond \subseteq X$, $c_\diamond = \mathrm{mid}(X_\diamond)$, $c = \mathrm{mid}(X)$, $r_\diamond = \mathrm{rad}(X_\diamond)$, $r = \mathrm{rad}(X)$. We show inclusion monotonicity of M_f, i.e. $M_f(X_\diamond) \subseteq M_f(X)$. The proof for M_f^* is analogous. As $X - c$ is a centered interval, it holds that

$$\begin{aligned} M_f(X) &= f(c) + \mathrm{mag}(H_{f'}(X))r[-1, 1] \\ M_f(X_\diamond) &= f(c_\diamond) + \mathrm{mag}(H_{f'}(X_\diamond))r_\diamond[-1, 1]. \end{aligned}$$

Hence, we have to show that

$$\begin{aligned} f(c_\diamond) + \mathrm{mag}(H_{f'}(X_\diamond))r_\diamond &\leq f(c) + \mathrm{mag}(H_{f'}(X))r \\ f(c_\diamond) - \mathrm{mag}(H_{f'}(X_\diamond))r_\diamond &\geq f(c) - \mathrm{mag}(H_{f'}(X))r. \end{aligned}$$

Note that

$$\begin{aligned} |c_\diamond - c| &= 1/2|(\overline{X_\diamond} + \underline{X_\diamond} - \overline{X} - \underline{X})| \\ &= 1/2 \max\{\overline{X_\diamond} + \underline{X_\diamond} - \overline{X} - \underline{X}, \overline{X} + \underline{X} - \overline{X_\diamond} - \underline{X_\diamond}\} \\ &\leq 1/2(\overline{X} + \underline{X_\diamond} - \overline{X_\diamond} - \underline{X}) \\ &= r - r_\diamond. \end{aligned}$$

Further, $H_{f'}(X_\diamond) \subseteq H_{f'}(X)$ by Theorem 3.1.4. Let $\xi \in X$ such that

$$f(c_\diamond) = f(c) + f'(\xi)(c_\diamond - c).$$

Then

$$\begin{aligned}
f(c_\diamond) + \mathrm{mag}(H_{f'}(X_\diamond))r_\diamond &= f(c) + f'(\xi)(c_\diamond - c) + \mathrm{mag}(H_{f'}(X_\diamond))r_\diamond \\
&\leq f(c) + |f'(\xi)||(c_\diamond - c)| + \mathrm{mag}(H_{f'}(X_\diamond))r_\diamond \\
&\leq f(c) + \mathrm{mag}(H_{f'}(X))(r - r_\diamond) + \mathrm{mag}(H_{f'}(X))r_\diamond \\
&= f(c) + \mathrm{mag}(H_{f'}(X))r \\
f(c_\diamond) - \mathrm{mag}(H_{f'}(X_\diamond))r_\diamond &= f(c) + f'(\xi)(c_\diamond - c) - \mathrm{mag}(H_{f'}(X_\diamond))r_\diamond \\
&\geq f(c) - |f'(\xi)||(c_\diamond - c)| - \mathrm{mag}(H_{f'}(X_\diamond))r_\diamond \\
&\geq f(c) - \mathrm{mag}(H_{f'}(X))(r - r_\diamond) - \mathrm{mag}(H_{f'}(X))r_\diamond \\
&= f(c) - \mathrm{mag}(H_{f'}(X))r. \ \square
\end{aligned}$$

Theorem 3.2.4 (Convergence) *M_f and M_f^* converge quadratically to f.* $\square$

Proof. Let $g(x,c) : \mathbb{R}^2 \to \mathbb{R}$ be the uniquely defined polynomial, such that

$$f(x) = f(c) + g(x,c)(x-c) \quad \text{for all} \quad x, c \in \mathbb{R}.$$

From the mean value Theorem it follows that

$$g(x, \mathrm{mid}(X)) \in f'(X) \subseteq H_{f'}^*(X) \quad \text{for all} \quad x \in X.$$

According to Theorem 1.3.19, $H_{f'}^*$ is Lipschitz. Hence M_f^* is a centered form and quadratically convergent by Definition 1.3.25 and Theorem 1.3.27. As $M_f(X) \subseteq M_f^*(X)$ for all $X \in \mathbb{IR}$, the quadratic convergence of M_f follows. $\square$

Algorithm 3.2.5 (MF) for evaluating the mean value form follows immediately from Definition 3.2.1. The algorithm requires evaluation of the Horner form of f'. As the coefficients of f' need not be floating point numbers we have to enclose them by intervals. Further, as the midpoint of X is computed using Algorithm 2.3.33 (MID), which gives only an approximation of $\mathrm{mid}(X)$, the output interval need not be a superset of $M_f^{*(\mathrm{s})}(X)$, but at least contains $f(X)$.

Algorithm 3.2.5 (MF) [Mean Value Form]

In: $f(x) = a_n x^{d_n} + a_{n-1} x^{d_{n-1}} + \ldots + a_1 x^{d_1} \in \mathbb{F}[x]$,
$X \in \mathbb{IF}$.

Out: $\mathrm{MF}(f,X) \in \mathbb{IF}$, $\mathrm{MF}(f,X) \approx M_f(X)$, $\mathrm{MF}(f,X) \supseteq f(X)$.

(1) [Midpoint.]
$c \longleftarrow \mathrm{MID}(X)$.

(2) [Evaluate $f(c)$.]
$Y \longleftarrow \mathrm{HF}(f(x), [c])$.

(3) [Inclusion of f'.]
if $d_1 \neq 0$
for $i = 1, \ldots, n$ do $A_i' \longleftarrow [d_i \mathbin{\check{*}} a_i, d_i \mathbin{\hat{*}} a_i]$.
$F'(x) \longleftarrow A_n' x^{d_n - 1} + A_{n-1}' x^{d_{n-1} - 1} + \ldots + A_1' x^{d_1 - 1}$.
else
for $i = 2, \ldots, n$ do $A_i' \longleftarrow [d_i \mathbin{\check{*}} a_i, d_i \mathbin{\hat{*}} a_i]$.
$F'(x) \longleftarrow A_n' x^{d_n - 1} + A_{n-1}' x^{d_{n-1} - 1} + \ldots + A_2' x^{d_2 - 1}$.

(4) [Evaluate $H_{f'}(X)$.]
$F' \longleftarrow \mathrm{HFI}(F'(x), X)$.

(5) [Return.]
return $Y + F' * (X - c)$.

Note that no invalid operation can occur in Step 3, because $d_i \neq 0$.

Remark. As $F' * (X - c) = \mathrm{mag}(F') * (X - c)$, it would be sufficient to compute "only" $\mathrm{mag}(F')$ instead of F'. However, it seems that there is no algorithm, which computes $\mathrm{mag}(H_{F'}(X))$ faster than $H_{F'}(X)$ for arbitrary X. (An exception is the case $\mathrm{mid}(X) = 0$, see Section 3.3.) □

Theorem 3.2.6 (Complexity) *Algorithm 3.2.5 (*MF*) costs*

$$\begin{array}{rl} 2n & \textit{interval power computations,} \\ n+1 & \textit{interval multiplications,} \\ 4n+1 & \textit{number multiplications, and} \\ 4n+2 & \textit{number additions.} \ \square \end{array}$$

Proof.

- Step 1 costs 1 number multiplication and 2 number additions.
- Step 2 costs n interval power computations, $2n$ number multiplications and $2n-2$ number additions (Theorem 3.1.9).
- Step 3 costs $2n$ number multiplications.
- Step 4 costs n interval power computations, n interval multiplications and $2n-2$ number additions (Theorem 3.1.51).
- Step 5 costs 1 interval multiplication and 4 number additions. □

3.2.1 Slope Form

The slope form can be viewed as an improvement of the mean value form, see for example [Hansen, 1978b], [Alefeld, 1981], [Krawczyk and Neumaier, 1985] [Neumaier, 1990], [Hansen, 1992]. It has the same "shape" as the mean value form, but the overestimation of $f'(X)$ in the mean value form is replaced by an overestimation of the set of slopes between x and the midpoint of X, where x ranges over X.

It is well known that for every $c \in \mathbb{R}$, there exists a polynomial $g_c \in \mathbb{R}[x]$ such that

$$f(x) = f(c) + g_c(x)(x - c). \tag{3.2.2}$$

for all $x \in \mathbb{R}$. Thus, a whole class of interval extensions of f is given by

$$f(X) \subseteq f(c) + G_c(X)(X - c), \tag{3.2.3}$$

where G_c is an interval extension of g_c. The slope form is a special case of (3.2.3), where G_c is the Horner form of g_c and c is the midpoint of X.

Definition 3.2.7 (Slope Form) *Let $c = \mathrm{mid}(X)$ and let g_c be the uniquely defined polynomial such that*

$$f(x) = f(c) + g_c(x)(x - c).$$

The slope form $M_f^{(\mathrm{s})} : \mathbb{IR} \to \mathbb{IR}$ is defined as

$$M_f^{(\mathrm{s})}(X) = f(c) + H_{g_c}(X)(X - c).$$

The dense slope form $M_f^{*(\mathrm{s})} : \mathbb{IR} \to \mathbb{IR}$ is defined as

$$M_f^{*(\mathrm{s})}(X) = f(c) + H_{g_c}^*(X)(X - c). \ \square$$

Obviously $M_f^{(s)}(X) \subseteq M_f^{*(s)}(X)$ for all $X \in \mathbb{IR}$. In the sequel let c and g_c as in Definition 3.2.7.

For the sake of clarity we give an explicit formula for g_c and $H^*_{g_c}$. Let a_i^* be the coefficient of x^i in f, let

$$\begin{aligned} b_n(c) &= a_n^* \\ b_i(c) &= b_{i+1}(c)c + a_i^*, \quad i = n-1, \ldots, 1 \end{aligned}$$

and let

$$g_c(x) = \sum_{i=1}^{n} b_i(c) x^{i-1}.$$

Then

$$\begin{aligned} f(c) + g_c(x)(x-c) &= f(c) + \sum_{i=1}^{n} b_i(c)x^{i-1}(x-c) \\ &= f(c) + \sum_{i=1}^{n} b_i(c)x^i - \sum_{i=1}^{n} b_i(c)x^{i-1}c \\ &= f(c) + \sum_{i=1}^{n} b_i(c)x^i - \sum_{i=0}^{n-1} b_{i+1}(c)x^i c \\ &= f(c) + \sum_{i=1}^{n} b_i(c)x^i - \sum_{i=1}^{n-1}(b_i(c) - a_i)x^i - b_1(c)c \\ &= a_0^* + a_n^* x^n + \sum_{i=1}^{n-1} a_i x^i \\ &= f(x). \end{aligned}$$

Further,

$$H^*_{g_c}(X) = \Big(\big(b_0(\mathrm{mid}(X))X + b_1(\mathrm{mid}(X))\big)X + \ldots + b_{n-2}(\mathrm{mid}(X))\Big)X + b_{n-1}(\mathrm{mid}(X)). \tag{3.2.4}$$

In contrast to the mean value form, neither the slope form nor the dense slope form are inclusion monotone.

Theorem 3.2.8 (Inclusion Monotonicity) *There exist polynomials f such that neither $M_f^{(s)}$ nor $M_f^{*(s)}$ are inclusion monotone.* □

Proof. Consider the example

$$f(x) = 2x^3 - x^2 - 2x, \quad X = [0, 0.9], \quad X_\diamond = [0, 0.5].$$

We obtain

$$\begin{aligned} M_f^{(s)}(f, X) &= M_f^{*(s)}(f, X) &&= [-1.881, 0.0405] \\ M_f^{(s)}(f, X_\diamond) &= M_f^{*(s)}(f, X_\diamond) &&= [-1.125, 0.0625], \end{aligned}$$

i.e.

$$\begin{aligned} \overline{M_f^{(s)}(f, X)} &< \overline{M_f^{(s)}(f, X_\diamond)} \\ \overline{M_f^{*(s)}(f, X)} &< \overline{M_f^{*(s)}(f, X_\diamond)}, \end{aligned}$$

although $X \supset X_\diamond$. □

Theorem 3.2.9 (Convergence) *$M_f^{(s)}$ and $M_f^{*(s)}$ converge quadratically to f.* □

Proof. From (3.2.4) and Corollary 1.3.19 it follows that $H^*_{g_c}$ is Lipschitz. Further, $g_c(x) \in H^*_{g_c}(X)$ for all $x \in X$. Hence $M_f^{*(s)}$ is a centered form of f and is quadratically convergent by Definition 1.3.25 and Theorem 1.3.27. As $M_f^{(s)}(X) \subseteq M_f^{*(s)}(X)$ for all $X \in \mathbb{IR}$, the quadratic convergence of $M_f^{*(s)}$ follows. □

The similarity of the slope form and the mean value form motivates a comparison.

Lemma 3.2.10 *For all $X \in \mathbb{IR}$ it holds that*

$$g_c(X) \subseteq f'(X). \ \square$$

Proof. Let $y \in g_c(X)$ arbitrary but fixed. We have to show that $y \in f'(X)$. Let $\hat{x} \in X$ such that $y = g_c(\hat{x})$.

- Assume $\hat{x} = c$. As $g_c(c) = f'(c)$ (see for example [Lipson, 1981]), it follows that $y \in f'(X)$.
- Assume $\hat{x} \neq c$. From the mean value Theorem it follows that there exists $\xi \in X$ such that

  $$f(\hat{x}) = f(c) + f'(\xi)(\hat{x} - c).$$

  According to (3.2.2),

  $$f(\hat{x}) = f(c) + y(\hat{x} - c).$$

  Hence $y = f'(\xi) \in f'(X)$. $\square$

Lemma 3.2.10 gives rise to the conjecture that the slope form yields always tighter inclusions than the mean value form. However, this is not true.

Theorem 3.2.11 *There exist polynomials f and intervals X such that*

$$M_f^{(\mathrm{s})}(X) \supset M_f(X). \ \square \tag{3.2.5}$$

Proof. Let

$$f(x) = x^3 - 2x, \quad X = [-0.5, 1] \tag{3.2.6}$$

We obtain

$$\begin{aligned} M_f(X) &= [-1.984375, 1.015625] \\ M_f^{(\mathrm{s})}(X) &= [-2.40625, 1.4375]. \ \square \end{aligned}$$

The reason why (3.2.5) holds in the example above is that f' is sparse, whereas g_c is dense. Therefore, the superiority of the interval power function over interval multiplication reveals in the Horner evaluation of f', but not in the Horner evaluation of g_c. This gives rise to the following conjecture.

Conjecture 3.2.12 *For all $X \in \mathbb{IR}$ it holds that*

$$M_f^{*(\mathrm{s})}(X) \subseteq M_f^*(X).$$

Justification. It suffices to show that

$$H_{g_c}^*(X) \subseteq H_{f'}^*(X). \tag{3.2.7}$$

Equation (3.2.7) can be written explicitly as

$$\begin{aligned} & \big((a_n^*X + a_n^*c + a_{n-1}^*)X + a_n^*c^2 + a_{n-1}^*c + a_{n-2}^*\big)X + \ldots + a_n^*c^{n-1} + a_{n-1}^*c^{n-2} + \ldots + a_2^*c + a_1^* \\ \subseteq \ & \big((na_n^*X + (n-1)a_{n-1}^*)X + (n-2)a_{n-2}^*)X + \ldots + 2a_2^*\big)X + a_1^*, \end{aligned}$$

where a_i^* is the coefficient of x^i in f. Experimental results indicate that the following inclusions are valid for arbitrary $c \in X$. We give formulas for the case $n = 5$, the generalization is straight forward.

$$\begin{aligned} & \Big(\big((a_5^*X + a_5^*c + a_4^*)X + a_5^*c^2 + a_4^*c + a_3^*\big)X + a_5^*c^3 + a_4^*c^2 + a_3^*c + a_2^*\Big)X \\ & + a_5^*c^4 + a_4^*c^3 + a_3^*c^2 + a_2^*c + a_1^* \\ \subseteq \ & \Big(\big((2a_5^*X + a_5^*c + 2a_4^*)X + a_5^*c^2 + a_4^*c + 2a_3^*\big)X + a_5^*c^3 + a_4^*c^2 + a_3^*c + 2a_2^*\Big)X + a_1^* \\ \subseteq \ & \Big(\big((3a_5^*X + a_5^*c + 3a_4^*)X + a_5^*c^2 + a_4^*c + 3a_3^*\big)X + 2a_2^*\Big)X + a_1^* \\ \subseteq \ & \Big(\big((4a_5^*X + a_5^*c + 4a_4^*)X + 3a_3^*\big)X + 2a_2^*\Big)X + a_1^* \\ \subseteq \ & \Big(\big((5a_5^*X + 4a_4^*)X + 3a_3^*\big)X + 2a_2^*\Big)X + a_1^*. \end{aligned}$$

Similar inclusions have been proved by [Alefeld, 1981]. However, for the case above, it seems that no proof is known. □

We give an algorithm for the dense slope form and not for the slope form because of the following reasons:

- Usually g_c is dense, even if f is sparse. This is in particular the case if the coefficients of g_c are computed by rounded interval arithmetic. Hence, in general $M_f^{(s)}(X) = M_f^{*(s)}(X)$.
- In a concrete implementation, an algorithm for $M_f^{*(s)}$ is usually faster than a corresponding algorithm for $M_f^{(s)}$, because some integer operations and the power function calls are saved.

Algorithm 3.2.13 (DSF) evaluates $M_f^{*(s)}(X)$. The coefficients of g_c and the function value of f at c are obtained simultaneously by the Horner scheme. As the midpoint of X is computed using Algorithm 2.3.33 (MID), which gives only an approximation of $\mathrm{mid}(X)$, the output interval need not be a superset of $M_f^{*(s)}(X)$, but at least contains $f(X)$.

Algorithm 3.2.13 (DSF) [Dense Slope Form]

In: $f(x) = a_n x^{d_n} + a_{n-1} x^{d_{n-1}} + \ldots + a_1 x^{d_1} \in \mathbb{F}[x]$,
$X \in \mathbb{IF}$.

Out: $\mathrm{DSF}(f, X) \in \mathbb{IF}$, $\mathrm{DSF}(f, X) \approx M_f^{*(s)}(X)$, $\mathrm{DSF}(f, X) \supseteq f(X)$.

(1) [Midpoint.]
$c \longleftarrow \mathrm{MID}(X)$.

(2) [Coefficients of $g_c(x)$ and $f(c)$.]
$k \longleftarrow d_n$.
$B_k \longleftarrow a_n$.
for $i = n-1, \ldots, 1$
  for $j = d_{i+1}, \ldots, d_i + 1$
    $k \longleftarrow k-1$, $B_k \longleftarrow B_{k+1} c$.
  $B_k \longleftarrow B_k + a_i$.
for $j = d_1, \ldots, 1$
  $k \longleftarrow k-1$, $B_k \longleftarrow B_{k+1} c$.

(3) [Evaluate $H_{g_c}^*(X)$.]
if $d_n < 1$ then $G \longleftarrow 0$, goto Step 4.
$G \longleftarrow B_{d_n}$.
for $i = d_n - 1, \ldots, 1$ do $G \longleftarrow GX + B_i$.

(4) [Return.]
return $B_0 + G * (X - c)$.

Theorem 3.2.14 (Complexity) *Algorithm 3.2.13 (DSF) costs*

d_n	*interval multiplications*
$2d_n + 1$	*number multiplication and*
$2d_n + 2n$	*number additions.* □

Proof.

- Step 1 costs 1 number multiplication and 2 number additions.
- Step 2 costs $2d_n$ number multiplications and $2n - 2$ number additions.
- Step 3 costs $d_n - 1$ interval multiplications and $2d_n - 2$ number additions.
- Step 4 costs 1 interval multiplication and 2 number additions. □

3.2.2 Bicentered Mean Value Form

The bicentered mean value form is a modification of the mean value form, which was introduced by [Baumann, 1988]. The basic idea is to evaluate the mean value form twice with different centers $c^{\uparrow}$ and $c^{\downarrow}$ and intersect the results. The centers are optimal in the sense that $c^{\uparrow}$ minimizes the upper bound and $c^{\downarrow}$ maximizes the lower bound. The bicentered mean value form is therefore particularly suitable for the separate computation of an upper or lower bound of f in X.

In the sequel let

$$M_f(X,c) = f(c) + H_{f'}(X)(X-c).$$

Example. Let $f(x) = x^2 - x - 5$ and let $X = [0,4]$. Then $H_{f'}(X) = [-1,7]$ and we obtain

$$M_f(X, \mathrm{mid}(X)) = M_f(X) = [-17, 11].$$

Let $c^{\uparrow} = 3.5$, $c^{\downarrow} = 0.5$. Then

$$\begin{aligned} M_f(X,c^{\uparrow}) &= [-20.75, 7.25] \\ M_f(X,c^{\downarrow}) &= [-8.75, 19.25] \\ M_f(X,c^{\uparrow}) \cap M_f(X,c^{\downarrow}) &= [-8.75, 7.25]. \end{aligned}$$

An illustration of this example is given in Figure 3.2.1. The dashed lines are the linear functions

$$f(c) + \overline{H_{f'}(X)}(x-c) \quad \text{and} \quad f(c) + \underline{H_{f'}(X)}(x-c) \tag{3.2.8}$$

where $c = \mathrm{mid}(X)$, $c = c^{\uparrow}$ and $c = c^{\downarrow}$ from left to right. From the picture it is clear, why $c^{\uparrow}$ and $c^{\downarrow}$ can be called optimal centers. Further, we can see the following important property, which will be used later on:

$$\begin{aligned} f(c^{\uparrow}) + \underline{H_{f'}(X)}(\underline{X} - c^{\uparrow}) &= f(c^{\uparrow}) + \overline{H_{f'}(X)}(\overline{X} - c^{\uparrow}) \\ f(c^{\downarrow}) + \underline{H_{f'}(X)}(\overline{X} - c^{\downarrow}) &= f(c^{\downarrow}) + \overline{H_{f'}(X)}(\underline{X} - c^{\downarrow}). \end{aligned}$$

(Note that this is only true if $0 \in H_{f'}(X)$).

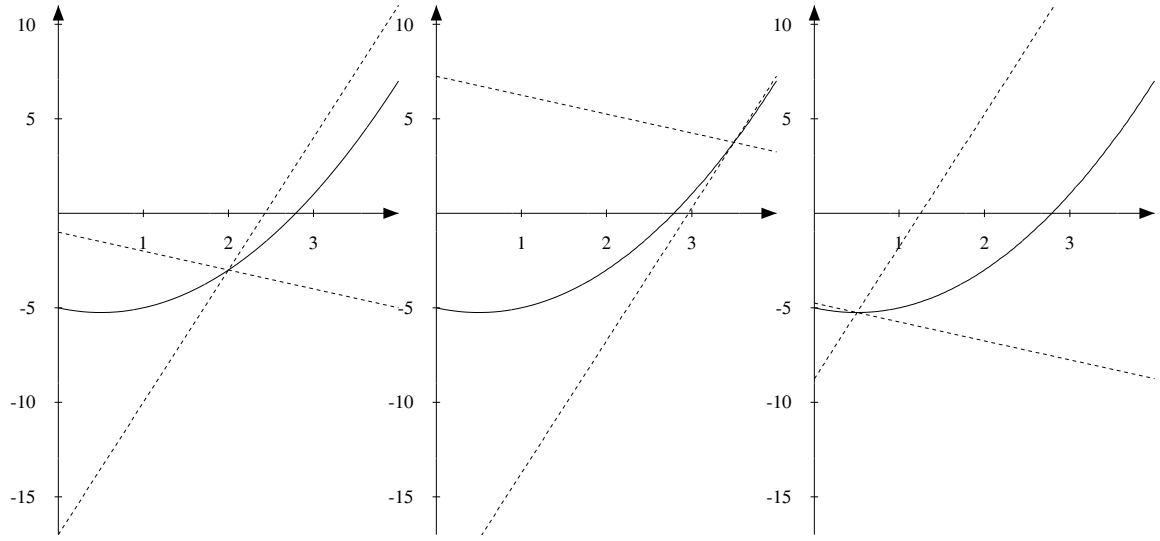


Figure 3.2.1: Geometric interpretation of the mean value form of $f(x) = x^2 - x - 5$ with different centers. From left to right: $M_f(X, \mathrm{mid}(X))$, $M_f(X, c^{\uparrow})$, $M_f(X, c^{\downarrow})$.

Definition 3.2.15 (Bicentered Mean Value Form) *The bicentered mean value form* $\ddot{M}_f : \mathbb{IR} \to \mathbb{IR}$ *is defined as*

$$\ddot{M}_f(X) = [\underline{M_f(X, c^{\downarrow})}, \overline{M_f(X, c^{\uparrow})}], \tag{3.2.9}$$

where the optimal centers $c^{\uparrow}$ *and* $c^{\downarrow}$ *are*

$$c^{\uparrow} = \begin{cases} \overline{X} & \text{if } \underline{F'} \geq 0 \\ \underline{X} & \text{if } \overline{F'} \leq 0 \\ (\overline{F'}\,\overline{X} - \underline{F'}\,\underline{X})/\mathrm{w}(F') & \text{else} \end{cases}$$

$$c^{\downarrow} = \begin{cases} \underline{X} & \text{if } \underline{F'} \geq 0 \\ \overline{X} & \text{if } \overline{F'} \leq 0 \\ (\overline{F'}\,\underline{X} - \underline{F'}\,\overline{X})/\mathrm{w}(F') & \text{else} \end{cases}$$

and $F' = H_{f'}(X)$. □

In the sequel let $c^{\uparrow}, c^{\downarrow}$ and F' as in Definition 3.2.15.

Theorem 3.2.16 (Interval Extension) $\ddot{M}_f$ *is an interval extension of* f. □

Proof. As $c^{\uparrow}, c^{\downarrow}$ are convex linear combinations of $\overline{X}$ and $\underline{X}$, it holds that $c^{\uparrow}, c^{\downarrow} \in X$. From (3.2.9) it follows that

$$\ddot{M}_f(X) \supseteq M_f(X, c^{\downarrow}) \cap M_f(X, c^{\uparrow}).$$

According to (3.2.1), $f(X) \subseteq M_f(X, c)$ for all $c \in X$. Hence,

$$f(X) \subseteq M_f(X, c^{\downarrow}) \cap M_f(X, c^{\uparrow}) \subseteq \ddot{M}_f(X).$$

Finally, if $X \in \mathbb{R}$ then obviously $\ddot{M}_f(X) \in \mathbb{R}$. □

Theorem 3.2.17 (Non–Overestimation) *If* $\underline{F'} \geq 0$ *or* $\overline{F'} \leq 0$ *then*

$$\ddot{M}_f(X) = f(X). \; \square$$

Proof.

- Assume $\underline{F'} \geq 0$.

$$\begin{aligned} \ddot{M}_f(X) &= [\underline{M_f(X, \underline{X})}, \overline{M_f(X, \overline{X})}] \\ &= [f(\underline{X}) + \underline{F' * (X - \underline{X})}, f(\overline{X}) + \overline{F' * (X - \overline{X})}] \\ &= [f(\underline{X}), f(\overline{X})]. \end{aligned}$$

- Assume $\overline{F'} \leq 0$.

$$\begin{aligned} \ddot{M}_f(X) &= [\underline{M_f(X, \overline{X})}, \overline{M_f(X, \underline{X})}] \\ &= [f(\underline{X}) + \underline{F' * (X - \overline{X})}, f(\underline{X}) + \overline{F' * (X - \underline{X})}] \\ &= [f(\underline{X}), f(\overline{X})]. \end{aligned}$$

In both cases f is monotone in X, hence $[f(\underline{X}), f(\overline{X})] = f(X)$. □

Theorem 3.2.18 (Optimality of Centers) *For all* $c \in X$ *it holds that*

$$\overline{M_f(X, c^{\uparrow})} \leq \overline{M_f(X, c)} \tag{3.2.10}$$

$$\underline{M_f(X, c^{\downarrow})} \geq \underline{M_f(X, c)}. \tag{3.2.11}$$

Proof. If $\underline{F'} \geq 0$ or $\overline{F'} \leq 0$ then (3.2.10) and (3.2.11) follow from Theorem 3.2.17. Hence assume $0 \in \mathrm{int}(F')$ and let $c \in X$ arbitrary but fixed. We show (3.2.10), the proof of (3.2.11) is similar.

$$\begin{aligned} \overline{M_f(X,c)} &= f(c) + \overline{F' * (X-c)} \\ &\geq f(c^\uparrow) + \underline{F' * (c-c^\uparrow)} + \overline{F' * (X-c)}. \end{aligned} \tag{3.2.12}$$

Next, we show that

$$\underline{F' * (c-c^\uparrow)} + \overline{F' * (X-c)} = \overline{F' * (X-c^\uparrow)}. \tag{3.2.13}$$

As $0 \in F'$ and $0 \in X - c$ we have

$$\overline{F' * (X-c)} = \max\{\underline{F'} * (\underline{X} - c), \overline{F'} * (\overline{X} - c)\} \tag{3.2.14}$$

and from the definition of $c^\uparrow$ it follows that

$$\overline{F' * (X-c^\uparrow)} = \underline{F'} * (\underline{X} - c^\uparrow) = \overline{F'} * (\overline{X} - c^\uparrow). \tag{3.2.15}$$

- Assume $c \geq c^\uparrow$. From (3.2.14) and (3.2.15) it follows that

  $$\begin{aligned} \overline{F' * (X-c)} &= \underline{F'} * (\underline{X} - c) \\ \underline{F' * (c-c^\uparrow)} &= \underline{F'} * (c - c^\uparrow). \end{aligned}$$

  Hence,

  $$\underline{F' * (c-c^\uparrow)} + \overline{F' * (X-c)} = \underline{F'} * (\underline{X} - c^\uparrow) = \overline{F' * (X-c^\uparrow)}.$$

- Assume $c < c^\uparrow$. From (3.2.14) and (3.2.15) it follows that

  $$\begin{aligned} \overline{F' * (X-c)} &= \overline{F'} * (\overline{X} - c) \\ \underline{F' * (c-c^\uparrow)} &= \overline{F'} * (c - c^\uparrow). \end{aligned}$$

  Hence,

  $$\underline{F' * (c-c^\uparrow)} + \overline{F' * (X-c)} = \overline{F'} * (\overline{X} - c^\uparrow) = \overline{F' * (X-c^\uparrow)}.$$

After having shown (3.2.13), we can continue with (3.2.12).

$$\begin{aligned} \overline{M_f(X,c)} &\geq f(c^\uparrow) + \overline{F' * (X-c^\uparrow)} \\ &= \overline{M_f(X,c^\uparrow)}. \ \square \end{aligned}$$

Corollary 3.2.19

$$\ddot{M}_f(X) = \bigcap_{c \in X} M_f(X,c). \ \square$$

Corollary 3.2.20 (Convergence) *$\ddot{M}_f$ converges quadratically to f.* $\square$

Proof. According to Corollary 3.2.19, $\ddot{M}_f(X) \subseteq M_f(X)$ for all $X \in \mathbb{IR}$, and M_f is quadratically convergent by Theorem 3.2.4. $\square$

Theorem 3.2.21 (Inclusion Monotonicity) *$\ddot{M}_f$ is inclusion monotone.* $\square$

Proof. Let $X_\diamond \subseteq X$, let $c^\uparrow_\diamond, c^\downarrow_\diamond$ be the optimal centers of f in $X_\diamond$, and let $F'_\diamond = H_{f'}(X_\diamond)$. We have to show that

$$\ddot{M}_f(X_\diamond) \subseteq \ddot{M}_f(X), \tag{3.2.16}$$

or equivalently

$$\overline{M_f(X_\diamond, c^\uparrow_\diamond)} \leq \overline{M_f(X, c^\uparrow)} \tag{3.2.17}$$

$$\underline{M_f(X_\diamond, c^\downarrow_\diamond)} \geq \underline{M_f(X, c^\downarrow)} \tag{3.2.18}$$

We give a proof of (3.2.17), the proof of (3.2.18) is analogous. If $\underline{F'_\diamond} \geq 0$ or $\overline{F'_\diamond} \leq 0$ then

$$\ddot{M}_f(X_\diamond) = f(X_\diamond) \subseteq f(X) \subseteq \ddot{M}_f(X)$$

by Theorem 3.2.17. In the sequel assume $0 \in \mathrm{int}(F'_\diamond)$. Inclusion monotonicity of the Horner form implies $0 \in \mathrm{int}(F')$.

- Assume $c^\uparrow_\diamond \geq c^\uparrow$. Then

$$f(c^\uparrow_\diamond) \leq f(c^\uparrow) + \overline{F'} * (c^\uparrow_\diamond - c^\uparrow). \tag{3.2.19}$$

Hence,

$$\begin{aligned}
\overline{M_f(X_\diamond, c^\uparrow_\diamond)} &= f(c^\uparrow_\diamond) + \overline{F'_\diamond * (X_\diamond - c^\uparrow_\diamond)} \\
&\overset{(3.2.15)}{=} f(c^\uparrow_\diamond) + \overline{F'_\diamond} * (\overline{X_\diamond} - c^\uparrow_\diamond) \\
&\overset{(3.2.19)}{\leq} f(c^\uparrow) + \overline{F'} * (c^\uparrow_\diamond - c^\uparrow) + \overline{F'_\diamond} * (\overline{X_\diamond} - c^\uparrow_\diamond) \\
&\leq f(c^\uparrow) + \overline{F'} * (c^\uparrow_\diamond - c^\uparrow) + \overline{F'} * (\overline{X} - c^\uparrow_\diamond) \\
&= f(c^\uparrow) + \overline{F'} * (\overline{X} - c^\uparrow) \\
&\overset{(3.2.15)}{=} f(c^\uparrow) + \overline{F' * (X - c^\uparrow)} \\
&= \overline{M_f(X, c^\uparrow)}. \ \square
\end{aligned}$$

- Assume $c^\uparrow_\diamond \leq c^\uparrow$. Then

$$f(c^\uparrow_\diamond) \leq f(c^\uparrow) + \underline{F'} * (c^\uparrow_\diamond - c^\uparrow). \tag{3.2.20}$$

Hence,

$$\begin{aligned}
\overline{M_f(X_\diamond, c^\uparrow_\diamond)} &= f(c^\uparrow_\diamond) + \overline{F'_\diamond * (X_\diamond - c^\uparrow_\diamond)} \\
&\overset{(3.2.15)}{=} f(c^\uparrow_\diamond) + \underline{F'_\diamond} * (\underline{X_\diamond} - c^\uparrow_\diamond) \\
&\overset{(3.2.20)}{\leq} f(c^\uparrow) + \underline{F'} * (c^\uparrow_\diamond - c^\uparrow) + \underline{F'_\diamond} * (\underline{X_\diamond} - c^\uparrow_\diamond) \\
&\leq f(c^\uparrow) + \underline{F'} * (c^\uparrow_\diamond - c^\uparrow) + \underline{F'} * (\underline{X} - c^\uparrow_\diamond) \\
&= f(c^\uparrow) + \underline{F'} * (\underline{X} - c^\uparrow) \\
&\overset{(3.2.15)}{=} f(c^\uparrow) + \overline{F' * (X - c^\uparrow)} \\
&= \overline{M_f(X, c^\uparrow)}. \ \square
\end{aligned}$$

In Figure 3.2.1 one sees that the width of the mean value form is the same for all centers c between $c^\uparrow$ and $c^\downarrow$. In this case the width is determined by the steeper dashed line. If the center is chosen outside $[c^\downarrow, c^\uparrow]$ then the width of the mean value form is larger. The following theorem is new.

Theorem 3.2.22 *Let $C = [c^\downarrow, c^\uparrow]$. For all $c \in X$ it holds that*

$$\mathrm{w}(M_f(X, c)) = \mathrm{mag}(F')\mathrm{w}(X) + \mathrm{w}(F')\mathrm{mig}(C - c). \ \square$$

Proof. Assume $0 \notin \mathrm{int}(F')$. Then $C = X$ and for all $c \in X$ it holds that $\mathrm{mig}(C - c) = 0$. Thus,

$$\begin{aligned}
\mathrm{w}(M_f(X, c)) &= \mathrm{w}(F' * (X - c)) \\
&= \mathrm{w}(\mathrm{mag}(F')(X - c)) \\
&= \mathrm{mag}(F')\mathrm{w}(X).
\end{aligned}$$

Assume $0 \in \mathrm{int}(F')$. Further, assume $\overline{F'} \geq -\underline{F'}$, i.e. $\mathrm{mag}(F') = \overline{F'}$. The case $\overline{F'} < -\underline{F'}$ can be treated similarly. Note that

$$\begin{aligned}
\overline{F'}\,\overline{X} - \underline{F'}\,\underline{X} &\geq \overline{F'}\,\overline{X} - \underline{F'}\,\underline{X} - (\overline{F'} + \underline{F'})(\overline{X} - \underline{X}) \\
&= \overline{F'}\,\underline{X} - \underline{F'}\,\overline{X}.
\end{aligned}$$

Hence, $c^{\downarrow} \leq c^{\uparrow}$. Further, for all $c \in X$

$$\overline{F' * (X - c)} = \max\{\underline{F'} * (\underline{X} - c), \overline{F'} * (\overline{X} - c)\} \quad (3.2.21)$$
$$\underline{F' * (X - c)} = \min\{\underline{F'} * (\overline{X} - c), \overline{F'} * (\underline{X} - c)\} \quad (3.2.22)$$

and

$$\underline{F'} * (\underline{X} - c^{\uparrow}) = \overline{F'} * (\overline{X} - c^{\uparrow}) \geq 0 \quad (3.2.23)$$
$$\underline{F'} * (\overline{X} - c^{\downarrow}) = \overline{F'} * (\underline{X} - c^{\downarrow}) \leq 0. \quad (3.2.24)$$

- Assume $c \in C$. Note that $\mathrm{mig}(C - c) = 0$. As $c \leq c^{\uparrow}$ and $c \geq c^{\downarrow}$, we obtain from (3.2.21) – (3.2.24)

$$\begin{aligned} \overline{F' * (X - c)} &= \overline{F'} * (\overline{X} - c) \\ \underline{F' * (X - c)} &= \overline{F'} * (\underline{X} - c). \end{aligned}$$

  Thus,

$$\begin{aligned} \mathrm{w}(M_f(X, c)) &= \overline{F' * (X - c)} - \underline{F' * (X - c)} \\ &= \overline{F'} * (\overline{X} - \underline{X}) \\ &= \overline{F'}\mathrm{w}(X). \end{aligned}$$

- Assume $c \notin C$. We give a proof only for the case $c > c^{\uparrow}$, the case $c < c^{\downarrow}$ is analogous. As $c > c^{\downarrow}$ and $c > c^{\uparrow}$ we obtain from (3.2.21) – (3.2.24)

$$\begin{aligned} \overline{F' * (X - c)} &= \underline{F'} * (\underline{X} - c) \\ \underline{F' * (X - c)} &= \overline{F'} * (\underline{X} - c). \end{aligned}$$

  Thus,

$$\begin{aligned} \mathrm{w}(M_f(X, c)) &= \overline{F' * (X - c)} - \underline{F' * (X - c)} \\ &= (\underline{F'} - \overline{F'})(\underline{X} - c) \\ &= (\underline{F'} - \overline{F'})(\underline{X} - c^{\uparrow}) + (\underline{F'} - \overline{F'})(c^{\uparrow} - c) \\ &= \underline{F'} * (\underline{X} - c^{\uparrow}) - \overline{F'} * (\underline{X} - c^{\uparrow}) + \mathrm{w}(F)\mathrm{mig}(C - c) \\ &= \overline{F'} * (\overline{X} - c^{\uparrow}) - \overline{F'} * (\underline{X} - c^{\uparrow}) + \mathrm{w}(F)\mathrm{mig}(C - c) \\ &= \overline{F'}\mathrm{w}(X) + \mathrm{w}(F)\mathrm{mig}(C - c). \ \square \end{aligned}$$

The midpoint of X is an optimal center for the mean value form in the following sense:

Corollary 3.2.23 (Optimality of the Midpoint for the Mean Value Form) *For all $c \in X$ it holds that*

$$\mathrm{w}(M_f(X)) \leq \mathrm{w}(M_f(X, c)). \ \square$$

Proof. According to Theorem 3.2.22, we have to show that $\mathrm{mid}(X) \in [c^{\downarrow}, c^{\uparrow}]$. In fact, we show that $\mathrm{mid}(X) = \mathrm{mid}([c^{\downarrow}, c^{\uparrow}])$. If $0 \notin \mathrm{int}(F')$ then $[c^{\downarrow}, c^{\uparrow}] = X$. If $0 \in \mathrm{int}(F')$ then

$$\begin{aligned} 1/2(c^{\downarrow} + c^{\uparrow}) &= \frac{\overline{F'}\overline{X} - \underline{F'}\underline{X} + \overline{F'}\underline{X} - \underline{F'}\overline{X}}{2\mathrm{w}(F')} \\ &= \frac{(\overline{F'} - \underline{F'})(\overline{X} + \underline{X})}{2(\overline{F'} - \underline{F'})} \\ &= 1/2(\overline{X} + \underline{X}) \\ &= \mathrm{mid}(X). \ \square \end{aligned}$$

A comparison between the slope form and the bicentered mean value form might be of interest. The bicentered mean value form returns the range without overestimation if $0 \notin \mathrm{int}(H_{f'}(X))$. This is not the case for the slope form, e.g. $f(x) = x^2 + 3x$, $X = [-1, 2]$. If $0 \in \mathrm{int}(H_{f'}(X))$ sometimes the slope form is better than the bicentered mean value form, sometimes vice versa:

- $f(x) = -x^3 + x^2 - x$, $X = [-1, 1]$.

$$M_f^{(s)}(X) = [-3, 3] \subset [-4.968, 5.048] = \ddot{M}_f(X)$$

- $f(x) = -x^3 + x^2 + x$, $X = [-0.4, 1]$.

$$M_f^{(s)}(X) = [-1.254, 1.98] \supset [-0.551826, 1.35286] = \ddot{M}_f(X).$$

Algorithm 3.2.26 (BMF) for evaluating the bicentered mean value form follows immediately from Definition 3.2.15. First, we give Algorithm 3.2.24 (OC) for computing the optimal centers. Note that an approximation of the optimal centers is sufficient for the correctness of Algorithm 3.2.26, provided that the approximations are elements of X.

Algorithm 3.2.24 (OC) [Optimal Centers]

In: $F' \in \mathbb{IF}$,
$X \in \mathbb{IF}$.

Out: $c^\uparrow, c^\downarrow \in X$, approximations of the optimal centers according to Definition 3.2.15.

(1) [Monotonicity test.]
if $\underline{\overline{F'}} \geq 0$ return $\overline{X}, \underline{\overline{X}}$.
if $\overline{\underline{F'}} \leq 0$ return $\underline{X}, \overline{\underline{X}}$.

(2) [Approximate.]
clear invalid operation flag
$c^\uparrow \longleftarrow (\overline{F'} \overset{\approx}{*} \overline{X} \overset{\approx}{-} \underline{F'} \overset{\approx}{*} \underline{X}) \overset{\approx}{/} (\overline{F'} \overset{\approx}{-} \underline{F'})$.
$c^\downarrow \longleftarrow (\overline{F'} \overset{\approx}{*} \underline{X} \overset{\approx}{-} \underline{F'} \overset{\approx}{*} \overline{X}) \overset{\approx}{/} (\overline{F'} \overset{\approx}{-} \underline{F'})$.

(3) [Check invalid operation and rounding errors.]
if invalid operation flag is raised then $c^\uparrow \longleftarrow \overline{X}$, $c^\downarrow \longleftarrow \underline{X}$.
if $c^\uparrow < \underline{X}$ then $c^\uparrow \longleftarrow \underline{\overline{X}}$.
if $c^\uparrow > \overline{X}$ then $c^\uparrow \longleftarrow \overline{X}$.
if $c^\downarrow < \underline{X}$ then $c^\downarrow \longleftarrow \underline{X}$.
if $c^\downarrow > \overline{X}$ then $c^\downarrow \longleftarrow \overline{\underline{X}}$.

(4) [Return.]
return $c^\uparrow, c^\downarrow$.

Note that in Step 2 $\overline{F} > 0$ and $\underline{F} < 0$, hence $\overline{F'} \overset{\approx}{-} \underline{F'} \neq 0$ and division by zero cannot occur.

Theorem 3.2.25 (Complexity) *Algorithm 3.2.24* (OC) *costs*

2 *number divisions,*
4 *number multiplications and*
3 *number additions.* □

Algorithm 3.2.26 (BMF) [Bicentered Mean Value Form]

In: $f(x) = a_n x^{d_n} + a_{n-1} x^{d_{n-1}} + \ldots + a_1 x^{d_1} \in \mathbb{F}[x]$,
$X \in \mathbb{IF}$.

Out: $\mathrm{BMF}(f, X) \in \mathbb{IF}$, $\mathrm{BMF}(f, X) \approx \ddot{M}_f(X)$, $\mathrm{BMF}(f, X) \supseteq f(X)$.

(1) [Inclusion of f'.]
$F'(x) \longleftarrow f'(x)$ (use interval arithmetic for coefficients).

(2) [Evaluate $f'(X)$.]
$F' \longleftarrow \mathrm{HFI}(F'(x), X)$.

(3) [Optimal centers.]
$c^{\uparrow}, c^{\downarrow} \longleftarrow \mathrm{OC}(F', X)$.

(4) [Evaluate at Centers.]
$y^{\uparrow} \longleftarrow \mathrm{HFUB}(f(x), [c^{\uparrow}]) \mathbin{\hat{+}} \overline{F' * (X - c^{\uparrow})}$.
$y^{\downarrow} \longleftarrow \mathrm{HFLB}(f(x), [c^{\downarrow}]) \mathbin{\check{+}} \underline{F' * (X - c^{\downarrow})}$.
return $[y^{\downarrow}, y^{\uparrow}]$.

Theorem 3.2.27 (Complexity) *Algorithm 3.2.26 (*BMF*) costs*

n	*interval power computations,*
$n+2$	*interval multiplications,*
$2n$	*number power computations,*
2	*number divisions,*
$4n+4$	*number multiplications and*
$4n+5$	*number additions.* □

Proof.

- Step 1 costs $2n$ number multiplications.
- Step 2 costs n interval power computations, n interval multiplications and $2n-2$ number additions (Theorem 3.1.8).
- Step 3 costs 2 number divisions, 4 number multiplications and 3 number additions (Theorem 3.2.25).
- Step 4 costs $2n$ number power computations, $2n$ number multiplications and $2n-2$ number additions during the calls of HFLB and HFUB (Theorem 3.1.37). Further, it takes 2 interval multiplications and 6 number additions. □

3.2.3 Experimental Results

An experimental comparison of the mean value form M_f, computed by Algorithm 3.2.5 (MF), the dense slope form $M_f^{*(\mathrm{s})}$, computed by Algorithm 3.2.13 (DSF) and the bicentered mean value form $\ddot{M}_f$, computed by Algorithm 3.2.26 (BMF) for dense and sparse polynomials with different degrees is given in Table 3.2.1, respectively Table 3.2.2. The coefficients of the polynomials and the endpoints of the input intervals are uniformly distributed in $[-1, 1]$. For each degree d_n, the average cost and overestimation error of 10^4 random polynomials is reported. The cost is the total number of arithmetic floating point instructions, including those which were executed during interval operations. As the forms are quadratically convergent, the distance to the range divided by $\mathrm{w}(X)^2$ was chosen to measure accuracy. In almost all cases $\ddot{M}_f$ is more accurate than $M_f^{(\mathrm{s})}$, which is significantly more accurate than M_f. The costs of $\ddot{M}_f$ and M_f are almost the same. If the polynomials are "sufficiently" sparse, then $\ddot{M}_f$ and M_f are more efficient than $M_f^{*(\mathrm{s})}$. In the dense case $M_f^{*(\mathrm{s})}$ is cheaper.

3.3 Taylor Form

The basic idea of the Taylor form is to translate f by $\mathrm{mid}(X)$ to the left and evaluate the Horner form of resulting polynomial on the centered interval $X - \mathrm{mid}(X)$. The Taylor form is a centered form and hence quadratically convergent.

The Taylor form for polynomials and rational functions was studied thoroughly in the literature, see e.g. [Alefeld and Herzberger, 1974], [Alefeld and Rokne, 1981], [Alefeld and Herzberger, 1983], [Hansen,

Degree of f	2	6	10	14	18	22	26	30
Flops for $M_f(X)$	28.6	71.2	113.7	156.2	198.8	241.3	283.7	325.8
Flops for $M_f^{*(\mathrm{s})}(X)$	20.3	54.2	88.5	122.7	157.1	191.5	225.8	259.7
Flops for $\ddot{M}_f(X)$	25.3	73.2	116.5	159.1	202.1	244.5	287.1	329.1
$\mathrm{q}(M_f(X), f(X))/\mathrm{w}(X)^2$	0.36	1.50	2.55	3.47	4.40	5.48	6.24	7.00
$\mathrm{q}(M_f^{*(\mathrm{s})}(X), f(X))/\mathrm{w}(X)^2$	0.23	0.70	1.05	1.30	1.56	1.87	2.07	2.16
$\mathrm{q}(\ddot{M}_f(X), f(X))/\mathrm{w}(X)^2$	0.07	0.43	0.77	1.04	1.34	1.66	1.90	2.24

Table 3.2.1: Comparison of M_f, $M_f^{*(\mathrm{s})}$ and $\ddot{M}_f$ for *dense* random polynomials f with different degrees and random intervals X. The coefficients of f and the endpoints of X are uniformly chosen in $[-1, 1]$.

Degree of f	2	6	10	14	18	22	26	30
Flops for $M_f(X)$	28.6	46.0	57.6	65.3	71.4	75.7	79.5	82.9
Flops for $M_f^{*(\mathrm{s})}(X)$	20.3	47.3	74.6	102.4	130.3	158.2	186.3	213.9
Flops for $\ddot{M}_f(X)$	25.3	45.0	56.5	64.1	70.5	74.6	78.4	81.6
$\mathrm{q}(M_f(X), f(X))/\mathrm{w}(X)^2$	0.36	0.95	1.26	1.42	1.55	1.72	1.71	1.90
$\mathrm{q}(M_f^{*(\mathrm{s})}(X), f(X))/\mathrm{w}(X)^2$	0.23	0.46	0.54	0.57	0.59	0.64	0.60	0.66
$\mathrm{q}(\ddot{M}_f(X), f(X))/\mathrm{w}(X)^2$	0.07	0.16	0.17	0.16	0.19	0.18	0.20	0.21

Table 3.2.2: Comparison of M_f, $M_f^{*(\mathrm{s})}$ and $\ddot{M}_f$ for *sparse* random polynomials f with different degrees and random intervals X. Each polynomial has only 3 non–zero coefficients. The coefficients of f and the endpoints of X are uniformly chosen in $[-1, 1]$.

1969], [Moore, 1966], [Rall, 1983], [Ratschek, 1978], [Ratschek, 1980a], [Ratschek, 1980b], [Ratschek and Rokne, 1980b], [Ratschek and Rokne, 1980a], [Ratschek and Schröder, 1981], [Ratschek and Rokne, 1984]. Interval polynomials are considered in [Rokne, 1981] and complex polynomials in [Rokne and Wu, 1982], [Rokne and Wu, 1983]. In Section 3.3.1 we give a new improvement of the Taylor form which is obtained through a bisection of the input interval. The overestimation error is thereby reduced at least by half.

According to Taylor's Theorem we can rewrite f for arbitrary $c \in \mathbb{R}$ as

$$f(x) = \sum_{i=0}^{d_n} \frac{f^{(i)}(c)}{i!}(x-c)^i.$$

If we translate f by c to the left, we obtain the polynomial

$$f^{(c)}(x) = \sum_{i=0}^{d_n} \frac{f^{(i)}(c)}{i!}x^i \tag{3.3.1}$$

and obviously

$$f(x) = f^{(c)}(x-c).$$

Thus, a whole class of interval extensions of f is given by

$$f(X) \subseteq F^{(c)}(X-c), \tag{3.3.2}$$

where $F^{(c)}$ is an interval extension of $f^{(c)}$. The Taylor form is a special case of (3.3.2), where $F^{(c)}$ is the Horner form of $f^{(c)}$ and c is the midpoint of X.

Definition 3.3.1 (Taylor Form) *The Taylor form $T_f : \mathbb{IR} \to \mathbb{IR}$ of f is defined as*

$$T_f(X) = H_{f(\mathrm{mid}(X))}(X - \mathrm{mid}(X)).$$

The dense Taylor form $T_f^* : \mathbb{IR} \to \mathbb{IR}$ of f is defined as

$$T_f^*(X) = H^*_{f(\mathrm{mid}(X))}(X - \mathrm{mid}(X)). \ \square$$

From Theorem 3.1.3 it follows that $T_f(X) \subseteq T_f^*(X)$ for all $X \in \mathbb{IR}$. The reason why it is still worthwhile to consider T_f^* is that it can be evaluated more efficiently than T_f. This is surprising, because the evaluation of H_f is usually cheaper than the evaluation of H_f^*. Another reason is that T_f^* is inclusion monotone, whereas T_f is not. Finally, $T_f(X) \subset T_f^*(X)$ only if a Taylor coefficient vanishes, which is "in general" not the case, especially if rounded interval arithmetic is used for their computation.

In the sequel let $c = \mathrm{mid}(X)$ and $f^{(c)}$ as in (3.3.1). Further, let

$$g^{(c)}(x) = \sum_{i=1}^{d_n} \frac{f^{(i)}(c)}{i!}x^{i-1}$$

and note that

$$T_f^*(X) = f(c) + H^*_{g^{(c)}}(X-c) * (X-c). \tag{3.3.3}$$

A corresponding representation is not possible for T_f if $f'(c) = 0$.

The following Theorem shows, why T_f^* can be evaluated faster than T_f. Further, it will be used to prove inclusion monotonicity of T_f^*.

Theorem 3.3.2 *Let $b_0, \ldots, b_{d_n-1}$ such that*

$$g^{(c)}(x) = \sum_{i=0}^{d_n-1} b_i x^i,$$

let

$$g_{||}^{(c)}(x) = \sum_{i=0}^{d_n-1} |b_i| x^i,$$

and let $r = \mathrm{rad}(X)$. Then

$$H^*_{g^{(c)}}(X-c)\,(X-c) = g_{||}^{(c)}(r)r[-1,1]. \ \square$$

Proof. Let r and $b_0, \ldots, b_{d_n-1}$ as in Theorem 3.3.2. As $X - c$ is a centered interval, it holds that

$$H^*_{g^{(c)}}(X-c)\,(X-c) = \mathrm{mag}(H^*_{g^{(c)}}(X-c))r[-1,1].$$

It remains to show that

$$\mathrm{mag}(H^*_{g^{(c)}}(X-c)) = g^{(c)}_{||}(r).$$

Let

$$\left.\begin{array}{rclrcl} P_{n-1} &=& b_{n-1}, & P_i &=& P_{i+1}(X-c)+b_i \\ p_{n-1} &=& |b_{n-1}|, & p_i &=& p_{i+1}r+|b_i| \end{array}\right\} i = n-2,\ldots,0.$$

Note that $P_0 = H^*_{g^{(c)}}(X-c)$ and $p_0 = g^{(c)}_{||}(r)$. Obviously $\mathrm{mag}(P_{n-1}) = p_{n-1}$. If $\mathrm{mag}(P_i) = p_i$ for some $1 \le i \le n-1$ then

$$\mathrm{mag}(P_{i-1}) = \mathrm{mag}(P_i * (X-c) + b_{i-1}) = \mathrm{mag}(P_i)r + |b_{i-1}| = p_{i-1},$$

hence $\mathrm{mag}(P_0) = p_0$. □

In [Caprani and Madsen, 1980] inclusion monotonicity of a centered form similar to the Taylor form (see [Ratschek, 1978], [Ratschek, 1980a]) is disproved by a counterexample. Apparently, inclusion monotonicity of the Taylor form was not studied before. In particular, Theorem 3.3.3 is new.

Theorem 3.3.3 (Inclusion Monotonicity of Dense Taylor Form) *T^*_f is inclusion monotone.* □

Proof. Let $X_\diamond \subseteq X$, $c_\diamond = \mathrm{mid}(X_\diamond)$, $r_\diamond = \mathrm{rad}(X_\diamond)$ and $r = \mathrm{rad}(X)$. Note that

$$\begin{array}{rcl} |c_\diamond - c| &=& 1/2|(\overline{X_\diamond} + \underline{X_\diamond} - \overline{X} - \underline{X})| \\ &=& 1/2\max\{\overline{X_\diamond} + \underline{X_\diamond} - \overline{X} - \underline{X}, \overline{X} + \underline{X} - \overline{X_\diamond} - \underline{X_\diamond}\} \\ &\le& 1/2(\overline{X} + \underline{X_\diamond} - \overline{X_\diamond} - \underline{X}) \\ &=& r - r_\diamond. \end{array}$$

We have to show that $T^*_f(X_\diamond) \subseteq T^*_f(X)$, which can be rewritten according to Theorem 3.3.2 as

$$f(c_\diamond) + [-1,1]\sum_{i=1}^{d_n} \frac{|f^{(i)}(c_\diamond)|}{i!} r_\diamond^i \subseteq f(c) + [-1,1]\sum_{i=1}^{d_n} \frac{|f^{(i)}(c)|}{i!} r^i. \tag{3.3.4}$$

The corresponding inequalities for the endpoints are

$$f(c_\diamond) + \sum_{i=1}^{d_n} \frac{|f^{(i)}(c_\diamond)|}{i!} r_\diamond^i \le f(c) + \sum_{i=1}^{d_n} \frac{|f^{(i)}(c)|}{i!} r^i \tag{3.3.5}$$

$$f(c_\diamond) - \sum_{i=1}^{d_n} \frac{|f^{(i)}(c_\diamond)|}{i!} r_\diamond^i \ge f(c) - \sum_{i=1}^{d_n} \frac{|f^{(i)}(c)|}{i!} r^i. \tag{3.3.6}$$

First, we express the derivatives of f at $c_\diamond$ in terms of derivatives of f at c.

$$f^{(i)}(c_\diamond) = \sum_{j=i}^{d_n} \frac{f^{(j)}(c)}{(j-i)!}(c_\diamond - c)^{j-i}.$$

Thus,

$$\begin{array}{rcl} \displaystyle\sum_{i=1}^{d_n} \frac{|f^{(i)}(c_\diamond)|}{i!} r_\diamond^i &=& \displaystyle\sum_{i=1}^{d_n} \left| \sum_{j=i}^{d_n} \frac{f^{(j)}(c)}{i!(j-i)!}(c_\diamond - c)^{j-i} \right| r_\diamond^i \\ &\le& \displaystyle\sum_{i=1}^{d_n} \sum_{j=i}^{d_n} \frac{|f^{(j)}(c)|}{i!(j-i)!} |(c_\diamond - c)|^{j-i} r_\diamond^i \end{array}$$

$$
\begin{aligned}
&\leq \sum_{i=1}^{d_n}\sum_{j=i}^{d_n}\frac{|f^{(j)}(c)|}{i!(j-i)!}(r-r_\diamond)^{j-i}r_\diamond^i \\
&= \sum_{j=1}^{d_n}\sum_{i=1}^{j}\frac{|f^{(j)}(c)|}{i!(j-i)!}(r-r_\diamond)^{j-i}r_\diamond^i \\
&= \sum_{j=1}^{d_n}\frac{|f^{(j)}(c)|}{j!}\sum_{i=1}^{j}\binom{j}{i}(r-r_\diamond)^{j-i}r_\diamond^i \\
&= \sum_{j=1}^{d_n}\frac{|f^{(j)}(c)|}{j!}\big(r^j-(r-r_\diamond)^j\big) \\
&= \sum_{j=1}^{d_n}\frac{|f^{(j)}(c)|}{j!}r^j-\sum_{j=1}^{d_n}\frac{|f^{(j)}(c)|}{j!}(r-r_\diamond)^j \\
&\leq \sum_{j=1}^{d_n}\frac{|f^{(j)}(c)|}{j!}r^j-\sum_{j=1}^{d_n}\frac{|f^{(j)}(c)|}{j!}|c_\diamond-c|^j \\
&\leq \sum_{j=1}^{d_n}\frac{|f^{(j)}(c)|}{j!}r^j-\left|\sum_{j=1}^{d_n}\frac{f^{(j)}(c)}{j!}(c_\diamond-c)^j\right| \\
&= \sum_{j=1}^{d_n}\frac{|f^{(j)}(c)|}{j!}r^j-|f(c_\diamond)-f(c)|.
\end{aligned}
$$

Finally, (3.3.5) and (3.3.6) follow from

$$
\begin{aligned}
f(c_\diamond)-|f(c_\diamond)-f(c)| &\leq f(c). \\
f(c_\diamond)+|f(c_\diamond)-f(c)| &\geq f(c). \ \square
\end{aligned}
$$

Theorem 3.3.4 *T_f is not inclusion monotone.* $\square$

Proof. Let $f(x)=x^3-3*x^2$.

- Let $X=[0,2]$, $c=1$. The Taylor coefficients are

$$
\begin{aligned}
f(c) &= -2 \\
f'(c)/1! &= -3 \\
f''(c)/2! &= 0 \\
f'''(c)/3! &= 1,
\end{aligned}
$$

  and

$$T_f(X)=-2+([-1,1]^2-3)[-1,1]=[-5,1].$$

  Note that $f''(c)/2!=0$, which caused that the interval square function was used instead of multiplication.

- Let $X_\diamond=[0,1.8]$, $c_\diamond=0.9$. The Taylor coefficients are

$$
\begin{aligned}
f(c_\diamond) &= -1.701 \\
f'(c_\diamond)/1! &= -2.97 \\
f''(c_\diamond)/2! &= -0.3 \\
f'''(c_\diamond)/3! &= 1,
\end{aligned}
$$

  and

$$T_f(X_\diamond)=-1.701+\big(([-0.9,0.9]-0.3)[-0.9,0.9]-2.97\big)[-0.9,0.9]=[-5.346,1.944].$$

Thus $X_\diamond \subset X$ but $T_f(X_\diamond) \supset T_f(X)$. □

Theorem 3.3.5 (Convergence) *T_f and T_f^* converge quadratically to f.* □

Proof. According to (3.3.3) the dense Taylor form can be written as

$$T_f^*(X) = f(c) + H^*_{g^{(c)}}(X-c)\,(X-c).$$

Let $G(X) = H^*_{g^{(c)}}(X-c)$. From Corollary 1.3.19 it follows that $H^*_{g^{(c)}}$ and hence G are Lipschitz. Further, $g^{(c)}(x-c) \in G(X)$ for all $x \in X$. Hence T_f^* is a centered form and quadratically convergent by Definition 1.3.25 and Theorem 1.3.27. As $T_f(X) \subseteq T_f^*(X)$ for all $X \in \mathbb{IR}$, T_f is also quadratically convergent. □

We give an algorithm for computing Taylor coefficients by the extended Horner scheme first.

Algorithm 3.3.6 (TC) [Taylor Coefficients]

In: $f(x) = a_n x^{d_n} + a_{n-1}x^{d_{n-1}} + \ldots + a_1 x^{d_1} \in \mathbb{F}[x]$,
$c \in \mathbb{F}$.

Out: $\mathrm{TC}(f,c) = (A_0, \ldots, A_{d_n})$, $A_i \in \mathbb{IF}$, $f^{(i)}(c)/i! \in A_i$, $i = 0, \ldots, d_n$.

(1) [Initialize.]
for $i = 0, \ldots, d_n$ do $A_i \longleftarrow$ coefficient of x^i in f.
$S \longleftarrow cA_{d_n}$.

(2) [Horner scheme.]
for $i = 0, \ldots, d_n - 1$
$A_{d_n-1} \longleftarrow A_{d_n-1} + S$.
for $j = d_n - 1, \ldots, i+1$ do $A_{j-1} \longleftarrow A_{j-1} + cA_j$.

(3) [Return.]
return $A_0, \ldots, A_{d_n}$.

Theorem 3.3.7 (Comlexity) *Algorithm 3.3.6 (*TC*) costs*

$d_n^2 - d_n + 2$ *number multiplications and*
$d_n^2 + d_n$ *number additions.* □

Proof.

- Step 1 costs 2 number multiplication.
- The i-th iteration in step 2 costs $2(d_n - i - 1)$ number multiplications and $2(d_n - i)$ number additions. Hence step 2 costs $d_n(d_n - 1)$ number multiplications and $d_n(d_n + 1)$ number additions.

Algorithm 3.3.8 evaluates the dense Taylor form. For the evaluation of the translated polynomial we use Theorem 3.3.2 instead of Horner form. This allows to replace interval operations by floating point operations.

Algorithm 3.3.8 (DTF) [Dense Taylor Form]

In: $f(x) = a_n x^{d_n} + a_{n-1}x^{d_{n-1}} + \ldots + a_1 x^{d_1} \in \mathbb{F}[x]$,
$X \in \mathbb{IF}$.

Out: $\mathrm{DTF}(f,X) \in \mathbb{IF}$, $\mathrm{DTF}(f,X) \supseteq T_f^*(X)$.

(1) [Special case $\overline{X} = \underline{X}$.]
if $\overline{X} = \underline{X}$ then return $H_f(X)$.

(2) [Midpoint and radius of X.]
$c \longleftarrow \mathrm{MID}(X)$.
$r \longleftarrow \max\{\overline{X} \mathbin{\hat{-}} c, c \mathbin{\hat{-}} \underline{X}\}$.

(3) [Taylor coefficients.]
$(A_0, \ldots, A_{d_n}) \longleftarrow \mathrm{TC}(f, c)$.

(4) [Magnitude of $H^*_{g^{(c)}}(X-c)(X-c)$.]
$m \longleftarrow \mathrm{mag}(A_{d_n}) \mathbin{\hat{*}} r$.
for $i = d_n - 1 \ldots 1$ do $m \longleftarrow (m \mathbin{\hat{+}} \mathrm{mag}(A_i)) \mathbin{\hat{*}} r$.

(5) [Return.]
return $A_0 + [-m, m]$.

The special case $\overline{X} = \underline{X}$ was excluded in Step 1, hence $r > 0$ in Step 4 and an invalid operation cannot occur.

Theorem 3.3.9 (Complexity) *If $\overline{X} \neq \underline{X}$ then Algorithm 3.3.8* (DTF) *costs*

$$\begin{array}{rl} d_n^2 + 3 & \textit{number multiplications and} \\ d_n^2 + 2d_n + 5 & \textit{number additions.} \ \square \end{array}$$

Proof.

- Step 2 costs 1 number multiplication and 4 number additions.
- Step 3 costs $d_n^2 - d_n + 2$ number multiplications and $d_n^2 + d_n$ number additions (Theorem 3.3.7).
- Step 4 costs d_n number multiplications and $d_n - 1$ number additions.
- Step 5 costs 2 number additions. $\square$

3.3.1 Bisection at Zero

The Taylor form reduces the computation of the range of f on an arbitrary interval X to the evaluation of the Horner form of the Taylor polynomial $f^{(c)}$ on a centered interval $X - c$. According to Theorem 3.1.45, the overestimation error of the Horner form can be reduced at least by half if $X - c$ is bisected at zero and Horner form is evaluated separately on both halves. In Section 3.1.3 it was pointed out that such a bisection is relatively inexpensive. The following new modification of the Taylor form is therefore slightly more expensive than T_f^*, but gives significantly tighter inclusions. In the sequel let

$$\check{H}_f^*(X) = \begin{cases} H_f^*([\underline{X}, 0]) \cup H_f([0, \overline{X}]) & \text{if } 0 \in \mathrm{int}(X) \\ H_f^*(X) & \text{else.} \end{cases}$$

Definition 3.3.10 (Dense Taylor Form with Bisection) *The dense Taylor form with bisection $\check{T}_f^* : \mathbb{IR} \to \mathbb{IR}$ of f is defined as*

$$\check{T}_f^*(X) = \check{H}^*_{f^{(\mathrm{mid}(X))}}(X - \mathrm{mid}(X)). \ \square$$

The following theorem states that the overestimation error of $\check{T}_f^*$ is at most half as big as the overestimation error of T_f^*.

Theorem 3.3.11 *For all $X \in \mathbb{IR}$ it holds that*

$$\begin{array}{rcl} |\overline{f(X)} - \overline{\check{T}_f^*(X)}| & \leq & 1/2|\overline{f(X)} - \overline{T_f^*(X)}| \\ |\underline{f(X)} - \underline{\check{T}_f^*(X)}| & \leq & 1/2|\underline{f(X)} - \underline{T_f^*(X)}|. \ \square \end{array}$$

Proof. Follows immediately from Theorem 3.1.45. □

Algorithm 3.3.12 (DTFBM) evaluates $\check{T}_f^*$. In order to improve efficiency, we apply the techniques developed in Algorithm 3.1.38 (HFBZ).

Algorithm 3.3.12 (DTFBM**)** [Dense Taylor Form with Bisection at Midpoint]

In: $f(x) = a_n x^{d_n} + a_{n-1} x^{d_{n-1}} + \ldots + a_1 x^{d_1} \in \mathbb{F}[x]$,
$X \in \mathbb{IF}$.

Out: $\mathrm{DTFBM}(f, X) \in \mathbb{IF}$, $\mathrm{DTFBM}(f, X) \approx \check{T}_f^*(X)$, $\mathrm{DTFBM}(f, X) \supseteq f(X)$.

(1) [Midpoint of X.]
$c \longleftarrow \mathrm{MID}(X)$.

(2) [Coefficients of $f^{(c)}(x)$.]
$(A_0, \ldots, A_{d_n}) \longleftarrow \mathrm{TC}(f, c)$.

(3) [Initialize floating point operations.]
clear invalid operation flag.

(4) [Right half.]
$x \longleftarrow \overline{X} \mathbin{\hat{-}} c$.
$R \longleftarrow A_{d_n}$.
for $i = d_n - 1, \ldots, 0$
if $\overline{R} > 0$ then $\overline{R} \longleftarrow \overline{R} \mathbin{\hat{*}} x \mathbin{\hat{+}} \overline{A}_i$, else $\overline{R} \longleftarrow \overline{A}_i$.
if $\underline{R} < 0$ then $\underline{R} \longleftarrow \underline{R} \mathbin{\check{*}} x \mathbin{\check{+}} \underline{A}_i$, else $\underline{R} \longleftarrow \underline{A}_i$.

(5) [Coefficients of $f^{(c)}(-x)$.]
for $i = 1, \ldots, d_n$ step 2 do $A_i \longleftarrow -A_i$.

(6) [Left half.]
$x \longleftarrow c \mathbin{\hat{-}} \underline{X}$.
$L \longleftarrow A_{d_n}$.
for $i = d_n - 1, \ldots, 0$
if $\overline{L} > 0$ then $\overline{L} \longleftarrow \overline{L} \mathbin{\hat{*}} x \mathbin{\hat{+}} \overline{A}_i$, else $\overline{L} \longleftarrow \overline{A}_i$.
if $\underline{L} < 0$ then $\underline{L} \longleftarrow \underline{L} \mathbin{\check{*}} x \mathbin{\check{+}} \underline{A}_i$, else $\underline{L} \longleftarrow \underline{A}_i$.

(7) [Check invalid operation and return.]
if invalid operation flag is raised, then return $[\bot, \top]$, else return $R \cup L$.

Theorem 3.3.13 (Complexity) *Algorithm* 3.3.12 *(*DTFBM*) costs*

$d_n^2 + 3d_n + 3$ *number multiplications and*
$d_n^2 + 5d_n + 4$ *number additions.* □

Proof.

- Step 1 costs 1 number multiplication and 2 number additions.
- Step 2 costs $d_n^2 - d_n + 2$ number multiplications and $d_n^2 + d_n$ number additions.
- Step 4 and 6 costs each $2d_n$ number multiplications and $2d_n + 1$ number additions. □

Hence, Algorithm 3.3.12 (DTFBM) costs $3d_n$ number multiplications and $3d_n - 1$ number additions more than Algorithm 3.3.8 (DTF).

3.3.2 Experimental Results

An experimental comparison between Algorithm 3.3.8 (DTF) and Algorithm 3.3.12 (DTFBM) for dense polynomials with different degrees is given in Table 3.3.1. The coefficients of the polynomials and the

Degree of f	2	6	10	14	18	22	26	30
Flops for $T_f^*(X)$	20.0	92.0	228.0	428.0	692.0	1020.0	1412.0	1868.0
Flops for $\check{T}_f^*(X)$	22.3	104.1	250.0	459.8	733.8	1071.7	1473.6	1939.1
$\mathrm{q}(T_f^*(X), f(X))/\mathrm{w}(X)^2$	0.23	0.70	1.06	1.31	1.58	1.90	2.10	2.22
$\mathrm{q}(\check{T}_f^*(X), f(X))/\mathrm{w}(X)^2$	0.11	0.26	0.36	0.43	0.52	0.60	0.66	0.69

Table 3.3.1: Comparison of T_f^* and $\check{T}_f^*$ for dense random polynomials f with different degrees and random intervals X. The coefficients of f and the endpoints of X are uniformly chosen in $[-1, 1]$.

endpoints of the intervals are uniformly distributed in $[-1, 1]$. For each degree d_n, the average cost and overestimation error of 10^4 random polynomials is reported. The cost is the total number of arithmetic floating point instructions, including those which were executed during interval operations. As the forms are quadratically convergent, the distance to the range divided by $\mathrm{w}(X)^2$ was chosen to measure accuracy.

While the measured costs of Algorithm 3.3.8 (DTF) correspond exactly to Theorem 3.3.9, the measured costs of Algorithm 3.3.12 (DTFBM) are significantly smaller than stated by Theorem 3.3.13. The reason is, that the worst case assumption of Theorem 3.3.13 does not hold in most cases.

3.4 Bernstein Form

It is well known that the $k+1$ Bernstein polynomials of order k form a basis of the vector space of polynomials of degree at most k. This means that every polynomial f with degree $\leq k$ can be written as a linear combination of k-th order Bernstein polynomials. The coefficients of such a linear combination have an important property, namely the largest and the smallest coefficient bound the range of f in the interval $[0, 1]$. This observation can be generalized to bound the range of f on arbitrary intervals X.

Experimental results indicate that the Bernstein form gives very tight inclusions compared to the other forms described so far. There are many situations when Bernstein form yields the range without overestimation. An a posteriori test, whether such a situation is given, does not require any additional computation.

The idea to use Bernstein polynomials for range computation goes back to [Cargo and Shisha, 1966] and [Rivlin, 1970]. The case $X \neq [0, 1]$ was considered first by [Rokne, 1977]. Efficient algorithms for real and for interval polynomials are presented in [Rokne, 1979a] respectively [Rokne, 1982]. Complex polynomials are studied in [Rokne, 1979b] and [Grassmann and Rokne, 1979]. The multivariate case is treated in [Garloff, 1985]. An application to real root isolation can be found in [Lane and Riesenfeld, 1981]. The inclusion monotonicity of the Bernstein form (Theorem 3.4.17) was proved by [Hong and Stahl, 1995]. A new criterion when the Bernstein form gives the range without overestimation is subject of Theorem 3.4.19. In Section 3.4.1 we give a new improvement of the Bernstein form which reduces both, the overestimation error and the computing time. The improvement is achieved through a bisection of the input interval at 0.

Let us first review some basic properties of Bernstein polynomials.

Definition 3.4.1 (Bernstein Polynomial) *The j-th Bernstein polynomial of order k is defined as*

$$p_j^{(k)}(x) = \binom{k}{j} x^j (1-x)^{k-j}, \quad k \geq 0, \quad j = 0, \ldots, k. \ \square$$

The Bernstein polynomials $\{p_j^{(k)}(x) \mid j = 0, \ldots, k\}$ form a basis of the vector space of polynomials of degree $\leq k$. All we have to show is that the elements x^i, $i = 0, \ldots, k$ of the power basis are linear combinations of Bernstein polynomials.

Lemma 3.4.2 *For all $0 \leq i \leq k$ it holds that*

$$x^i = \sum_{j=i}^{k} \frac{\binom{j}{i}}{\binom{k}{i}} p_j^{(k)}(x). \ \square$$

Proof.

$$\begin{aligned} x^i &= x^i(x + (1-x))^{k-i} \\ &= \sum_{r=0}^{k-i} \binom{k-i}{r} x^{i+r}(1-x)^{k-i-r} \\ &= \sum_{j=i}^{k} \binom{k-i}{j-i} x^j (1-x)^{k-j} \\ &= \sum_{j=i}^{k} \frac{\binom{j}{i}}{\binom{k}{i}} \binom{k}{j} x^j (1-x)^{k-j} \\ &= \sum_{j=i}^{k} \frac{\binom{j}{i}}{\binom{k}{i}} p_j^{(k)}(x). \ \square \end{aligned}$$

In the sequel let $k \geq d_n$ and let a_i^* be the coefficient of x^i in f, $i = 0, \ldots, k$. According to Lemma 3.4.2 there exist coefficients $b_j^{(k)}$ such that

$$f(x) = \sum_{j=0}^{k} b_j^{(k)} p_j^{(k)}(x).$$

Using Lemma 3.4.2 we can derive the coefficients as follows:

$$\begin{aligned} f(x) &= \sum_{i=0}^{d_n} a_i^* x^i \\ &= \sum_{i=0}^{d_n} a_i^* \sum_{j=i}^{k} \frac{\binom{j}{i}}{\binom{k}{i}} p_j^{(k)}(x) \\ &= \sum_{j=0}^{k} \sum_{i=0}^{j} a_i^* \frac{\binom{j}{i}}{\binom{k}{i}} p_j^{(k)}(x). \end{aligned}$$

This motivates the following definition.

Definition 3.4.3 (Bernstein Coefficient) *The j-th Bernstein coefficient of f of order k is defined as*

$$b_j^{(k)} = \sum_{i=0}^{j} a_i^* \frac{\binom{j}{i}}{\binom{k}{i}}, \quad j = 0, \ldots, k. \ \square$$

The following theorem shows the relationship between Bernstein coefficients and the range of polynomials over $[0, 1]$.

Theorem 3.4.4 (Range Overestimation)

$$f([0,1]) \subseteq [\{b_j^{(k)} \mid j = 0, \ldots, k\}]. \ \square \tag{3.4.1}$$

Proof. Note that

$$p_j^{(k)}(x) \geq 0 \ \text{ for all } \ x \in [0,1], \quad j = 0, \ldots, k.$$

Further,

$$\begin{aligned}\sum_{j=0}^{k} p_j^{(k)}(x) &= \sum_{j=0}^{k} \binom{k}{j} x^j (1-x)^{k-j} \\ &= (x+(1-x))^k \\ &\equiv 1.\end{aligned}$$

Hence, for all $x \in X$

$$\begin{aligned}f(x) &= \sum_{j=0}^{k} b_j^{(k)} p_j^{(k)}(x) \\ &\leq \max_j b_j^{(k)} \sum_{j=0}^{k} p_j^{(k)}(x) \\ &= \max_j b_j^{(k)} \\ f(x) &= \sum_{j=0}^{k} b_j^{(k)} p_j^{(k)}(x) \\ &\geq \min_j b_j^{(k)} \sum_{j=0}^{k} p_j^{(k)}(x) \\ &= \min_j b_j^{(k)}. \ \square\end{aligned}$$

The following theorem gives a necessary and sufficient condition when equality holds in (3.4.1).

Theorem 3.4.5 (Non–Overestimation)

$$\begin{aligned}\overline{f([0,1])} &= \max\{b_j^{(k)} \mid j=0,\ldots,k\} && (3.4.2)\\ \text{iff} \quad & \max\{b_j^{(k)} \mid j=0,\ldots,k\} \in \{b_0^{(k)}, b_k^{(k)}\} \\ \underline{f([0,1])} &= \min\{b_j^{(k)} \mid j=0,\ldots,k\} && (3.4.3)\\ \text{iff} \quad & \min\{b_j^{(k)} \mid j=0,\ldots,k\} \in \{b_0^{(k)}, b_k^{(k)}\}. \ \square\end{aligned}$$

Proof. We give a proof of (3.4.2), the proof of (3.4.3) is analogous.

"$\Leftarrow$" Assume $\max\{b_j^{(k)} \mid j=0,\ldots,k\} = \max\{b_0^{(k)}, b_k^{(k)}\}$. As

$$\begin{aligned}b_0^{(k)} &= \sum_{i=0}^{0} a_i^* \frac{\binom{0}{i}}{\binom{k}{i}} = a_0^* = f(0) \\ b_k^{(k)} &= \sum_{i=0}^{k} a_i^* \frac{\binom{k}{i}}{\binom{k}{i}} = \sum_{i=0}^{k} a_i^* = f(1),\end{aligned}$$

it holds that $\max\{b_j^{(k)} \mid j=0,\ldots,k\} \leq \overline{f([0,1])}$. Hence, by Theorem 3.4.4,

$$\max\{b_j^{(k)} \mid j=0,\ldots,k\} = \overline{f([0,1])}.$$

"$\Rightarrow$" Assume $\overline{f([0,1])} = \max\{b_j^{(k)} \mid j=0,\ldots,k\}$.

- If $b_0^{(k)} = b_1^{(k)} = \ldots = b_k^{(k)}$ then trivially $\max\{b_j^{(k)} \mid j=0,\ldots,k\} = \max\{b_0^{(k)}, b_k^{(k)}\}$.

– Otherwise, if $0 < x < 1$ then $0 < p_j^{(k)}(x) < 1$ for all j and

$$\begin{aligned} f(x) &= \sum_{j=0}^{k} b_j^{(k)} p_j^{(k)}(x) \\ &< \max_j b_j^{(k)} \sum_{j=0}^{k} p_j^{(k)}(x) \\ &= \max_j b_j^{(k)}. \end{aligned}$$

Hence, f achieves its maximum in $[0,1]$ either for $x = 0$ or for $x = 1$ and

$$\max\{b_j^{(k)} \mid j = 0, \ldots, k\} = \overline{f([0,1])} = \max\{f(0), f(1)\} = \max\{b_0^{(k)}, b_k^{(k)}\}. \ \square$$

So far we restricted our considerations to the case $X = [0,1]$. In order to extend Theorem 3.4.4 and Theorem 3.4.5 to arbitrary intervals, we introduce Bernstein basis polynomials $p_j^{(k,X)}$ which depend on X. The relevant properties which were needed for the case $X = [0,1]$ are supposed to carry over to the generalized case. Hence, we expect

$$\begin{aligned} p_j^{(k,X)}(x) &\geq 0 \quad \text{for all} \quad x \in X, \\ \sum_{j=0}^{k} p_j^{(k,X)}(x) &\equiv 1. \end{aligned}$$

In the sequel let $X \in \mathbb{IR}$, $\mathrm{w}(X) \neq 0$.

Definition 3.4.6 (Generalized Bernstein Polynomial) *The j-th Bernstein polynomial of order k over X is defined as*

$$p_j^{(k,X)}(x) = \binom{k}{j} \frac{(x - \underline{X})^j (\overline{X} - x)^{k-j}}{\mathrm{w}(X)^k}. \ \square$$

As in the case $X = [0,1]$, we show that $\{p_j^{(k,X)}(x) \mid j = 0, \ldots, k\}$ forms a basis of the vector space of polynomials of degree $\leq k$.

Lemma 3.4.7 *For every $k \geq i \geq 0$ it holds that*

$$\begin{aligned} (x - \underline{X})^i &= \mathrm{w}(X)^i \sum_{j=i}^{k} \frac{\binom{j}{i}}{\binom{k}{i}} p_j^{(k,X)}(x). \\ (x - \overline{X})^i &= -\mathrm{w}(X)^i \sum_{j=0}^{k-i} \frac{\binom{k-j}{i}}{\binom{k}{i}} p_j^{(k,X)}(x). \ \square \end{aligned}$$

Proof.

$$\begin{aligned} (x - \underline{X})^i &= \frac{1}{\mathrm{w}(X)^{k-i}} (x - \underline{X})^i (x - \underline{X} + \overline{X} - x)^{k-i} \\ &= \frac{1}{\mathrm{w}(X)^{k-i}} \sum_{r=0}^{k-i} \binom{k-i}{r} (x - \underline{X})^{i+r} (\overline{X} - x)^{k-i-r} \\ &= \frac{1}{\mathrm{w}(X)^{k-i}} \sum_{j=i}^{k} \binom{k-i}{j-i} (x - \underline{X})^j (\overline{X} - x)^{k-j} \\ &= \frac{1}{\mathrm{w}(X)^{k-i}} \sum_{j=i}^{k} \frac{\binom{j}{i}}{\binom{k}{i}} \binom{k}{j} (x - \underline{X})^j (\overline{X} - x)^{k-j} \end{aligned}$$

$$
\begin{aligned}
&= \mathrm{w}(X)^i \sum_{j=i}^{k} \frac{\binom{j}{i}}{\binom{k}{i}} p_j^{(k,X)}(x) \\
(x-\overline{X})^i &= \frac{(-1)^i}{\mathrm{w}(X)^{k-i}} (\overline{X}-x)^i (x-\underline{X}+\overline{X}-x)^{k-i} \\
&= \frac{(-1)^i}{\mathrm{w}(X)^{k-i}} \sum_{r=0}^{k-i} \binom{k-i}{r} (\overline{X}-x)^{i+r} (x-\underline{X})^{k-i-r} \\
&= \frac{(-1)^i}{\mathrm{w}(X)^{k-i}} \sum_{j=0}^{k-i} \binom{k-i}{k-j-i} (x-\underline{X})^j (\overline{X}-x)^{k-j} \\
&= \frac{(-1)^i}{\mathrm{w}(X)^{k-i}} \sum_{j=0}^{k-i} \frac{\binom{k-j}{i}}{\binom{k}{i}} \binom{k}{j} (x-\underline{X})^j (\overline{X}-x)^{k-j} \\
&= (-\mathrm{w}(X))^i \sum_{j=0}^{k-i} \frac{\binom{k-j}{i}}{\binom{k}{i}} p_j^{(k,X)}(x). \ \square
\end{aligned}
$$

Let us transform f from the power basis to the generalized Bernstein basis. For that purpose, it is useful to introduce Taylor coefficients first: In the sequel let

$$
\begin{aligned}
a_i^{(\underline{X})} &= f^{(i)}(\underline{X})/i! \\
a_i^{(\overline{X})} &= f^{(i)}(\overline{X})/i!
\end{aligned}
$$

be the i-th Taylor coefficient of f with developing point $\underline{X}$ respectively $\overline{X}$, i.e.

$$
\begin{aligned}
f(x) &= \sum_{i=0}^{d_n} a_i^{(\underline{X})} (x-\underline{X})^i \\
&= \sum_{i=0}^{d_n} a_i^{(\overline{X})} (x-\overline{X})^i.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
f(x) &= \sum_{i=0}^{d_n} a_i^{(\underline{X})} (x-\underline{X})^i \\
&= \sum_{i=0}^{d_n} a_i^{(\underline{X})} \sum_{j=i}^{k} \frac{\binom{j}{i}}{\binom{k}{i}} \mathrm{w}(X)^i p_j^{(k,X)}(x) \\
&= \sum_{j=0}^{k} \sum_{i=0}^{j} \frac{\binom{j}{i}}{\binom{k}{i}} a_i^{(\underline{X})} \mathrm{w}(X)^i p_j^{(k,X)}(x) \\
f(x) &= \sum_{i=0}^{d_n} a_i^{(\overline{X})} (x-\overline{X})^i \\
&= \sum_{i=0}^{d_n} a_i^{(\overline{X})} \sum_{j=0}^{k-i} \frac{\binom{k-j}{i}}{\binom{k}{i}} (-\mathrm{w}(X))^i p_j^{(k,X)}(x) \\
&= \sum_{j=0}^{k} \sum_{i=0}^{k-j} \frac{\binom{k-j}{i}}{\binom{k}{i}} a_i^{(\overline{X})} (-\mathrm{w}(X))^i p_j^{(k,X)}(x).
\end{aligned}
$$

This motivates the following definition.

Definition 3.4.8 (Generalized Bernstein Coefficient) *The j-th Bernstein coefficient of f of order*

k over X is defined as

$$b_j^{(k,X)} = \sum_{i=0}^{j} \frac{\binom{j}{i}}{\binom{k}{i}} a_i^{(\underline{X})} \mathrm{w}(X)^i \tag{3.4.4}$$

$$= \sum_{i=0}^{k-j} \frac{\binom{k-j}{i}}{\binom{k}{i}} a_i^{(\overline{X})} (-\mathrm{w}(X))^i \tag{3.4.5}$$

for $j = 0, \ldots, k$. □

Thus, Theorem 3.4.4 can be generalized as follows:

Theorem 3.4.9 (Range Overestimation)

$$f(X) \subseteq [\{b_j^{(k,X)} \mid j = 0, \ldots, k\}]. \ \square \tag{3.4.6}$$

Proof. Note that

$$p_j^{(k,X)}(x) \geq 0 \quad \text{for all} \quad x \in X, \quad j = 0, \ldots, k.$$

Further,

$$\begin{aligned} \sum_{j=0}^{k} p_j^{(k,X)}(x) &= \frac{1}{\mathrm{w}(X)^k} \sum_{j=1}^{k} \binom{k}{j} (x - \underline{X})^j \, (\overline{X} - x)^{k-j} \\ &= \frac{1}{\mathrm{w}(X)^k} (x - \underline{X} + \overline{X} - x)^k \\ &= 1. \end{aligned}$$

Hence, for all $x \in X$

$$\begin{aligned} f(x) &= \sum_{j=0}^{k} b_j^{(k,X)} p_j^{(k,X)}(x) \\ &\leq \max_j b_j^{(k,X)} \sum_{j=0}^{k} p_j^{(k,X)}(x) \\ &= \max_j b_j^{(k,X)} \\ f(x) &= \sum_{j=0}^{k} b_j^{(k,X)} p_j^{(k,X)}(x) \\ &\geq \min_j b_j^{(k,X)} \sum_{j=0}^{k} p_j^{(k,X)}(x) \\ &= \min_j b_j^{(k,X)}. \ \square \end{aligned}$$

The generalization of Theorem 3.4.5 is as expected:

Theorem 3.4.10 (Non–Overestimation)

$$\begin{aligned} \overline{f(X)} &= \max\{b_j^{(k,X)} \mid j = 0, \ldots, k\} \\ &\text{iff} \quad \max\{b_j^{(k,X)} \mid j = 0, \ldots, k\} \in \{b_0^{(k,X)}, b_k^{(k,X)}\} \qquad (3.4.7) \\ \underline{f(X)} &= \min\{b_j^{(k,X)} \mid j = 0, \ldots, k\} \\ &\text{iff} \quad \min\{b_j^{(k,X)} \mid j = 0, \ldots, k\} \in \{b_0^{(k,X)}, b_k^{(k,X)}\}. \ \square \qquad (3.4.8) \end{aligned}$$

Proof. We give a proof of (3.4.7), the proof of (3.4.8) is analogous.

"$\Leftarrow$" Assume $\max\{b_j^{(k,X)} \mid j = 0, \ldots, k\} = \max\{b_0^{(k,X)}, b_k^{(k,X)}\}$. As

$$b_0^{(k,X)} = \sum_{i=0}^{0} \frac{\binom{0}{i}}{\binom{0}{i}} a_i^{(\underline{X})} \mathrm{w}(X)^i = a_0^{(\underline{X})} = f(\underline{X}) \tag{3.4.9}$$

$$b_k^{(k,X)} = \sum_{i=0}^{k} \frac{\binom{k}{i}}{\binom{k}{i}} a_i^{(\underline{X})} \mathrm{w}(X)^i = \sum_{i=0}^{k} a_i^{(\underline{X})} \mathrm{w}(X)^i = f(\overline{X}), \tag{3.4.10}$$

it holds that $\max\{b_j^{(k,X)} \mid j = 0, \ldots, k\} \leq \overline{f(X)}$. Hence, by Theorem 3.4.9,

$$\max\{b_j^{(k,X)} \mid j = 0, \ldots, k\} = \overline{f(X)}.$$

"$\Rightarrow$" Assume $\overline{f(X)} = \max\{b_j^{(k,X)} \mid j = 0, \ldots, k\}$.

- If $b_0^{(k,X)} = b_1^{(k,X)} = \ldots = b_k^{(k,X)}$ then trivially $\max\{b_j^{(k,X)} \mid j = 0, \ldots, k\} = \max\{b_0^{(k,X)}, b_k^{(k,X)}\}$.
- Otherwise, if $\underline{X} < x < \overline{X}$ then $0 < p_j^{(k,X)}(x) < 1$ for all j and

$$\begin{aligned} f(x) &= \sum_{j=0}^{k} b_j^{(k,X)} p_j^{(k,X)}(x) \\ &< \max_j b_j^{(k,X)} \sum_{j=0}^{k} p_j^{(k,X)}(x) \\ &= \max_j b_j^{(k,X)}. \end{aligned}$$

Hence, f achieves its maximum in X either for $x = \underline{X}$ or for $x = \overline{X}$ and

$$\max\{b_j^{(k,X)} \mid j = 0, \ldots, k\} = \overline{f(X)} = \max\{f(\underline{X}), f(\overline{X})\} = \max\{b_0^{(k,X)}, b_k^{(k,X)}\}. \ \square$$

The range of f is bounded by the smallest and the largest Bernstein coefficient. This defines an interval extension of f, which is called Bernstein form.

Definition 3.4.11 (Bernstein Form) *The k-th order Bernstein form $B_f^{(k)} : \mathbb{IR} \to \mathbb{IR}$ of f is defined as*

$$B_f^{(k)}(X) = [\{b_j^{(k,X)} \mid j = 0, \ldots, k\}]. \ \square$$

Theorem 3.4.12 (Interval Extension) *$B_f^{(k)}$ is an interval extension of f.* $\square$

Proof. From Theorem 3.4.9 it follows that $f(X) \subseteq B_f^{(k)}(X)$ for all $X \in \mathbb{IR}$. If $X \in \mathbb{R}$ then $b_j^{(k,X)} = f(\underline{X}) = f(\overline{X})$ for all j, hence $B_f^{(k)}(X) \in \mathbb{R}$. $\square$

For the parameter k it was required so far merely that it is greater or equal the degree of f. By incrementing k, usually better approximations of the range are obtained.

Theorem 3.4.13 (Overestimation Error Bounds)

$$\mathrm{q}(f(X), B_f^{(k)}(X)) \leq \frac{1}{k} \sum_{i=2}^{d_n} (i-1)^2 |a_i^{(\underline{X})}| \mathrm{w}(X)^i. \tag{3.4.11}$$

$$\mathrm{q}(f(X), B_f^{(k)}(X)) \leq \frac{1}{k} \sum_{i=2}^{d_n} (i-1)^2 |a_i^{(\overline{X})}| \mathrm{w}(X)^i. \ \square \tag{3.4.12}$$

The proof of Theorem 3.4.13 requires some preparation and will be given after Lemma 3.4.15. As the proofs of (3.4.11) and (3.4.12) are analogous, we give only the proof of (3.4.11).

Lemma 3.4.14 *Let* ${}^ic_j^{(k,X)} \in \mathbb{R}$, $i = 0, \ldots, k$, $j = 0, \ldots, k$ *such that*

$$\sum_{j=0}^{k} (j/k \ \mathrm{w}(X))^i p_j^{(k,X)} - (x - \underline{X})^i = \sum_{j=0}^{k} {}^ic_j^{(k,X)} p_j^{(k,X)}.$$

Then

$$0 \le {}^ic_j^{(k,X)} \le \frac{(i-1)^2}{k} \mathrm{w}(X)^i. \ \square$$

Proof. As $p_j^{(k,X)}$, $j = 0, \ldots, k$ span the space of polynomials of degree $\le k$, the coefficients ${}^ic_j^{(k,X)}$ exist and are uniquely defined.

For $i = 0$ and $i = 1$ we obtain

$${}^0c_j^{(k,X)} = {}^1c_j^{(k,X)} = 0, \quad j = 0, \ldots, k.$$

For $i > 1$ it holds that

$$\begin{aligned}
&\sum_{j=0}^{k} (j/k \ \mathrm{w}(X))^i p_j^{(k,X)} - (x - \underline{X})^i \\
&= \sum_{j=0}^{k} (j/k \ \mathrm{w}(X))^i p_j^{(k,X)} - \sum_{j=i}^{k} \frac{\binom{j}{i}}{\binom{k}{i}} \mathrm{w}(X)^i p_j^{(k,X)} \\
&= \sum_{j=0}^{i-1} (j/k \ \mathrm{w}(X))^i p_j^{(k,X)} - \sum_{j=i}^{k} \left((j/k \ \mathrm{w}(X))^i - \frac{\binom{j}{i}}{\binom{k}{i}} \mathrm{w}(X)^i \right) p_j^{(k,X)} \\
&= \sum_{j=0}^{k} {}^ic_j^{(k,X)} p_j^{(k,X)}.
\end{aligned}$$

Comparing the coefficients of the Bernstein polynomials we get for $j < i$

$$0 \le {}^ic_j^{(k,X)} = (j/k \ \mathrm{w}(X))^i \le \frac{(i-1)^2}{k} \mathrm{w}(X)^i$$

and for $j \ge i$

$${}^ic_j^{(k,X)} = \left((j/k)^i - \frac{\binom{j}{i}}{\binom{k}{i}} \right) \mathrm{w}(X)^i.$$

It remains to show for $j \ge i > 1$

$$0 \le (j/k)^i - \frac{\binom{j}{i}}{\binom{k}{i}} \le \frac{(i-1)^2}{k}.$$

$$\begin{aligned}
(j/k)^i - \frac{\binom{j}{i}}{\binom{k}{i}} &= (j/k)^i - \frac{j!(k-i)!}{k!(j-i)!} \\
&= (j/k)^i - \frac{j(j-1)\cdots(j-i+1)}{k(k-1)\cdots(k-i+1)} \\
&= (j/k)^i \left(1 - \frac{\left(1 - \frac{1}{j}\right) \cdots \left(1 - \frac{i-1}{j}\right)}{\left(1 - \frac{1}{k}\right) \cdots \left(1 - \frac{i-1}{k}\right)} \right) \qquad (3.4.13) \\
&\le (j/k)^i \left(1 - \left(1 - \frac{1}{j}\right) \cdots \left(1 - \frac{i-1}{j}\right) \right) \\
&\le (j/k)^i \left(1 - \left(1 - \frac{i-1}{j}\right)^{i-1} \right).
\end{aligned}$$

Applying the mean value Theorem to $(1-x)^{i-1}$, we obtain

$$1-\left(1-\frac{i-1}{j}\right)^{i-1} \leq \frac{(i-1)^2}{j},$$

hence

$$(j/k)^i \frac{(i-1)^2}{j} = (j/k)^{i-1}\frac{(i-1)^2}{k} \leq \frac{(i-1)^2}{k}.$$

Finally,

$$0 \leq (j/k)^i - \frac{\binom{j}{i}}{\binom{k}{j}}$$

follows from comparison of the factors in the numerator and the denominator of (3.4.13). □

The following lemma gives a bound for the difference between a Bernstein coefficient and $f(x)$ for certain $x \in X$.

Lemma 3.4.15 *Let*

$$c_j^{(k,X)} = f(j/k\ \mathrm{w}(X) + \underline{X}) - b_j^{(k,X)} \quad j = 0,\ldots,k.$$

Then

$$|c_j^{(k,X)}| \leq \frac{1}{k}\sum_{i=2}^{d_n}(i-1)^2 a_i^{(\underline{X})}\mathrm{w}(X)^i. \ \square$$

Proof. With the definition of $c_j^{(k,X)}$ as in Lemma 3.4.15 it holds that

$$\begin{aligned}
&\sum_{j=0}^{k} f(j/k\ \mathrm{w}(X) + \underline{X})p_j^{(k,X)}(x) - f(x) \\
&\quad = \sum_{j=0}^{k}(c_j^{(k,X)} + b_j^{(k,X)})p_j^{(k,X)}(x) - \sum_{j=0}^{k} b_j^{(k,X)} p_j^{(k,X)} \\
&\quad = \sum_{j=0}^{k} c_j^{(k,X)} p_j^{(k,X)}(x). \qquad (3.4.14)
\end{aligned}$$

Using the coefficient ${}^i c_j^{(k,X)}$ of Lemma 3.4.14 we obtain

$$\begin{aligned}
&\sum_{j=0}^{k} f(j/k\ \mathrm{w}(X) + \underline{X})p_j^{(k,X)}(x) - f(x) \\
&\quad = \sum_{j=0}^{k}\sum_{i=0}^{d_n} a_i^*(j/k\ \mathrm{w}(X) + \underline{X})^i p_j^{(k,X)}(x) - \sum_{i=0}^{d_n} a_i^* x^i \\
&\quad = \sum_{j=0}^{k}\sum_{i=0}^{d_n} a_i^{(\underline{X})}(j/k\ \mathrm{w}(X))^i p_j^{(k,X)}(x) - \sum_{i=0}^{d_n} a_i^{(\underline{X})}(x-\underline{X})^i \\
&\quad = \sum_{i=0}^{d_n} a_i^{(\underline{X})}\sum_{j=0}^{k}(j/k\ \mathrm{w}(X))^i p_j^{(k,X)}(x) - (x-\underline{X})^i \\
&\quad = \sum_{i=0}^{d_n} a_i^{(\underline{X})}\sum_{j=0}^{k} {}^i c_j^{(k,X)} p_j^{(k,X)}(x) \\
&\quad = \sum_{j=0}^{k}\sum_{i=0}^{d_n} a_i^{(\underline{X})} {}^i c_j^{(k,X)} p_j^{(k,X)}(x). \qquad (3.4.15)
\end{aligned}$$

Comparing the coefficients of $p_j^{(k,X)}$ in (3.4.14) and (3.4.15) we obtain

$$c_j^{(k,X)} = \sum_{i=0}^{d_n} a_i^{(\underline{X})} {}^i c_j^{(k,X)}.$$

From Lemma 3.4.14 and ${}^0c_j^{(k,X)} = {}^1c_j^{(k,X)} = 0$ it follows that

$$\begin{aligned} |c_j^{(k,X)}| &\leq \sum_{i=0}^{d_n} |a_i^{(\underline{X})}| {}^i c_j^{(k,X)} \\ &\leq \frac{1}{k} \sum_{i=2}^{d_n} |a_i^{(\underline{X})}| (i-1)^2 \mathrm{w}(X)^i \\ &= \frac{1}{k} \sum_{i=2}^{d_n} |a_i^{(\underline{X})}| (i-1)^2 \mathrm{w}(X)^i. \ \square \end{aligned}$$

Proof of Theorem 3.4.13. Let $j_{\min}$, $j_{\max}$ be the index of a smallest respectively largest Bernstein coefficient of f in X. From Theorem 3.4.9 and Lemma 3.4.15 it follows that

$$\begin{aligned} \overline{B_f^{(k)}(X)} - \overline{f(X)} &\leq |f(j_{\max}/k \ \mathrm{w}(X) + \underline{X}) - b_{j_{\max}}^{(k,X)}| \\ &\leq \frac{1}{k} \sum_{i=2}^{d_n} (i-1)^2 |a_i^{(\underline{X})}| \mathrm{w}(X)^i \\ \underline{f(X)} - \underline{B_f^{(k)}(X)} &\leq |f(j_{\min}/k \ \mathrm{w}(X) + \underline{X}) - b_{j_{\min}}^{(k,X)}| \\ &\leq \frac{1}{k} \sum_{i=2}^{d_n} (i-1)^2 |a_i^{(\underline{X})}| \mathrm{w}(X)^i. \ \square \end{aligned}$$

Corollary 3.4.16 (Convergence) *$B_f^{(k)}$ converges quadratically to f.* $\square$

Proof. Let $A \in \mathbb{IR}$ arbitrary but fixed. Let

$$\hat{a_i}^{(A)} = \max\{|a_i^{(\underline{X})}| \mid X \in \mathbb{I}A\}.$$

The maximum exists because A is bounded and all derivatives of f are continuous. From Theorem 3.4.13 it follows that

$$\begin{aligned} \mathrm{q}(f(X), B_f^{(k)}(X)) &\leq \left(\frac{1}{k} \sum_{i=2}^{d_n} (i-1)^2 \hat{a_i}^{(A)} \mathrm{w}(X)^{i-2} \right) \mathrm{w}(X)^2 \\ &\leq \left(\frac{1}{k} \sum_{i=2}^{d_n} (i-1)^2 \hat{a_i}^{(A)} \mathrm{w}(A)^{i-2} \right) \mathrm{w}(X)^2 \\ &= \lambda_{B_f^{(k)},A} \mathrm{w}(X)^2, \end{aligned}$$

where

$$\lambda_{B_f^{(k)},A} = \frac{1}{k} \sum_{i=2}^{d_n} (i-1)^2 \hat{a_i}^{(A)} \mathrm{w}(A)^{i-2}. \ \square$$

The following theorem is taken from [Hong and Stahl, 1994a].

Theorem 3.4.17 (Inclusion Monotonicity) *$B_f^{(k)}$ is inclusion monotone.* $\square$

Before giving a formal proof of Theorem 3.4.17 we explain the key idea at an example. Consider the polynomial

$$f(x) = x^5 + 0.2x^4 + 0.5x^3 + 0.7x^2 - x,$$

and let $X, Y \in \mathbb{IR}$ such that $X \subseteq Y$. For simplicity, we first consider the special case where Y has one endpoint in common with X. Actually, this is already sufficient for the general case because the general case can be viewed as applying the special case twice on each endpoint. Precisely, consider the interval $[\underline{Y}, \overline{X}]$. This interval shares an endpoint with X and Y. Thus, if the special case is true, we have

$$B_f(X) \subseteq B_f([\underline{Y}, \overline{X}]) \subseteq B_f(Y).$$

In Figure 3.4.1 we show the Bernstein coefficients of order $k = 5$ of f over all X, where $\underline{X} = -1$ and $\overline{X}$ varies between -1 and 1. The curve for $b_j^{(5,X)}$ is labeled by j. By careful inspection of this picture one makes the following observations:

- The enveloping curves (bold lines) are monotone, i.e. $\max_j b_j^{(5,X)}$ increases and $\min_j b_j^{(5,X)}$ decreases as $\overline{X}$ grows. This precisely means inclusion monotonicity.
- Whenever the curve for some $b_j^{(5,X)}$ has a local extremum, it is intersected by the curve for $b_{j-1}^{(5,X)}$. Further, the curve for $b_0^{(5,X)}$ is constant. This observation is the key for the proof of inclusion monotonicity because, as one sees, it implies immediately monotonicity of the enveloping curves.

Next, we fix the right endpoint $\overline{X}$ to 1 and observe the Bernstein coefficients as $\underline{X}$ varies between -1 and 1 in Figure 3.4.2. Again, the enveloping curves are monotone, but this time, if $b_j^{(5,X)}$ has a local extremum, it is intersected by $b_{j+1}^{(5,X)}$ and $b_5^{(5,X)}$ is constant.

Lemma 3.4.18 makes the experimental observations more precise. From now on we view $b_j^{(k,X)}$ as a function in the variables $\underline{X}$, $\overline{X}$.

Lemma 3.4.18

$$b_j^{(k,X)} - b_{j-1}^{(k,X)} = \frac{\mathrm{w}(X)}{j} \frac{\partial b_j^{(k,X)}}{\partial \overline{X}}, \quad j = 1, \ldots, k \tag{3.4.16}$$

$$b_{j+1}^{(k,X)} - b_j^{(k,X)} = \frac{\mathrm{w}(X)}{k-j} \frac{\partial b_j^{(k,X)}}{\partial \underline{X}}, \quad j = 0, \ldots, k-1. \ \square \tag{3.4.17}$$

Proof.

- Proof of (3.4.16). Let $1 \le j \le k$ and note that

$$\begin{aligned} \binom{j}{i} - \binom{j-1}{i} &= \binom{j-1}{i-1} \\ i\binom{j}{i} &= j\binom{j-1}{i-1}. \end{aligned}$$

$$\begin{aligned} b_j^{(k,X)} - b_{j-1}^{(k,X)} &\overset{(3.4.4)}{=} \sum_{i=0}^{j} \frac{\binom{j}{i}}{\binom{k}{i}} a_i^{(\underline{X})} \mathrm{w}(X)^i - \sum_{i=0}^{j-1} \frac{\binom{j-1}{i}}{\binom{k}{i}} a_i^{(\underline{X})} \mathrm{w}(X)^i \\ &= \sum_{i=0}^{j-1} \frac{\binom{j}{i} - \binom{j-1}{i}}{\binom{k}{i}} a_i^{(\underline{X})} \mathrm{w}(X)^i + \frac{1}{\binom{k}{j}} a_j^{(\underline{X})} \mathrm{w}(X)^j \\ &= \sum_{i=1}^{j-1} \frac{\binom{j-1}{i-1}}{\binom{k}{i}} a_i^{(\underline{X})} \mathrm{w}(X)^i + \frac{1}{\binom{k}{j}} a_j^{(\underline{X})} \mathrm{w}(X)^j \\ &= \sum_{i=1}^{j} \frac{\binom{j-1}{i-1}}{\binom{k}{i}} a_i^{(\underline{X})} \mathrm{w}(X)^i. \end{aligned} \tag{3.4.18}$$

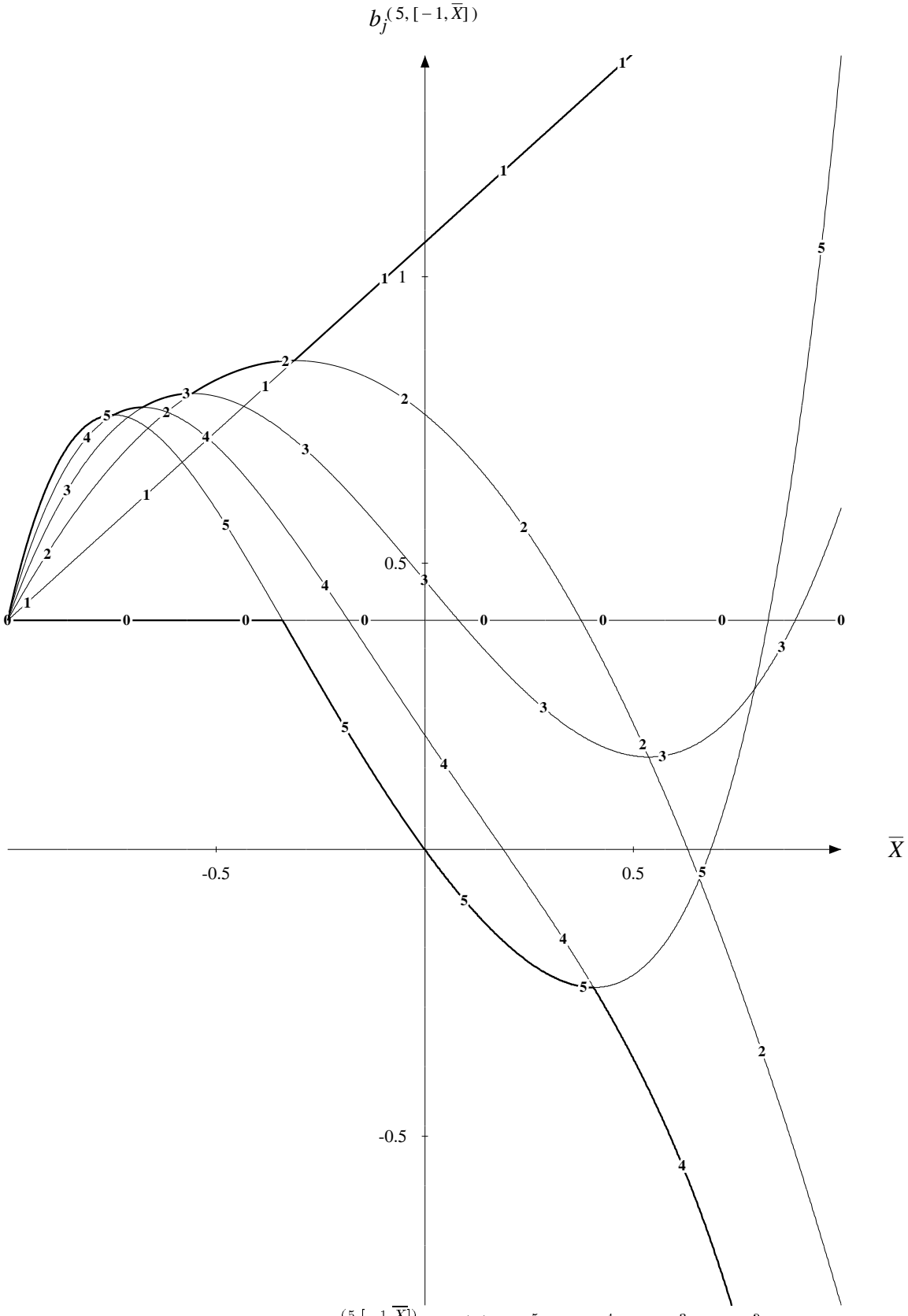


Figure 3.4.1: Bernstein coefficients $b_j^{(5,[-1,\overline{X}])}$ of $f(x) = x^5 + 0.2x^4 + 0.5x^3 + 0.7x^2 - x$ in dependence of $\overline{X}$.

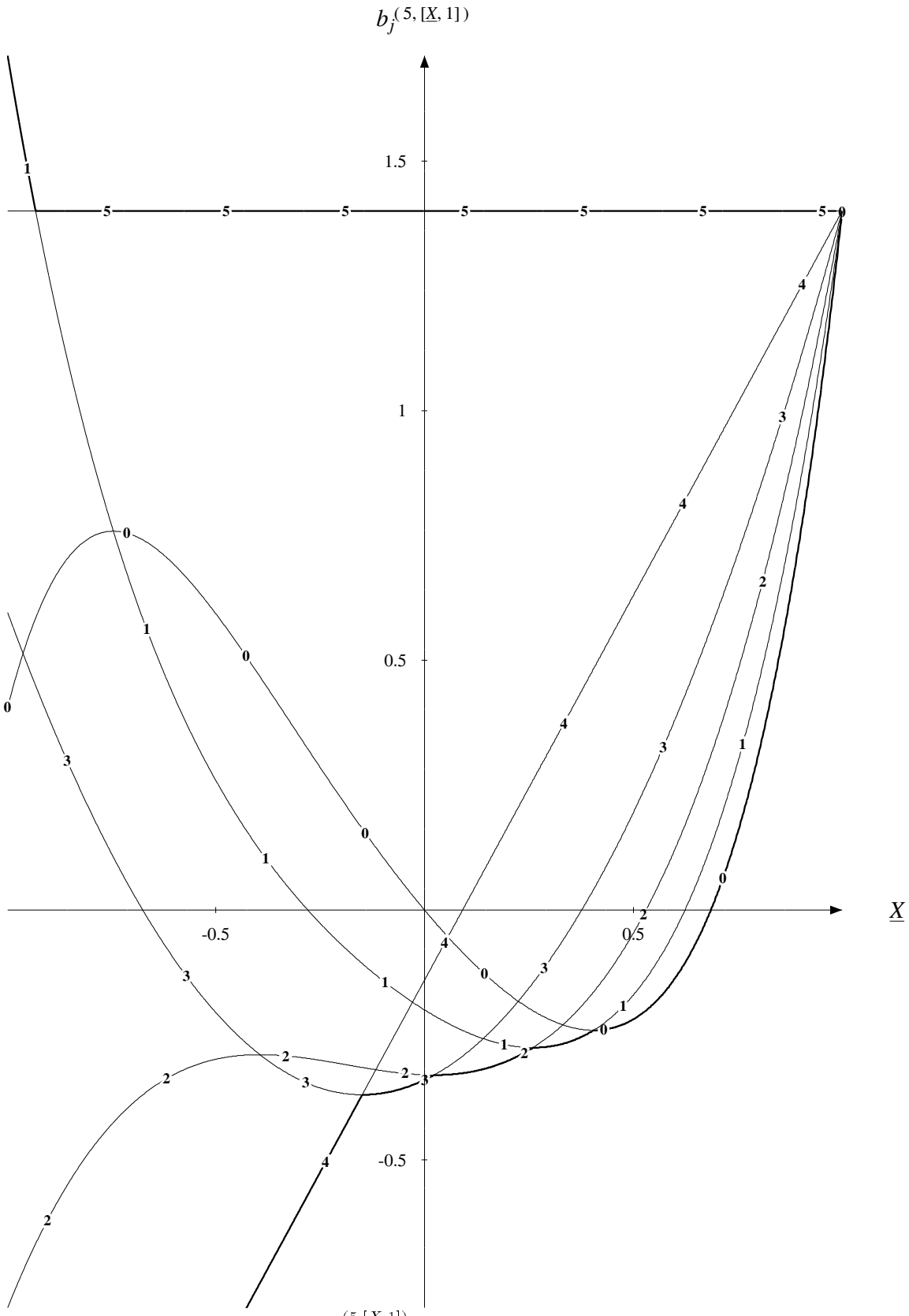


Figure 3.4.2: Bernstein coefficients $b_j^{(5,[\underline{X},1])}$, of $f(x) = x^5 + 0.2x^4 + 0.5x^3 + 0.7x^2 - x$ in dependence of $\underline{X}$.

$$\begin{aligned}\frac{\partial b_j^{(k,X)}}{\partial \overline{X}} &\overset{(3.4.4)}{=} \frac{\partial}{\partial \overline{X}} \sum_{i=0}^{j} \frac{\binom{j}{i}}{\binom{k}{i}} a_i^{(\underline{X})} (\overline{X} - \underline{X})^i \\ &= \sum_{i=1}^{j} \frac{\binom{j}{i}}{\binom{k}{i}} a_i^{(\underline{X})} i \mathrm{w}(X)^{i-1} \\ &= j \sum_{i=1}^{j} \frac{\binom{j-1}{i-1}}{\binom{k}{i}} a_i^{(\underline{X})} \mathrm{w}(X)^{i-1}.\end{aligned}$$

- Proof of (3.4.17). Let $0 \le j \le k-1$ and note that

$$\begin{aligned}\binom{k-j-1}{i} - \binom{k-j}{i} &= -\binom{k-j-1}{i-1} \\ i\binom{k-j}{i} &= (k-j)\binom{k-j-i}{i-1}.\end{aligned}$$

$$\begin{aligned}b_{j+1}^{(k,X)} - b_j^{(k,X)} &\overset{(3.4.5)}{=} \sum_{i=0}^{k-j-1} \frac{\binom{k-j-1}{i}}{\binom{k}{i}} a_i^{(\overline{X})} (-\mathrm{w}(X))^i - \sum_{i=0}^{k-j} \frac{\binom{k-j}{i}}{\binom{k}{i}} a_i^{(\overline{X})} (-\mathrm{w}(X))^i \\ &= -\sum_{i=1}^{k-j-1} \frac{\binom{k-j-1}{i} - \binom{k-j}{i}}{\binom{k}{i}} a_i^{(\overline{X})} (-\mathrm{w}(X))^i + \frac{1}{\binom{k}{k-j}} a_j^{(\overline{X})} (-\mathrm{w}(X))^j \\ &= -\sum_{i=1}^{k-j} \frac{\binom{k-j-1}{i-1}}{\binom{k}{i}} a_i^{(\overline{X})} (-\mathrm{w}(X))^i. \qquad (3.4.19)\end{aligned}$$

$$\begin{aligned}\frac{\partial b_j^{(k,X)}}{\partial \underline{X}} &\overset{(3.4.5)}{=} \frac{\partial}{\partial \underline{X}} \sum_{i=0}^{k-j} \frac{\binom{k-j}{i}}{\binom{k}{i}} a_i^{(\overline{X})} (\underline{X} - \overline{X})^i \\ &= \sum_{i=1}^{k-j} \frac{\binom{k-j}{i}}{\binom{k}{i}} a_i^{(\overline{X})} i (-\mathrm{w}(X))^{i-1} \\ &= (k-j) \sum_{i=1}^{k-j} \frac{\binom{k-j-1}{i-1}}{\binom{k}{i}} a_i^{(\overline{X})} (-\mathrm{w}(X))^{i-1}. \ \square\end{aligned}$$

Finally, we come to the proof of Theorem 3.4.17.

Proof of Theorem 3.4.17 According to the remarks above, we have to show the following:

$$\min_j b_j^{(k,X)} \quad \text{is monotonically decreasing in } \overline{X} \qquad (3.4.20)$$

$$\max_j b_j^{(k,X)} \quad \text{is monotonically increasing in } \overline{X} \qquad (3.4.21)$$

$$\min_j b_j^{(k,X)} \quad \text{is monotonically increasing in } \underline{X} \qquad (3.4.22)$$

$$\max_j b_j^{(k,X)} \quad \text{is monotonically decreasing in } \underline{X}. \qquad (3.4.23)$$

Note that for all $X \in \mathbb{IR}$

$$\begin{aligned}b_0^{(k,X)} &= f(\underline{X}) \\ b_k^{(k,X)} &= f(\overline{X}).\end{aligned}$$

The proofs of (3.4.20) – (3.4.23) are similar, but for the sake of completeness we give them all in detail. Let $X \in \mathbb{IR}$ arbitrary but fixed.

- Proof of (3.4.20). Let $j_{\min}$ such that

$$b_{j_{\min}}^{(k,X)} \le b_j^{(k,X)} \quad \text{for all} \quad j. \qquad (3.4.24)$$

We have to show that

$$\frac{\partial b_{j_{\min}}^{(k,X)}}{\partial \overline{X}} \leq 0.$$

As $b_0^{(k,X)}$ does not depend on $\overline{X}$, this is obviously true if $j_{\min} = 0$. Hence, assume $j_{\min} \neq 0$ and

$$\frac{\partial b_{j_{\min}}^{(k,X)}}{\partial \overline{X}} > 0.$$

From 3.4.16 we obtain

$$b_{j_{\min}}^{(k,X)} - b_{j_{\min}-1}^{(k,X)} = \frac{(\overline{X} - \underline{X})}{j_{\min}} \frac{\partial b_{j_{\min}}^{(k,X)}}{\partial \overline{X}} > 0,$$

which contradicts (3.4.24).

- Proof of (3.4.21). Let $j_{\max}$ such that

$$b_{j_{\max}}^{(k,X)} \geq b_j^{(k,X)} \quad \text{for all} \quad j. \tag{3.4.25}$$

We have to show that

$$\frac{\partial b_{j_{\max}}^{(k,X)}}{\partial \overline{X}} \geq 0.$$

As $b_0^{(k,X)}$ does not depend on $\overline{X}$, this is obviously true if $j_{\max} = 0$. Hence, assume $j_{\max} \neq 0$ and

$$\frac{\partial b_{j_{\max}}^{(k,X)}}{\partial \overline{X}} < 0.$$

From 3.4.16 we obtain

$$b_{j_{\max}}^{(k,X)} - b_{j_{\max}-1}^{(k,X)} = \frac{(\overline{X} - \underline{X})}{j_{\max}} \frac{\partial b_{j_{\max}}^{(k,X)}}{\partial \overline{X}} < 0,$$

which contradicts (3.4.25).

- Proof of (3.4.22). Let $j_{\min}$ such that

$$b_{j_{\min}}^{(k,X)} \leq b_j^{(k,X)} \quad \text{for all} \quad j. \tag{3.4.26}$$

We have to show that

$$\frac{\partial b_{j_{\min}}^{(k,X)}}{\partial \underline{X}} \geq 0.$$

As $b_k^{(k,X)}$ does not depend on $\underline{X}$, this is obviously true if $j_{\min} = k$. Hence, assume $j_{\min} \neq k$ and

$$\frac{\partial b_{j_{\min}}^{(k,X)}}{\partial \underline{X}} < 0.$$

From 3.4.17 we obtain

$$b_{j_{\min}+1}^{(k,X)} - b_{j_{\min}}^{(k,X)} = \frac{(\overline{X} - \underline{X})}{k - j_{\min}} \frac{\partial b_{j_{\min}}^{(k,X)}}{\partial \underline{X}} < 0,$$

which contradicts (3.4.26).

- Proof of (3.4.23). Let $j_{\max}$ such that

$$b_{j_{\max}}^{(k,X)} \geq b_j^{(k,X)} \quad \text{for all} \quad j. \tag{3.4.27}$$

We have to show that

$$\frac{\partial b_{j_{\max}}^{(k,X)}}{\partial \underline{X}} \leq 0.$$

As $b_k^{(k,X)}$ does not depend on $\underline{X}$, this is obviously true if $j_{\max} = k$. Hence, assume $j_{\max} \neq k$ and

$$\frac{\partial b_{j_{\max}}^{(k,X)}}{\partial \underline{X}} > 0.$$

From 3.4.17 we obtain

$$b_{j_{\max}+1}^{(k,X)} - b_{j_{\max}}^{(k,X)} = \frac{(\overline{X} - \underline{X})}{k - j_{\max}} \frac{\partial b_{j_{\max}}^{(k,X)}}{\partial \underline{X}} > 0,$$

which contradicts (3.4.27). □

Figure 3.4.1 reveals another property of the Bernstein coefficients. Let $\xi \in X$. If $\overline{X} \approx -1$, then

$$b_0^{(k,X)} \leq b_j^{(k,X)} \leq b_k^{(k,X)} \quad \text{for all} \quad j.$$

Similarly, Figure 3.4.2 shows that if $\underline{X} \approx 1$, then

$$b_0^{(k,X)} \leq b_j^{(k,X)} \leq b_k^{(k,X)} \quad \text{for all} \quad j.$$

According to Theorem 3.4.10, this gives rise to the conjecture that $B_f^{(k)}(X) = f(X)$ if $\mathrm{w}(X)$ is small enough. Yet, this conjecture is true only if we make the notion "small enough" more precise. This is done in the following Theorem, which is new.

Theorem 3.4.19

(i) *For all $\underline{X} \in \mathbb{R}$ there exists $\varepsilon > 0$ such that for all $\overline{X} \in [\underline{X}, \underline{X} + \varepsilon]$ it holds that*

$$B_f^{(k)}(X) = f(X).$$

(ii) *For all $\overline{X} \in \mathbb{R}$ there exists $\varepsilon > 0$ such that for all $\underline{X} \in [\overline{X} - \varepsilon, \overline{X}]$ it holds that*

$$B_f^{(k)}(X) = f(X). \ \square$$

Proof. The proof of (i) and (ii) is analogous but for the sake of completeness we give both of them.

(i) Let $\underline{X} \in \mathbb{R}$ arbitrary but fixed. According to Theorem 3.4.10 it suffices to show that there exists $\varepsilon > 0$ such that for all $\overline{X} \in [\underline{X}, \underline{X} + \varepsilon]$

$$b_j^{(k,X)} \geq b_{j-1}^{(k,X)} \quad \text{for all} \quad j = 1, \ldots, k \quad \text{or} \tag{3.4.28}$$

$$b_j^{(k,X)} \leq b_{j-1}^{(k,X)} \quad \text{for all} \quad j = 1, \ldots, k. \tag{3.4.29}$$

From (3.4.18) it follows that for all $\overline{X} \geq \underline{X}$

$$b_j^{(k,X)} - b_{j-1}^{(k,X)} = \sum_{i=1}^{j} \frac{\binom{j-1}{i-1}}{\binom{k}{i}} a_i^{(\underline{X})} \mathrm{w}(X)^i.$$

Let $1 \leq l \leq k$ be the smallest index such that $a_l^{(\underline{X})} \neq 0$. If no such l exists, then let $l = k+1$. Thus,

$$\begin{aligned} b_j^{(k,X)} - b_{j-1}^{(k,X)} &= \sum_{i=l}^{j} \frac{\binom{j-1}{i-1}}{\binom{k}{i}} a_i^{(\underline{X})} \mathrm{w}(X)^i \\ &= \mathrm{w}(X)^l \sum_{i=0}^{j-l} \frac{\binom{j-1}{i+l-1}}{\binom{k}{i+l}} a_{i+l}^{(\underline{X})} \mathrm{w}(X)^i. \end{aligned}$$

Depending on the sign of $a_l^{(\underline{X})}$ we show that either (3.4.28) or (3.4.29) holds.

– Assume $a_l^{(\underline{X})} > 0$. Then for all $j = 1, \ldots, k$ there exists $\varepsilon_j > 0$ such that for all $\mathrm{w}(X) \leq \varepsilon_j$

$$\sum_{i=0}^{j-l} \frac{\binom{j-1}{i+l-1}}{\binom{k}{i+l}} a_{i+l}^{(\underline{X})} \mathrm{w}(X)^i \geq 0.$$

Let $\varepsilon = \min_j \varepsilon_j$. Then for all $\overline{X} \in [\underline{X}, \underline{X} + \varepsilon]$

$$b_j^{(k,X)} - b_{j-1}^{(k,X)} \geq 0.$$

– Assume $a_l^{(\underline{X})} < 0$. Then for all $j = 1, \ldots, k$ there exists $\varepsilon_j > 0$ such that for all $\mathrm{w}(X) \leq \varepsilon_j$

$$\sum_{i=0}^{j-l} \frac{\binom{j-1}{i+l-1}}{\binom{k}{i+l}} a_{i+l}^{(\underline{X})} \mathrm{w}(X)^i \leq 0.$$

Let $\varepsilon = \min_j \varepsilon_j$. Then for all $\overline{X} \in [\underline{X}, \underline{X} + \varepsilon]$

$$b_j^{(k,X)} - b_{j-1}^{(k,X)} \leq 0.$$

(ii) Let $\overline{X} \in \mathbb{R}$ arbitrary but fixed. According to Theorem 3.4.10 it suffices to show that there exists $\varepsilon > 0$ such that for all $\underline{X} \in [\overline{X} - \varepsilon, \overline{X}]$

$$b_{j+1}^{(k,X)} \geq b_j^{(k,X)} \text{ for all } j = 0, \ldots, k-1 \quad \text{or} \tag{3.4.30}$$

$$b_{j+1}^{(k,X)} \leq b_j^{(k,X)} \text{ for all } j = 0, \ldots, k-1. \tag{3.4.31}$$

From (3.4.19) it follows that for all $\underline{X} \leq \overline{X}$

$$b_{j+1}^{(k,X)} - b_j^{(k,X)} = -\sum_{i=1}^{k-j} \frac{\binom{k-j-1}{i-1}}{\binom{k}{i}} a_i^{(\overline{X})} (-\mathrm{w}(X))^i.$$

Let $0 \leq l \leq k-1$ be the smallest index such that $a_l^{(\overline{X})} \neq 0$. If no such l exists, then let $l = k+1$. Thus,

$$\begin{aligned} b_{j+1}^{(k,X)} - b_j^{(k,X)} &= -\sum_{i=l}^{k-j} \frac{\binom{k-j-1}{i-1}}{\binom{k}{i}} a_i^{(\overline{X})} (-\mathrm{w}(X))^i \\ &= -\mathrm{w}(X)^l \sum_{i=0}^{k-j-l} \frac{\binom{k-j-1}{i+l-1}}{\binom{k}{i+l}} (-1)^l a_{i+l}^{(\overline{X})} (-\mathrm{w}(X))^i. \end{aligned}$$

Depending on the sign of $(-1)^l a_l^{(\overline{X})}$ we show that either (3.4.30) or (3.4.31) holds.

– Assume $(-1)^l a_l^{(\overline{X})} > 0$. Then for all $j = 0, \ldots, k-1$ there exists $\varepsilon_j > 0$ such that for all $\mathrm{w}(X) \leq \varepsilon_j$

$$\sum_{i=0}^{k-j-l} \frac{\binom{k-j-1}{i+l-1}}{\binom{k}{i+l}} (-1)^l a_{i+l}^{(\overline{X})} (-\mathrm{w}(X))^i \geq 0.$$

Let $\varepsilon = \min_j \varepsilon_j$. Then for all $\underline{X} \in [\overline{X} - \varepsilon, \overline{X}$

$$b_{j+1}^{(k,X)} - b_j^{(k,X)} \leq 0.$$

– Assume $(-1)^l a_l^{(\overline{X})} < 0$. Then for all $j = 0, \ldots, k-1$ there exists $\varepsilon_j > 0$ such that for all $\mathrm{w}(X) \leq \varepsilon_j$

$$\sum_{i=0}^{k-j-l} \frac{\binom{k-j-1}{i+l-1}}{\binom{k}{i+l}} (-1)^l a_{i+l}^{(\overline{X})} (-\mathrm{w}(X))^i \leq 0.$$

Let $\varepsilon = \min_j \varepsilon_j$. Then for all $\underline{X} \in [\overline{X} - \varepsilon, \overline{X}]$

$$b_{j+1}^{(k,X)} - b_j^{(k,X)} \geq 0.$$

Remark. Note that it is *not* true that for all $c \in \mathbb{R}$ there exists $\varepsilon > 0$ such that for all $0 \le r \le \varepsilon$

$$B_f^{(k)}([c-r, c+r]) = f([c-r, c+r]).$$

As a counter example consider $f(x) = x^2$ and $c = 0$. The Bernstein coefficients in dependence of r are

$$\begin{aligned} b_0^{(2,[-r,r])} &= r^2 \\ b_1^{(2,[-r,r])} &= -r^2 \\ b_2^{(2,[-r,r])} &= r^2, \end{aligned}$$

Hence, $b_1^{(2,[-r,r])}$ is the smallest Bernstein coefficient, and according to Theorem 3.4.10,

$$B_f^{(2)}([-r, r]) \neq f([-r, r])$$

for all $r \neq 0$. □

Remark. If $|f'(\underline{X})|$ is sufficiently large compared to $|f^{(i)}(\underline{X})|$ for all $i > 1$ and $\mathrm{w}(X)$ is sufficiently small, then the sign of $b_j^{(k,X)} - b_{j-1}^{(k,X)}$ is the same for all j. Similarly, if $|f'(\overline{X})|$ is sufficiently large compared to $|f^{(i)}(\overline{X})|$ for all $i > 1$ and $\mathrm{w}(X)$ is sufficiently small, then the sign of $b_{j+1}^{(k,X)} - b_j^{(k,X)}$ is the same for all j. In both cases we obtain $B_f^{(k)}(X) = f(X)$. This qualitative result is illustrated in Figure 3.4.3, where the overestimation error of the Bernstein form is computed for $f(x) = (x-1)(x-2)(x-3)(x-4)$, $\mathrm{w}(X) \equiv 0.5$ and $k = 4$, 8 and 16. One sees that the Bernstein form is exact if X is not close to a local extremum of f. However, it is *not* true, that monotonicity of f in X is a sufficient condition for non–overestimation, although this is claimed sometimes in the literature. For example let $f(x) = -x^3 + 0.7x^2 - 0.2x + 0.9$, $k = 3$ and $X = [0, 1]$. Note that f is monotone in X because $f'(x) < 0$ for all $x \in \mathbb{R}$. The Bernstein coefficients are

$$\begin{aligned} b_0^{(k,X)} &= 0.9, \\ b_1^{(k,X)} &= 0.8333, \\ b_2^{(k,X)} &= 1.0, \\ b_3^{(k,X)} &= 0.4. \end{aligned}$$

Thus, we obtain $B_f^{(k)}(X) = [0.4, 1.0]$, but $f(X) = [0.4, 0.9]$. □

Algorithm 3.4.23 (BF) for evaluating the Bernstein form proceeds in 3 steps: Compute the Taylor coefficients $a_i^{(\underline{X})}$, compute the Bernstein coefficients $b_j^{(k,X)}$ according to (3.4.4), find the largest and the smallest Bernstein coefficient. Equivalently, one could compute the Taylor coefficients $a_i^{(\overline{X})}$ and the Bernstein coefficients according to (3.4.5). The Taylor coefficients are obtained by the extended Horner scheme. An efficient method for computing the Bernstein coefficients is given by the following lemma [Rokne, 1979a].

Lemma 3.4.20 *Let* ${}^{l}b_j^{(k,X)}$, $l = 0, \ldots, d_n$, $j = 0, \ldots, k-l$ *be defined recursively as*

$$ {}^{l}b_0^{(k,X)} = \frac{\mathrm{w}(X)^l}{\binom{k}{l}} a_l^{(\underline{X})} \quad \text{for } l = 0, \ldots, d_n \tag{3.4.32} $$

$$ {}^{d_n}b_j^{(k,X)} = {}^{d_n}b_0^{(k,X)} \quad \text{for } j = 1, \ldots, k - d_n \tag{3.4.33} $$

$$ {}^{l}b_j^{(k,X)} = {}^{l}b_{j-1}^{(k,X)} + {}^{l+1}b_{j-1}^{(k,X)} \quad \text{for } j = 1, \ldots, k, \quad l = 0, \ldots, \min\{d_n - 1, k - j\}. \tag{3.4.34} $$

Then

$$b_j^{(k,X)} = {}^{0}b_j^{(k,X)} \text{ for } j = 0, \ldots, k. \ \square$$

Figure 3.4.4 illustrates the computation of the Bernstein coefficients according to Lemma 3.4.20. The left column is computed by (3.4.32), the upper row by (3.4.33) and the remaining points by (3.4.34). The Bernstein coefficients appear in the last row.

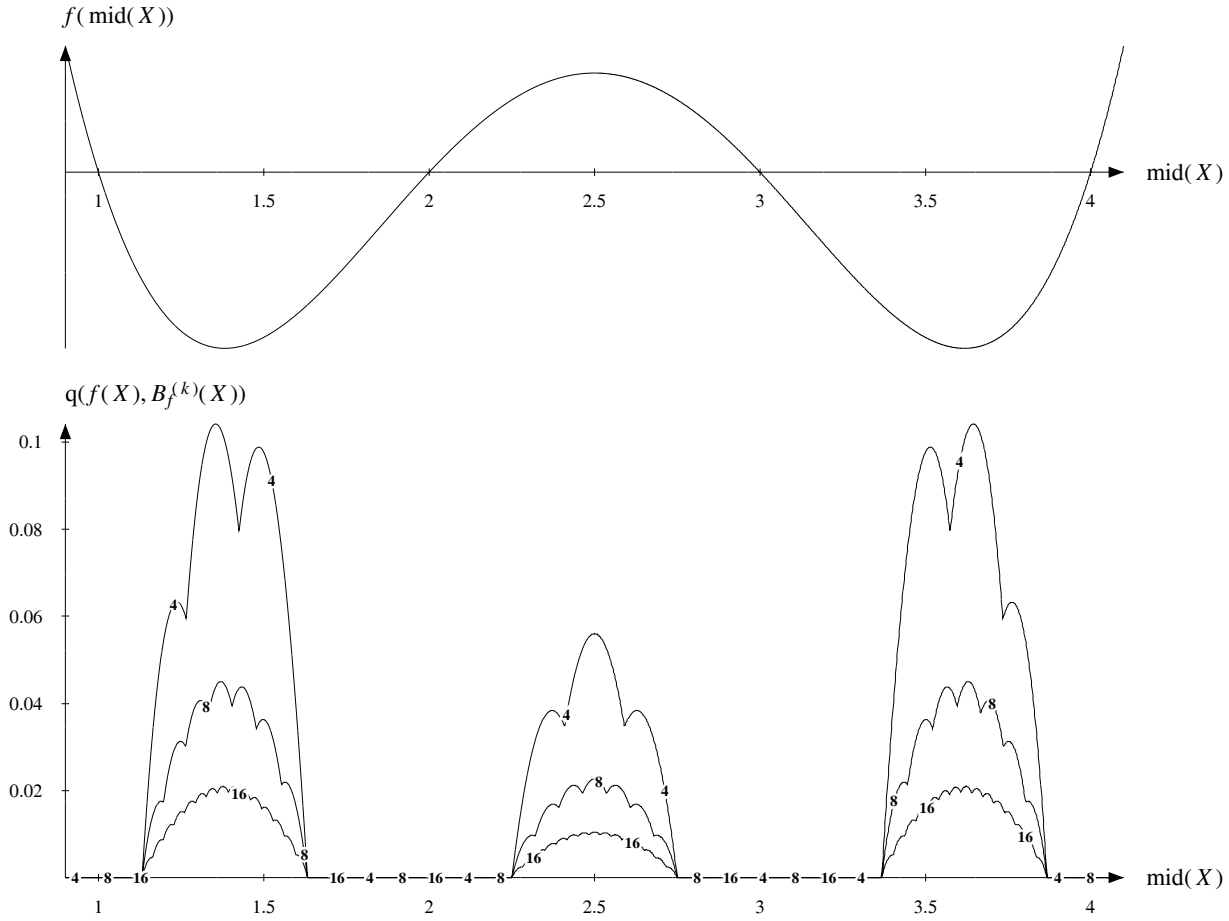


Figure 3.4.3: Overestimation error of the Bernstein form for $f(x) = (x-1)(x-2)(x-3)(x-4)$, in dependence of $\mathrm{mid}(X)$, where $\mathrm{w}(X) \equiv 0.5$ and $k = 4, 8, 16$.

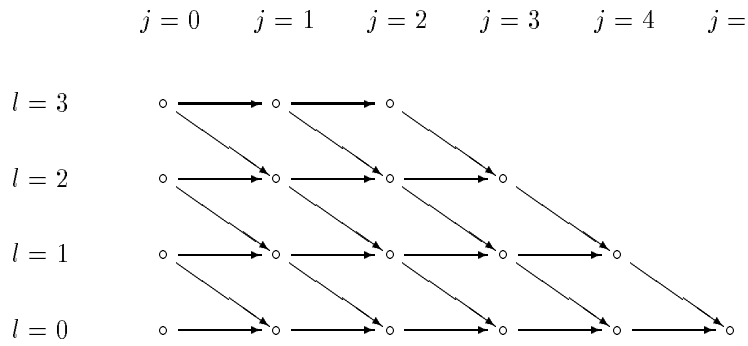


Figure 3.4.4: Computation of the Bernstein coefficients for $d_n = 3$ and $k = 5$.

Proof. Inductively we show for $l = 0, \ldots, d_n$, $j = 0, \ldots, k-l$ that

$$ {}^{l}b_j^{(k,X)} = \sum_{i=0}^{j} \frac{\binom{j}{i}}{\binom{k}{i+l}} a_{i+l}^{(\underline{X})} \mathrm{w}(X)^{i+l}. \tag{3.4.35} $$

Obviously (3.4.35) holds for $j = 0, l = 0, \ldots, d_n$ and for $l = d_n$, $j = 1, \ldots, k-d_n$. Further, for $j = 1, \ldots, k$, $l = 0, \ldots, \min\{d_n - 1, k - j\}$

$$ \begin{aligned} {}^{l}b_j^{(k,X)} &= {}^{l}b_{j-1}^{(k,X)} + {}^{l+1}b_{j-1}^{(k,X)} \\ &= \sum_{i=0}^{j-1} \frac{\binom{j-1}{i}}{\binom{k}{i+l}} a_{i+l}^{(\underline{X})} \mathrm{w}(X)^{i+l} + \sum_{i=0}^{j-1} \frac{\binom{j-1}{i}}{\binom{k}{i+l+1}} a_{i+l+1}^{(\underline{X})} \mathrm{w}(X)^{i+l+1} \\ &= \sum_{i=0}^{j-1} \frac{\binom{j-1}{i}}{\binom{k}{i+l}} a_{i+l}^{(\underline{X})} \mathrm{w}(X)^{i+l} + \sum_{i=1}^{j} \frac{\binom{j-1}{i-1}}{\binom{k}{i+l}} a_{i+l}^{(\underline{X})} \mathrm{w}(X)^{i+l} \\ &= \sum_{i=1}^{j-1} \frac{\binom{j-1}{i} + \binom{j-1}{i-1}}{\binom{k}{i+l}} a_{i+l}^{(\underline{X})} \mathrm{w}(X)^{i+l} + \frac{1}{\binom{k}{l}} a_l^{(\underline{X})} \mathrm{w}(X)^{l} + \frac{1}{\binom{k}{j+l}} a_{j+l}^{(\underline{X})} \mathrm{w}(X)^{j+l} \\ &= \sum_{i=1}^{j-1} \frac{\binom{j}{i}}{\binom{k}{i+l}} a_{i+l}^{(\underline{X})} \mathrm{w}(X)^{i+l} + \frac{1}{\binom{k}{l}} a_l^{(\underline{X})} \mathrm{w}(X)^{l} + \frac{1}{\binom{k}{j+l}} a_{j+l}^{(\underline{X})} \mathrm{w}(X)^{j+l} \\ &= \sum_{i=0}^{j} \frac{\binom{j}{i}}{\binom{k}{i+l}} a_{i+l}^{(\underline{X})} \mathrm{w}(X)^{i+l}. \end{aligned} $$

In particular, if $l = 0$ we obtain from (3.4.35)

$$ {}^{0}b_j^{(k,X)} = \sum_{i=0}^{j} \frac{\binom{j}{i}}{\binom{k}{i}} a_i^{(\underline{X})} \mathrm{w}(X)^{i} = b_j^{(k,X)}. \ \square $$

Algorithm 3.4.21 (BCB) takes the Taylor coefficients as input and returns the smallest and the largest Bernstein coefficient.

Algorithm 3.4.21 (BCB**)** [Bernstein Coefficient Bounds]

In: $X \in \mathbb{IF}$,
$A_0, \ldots, A_{d_n} \in \mathbb{IF}$, $f^{(i)}(\underline{X})/i! \in A_i$, $i = 0, \ldots, d_n$,
$k \in \mathbb{N}$, $k \geq d_n$.

Out: $\mathrm{BCB}(X, A_0, \ldots, A_n, k) \in \mathbb{IF}$,
$\underline{\mathrm{BCB}(X, A_0, \ldots, A_n, k)} \leq b_j^{(k,X)} \leq \overline{\mathrm{BCB}(X, A_0, \ldots, A_n, k)}$, for all $j = 0, \ldots, k$.

(1) [Left column.]
$w \longleftarrow \mathrm{WIDTH}(X)$.
$W \longleftarrow w$.
${}^{0}B \longleftarrow 1$.
for $l = 1, \ldots, d_n$ do ${}^{l}B \longleftarrow {}^{l-1}B \; W/(k-l+1)$, $W \longleftarrow W + w$.
for $l = 0, \ldots, d_n$ do ${}^{l}B \longleftarrow {}^{l}B \; A_l$.

(2) [Inner points and bounds for Bernstein coefficients]
$F \longleftarrow {}^{0}B$.
for $j = 1, \ldots, k$
for $i = 0, \ldots, \min\{d_n - 1, k - j\}$ do ${}^{l}B \longleftarrow {}^{l}B + {}^{l+1}B$.
$\underline{F} \longleftarrow \min\{{}^{0}\underline{B}, \underline{F}\}$, $\overline{F} \longleftarrow \max\{{}^{0}\overline{B}, \overline{F}\}$.

(3) [Return.]
return F.

Theorem 3.4.22 (Complexity) *Algorithm 3.4.21 (*BCB*) costs*

$$\begin{array}{rl} 2d_n & \textit{number divisions,} \\ 4d_n + 2 & \textit{number multiplications and} \\ d_n(2k - d_n + 3) + 1 & \textit{number additions.} \ \square \end{array}$$

Proof.

- Step 1 costs $2d_n$ number divisions, $2d_n + 1$ interval multiplications and $2d_n + 1$ number additions. As $0 \notin \mathrm{int}({}^lB)$ for all l, each interval multiplication costs 2 number multiplications.
- Step 2 costs $d_n(2k - d_n + 1)$ number additions. $\square$

An algorithm for computing the Bernstein form is now straight forward.

Algorithm 3.4.23 (BF**)** [Bernstein Form]

In: $f(x) = a_n x^{d_n} + a_{n-1} x^{d_{n-1}} + \ldots + a_1 x^{d_1} \in \mathbb{F}[x]$,
$X \in \mathbb{IF}$,
$k \in \mathbb{N}$, $k \geq d_n$.

Out: $\mathrm{BF}(f, X, k) \in \mathbb{IF}, \mathrm{BF}(f, X, k) \supseteq B_f^{(k)}(X)$.

(1) [Taylor coefficients.]
$(A_0, \ldots, A_{d_n}) \longleftarrow \mathrm{TC}(f, \underline{X})$.

(2) [Bounds for Bernstein coefficients.]
$F \longleftarrow \mathrm{BCB}(X, A_0, \ldots, A_{d_n}, k)$.

(3) [Return.]
return F.

Theorem 3.4.24 (Complexity) *Algorithm 3.4.23 (*BF*) costs*

$$\begin{array}{rl} 2d_n & \textit{number divisions,} \\ d_n^2 + 3d_n + 4 & \textit{number multiplications and} \\ 2kd_n + 4d_n + 1 & \textit{number additions.} \ \square \end{array}$$

Proof.

- Step 1 costs $d_n^2 - d_n + 2$ number multiplications $d_n^2 + d_n$ number additions (Theorem 3.3.7).
- Step 2 costs $2d_n$ number divisions, $4d_n + 2$ number multiplications and $d_n(2k - d_n + 3) + 1$ number additions (Theorem 3.4.22). $\square$

3.4.1 Bisection at Zero

If $0 \in X$ then we can avoid the expensive computation of Taylor coefficients in Algorithm 3.4.23 if we bisect X at 0 and compute the Bernstein coefficients separately for both halves of X. Note that the Taylor coefficients with developing point 0 are precisely the power basis coefficients. This reduces overall computing time if d_n is large and $k \approx d_n$. Further, due to the inclusion monotonicity of the Bernstein form, bisection leads usually to a reduction of the overestimation error. The content of this section is new.

Thus, in the sequel let

$$\check{B}_f^{(k)}(X) = \begin{cases} B_f^{(k)}([\underline{X}, 0]) \cup B_f^{(k)}([0, \overline{X}]) & \text{if } 0 \in X \\ B_f^{(k)}(X) & \text{else.} \end{cases}$$

The following algorithm computes $\check{B}_f^{(k)}(X)$.

Algorithm 3.4.25 (BFBZ) [Bernstein Form with Bisection at Zero]

In: $f(x) = a_n x^{d_n} + a_{n-1} x^{d_{n-1}} + \ldots + a_1 x^{d_1} \in \mathbb{F}[x]$,
$X \in \mathbb{IF}$,
$k \in \mathbb{N}$, $k \geq d_n$.

Out: $\mathrm{BFBZ}(f, X, k) \in \mathbb{IF}$, $\mathrm{BFBZ}(f, X, k) \supseteq \check{B}_f^{(k)}(X)$.

(1) [Test $0 \in X$.]
if $\underline{X} > 0$ or $\overline{X} < 0$ then return $\mathrm{BF}(f, X, k)$.

(1) [Coefficients of $f(x)$.]
for $i = 0, \ldots, d_n$ do $A_i \longleftarrow$ coefficient of x^i in f.

(2) [Bounds for Bernstein coefficients in $[0, \overline{X}]$.]
$F^+ \longleftarrow \mathrm{BCB}([0, \overline{X}], A_0, \ldots, A_{d_n}, k)$.

(3) [Coefficients of $f(-x)$.]
for $i = 1, \ldots, d_n$ step 2 do $A_i \longleftarrow -A_i$.

(4) [Bounds for Bernstein coefficients in $[0, -\underline{X}]$.]
$F^- \longleftarrow \mathrm{BCB}([0, -\underline{X}], A_0, \ldots, A_{d_n}, k)$.

(5) [Return.]
return $F^+ \cup F^-$.

Theorem 3.4.26 (Complexity)

- *If $0 \notin X$ then Algorithm 3.4.25 costs*

$$\begin{array}{rl} 2d_n & \textit{number divisions,} \\ d_n^2 + 3d_n + 4 & \textit{number multiplications and} \\ 2kd_n + 4d_n + 1 & \textit{number additions.} \end{array}$$

- *If $0 \in X$ then Algorithm 3.4.25 costs*

$$\begin{array}{rl} 4d_n & \textit{number divisions,} \\ 8d_n + 4 & \textit{number multiplications and} \\ 2d_n(2k - d_n + 3) + 2 & \textit{number additions.} \ \square \end{array}$$

Proof.

- If $0 \notin X$ then Algorithm 3.4.25 (BFBZ) and Algorithm 3.4.23 (BF) are identical.
- If $0 \in X$ then step 2 and step 4 cost each $2d_n$ number divisions, $4d_n + 2$ number multiplications and $d_n(2k - d_n + 3) + 1$ number additions. $\square$

Thus, if $0 \in X$ then Algorithm 3.4.25 (BFBZ) costs $d_n^2 - 5d_n$ number multiplications less, but $2d_n$ number divisions and $2d_n(k - d_n + 1) + 1$ number additions more than Algorithm BF. In particular, if $k = d_n$, the difference in number additions is only $2d_n + 1$. In this case BFBZ costs $d_n^2 - 9d_n - 1$ floating point operations less than BF, hence BFBZ is cheaper than BF if $d_n > 9$.

3.4.2 Experimental Results

An experimental comparison of Algorithm 3.4.23 (BF) and Algorithm 3.4.25 (BFBZ) for dense polynomials with different degrees is given in Table 3.3.1. The coefficients of the polynomials are uniformly distributed in $[-1, 1]$ and the order of the Bernstein form is $k \equiv d_n$. Further, $\underline{X}$ and $\overline{X}$ are uniformly chosen in $[-1, 0]$ respectively $[0, 1]$, such that in all cases $0 \in X$. For each degree d_n the average cost and overestimation error of 10^4 random polynomials is reported. The cost is the total number of arithmetic

Degree d_n of f	2	6	10	14	18	22	26	30
Flops for $B_f^{(d_n)}(X)$	35.0	167.0	395.0	719.0	1139.0	1655.0	2267.0	2975.0
Flops for $\check{B}_f^{(d_n)}(X)$	50.0	186.0	386.0	650.0	978.0	1370.0	1826.0	2346.0
$\mathrm{q}(B_f^{(d_n)}(X), f(X))/\mathrm{w}(X)^2$	0.0523	0.0363	0.0252	0.0205	0.0174	0.0154	0.0137	0.0133
$\mathrm{q}(\check{B}_f^{(d_n)}(X), f(X))/\mathrm{w}(X)^2$	0.0194	0.0120	0.0092	0.0079	0.0069	0.0062	0.0057	0.0054

Table 3.4.1: Comparison of Bernstein form $B_f^{(d_n)}$ and Bernstein form with bisection at zero $\check{B}_f^{(d_n)}(X)$ for random polynomials f with different degrees and random intervals X, where $0 \in X$.

floating point instructions, including those which were executed during interval operations. As the Bernstein form is quadratically convergent, the distance to the range divided by $\mathrm{w}(X)^2$ was chosen to measure accuracy. According to Table 3.4.1, the error of $B_f^{(d_n)}$ is between one half and one third of the error of $\check{B}_f^{(d_n)}$. The measured cost of both algorithms corresponds exactly to Theorem 3.4.24 respectively Theorem 3.4.26. Note that the overestimation error gets smaller as the degree of the polynomials increases. The reason is that k is equal to the degree. If k would be the same for all polynomials, then the error would grow with increasing degree.

3.5 Interpolation Form

Interpolation forms were introduced by [Cornelius and Lohner, 1984]. They are the first interval extensions which have convergence order higher than 2. The original definition of interpolation forms comprises a whole class of interval extensions with different convergence orders. A particular instance, which is cubically convergent, and which allows efficient evaluation, is suggested in [Cornelius and Lohner, 1984]. In this section, we consider only this instance and call it "the" interpolation form. Section 3.5.1 and Section 3.5.2 contain some modifications, which are new. Another cubically convergent interpolation form, the parabolic boundary value form, was introduced by [Neumaier, 1990] and is reviewed in Section 3.5.3.

The basic idea of interpolation forms is to approximate f by a low degree polynomial g such that

- the range of g can be computed without overestimation,
- the remainder $f - g$ can be enclosed with high convergence order.

Let us now derive the interpolation form. Using Taylor expansion of f at c we obtain

$$f(x) = f(c) + f'(c)(x-c) + 1/2 f''(\xi)(x-c)^2$$

for some ξ between c and x. For arbitrary $m \in \mathbb{R}$ we can continue with

$$f(x) = \underbrace{f(c) + f'(c)(x-c) + 1/2 m(x-c)^2}_{g_{c,m}(x)} + \underbrace{1/2(f''(\xi) - m)(x-c)^2}_{r_{c,m}(x)}.$$

Thus, a class of interval extensions is defined by

$$f(X) \subseteq g_{c,m}(X) + R_{c,m}(X), \tag{3.5.1}$$

where $R_{c,m}$ is an interval extension of $r_{c,m}$. Note that $g_{c,m}$ is a polynomial of degree two, hence $g_{c,m}(X)$ can be computed cheaply. The interpolation form is an instance of (3.5.1), where $c = \mathrm{mid}(X)$, $m = \mathrm{mid}(H_{f''}(X))$ and

$$R_{c,m}(X) = 1/2(H_{f''}(X) - m)(X-c)^2.$$

Definition 3.5.1 (Interpolation form) *The interpolation form $I_f : \mathbb{IR} \to \mathbb{IR}$ is defined as*

$$I_f(X) = g_{c,m}(X) + 1/2(H_{f''}(X) - m)(X - c)^2,$$

where

$$\begin{aligned} g_{c,m}(x) &= f(c) + f'(c)(x - c) + 1/2m(x - c)^2 \\ c &= \mathrm{mid}(X) \\ m &= \mathrm{mid}(H_{f''}(X)). \ \square \end{aligned}$$

In the sequel let c, m, and $g_{c,m}$ as in Definition 3.5.1 and let $r_{c,m} = f - g_{c,m}$.

Theorem 3.5.2 (Convergence) *I_f converges cubically to f.* $\square$

Proof. Let $A \in \mathbb{IR}$ arbitrary but fixed and let X range over subintervals of A. Note that for all $x \in X$ it holds that

$$\begin{aligned} |f(x) - g_{c,m}(x)| &= |r_{c,m}(x)| \\ &\leq \mathrm{mag}\left(1/2(H_{f''}(X) - m)(X - c)^2\right). \end{aligned}$$

Thus,

$$\mathrm{q}(f(X), g_{c,m}(X)) \leq \mathrm{mag}\left(1/2(H_{f''}(X) - m)(X - c)^2\right)$$

and

$$\begin{aligned} \mathrm{q}(f(X), I_f(X)) &= \mathrm{q}\big(f(X), g_{c,m}(X) + 1/2(H_{f''}(X) - m)(X - c)^2\big) \\ &\leq \mathrm{mag}\left((H_{f''}(X) - m)(X - c)^2\right). \end{aligned}$$

Hence, it suffices to find a constant λ such that for all $X \in \mathbb{I}A$

$$\mathrm{mag}\left((H_{f''}(X) - m)(X - c)^2\right) \leq \lambda \mathrm{w}(X)^3.$$

As $H_{f''}(X) - m$ is a centered interval we have

$$\begin{aligned} \mathrm{mag}\left((H_{f''}(X) - m)(X - c)^2\right) &= \mathrm{mag}([-1, 1]\mathrm{rad}(H_{f''}(X))\mathrm{rad}(X)^2) \\ &= 1/2\mathrm{w}(H_{f''}(X))\mathrm{rad}(X)^2 \\ &\leq 1/2\lambda_{H_{f''},A}\mathrm{w}(X)\mathrm{rad}(X)^2 \\ &= \frac{1}{8}\lambda_{H_{f''},A}\mathrm{w}(X)^3, \end{aligned}$$

where $\lambda_{H_{f''},A}$ is a Lipschitz constant of $H_{f''}$ in A. $\square$

The interpolation form requires exact range computation of parabolas. The following theorem ([Apostolatos and Kulisch, 1967]) shows, how this can be done efficiently.

Theorem 3.5.3 (Range of Parabola) *Let $g(x) = a_2x^2 + a_1x + a_0$. Then*

$$g(X) = \begin{cases} a_1X + a_0 & \text{if } a_2 = 0 \\ \frac{1}{a_2}\left(\left(a_2X + \frac{a_1}{2}\right)^2 - \left(\frac{a_1}{2}\right)^2\right) + a_0 & \text{if } a_2 \neq 0. \ \square \end{cases} \tag{3.5.2}$$

Proof. The case $a_2 = 0$ is trivial, hence assume $a_2 \neq 0$. Note that

$$\frac{1}{a_2}\left(\left(a_2X + \frac{a_1}{2}\right)^2 - \left(\frac{a_1}{2}\right)^2\right) + a_0 = \left\{\frac{1}{a_2}\left(\left(a_2x + \frac{a_1}{2}\right)^2 - \left(\frac{a_1}{2}\right)^2\right) + a_0 \mid x \in X\right\},$$

because X occurs only once in the expression on the left hand side. Further

$$\begin{aligned} \frac{1}{a_2}\left(\left(a_2x + \frac{a_1}{2}\right)^2 - \left(\frac{a_1}{2}\right)^2\right) + a_0 &= \frac{1}{a_2}\left(a_2^2x^2 + a_2a_1x + \left(\frac{a_1}{2}\right)^2 - \left(\frac{a_1}{2}\right)^2\right) + a_0 \\ &= a_2x^2 + a_1x + a_0. \ \square \end{aligned}$$

An implementation of (3.5.2) using floating point numbers is numerically unstable if $a_2 \approx 0$. Therefore, we prefer to evaluate g at the endpoints of X, test whether g has a local extremum in X, and if so, evaluate g at the extremum. Algorithm 3.5.4 (PE) computes bounds for the range of parabolas $A_2x^2 + A_1x + A_0$, where A_2, A_1 and A_0 are intervals. However, it is assumed that the width of A_2 and A_1 is small, which will be the case when PE is called later on by other algorithms. Otherwise, the overestimation error of PE might be large. If $0 \in A_2$, then PE evaluates the Horner scheme. This does not lead to big overestimation errors if $w(A_2)$ is small.

Algorithm 3.5.4 (PE) [Parabola Evaluation]

In: $A_2, A_1, A_0 \in \mathbb{IF}$,
$X \in \mathbb{IF}$.

Out: $\mathrm{PE}(A_2, A_1, A_0, X) \in \mathbb{IF}$, $\mathrm{PE}(A_2, A_1, A_0, X) \supseteq \{A_2x^2 + A_1x + A_0 \mid x \in X\}$.

(1) [Case $0 \in A_2$.]
if $0 \in A_2$ then return $(A_2X + A_1)X + A_0$.

(2) [Evaluate at endpoints.]
$Y \longleftarrow [(A_2\underline{X} + A_1)\underline{X}, A_2\overline{X} + A_1)\overline{X}]$.

(3) [Extremum.]
$B_1 \longleftarrow 0.5A_1$.
$E \longleftarrow (-B_1/A_2) \cap X$.
if $E \neq \emptyset$ then $Y \longleftarrow [Y \cup B_1E]$.

(4) [Return.]
return $Y + A_0$.

Theorem 3.5.5 (Complexity) *Algorithm 3.5.4 costs*

1 *interval multiplication,*
2 *number divisions,*
10 *number multiplications and*
6 *number additions.* □

Proof. If $0 \in A_2$ then Algorithm 3.5.4 (PE) costs 2 interval multiplications and 2 interval additions. Assume $0 \notin A_2$.

- Step 2 costs 8 number multiplications and 4 number additions.
- Step 3 costs 1 interval multiplication, 2 number divisions and 2 number multiplications.
- Step 4 costs 2 number additions. □

Algorithm 3.5.6 (IF) evaluates the interpolation form I_f. We assume that f is sparse, hence we do not use the extended Horner scheme for evaluating f' at c but compute the coefficients of f' explicitly. A modification of the algorithm for dense polynomials is given in Section 3.5.2. The coefficients of $g_{c,m}$ are

$$g_{c,m}(x) = 1/2mx^2 + (f'(c) - mc)x + f(c) - f'(c)c + 1/2mc^2.$$

Algorithm 3.5.6 (IF) [Interpolation Form]

In: $f(x) = a_nx^{d_n} + a_{n-1}x^{d_{n-1}} + \ldots + a_1x^{d_1} \in \mathbb{F}[x]$,
$X \in \mathbb{IF}$.

Out: $\mathrm{IF}(f, X) \in \mathbb{IF}$, $\mathrm{IF}(f, X) \approx I_f(X)$, $\mathrm{IF}(f, X) \supseteq f(X)$.

(1) [Midpoint.]
$c \longleftarrow \mathrm{MID}(X)$.

(2) [Derivatives of f, use rounded interval arithmetic for the coefficients.]
$F'(x) \longleftarrow$ formal derivative of f.
$F''(x) \longleftarrow$ formal derivative of $F'(x)$.

(3) [$f(c), f'(c), f''(X)$.]
$F \longleftarrow \mathrm{HF}(f(x), [c])$.
$F' \longleftarrow \mathrm{HFI}(F'(x), [c])$.
$F'' \longleftarrow \mathrm{HFI}(F''(x), X)$.

(4) [Coefficients of $g_{c,m}$.]
$M \longleftarrow [\mathrm{MID}(F'')]$.
$A_2 \longleftarrow 0.5M$.
$A_1 \longleftarrow F' - Mc$.
$A_0 \longleftarrow (A_2 c - F')c + F$.

(5) [Evaluation of parabola.]
$Y \longleftarrow \mathrm{PE}(A_2, A_1, A_0, X)$.

(6) [Add remainder and return.]
$r \longleftarrow \mathrm{mag}(X - c)$.
return $Y + (F'' - M)[0, 0.5 \hat{*} r \hat{*} r]$.

Theorem 3.5.7 (Complexity) *Algorithm 3.5.6 (*IF*) costs*

$3n$	*interval power computations,*
$n+1$	*interval multiplications,*
2	*number divisions,*
$8n+23$	*number multiplications and*
$6n+16$	*number additions.* □

Proof.

- Step 1 costs at 1 number multiplication and 2 number additions.
- Step 2 costs $4n$ number multiplications.
- Step 3 costs $3n$ interval power computations, n interval multiplications, $4n$ number multiplications and $6n-6$ number additions.
- Step 4 costs 9 number multiplication and 8 number additions.
- Step 5 costs 1 interval multiplication 2 number division, 10 number multiplications and 6 number additions.
- Step 6 costs 3 number multiplications and 6 number additions. □

3.5.1 Reduction of the Overestimation Error

In this section we present an improvement $\ddot{I}_f$ of the interpolation form I_f. The costs for evaluating $\ddot{I}_f$ and I_f are comparable (Theorem 3.5.12), but $\ddot{I}_f$ gives always at least as tight bounds as I_f (Theorem 3.5.10). The idea is to evaluate I_f twice with different choices of m and intersect the result. The content of this section is new.

Definition 3.5.8 *The modified interpolation form* $\ddot{I}_f : \mathbb{IR} \to \mathbb{IR}$ *of* f *is defined as*

$$\ddot{I}_f(X) = [\underline{p_c^{\downarrow}(X)}, \overline{p_c^{\uparrow}(X)}],$$

where

$$\begin{aligned} p_c^{\uparrow}(x) &= f(c) + f'(c)(x-c) + 1/2\overline{F''} * (x-c)^2 \\ p_c^{\downarrow}(x) &= f(c) + f'(c)(x-c) + 1/2\underline{F''} * (x-c)^2, \end{aligned}$$

$F'' = H_{f''}(X)$ *and* $c = \mathrm{mid}(X)$. □

In the sequel let F'', c, $p_c^{\uparrow}$ and $p_c^{\downarrow}$ as in Definition 3.5.8.

Remark. If $I_f(X)$ is evaluated twice with $m = \underline{F''}$ and $m = \overline{F''}$ and the results are intersected, then $\ddot{I}_f(X)$ is obtained. □

Theorem 3.5.9 *$\ddot{I}_f$ is an interval extension of f.* □

Proof. Let $x \in X$ arbitrary but fixed and let $\xi \in X$ such that

$$f(x) = f(c) + f'(c)(x-c) + 1/2 f''(\xi)(x-c)^2.$$

Then

$$\begin{aligned} \overline{p_c^{\uparrow}(X)} &\geq f(c) + f'(c)(x-c) + 1/2\overline{F''} * (x-c)^2 \\ &\geq f(c) + f'(c)(x-c) + 1/2 f''(\xi)(x-c)^2 \\ &= f(x) \\ \underline{p_c^{\downarrow}(X)} &\leq f(c) + f'(c)(x-c) + 1/2\underline{F''} * (x-c)^2 \\ &\leq f(c) + f'(c)(x-c) + 1/2 f''(\xi)(x-c)^2 \\ &= f(x), \end{aligned}$$

i.e. $f(X) \subseteq \ddot{I}_f(X)$. Further, $\ddot{I}_f(X) \in \mathbb{R}$ if $X \in \mathbb{R}$. □

Theorem 3.5.10 *For all $X \in \mathbb{IR}$ it holds that*

$$\ddot{I}_f(X) \subseteq I_f(X).$$ □

Proof. Note that

$$\overline{p_c^{\uparrow}(X)} \geq \underline{p_c^{\downarrow}(X)}.$$

Hence, we have to show that

$$\overline{I_f(X)} \geq \overline{p_c^{\uparrow}(X)} \tag{3.5.3}$$

$$\underline{I_f(X)} \leq \underline{p_c^{\downarrow}(X)}. \tag{3.5.4}$$

- We begin with (3.5.3). Let $g_{c,m}$ as in Definition 3.5.1 and let $a, b \in X$ such that

$$\begin{aligned} p_c^{\uparrow}(a) &= \max_{x \in X} p_c^{\uparrow}(x) \\ g_{c,m}(b) &= \max_{x \in X} g_{c,m}(x). \end{aligned}$$

  Then

$$\begin{aligned} \overline{p^{\uparrow}(X)} &= p^{\uparrow}(a) \\ &= f(c) + f'(a-c) + 1/2\overline{F''}(a-c)^2 \\ &= f(c) + f'(a-c) + 1/2\mathrm{mid}(F'')(a-c)^2 + 1/2\mathrm{rad}(F'')(a-c)^2 \\ &= g_{c,m}(a) + 1/2\mathrm{rad}(F'')(a-c)^2 \\ &\leq g_{c,m}(b) + 1/2\mathrm{rad}(F'')(\overline{X}-c)^2 \\ &= \overline{I_f(X)}. \end{aligned}$$

- Similarly, for (3.5.4), let $g_{c,m}$ as in Definition 3.5.1 and let $a, b \in X$, such that

$$\begin{aligned} p_c^{\downarrow}(a) &= \min_{x \in X} p_c^{\downarrow}(x) \\ g_{c,m}(b) &= \min_{x \in X} g_{c,m}(x). \end{aligned}$$

Then

$$\begin{aligned} \underline{p^{\downarrow}(X)} &= p^{\downarrow}(a) \\ &= f(c) + f'(a-c) + 1/2\underline{F''}(a-c)^2 \\ &= f(c) + f'(a-c) + 1/2\mathrm{mid}(F'')(a-c)^2 - 1/2\mathrm{rad}(F'')(a-c)^2 \\ &= g_{c,m}(a) - 1/2\mathrm{rad}(F'')(a-c)^2 \\ &\geq g_{c,m}(b) - 1/2\mathrm{rad}(F'')(\overline{X}-c)^2 \\ &= \underline{I_f(X)}. \ \square \end{aligned}$$

The following algorithm evaluates $\ddot{I}_f$. As in Algorithm 3.5.6 (IF) we assume that f is sparse. The coefficients of $p_c^{\uparrow}$ and $p_c^{\downarrow}$ are

$$\begin{aligned} p_c^{\uparrow}(x) &= 1/2\overline{F''}x^2 + (f'(c) - \overline{F''}c)x + f(c) - f'(c)c + 1/2\overline{F''}c^2 \\ p_c^{\downarrow}(x) &= 1/2\underline{F''}x^2 + (f'(c) - \underline{F''}c)x + f(c) - f'(c)c + 1/2\underline{F''}c^2. \end{aligned}$$

Algorithm 3.5.11 (MIF) [Modified Interpolation Form]

In: $f(x) = a_n x^{d_n} + a_{n-1}x^{d_{n-1}} + \ldots + a_1 x^{d_1} \in \mathbb{F}[x]$,
$X \in \mathbb{IF}$.

Out: $\mathrm{MIF}(f, X) \in \mathbb{IF}$, $\mathrm{MIF}(f, X) \approx \ddot{I}_f(X)$, $\mathrm{MIF}(f, X) \supseteq \ddot{I}_f(X)$.

(1) [Midpoint.]
$c \longleftarrow \mathrm{MID}(X)$.

(2) [Derivatives of f, use rounded interval arithmetic for the coefficients.]
$F'(x) \longleftarrow$ formal derivative of $f(x)$.
$F''(x) \longleftarrow$ formal derivative of $F'(x)$.

(3) [$f(c), f'(c), f''(X)$.]
$F \longleftarrow \mathrm{HF}(f(x), [c])$.
$F' \longleftarrow \mathrm{HFI}(F'(x), [c])$.
$F'' \longleftarrow \mathrm{HFI}(F''(x), X)$.

(4) [Coefficients of $p_c^{\uparrow}$ and $p_c^{\downarrow}$.]
$A_2 \longleftarrow 0.5F''$.
$A_2^{\uparrow} \longleftarrow \overline{A}_2$, $A_1^{\uparrow} \longleftarrow F' - [\overline{F}''] * c$, $A_0^{\uparrow} \longleftarrow (A_2^{\uparrow}c - F')c$.
$A_2^{\downarrow} \longleftarrow \underline{A}_2$, $A_1^{\downarrow} \longleftarrow F' - [\underline{F}''] * c$, $A_0^{\downarrow} \longleftarrow (A_2^{\downarrow}c - F')c$.

(5) [Evaluate parabolas.]
$Y \longleftarrow [\mathrm{PE}(A_2^{\uparrow}, A_1^{\uparrow}, A_0^{\uparrow}, X), \mathrm{PE}(A_2^{\downarrow}, A_1^{\downarrow}, A_0^{\downarrow}, X)] + F$.

(6) [Return.]
return Y.

Theorem 3.5.12 (Complexity) *Algorithm 3.5.11* (MIF) *costs*

$$\begin{aligned} 3n &\quad \textit{interval power computations,} \\ n+2 &\quad \textit{interval multiplications,} \\ 4 &\quad \textit{number divisions,} \\ 8n+35 &\quad \textit{number multiplication and} \\ 6n+18 &\quad \textit{number additions.} \ \square \end{aligned}$$

Proof.

- Step 1 costs 1 number multiplication and 2 number additions.
- Step 2 costs $4n$ number multiplications.
- Step 3 costs $3n$ interval power computations, n interval multiplications, $4n$ number multiplications and $6n-6$ number additions.
- Step 4 costs 14 number multiplications and 8 number additions.
- Step 5 costs 2 interval multiplications, 4 number divisions, 20 number multiplications 14 number additions. □

3.5.2 Slopes Instead of Derivatives

The comparison of the mean value form and the slope form in Section 3.2.1 turned out, that using slopes instead of derivatives improves efficiency and accuracy for dense polynomials. In the following, we apply this idea to the modified interpolation form.

Definition 3.5.13 (Slope Interpolation Form) *The slope interpolation form* $\ddot{I}_f^{(s)} : \mathbb{IR} \to \mathbb{IR}$ *of* f *is defined as*

$$\ddot{I}_f^{(s)}(X) = [\underline{p_c^{\downarrow}(X)}, \overline{p_c^{\uparrow}(X)}],$$

where $c = \mathrm{mid}(X)$,

$$\begin{aligned} p_c^{\uparrow}(x) &= f(c) + f'(c)(x-c) + \overline{H}(x-c)^2 \\ p_c^{\downarrow}(x) &= f(c) + f'(c)(x-c) + \underline{H}(x-c)^2 \\ H &= H^*_{h_c}(X) \end{aligned}$$

and h_c *is the uniquely defined polynomial such that*

$$f(x) = f(c) + f'(c)(x-c) + h_c(x)(x-c)^2. \ \square$$

In the sequel let $p_c^{\uparrow}$, $p_c^{\downarrow}$, h_c and H as in Definition 3.5.13. The proof that $\ddot{I}_f^{(s)}$ is an interval extension of f, is analogous to the proof of Theorem 3.5.9.

Theorem 3.5.14 (Convergence) $\ddot{I}_f^{(s)}$ *converges cubically to* f. □

Proof. Let $A \in \mathbb{IR}$ and let X range over subintervals of A. For all $x \in X$ it holds that

$$\begin{aligned} |f(x) - p_c^{\uparrow}(x)| &= |\overline{H} - h_c(x)|(x-c)^2 \leq \mathrm{w}(H)\mathrm{w}(X)^2 \\ |f(x) - p_c^{\downarrow}(x)| &= |\underline{H} - h_c(x)|(x-c)^2 \leq \mathrm{w}(H)\mathrm{w}(X)^2. \end{aligned}$$

According to Corollary 1.3.19, $H^*_{h_{\mathrm{mid}(\cdot)}}$ is Lipschitz, and there exists $\lambda_{H,A} \in \mathbb{R}$ such that

$$\mathrm{w}(H^*_{h_c}(X) \leq \lambda_{H,A}\mathrm{w}(X)$$

for all $X \in \mathbb{I}A$. Hence,

$$\begin{aligned} \mathrm{q}(f(X), \ddot{I}_f^{(s)}(X)) &\leq \mathrm{w}(H)\mathrm{w}(X)^2 \\ &\leq \lambda_{H,A}\mathrm{w}(X)^3. \ \square \end{aligned}$$

The following algorithm evaluates $\ddot{I}_f^{(s)}$. For the computation of $f(c), f'(c)$ and the coefficients of h_c we use the extended Horner scheme. The coefficients of $p_c^{\uparrow}$ and $p_c^{\downarrow}$ are

$$\begin{aligned} p_c^{\uparrow}(x) &= \overline{H}x^2 + (f'(c) - 2\overline{H}c)x + f(c) - f'(c)c + \overline{H}c^2 \\ p_c^{\downarrow}(x) &= \underline{H}x^2 + (f'(c) - 2\underline{H}c)x + f(c) - f'(c)c + \underline{H}c^2. \end{aligned}$$

Algorithm 3.5.15 (SIF) [Slope Interpolation Form]

In: $f(x) = a_n x^{d_n} + a_{n-1} x^{d_{n-1}} + \ldots + a_1 x^{d_1} \in \mathbb{F}[x]$,
$X \in \mathbb{IF}$.

Out: $\mathrm{SIF}(f,X) \in \mathbb{IR}$, $\mathrm{SIF}(f,X) \approx \ddot{I}_f^{(s)}(X)$, $\mathrm{SIF}(f,X) \supseteq f(X)$.

(1) [Trival cases.]
if $d_n < 2$ return $\mathrm{HF}(f,X)$.

(2) [Midpoint.]
$c \longleftarrow \mathrm{MID}(X)$.

(3) [Dense coefficients.]
for $i = 0, \ldots, d_n$ do $A_i \longleftarrow$ coefficient of x^i in f.

(4) [Horner scheme, $f(c)$, $f'(c)$, h_c.]
for $i = d_n, \ldots, 1$ do $A_{i-1} \longleftarrow A_{i-1} + cA_i$.
for $i = d_n, \ldots, 2$ do $A_{i-1} \longleftarrow A_{i-1} + cA_i$.

(5) [Compute $H^*_{h_c}(X)$.]
$H \longleftarrow A_{d_n}$.
for $i = d_n - 1, \ldots, 2$ do $H \longleftarrow HX + A_i$.

(6) [Coefficients of parabolas.]
$H_c^{\uparrow} \longleftarrow [\overline{H}] * c$, $H_1^{\uparrow} \longleftarrow A_1 - H_c^{\uparrow}$, $P_1^{\uparrow} \longleftarrow H_1^{\uparrow} - H_c^{\uparrow}$, $P_0^{\uparrow} \longleftarrow -H_1^{\uparrow} c$.
$H_c^{\downarrow} \longleftarrow [\underline{H}] * c$, $H_1^{\downarrow} \longleftarrow A_1 - H_c^{\downarrow}$, $P_1^{\downarrow} \longleftarrow H_1^{\downarrow} - H_c^{\downarrow}$, $P_0^{\downarrow} \longleftarrow -H_1^{\downarrow} c$.

(7) [Evaluate parabolas.]
$Y \longleftarrow [\mathrm{PE}(\overline{H}, P_1^{\uparrow}, P_0^{\uparrow}, X), \mathrm{PE}(\underline{H}, P_1^{\downarrow}, P_0^{\downarrow}, X)] + A_0$.

(8) [Return.]
return Y.

Theorem 3.5.16 (Complexity) *Algorithm 3.5.15 costs*

d_n	*interval multiplications,*
4	*number divisions,*
$4d_n + 27$	*number multiplication and*
$6d_n + 18$	*number additions.* □

Proof.

- Step 2 costs 1 number multiplication and 2 number additions.
- Step 4 costs $4d_n - 2$ number multiplications and $4d_n - 2$ number additions.
- Step 5 costs $d_n - 2$ interval multiplications and $2d_n - 4$ number additions.
- Step 6 costs 8 number multiplications and 8 number additions.
- Step 7 costs 4 number division, 2 interval multiplications, 20 number multiplications and 14 number additions. □

3.5.3 Parabolic Boundary Value Form

The parabolic boundary value form is a modification of the interpolation form, which was introduced by [Neumaier, 1990]. For the approximation of f, the values $f(\underline{X})$, $f(\overline{X})$ are used, whereas the interpolation forms in the previous sections used $f(c)$ and $f'(c)$.

Definition 3.5.17 (Parabolic Boundary Value Form) *The parabolic boundary value form* $\tilde{I}_f^{(s)} : \mathbb{IR} \to \mathbb{IR}$ *is defined as*

$$\tilde{I}_f^{(s)}(X) = [\underline{p_X^{\downarrow}(X)}, \overline{p_X^{\uparrow}(X)}],$$

where

$$\begin{aligned} p_X^{\uparrow}(x) &= a_X + b_X x + \underline{H}(x - \underline{X})(x - \overline{X}) \\ p_X^{\downarrow}(x) &= a_X + b_X x + \overline{H}(x - \underline{X})(x - \overline{X}) \\ H &= H_{h_X}^*(X) \end{aligned}$$

and a_X, b_X, h_X *are uniquely defined by*

$$f(x) = a_X + b_X x + h_X(x)(x - \underline{X})(x - \overline{X}). \ \square$$

In the sequel let a_X, b_X and h_X as in Definition 3.5.17.

Theorem 3.5.18 (Convergence) $\tilde{I}_f^{(s)}$ *converges cubically to* f. $\square$

Proof. See [Neumaier, 1990]. $\square$

For the computation of a_X, b_X and the coefficients of h_X we use the extended Horner scheme. More precisely, we compute first $f(\underline{X})$ and the coefficients of the polynomial g_X such that

$$f(x) = f(\underline{X}) + g_X(x)(x - \underline{X}).$$

Next, we compute $g_X(\overline{X})$ and the coefficients of h_X, where

$$g_X(x) = g_X(\overline{X}) + h_X(x)(x - \overline{X}).$$

Thus,

$$\begin{aligned} f(x) &= f(\underline{X}) + \big(g_X(\overline{X}) + h_X(x)(x - \overline{X})\big)(x - \underline{X}) \\ &= \underbrace{f(\underline{X}) - g_X(\overline{X})\underline{X}}_{a_X} + \underbrace{g_X(\overline{X})}_{b_X} x + h_X(x)(x - \underline{X})(x - \overline{X}). \end{aligned}$$

The coefficients of $p_X^{\uparrow}$ and $p_X^{\downarrow}$ are

$$\begin{aligned} p_X^{\uparrow}(x) &= f(\underline{X}) - g_X(\overline{X})\underline{X} + g_X(\overline{X})x + \underline{H}(x - \underline{X})(x - \overline{X}) \\ &= \underline{H}x^2 + \big(g_X(\overline{X}) - \underline{H}(\overline{X} + \underline{X})\big)x + f(\underline{X}) - g_X(\overline{X})\underline{X} + \underline{H}\,\underline{X}\overline{X} \\ p_X^{\downarrow}(x) &= f(\underline{X}) - g_X(\overline{X})\underline{X} + g_X(\overline{X})x + \overline{H}(x - \underline{X})(x - \overline{X}) \\ &= \overline{H}x^2 + \big(g_X(\overline{X}) - \overline{H}(\overline{X} + \underline{X})\big)x + f(\underline{X}) - g_X(\overline{X})\underline{X} + \overline{H}\,\underline{X}\overline{X}. \end{aligned}$$

Algorithm 3.5.19 (PBF) [Parabolic Boundary Value Form]

In: $f(x) = a_n x^{d_n} + a_{n-1} x^{d_{n-1}} + \ldots + a_1 x^{d_1} \in \mathbb{F}[x]$,
$X \in \mathbb{IF}$.

Out: $\mathrm{PBF}(f, X) \in \mathbb{IR}$, $\mathrm{PBF}(f, X) \supseteq \tilde{I}_f^{(s)}(X)$.

(1) [Trivial cases.]
if $d_n < 2$ return $\mathrm{HF}(f, X)$.

(2) [Dense coefficients.]
for $i = 0, \ldots, d_n$ do $A_i \longleftarrow$ coefficient of x^i in f.

(3) [Horner scheme, $f(\underline{X})$, $g_X(\overline{X})$, h_X.]
for $i = d_n, \ldots, 1$ do $A_{i-1} \longleftarrow A_{i-1} + \underline{X} A_i$.
for $i = d_n, \ldots, 2$ do $A_{i-1} \longleftarrow A_{i-1} + \overline{X} A_i$.

(4) [Compute $H^*_{h_X}(X)$.]
$H \longleftarrow A_{d_n}$.
for $i = d_n - 1, \ldots, 2$ do $H \longleftarrow HX + A_i$.

(5) [Coefficients of parabolas.]
$X^+ \longleftarrow [\underline{X}] + [\overline{X}]$.
$X^* \longleftarrow [\underline{X}] * [\overline{X}]$.
$P_1^{\uparrow} \longleftarrow A_1 - \underline{H}X^+$, $P_0^{\uparrow} \longleftarrow \underline{H}X^*$.
$P_1^{\downarrow} \longleftarrow A_1 - \overline{H}X^+$, $P_0^{\downarrow} \longleftarrow \overline{H}X^*$.

(6) [Evaluate parabolas.]
$Y \longleftarrow [\mathrm{PE}(\underline{H}, P_1^{\uparrow}, P_0^{\uparrow}, X), \mathrm{PE}(\overline{H}, P_1^{\downarrow}, P_0^{\downarrow}, X)] + A_0$.

(7) [Return.]
return Y.

Theorem 3.5.20 (Complexity) *Algorithm 3.5.19 (*PBF*) costs*

d_n	*interval multiplications,*
4	*number divisions,*
$4d_n + 28$	*number multiplications and*
$6d_n + 12$	*number additions.* □

Proof.

- Step 3 costs $4d_n - 2$ number multiplications and $4d_n - 2$ number additions.
- Step 4 costs $d_n - 2$ interval multiplications and $2d_n - 4$ number additions.
- Step 5 costs 10 number multiplications and 6 number additions.
- Step 6 costs 2 interval multiplications, 4 number divisions, 20 number multiplications and 12 number additions. □

3.5.4 Experimental Results

An experimental comparison of the interpolation form I_f, computed by Algorithm 3.5.6 (IF), the modified interpolation form $\ddot{I}_f$, computed by Algorithm 3.5.11 (MIF), the slope interpolation form $\ddot{I}_f^{(\mathrm{s})}$, computed by Algorithm 3.5.15 (SIF), and the parabolic boundary value form $\tilde{I}_f^{(\mathrm{s})}$, computed by Algorithm 3.5.19 (PBF), for dense and sparse polynomials with different degrees is given in Table 3.5.1, respectively Table 3.5.2. The coefficients of the polynomials and the endpoints of the input intervals are uniformly distributed in $[-1, 1]$. For each degree d_n the average cost and overestimation error of 10^4 random polynomials is reported. The cost is the total number of arithmetic floating point instructions, including those which were executed during interval operations. As the forms are cubically convergent, the distance to the range divided by $\mathrm{w}(X)^3$ was chosen to measure accuracy.

- The accuracy of I_f and $\ddot{I}_f$ is almost the same, but $\ddot{I}_f$ is more expensive.
- In all cases $\ddot{I}_f^{(\mathrm{s})}$ and $\tilde{I}_f^{(\mathrm{s})}$ return significantly tighter inclusions than I_f and $\ddot{I}_f$. If f is dense, then $\ddot{I}_f^{(\mathrm{s})}$ is cheaper than I_f and $\ddot{I}_f$.
- $\ddot{I}_f^{(\mathrm{s})}$ gives usually tighter inclusions than $\tilde{I}_f^{(\mathrm{s})}$. However, the following experiment shows that this depends on $\mathrm{w}(X)$.

Figure 3.5.1 shows a comparison of the slope interpolation form $\ddot{I}_f^{(\mathrm{s})}$ and the parabolic boundary value form $\tilde{I}_f^{(\mathrm{s})}$ for dense polynomials of degree 10. In the left graph, the input intervals X are chosen such that $\mathrm{mid}(X) \equiv 0.5$ and $\mathrm{w}(X)$ varies between 0 and 1, in the right graph $\mathrm{mid}(X) \equiv 0$ and $\mathrm{w}(X)$ varies between 0 and 2. The curves are averages over 1000 random polynomials with coefficients in $[-1, 1]$.

Degree of f (dense)	2	6	10	14	18	22	26	30
Flops for $I_f(X)$	68.6	136.9	203.9	270.6	337.4	404.3	470.7	537.4
Flops for $\ddot{I}_f(X)$	87.2	154.2	221.3	288.1	354.9	421.8	488.2	554.9
Flops for $\ddot{I}_f^{(\mathrm{s})}(X)$	69.2	118.9	169.2	219.5	270.0	320.6	370.7	421.2
Flops for $\tilde{I}_f^{(\mathrm{s})}(X)$	71.2	120.8	171.0	221.2	271.7	322.1	372.2	422.7
$\mathrm{q}(I_f(X), f(X))/\mathrm{w}(X)^3$	0.00	1.14	3.26	6.26	10.52	14.73	19.82	26.92
$\mathrm{q}(\ddot{I}_f(X), f(X))/\mathrm{w}(X)^3$	0.00	1.14	3.26	6.25	10.51	14.73	19.81	26.92
$\mathrm{q}(\ddot{I}_f^{(\mathrm{s})}(X), f(X))/\mathrm{w}(X)^3$	0.00	0.15	0.26	0.37	0.49	0.59	0.72	0.87
$\mathrm{q}(\tilde{I}_f^{(\mathrm{s})}(X), f(X))/\mathrm{w}(X)^3$	0.00	0.11	0.32	0.59	0.97	1.23	1.66	2.26

Table 3.5.1: Comparison of interpolation form I_f, modified interpolation form $\ddot{I}_f$, slope interpolation form $\ddot{I}_f^{(\mathrm{s})}$ and parabolic boundary value form $\tilde{I}_f^{(\mathrm{s})}$ for *dense* random polynomials f with different degrees and random intervals X. The coefficients of f and the endpoints of X are uniformly chosen in $[-1, 1]$.

Degree of f (sparse)	2	6	10	14	18	22	26	30
Flops for $I_f(X)$	68.6	101.4	120.9	133.5	143.4	150.5	156.4	162.1
Flops for $\ddot{I}_f(X)$	87.2	117.2	136.4	148.9	158.6	165.4	171.1	176.7
Flops for $\ddot{I}_f^{(\mathrm{s})}(X)$	69.2	119.8	171.0	222.8	274.7	326.5	378.6	430.2
Flops for $\tilde{I}_f^{(\mathrm{s})}(X)$	71.2	121.3	172.1	223.4	274.8	326.4	378.0	429.3
$\mathrm{q}(I_f(X), f(X))/\mathrm{w}(X)^3$	0.00	0.79	1.85	2.95	4.19	5.71	6.82	8.94
$\mathrm{q}(\ddot{I}_f(X), f(X))/\mathrm{w}(X)^3$	0.00	0.79	1.85	2.95	4.19	5.71	6.81	8.94
$\mathrm{q}(\ddot{I}_f^{(\mathrm{s})}(X), f(X))/\mathrm{w}(X)^3$	0.00	0.08	0.12	0.14	0.16	0.18	0.25	0.31
$\mathrm{q}(\tilde{I}_f^{(\mathrm{s})}(X), f(X))/\mathrm{w}(X)^3$	0.00	0.10	0.22	0.32	0.44	0.56	0.65	0.87

Table 3.5.2: Comparison of interpolation form I_f, modified interpolation form $\ddot{I}_f$, slope interpolation form $\ddot{I}_f^{(\mathrm{s})}$ and parabolic boundary value form $\tilde{I}_f^{(\mathrm{s})}$ for *sparse* random polynomials f with different degrees and random intervals X. Each polynomial has only 3 non–zero coefficients. The coefficients of f and the endpoints of X are uniformly chosen in $[-1, 1]$.

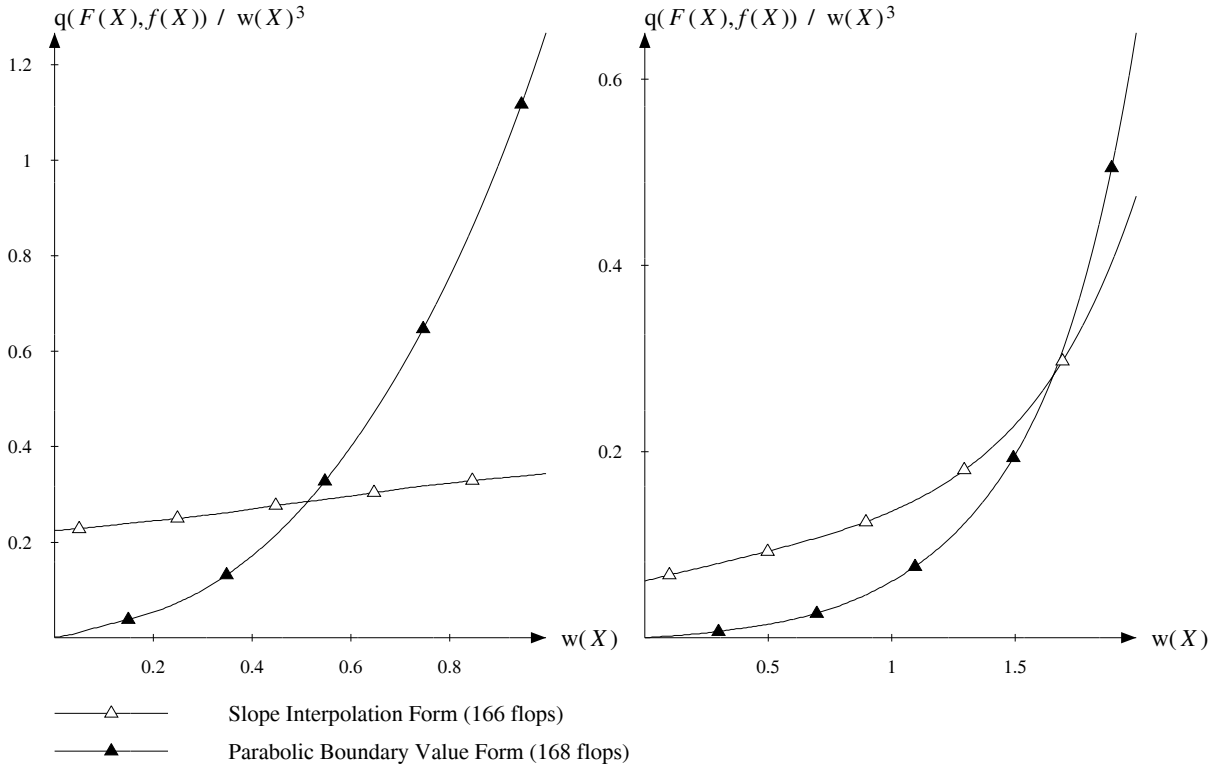


Figure 3.5.1: Comparison of $\ddot{I}_f^{(s)}$ and $\tilde{I}_f^{(s)}$ for dense random polynomials of degree 10. In the left graph the input intervals X are such that $\mathrm{mid}(X) \equiv 0.5$ and $\mathrm{w}(X)$ varies between 0 and 1, in the right graph $\mathrm{mid}(X) \equiv 0$ and $\mathrm{w}(X)$ varies between 0 and 2. In both cases $\ddot{I}_f^{(s)}$ is more accurate than $\tilde{I}_f^{(s)}$ for large intervals whereas $\tilde{I}_f^{(s)}$ is better than $\ddot{I}_f^{(s)}$ for small intervals.

3.6 Experimental Comparison

After having defined several interval extensions of univariate polynomials in the previous sections, we now come to a comparison. The computational cost for the evaluation of the interval extensions can be compared easily because it depends basically only on the degree and on the sparsity of the input polynomial. A comparison of the overestimation errors is more involved, because the overestimation error depends on the actual value of the polynomial coefficients and on the input interval.

The main theoretical results of the previous sections concerning accuracy can be summarized as follows.

- For input intervals with small width, interpolation forms are superior because of their cubic convergence (Theorem 3.5.2, 3.5.14, 3.5.18), the Horner form is worst because it is only linearly convergent (Theorem 3.1.5), and the chances, that the bicentered mean value form and the Bernstein form are exact are high (Theorem 3.2.17, 3.4.19).
- If the Horner form of f' evaluated on X does not contain zero, then the bicentered mean value form is exact (Theorem 3.2.17). Note that every interval extension can be modified to be exact in this case, simply by executing the monotonicity test first.
- The bicentered mean value form gives always at least as tight inclusions as the mean value form (Corollary 3.2.19).
- The dense slope form is at least as accurate as the dense mean value form (Conjecture 3.2.12).
- If $\mathrm{mig}(X)$ is large enough, then the Horner forms are exact (Theorem 3.1.12).
- If $\mathrm{mid}(X) = 0$, then the (dense) Horner form, the (dense) Taylor form and the (dense) slope form are equivalent.
- The overestimation error of the dense Taylor form with bisection is at most half as large as the overestimation error of the dense Taylor form (Theorem 3.3.11).
- If $0 \in \mathrm{int}(X)$ then a bisection at zero is inexpensive for the Horner forms (Algorithm 3.1.38) and the Bernstein form (Algorithm 3.4.25), and usually reduces the overestimation error.

Section 3.6.1 contains an experimental comparison of interval extensions for various classes of random polynomials and a fixed input interval X. This kind of comparison shows statistically, which interval extension is best for a given input interval X.

Every interval extension is a compromise between computational cost and accuracy. For individual algorithms, which evaluate interval extensions as subroutines, the optimal choice may therefore be different. For example, in an iterative algorithm it may be cheaper to choose an inaccurate and inexpensive interval extension (like the Horner form) and execute some more iterations, or to choose an accurate and expensive interval extension (like the Bernstein form) and save some iterations. In Section 3.6.2 and 3.6.3 we tackle this problem for two typical applications of interval arithmetic, namely finding the roots and the global minimum of univariate polynomials.

In the sequel, when we refer to the Bernstein form of a polynomial, we implicitly assume, that the order of the Bernstein form is equal to the degree of the polynomial.

3.6.1 Efficiency and Accuracy for Random Polynomials

In this section we compare the average overestimation error of some interval extensions for 3 classes of random polynomials and for 2 classes of input intervals.

- The classes of random polynomials f are:
  - Dense polynomials with degree 5 (Figure 3.6.1, 3.6.2).
  - Dense polynomials with degree 15 (Figure 3.6.3, 3.6.4).

  - Sparse polynomials with degree 10, where only 3 coefficients are non–zero (Figure 3.6.5, 3.6.6).

  The coefficients are uniformly chosen from $[-1, 1]$.

- The input intervals X are as follows:
  - $\mathrm{mid}(X) \equiv 0$, $0 \leq \mathrm{w}(X) \leq 1.0$ (Figure 3.6.1, 3.6.3, 3.6.5).
  - $\mathrm{mid}(X) \equiv 0.3$, $0 \leq \mathrm{w}(X) \leq 1.0$ (Figure 3.6.2, 3.6.4, 3.6.6).

In order to measure the overestimation error of some interval extension F of f we use

$$\frac{\mathrm{q}(F(X), f(X))}{\mathrm{w}(X)}.$$

In each experiment we report the average overestimation error of 1000 random polynomials in dependence of the input interval as well as the average number of floating point operations for the evaluation.

3.6.2 Newton's Method

The (one–dimensional) interval Newton method is an iterative algorithm, which computes inclusions of all roots of a polynomial f in an interval A. It was studied thoroughly in the literature, see for example [Moore, 1966], [Krawczyk, 1969], [Alefeld, 1970], [Hansen, 1978a], [Hansen, 1978b], [Hansen, 1979], [Alefeld and Herzberger, 1983], [Hansen, 1992], [Hammer *et al.*, 1993] and many others. In each iteration, an interval extension of f' has to be evaluated. In this section we compare the efficiency of the interval Newton method in dependence of the chosen interval extension. The number of Newton iterations depends on the accuracy of the interval extension. Thus, a more accurate and more expensive interval extension reduces the number of Newton iterations, but increases the cost of each iteration.

First, we give a skeleton algorithm for Newton's method. In the algorithm, F' denotes an interval extension of f'.

Algorithm 3.6.1 (NEWTON) [Newton Method for Univariate Polynomials]

In: $f(x) \in \mathbb{F}[x]$,
$A \in \mathbb{IF}$,
$\varepsilon \in \mathbb{F}$, $\varepsilon > 0$.

Out: $\mathcal{Z} \in \mathcal{P}(\mathbb{IF})$ such that
- for every $X \in \mathcal{Z}$ it holds that $\mathrm{w}(X) \leq \varepsilon$ or $\mathrm{w}(X)$ is so small that X cannot be bisected,
- if $f(x) = 0$ for some $x \in A$ then $x \in \bigcup \mathcal{Z}$.

(1) [Initialize.]
$\mathcal{S} \longleftarrow$ empty stack.
$\mathcal{Z} \longleftarrow \{\}$.
compute the coefficients of f'.
push A on $\mathcal{S}$.

(2) [Iterate.]
if $\mathcal{S}$ is empty, then return $\mathcal{Z}$.
pop X from $\mathcal{S}$.

(3) [Test if $\mathrm{w}(X)$ is too small.]
$c \longleftarrow \mathrm{MID}(X)$.
if $\mathrm{WIDTH}(X) \leq \varepsilon$ or $c \notin \mathrm{int}(X)$ then $\mathcal{Z} \longleftarrow \mathcal{Z} \cup \{X\}$, goto Step 2.

(4) [Compute interval extensions.]
$Y \longleftarrow \mathrm{HF}(f, [c])$.
$D \longleftarrow F'(X)$.

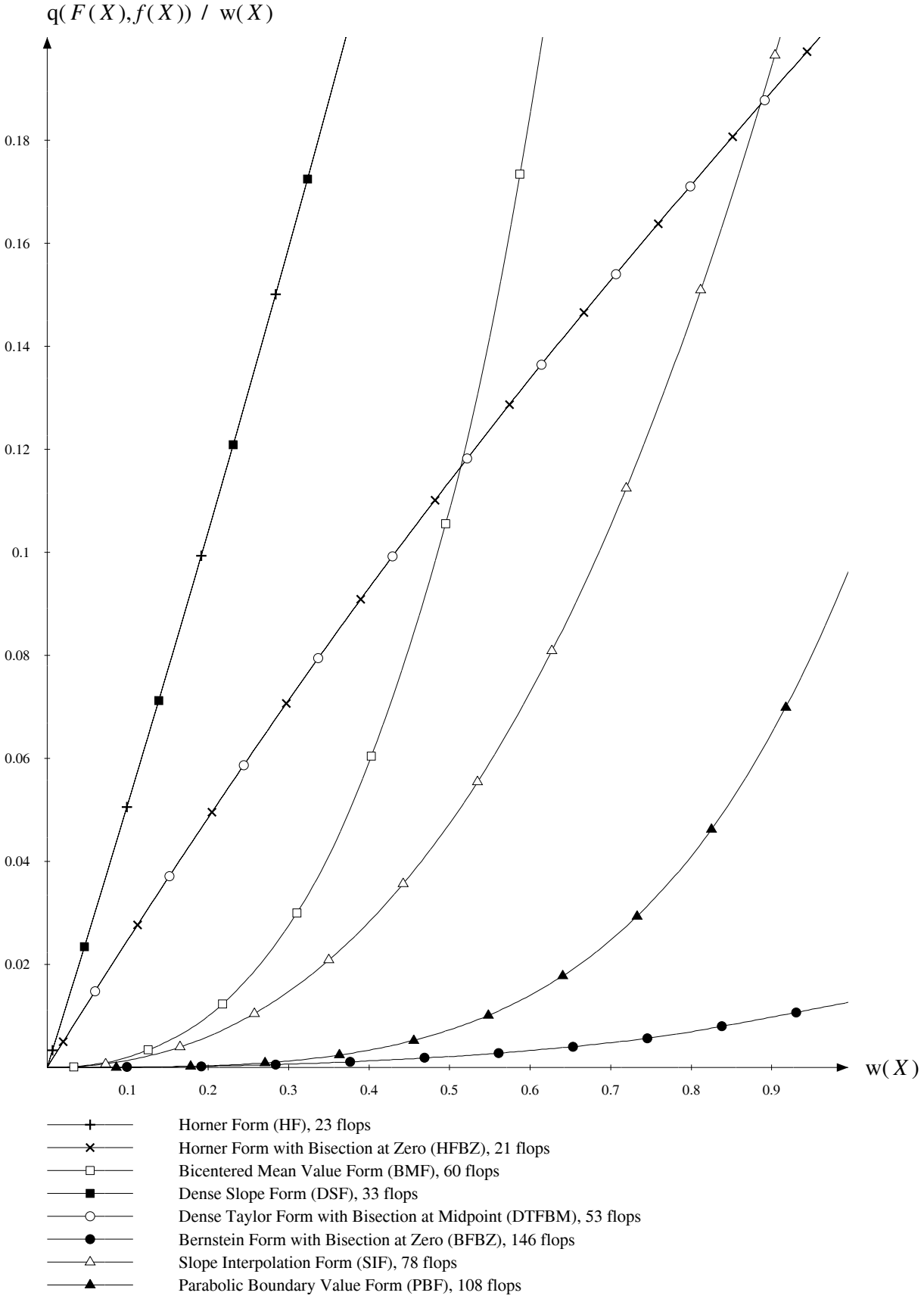


Figure 3.6.1: Average overestimation error for dense polynomials of degree 5, $\mathrm{mid}(X) \equiv 0$.

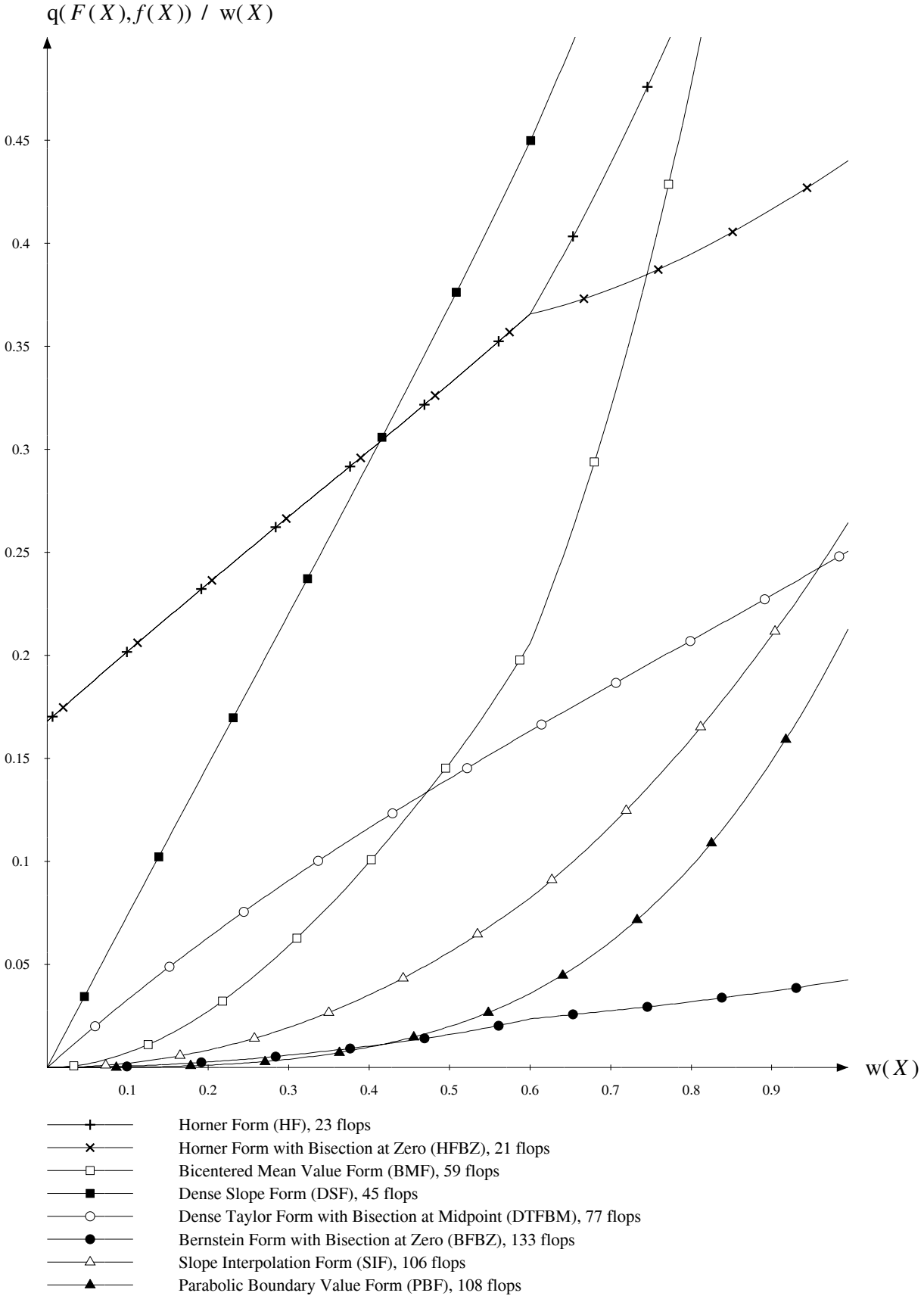


Figure 3.6.2: Average overestimation error for dense polynomials of degree 5, $\mathrm{mid}(X) \equiv 0.3$.

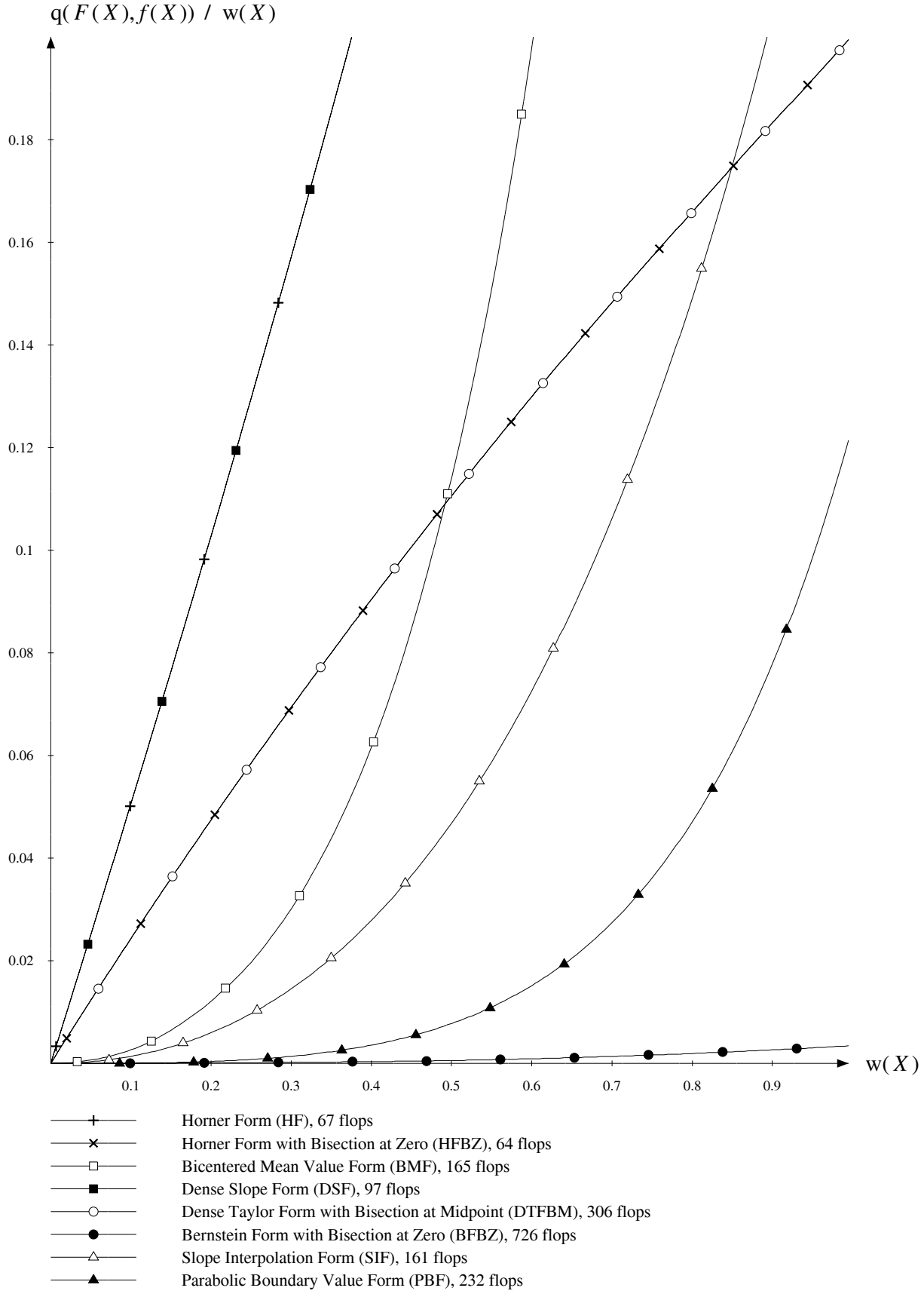


Figure 3.6.3: Average overestimation error for dense polynomials of degree 15, $\mathrm{mid}(X) \equiv 0.0$.

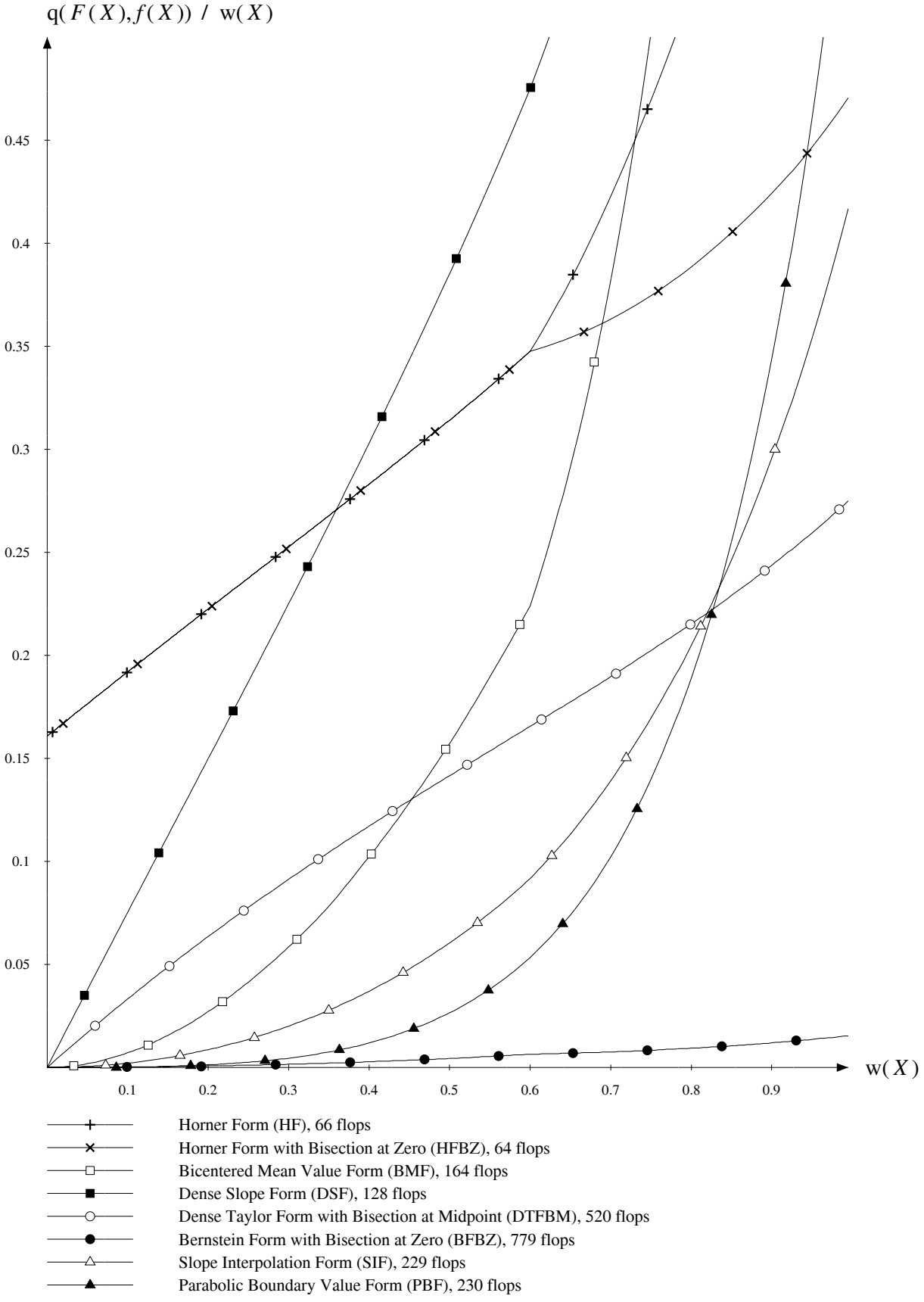


Figure 3.6.4: Average overestimation error for dense polynomials of degree 15, $\mathrm{mid}(X) \equiv 0.3$.

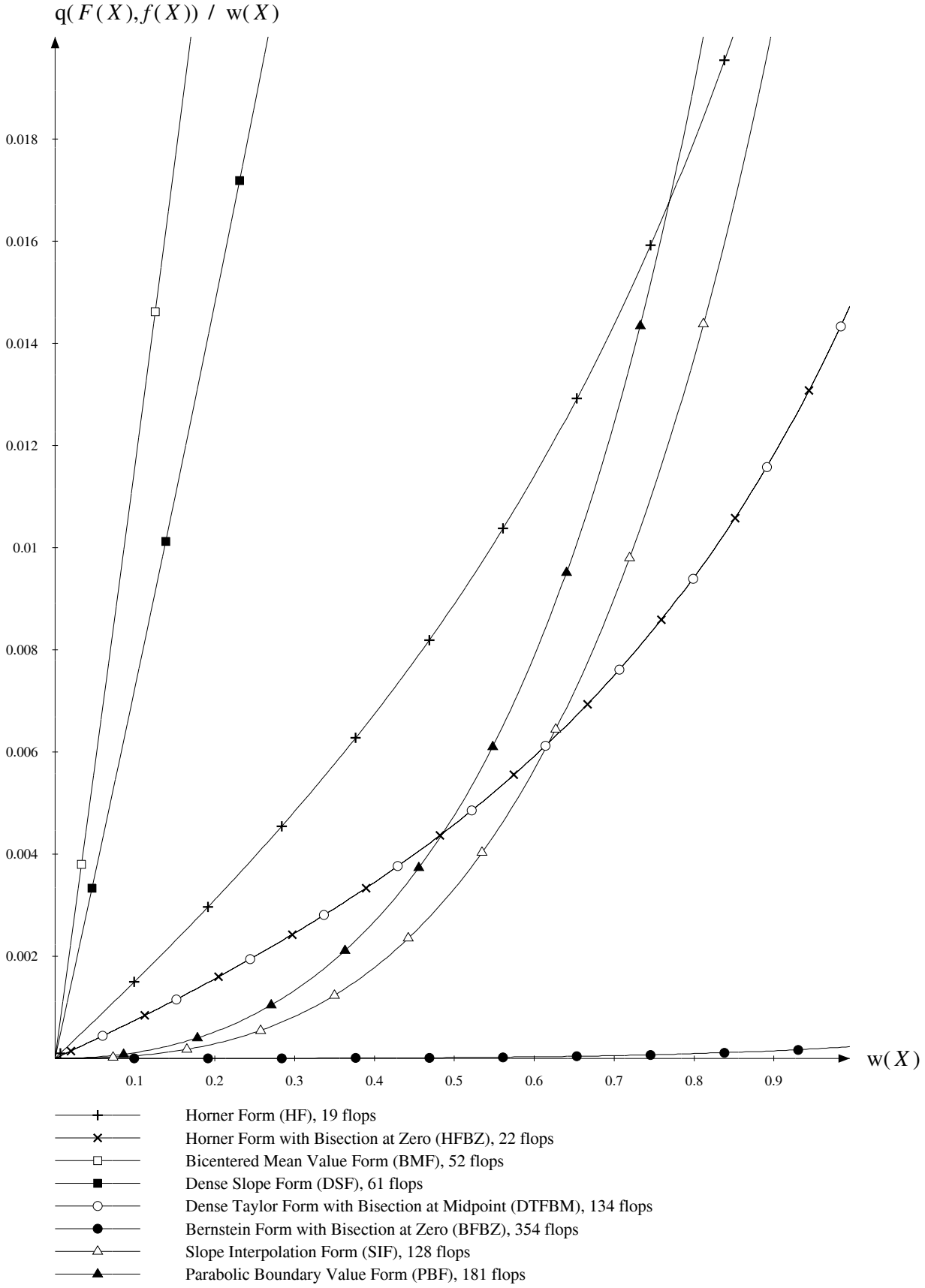


Figure 3.6.5: Average overestimation error for sparse polynomials of degree 10, $\mathrm{mid}(X) \equiv 0.0$.

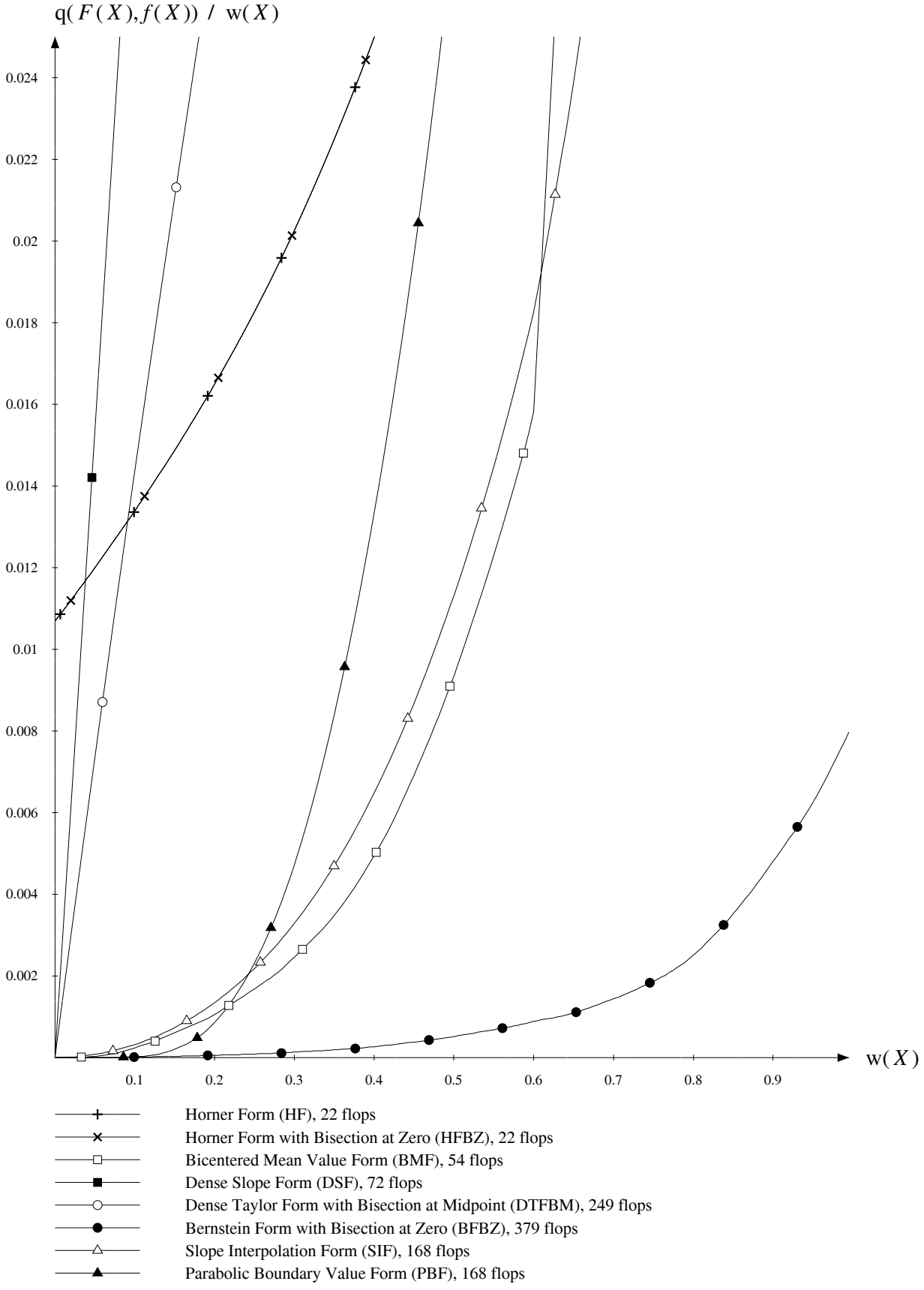


Figure 3.6.6: Average overestimation error for sparse polynomials of degree 10, $\mathrm{mid}(X) \equiv 0.3$.

(5) [Newton step if $0 \notin D$.]
if $0 \notin D$
$X' \longleftarrow (c - Y/D) \cap X$.
if $X' = \emptyset$ then goto Step 2.
if $X' \neq X$ then push X' on $\mathcal{S}$ and goto Step 2.
push $[\underline{X}, c]$ and $[c, \overline{X}]$ on $\mathcal{S}$ and goto Step 2.

(6) [Split X if $0 \in D$.]
$X_1, X_2 \longleftarrow \mathrm{GDIV}(-Y, D, X - c)$
if $X_1 \neq \emptyset$ then $X_1 \longleftarrow (X_1 + c) \cap X$.
if $X_2 \neq \emptyset$ then $X_2 \longleftarrow (X_2 + c) \cap X$.
if $X_1 \neq \emptyset$ and $X_2 \neq \emptyset$ and $X_1 \cap X_2 = \emptyset$.
if $X_1 \neq X$ and $X_2 \neq X$ then push X_1, X_2 on $\mathcal{S}$ and goto Step 2.
push $[\underline{X}, c]$ and $[c, \overline{X}]$ on $\mathcal{S}$ and goto Step 2.
else
$X' \longleftarrow X_1 \cup X_2$.
if $X' = \emptyset$ goto Step 2.
if $X' = X$ then push $[\underline{X}, c]$ and $[c, \overline{X}]$ on $\mathcal{S}$ and goto Step 2.
push X' on $\mathcal{S}$ and goto Step 2.

Algorithm 3.6.1 (NEWTON) is modified slightly if the bicentered mean value form or the dense slope form is used.

- During the evaluation of the bicentered mean value form of f', the coefficients of f'' have to be computed. This is done only once in Step 1.
- As a side product of the evaluation of the the dense linear factor form of f on X, we obtain an overestimation of $f(c)$ and an overestimation of the set of slopes of f between x and c for all $x \in X$ and $c = \mathrm{mid}(X)$. This set of slopes is used instead of $M_{f'}^{*(\mathrm{s})}(X)$. Thus, in Step 4 the computation of $f(c)$ and $M_{f'}^{*(\mathrm{s})}(X)$ is basically replaced by the evaluation of $M_f^{*(\mathrm{s})}(X)$.

Finally, one should mention, that the coefficients of f' in Step 1 have to be computed by rounded interval arithmetic. Hence, for the computation of D in Step 4, we have to overestimate the range of an interval polynomial. The necessary modifications of the algorithms are straight forward.

In the experiments we compute the roots of random polynomials in $[-1, 1]$. We consider three classes of polynomials:

- Dense random polynomials of degree 10. The coefficients are uniformly chosen from $[-1, 1]$ (Table 3.6.1, left chart).
- Sparse random polynomials of degree 10. Only 3 coefficients are non–zero and chosen uniformly from $[-1, 1]$. Note that these polynomials have a multiple root at zero if the constant coefficient is zero (Table 3.6.1, middle chart).
- Dense random polynomials of degree 10 with at least 5 roots in $[0.5, 0.6]$. The polynomials are generated as follows: 5 points are randomly chosen from $[0.5, 0.6]$ and the unique polynomial of degree 5 is computed which vanishes on these points. The polynomial is multiplied by a random polynomial of degree 5, whose coefficients are uniformly chosen from $[-1, 1]$ (Table 3.6.1, right chart).

In Table 3.6.1 we report the number of floating point operations and iterations of Algorithm 3.6.1 (NEWTON), averaged over 1000 random input polynomials. The accuracy is $\varepsilon = 10^{-6}$ in all cases. The results can be summarized and explained as follows.

- For dense random polynomials (left chart) the dense slope form ($M_{f'}^{*(\mathrm{s})}$) is best. This is mainly due to the modification described above, i.e. instead of overestimating $f'(X)$ in Newton's method,

Dense

	Kflops	Iterations
$\check{H}_{f'}$	0.95	10.83
$\ddot{M}_{f'}$	1.61	12.08
$M_{f'}^{*(\mathrm{s})}$	0.62	7.53
$\check{T}_{f'}^{*}$	2.67	10.68
$\check{B}_{f'}^{(k)}$	3.19	8.41
$\ddot{I}_{f'}^{(\mathrm{s})}$	1.91	9.56
$\tilde{I}_{f'}^{(\mathrm{s})}$	2.45	11.89

Sparse

	Kflops	Iterations
$\check{H}_{f'}$	1.32	30.21
$\ddot{M}_{f'}$	2.06	30.80
$M_{f'}^{*(\mathrm{s})}$	1.52	25.45
$\check{T}_{f'}^{*}$	14.99	63.56
$\check{B}_{f'}^{(k)}$	9.14	28.71
$\ddot{I}_{f'}^{(\mathrm{s})}$	7.06	39.98
$\tilde{I}_{f'}^{(\mathrm{s})}$	10.11	58.09

Root Cluster

	Kflops	Iterations
$\check{H}_{f'}$	1224.54	13957.25
$\ddot{M}_{f'}$	99.01	686.76
$M_{f'}^{*(\mathrm{s})}$	564.09	6596.54
$\check{T}_{f'}^{*}$	19.53	76.82
$\check{B}_{f'}^{(k)}$	20.68	54.57
$\ddot{I}_{f'}^{(\mathrm{s})}$	19.07	93.67
$\tilde{I}_{f'}^{(\mathrm{s})}$	18.91	91.84

Table 3.6.1: Newton's method for finding the roots of f in $[-1,1]$ in dependence of the chosen interval extension for overestimating $f'(X)$. In each chart, the average of 1000 random polynomials of degree 10 is reported. In the left chart, the polynomials are dense, in the middle chart, the polynomials are sparse with only three non–zero coefficients, and in the right chart, the polynomials have a cluster of at least 5 roots in $[0.5, 0.6]$.

merely an overestimation of the set of slopes of f between $\mathrm{mid}(X)$ and arbitrary points in X is used. This explains also the small number of iterations of the slope form.

- For sparse polynomials (middle chart), the Horner form with bisection at zero ($\check{H}_{f'}$) is comparable to $M_{f'}^{*(\mathrm{s})}$. The reason is that $\check{H}_{f'}$ exploits the sparsity of f' for improving both efficiency and accuracy, whereas the accuracy of $M_{f}^{*(\mathrm{s})}$ is not affected, and the costs are reduced by a smaller amount as compared to $\check{H}_{f'}$.
- The high accuracy of more expensive interval extensions pays only in the presence of root clusters (right chart). Here, most iterations are done when $\mathrm{w}(X)$ is very small. This explains the bad performance of the only linearly convergent Horner form. Further, it explains why the bicentered mean value form ($\ddot{M}_{f'}$) is now much better than $M_{f'}^{*(\mathrm{s})}$: If $\mathrm{w}(X)$ is small, then the chances are high, that $0 \notin H_{f''}(X)$ and $\ddot{M}_{f'}$ is exact. The smallest number of iterations is achieved by the Bernstein form with bisection at zero ($\check{B}_{f'}^{(k)}$), but as it is more expensive than the interpolation forms and the dense Taylor form with bisection ($\check{T}_{f'}^{*}$), the methods perform about equally well. It is interesting to note that $\check{T}_{f'}^{*}$, which is relatively inefficient for dense and sparse random polynomials, is one of the best in the presence of root clusters.

From our experimental observations we conclude that it is very important to choose an appropriate interval extension for overestimating $f'(X)$ in Newton's method. It seems that the following strategy can be recommended:

- If $\mathrm{w}(X)$ is large and f is dense, then use the dense linear factor form.
- If $\mathrm{w}(X)$ is large and f is sparse, then use the Horner form with bisection at zero.
- If $\mathrm{w}(X)$ is small and $\mathrm{mid}(X) \approx 0$ then use the Horner form with bisection at zero.
- If $\mathrm{w}(X)$ is small and $\mathrm{mid}(X) \not\approx 0$ then use the Bernstein form, the Taylor form with bisection or an interpolation form.

3.6.3 Global Optimization

In this section we compare interval extensions at the problem of finding the global minimum with given accuracy of a univariate polynomial f in an interval A. The problem is solved by an iterative algorithm,

which evaluates an interval extension of f on a subinterval of A in each iteration. As in the case of Newton's method, the number of iterations and the cost of each iteration depend on the chosen interval extension.

Global optimization is one of the standard applications of interval arithmetic and was studied thoroughly in the literature, see for example [Skelboe, 1974], [Moore, 1976], [Ichida and Fujii, 1979], [Hansen, 1979], [Hansen, 1980], [Hansen and Sengupta, 1980], [Hansen, 1988], [Ratschek and Rokne, 1988], [Hansen, 1992], [Hammer *et al.*, 1993] and many others.

In order to facilitate the interpretation of the experimental results, we use a very simple optimization algorithm and omit all kinds of accelerating devices. In the experiments below, we compute the global minimum of random polynomials in $[-1, 1]$. We consider the same classes of random polynomials as in the previous section on Newton's method, but multiply each polynomial by x^2-1 Without this multiplication, the probability that f achieves its minimum at the boundary of X is very high, and the experimental results might not be meaningful.

The comparison is done using the following optimization algorithm. As usual, F denotes an interval extension of f. The algorithm keeps a list $\mathcal{L}$ of intervals which contains initially only the input interval A. In each iteration, we chose the interval X of $\mathcal{L}$ for which $\underline{F(X)}$ is smallest, bisect it in the middle and add both halves to $\mathcal{L}$. Obviously, $F(X)$ encloses the global minimum of f in A. The algorithm terminates if $\mathrm{w}(F(X)) < \varepsilon$, where ε is the desired accuracy of the result, or if X cannot be bisected any more because of finite precision arithmetic. For efficiency reason it is best to store pairs $(X, F(X))$ in $\mathcal{L}$ and sort them according to increasing $\underline{F(X)}$.

Algorithm 3.6.2 (MIN) [Global Minimum of Univariate Polynomials]

In: $f(x) \in \mathbb{F}[x]$,
$A \in \mathbb{IF}$,
$\varepsilon \in \mathbb{F}$, $\varepsilon > 0$.

Out: $Y \in \mathbb{IF}$ such that
- $\min_{x \in A} f(x) \in Y$
- $\mathrm{w}(Y) \leq \varepsilon$ if the precision of the floating point numbers is sufficiently high.

(1) [Initialize.]
$\mathcal{L} \longleftarrow \langle (A, F(A)) \rangle$.

(2) [Iterate.]
pop (X, Y) from $\mathcal{L}$.

(3) [Terminate.]
if $\mathrm{WIDTH}(Y) \leq \varepsilon$ then return Y.
$c \longleftarrow \mathrm{MID}(X)$.
if $c \notin \mathrm{int}(X)$ then return Y.

(4) [Bisect.]
$X_1 \longleftarrow [\underline{X}, c]$.
$X_2 \longleftarrow [c, \overline{X}]$.

(5) [Insert.]
insert $(X_1, F(X_1))$ and $(X_2, F(X_2))$ in $\mathcal{L}$ such that the elements (X, Y) in $\mathcal{L}$ are sorted with increasing $\underline{Y}$.
goto Step 2.

Algorithm 3.6.2 is modified slightly in dependence of the chosen interval extension F:

- The bicentered mean value form requires the computation of the coefficients of f'. Apparently, this should be done only once in Step 1. Further, instead of $\ddot{M}_f(X)$, the smaller interval

$$[\underline{M_f(X, c^{\downarrow})}, H_f(c^{\downarrow})]$$

Dense

	Kflops	Iterations
$\check{H}_f$	10.57	107.90
$\ddot{M}_f$	2.45	11.98
$M_f^{*(\mathrm{s})}$	3.07	15.83
$\check{T}_f^*$	7.61	11.68
$B_f^{(k)}$	5.25	5.53
$\ddot{I}_f^{(\mathrm{s})}$	3.06	8.79
$\tilde{I}_f^{(\mathrm{s})}$	2.20	6.40

Sparse

	Kflops	Iterations
$\check{H}_f$	5.78	86.33
$\ddot{M}_f$	1.64	11.56
$M_f^{*(\mathrm{s})}$	2.56	15.50
$\check{T}_f^*$	6.95	10.75
$B_f^{(k)}$	5.17	5.44
$\ddot{I}_f^{(\mathrm{s})}$	2.99	8.61
$\tilde{I}_f^{(\mathrm{s})}$	2.15	6.25

Root Cluster

	Kflops	Iterations
$\check{H}_f$	73.22	740.69
$\ddot{M}_f$	6.31	29.44
$M_f^{*(\mathrm{s})}$	4.22	21.32
$\check{T}_f^*$	7.33	11.26
$B_f^{(k)}$	6.07	6.22
$\ddot{I}_f^{(\mathrm{s})}$	3.25	9.24
$\tilde{I}_f^{(\mathrm{s})}$	2.92	8.17

Table 3.6.2: Computation of the global minimum of f in $[-1,1]$ in dependence of the chosen interval extension for overestimating $f(X)$. In each chart, the average of 1000 random polynomials of degree 12 is reported. In the left chart, the polynomials are dense, in the middle chart, the polynomials are sparse with at most 6 non–zero coefficients, and in the right chart, the polynomials have a cluster of at least 5 roots in $[0.5, 0.6]$. Every polynomial vanishes at -1 and 1.

is used in Step 1 and Step 5. This saves the computation of $c^\uparrow$ and $H_f(c^\uparrow)$, see Section 3.2.2.

- As a side product during the evaluation of $M_f^{*(\mathrm{s})}(X)$, $\check{T}_f^*(X)$ and $\ddot{I}_f^{(\mathrm{s})}(X)$ we obtain an overestimation Y of $f(\mathrm{mid}(X))$. Thus, in Step 1 and Step 5 we use the smaller interval $[\underline{F(X)}, \overline{Y}]$ instead of $F(X)$, where F is $M_f^{*(\mathrm{s})}$, $\check{T}_f^*$ and $\ddot{I}_f^{(\mathrm{s})}$ respectively.
- Similarly, during the evaluation of $B_f^{(k)}(X)$ and $\tilde{I}_f^{(\mathrm{s})}(X)$ we obtain overestimations L, R of $f(\underline{X})$ respectively $f(\overline{X})$. Thus, in Step 1 and Step 5 we use the smaller interval $[\underline{F(X)}, \min\{\overline{L}, \overline{R}\}]$ instead of $F(X)$, where F is $B_f^{(k)}$ and $\tilde{I}_f^{(\mathrm{s})}$ respectively.

In Table 3.6.2 we report the number of floating point operations and iterations of Algorithm 3.6.2 (MIN), averaged over 1000 random input polynomials. The accuracy is $\varepsilon = 10^{-3}$ in all cases. It does not seem reasonable to choose a higher accuracy unless quadratically convergent accelerators like Newton's method are used.

The results can be summarized and explained as follows.

- For dense random polynomials (left chart) the parabolic boundary value form $\tilde{I}_f^{(\mathrm{s})}$ is best, followed by the bicentered mean value form $\ddot{M}_f$. The good performance of $\ddot{M}_f$ is mainly due to the modification described above. The Horner form with bisection at zero $\check{H}_f$ is significantly worse than the other methods. Whereas all methods could be improved because an upper bound of f at the center or the boundary of X was available, this was not possible for the Horner form.
- For sparse random polynomials (middle chart), $\ddot{M}_f$ is better than $\tilde{I}_f^{(\mathrm{s})}$, because the former exploits sparsity to improve both efficiency and accuracy, whereas the latter does not. This explains also, why $\check{H}_f$ is much better compared to the dense case, whereas all other methods remain basically the same.
- In the presence of a root cluster, interpolation forms are best. A closer analysis shows that this is only true if the global minimum is near the root cluster. In this case many iterations are executed when the width of X is small, and the cubic convergence order of interpolation forms pays.

Note that the smallest number of iterations is achieved by the Bernstein form in all cases. However, as the evaluation of the Bernstein form is relatively expensive, its overall performance is not good. Further, the experiments show that $\check{H}_f$ is not a good choice for global optimization problems. Not only the number of floating point operations is large, but each iteration requires insertion of elements in a sorted list, which causes additional overhead.

Chapter 4

Inclusion of the Range of Multivariate Polynomials

In this chapter we consider the problem of finding an overestimation of the range of multivariate polynomials over an interval vector. Most of the methods for univariate polynomials presented in Chapter 3 can easily be generalized. However, for efficiency reasons, it is very important to exploit sparsity in the multivariate case. Hence, we restrict our considerations to methods which fulfill this condition.

In the following let $f(x_1, x_2, \ldots, x_n) \in \mathbb{R}[x_1, \ldots, x_n]$

$$f(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{m} a_i \, x_1^{d_{i,1}} x_2^{d_{i,2}} \cdots x_n^{d_{i,n}}$$

be a multivariate real polynomial in $n \geq 0$ variables, such that $a_i \neq 0$ for all i and

$$(d_{i,1}, d_{i,2}, \ldots, d_{i,n}) \succ (d_{j,1}, d_{j,2}, \ldots, d_{j,n}) \quad \text{for all} \quad 1 \leq i < j \leq m,$$

where $\succ$ means lexicographically greater. For the zero polynomial we define $m = 1$ and $a_1 = 0$. The term

$$a_i \, x_1^{d_{i,1}} x_2^{d_{i,2}} \cdots x_n^{d_{i,n}}$$

is called i-th monomial of f.

4.1 Horner Form

The Horner form of multivariate polynomials is defined recursively: If $n = 0$ then $f = a_1$ is a real number and the Horner form of f is a_1. Otherwise, we consider $f(x_1, \ldots, x_n)$ as a univariate polynomial in x_1 with coefficients in $\mathbb{R}[x_2, \ldots, x_n]$. By evaluating the Horner form of the coefficients recursively, we obtain a univariate interval polynomial $F(x_1) \in \mathbb{IR}[x_1]$, and the Horner form of the multivariate polynomial f is the Horner form of the univariate interval polynomial F (see Definition 3.1.48). Before giving a formal definition of the multivariate Horner form, we introduce some basic notions.

Definition 4.1.1 (Coefficient) *The coefficient of x_1^d in f is defined as*

$$\operatorname{coef}(f, d) = \sum_{\substack{i=1 \\ d_{i,1}=d}}^{m} a_i \, x_2^{d_{i,2}} x_3^{d_{i,3}} \cdots x_n^{d_{i,n}} \in \mathbb{R}[x_2, \ldots, x_n]. \; \square$$

Definition 4.1.2 (Degree) *The degree of f is defined as*

$$\deg(f) = \begin{cases} d_{1,1} & \text{if } n > 0 \\ 0 & \text{else. } \square \end{cases}$$

Definition 4.1.3 (Multivariate Horner Form) *The multivariate Horner form $H_f : \mathbb{IR}^n \to \mathbb{IR}$ of f is defined recursively as*

$$H_f(X_1, \ldots, X_n) = \begin{cases} a_1 & \text{if } n = 0 \\ H_F(X_1) & \text{else,} \end{cases}$$

where

$$F(x_1) = \sum_{i=0}^{\deg(f)} H_{\mathrm{coef}(f,i)}(X_2, \ldots, X_n)\, x_1^i \ \in \mathbb{IR}[x_1]. \ \square$$

We are using the same symbol H for the univariate and the multivariate Horner form, because the forms coincide if the number of variables is 1.

Theorem 4.1.4 (Inclusion Monotonicity) *H_f is an inclusion monotone interval extension of f.* $\square$

Proof. Follows from Corollary 1.3.5 and Theorem 1.3.22. $\square$

Theorem 4.1.5 (Convergence) *H_f converges linearly to f.* $\square$

Proof. H_f is an interval extensions of f (Theorem 4.1.4) and Lipschitz (Corollary 1.3.19). Hence H_f converges linearly to f (Theorem 1.3.24). $\square$

From Definition 4.1.3 we obtain the following recursive algorithm for evaluating the multivariate Horner form.

Algorithm 4.1.6 (HFM) [Multivariate Horner Form]

In: $f(x_1, \ldots, x_n) = \sum_{i=1}^{m} a_i\, x_1^{d_{i,1}} x_2^{d_{i,2}} \cdots x_n^{d_{i,n}} \in \mathbb{F}[x_1, \ldots, x_n]$,
$\boldsymbol{X} \in \mathbb{IF}^n$.

Out: $\mathrm{HFM}(f, \boldsymbol{X}) \in \mathbb{IF}$, $\mathrm{HFM}(f, \boldsymbol{X}) \supseteq H_f(\boldsymbol{X})$.

(1) [Recursion base.]
if $n = 0$ return a_1.

(2) [Coefficients of univariate interval polynomial.]
for $i = 0 \ldots \deg(f)$
$g_i(x_2, \ldots, x_n) \longleftarrow \mathrm{coef}(f, i)$.
$A_i \longleftarrow \mathrm{HFM}(g_i(x_2, \ldots, x_n), X_2, \ldots, X_n)$.

(3) [Evaluate univariate interval polynomial.]
$F(x_1) \longleftarrow \sum_{i=0}^{\deg(f)} A_i x_1^i$.
return $\mathrm{HFI}(F(x_1), X_1)$.

Theorem 4.1.7 (Complexity) *Algorithm 4.1.6 (*HFM*) costs at most*

nm	*interval power computations,*
nm	*interval multiplications and*
$2m - 2$	*number additions.* $\square$

Proof. Let $\pi(m, n)$ be the number of interval power computations and interval multiplications, and let $\alpha(m, n)$ be the number of interval additions of Algorithm 4.1.6 (HFM). By induction on n we show that

$$\pi(m, n) \le mn, \quad \alpha(m, n) \le 2m - 2 \tag{4.1.1}$$

for all $m \ge 1$ and $n \ge 0$. Obviously (4.1.1) holds for $n = 0$. Let n arbitrary but fixed and assume

$$\pi(m, n-1) \le m(n-1), \quad \alpha(m, n-1) \le 2m - 2$$

for all m. Let k be the number of different powers of x_1 occuring in f and let m_i be the number of monomials in the i-th coefficient, $i = 1, \ldots, k$. Step 2 costs $\sum_{i=1}^{k} \pi(m_i, n-1)$ power computations and multiplications and $\sum_{i=1}^{k} \alpha(m_i, n-1)$ additions. Step 3 costs at most k power computations and multiplications and $k-1$ additions. Hence, by induction on n we obtain

$$\begin{aligned}
\pi(m,n) &\leq k + \sum_{i=1}^{k} \pi(m_i, n-1) \\
&\leq k + \sum_{i=1}^{k} m_i(n-1) \\
&= m(n-1) + k \\
&\leq nm \\
\alpha(m,n) &\leq k - 1 + \sum_{i=1}^{k} \alpha(m_i, n-1) \\
&\leq k - 1 + \sum_{i=1}^{k} (m_i - 1) \\
&= m - 1. \ \square
\end{aligned}$$

Theorem 4.1.8 *If $0 \notin \mathrm{int}(X_i)$ for all i, then Algorithm 4.1.6 (*HFM*) costs*

$$\begin{aligned}
nm &\quad \textit{interval power computations,} \\
2nm &\quad \textit{number multiplications and} \\
2m-2 &\quad \textit{number additions.} \ \square
\end{aligned}$$

4.1.1 Nested Form

A simple way to obtain interval extensions F of f is as follows:

- Find an arithmetic expression e for f.
- Let $F(\boldsymbol{X})$ be the interval which is obtained by evaluating e on $\boldsymbol{X}$ using interval arithmetic.

The Horner form is obtained in this way, where the expression e has a particular shape. For example, let

$$f(x_1, x_2) = x_1^2 x_2^2 + x_1 x_2^2 + x_1 x_2.$$

Then

$$\mathsf{e} = x_1^2 x_2^2 + x_1(x_2 + 1)x_2$$

is an expression for f, and the Horner form of f is the interval evaluation of e. Another expression for f is

$$\mathsf{e}' = x_1 x_2 (x_2(x_1 + 1) + 1).$$

Usually, the interval evaluation of e and e' yield different results. Further, it depends on the value of $\boldsymbol{X}$ which expression gives a tighter interval. For example let

$$\boldsymbol{X} = ([0,1], [-1,1])$$

then

$$\mathsf{e}(\boldsymbol{X}) = [-2,3] \subset [-3,3] = \mathsf{e}'(\boldsymbol{X}),$$

whereas if

$$\boldsymbol{X} = ([-1,0], [-1,1])$$

then

$$\mathsf{e}(\boldsymbol{X}) = [-2,3] \supset [-2,2] = \mathsf{e}'(\boldsymbol{X}).$$

It is non–trivial to find an expression for given f and $\boldsymbol{X}$ which yields an interval with smallest width efficiently. In the literature on optimizing compilers a similar problem is treated. For a given expression an equivalent expression is searched, which allows evaluation with a minimal number of machine instructions, see e.g. [Sethi and Ullman, 1970], [Ullman, 1973], [Aho *et al.*, 1977]. However, the arithmetic operations are not assumed to be commutative or associative but mainly the register usage is optimized. Thus, the results cannot be applied here. Our experience is that searching an expression which minimizes the overestimation error is very expensive and other interval extensions with comparable costs give usually tighter inclusions. Thus, instead of searching an optimal expression we are content with a suboptimum. In applications it is often the case that the same polynomial has to be evaluated over many different intervals. As the determination of a suboptimal expression can be costly, it should be done only once and not each time the polynomial is evaluated. Therefore, we are interested in an expression which gives tight inclusions for arbitrary input intervals.

Having in mind the subdistributivity law of interval arithmetic, one might be tempted to say that e' is always better than e. However, this is wrong because the interval power function gives tighter inclusions than iterated multiplication. On the other hand, there are many cases where an application of the distributivity rule

$$Ax^a + BX^b \longrightarrow (AX^{a-c} + BX^{b-c})X^c, \quad 0 < c \leq \min(a,b)$$

leads to better expressions. The following theorem is new.

Theorem 4.1.9 *Let $A, B, X \in \mathbb{IR}$, let $a, b, c \in \mathbb{N}$ such that $0 < c < \min(a,b)$. Let*

$$\begin{aligned} Y_1 &= AX^a + BX^b \\ Y_2 &= (AX^{a-c} + BX^{b-c})X^c \\ Y_3 &= (AX^{a-\min(a,b)} + BX^{b-\min(a,b)})X^{\min(a,b)}. \end{aligned}$$

(i) *If $0 \notin \mathrm{int}(X)$ then $Y_3 \subseteq Y_2 \subseteq Y_1$.*

(ii) *If a, b are both even then $Y_3 \subseteq Y_1$.*

(iii) *If a, b are both odd and $Y_2 \subset Y_1$ then $Y_3 \subset Y_2$.* □

Intuitively (iii) can be read as "if factoring out something gives an improvement, then factoring out as much as possible is best".

Proof.

(i) Assume $0 \notin \mathrm{int}(X)$. Then the d-th power of X is equal to d times the product of X with itself for all $d \in \mathbb{N}$ and $Y_3 \subseteq Y_2 \subseteq Y_1$ follows from the subdistributivity law (Theorem 1.1.3).

(ii) Assume a, b are even and without loss of generality $b > a$. Then $b - a$ is even and

$$X^b = X^{b-a}X^a$$

and thus

$$\begin{aligned} Y_3 &= (A + BX^{b-a})X^a \\ &\subseteq AX^a + BX^{b-a}X^a \\ &= AX^a + BX^b \\ &= Y_1. \end{aligned}$$

(iii) Assume a, b are odd, $Y_2 \subset Y_1$ and without loss of generality $b > a$. As A, B occur only once in each expression of Theorem 4.1.9, it suffices to consider the case when $A = \alpha \in \mathbb{R}$ and $B = \beta \in \mathbb{R}$. Further, the case $\alpha < 0$ can be reduced to the case $\alpha \geq 0$, hence we may assume $\alpha \geq 0$. If $0 \notin \mathrm{int}(X)$ then (iii) follows from (i), hence in the following assume $0 \in \mathrm{int}(X)$. The case $\overline{X} < -\underline{X}$ can be reduced to the case $\overline{X} \geq -\underline{X}$ by replacing X by $-X$, hence we may assume $\overline{X} \geq -\underline{X}$. We distinguish two cases depending on the sign of β.

- Assume $\beta \geq 0$. Then

$$\begin{aligned}\overline{Y}_2 &= (\alpha\overline{X}^{a-c} + \beta\overline{X}^{b-c})\overline{X}^c \\ &= (\alpha + \beta\overline{X}^{b-a})\overline{X}^a \\ &= \overline{Y}_3.\end{aligned}$$

  - If c is odd then $a-c$, $b-c$ are even and

$$\begin{aligned}\underline{Y}_2 &= (\alpha\overline{X}^{a-c} + \beta\overline{X}^{b-c})\underline{X}^c \\ &\leq (\alpha + \beta\overline{X}^{b-a})\underline{X}^a \\ &= \underline{Y}_3.\end{aligned}$$

  - If c is even then $a-c$, $b-c$ are odd and

$$\begin{aligned}\underline{Y}_2 &= (\alpha\underline{X}^{a-c} + \beta\underline{X}^{b-c})\overline{X}^c \\ &\leq (\alpha\underline{X}^a + \beta\underline{X}^b) \\ &= \underline{Y}_1,\end{aligned}$$

    i.e. the assumption $Y_2 \subset Y_1$ is not satisfied.

- Assume $\beta < 0$.
  - If c is odd then $a-c$, $b-c$ are even and

$$\begin{aligned}\overline{Y}_2 &= \max\{(\alpha\overline{X}^{a-c} + \beta 0)\overline{X}^c, (\alpha 0 + \beta\overline{X}^{b-c})\underline{X}^c\} \\ &= \max\{\alpha\overline{X}^a, \beta\overline{X}^{b-c}\underline{X}^c\} \\ &\geq \max\{\alpha\overline{X}^a, \beta\overline{X}^{b-a}\underline{X}^a\} \\ &= \overline{Y}_3, \\ \underline{Y}_2 &= \min\{(\alpha\overline{X}^{a-c} + \beta 0)\underline{X}^c, (\alpha 0 + \beta\overline{X}^{b-c})\overline{X}^c\} \\ &= \min\{\alpha\overline{X}^{a-c}\underline{X}^c, \beta\overline{X}^b\} \\ &\leq \min\{\alpha\underline{X}^a, \beta\overline{X}^{b-a}\underline{X}^a\} \\ &= \underline{Y}_3.\end{aligned}$$

  - If c is even then $a-c$, $b-c$ is odd and

$$\begin{aligned}\overline{Y}_2 &= (\alpha\overline{X}^{a-c} + \beta\underline{X}^{b-c})\overline{X}^c \\ &\geq (\alpha\overline{X}^a + \beta\underline{X}^b) \\ &= \overline{Y}_1, \\ \underline{Y}_2 &= (\alpha\underline{X}^{a-c} + \beta\overline{X}^{b-c})\overline{X}^c \\ &\leq (\alpha\underline{X}^a + \beta\overline{X}^b) \\ &= \underline{Y}_1,\end{aligned}$$

    i.e. the assumption $Y_2 \subset Y_1$ is not satisfied.

Thus, it seems reasonable to search for expressions which are highly nested. Although intuitively it is clear what is meant by nested, this notion does not define a unique expression. For example, let

$$f(x_1, x_2) = x_1x_2 + x_1 + x_2.$$

Then

$$x_1(x_2 + 1) + x_2, \quad x_2(x_1 + 1) + x_1$$

are two locally optimal expressions for f. This means that even after rewriting the expressions using the commutativity or associativity law, the distributivity law cannot be applied in the direction of higher nesting.

We give a simple algorithm which computes such a locally optimal expression for a given polynomial.

Algorithm 4.1.10 (NEP) [Nested Expression for Polynomial]

In: $\mathsf{e} = \sum_{i=1}^{m} a_i\, x_1^{d_{i,1}} x_2^{d_{i,2}} \cdots x_n^{d_{i,n}}$, an expression.

Out: NEP(e), an expression equivalent to e up to commutativity, associativity and distributivity.

(1) [Initialize.]
$\mathcal{L} \longleftarrow \mathrm{List}(a_i\, x_1^{d_{i,1}} x_2^{d_{i,2}} \cdots x_n^{d_{i,n}} \mid i = 1, \ldots, m)$.

(2) [Iterate.]
The elements of $\mathcal{L}$ are expressions of the form $\mathsf{a}\, x_1^{d_1} x_2^{d_2}, \cdots x_n^{d_n}$ where a is a (compound) expression.
while $\mathrm{Length}(\mathcal{L}) > 1$

(2.1) [Search.]
Choose $\mathsf{e}' = \mathsf{a}' x_1^{d_1'} x_2^{d_2'}, \cdots x_n^{d_n'}$ and $\mathsf{e}'' = \mathsf{a}'' x_1^{d_1''} x_2^{d_2''}, \cdots x_n^{d_n''}$ from $\mathcal{L}$ such that $\sum_{i=1}^{n} \min(d_i', d_i'')$ is maximal. Remove e' and e'' from $\mathcal{L}$.

(2.3) [Combine.]
for $i = 1, \ldots n$ do $d_i \longleftarrow \min(d_i', d_i'')$.
$\mathsf{e} \longleftarrow (\mathsf{a}' \prod_{\substack{i=1 \\ d_i' \neq d_i}}^{n} x_i^{d_i' - d_i} + \mathsf{a}' \prod_{\substack{i=1 \\ d_i'' \neq d_i}}^{n} x_i^{d_i'' - d_i}) x_1^{d_1} x_2^{d_2} \cdots x_n^{d_n}$.
Add e to $\mathcal{L}$.

(3) [Return.]
return $\mathrm{First}(\mathcal{L})$.

Note that Step 2.1 is non–deterministic. For example, if $\mathsf{e} = x_1 x_2 + x_1 + x_2$ then Algorithm 4.1.10 (NEP) may either return $x_1(x_2 + 1) + x_2$ or $x_2(x_1 + 1) + x_1$. In the sequel we assume that an arbitrary but fixed strategy is used.

Definition 4.1.11 (Nested Form) *The nested form $N_f : \mathbb{IR}^n \to \mathbb{IR}$ of f is the interval evaluation of the expression obtained by Algorithm 4.1.10* (NEP) *with input expression*

$$\sum_{i=1}^{m} a_i\, x_1^{d_{i,1}} x_2^{d_{i,2}} \cdots x_n^{d_{i,n}}. \quad \square$$

An algorithm for evaluating the nested form is now straight forward.

Algorithm 4.1.12 (NF) [Nested Form]

In: $f(x_1, \ldots, x_n) = \sum_{i=1}^{m} a_i\, x_1^{d_{i,1}} x_2^{d_{i,2}} \cdots x_n^{d_{i,n}} \in \mathbb{F}[x_1, \ldots, x_n]$,
$\boldsymbol{X} \in \mathbb{IF}^n$.

Out: $\mathrm{NF}(f, \boldsymbol{X}) \in \mathbb{IF}$, $\mathrm{NF}(f, \boldsymbol{X}) \supseteq N_f(\boldsymbol{X})$.

(1) [Nested expression.]
$\mathsf{e} \longleftarrow \mathrm{NF}(\sum_{i=1}^{m} a_i\, x_1^{d_{i,1}} x_2^{d_{i,2}} \cdots x_n^{d_{i,n}})$.

(2) [Evaluate.]
$Y \longleftarrow$ evaluation of e on $\boldsymbol{X}$ using floating point interval arithmetic.

(3) [Return.]
return Y.

Remark. In the univariate case the nested form and the Horner form are identical.

Remark. If N_f has to be computed repeatedly for different intervals $\boldsymbol{X}$ using Algorithm 4.1.12 (NF), then Step 1 has to be executed only once, i.e. e can be pre–computed.

Theorem 4.1.13 (Complexity) *Algorithm 4.1.12 (*NF*) costs*

$$\begin{array}{rl} nm & \textit{interval power computations,} \\ nm & \textit{interval multiplications and} \\ 2m-2 & \textit{number additions.} \ \square \end{array}$$

Proof. Step 1 involves merely symbolic and integer operations. For the cost of Step 2 we have to count the number of operation symbols in the expression `e`. Therefore, we count the number of operation symbols of all expressions in $\mathcal{L}$ of Algorithm 4.1.10 (NEP). Upon termination of this algorithm $\mathcal{L}$ contains the single element `e`. In Step 1, every element of $\mathcal{L}$ comprises n power computations and n multiplications, hence there are nm power computations and nm multiplications. In each iteration in Step 2 we remove in Step 2.1 the operations for $\mathtt{a}', \mathtt{a}''$, $2n$ power computations and $2n$ multiplications and add in Step 2.2 the operations for $\mathtt{a}', \mathtt{a}''$, at most $2n$ power computations, at most $2n$ multiplications and 1 addition. Hence, in each iteration in Step 2 we add 1 addition and, in the worst case, keep the number of power computations and multiplications constant. As $m-1$ iterations are performed, we add $m-1$ additions. Hence, upon termination there are at most nm power computations, nm multiplications and $m-1$ additions in $\mathcal{L}$. $\square$

4.1.2 Experimental Results

In this section we compare the Horner form H_f, the nested form N_f and the distributed form D_f which is defined as

$$D_f(\boldsymbol{X}) = \sum_{i=1}^{n} a_i \, X_1^{d_{i,1}} X_2^{d_{i,2}} \cdots X_n^{d_{i,n}}.$$

The distributed form is the evaluation of the "least nested" expression for f, hence D_f and N_f can be viewed as two extremes where H_f is in between.

As the computation of the exact range of a multivariate polynomial is expensive, we do not report overestimation errors. Instead, we compare the width of the nested and the dense form with the width of the Horner form. The costs are given in terms of arithmetic floating point instructions. The coefficients of the test polynomials and the endpoints of the components of the input interval vectors are uniformly distributed in $[-1, 1]$. We report the average over 1000 random polynomials and input interval vectors. Table 4.1.1 shows a comparison for random polynomials with different numbers of variables n and monomials m.

- In the first experiment (Table 4.1.1) each degree vector contains n non–zero components and the degree in each variable is at most 5. The accuracy of all three forms is roughly the same. However, the cost of N_f is significantly smaller than for H_f, in particular if the number of monomials is large.

  In fact, we expected that the accuracy of N_f would be much better than the accuracy of H_f. According to Theorem 4.1.9, N_f is at least as accurate as D_f if $0 \notin X_i$ for all i.

- Thus, in Table 4.1.2 we repeat the experiment, but consider only cases where $0 \notin X_i$ for all i. The result is disappointing. In general, N_f is only slightly more accurate than D_f and H_f.

- Table 4.1.3 shows the same experiment as in Table 4.1.1, but this time each degree vector has at most 2 non–zero components. Qualitatively, we obtain the same result concerning accuracy and costs.

- Finally, in Table 4.1.4 we repeat the experiment of Table 4.1.3, but this time the degree in each variable is up to 15. Again, the result is qualitatively the same as in the previous experiments.

Thus, we conclude that N_f is in general not more accurate than H_f but costs significantly less in terms of arithmetic floating point operations. The non–numeric overhead for computing the nested expression for N_f pays only if f has to be evaluated several times for different input intervals.

Degree 5, Dense Monomials

Variables n, Monomials m	$2,3$	$2,6$	$4,4$	$4,8$	$8,5$	$8,10$	$16,6$	$16,12$
Flops for $H_f(\boldsymbol{X})$	31.8	53.6	87.5	157.7	232.1	440.8	589.5	1148.6
Flops for $N_f(\boldsymbol{X})$	22.8	36.3	53.2	86.9	127.0	215.9	300.9	531.1
Flops for $D_f(\boldsymbol{X})$	40.9	82.6	104.5	208.0	258.7	515.9	625.8	1251.5
$\mathrm{w}(N_f(\boldsymbol{X}))/\mathrm{w}(H_f(\boldsymbol{X}))$	0.960	0.937	1.016	0.996	1.066	1.051	1.061	1.056
$\mathrm{w}(D_f(\boldsymbol{X}))/\mathrm{w}(H_f(\boldsymbol{X}))$	1.086	1.205	0.989	0.997	0.985	0.982	0.988	0.990

Table 4.1.1: Comparison of H_f, N_f and D_f for random polynomials Each degree vector contains n non–zero components and the degree in each variable is at most 5.

Degree 5, Dense Monomials, $0 \notin X_i$

Variables n, Monomials m	$2,3$	$2,6$	$4,4$	$4,8$	$8,5$	$8,10$	$16,6$	$16,12$
Flops for $H_f(\boldsymbol{X})$	33.2	54.6	90.7	162.5	237.6	449.7	588.1	1146.2
Flops for $N_f(\boldsymbol{X})$	23.5	35.9	52.7	84.3	121.9	204.4	283.1	495.3
Flops for $D_f(\boldsymbol{X})$	44.6	88.7	110.6	221.0	265.9	531.5	626.8	1253.3
$\mathrm{w}(N_f(\boldsymbol{X}))/\mathrm{w}(H_f(\boldsymbol{X}))$	0.903	0.882	0.907	0.864	0.971	0.962	0.998	0.997
$\mathrm{w}(D_f(\boldsymbol{X}))/\mathrm{w}(H_f(\boldsymbol{X}))$	1.166	1.334	1.029	1.054	1.000	1.001	1.000	1.000

Table 4.1.2: Same experiment as in Table 4.1.1, but this time $0 \notin X_i$ for all i.

Degree 5, Sparse Monomials

Variables n, Monomials m	$2,3$	$2,6$	$4,4$	$4,8$	$8,5$	$8,10$	$16,6$	$16,12$
Flops for $H_f(\boldsymbol{X})$	31.8	53.6	62.3	113.7	114.9	211.8	218.7	404.2
Flops for $N_f(\boldsymbol{X})$	22.8	36.3	47.0	77.5	70.7	116.1	102.8	163.3
Flops for $D_f(\boldsymbol{X})$	40.9	82.6	69.5	140.1	124.1	248.1	238.8	476.1
$\mathrm{w}(N_f(\boldsymbol{X}))/\mathrm{w}(H_f(\boldsymbol{X}))$	0.960	0.937	0.993	0.982	0.993	0.982	0.995	0.991
$\mathrm{w}(D_f(\boldsymbol{X}))/\mathrm{w}(H_f(\boldsymbol{X}))$	1.086	1.205	1.020	1.054	1.012	1.019	1.009	1.010

Table 4.1.3: Same experiment as in Table 4.1.1, but this time each degree vector has at most 2 non–zero components.

Degree 15, Sparse Monomials

Variables n, Monomials m	$2,3$	$2,6$	$4,4$	$4,8$	$8,5$	$8,10$	$16,6$	$16,12$
Flops for $H_f(\boldsymbol{X})$	53.0	93.5	90.9	166.7	149.6	278.7	259.2	483.5
Flops for $N_f(\boldsymbol{X})$	43.2	72.1	74.7	130.1	105.6	184.0	144.3	246.1
Flops for $D_f(\boldsymbol{X})$	62.3	125.4	98.4	195.6	159.5	317.8	278.2	556.6
$\mathrm{w}(N_f(\boldsymbol{X}))/\mathrm{w}(H_f(\boldsymbol{X}))$	0.978	0.978	0.991	0.991	0.998	0.995	1.004	1.003
$\mathrm{w}(D_f(\boldsymbol{X}))/\mathrm{w}(H_f(\boldsymbol{X}))$	1.014	1.028	0.995	1.007	1.000	1.000	0.999	0.998

Table 4.1.4: Same experiment as in Table 4.1.3, but this time the degree in each variable is at most 15 and each degree vector has at most 2 non–zero components.

4.2 Mean Value Form

The mean value form of multivariate polynomials is a straight forward generalization of the univariate case. It is quadratically convergent ([Alefeld and Herzberger, 1974], [Skelboe, 1974], [Caprani and Madsen, 1980], [Alefeld and Herzberger, 1983]) and inclusion monotone ([Caprani and Madsen, 1980]). An important improvement of the mean value form, called successive mean value form [Hansen, 1968] is subject of Section 4.2.1. The successive mean value form with slopes instead of derivatives is studied in Section 4.2.2. In Section 4.2.3 we generalize the bicentered mean value form to the multivariate case. A new combination of the successive mean value form and the bicentered mean value form is introduced in Section 4.2.4. Finally, in Section 4.2.5 we compare the different mean value forms experimentally.

The multivariate mean value form can be derived easily from the mean value Theorem: For all $\boldsymbol{X} \in \mathbb{IR}^n$ and for all $\boldsymbol{x}, \boldsymbol{c} \in \boldsymbol{X}$ there exists $\boldsymbol{\xi} \in \boldsymbol{X}$ such that

$$f(\boldsymbol{x}) = f(\boldsymbol{c}) + \sum_{i=1}^{n} \partial_i f(\boldsymbol{\xi})(x_i - c_i),$$

where

$$\partial_i f = \frac{\partial f}{\partial x_i}.$$

Hence,

$$f(\boldsymbol{x}) \in f(\boldsymbol{c}) + \sum_{i=1}^{n} \partial_i f(\boldsymbol{X})(X_i - c_i).$$

Thus, every interval extension F_i of $\partial_i f$, $i = 1, \ldots, n$ and any choice of $\boldsymbol{c} \in \boldsymbol{X}$ give rise to an interval extension F of f:

$$F(\boldsymbol{X}) = f(\boldsymbol{c}) + \sum_{i=1}^{n} F_i(\boldsymbol{X})(X_i - c_i) \supseteq f(\boldsymbol{X}). \tag{4.2.1}$$

The mean value form is a special case of (4.2.1), where F_i is the Horner form of $\partial_i f$ and $\boldsymbol{c} = \mathrm{mid}(\boldsymbol{X})$.

Definition 4.2.1 (Multivariate Mean Value Form) *The multivariate mean value form* $M_f : \mathbb{IR}^n \to \mathbb{IR}$ *is defined as*

$$M_f(\boldsymbol{X}) = f(\mathrm{mid}(\boldsymbol{X})) + \sum_{i=1}^{n} H_{\partial_i f}(\boldsymbol{X})(X_i - \mathrm{mid}(X_i)). \ \square$$

Theorem 4.2.2 (Interval Extension) *M_f is an interval extension of f.* $\square$

Proof. Follows from (4.2.1). $\square$

Theorem 4.2.3 (Convergence) *M_f converges quadratically to f.* $\square$

Proof. Straight forward generalization of Theorem 3.2.4. $\square$

The following theorem is taken from [Caprani and Madsen, 1980].

Theorem 4.2.4 (Inclusion Monotonicity) *M_f is inclusion monotone.* $\square$

Proof. The proof is a straight forward generalization of the proof of Theorem 3.2.3. Let $\boldsymbol{X}_\diamond \subseteq \boldsymbol{X}$, $\boldsymbol{c}_\diamond = \mathrm{mid}(\boldsymbol{X}_\diamond)$, $\boldsymbol{c} = \mathrm{mid}(\boldsymbol{X})$, $\boldsymbol{r}_\diamond = \mathrm{rad}(\boldsymbol{X}_\diamond)$, $\boldsymbol{r} = \mathrm{rad}(\boldsymbol{X})$. We have to show that $M_f(\boldsymbol{X}_\diamond) \subseteq M_f(\boldsymbol{X})$. As $X_i - c_i$ and $X_{i\diamond} - c_{i\diamond}$ are centered intervals for all i, it holds that

$$\begin{aligned} M_f(\boldsymbol{X}) &= f(\boldsymbol{c}) + \sum_{i=1}^{n} \mathrm{mag}(H_{\partial_i f}(\boldsymbol{X}))r_i[-1,1] \\ M_f(\boldsymbol{X}_\diamond) &= f(\boldsymbol{c}_\diamond) + \sum_{i=1}^{n} \mathrm{mag}(H_{\partial_i f}(\boldsymbol{X}_\diamond))r_{i\diamond}[-1,1]. \end{aligned}$$

Hence, we have to show that

$$\begin{aligned} f(\boldsymbol{c}_\diamond) + \sum_{i=1}^{n} \mathrm{mag}(H_{\partial_i f}(\boldsymbol{X}_\diamond))r_{i\diamond} &\leq f(\boldsymbol{c}) + \sum_{i=1}^{n} \mathrm{mag}(H_{\partial_i f}(\boldsymbol{X}))r_i \\ f(\boldsymbol{c}_\diamond) - \sum_{i=1}^{n} \mathrm{mag}(H_{\partial_i f}(\boldsymbol{X}_\diamond))r_{i\diamond} &\geq f(\boldsymbol{c}) - \sum_{i=1}^{n} \mathrm{mag}(H_{\partial_i f}(\boldsymbol{X}))r_i. \end{aligned}$$

Note that for all i

$$\begin{aligned} |c_{i\diamond} - c_i| &= 1/2|(\overline{X_{i\diamond}} + \underline{X_{i\diamond}} - \overline{X}_i - \underline{X}_i)| \\ &= 1/2 \max\{\overline{X_{i\diamond}} + \underline{X_{i\diamond}} - \overline{X}_i - \underline{X}_i, \overline{X}_i + \underline{X}_i - \overline{X_{i\diamond}} - \underline{X_{i\diamond}}\} \\ &\leq 1/2(\overline{X}_i + \underline{X_{i\diamond}} - \overline{X_{i\diamond}} - \underline{X}_i) \\ &= r_i - r_{i\diamond}. \end{aligned}$$

Further, $H_{\partial_i f}(\boldsymbol{X}_\diamond) \subseteq H_{\partial_i f}(\boldsymbol{X})$ by Theorem 3.1.4. Let $\boldsymbol{\xi} \in \boldsymbol{X}$ such that

$$f(\boldsymbol{c}_\diamond) = f(\boldsymbol{c}) + \sum_{i=1}^{n} \partial_i f(\boldsymbol{\xi})(c_{i\diamond} - c_i).$$

Then

$$\begin{aligned} f(\boldsymbol{c}_\diamond) + \sum_{i=1}^{n} \mathrm{mag}(H_{\partial_i f}(\boldsymbol{X}_\diamond))r_{i\diamond} &= f(\boldsymbol{c}) + \sum_{i=1}^{n} \partial_i f(\boldsymbol{\xi})(c_{i\diamond} - c_i) + \mathrm{mag}(H_{\partial_i f}(\boldsymbol{X}_\diamond))r_{i\diamond} \\ &\leq f(\boldsymbol{c}) + \sum_{i=1}^{n} |\partial_i f(\boldsymbol{\xi})||(c_{i\diamond} - c_i)| + \mathrm{mag}(H_{\partial_i f}(\boldsymbol{X}_\diamond))r_{i\diamond} \\ &\leq f(\boldsymbol{c}) + \sum_{i=1}^{n} \mathrm{mag}(H_{\partial_i f}(\boldsymbol{X}))(r_i - r_{i\diamond}) + \mathrm{mag}(H_{\partial_i f}(\boldsymbol{X}))r_{i\diamond} \\ &= f(\boldsymbol{c}) + \sum_{i=1}^{n} \mathrm{mag}(H_{\partial_i f}(\boldsymbol{X}))r_i \\ f(\boldsymbol{c}_\diamond) - \sum_{i=1}^{n} \mathrm{mag}(H_{\partial_i f}(\boldsymbol{X}_\diamond))r_{i\diamond} &= f(\boldsymbol{c}) + \sum_{i=1}^{n} \partial_i f(\boldsymbol{\xi})(c_{i\diamond} - c_i) - \mathrm{mag}(H_{\partial_i f}(\boldsymbol{X}_\diamond))r_{i\diamond} \\ &\geq f(\boldsymbol{c}) + \sum_{i=1}^{n} -|\partial_i f(\boldsymbol{\xi})||(c_{i\diamond} - c_i)| - \mathrm{mag}(H_{\partial_i f}(\boldsymbol{X}_\diamond))r_{i\diamond} \\ &\geq f(\boldsymbol{c}) + \sum_{i=1}^{n} -\mathrm{mag}(H_{\partial_i f}(\boldsymbol{X}))(r_i - r_{i\diamond}) - \mathrm{mag}(H_{\partial_i f}(\boldsymbol{X}))r_{i\diamond} \\ &= f(\boldsymbol{c}) - \sum_{i=1}^{n} \mathrm{mag}(H_{\partial_i f}(\boldsymbol{X}))r_i. \ \square \end{aligned}$$

Algorithm 4.2.5 (MFM) for evaluating the multivariate mean value form follows immediately from Definition 4.2.1. The algorithm requires evaluation of the Horner form of the partial derivatives $\partial_i f$, $i = 1, \ldots, n$. As the coefficients of $\partial_i f$ need not be floating point numbers, we have to enclose them by intervals. The Horner form of multivariate interval polynomials is computed by Algorithm HFMI, which is a trivial modification of Algorithm 4.1.6 (HFM) and which is therefore not given explicitly. Further, we extend Algorithm 2.3.33 MID and Algorithm 2.3.44 (CONVERT) to interval vectors component wise.

Algorithm 4.2.5 (MFM) [Multivariate Mean Value Form]

In: $f(\boldsymbol{x}) = \sum_{i=1}^{m} a_i x_1^{d_{i,1}} x_2^{d_{i,2}} \cdots x_n^{d_{i,n}} \in \mathbb{F}[x_1, \ldots, x_n]$,
$\boldsymbol{X} \in \mathbb{IF}^n$.

Out: $\mathrm{MFM}(f, \boldsymbol{X}) \in \mathbb{IF}$, $\mathrm{MFM}(f, \boldsymbol{X}) \approx M_f(\boldsymbol{X})$, $\mathrm{MFM}(f, \boldsymbol{X}) \supseteq f(X)$.

(1) [Midpoint.]
$\boldsymbol{c} \longleftarrow \mathrm{MID}(\boldsymbol{X})$.

(2) [Evaluate $f(\boldsymbol{c})$.]
$Y \longleftarrow \mathrm{HFM}(f(\boldsymbol{x}), [\boldsymbol{c}])$.

(3) [Partial derivatives.]
for $i = 1, \ldots, n$ do $F_i(\boldsymbol{x}) \longleftarrow \frac{\partial f}{\partial x_i}$ (interval arithmetic for coefficients).

(4) [Evaluate partial derivatives.]
for $i = 1, \ldots, n$ do $F_i \longleftarrow \mathrm{HFMI}(F_i(\boldsymbol{x}), \boldsymbol{X})$.

(5) [Accumulate.]
for $i = 1, \ldots, n$ do $Y \longleftarrow Y + F_i * (X_i - c_i)$.

(6) [Return.]
return Y.

Theorem 4.2.6 (Complexity) *Algorithm 4.2.5 (*MFM*) costs*

$$\begin{aligned} n^2m + nm &\quad \textit{interval power computations,} \\ n^2m + n &\quad \textit{interval multiplications,} \\ 4nm + n &\quad \textit{number multiplications and} \\ 2nm + 4n + 2m - 2 &\quad \textit{number additions.} \ \square \end{aligned}$$

Proof.

- Step 1 costs n number multiplications and $2n$ number additions.
- Step 2 costs nm interval power computations, $2nm$ number multiplications and $2m - 2$ number additions.
- Step 3 costs $2nm$ number multiplications.
- Step 4 costs n^2m interval power computations, n^2m interval multiplications and $2nm - 2n$ number additions.
- Step 5 costs n interval multiplications and $4n$ number additions. $\square$

In Algorithm 4.2.5 (MFM) we explicitly computed polynomials for the partial derivatives of f and evaluated all of them on $\boldsymbol{X}$. A new possibility for obtaining the value of the i-th partial derivative is to evaluate f first in the variables $x_{i+1}, \ldots, x_n$, next compute the formal i-th partial derivative of this i-variate interval polynomial and evaluate it. This can be done successively as follows. Let

$$F^{(n)}(x_1, \ldots, x_n) = f(x_1, \ldots, x_n).$$

For $i = n - 1 \ldots 0$ evaluate $F^{(i+1)}(x_1, \ldots, x_{i+1})$ in its last variable x_i by X_i using the Horner scheme and obtain an i-variate interval polynomial $F^{(i)}(x_1, \ldots, x_i)$.

$$F^{(i)}(x_1, \ldots, x_i) = F^{(i+1)}(x_1, \ldots, x_i, X_{i+1}).$$

Next, let $F_i^{(i)}$ be the formal i-th partial derivative of $F^{(i)}$ and evaluate $F_i^{(i)}$ on $X_1, \ldots, X_n$ using the Horner scheme. One checks that

$$H_{F_i^{(i)}}(X_1, \ldots, X_i) = H_{\partial_i f}(X_1, \ldots, X_n).$$

An advantage of this method is that $F^0() = H_f(\boldsymbol{X})$, i.e. we obtain the Horner form as a side product with minimal additional costs, such that we can compute easily $M_f(\boldsymbol{X}) \cap H_f(\boldsymbol{X})$.

Definition 4.2.7 (Mean Value – Horner Form) *The mean value – Horner form* $M_f^{(\mathrm{H})} : \mathbb{IR}^n \to \mathbb{IR}$ *is defined as*

$$M_f^{(\mathrm{H})}(\boldsymbol{X}) = M_f(\boldsymbol{X}) \cap H_f(\boldsymbol{X}). \ \square$$

We formalize the procedure described above in Algorithm 4.2.11 (MHF) and compare its complexity with Algorithm 4.2.5 (MFM). First, we give an auxiliary algorithm for evaluating an interval polynomial in its last variable.

Algorithm 4.2.8 (HFIL) [Horner Form in Last Variable]

In: $F(\boldsymbol{x}) = \sum_{i=1}^{m} A_i \, x_1^{d_{i,1}} x_2^{d_{i,2}} \cdots x_n^{d_{i,n}} \in \mathbb{IF}[x_1, \ldots, x_n]$
such that $(d_{i,1}, d_{i,2}, \ldots, d_{i,n}) \succ (d_{j,1}, d_{j,2}, \ldots, d_{j,n})$ for $i < j$,
$X \in \mathbb{IF}$.

Out: $\mathrm{HFIL}(F, X) \in \mathbb{IF}[x_1, \ldots, x_{n-1}]$, $\mathrm{HFIL}(F, X) \supseteq H_F(x_1, \ldots, x_{n-1}, X)$.

(1) [Partition f.]
Let

$$k = |\{(d_{i,1}, d_{i,2}, \ldots, d_{i,n-1}) \mid i = 1, \ldots, m\}|.$$

and let $1 = m_1 < \ldots < m_{k+1} = m + 1$ such that

$$F_j(x_n) = \sum_{i=m_j}^{m_{j+1}-1} A_i x_n^{d_{i,n}} \in \mathbb{IF}[x_n], \quad j = 1, \ldots, k$$

and

$$F(\boldsymbol{x}) = \sum_{j=1}^{k} F_j(x_n) \, x_1^{d_{m_j,1}} x_2^{d_{m_j,2}} \cdots x_{n-1}^{d_{m_j,n-1}}.$$

(2) [Evaluate.]
for $j = 1, \ldots, k$ do $B_j \longleftarrow \mathrm{HFI}(F_j(x_n), X)$

(3) [Return.]
return $\sum_{j=1}^{k} B_j \, x_1^{d_{m_j,1}} x_2^{d_{m_j,2}} \cdots x_{n-1}^{d_{m_j,n-1}}$

Theorem 4.2.9 (Complexity) *Algorithm 4.2.8 (*HFIL*) costs*

m	*interval power computations,*
m	*interval multiplications and*
$2m - 2k$	*number additions,*

where

$$k = |\{(d_{i,1}, d_{i,2}, \ldots, d_{i,n-1}) \mid i = 1, \ldots, m\}|. \ \square$$

Proof. The computation of $\mathrm{HFI}(F_j, X)$ in Step 2 costs $m_{j+1} - m_j$ interval power computations, $m_{j+1} - m_j$ multiplications and $m_{j+1} - m_j - 1$ interval additions (Theorem 3.1.51). Hence, Step 2 costs

$$\sum_{j=1}^{k} m_{j+1} - m_j = m_{k+1} - m_1 = m$$

interval power computations and multiplications and

$$\sum_{j=1}^{k} m_{j+1} - m_j - 1 = m_{k+1} - m_1 - k = m - k$$

interval additions. $\square$

Theorem 4.2.10 (Complexity) *If* $0 \notin \mathrm{int}(X)$ *then Algorithm 4.2.8* (HFIL) *costs*

$$\begin{array}{rl} m & \textit{interval power computations,} \\ 2m & \textit{number multiplications and} \\ 2m-2k & \textit{number additions.} \ \square \end{array}$$

Algorithm 4.2.11 (MHF) [Mean Value – Horner Form]

In: $f(\boldsymbol{x}) = \sum_{i=1}^{m} a_i\, x_1^{d_{i,1}} x_2^{d_{i,2}} \cdots x_n^{d_{i,n}} \in \mathbb{F}[x_1, \ldots, x_n]$,
$\boldsymbol{X} \in \mathbb{IF}^n$.

Out: $\mathrm{MHF}(f, \boldsymbol{X}) \in \mathbb{IF}$, $\mathrm{MHF}(f, \boldsymbol{X}) \approx M_f^{(\mathrm{H})}(\boldsymbol{X})$, $\mathrm{MHF}(f, \boldsymbol{X}) \supseteq f(\boldsymbol{X})$.

(1) [Midpoint.]
$\boldsymbol{c} \longleftarrow \mathrm{MID}(\boldsymbol{X})$.

(2) [Evaluate $f(\boldsymbol{c})$.]
$Y \longleftarrow \mathrm{HFM}(f(\boldsymbol{x}), [\boldsymbol{c}])$.

(3) [Convert f to an interval polynomial.]
$F^{(n)}(x_1, \ldots, x_n) \longleftarrow f(x_1, \ldots, x_n)$.

(4) [Evaluate f successively in $\boldsymbol{X}$.]
for $i = n, \ldots, 1$ do $F^{(i-1)}(x_1, \ldots, x_{i-1}) \longleftarrow \mathrm{HFIL}(F^{(i)}(x_1, \ldots, x_i), X_i)$.

(5) [Partial derivatives.]
for $i = n, \ldots, 1$ do $F_i^{(i)}(x_1, \ldots, x_i) \longleftarrow \frac{\partial}{\partial x_i} F^{(i)}(x_1, \ldots, x_i)$.

(6) [Evaluate partial derivatives.]
for $i = n, \ldots, 1$ do $F_i \longleftarrow \mathrm{HFMI}(F_i^{(i)}(x_1, \ldots, x_i), X_1, \ldots, X_i)$.

(7) [Accumulate.]
for $i = n, \ldots, 1$ do $Y \longleftarrow Y + F_i * (X_i - c_i)$.

(8) [Intersect with $H_f(\boldsymbol{X})$.]
$Y \longleftarrow Y \cap F^{(0)}()$.

(9) [Return.]
return Y.

Theorem 4.2.12 (Complexity) *Algorithm 4.2.11* (MHF) *costs*

$$\begin{array}{rl} 1/2n^2m + 5/2nm & \textit{interval power computations,} \\ 1/2n^2m + 3/2nm + n & \textit{interval multiplications,} \\ 4nm + n & \textit{number multiplications and} \\ 2nm + 4n + 4m - 4 & \textit{number additions.} \ \square \end{array}$$

Proof. We assume the worst case where all $F^{(i)}$ have m monomials, i.e. no two degree vectors of f have a common prefix.

- Step 1 costs n number multiplications and $2n$ number additions.
- Step 2 costs nm interval power computations, $2nm$ number multiplications and $2m-2$ number additions.
- Step 4 costs nm interval power computations, nm interval multiplications and $2m-2$ number additions.
- Step 5 costs $2nm$ number multiplications.
- The i-th iteration in Step 6 costs im interval power computations, im interval multiplications and $2m-2$ number additions. Hence, Step 6 costs $1/2n(n+1)m$ interval power computations, $1/2n(n+1)m$ interval multiplications and $2nm-2n$ number additions.

- Step 7 costs n interval multiplications and $4n$ number additions. □

Thus, asymptotically Algorithm 4.2.11 (MHF) is half as expensive as Algorithm 4.2.5 (MFM).

4.2.1 Successive Mean Value Form

An important improvement of the mean value form is presented in [Hansen, 1968]. Recall that for the mean value form we have to evaluate the partial derivatives of f on $(X_1, \ldots, X_n)$. If, instead, we evaluate the i-th partial derivative only on

$$(X_1, \ldots, X_i, c_{i+1}, \ldots, c_n) \subseteq (X_1, \ldots, X_n),$$

we still obtain an interval extension, which is called successive mean value form. Obviously, the successive mean value form is usually tighter than the mean value form.

In order to show that the successive mean value form is an interval extension, we apply the mean value Theorem successively to the variables of f. Let $\boldsymbol{X} \in \mathbb{IR}^n$ and let $\boldsymbol{x}, \boldsymbol{c} \in \boldsymbol{X}$. First, consider $f(x_1, \ldots, x_n)$ as a univariate function in x_n. According to the mean value Theorem there exists $\xi_n \in X_n$ such that

$$f(x_1, \ldots, x_n) = f(x_1, \ldots, x_{n-1}, c_n) + \partial_n f(x_1, \ldots, x_{n-1}, \xi_n)(x_n - c_n).$$

Next, consider $f(x_1, \ldots, x_{n-1}, c_n)$ as a univariate function in x_{n-1}. Applying the mean value Theorem again, we obtain

$$f(x_1, \ldots, x_{n-1}, c_n) = f(x_1, \ldots, x_{n-2}, c_{n-1}, c_n) + \partial_{n-1} f(x_1, \ldots, x_{n-2}, \xi_{n-1}, c_n)(x_{n-1} - c_{n-1})$$

for some $\xi_{n-1} \in X_{n-1}$. Hence,

$$\begin{array}{rcl} f(x_1, \ldots, x_n) & = & f(x_1, \ldots, x_{n-2}, c_{n-1}, c_n) \\ & + & \partial_{n-1} f(x_1, \ldots, x_{n-2}, \xi_{n-1}, c_n)(x_{n-1} - c_{n-1}) \\ & + & \partial_n f(x_1, \ldots, x_{n-1}, \xi_n)(x_n - c_n). \end{array}$$

Next, we apply the mean value Theorem to $f(x_1, \ldots, x_{n-2}, c_{n-1}, c_n)$ as a univariate function in x_{n-2}, and so on. Finally, we obtain

$$f(x_1, \ldots, x_n) = f(c_1, \ldots, c_n) + \sum_{i=1}^{n} \partial_i f(x_1, \ldots, x_{i-1}, \xi_i, c_{i+1}, \ldots, c_n)(x_i - c_i)$$

for some $\boldsymbol{\xi} \in \boldsymbol{X}$ and

$$f(x_1, \ldots, x_n) \in f(c_1, \ldots, c_n) + \sum_{i=1}^{n} \partial_i f(X_1, \ldots, X_i, c_{i+1}, \ldots, c_n)(X_i - c_i).$$

Thus, every interval extension F_i of $\partial_i f$, $i = 1, \ldots, n$ and every choice of $\boldsymbol{c} \in \boldsymbol{X}$ give rise to an interval extension F of f:

$$F(\boldsymbol{X}) = f(\boldsymbol{c}) + \sum_{i=1}^{n} F_i(X_1, \ldots, X_i, c_{i+1}, \ldots, c_n)(X_i - c_i) \supseteq f(\boldsymbol{X}). \tag{4.2.2}$$

The successive mean value form is a special case of (4.2.2), where F_i is the Horner form of $\partial_i f$ and $\boldsymbol{c} = \mathrm{mid}(\boldsymbol{X})$.

Definition 4.2.13 (Successive Mean Value Form) *The successive mean value form* $M_f^{\sim} : \mathbb{IR}^n \to \mathbb{IR}$ *is defined as*

$$M_f^{\sim}(\boldsymbol{X}) = f(\boldsymbol{c}) + \sum_{i=1}^{n} H_{\partial_i f}(X_1, \ldots, X_i, c_{i+1}, \ldots, c_n)(X_i - c_i),$$

where $\boldsymbol{c} = \mathrm{mid}(\boldsymbol{X})$. □

Theorem 4.2.14 (Interval Extension) *$M_f^{\curvearrowright}$ is an interval extension of f.* □

Proof. Follows from (4.2.2). □

Theorem 4.2.15 (Accuracy) *For all $\boldsymbol{X} \in \mathbb{IR}^n$ it holds that*

$$M_f^{\curvearrowright}(X) \subseteq M_f(X). \quad \square$$

Proof. Obvious. □

Corollary 4.2.16 (Convergence) *$M_f^{\curvearrowright}$ converges quadratically to f.* □

Theorem 4.2.17 (Inclusion Monotonicity) *$M_f^{\curvearrowright}$ is* not *inclusion monotone.* □

Proof. Consider the polynomial $f(x_1, x_2) = x_1^2 x_2$. We obtain

$$M_f^{\curvearrowright}(X_1, X_2) = \mathrm{mid}(X_1)^2 \mathrm{mid}(X_2) + 2X_1 \mathrm{mid}(X_2)(X_1 - \mathrm{mid}(X_1)) + X_1^2(X_2 - \mathrm{mid}(X_2)).$$

Obviously

$$([-1,1],[0,2])^{\mathrm{T}} \supset ([-1,1],[2,2])^{\mathrm{T}},$$

but

$$M_f^{\curvearrowright}([-1,1],[0,2]) = [-3,3] \subset [-4,4] = M_f^{\curvearrowright}([-1,1],[2,2]). \quad \square$$

Algorithm 4.2.18 (SMF) for evaluating the successive mean value form follows immediately from Definition 4.2.13. As in Algorithm 4.2.11 (MHF), we do *not* compute the i-th partial derivative of f and evaluate it on $(X_1, \ldots, X_i, c_{i+1}, \ldots, c_n)$. Instead, we first compute successively i-variate polynomials

$$f^{(i)}(x_1, \ldots, x_i) = f(x_1, \ldots, x_i, c_{i+1}, \ldots, c_n)$$

and then evaluate the i-th partial derivative of $f^{(i)}$ on $(X_1, \ldots, X_i)$. As a side product, we obtain $f(c_1, \ldots, c_n)$ which is needed anyways. As usual, the intermediate polynomials have to be enclosed by interval polynomials because of rounding errors.

Algorithm 4.2.18 (SMF) [Successive Mean Value Form]

In: $f(\boldsymbol{x}) = \sum_{i=1}^{m} a_i\, x_1^{d_{i,1}} x_2^{d_{i,2}} \cdots x_n^{d_{i,n}} \in \mathbb{F}[x_1, \ldots, x_n]$,
$\boldsymbol{X} \in \mathbb{IF}^n$.

Out: $\mathrm{SMF}(f, \boldsymbol{X}) \in \mathbb{IF}$, $\mathrm{SMF}(f, \boldsymbol{X}) \approx M_f^{\curvearrowright}(\boldsymbol{X})$, $\mathrm{SMF}(f, \boldsymbol{X}) \supseteq f(\boldsymbol{X})$.

(1) [Midpoint.]
$\boldsymbol{c} \longleftarrow \mathrm{MID}(\boldsymbol{X})$.

(2) [Convert f to an interval polynomial.]
$F^{(n)}(x_1, \ldots, x_n) \longleftarrow f(x_1, \ldots, x_n)$.

(3) [Evaluate f successively in $\boldsymbol{c}$.]
for $i = n, \ldots, 1$ do $F^{(i-1)}(x_1, \ldots, x_{i-1}) \longleftarrow \mathrm{HFIL}(F^{(i)}(x_1, \ldots, x_i), [c_i])$.

(4) [Partial derivatives.]
for $i = n, \ldots, 1$ do $F_i^{(i)}(x_1, \ldots, x_i) \longleftarrow \frac{\partial}{\partial x_i} F^{(i)}(x_1, \ldots, x_i)$.

(5) [Evaluate partial derivatives.]
for $i = n, \ldots, 1$ do $F_i \longleftarrow \mathrm{HFMI}(F_i^{(i)}(x_1, \ldots, x_i), X_1, \ldots, X_i)$.

(6) [Accumulate.]
$Y \longleftarrow F^{(0)}()$.
for $i = n, \ldots, 1$ do $Y \longleftarrow Y + F_i * (X_i - c_i)$.

(7) [Return.]
return Y.

Theorem 4.2.19 (Complexity) *Algorithm 4.2.18 (*SMF*) costs*

$$\begin{aligned} 1/2n^2m + 3/2nm &\quad \textit{interval power computations,} \\ 1/2n^2m + 1/2nm + n &\quad \textit{interval multiplications,} \\ 4nm + n &\quad \textit{number multiplications and} \\ 2nm + 4n + 2m - 2 &\quad \textit{number additions.} \ \square \end{aligned}$$

Proof. We assume the worst case where all $F^{(i)}$ have m monomials, i.e. no two degree vectors of f have a common prefix.

- Step 1 costs n number multiplications and $2n$ number additions.
- Step 3 costs nm interval power computations, $2nm$ number multiplications and $2m-2$ number additions.
- Step 4 costs $2nm$ number multiplications.
- The i-th iteration in Step 5 costs im interval power computations, im interval multiplications and $2m-2$ number additions. Hence, Step 5 costs $1/2n(n+1)m$ interval power computations, $1/2n(n+1)m$ interval multiplications and $2nm-2n$ number additions.
- Step 6 costs n interval multiplications and $4n$ number additions. $\square$

4.2.2 Successive Slope Form

The univariate mean value form was improved significantly by replacing the derivative by a slope. Now, we apply this idea to the multivariate case. The interval extension which we derive can also be obtained by applying the formulas of [Krawczyk and Neumaier, 1985] for rational expressions in a specific way. However, for polynomials there is a more direct method, which reveals possibilities to optimize a corresponding algorithm.

Generalizing the univariate case means finding polynomials $g_i(\boldsymbol{x}, \boldsymbol{c})$, $i = 1, \ldots, n$ such that

$$f(\boldsymbol{x}) = f(\boldsymbol{c}) + \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{c})(\boldsymbol{x} - \boldsymbol{c}) \tag{4.2.3}$$

for all $\boldsymbol{x}$, $\boldsymbol{c}$. As the g_i are not uniquely determined by (4.2.3) we have to impose a further condition: We require that g_i depends only on $x_1, \ldots, x_i, c_i, c_{i+1}, \ldots, c_n$. Therefore, we write

$$g_i(x_1, \ldots, x_i, c_i, c_{i+1}, \ldots, c_n)$$

instead of $g_i(\boldsymbol{x}, \boldsymbol{c})$.

The polynomials g_i can be constructed as follows. First, consider $f(x_1, \ldots, x_n)$ as a univariate function in x_n. Let $g_n(x_1, \ldots, x_n, c_n)$ be the uniquely defined polynomial such that

$$f(x_1, \ldots, x_n) = f(x_1, \ldots, x_{n-1}, c_n) + g_n(x_1, \ldots, x_n, c_n)(x_n - c_n).$$

Next, consider $f(x_1, \ldots, x_{n-1}, c_n)$ as a univariate function in x_{n-1} and let $g_{n-1}(x_1, \ldots, x_{n-1}, c_{n-1}, c_n)$ be the uniquely defined polynomial such that

$$f(x_1, \ldots, x_{n-1}, c_n) = f(x_1, \ldots, x_{n-2}, c_{n-1}, c_n) + g_{n-1}(x_1, \ldots, x_{n-1}, c_{n-1}, c_n)(x_{n-1} - c_{n-1}).$$

Hence,

$$\begin{aligned} f(x_1, \ldots, x_n) &= f(x_1, \ldots, x_{n-2}, c_{n-1}, c_n) \\ &+ g_{n-1}(x_1, \ldots, x_{n-1}, c_{n-1}, c_n)(x_{n-1} - c_{n-1}) \\ &+ g_n(x_1, \ldots, x_n, c_n)(x_n - c_n). \end{aligned}$$

Applying this procedure successively to $x_{n-2}, \ldots, x_1$ we obtain polynomials $g_i(x_1, \ldots, x_i, c_i, c_{i+1}, \ldots, c_n)$ such that

$$f(x_1, \ldots, x_n) = \sum_{i=1}^{n} g_i(x_1, \ldots, x_i, c_i, c_{i+1}, \ldots, c_n)(x_i - c_i). \tag{4.2.4}$$

Thus, every interval extension G_i of g_i, $i = 1, \ldots, n$ and any choice of $\boldsymbol{c} \in \mathbb{R}^n$ gives rise to an interval extension F of f:

$$F(\boldsymbol{X}) = f(\boldsymbol{c}) + \sum_{i=1}^{n} G_i(X_1, \ldots, X_i, c_i, c_{i+1}, \ldots, c_n)(X_i - c_i) \supseteq f(\boldsymbol{X}). \tag{4.2.5}$$

The successive slope form is a special case of (4.2.5) where G_i is the Horner form of g_i.

Definition 4.2.20 (Successive Slope Form) *The successive slope form $M_f^{(s)} : \mathbb{IR}^n \to \mathbb{IR}$ is defined as*

$$M_f^{(s)}(\boldsymbol{X}) = f(\boldsymbol{c}) + \sum_{i=1}^{n} H_{g_i}(X_1, \ldots, X_i, c_i, c_{i+1}, \ldots, c_n)(X_i - c_i),$$

where $\boldsymbol{c} = \mathrm{mid}(\boldsymbol{X})$ and g_i is as in (4.2.4). □

Theorem 4.2.21 (Convergence) *$M_f^{(s)}$ converges quadratically to f.* □

Proof. Straight forward generalization of Theorem 3.2.9. □

Before we devise an algorithm for $M_f^{(s)}$ we give formulas for the polynomials g_i.

Theorem 4.2.22 *Let*

$$g_i(x_1, \ldots, x_i, c_i, c_{i+1}, \ldots, c_n) = \sum_{j=1}^{m} a_j x_1^{d_{j,1}} \cdots x_{i-1}^{d_{j,i-1}} \, h^{(d_{j,i})}(x_i, c_i) \, c_{i+1}^{d_{j,i+1}} \cdots c_n^{d_{j,n}},$$

where

$$h^{(d)}(x_i, c_i) = \sum_{\substack{u+v=d-1 \\ u \geq 0, v \geq 0}} x_i^u c_i^v.$$

Then

$$f(\boldsymbol{x}) = f(\boldsymbol{c}) + \sum_{i=1}^{n} g_i(x_1, \ldots, x_i, c_i, c_{i+1}, \ldots, c_n)(x_i - c_i). \; \square$$

Proof. We have to show that for all i

$$f(x_1, \ldots, x_i, c_{i+1}, \ldots, c_n) - f(x_1, \ldots, x_{i-1}, c_i, \ldots, c_n) = g_i(x_1, \ldots, x_i, c_i, c_{i+1}, \ldots, c_n)(x_i - c_i).$$

Note that

$$h^{(d)}(x_i, c_i)(x_i - c_i) = x_i^d - c_i^d.$$

$$\begin{aligned}
& f(x_1, \ldots, x_i, c_{i+1}, \ldots, c_n) - f(x_1, \ldots, x_{i-1}, c_i, \ldots, c_n) \\
&= \sum_{j=1}^{m} a_j \, x_1^{d_{j,1}} \cdots x_i^{d_{j,i}} c_{i+1}^{d_{j,i+1}} \cdots c_n^{d_{j,n}} - \sum_{j=1}^{m} a_j \, x_1^{d_{j,1}} \cdots x_{i-1}^{d_{j,i-1}} c_i^{d_{j,i}} \cdots c_n^{d_{j,n}} \\
&= \sum_{j=1}^{m} a_j \, x_1^{d_{j,1}} \cdots x_{i-1}^{d_{j,i-1}} c_{i+1}^{d_{j,i+1}} \cdots c_n^{d_{j,n}} (x_i^{d_{j,i}} - c_i^{d_{j,i}}) \\
&= \sum_{j=1}^{m} a_j \, x_1^{d_{j,1}} \cdots x_{i-1}^{d_{j,i-1}} c_{i+1}^{d_{j,i+1}} \cdots c_n^{d_{j,n}} h^{(d_{j,i})}(x_i, c_i)(x_i - c_i) \\
&= g_i(x_1, \ldots, x_i, c_i, c_{i+1}, \ldots, c_n)(x_i - c_i). \; \square
\end{aligned}$$

The following observations are important for the efficiency of an implementation of $M_f^{(s)}$.

- First, the polynomials $h^{(d)}$ have the property

$$h^{(d+1)}(x_i, c_i) = h^{(d)}(x_i, c_i)x_i + c_i^d.$$

As the intervals $H_{h^{(d_{j,i})}}(X_i, c_i)$ have to be computed for $j = 1, \ldots, m$, it is usually advantageous to pre–compute the intervals

$$\begin{aligned} H_i^{(0)} &= 0 \\ H_i^{(d)} &= H_i^{(d-1)} X_i + c_i^{d-1} \text{ for } d = 1, \ldots, \max_j d_{j,i} \end{aligned}$$

as well as the powers c_i^d, $d = 0, \ldots, \max_j d_{j,i} - 1$ where $\boldsymbol{X}$ is the input interval and $\boldsymbol{c} = \mathrm{mid}(\boldsymbol{X})$.

- Further, by evaluating f successively in its last variable on c_i, $i = n, \ldots, 1$ we obtain polynomials

$$\begin{aligned} f^{(n)}(x_1, \ldots, x_n) &= f(x_1, \ldots, x_n) \\ f^{(i)}(x_1, \ldots, x_i) &= f^{(i+1)}(x_1, \ldots, x_i, c_i) \\ &= \sum_{j=1}^{m^{(i)}} a_j^{(i)} x_1^{d_{j,1}^{(i)}} \cdots x_i^{d_{j,i}^{(i)}} \end{aligned}$$

The coefficients of these polynomials can be used for the evaluation of g_i, as

$$g_i(x_1, \ldots, x_i, c_i, c_{i+1}, \ldots, c_n) = \sum_{j=1}^{m^{(i)}} a_j^{(i)} x_1^{d_{j,1}^{(i)}} \cdots x_{i-1}^{d_{j,i-1}^{(i)}} h^{(d_{j,i}^{(i)})}(x_i, c_i).$$

- Finally, let a_i^*, $i = 0, \ldots, \deg(f^{(1)})$ such that

$$f^{(1)}(x_1) = \sum_{i=0}^{\deg(f^{(1)})} a_i^* x_1^i.$$

Then

$$g_1(x_1, c_1, c_2, \ldots, c_n) = \sum_{i=1}^{\deg(f^{(1)})} b_i^* x_1^{i-1}$$

where

$$b_i^* = \begin{cases} a_i^* & \text{if } i = \deg(f^{(1)}) \\ b_{i+1}^* c_1 + a_i^* & \text{else.} \end{cases}$$

Thus,

$$f(c_1, \ldots, c_n) + H_{g_1}(X_1, c_1, c_2, \ldots, c_n)(X_1 - c_1) = M_{f^{(1)}}^{(s)}(X_1),$$

i.e. we end up with the univariate slope form (see Definition 3.2.7).

Remark. The polynomials $h^{(d)}(x_i, c_i)$ have an interesting structure and one may expect that the computation of the exact range $h(X_i, \mathrm{mid}(X_i))$ is cheap. In fact, from Theorem 3.1.12 it follows that if $0 \notin \mathrm{int}(X_i)$ then

$$H_i^{(d)} = h^{(d)}(X_i, \mathrm{mid}(X_i)).$$

Further, if d is even, then $h^{(d)}(x_i, c_i)$ is monotone in x_i, hence

$$[h^{(d)}(\underline{X}_i, \mathrm{mid}(X_i)), h^{(d)}(\overline{X}_i, \mathrm{mid}(X_i))] = h^{(d)}(X_i, \mathrm{mid}(X_i)).$$

However, this would introduce additional costs and we will not make use of it. □

Algorithm 4.2.23 (SSF) [Successive Slope Form]

In: $f(\boldsymbol{x}) = \sum_{i=1}^m a_i x_1^{d_{i,1}} x_2^{d_{i,2}} \cdots x_n^{d_{i,n}} \in \mathbb{F}[x_1, \ldots, x_n]$,
$\boldsymbol{X} \in \mathbb{IF}^n$.

Out: $\mathrm{SSF}(f, \boldsymbol{X}) \in \mathbb{IF}$, $\mathrm{SSF}(f, \boldsymbol{X}) \approx M_f^{(s)}(\boldsymbol{X})$, $\mathrm{SSF}(f, \boldsymbol{X}) \supseteq f(\boldsymbol{X})$.

(1) [Powers of c_i and $H_i^{(j)}$.]
for $i = 2, \ldots, n$

(1.1) [Degree.]
$d \longleftarrow$ degree of f in x_i.

(1.2) [Powers of c_i.]
$c_i \longleftarrow \mathrm{MID}(X_i)$.
$C_i^{(0)} \longleftarrow 1$.
for $j = 1, \ldots, d-1$ do $C_i^{(j)} \longleftarrow C_i^{(j-1)} * [c_i]$.

(1.3) [Compute $H_i^{(d)}$.]
$H_i^{(0)} \longleftarrow 0$.
$H_i^{(1)} \longleftarrow 1$.
for $j = 2, \ldots, d$ do $H_i^{(j)} \longleftarrow H_i^{(j-1)} X_i + C_i^{(j-1)}$.

(2) [Sucessively evaluate slopes.]
$F^{(n)}(x_1, \ldots, x_n) \longleftarrow f(x_1, \ldots, x_n)$.
for $i = n, \ldots, 2$

(2.1) [Slope.]
Let $F^{(i)}(x_1, \ldots, x_i) = \sum_{j=1}^{m^{(i)}} a_j^{(i)} x_1^{d_{j,1}^{(i)}} \cdots x_i^{d_{j,i}^{(i)}}$.
$\tilde{F}^{(i)}(x_1, \ldots, x_{i-1}) \longleftarrow \sum_{j=1}^{m^{(i)}} a_j^{(i)} x_1^{d_{j,1}^{(i)}} \cdots x_{i-1}^{d_{j,i-1}^{(i)}} H_i^{d_{j,i}^{(i)}}$.
$F_i \longleftarrow \mathrm{HFMI}(\tilde{F}^{(i)}(x_1, \ldots, x_{i-1}), X_1, \ldots, X_{i-1})$.

(2.2) [Successively evaluate f.]
$F^{(i-1)} \longleftarrow \mathrm{HFIL}(F^{(i)}(x_1, \ldots, x_i), [c_i])$.

(3) [Univarate slope form.]
$Y \longleftarrow \mathrm{DSF}(F^{(1)}(x_1), X_1)$.

(4) [Add Slopes.]
for $i = 2, \ldots, n$ do $Y \longleftarrow Y + F_i * (X_i - c_i)$.

(5) [Return.]
return Y.

Remark. In Step 3 we call Algorithm 3.2.13 (DSF) for evaluating the dense univariate slope form. As the first argument of DSF is an interval polynomial, we have to modify Algorithm 3.2.13 (DSF). However, the modification is trivial and the complexity remains the same. In order to conform precisely to Definition 4.2.20, we would have to evaluate the slope form instead of the dense slope form. But for the reasons discussed in Section 3.2.1, we use the dense slope form. □

Theorem 4.2.24 (Complexity) *Let d be the maximum degree of f in each variable. Then Algorithm 4.2.23 (*SSF*) costs*

$$\begin{array}{rl} 1/2n^2m + 1/2nm - m & \textit{interval power computations} \\ 1/2n^2m + 1/2nm + nd - m + d - 1 & \textit{interval multiplications} \\ 2nm + 2nd - n - 2 * m + 2 * d + 1 & \textit{number multiplications} \\ 2nm + 2nd + 2n + 2m + 2d - 4 & \textit{number additions.} \ \square \end{array}$$

Proof.

- Let d be the maximum degree of f in each variable. First, we compute the cost of the i-th iteration of Step 1. Step 1.2 costs $2d - 1$ number multiplication and 2 number additions. Step 1.3 costs

$d-1$ interval multiplications and $2d-2$ number additions. Thus, Step 1 costs $nd-n$ interval multiplications, $2nd-n$ number multiplications and $2nd$ number additions.

- The i-th iteration of Step 2.1 costs $m^{(i)}$ interval multiplications $2(m^{(i)}-m^{(i-1)})$ number additions for the computation of the coefficients of $\tilde{F}^{(i)}$ and $(i-1)m^{(i-1)}$ interval power computations, $(i-1)m^{(i-1)}$ interval multiplications and $2m^{(i-1)}-2$ number additions for its evaluation. In the worst case $m^{(i)}=m$ for all $i>0$ and we obtain $1/2(n^2-n)m$ interval power computations, $1/2(n^2+n-2)m$ interval multiplications and $2(n-1)(m-1)$ number additions for all iterations of Step 2.1. The total cost of Step 2.2 is bounded by $(n-1)m$ interval power computations, $2(n-1)m$ number multiplications and $2m-2$ number additions. So, Step 2 costs $1/2(n^2+n-2)m$ interval power computations $1/2(n^2+n-2)m$ interval multiplications, $2(n-1)m$ number multiplications and $2n(m-1)$ number additions.
- Step 3 costs d interval multiplications, $2d+1$ number multiplications and $2d+2m$ number additions.
- Step 4 costs $n-1$ interval multiplications and $4(n-1)$ number additions. □

4.2.3 Bicentered Mean Value Form

The multivariate bicentered mean value form is a straight forward generalization of the univariate case, see Section 3.2.2. Both have been introduced by [Baumann, 1988].

The basic idea is to evaluate the mean value form twice with different centers $\boldsymbol{c}^{\uparrow}$ and $\boldsymbol{c}^{\downarrow}$ and intersect the results. The centers are optimal in the sense that $\boldsymbol{c}^{\uparrow}$ minimizes the upper bound and $\boldsymbol{c}^{\downarrow}$ maximizes the lower bound among all centers $\boldsymbol{c}\in\boldsymbol{X}$. The bicentered mean value form is therefore particularly suitable for the separate computation of an upper or lower bound of f in $\boldsymbol{X}$.

In the sequel let

$$M_f(\boldsymbol{X},\boldsymbol{c})=f(\boldsymbol{c})+\sum_{i=1}^{n}H_{\partial_i f}(\boldsymbol{X})(X_i-c_i).$$

Definition 4.2.25 (Multivariate Bicentered Mean Value Form) *The multivariate bicentered mean value form $\ddot{M}_f:\mathbb{IR}^n\to\mathbb{IR}$ is defined as*

$$\ddot{M}_f(\boldsymbol{X})=[\underline{M_f(\boldsymbol{X},\boldsymbol{c}^{\downarrow})},\overline{M_f(\boldsymbol{X},\boldsymbol{c}^{\uparrow})}],$$

where the optimal centers $\boldsymbol{c}^{\uparrow}$ and $\boldsymbol{c}^{\downarrow}$ are

$$c_i^{\uparrow} = \begin{cases} \overline{X}_i & \text{if } \underline{F}_i\geq 0\\ \underline{X}_i & \text{if } \overline{F}_i\leq 0\\ (\overline{F}_i\,\overline{X}_i-\underline{F}_i\,\underline{X}_i)/\mathrm{w}(F_i) & \text{else}\end{cases}$$

$$c_i^{\downarrow} = \begin{cases} \underline{X}_i & \text{if } \underline{F}_i\geq 0\\ \overline{X}_i & \text{if } \overline{F}_i\leq 0\\ (\overline{F}_i\,\underline{X}_i-\underline{F}_i\,\overline{X}_i)/\mathrm{w}(F_i) & \text{else}\end{cases}$$

and $F_i=H_{\partial_i f}(\boldsymbol{X})$. □

In the sequel let $\boldsymbol{c}^{\uparrow},\boldsymbol{c}^{\downarrow}$ and F_i as in Definition 4.2.25.

The following theorems are generalizations of the corresponding theorems for the univariate case. As the proofs are completely analogous, we omit them.

Theorem 4.2.26 (Interval Extension) *$\ddot{M}_f$ is an interval extension of f.* □

Theorem 4.2.27 (Non–Overestimation) *If $0\notin\mathrm{int}(F_i)$ for all i then*

$$\ddot{M}_f(\boldsymbol{X})=f(\boldsymbol{X}).$$ □

Theorem 4.2.28 (Optimality of Centers) *For all $\boldsymbol{c} \in \boldsymbol{X}$ it holds that*

$$\begin{array}{rcl} \overline{M_f(\boldsymbol{X}, \boldsymbol{c}^{\uparrow})} & \leq & \overline{M_f(\boldsymbol{X}, \boldsymbol{c})} \\ \underline{M_f(\boldsymbol{X}, \boldsymbol{c}^{\downarrow})} & \geq & \underline{M_f(\boldsymbol{X}, \boldsymbol{c})}. \end{array} \quad \square$$

Corollary 4.2.29

$$\ddot{M}_f(\boldsymbol{X}) = \bigcap_{\boldsymbol{c} \in \boldsymbol{X}} M_f(\boldsymbol{X}, \boldsymbol{c}). \quad \square$$

Corollary 4.2.30 (Convergence) *$\ddot{M}_f$ is quadratically convergent.* $\square$

Theorem 4.2.31 (Inclusion Monotonicity) *$\ddot{M}_f$ is inclusion monotone.* $\square$

Theorem 4.2.32 *Let $C_i = [c_i^{\downarrow}, c_i^{\uparrow}]$ and let $\boldsymbol{c} \in \boldsymbol{X}$.*

$$\mathrm{w}(M_f(\boldsymbol{X}, \boldsymbol{c})) = \sum_{i=1}^{n} \mathrm{mag}(F_i)\mathrm{w}(X_i) + \mathrm{w}(F_i)\mathrm{mig}(C_i - c_i). \quad \square$$

Corollary 4.2.33 (Optimality of the Midpoint for the Mean Value Form) *For all $\boldsymbol{c} \in \boldsymbol{X}$ it holds that*

$$\mathrm{w}(M_f(\boldsymbol{X})) \leq \mathrm{w}(M_f(\boldsymbol{X}, \boldsymbol{c})). \quad \square$$

Algorithm 4.2.34 (BMFM) for evaluating the bicentered mean value form follows immediately from Definition 4.2.25. As in the univariate case, we use Algorithm 3.2.24 (OC) for approximating the optimal centers $\boldsymbol{c}^{\uparrow}$ and $\boldsymbol{c}^{\downarrow}$.

Algorithm 4.2.34 (BMFM) [Multivariate Bicentered Mean Value Form]

In: $f(\boldsymbol{x}) = \sum_{i=1}^{m} a_i x_1^{d_{i,1}} x_2^{d_{i,2}} \cdots x_n^{d_{i,n}} \in \mathbb{F}[x_1, \ldots, x_n]$,
$\boldsymbol{X} \in \mathbb{IF}^n$.

Out: $\mathrm{BMFM}(f, \boldsymbol{X}) \in \mathbb{IF}$, $\mathrm{BMFM}(f, \boldsymbol{X}) \approx \ddot{M}_f(\boldsymbol{X})$, $\mathrm{BMFM}(f, \boldsymbol{X}) \supseteq f(\boldsymbol{X})$.

(1) [Initialize.]
$y^{\uparrow} \longleftarrow 0$, $y^{\downarrow} \longleftarrow 0$.

(2) [Iterate over all variables.]
for $i = 1, \ldots, n$ do

(2.1) [Partial derivative.]
$F_i(\boldsymbol{x}) \longleftarrow \partial f(\boldsymbol{x})/\partial x_i$ (interval arithmetic for coefficients).

(2.2) [Evaluate partial derivative.]
$F_i \longleftarrow \mathrm{HFMI}(F_i(\boldsymbol{x}), \boldsymbol{X})$.

(2.3) [Optimal centers.]
$c_i^{\uparrow}, c_i^{\downarrow} \longleftarrow \mathrm{OC}(F_i, X_i)$.

(2.4) [Accumulate.]
$y^{\uparrow} \longleftarrow y^{\uparrow} \mathbin{\hat{+}} \overline{F_i * (X_i - c_i^{\uparrow})}$.
$y^{\downarrow} \longleftarrow y^{\downarrow} \mathbin{\check{+}} \underline{F_i * (X_i - c_i^{\downarrow})}$.

(3) [Evaluate at centers.]
$y^{\uparrow} \longleftarrow y^{\uparrow} \mathbin{\hat{+}} \overline{\mathrm{HFM}(f(\boldsymbol{x}), [\boldsymbol{c}^{\uparrow}])}$.
$y^{\downarrow} \longleftarrow y^{\downarrow} \mathbin{\check{+}} \underline{\mathrm{HFM}(f(\boldsymbol{x}), [\boldsymbol{c}^{\downarrow}])}$.
return $[y^{\downarrow}, y^{\uparrow}]$.

Theorem 4.2.35 (Complexity) *Algorithm 4.2.34 (*BMFM*) costs*

$$\begin{aligned} n^2m + 2nm &\quad \textit{interval power computations,}\\ n^2m + 2n &\quad \textit{interval multiplications,}\\ 2n &\quad \textit{number divisions,}\\ 6nm + 4n &\quad \textit{number multiplications and}\\ 2nm + 7n + 4m - 2 &\quad \textit{number additions.}\ \square \end{aligned}$$

Proof. The costs for one iteration of Step 2 are as follows:

- Step 2.1 costs $2m$ number multiplications.
- Step 2.2 costs nm interval power computations, nm interval multiplications and $2m-2$ number additions.
- Step 2.3 costs 2 number divisions, 4 number multiplications and 3 number additions (Theorem 3.2.25).
- Step 2.4 costs 2 interval multiplications and 6 number additions.

Summarizing, Step 2 costs n^2m interval power computations, n^2m+2n interval multiplications, $2n$ number divisions, $2nm+4n$ number multiplications and $2nm+7n$ number additions. Further, Step 3 costs $2nm$ interval power computations, $4nm$ number multiplications and $4m-2$ number additions. $\square$

As in the case of the mean value form, we can evaluate the partial derivatives successively and intersect with $H_f(\boldsymbol{X})$ which is obtained as a side product.

Definition 4.2.36 (Bicentered Mean Value – Horner Form) *The bicentered mean value – Horner form* $\ddot{M}_f^{(\mathrm{H})} : \mathbb{IR}^n \to \mathbb{IR}$ *is defined as*

$$\ddot{M}_f^{(\mathrm{H})}(\boldsymbol{X}) = \ddot{M}_f(\boldsymbol{X}) \cap H_f(\boldsymbol{X}).\ \square$$

Algorithm 4.2.37 (BMHF) [Bicentered Mean Value – Horner Form]

In: $f(\boldsymbol{x}) = \sum_{i=1}^{m} a_i\, x_1^{d_{i,1}} x_2^{d_{i,2}} \cdots x_n^{d_{i,n}} \in \mathbb{F}[x_1,\ldots,x_n]$,
$\boldsymbol{X} \in \mathbb{IF}^n$.

Out: $\mathrm{BMHF}(f,\boldsymbol{X}) \in \mathbb{IF}$, $\mathrm{BMHF}(f,\boldsymbol{X}) \approx \ddot{M}_f^{(\mathrm{H})}(\boldsymbol{X})$, $\mathrm{BMHF}(f,\boldsymbol{X}) \supseteq f(\boldsymbol{X})$.

(1) [Initialize.]
$y^{\uparrow} \longleftarrow 0$, $y^{\downarrow} \longleftarrow 0$.
$F^{(n)}(x_1,\ldots,x_n) \longleftarrow f(x_1,\ldots,x_n)$.

(2) [Iterate over all variables.]
for $i = n,\ldots,1$ do

(2.1) [Partial Derivative.]
$F_i^{(i)}(x_1,\ldots,x_i) \longleftarrow \partial F^{(i)}(x_1,\ldots,x_i)/\partial x_i$.

(2.2) [Evaluate partial derivative.]
$F_i \longleftarrow \mathrm{HFMI}(F_i^{(i)}(x_1,\ldots,x_i), X_1,\ldots,X_i)$.

(2.3) [Optimal centers.]
$c_i^{\uparrow}, c_i^{\downarrow} \longleftarrow \mathrm{OC}(F_i, X_i)$.

(2.4) [Accumulate.]
$y^{\uparrow} \longleftarrow y^{\uparrow} \mathbin{\hat{+}} \overline{F_i * (X_i - c_i^{\uparrow})}$.
$y^{\downarrow} \longleftarrow y^{\downarrow} \mathbin{\check{+}} \underline{F_i * (X_i - c_i^{\downarrow})}$.

(2.5) [Evaluate f successively.]
$F^{(i-1)}(x_1,\ldots,x_{i-1}) \longleftarrow \mathrm{HFIL}(F^{(i)}(x_1,\ldots,x_i), X_i)$.

(3) [Evaluate at centers.]
$y^{\uparrow} \longleftarrow y^{\uparrow} \mathbin{\hat{+}} \overline{\mathrm{HFM}(f(\boldsymbol{x}), [\boldsymbol{c}^{\uparrow}])}$.
$y^{\downarrow} \longleftarrow y^{\downarrow} \mathbin{\check{+}} \underline{\mathrm{HFM}(f(\boldsymbol{x}), [\boldsymbol{c}^{\downarrow}])}$.
$Y \longleftarrow [y^{\downarrow}, y^{\uparrow}]$.

(4) [Intersect with $H_f(\boldsymbol{X})$.]
$Y \longleftarrow Y \cap F^{(0)}()$.

(5) [Return.]
return Y.

Theorem 4.2.38 (Complexity) *Algorithm 4.2.37 (*BMHF*) costs*

$$\begin{array}{rl} 1/2n^2m + 9/2nm & \textit{interval power computations,} \\ 1/2n^2m + 5/2nm + 2n & \textit{interval multiplications,} \\ 2n & \textit{number divisions,} \\ 6nm + 4n & \textit{number multiplications and} \\ 2nm + 7n + 8m - 6 & \textit{number additions.} \ \square \end{array}$$

Proof. As in the proof of Theorem 4.2.12 we assume the worst case where all $F^{(i)}$ have m monomials. Hence, the successive evaluation of f in Step 2.5 costs $2nm$ interval power computations, $2nm$ interval multiplications and $4m - 4$ number additions. The remaining costs for the i-th iteration of Step 2 are:

- Step 2.1 costs $2m$ number multiplications.
- Step 2.2 costs im interval power computations, im interval multiplications and $2m - 2$ number additions.
- Step 2.3 costs 2 number divisions, 4 number multiplications and 3 number additions (Theorem 3.2.25).
- Step 2.4 costs 2 interval multiplications and 6 number additions.

Summarizing, Step 2 costs $1/2(n^2 + 5n)m$ interval power computations, $1/2(n + 5)nm + 2n$ interval multiplications, $2n$ number divisions, $2nm + 4n$ number multiplications and $2nm + 7n + 4m - 4$ number additions. Further, Step 3 costs $2nm$ interval power computations, $4nm$ number multiplications and $4m - 2$ number additions. $\square$

4.2.4 Successive Bicentered Mean Value Form

In this section we present a new interval extension which combines the ideas of the successive and the bicentered mean value form: The successive mean value form is evaluated twice with different centers and the results are intersected. The centers are chosen in a similar way as for the bicentered mean value form. However, the centers are not optimal and it can even happen that the successive mean value form gives more accurate results than the successive bicentered mean value form. On the other hand, the successive bicentered mean value form is always at least as accurate as the bicentered mean value form.

In the sequel let

$$M_f^{\sim}(\boldsymbol{X}, \boldsymbol{c}) = f(\boldsymbol{c}) + \sum_{i=1}^{n} H_{\partial_i f}(X_1, \ldots, X_i, c_{i+1}, \ldots, c_n).$$

Definition 4.2.39 (Successive Bicentered Mean Value Form) *The successive bicentered mean value form $\ddot{M}_f^{\sim} : \mathbb{IR}^n \to \mathbb{IR}$ is defined as*

$$\ddot{M}_f^{\sim}(\boldsymbol{X}) = [\underline{M_f^{\sim}(\boldsymbol{X}, \boldsymbol{c}^{\downarrow})}, \overline{M_f^{\sim}(\boldsymbol{X}, \boldsymbol{c}^{\uparrow})}], \tag{4.2.6}$$

where the centers $\boldsymbol{c}^{\uparrow}$ *and* $\boldsymbol{c}^{\downarrow}$ *are defined recursively as*

$$c_i^{\uparrow} = \begin{cases} \overline{X}_i & \text{if } \underline{F}_i^{\uparrow} \geq 0 \\ \underline{X}_i & \text{if } \overline{F}_i^{\uparrow} \leq 0 \\ (\overline{F}_i^{\uparrow}\,\overline{X}_i - \underline{F}_i^{\uparrow}\,\underline{X}_i)/\mathrm{w}(F_i^{\uparrow}) & \text{else} \end{cases}$$

$$c_i^{\downarrow} = \begin{cases} \underline{X}_i & \text{if } \underline{F}_i^{\downarrow} \geq 0 \\ \overline{X}_i & \text{if } \overline{F}_i^{\downarrow} \leq 0 \\ (\overline{F}_i^{\downarrow}\,\underline{X}_i - \underline{F}_i^{\downarrow}\,\overline{X}_i)/\mathrm{w}(F_i^{\downarrow}) & \text{else} \end{cases}$$

and

$$\begin{aligned} F_i^{\uparrow} &= H_{\partial_i f}(X_1, \ldots, X_i, c_{i+1}^{\uparrow}, \ldots, c_n^{\uparrow}) \\ F_i^{\downarrow} &= H_{\partial_i f}(X_1, \ldots, X_i, c_{i+1}^{\downarrow}, \ldots, c_n^{\downarrow}). \ \square \end{aligned}$$

In the sequel let $\boldsymbol{c}^{\uparrow}, \boldsymbol{c}^{\downarrow}$ and $\boldsymbol{F}^{\uparrow}, \boldsymbol{F}^{\downarrow}$ as in Definition 4.2.39.

Theorem 4.2.40 (Interval Extension) $\ddot{M}_f^{\curvearrowright}$ *is an interval extension of* f. $\square$

Proof. As $c_i^{\uparrow}, c_i^{\downarrow}$ are convex linear combinations of $\overline{X}_i$ and $\underline{X}_i$, it holds that $\boldsymbol{c}^{\uparrow}, \boldsymbol{c}^{\downarrow} \in \boldsymbol{X}$. From (4.2.6) it follows that

$$M_f^{\curvearrowright}(\boldsymbol{X}, \boldsymbol{c}^{\downarrow}) \cap M_f^{\curvearrowright}(\boldsymbol{X}, \boldsymbol{c}^{\uparrow}) \subseteq \ddot{M}_f^{\curvearrowright}(X).$$

According to (4.2.2)

$$f(\boldsymbol{X}) \subseteq M_f^{\curvearrowright}(\boldsymbol{X}, \boldsymbol{c})$$

for all $\boldsymbol{c} \in \boldsymbol{X}$. Hence,

$$f(X) \subseteq M_f^{\curvearrowright}(\boldsymbol{X}, \boldsymbol{c}^{\downarrow}) \cap M_f^{\curvearrowright}(\boldsymbol{X}, \boldsymbol{c}^{\uparrow}) \subseteq \ddot{M}_f^{\curvearrowright}(\boldsymbol{X}).$$

Further, $\ddot{M}_f^{\curvearrowright}(\boldsymbol{x}) = f(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^n$. $\square$

Theorem 4.2.41 (Non–Overestimation)

- *If* $0 \notin \mathrm{int}(F_i^{\uparrow})$ *for all* i *then*

$$\overline{\ddot{M}_f^{\curvearrowright}(\boldsymbol{X})} = \overline{f(\boldsymbol{X})}. \tag{4.2.7}$$

- *If* $0 \notin \mathrm{int}(F_i^{\downarrow})$ *for all* i *then*

$$\underline{\ddot{M}_f^{\curvearrowright}(\boldsymbol{X})} = \underline{f(\boldsymbol{X})}. \ \square \tag{4.2.8}$$

Proof. We give a proof of (4.2.7), the proof of (4.2.8) is analogous. According to Definition 4.2.39

$$\begin{aligned} \overline{\ddot{M}_f^{\curvearrowright}(\boldsymbol{X})} &= \overline{M_f^{\curvearrowright}(\boldsymbol{X}, \boldsymbol{c}^{\uparrow})} \\ &= f(\boldsymbol{c}^{\uparrow}) + \sum_{i=1}^{n} \overline{H_{\partial_i f}(X_1, \ldots, X_i, c_{i+1}^{\uparrow}, \ldots, c_n^{\uparrow})(X_i - c_i^{\uparrow})} \\ &= f(\boldsymbol{c}^{\uparrow}) + \sum_{i=1}^{n} \overline{F_i^{\uparrow} * (X_i - c_i^{\uparrow})}. \end{aligned}$$

Assume $0 \notin \mathrm{int}(F_i^{\uparrow})$ for all i. If $\underline{F}_i^{\uparrow} \geq 0$ then $c_i^{\uparrow} = \overline{X}_i$. If $\overline{F}_i^{\uparrow} \leq 0$ then $c_i^{\uparrow} = \underline{X}_i$. In both cases

$$\overline{F_i^{\uparrow} * (X_i - c_i^{\uparrow})} = 0,$$

hence, as $\boldsymbol{c}^{\uparrow} \in \boldsymbol{X}$,

$$\overline{\ddot{M}_f^{\curvearrowright}(\boldsymbol{X})} = f(\boldsymbol{c}^{\uparrow}) \in f(\boldsymbol{X}).$$

From $f(\boldsymbol{X}) \subseteq \ddot{M}_f^{\curvearrowright}(\boldsymbol{X})$, it follows that

$$\overline{\ddot{M}_f^{\curvearrowright}(\boldsymbol{X})} = \overline{f(\boldsymbol{X})}. \ \square$$

The centers $\boldsymbol{c}^{\uparrow}, \boldsymbol{c}^{\downarrow}$ are not optimal in the sense that $\boldsymbol{c}^{\uparrow}$ minimizes the upper bound of the successive mean value form and $\boldsymbol{c}^{\downarrow}$ maximizes the lower bound. In particular there are cases where $M_f^{\curvearrowright}$ is more accurate than $\ddot{M}_f^{\curvearrowright}$, i.e. $\boldsymbol{c} = \mathrm{mid}(\boldsymbol{X})$ leads to a smaller upper bound than $\boldsymbol{c}^{\uparrow}$ and to a larger lower bound than $\boldsymbol{c}^{\downarrow}$.

Theorem 4.2.42 (Comparison with Successive Mean Value Form) *There exist $f \in \mathbb{R}[x_1, \ldots, x_n]$ and $\boldsymbol{X} \in \mathbb{IR}^n$ such that*

$$M_f^{\curvearrowright}(\boldsymbol{X}) \subset \ddot{M}_f^{\curvearrowright}(\boldsymbol{X}). \ \square$$

Proof. Let $f(x_1, x_2) = x_1^2 x_2$ and let $\boldsymbol{X} = ([-1, 1], [-1, 1])$. Then

$$M_f^{\curvearrowright}(\boldsymbol{X}) = [-1, 1] \subset [-2, 2] = \ddot{M}_f^{\curvearrowright}(\boldsymbol{X}). \ \square$$

On the other hand, $\ddot{M}_f^{\curvearrowright}$ is always at least as accurate as $\ddot{M}_f$.

Theorem 4.2.43 (Comparison with Bicentered Mean Value Form) *For all $f \in \mathbb{R}[x_1, \ldots, x_n]$ and for all $\boldsymbol{X} \in \mathbb{IR}^n$ it holds that*

$$\ddot{M}_f^{\curvearrowright}(\boldsymbol{X}) \subseteq \ddot{M}_f(\boldsymbol{X}). \ \square$$

Proof. Let $f \in \mathbb{R}[x_1, \ldots, x_n]$ and let $\boldsymbol{X} \in \mathbb{IR}^n$ arbitrary but fixed. We have to show that

$$\overline{\ddot{M}_f^{\curvearrowright}(\boldsymbol{X})} \leq \overline{\ddot{M}_f(\boldsymbol{X})} \tag{4.2.9}$$

$$\underline{\ddot{M}_f^{\curvearrowright}(\boldsymbol{X})} \geq \underline{\ddot{M}_f(\boldsymbol{X})}. \tag{4.2.10}$$

We give a proof of (4.2.9), the proof of (4.2.10) is analogous. In order to avoid confusion of notation, we replace $c_i^{\uparrow}$ by $d_i^{\uparrow}$ and $F_i^{\uparrow}$ by $G_i^{\uparrow}$ for the successive bicentered mean value form and keep the symbols $c_i^{\uparrow}$, F_i for the bicentered mean value form. Thus,

$$d_i^{\uparrow} = \begin{cases} \overline{X}_i & \text{if } \underline{G_i^{\uparrow}} \geq 0 \\ \underline{X}_i & \text{if } \overline{G_i^{\uparrow}} \leq 0 \\ (\overline{G_i^{\uparrow}}\,\overline{X}_i - \underline{G_i^{\uparrow}}\,\underline{X}_i)/\mathrm{w}(G_i^{\uparrow}) & \text{else} \end{cases}$$

where

$$G_i^{\uparrow} = H_{\partial_i f}(X_1, \ldots, X_i, d_{i+1}^{\uparrow}, \ldots, d_n^{\uparrow})$$

and

$$c_i^{\uparrow} = \begin{cases} \overline{X}_i & \text{if } \underline{F}_i \geq 0 \\ \underline{X}_i & \text{if } \overline{F}_i \leq 0 \\ (\overline{F}_i\,\overline{X}_i - \underline{F}_i\,\underline{X}_i)/\mathrm{w}(F_i) & \text{else} \end{cases}$$

where

$$F_i = H_{\partial_i f}(X_1, \ldots, X_n).$$

We have to show that

$$f(\boldsymbol{d}^{\uparrow}) + \sum_{i=1}^{n} \overline{G_i^{\uparrow} * (X_i - d_i^{\uparrow})} \leq f(\boldsymbol{c}^{\uparrow}) + \sum_{i=1}^{n} \overline{F_i * (X_i - c_i^{\uparrow})}.$$

From the mean value Theorem it follows that there exists $\boldsymbol{\xi} \in \boldsymbol{X}$ such that

$$f(\boldsymbol{d}^{\uparrow}) = f(\boldsymbol{c}^{\uparrow}) + \sum_{i=1}^{n} \partial_i f(\boldsymbol{\xi})(d_i^{\uparrow} - c_i^{\uparrow}).$$

Thus,

$$\begin{aligned} f(\boldsymbol{d}^{\uparrow}) + \sum_{i=1}^{n} \overline{G_i^{\uparrow} * (X_i - d_i^{\uparrow})} &= f(\boldsymbol{c}^{\uparrow}) + \sum_{i=1}^{n} \partial_i f(\boldsymbol{\xi})(d_i^{\uparrow} - c_i^{\uparrow}) + \overline{G_i^{\uparrow} * (X_i - d_i^{\uparrow})} \\ &\leq f(\boldsymbol{c}^{\uparrow}) + \sum_{i=1}^{n} \overline{F_i * (d_i^{\uparrow} - c_i^{\uparrow})} + \overline{G_i^{\uparrow} * (X_i - d_i^{\uparrow})}. \end{aligned}$$

In order to complete the proof, we show for each i that

$$\overline{F_i * (d_i^{\uparrow} - c_i^{\uparrow})} + \overline{G_i^{\uparrow} * (X_i - d_i^{\uparrow})} \leq \overline{F_i * (X_i - c_i^{\uparrow})}.$$

Hence, let i arbitrary but fixed. We distinguish the cases whether or not $0 \in \operatorname{int}(G_i^{\uparrow})$.

- Assume $0 \notin \operatorname{int}(G_i^{\uparrow})$. From the definition of $d_i^{\uparrow}$ it follows that

  $$\overline{G_i^{\uparrow} * (X_i - d_i^{\uparrow})} = 0,$$

  hence

  $$\begin{aligned} \overline{F_i * (d_i^{\uparrow} - c_i^{\uparrow})} + \overline{G_i^{\uparrow} * (X_i - d_i^{\uparrow})} &= \overline{F_i * (d_i^{\uparrow} - c_i^{\uparrow})} \\ &\leq \overline{F_i * (X_i - c_i^{\uparrow})}. \end{aligned}$$

- Assume $0 \in \operatorname{int}(G_i^{\uparrow})$. As $G_i^{\uparrow} \subseteq F_i$, it holds that $0 \in \operatorname{int}(F_i)$ and from the definition of $d_i^{\uparrow}$ and $c_i^{\uparrow}$ it follows that

  $$\begin{aligned} \overline{G_i^{\uparrow} * (X_i - d_i^{\uparrow})} &= \overline{G}_i^{\uparrow} * (\overline{X}_i - d_i^{\uparrow}) &&= \underline{G}_i^{\uparrow} * (\underline{X}_i - d_i^{\uparrow}) \\ \overline{F_i * (X_i - c_i^{\uparrow})} &= \overline{F}_i * (\overline{X}_i - c_i^{\uparrow}) &&= \underline{F}_i * (\underline{X}_i - c_i^{\uparrow}). \end{aligned}$$

  Next, we distinguish the cases $d_i > c_i$ and $d_i \leq c_i$.

  - Assume $d_i > c_i$.

    $$\begin{aligned} \overline{F_i * (d_i^{\uparrow} - c_i^{\uparrow})} + \overline{G_i^{\uparrow} * (X_i - d_i^{\uparrow})} &= \overline{F}_i * (d_i^{\uparrow} - c_i^{\uparrow}) + \overline{G}_i^{\uparrow} * (\overline{X}_i - d_i^{\uparrow}) \\ &\leq \overline{F}_i * (d_i^{\uparrow} - c_i^{\uparrow}) + \overline{F}_i * (\overline{X}_i - d_i^{\uparrow}) \\ &= \overline{F}_i * (\overline{X}_i - c_i^{\uparrow}) \\ &= \overline{F_i * (X_i - c_i^{\uparrow})}. \end{aligned}$$

  - Assume $d_i \leq c_i$.

    $$\begin{aligned} \overline{F_i * (d_i^{\uparrow} - c_i^{\uparrow})} + \overline{G_i^{\uparrow} * (X_i - d_i^{\uparrow})} &= \underline{F}_i * (d_i^{\uparrow} - c_i^{\uparrow}) + \underline{G}_i^{\uparrow} * (\underline{X}_i - d_i^{\uparrow}) \\ &\leq \underline{F}_i * (d_i^{\uparrow} - c_i^{\uparrow}) + \underline{F}_i * (\underline{X}_i - d_i^{\uparrow}) \\ &= \underline{F}_i * (\underline{X}_i - c_i^{\uparrow}) \\ &= \overline{F_i * (X_i - c_i^{\uparrow})}. \quad \square \end{aligned}$$

Corollary 4.2.44 (Convergence) *$\ddot{M}_f^{\curvearrowright}$ converges quadratically to f.* $\square$

Theorem 4.2.45 (Inclusion Monotonicity) *$\ddot{M}_f^{\curvearrowright}$ is* not *inclusion monotone.* $\square$

Proof. Consider the polynomial $f(x_1, x_2) = x_1^2 x_2$. Obviously

$$([-1,1],[0,2])^{\mathrm{T}} \supset ([-1,1],[2,2])^{\mathrm{T}},$$

but

$$\ddot{M}_f^{\curvearrowright}([-1,1],[0,2]) = [0,2] \subset [-4,4] = \ddot{M}_f^{\curvearrowright}([-1,1],[2,2]). \quad \square$$

Algorithm 4.2.48 (SBM) computes the successive bicentered mean value form. First, we have to modify Algorithm 3.2.24 (OC), because now the centers $c_i^{\uparrow}$ and $c_i^{\downarrow}$ depend on different derivatives $F_i^{\uparrow}$ and $F_i^{\downarrow}$.

Algorithm 4.2.46 (OC') [Centers for Successive Bicentered Mean Value Form]

In: $F^{\uparrow}, F^{\downarrow} \in \mathbb{IF}$,
$X \in \mathbb{IF}$.

Out: $c^{\uparrow}, c^{\downarrow} \in X$, approximations of the centers according to Definition 4.2.39.

(1) [Computation of $c^{\uparrow}$.]

(1.1) [Monotonicity test.]
if $\underline{F^{\uparrow}} \geq 0$ then $c^{\uparrow} \longleftarrow \overline{X}$, goto Step 2.
if $\overline{F^{\uparrow}} \leq 0$ then $c^{\uparrow} \longleftarrow \underline{X}$, goto Step 2.

(1.2) [Approximate.]
clear invalid operation flag.
$c^{\uparrow} \longleftarrow (\overline{F}^{\uparrow} \overset{\approx}{*} \overline{X} \overset{\approx}{-} \underline{F}^{\uparrow} \overset{\approx}{*} \underline{X}) \overset{\approx}{/} (\overline{F}^{\uparrow} \overset{\approx}{-} \underline{F}^{\uparrow})$.

(1.3) [Check floating point errors.]
if invalid operation flag is raised then $c^{\uparrow} \longleftarrow \overline{X}$.
if $c^{\uparrow} < \underline{X}$ then $c^{\uparrow} \longleftarrow \underline{X}$.
if $c^{\uparrow} > \overline{X}$ then $c^{\uparrow} \longleftarrow \overline{X}$.

(2) [Computation of $c^{\downarrow}$.]

(2.1) [Monotonicity test.]
if $\underline{F^{\downarrow}} \geq 0$ then $c^{\downarrow} \longleftarrow \underline{X}$, goto Step 3.
if $\overline{F^{\downarrow}} \leq 0$ then $c^{\downarrow} \longleftarrow \overline{X}$, goto Step 3.

(2.2) [Approximate.]
clear invalid operation flag
$c^{\downarrow} \longleftarrow (\overline{F}^{\downarrow} \overset{\approx}{*} \underline{X} \overset{\approx}{-} \underline{F}^{\downarrow} \overset{\approx}{*} \overline{X}) \overset{\approx}{/} (\overline{F}^{\downarrow} \overset{\approx}{-} \underline{F}^{\downarrow})$.

(2.3) [Check floating point errors.]
if invalid operation flag is raised then $c^{\downarrow} \longleftarrow \underline{X}$.
if $c^{\downarrow} < \underline{X}$ then $c^{\downarrow} \longleftarrow \underline{X}$.
if $c^{\downarrow} > \overline{X}$ then $c^{\downarrow} \longleftarrow \overline{X}$.

(3) [Return.]
return $c^{\uparrow}, c^{\downarrow}$.

Theorem 4.2.47 (Complexity) *Algorithm 4.2.46 (OC') costs*

2 *number divisions,*
4 *number multiplications and*
4 *number additions.* □

In Algorithm 4.2.48 (SBM) the n-th variable is handled separately because $F_n^{\uparrow} = F_n^{\downarrow}$ and therefore the n-th partial derivative has to be computed only once.

Algorithm 4.2.48 (SBM) [Successive Bicentered Mean Value Form]

In: $f(\boldsymbol{x}) = \sum_{i=1}^{m} a_i \, x_1^{d_{i,1}} x_2^{d_{i,2}} \cdots x_n^{d_{i,n}} \in \mathbb{F}[x_1, \ldots, x_n]$,
$\boldsymbol{X} \in \mathbb{IF}^n$.

Out: $\mathrm{SBM}(f, \boldsymbol{X}) \in \mathbb{IF}$, $\mathrm{SBM}(f, \boldsymbol{X}) \approx \ddot{M}_f^{\sim}(\boldsymbol{X})$, $\mathrm{SBM}(f, \boldsymbol{X}) \supseteq f(\boldsymbol{X})$.

(1) [Last variable.]

(1.1) [n-th partial derivative.]
$F_n(x_1, \ldots, x_n) \longleftarrow \partial f(x_1, \ldots, x_n)/\partial x_n$.

(1.2) [Evaluate n-th partial derivative.]
$F_n \longleftarrow \mathrm{HFMI}(F_n(x_1,\ldots,x_n), X_1,\ldots,X_n)$.

(1.3) [Centers.]
$c_n^{\uparrow}, c_n^{\downarrow} \longleftarrow \mathrm{OC}(F_n, X_n)$.

(1.4) [Accumulate.]
$y^{\uparrow} \longleftarrow \overline{F_n * (X_n - c_n^{\uparrow})}$.
$y^{\downarrow} \longleftarrow \underline{F_n * (X_n - c_n^{\downarrow})}$.

(1.5) [Evaluate successively.]
$F^{\uparrow}(x_1,\ldots,x_{n-1}) \longleftarrow \mathrm{HFIL}(f(x_1,\ldots,x_n), [c_n^{\uparrow}])$.
$F^{\downarrow}(x_1,\ldots,x_{n-1}) \longleftarrow \mathrm{HFIL}(f(x_1,\ldots,x_n), [c_n^{\downarrow}])$.

(2) [Other variables.]
for $i = n-1,\ldots,1$ do

(2.1) [i-th partial derivative.]
$F_i^{\uparrow}(x_1,\ldots,x_i) \longleftarrow \partial F^{\uparrow}(x_1,\ldots,x_i)/\partial x_i$.
$F_i^{\downarrow}(x_1,\ldots,x_i) \longleftarrow \partial F^{\downarrow}(x_1,\ldots,x_i)/\partial x_i$.

(2.2) [Evaluate i-th partial derivative.]
$F_i^{\uparrow} \longleftarrow \mathrm{HFMI}(F_i^{\uparrow}(x_1,\ldots,x_i), X_1,\ldots,X_i)$.
$F_i^{\downarrow} \longleftarrow \mathrm{HFMI}(F_i^{\downarrow}(x_1,\ldots,x_i), X_1,\ldots,X_i)$.

(2.3) [Centers.]
$c_i^{\uparrow}, c_i^{\downarrow} \longleftarrow \mathrm{OC'}(F_i^{\uparrow}, F_i^{\downarrow}, X_i)$.

(2.4) [Accumulate.]
$y^{\uparrow} \longleftarrow y^{\uparrow} \mathbin{\hat{+}} \overline{F_i^{\uparrow} * (X_i - c_i^{\uparrow})}$.
$y^{\downarrow} \longleftarrow y^{\downarrow} \mathbin{\check{+}} \underline{F_i^{\downarrow} * (X_i - c_i^{\downarrow})}$.

(2.5) [Evaluate successively.]
$F^{\uparrow}(x_1,\ldots,x_{i-1}) \longleftarrow \mathrm{HFIL}(F^{\uparrow}(x_1,\ldots,x_i), [c_i^{\uparrow}])$.
$F^{\downarrow}(x_1,\ldots,x_{i-1}) \longleftarrow \mathrm{HFIL}(F^{\uparrow}(x_1,\ldots,x_i), [c_i^{\downarrow}])$.

(3) [Add $f(\boldsymbol{c}^{\uparrow}), f(\boldsymbol{c}^{\downarrow})$.]
$y^{\uparrow} \longleftarrow y^{\uparrow} \mathbin{\hat{+}} F^{\uparrow}()$.
$y^{\downarrow} \longleftarrow y^{\downarrow} \mathbin{\check{+}} F^{\downarrow}()$.
return $[y^{\downarrow}, y^{\uparrow}]$.

Theorem 4.2.49 (Complexity) *Algorithm 4.2.48 (*SBM*) costs*

$n^2m + 2nm$	*interval power computations,*
$n^2m + 2n$	*interval multiplications,*
$2n$	*number divisions,*
$8nm + 4n - 2m$	*number multiplications and*
$4nm + 6n + 2m - 3$	*number additions.* □

Proof. As in the proof of Theorem 4.2.12 we assume the worst case that during the successive evaluation of f in its last variable the number of monomials does not decrease. Hence, the successive evaluation of f in Step 1.5 and Step 2.5 costs $2nm$ interval power computations, $4nm$ number multiplications and $4m-4$ number additions. The remaining costs for Step 1 are:

- Step 1.1 costs $2m$ number multiplications.
- Step 1.2 costs nm interval power computations, nm interval multiplications and $2m-2$ number additions (Theorem 4.1.7).
- Step 1.3 costs 2 number divisions, 4 number multiplications and 3 number additions (Theorem 3.2.25).

- Step 1.4 costs 2 interval multiplications and 4 number additions.

Summarizing, Step 1.1 – Step 1.4 cost nm interval power computations, $nm+2$ interval multiplications, 2 number divisions, $2m+4$ number multiplications and $2m+5$ number additions. In the i-th iteration, the costs for Step 2.1 – Step 2.4 are:

- Step 2.1 costs $4m$ number multiplications.
- Step 2.2 costs $2im$ interval power computations, $2im$ interval multiplications and $4m-4$ number additions (Theorem 4.1.7).
- Step 2.3 costs 2 number divisions, 4 number multiplications and 4 number additions (Theorem 3.2.25).
- Step 2.4 costs 2 interval multiplications and 6 number additions.

Summarizing, in the i-th iteration Step 2.1 – Step 2.4 cost $2im$ interval power computations, $2im+2$ interval multiplications, 2 number divisions, $4m+4$ number multiplications and $4m+6$ number additions. Summing up for $i = n-1, \ldots, 1$ we obtain $n^2m - nm$ interval power computations, $n^2m - nm + 2n - 2$ interval multiplications, $2n-2$ number divisions, $4nm + 4n - 4m - 4$ number multiplications and $4nm + 6n - 4m - 6$ number additions. Finally, Step 3 costs 2 number additions. □

4.2.5 Experimental Results

We compare the mean value forms which were described in the previous sections experimentally at some classes of random polynomials and input intervals. As the computation of the range of a multivariate polynomial is expensive, we do not report overestimation errors. Instead, we compare the width of the refined mean value forms and the width of the ordinary mean value form. The cost for evaluating each mean value form is given in terms of arithmetic floating point instructions. The coefficients of the test polynomials are uniformly distributed in $[-1,1]$. The midpoints of the components of the input interval vectors are uniformly distributed in $[-1,1]$, the widths are uniformly distributed in $[0, 0.1]$. The reason why we use relatively small widths is that otherwise the Horner form or the nested form are preferable, as will be shown in Section 4.3. In each experiment we report the average over 1000 random polynomials and input interval vectors. Table 4.2.1 shows a comparison for random polynomials with different numbers of variables n and monomials m. Each degree vector contains n non–zero components and the degree in each variable is at most 5.

Accuracy. The intersection with H_f leads to a significant improvement of $M_f^{(\mathrm{H})}$ and $\ddot{M}_f^{(\mathrm{H})}$ compared to M_f respectively $\ddot{M}_f$, especially if the number of variables is large. In all cases $\ddot{M}_f^{\sim}$ is more accurate than $M_f^{\sim}$, $M_f^{(\mathrm{s})\sim}$ and $\ddot{M}_f$. The accuracy of $\ddot{M}_f$ is usually better than the accuracy of $M_f^{\sim}$, but with increasing n, the difference shrinks. In the last column, $M_f^{\sim}$ is more accurate than $\ddot{M}_f$.

Cost. The successive evaluation reduces the cost of $M_f^{(\mathrm{H})}$ and $\ddot{M}_f^{(\mathrm{H})}$ almost by half compared to M_f and $\ddot{M}_f$ respectively if the number of variables is large. Otherwise, the costs are comparable. The evaluation of $M_f^{\sim}$ is only about half as expensive as M_f, $\ddot{M}_f$ and $\ddot{M}_f^{\sim}$. These observations conform to the complexity considerations in the previous sections. The evaluation of $M_f^{(\mathrm{s})\sim}$ is even a bit cheaper than $M_f^{\sim}$.

Table 4.2.2 shows the same experiment, but this time the polynomials are such that each degree vector has only 2 non–zero components.

Accuracy. The accuracy is almost independent of the number of variables and monomials in all cases. Again, $\ddot{M}_f$ is significantly more accurate than $M_f^{\sim}$, even if the number of variables is high. The accuracy of $M_f^{(\mathrm{s})\sim}$ is between $M_f^{\sim}$ and $\ddot{M}_f$. The accuracies of $\ddot{M}_f^{\sim}$, $\ddot{M}_f$ and $\ddot{M}_f^{(\mathrm{H})}$ are comparable. Intersection with $H_f(\boldsymbol{X})$ leads to an improvement of $M_f^{(\mathrm{H})}$ over M_f but not of $\ddot{M}_f^{(\mathrm{H})}$ over $\ddot{M}_f$.

Cost. In contrast to the previous experiment, the costs of M_f and $M_f^{(\mathrm{H})}$ and the costs of $\ddot{M}_f$ and $\ddot{M}_f^{(\mathrm{H})}$ are comparable. As the degree vectors are sparse, the worst case assumptions in the complexity considerations in Theorem 4.2.12 and Theorem 4.2.38 are not adequate. Again, $M_f^{\curvearrowright}$ and $M_f^{\overset{(\mathrm{s})}{\curvearrowright}}$ are about half as expensive as $\ddot{M}_f$ and $\ddot{M}_f^{\curvearrowright}$.

4.3 Experimental Comparison

4.3.1 Efficiency and Accuracy for Random Polynomials

In this section we compare the accuracy of $H_f, N_f, M_f, \ddot{M}_f, M_f^{\curvearrowright}$ and $\ddot{M}_f^{\curvearrowright}$ experimentally. As the computation of the exact range of a multivariate polynomial is expensive, we compare the widths of the interval extensions and the width of the Horner form. More precisely, we report the average over 1000 random polynomials and inputs of

$$\mathrm{w}(F(\boldsymbol{X}))/\mathrm{w}(H_f(\boldsymbol{X})),$$

where F is one of $N_f, M_f, \ddot{M}_f, M_f^{\curvearrowright}$ or $\ddot{M}_f^{\curvearrowright}$.

For the experiments, we use 2 classes of random polynomials and 2 classes of input interval vectors.

- In both classes of random polynomials the number of variables is $n = 8$, each polynomial consists of $m = 10$ monomials and the degree in each variable is at most 5.
  - In the first class of random polynomials, all components of the degree vectors are randomly chosen (Figure 4.3.1,4.3.2).
  - In the second class, each degree vector has only two non–zero components (Figure 4.3.3,4.3.4).
- The input interval vectors $\boldsymbol{X}$ are such that $\mathrm{w}(X_i)$ is randomly chosen from $[0, w]$ for all i where w is a parameter.
  - In the first class of random input interval vectors we fix $\mathrm{mid}(X_i) \equiv 0.5$ for all i (Figure 4.3.1,4.3.3).
  - In the second class we fix $\mathrm{mid}(X_i) \equiv 0$ for all i (Figure 4.3.2,4.3.4).

It seems that the results of these and several other experiments can be generalized as follows:

- The accuracy of N_f is usually slightly better than the accuracy of H_f.
- If $\mathrm{mid}(X_i) \approx 0$ (Figure 4.3.2,4.3.4) then H_f, N_f, $M_f^{(\mathrm{H})}$ and $\ddot{M}_f^{(\mathrm{H})}$ are best. An explanation is that if $\mathrm{mid}(X_i) = 0$ then H_f is equivalent to the quadratically convergent multivariate version of the Taylor form. $M_f^{\overset{(\mathrm{s})}{\curvearrowright}}$ is only slightly less accurate.
- If $\mathrm{mid}(X_i) \approx 0$ (Figure 4.3.2,4.3.4) then $M_f^{\overset{(\mathrm{s})}{\curvearrowright}}$ is best among all mean value forms which are not intersected with the Horner form. In particular, $M_f^{\overset{(\mathrm{s})}{\curvearrowright}}$ and even $M_f^{\curvearrowright}$ are better than $\ddot{M}_f^{\curvearrowright}$. If all degree vectors are dense (Figure 4.3.2). and if $\mathrm{mid}(X_i) = 0$ then $c_i = 0$ and $\partial_i f(X_1, \ldots, X_i, c_{i+1}, \ldots, c_n) = 0$ for $i \neq n$. Hence,

  $$M_f^{\curvearrowright}(\boldsymbol{X}) = f(0, \ldots, 0) + H_{\partial_n f}(X_1, \ldots, X_n)(X_n - c_n)$$

  which explains the good performance of $M_f^{\curvearrowright}$ compared to the bicentered forms.
- If $\mathrm{mid}(X_i) \neq 0$ (Figure 4.3.1,4.3.3) then $\ddot{M}_f^{\curvearrowright}$ is more accurate than M_f, $M_f^{\curvearrowright}$, $M_f^{\overset{(\mathrm{s})}{\curvearrowright}}$ and $\ddot{M}_f$. If $\mathrm{w}(\boldsymbol{X})$ is small, then $\ddot{M}_f^{\curvearrowright}$ is more accurate than H_f and N_f because of its quadratic convergence. If the degree vectors are sparse (Figure 4.3.3) then $\ddot{M}_f$ is better than $M_f^{\curvearrowright}$.

Dense Monomials

Variables n, Monomials m	$2,3$	$2,6$	$4,4$	$4,8$	$8,5$	$8,10$	$16,6$	$16,12$
Flops for $M_f(\boldsymbol{X})$	124.0	197.9	500.5	878.4	2215.9	4147.4	10087.4	19530.9
Flops for $M_f^{(\mathrm{H})}(\boldsymbol{X})$	134.2	205.7	431.8	724.0	1555.4	2808.8	6103.0	11527.8
Flops for $M_f^{\sim}(\boldsymbol{X})$	102.6	152.9	342.8	563.4	1321.5	2366.3	5541.2	10442.2
Flops for $M_f^{(\mathrm{s})\sim}(\boldsymbol{X})$	92.9	123.0	300.4	437.3	1170.5	1941.9	5093.0	9241.7
Flops for $\ddot{M}_f(\boldsymbol{X})$	169.1	265.3	622.0	1075.5	2534.9	4686.5	10895.3	20908.6
Flops for $\ddot{M}_f^{(\mathrm{H})}(\boldsymbol{X})$	179.3	273.1	553.4	921.1	1874.3	3347.9	6910.9	12905.5
Flops for $\ddot{M}_f^{\sim}(\boldsymbol{X})$	155.2	231.6	574.2	944.0	2382.7	4262.7	10493.3	19748.4
$\mathrm{w}(M_f^{(\mathrm{H})}(\boldsymbol{X}))/\mathrm{w}(M_f(\boldsymbol{X}))$	0.810	0.876	0.617	0.697	0.348	0.393	0.160	0.176
$\mathrm{w}(M_f^{\sim}(\boldsymbol{X}))/\mathrm{w}(M_f(\boldsymbol{X}))$	0.919	0.909	0.767	0.764	0.558	0.572	0.319	0.330
$\mathrm{w}(M_f^{(\mathrm{s})\sim}(\boldsymbol{X}))/\mathrm{w}(M_f(\boldsymbol{X}))$	0.818	0.822	0.671	0.679	0.479	0.502	0.264	0.279
$\mathrm{w}(\ddot{M}_f(\boldsymbol{X}))/\mathrm{w}(M_f(\boldsymbol{X}))$	0.721	0.739	0.561	0.591	0.415	0.457	0.326	0.370
$\mathrm{w}(\ddot{M}_f^{(\mathrm{H})}(\boldsymbol{X}))/\mathrm{w}(M_f(\boldsymbol{X}))$	0.714	0.730	0.536	0.569	0.333	0.371	0.160	0.176
$\mathrm{w}(\ddot{M}_f^{\sim}(\boldsymbol{X}))/\mathrm{w}(M_f(\boldsymbol{X}))$	0.710	0.719	0.516	0.532	0.323	0.347	0.168	0.183

Table 4.2.1: Comparison of mean value forms for random polynomials with different numbers of variables and monomials. Each degree vector has n non–zero components and the degree in each variable is at most 5.

Sparse Monomials

Variables n, Monomials m	$2,3$	$2,6$	$4,4$	$4,8$	$8,5$	$8,10$	$16,6$	$16,12$
Flops for $M_f(\boldsymbol{X})$	124.0	197.9	239.1	416.3	437.8	788.2	841.1	1532.5
Flops for $M_f^{(\mathrm{H})}(\boldsymbol{X})$	134.2	205.7	256.6	434.3	454.3	800.7	841.7	1503.5
Flops for $M_f^{\sim}(\boldsymbol{X})$	102.6	152.9	192.7	318.6	335.8	581.4	614.9	1082.5
Flops for $M_f^{(\mathrm{s})\sim}(\boldsymbol{X})$	92.9	123.0	190.1	286.4	348.8	569.8	645.3	1101.2
Flops for $\ddot{M}_f(\boldsymbol{X})$	169.1	265.3	321.7	553.8	584.3	1041.6	1114.0	2006.0
Flops for $\ddot{M}_f^{(\mathrm{H})}(\boldsymbol{X})$	179.3	273.1	339.2	571.8	600.9	1054.1	1114.6	1976.9
Flops for $\ddot{M}_f^{\sim}(\boldsymbol{X})$	155.2	231.6	325.7	547.8	595.8	1057.1	1117.7	2020.3
$\mathrm{w}(M_f^{(\mathrm{H})}(\boldsymbol{X}))/\mathrm{w}(M_f(\boldsymbol{X}))$	0.810	0.876	0.829	0.884	0.825	0.879	0.818	0.860
$\mathrm{w}(M_f^{\sim}(\boldsymbol{X}))/\mathrm{w}(M_f(\boldsymbol{X}))$	0.919	0.909	0.935	0.927	0.938	0.935	0.944	0.941
$\mathrm{w}(M_f^{(\mathrm{s})\sim}(\boldsymbol{X}))/\mathrm{w}(M_f(\boldsymbol{X}))$	0.818	0.822	0.859	0.850	0.867	0.866	0.878	0.875
$\mathrm{w}(\ddot{M}_f(\boldsymbol{X}))/\mathrm{w}(M_f(\boldsymbol{X}))$	0.721	0.739	0.758	0.746	0.767	0.764	0.781	0.776
$\mathrm{w}(\ddot{M}_f^{(\mathrm{H})}(\boldsymbol{X}))/\mathrm{w}(M_f(\boldsymbol{X}))$	0.714	0.730	0.755	0.745	0.766	0.763	0.780	0.776
$\mathrm{w}(\ddot{M}_f^{\sim}(\boldsymbol{X}))/\mathrm{w}(M_f(\boldsymbol{X}))$	0.710	0.719	0.755	0.740	0.765	0.760	0.780	0.774

Table 4.2.2: Same experiment as before, but this time each degree vector has only 2 non–zero components.

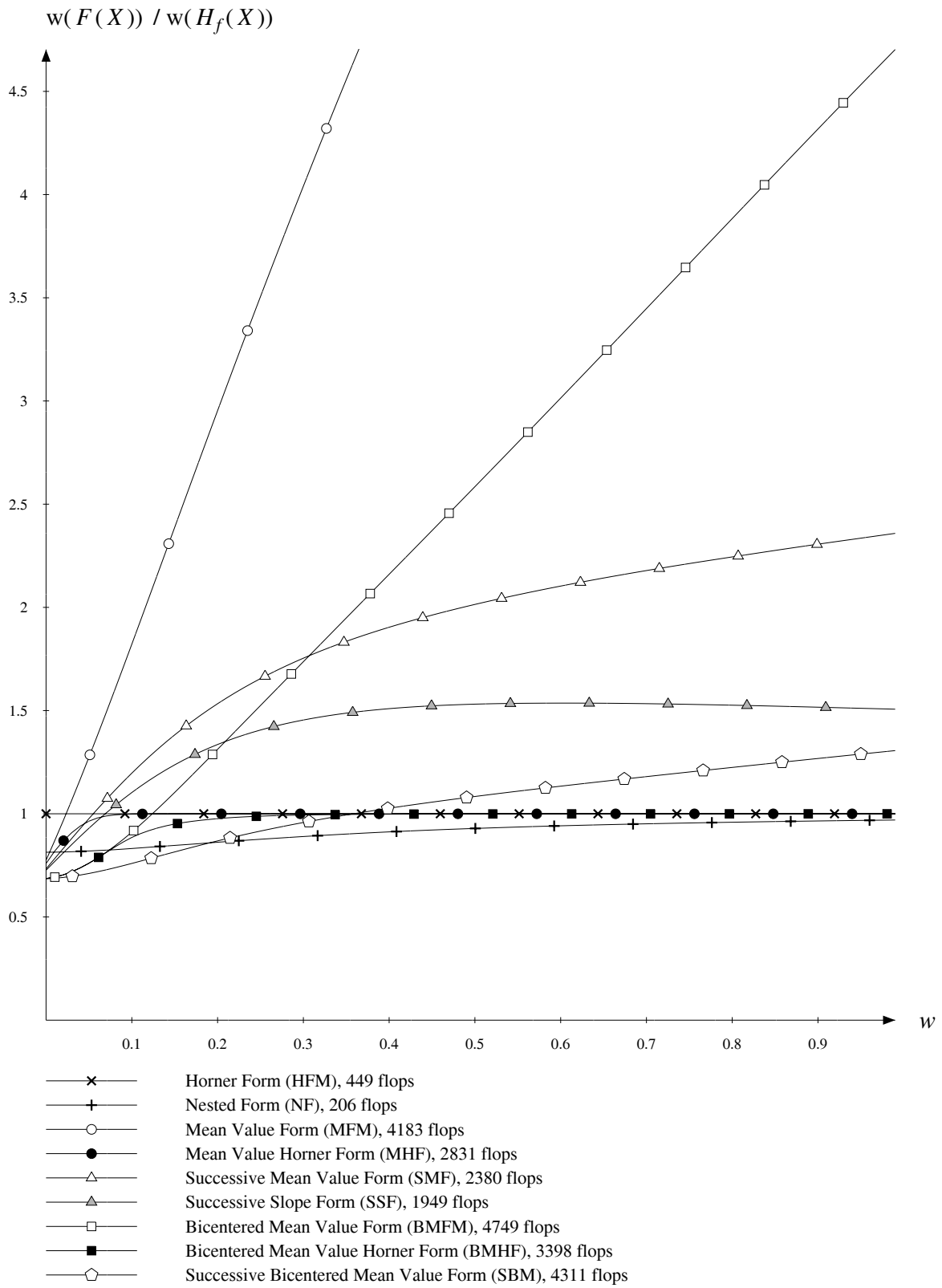


Figure 4.3.1: Accuracy for random polynomials with $n = 8$ variables, $m = 10$ monomials and degree ≤ 5 in each variable. The degree vectors have 8 non–zero components. The widths of the X_i are randomly chosen from $[0, w]$ and $\mathrm{mid}(X_i) \equiv 0.5$ for all i.

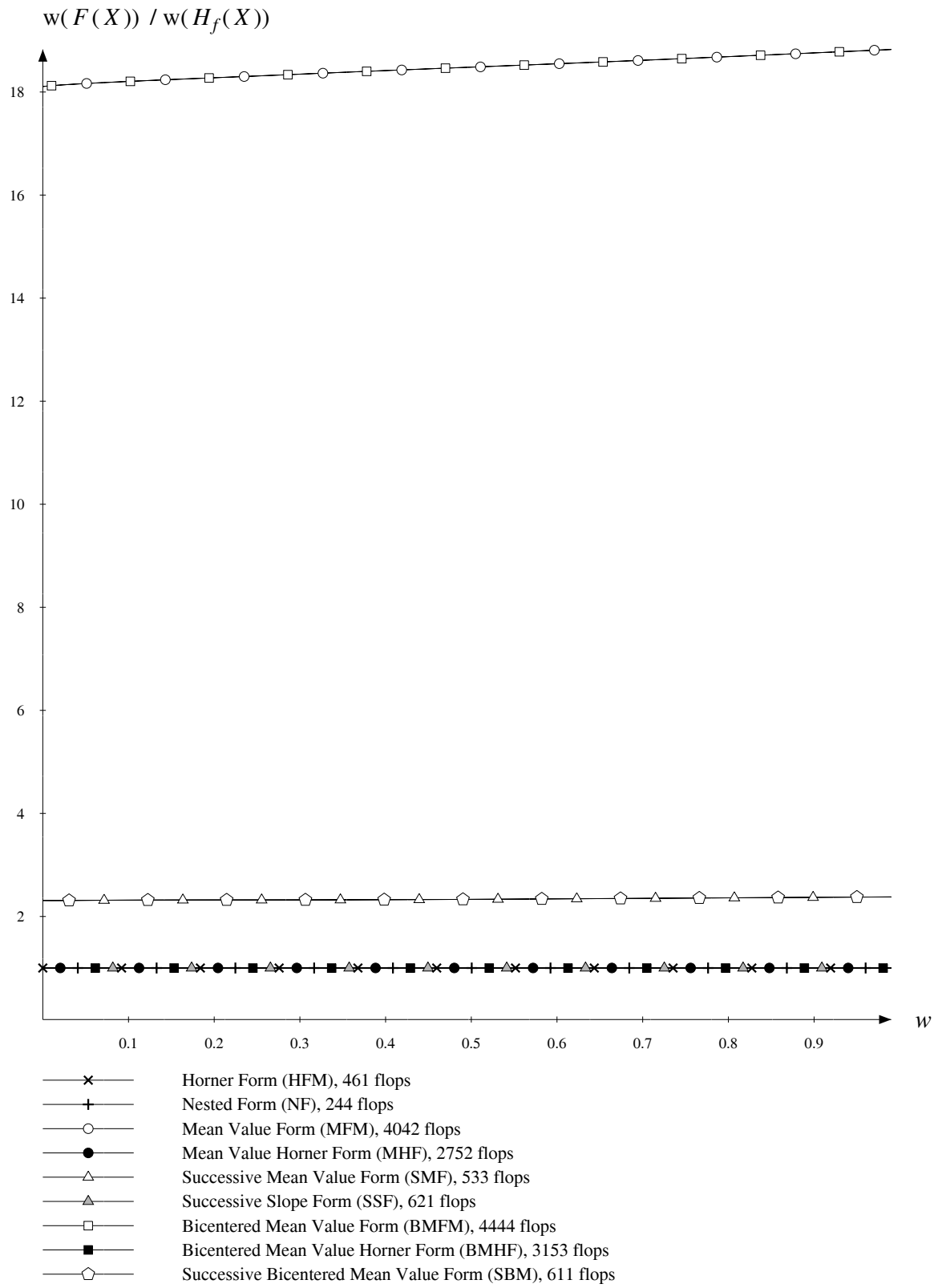


Figure 4.3.2: Accuracy for random polynomials with $n = 8$ variables, $m = 10$ monomials and degree ≤ 5 in each variable. The degree vectors have 8 non–zero components. The widths of the X_i are randomly chosen from $[0, w]$ and $\mathrm{mid}(X_i) \equiv 0$ for all i.

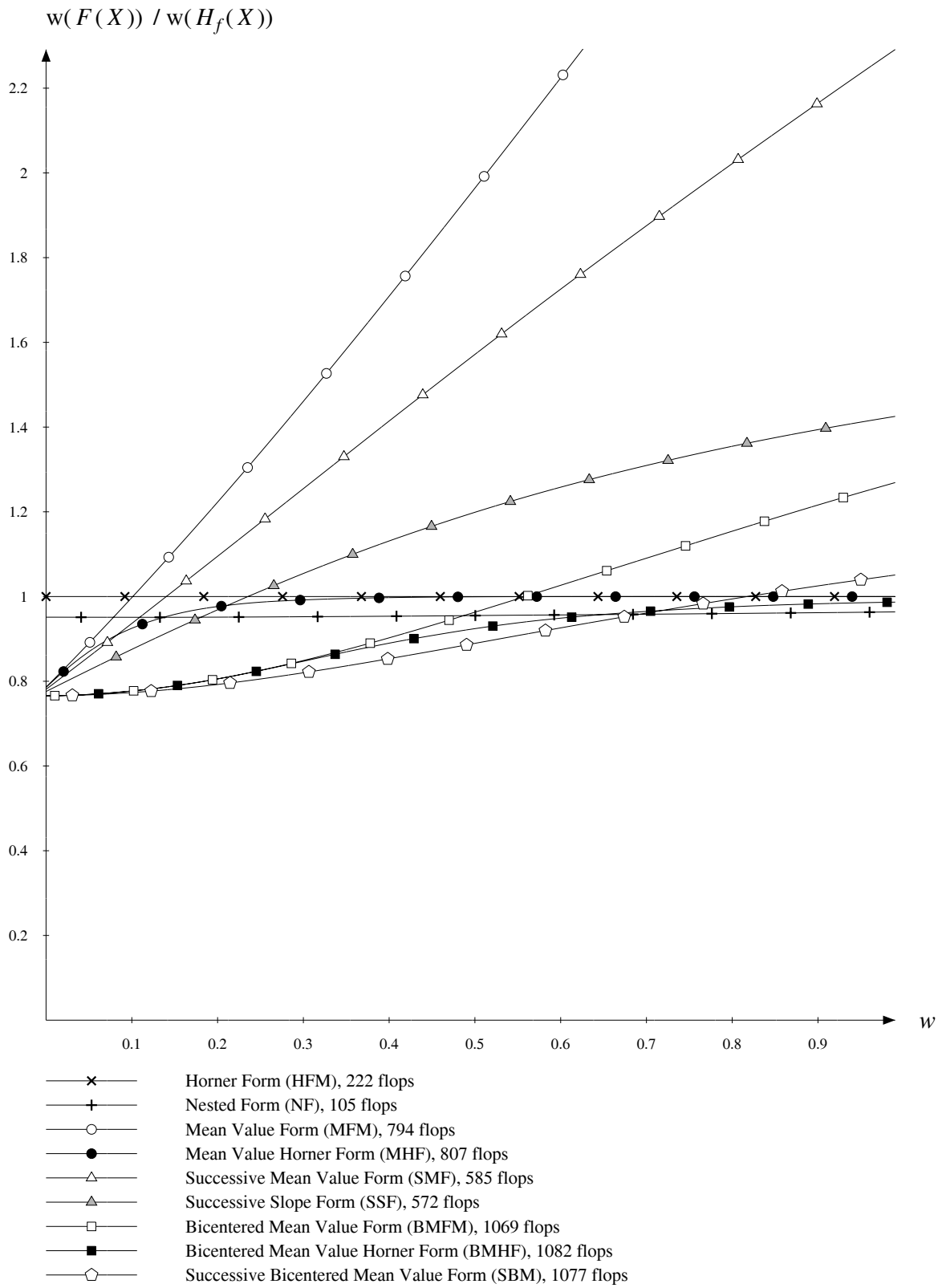


Figure 4.3.3: Accuracy for random polynomials with $n = 8$ variables, $m = 10$ monomials and degree ≤ 5 in each variable. The degree vectors have 2 non–zero components. The widths of the X_i are randomly chosen from $[0, w]$ and $\mathrm{mid}(X_i) \equiv 0.5$ for all i.

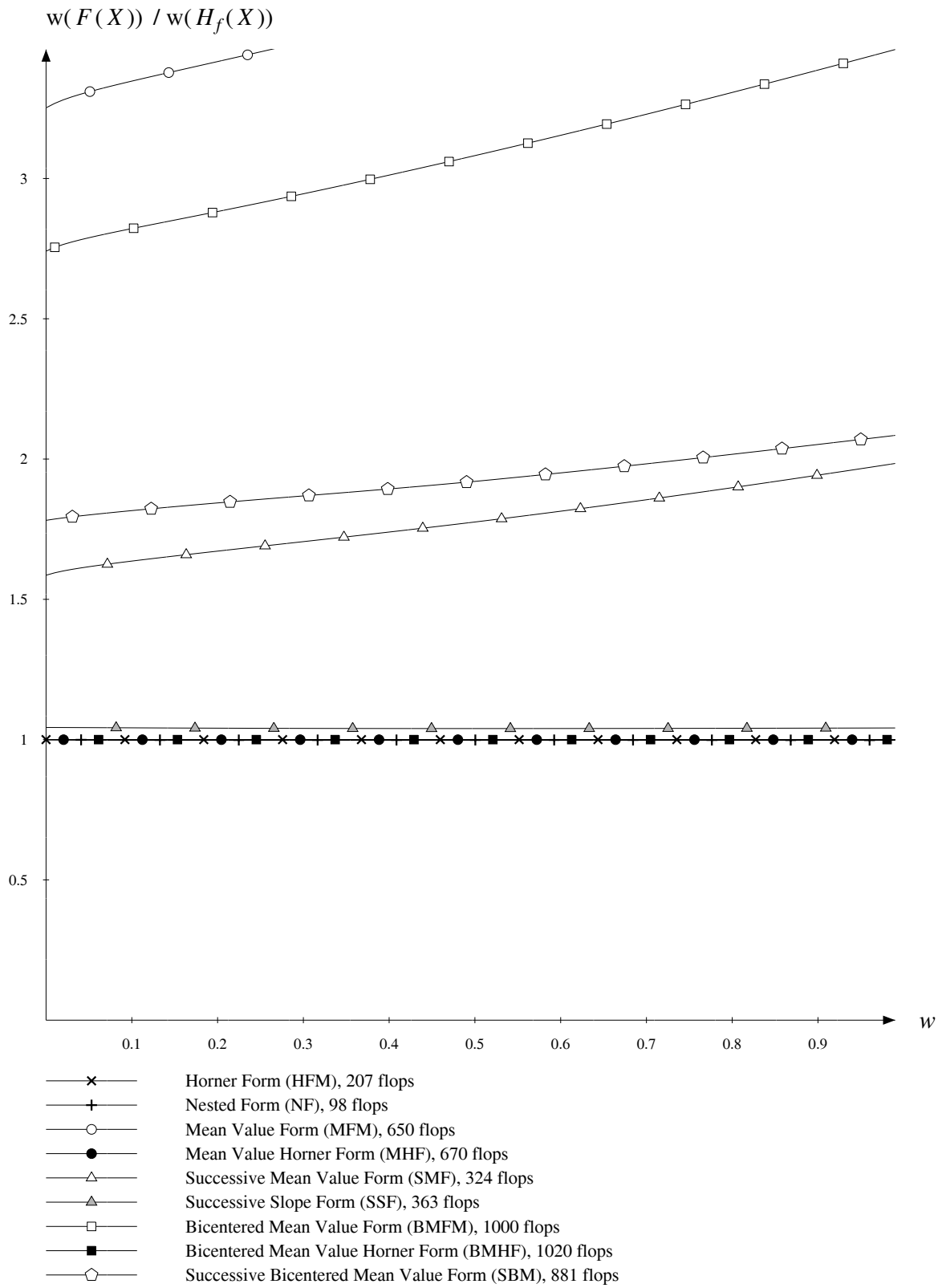


Figure 4.3.4: Accuracy for random polynomials with $n = 8$ variables, $m = 10$ monomials and degree ≤ 5 in each variable. The degree vectors have 2 non–zero components. The widths of the X_i are randomly chosen from $[0, w]$ and $\mathrm{mid}(X_i) \equiv 0$ for all i.

4.3.2 Global Optimization

In this section we compare interval extensions at the problem of finding the global minimum of a multivariate polynomial f in a given box $\boldsymbol{A}$. The problem is solved by a straight forward generalization of Algorithm 3.6.2 to the multivariate case. In Step 4 of this algorithm, we bisect in a coordinate direction where the width of $\boldsymbol{X}$ is maximal. Algorithm 3.6.2 is modified slightly in dependence of the interval extension which is used to bound $f(\boldsymbol{X})$:

- For M_f and $\ddot{M}_f$ the coefficients of the partial derivatives are computed only once and not in every iteration.
- Instead of $\ddot{M}_f(\boldsymbol{X})$, $\ddot{M}_f^{(\mathrm{H})}(\boldsymbol{X})$ and $\ddot{M}_f^{\sim}(\boldsymbol{X})$ we use the intervals

$$\begin{aligned}&[\underline{M_f(\boldsymbol{X}, \boldsymbol{c}^{\downarrow})}, H_f(\boldsymbol{c}^{\downarrow})]\\ &[\underline{M_f(\boldsymbol{X}, \boldsymbol{c}^{\downarrow})}, H_f(\boldsymbol{c}^{\downarrow})] \cap H_f(\boldsymbol{X})\\ &[M_f^{\sim}(\boldsymbol{X}, \boldsymbol{c}^{\downarrow}), H_f(\boldsymbol{c}^{\downarrow})],\end{aligned}$$

  where $\boldsymbol{c}^{\downarrow}$ is as in Definition 4.2.25 respectively Definition 4.2.39. In the case of $\ddot{M}_f^{\sim}$, this reduces the costs approximately by half.

- Instead of $M_f(\boldsymbol{X})$, $M_f^{(\mathrm{H})}(\boldsymbol{X})$, $M_f^{\sim}(\boldsymbol{X})$ and $M_f^{(\mathrm{s})\sim}(\boldsymbol{X})$ we use the intervals

$$\begin{aligned}&[\underline{M_f(\boldsymbol{X})}, H_f(\boldsymbol{c})]\\ &[\underline{M_f(\boldsymbol{X})}, H_f(\boldsymbol{c})] \cap H_f(\boldsymbol{X})\\ &[\underline{M_f^{\sim}(\boldsymbol{X})}, H_f(\boldsymbol{c})]\\ &[\underline{M_f^{(\mathrm{s})\sim}(\boldsymbol{X})}, H_f(\boldsymbol{c})],\end{aligned}$$

  which give tighter inclusions of the minimum with no additional costs.

The algorithms can be improved if intermediate results such as partially evaluated polynomials are memorized between the iterations. However, in order to keep the presentation simple we will not make use of this. In Figure 4.3.5 – 4.3.6 we trace the width of the inclusion of the global minimum in dependence of the number of executed arithmetic floating point operations. The figures show the average over 1000 random polynomials where each polynomial has $m = 5$ monomials and the degree in each variable is at most 5. The search interval $\boldsymbol{A}$ is in all cases the n-cube $[-1, 1] \times \cdots \times [-1, 1]$.

- In Figure 4.3.5 the polynomials have 3 variables and each degree vector has 2 non–zero components. The computation is stopped after 20 Kflop. The bicentered forms are best in the order $\ddot{M}_f^{\sim}$, $\ddot{M}_f^{(\mathrm{H})}$, $\ddot{M}_f$. Next comes N_f, which is slightly better than H_f and $M_f^{(\mathrm{s})\sim}$.
- In Figure 4.3.6 the polynomials have $n = 6$ variables and each degree vector has 2 non–zero components. The computation is stopped after 150 Kflop. As in Figure 4.3.5, the bicentered forms are best, but this time $\ddot{M}_f^{(\mathrm{H})}$ is slightly better than $\ddot{M}_f^{\sim}$. Intersection with the Horner form gives a significant improvement.
- In Figure 4.3.7 we repeat the experiment of Figure 4.3.6 but this time the degree vectors are dense. Here, N_f is clearly best, followed by H_f and $\ddot{M}_f^{(\mathrm{H})}$.

4.3.3 Solution of Systems of Nonlinear Equations

In this section we compare interval extensions at the problem of finding all solutions of a system of polynomial equations $\boldsymbol{f}(\boldsymbol{x}) : \mathbb{IR}^n \to \mathbb{IR}^n$ in a box $\boldsymbol{A}$. The algorithm is a simple bisection and range test algorithm in the case of H_f and N_f. Every mean value form gives rise to a Newton operator, where the

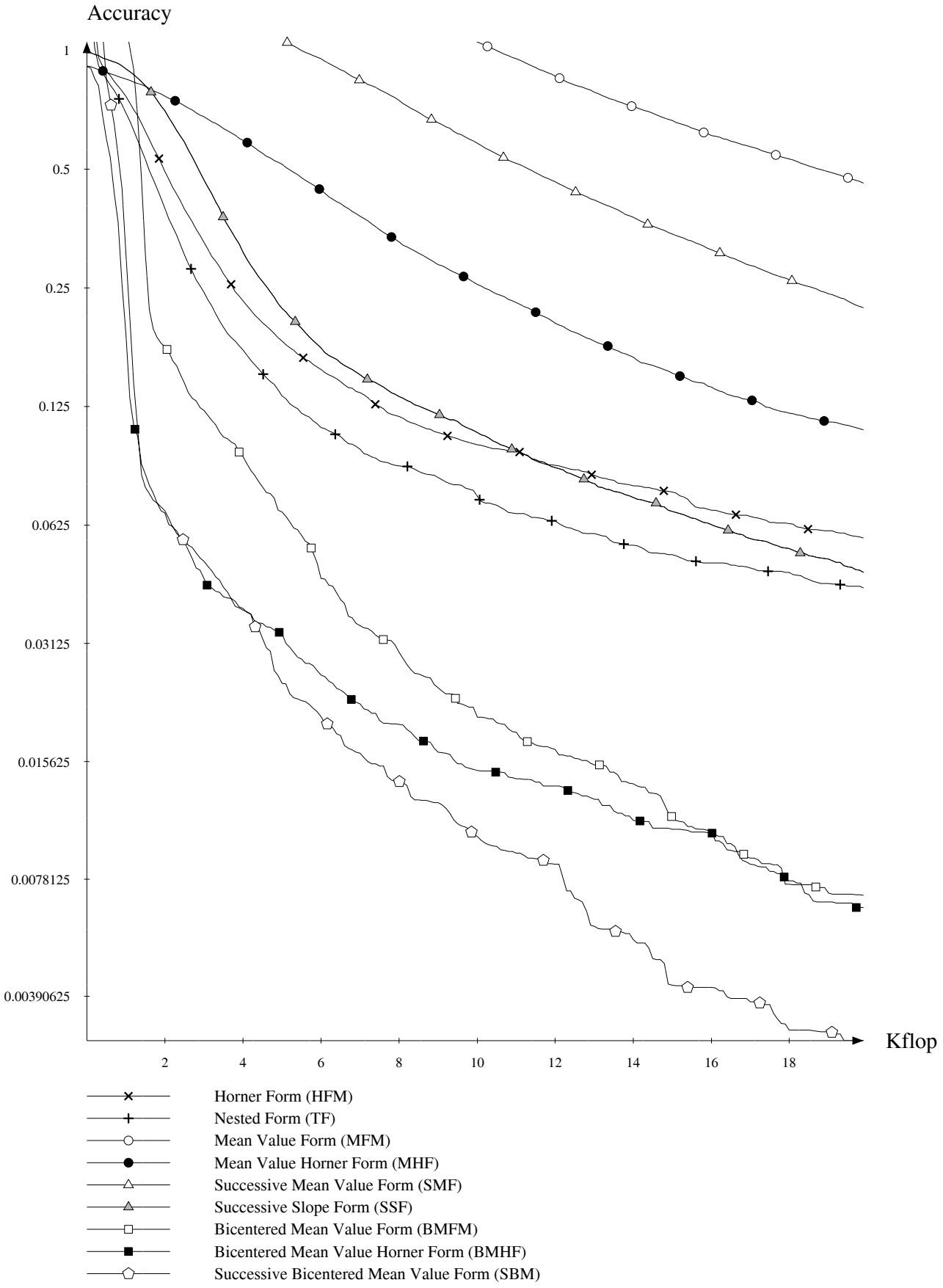


Figure 4.3.5: Optimization of random polynomials with $n = 3$ variables, $m = 5$ monomials and degree ≤ 5 in each variable. The degree vectors have 2 non–zero components.

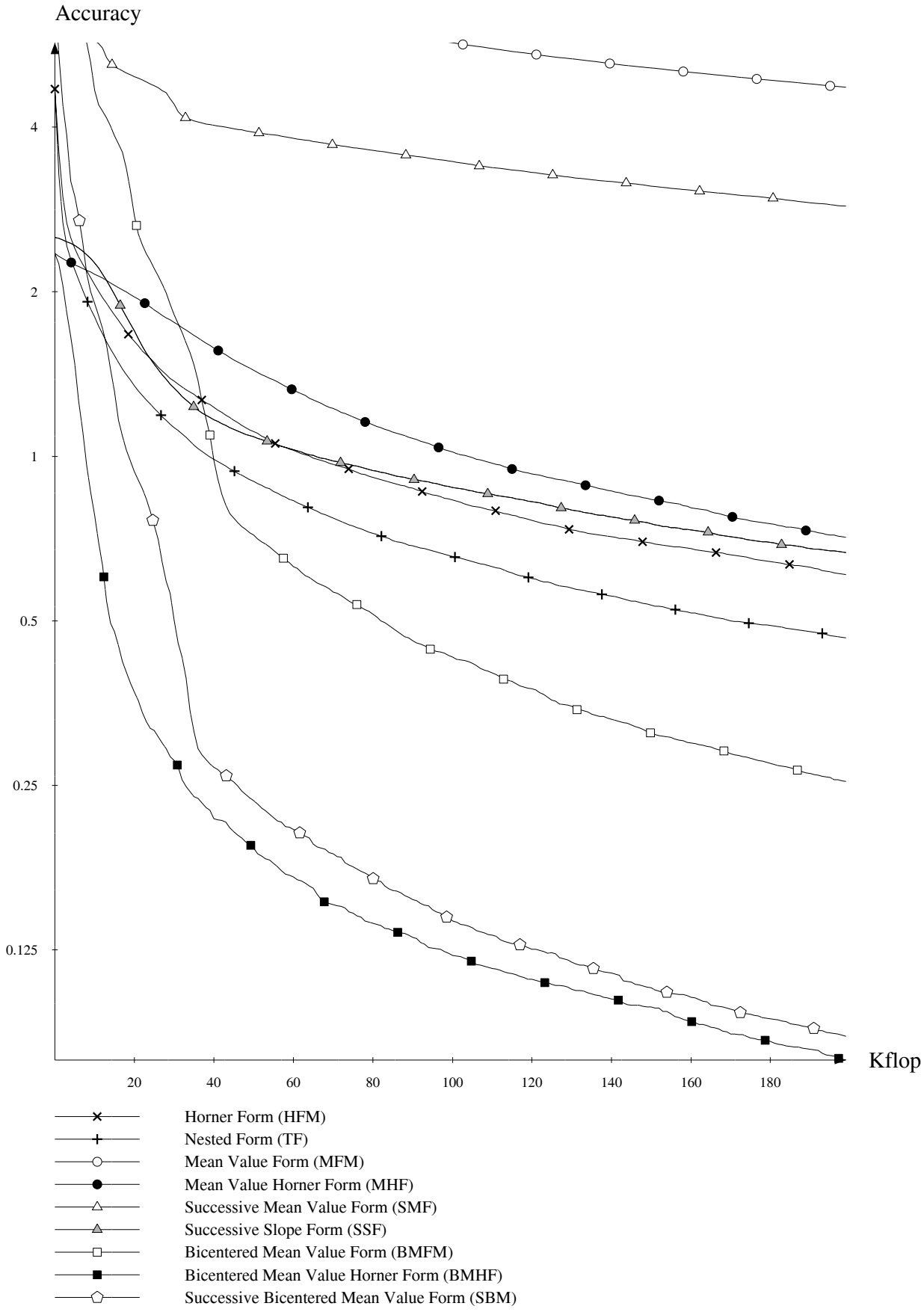


Figure 4.3.6: Optimization of random polynomials with $n = 6$ variables, $m = 5$ monomials and degree ≤ 5 in each variable. The degree vectors have 2 non–zero components.

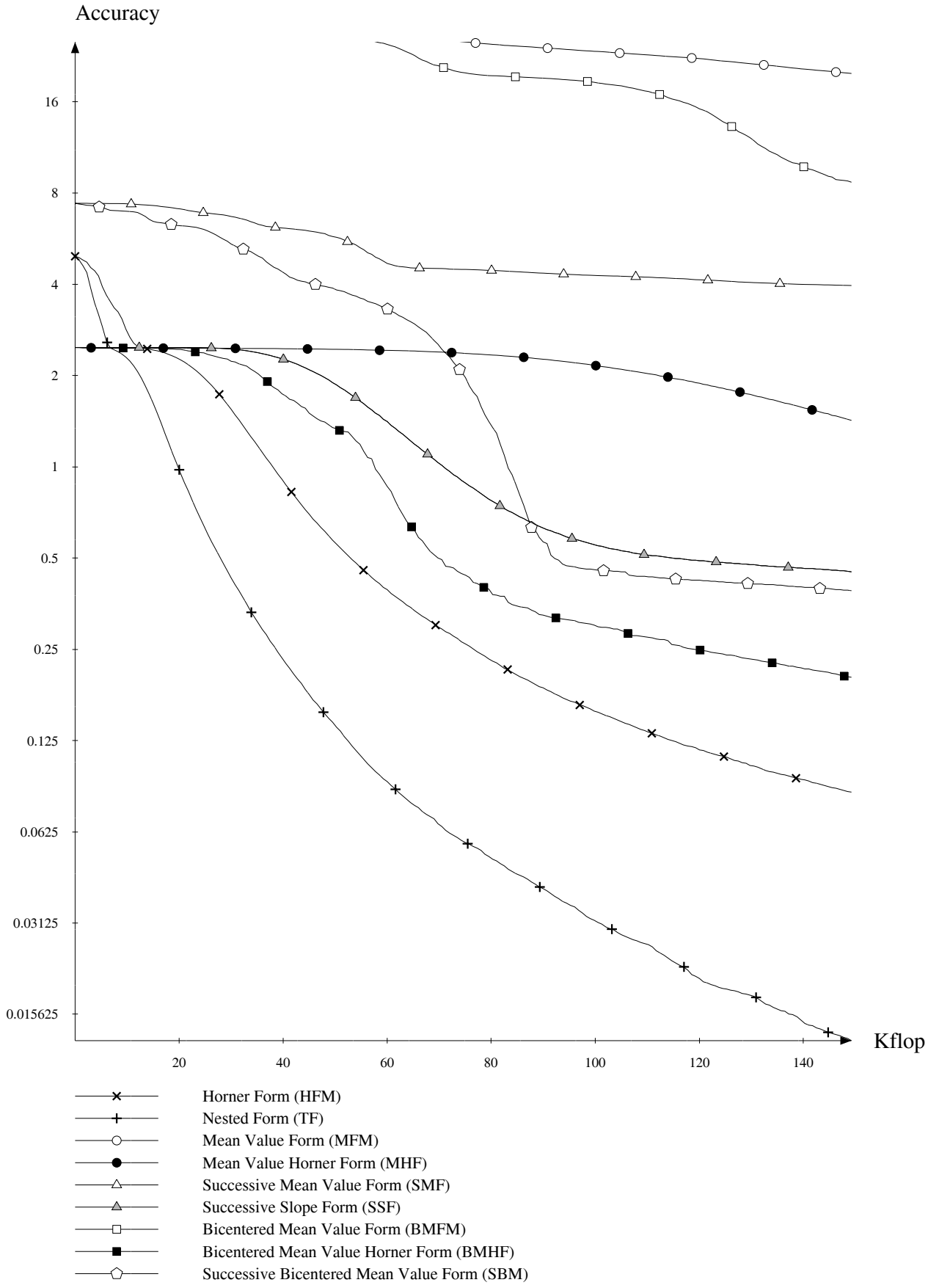


Figure 4.3.7: Optimization of random polynomials with $n = 6$ variables, $m = 5$ monomials and degree ≤ 5 in each variable. The degree vectors have 6 non–zero components.

Jacobian matrix is obtained in the same way as the partial derivatives or slopes for the corresponding mean value form. We are using the Hansen–Sengupta operator [Hansen and Sengupta, 1981], which is an interval version of the Newton–Gauss–Seidel method.

The bicentered mean value forms are problematic. First, it is not clear which center to use during the Newton step. Second, for each component of $\boldsymbol{f}$ the optimal centers are usually different, which makes preconditioning as it is used for the other forms impossible. Therefore, as in the case of the ordinary mean value form, we are using the midpoint for the Newton step and take the optimal centers only for a range test. Note that this is not possible for the successive bicentered mean value form because here we have slopes at the optimal centers and not a derivative.

First, we give an algorithm for the simple bisection and range test method. The algorithm terminates if all solutions are enclosed by interval vectors whose total volume is less then a given constant ε. The volume $\mathrm{w}^*(\boldsymbol{X})$ of an interval vector $\boldsymbol{X} \in \mathbb{IR}^n$ is defined as

$$\mathrm{w}^*(\boldsymbol{X}) = \prod_{i=1}^{n} \mathrm{w}(X_i).$$

The volume was chosen as a termination criterion because the sum of volumes of a set of interval vectors remains unchanged if some element is bisected.

Algorithm 4.3.1 (SBR) [Solving with Bisection and Range Test]

In: $\boldsymbol{f}(\boldsymbol{x}) \in \mathbb{IF}[x_1, \ldots, x_n]^n$,
$\boldsymbol{A} \in \mathbb{IF}^n$ such that $\boldsymbol{f}$ has only finitely many solutions in $\boldsymbol{A}$,
$\varepsilon \in \mathbb{F}$, $\varepsilon > 0$.

Out: $\mathcal{L} = \{\boldsymbol{Z}^{(j)} \in \mathbb{IF}^n \mid j = 1, \ldots, k\}$, such that

- $\{\boldsymbol{z} \in \boldsymbol{A} \mid \boldsymbol{f}(\boldsymbol{z}) = 0\} \subseteq \bigcup_{j=1}^{k} \boldsymbol{Z}^{(j)}$
- $\sum_{j=1}^{k} \mathrm{w}^*(\boldsymbol{Z}^{(k)}) \leq \varepsilon$.

(1) [Initialize.]
$\mathcal{L} \longleftarrow \{\boldsymbol{A}\}$.

(2) [Termination test.]
if $\sum_{\boldsymbol{Z} \in \mathcal{L}} \mathrm{w}^*(\boldsymbol{Z}) \leq \varepsilon$ then return $\mathcal{L}$.

(3) [Take largest box.]
$\boldsymbol{X} \longleftarrow$ an element of $\mathcal{L}$ with largest volume.
remove $\boldsymbol{X}$ from $\mathcal{L}$.

(4) [Range test.]
for $i = 1, \ldots, n$
$Y \longleftarrow$ overestimation of $f_i(\boldsymbol{X})$.
if $0 \notin Y$ goto Step 2.

(5) [Bisect.]
bisect $\boldsymbol{X}$ at the midpoint in a direction where its width is largest into $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$.

(6) [Store the sub–boxes.]
$\mathcal{L} \longleftarrow \mathcal{L} \cup \{\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}\}$.
goto Step 2.

Next, we give an algorithm which solves systems using bisection, range test and Newton's method, see [Hansen and Sengupta, 1981], [Neumaier, 1990]. The specification of the algorithm is the same as before.

Algorithm 4.3.2 (SBRN) [Solving with Bisection, Range Test and Newton's Method]

In: $\boldsymbol{f}(\boldsymbol{x}) \in \mathbb{IF}[x_1, \ldots, x_n]^n$,
$\boldsymbol{A} \in \mathbb{IF}^n$ such that $\boldsymbol{f}$ has only finitely many solutions in $\boldsymbol{A}$,
$\varepsilon \in \mathbb{F}$, $\varepsilon > 0$.

Out: $\mathcal{L} = \{\boldsymbol{Z}^{(j)} \in \mathbb{IF}^n \mid j = 1, \ldots, k\}$, such that

- $\{\boldsymbol{z} \in \boldsymbol{A} \mid \boldsymbol{f}(\boldsymbol{z}) = 0\} \subseteq \bigcup_{j=1}^{k} \boldsymbol{Z}^{(j)}$
- $\sum_{j=1}^{k} \mathrm{w}^*(\boldsymbol{Z}^{(k)}) \leq \varepsilon$.

(1) [Initialize.]
$\mathcal{L} \longleftarrow \{\boldsymbol{A}\}$

(2) [Termination test.]
if $\sum_{\boldsymbol{Z} \in \mathcal{L}} \mathrm{w}^*(\boldsymbol{Z}) \leq \varepsilon$ then return $\mathcal{L}$.

(3) [Take largest box.]
$\boldsymbol{X} \longleftarrow$ an element of $\mathcal{L}$ with largest volume.
Remove $\boldsymbol{X}$ from $\mathcal{L}$

(5) [Gradients and range test.]
$\boldsymbol{c} \longleftarrow \mathrm{mid}(\boldsymbol{X})$
for $i = 1, \ldots, n$

(5.1) [Gradients.]
for $j = 1, \ldots, n$ do $J_{ij} \longleftarrow$ overestimation of derivative or slope of f_i w.r.t. x_j evaluated at $\boldsymbol{X}$.

(5.2) [Range test.]
$Y_i \longleftarrow$ overestimation of $f_i(\boldsymbol{c})$.
if $0 \notin Y_i + \sum_{j=1}^{n} J_{ij}(X_j - c_j)$ then goto Step 2.

(6) [Solve linear system.]
$\boldsymbol{X}' \longleftarrow \boldsymbol{X}$.

(6.1) [Precondition.]
$\mathfrak{m} \longleftarrow$ approximation of $\mathrm{mid}(\boldsymbol{J})^{-1}$.
$\boldsymbol{J} \longleftarrow \mathfrak{m}\boldsymbol{J}$.
$\boldsymbol{Y} \longleftarrow \mathfrak{m}\boldsymbol{Y}$.

(6.2) [Gauss–Seidel iteration.]
for $i = 1, \ldots, n$
$Q_1, Q_2 \longleftarrow \mathrm{GDIV}(-Y_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} J_{ij}(X_j - c_j),\ J_{ii}, X_i - c_i)$.
$Q_1 \longleftarrow (Q_1 + c_i) \cap X_i$.
$Q_2 \longleftarrow (Q_2 + c_i) \cap X_i$.
if $Q_1 \neq \emptyset$ and $Q_2 \neq \emptyset$ then $G_i = [\overline{Q_1}, \underline{Q_2}]$, else $G_i = \emptyset$.
if $Q_1 = \emptyset$ and $Q_2 = \emptyset$ then goto Step 2.
$X_i \longleftarrow [Q_1 \cup Q_2]$.

(7) [No bisection if sufficient improvement.]
if $\mathrm{w}^+(\boldsymbol{X}) \leq 0.9\mathrm{w}^+(\boldsymbol{X}')$ then $\mathcal{L} \longleftarrow \mathcal{L} \cup \{\boldsymbol{X}\}$, goto Step 2.

(8) [Bisect.]
if $G_i = \emptyset$ for all $i = 1, \ldots, n$ then
bisect $\boldsymbol{X}$ at the midpoint in a direction of largest width into $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$.
else
let i such that $\mathrm{w}(G_i)$ is maximal.
$\boldsymbol{X}^{(1)} \longleftarrow (X_1, \ldots, [\underline{X_i}, \underline{G_i}], \ldots, X_n)$.
$\boldsymbol{X}^{(2)} \longleftarrow (X_1, \ldots, [\overline{G_i}, \overline{X_i}], \ldots, X_n)$.
$\mathcal{L} \longleftarrow \mathcal{L} \cup \{\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}\}$.
goto Step 2.

The value of J_{ij} in Step 5.1 corresponds to the the partial derivatives or slopes of the chosen interval extension.

The pictures on the following pages show the total volume of the boxes in the list $\mathcal{L}$ of Algorithm 4.3.1 (SBR) and Algorithm 4.3.2 (SBRN) in dependence of the number of executed floating point instructions at some sample polynomial systems from the literature. The costs of Step 2 for computing the total volume of all boxes in the list is neglected in both algorithms.

- Brown's almost linear system:

$$\begin{aligned} x_1 + x_1 + x_2 + x_3 + x_4 + x_5 - 6 &= 0 \\ x_2 + x_1 + x_2 + x_3 + x_4 + x_5 - 6 &= 0 \\ x_3 + x_1 + x_2 + x_3 + x_4 + x_5 - 6 &= 0 \\ x_4 + x_1 + x_2 + x_3 + x_4 + x_5 - 6 &= 0 \\ x_1 x_2 x_3 x_4 x_5 - 1 &= 0 \end{aligned}$$

  Initial box: $[-2, 2]^5$.
  There are two solutions within the box. The Jacobian is ill–conditioned at these roots, see [Kearfott, 1987], [Morgan, 1983], [Morgan and Shapiro, 1987]. (Figure 4.3.8)

- Combustion chemistry problem.

$$\begin{aligned} a_1 x_2 x_4 + a_2 x_2 + a_3 x_1 x_4 + a_4 x_1 + a_5 x_4 &= 0 \\ b_1 x_2 x_4 + b_2 x_1 x_3 + b_3 x_1 x_4 + b_4 x_3 x_4 + b_5 x_3 + b_6 x_4 + b_7 &= 0 \\ x_1^2 - x_2 &= 0 \\ x_4^2 - x_3 &= 0 \end{aligned}$$

$$\begin{array}{llllll} a_1 = -1.697e7 & a_2 = 2.177 \cdot 10^7 & a_3 = 0.55 \\ a_4 = 0.45 & a_5 = -1.0 & b_1 = 1.585 \cdot 10^{14} \\ b_2 = 4.126 \cdot 10^7 & b_3 = -8.285 \cdot 10^6 & b_4 = 2.284 \cdot 10^7 \\ b_5 = -1.918 \cdot 10^7 & b_6 = 48.4 & b_7 = -27.73 \end{array}$$

  Initial box: $[0, 1]^4$.
  There is a unique solution within the box, see [Kearfott, 1987], [Morgan and Shapiro, 1987]. (Figure 4.3.9)

- Robot kinematics problem.

$$\begin{aligned} A_1 x_1 x_3 + A_2 x_2 x_3 + A_3 x_1 + A_4 x_2 + A_5 x_4 + A_6 x_7 + A_7 &= 0 \\ A_8 x_1 x_3 + A_9 x_2 x_3 + A_{10} x_1 + A_{11} x_2 + A_{12} x_4 + A_{13} &= 0 \\ A_{14} x_6 x_8 + A_{15} x_1 + A_{16} x_2 &= 0 \\ A_{17} x_1 + A_{18} x_2 + A_{19} &= 0 \\ x_1^2 + x_2^2 - 1 &= 0 \\ x_3^2 + x_2^4 - 1 &= 0 \\ x_5^2 + x_2^6 - 1 &= 0 \\ x_7^2 + x_2^8 - 1 &= 0 \end{aligned}$$

$$\begin{array}{llllll} A_1 = 4.731 \cdot 10^{-3} & A_2 = -0.3578 & A_3 = -0.1238 \\ A_4 = -1.637 \cdot 10^{-3} & A_5 = -0.9338 & A_6 = 1.0 \\ A_7 = -0.3571 & A_8 = 0.2238 & A_9 = 0.7623 \\ A_{10} = 0.2638 & A_{11} = -0.7745 \cdot 10^{-1} & A_{12} = -0.6734 \\ A_{13} = -0.6022 & A_{14} = 1.0 & A_{15} = 0.3578 \\ A_{16} = 4.731 \cdot 10^{-3} & A_{17} = -0.7623 & A_{18} = 0.2238 \\ A_{19} = 0.3461 \end{array}$$

  Initial box: $[-1, 1]^8$. There are 16 solutions within the box, see, [Kearfott, 1987], [Tsai and Morgan, 1984]. (Figure 4.3.10)

In the first two examples, $M_f^{\sim}$ and $M_f^{(s)\sim}$ are clearly best. In the robot kinematics problem N_f is best, followed by H_f, at least after the first 50 Mflops. At this time the inclusion of the solutions is still not very precise.

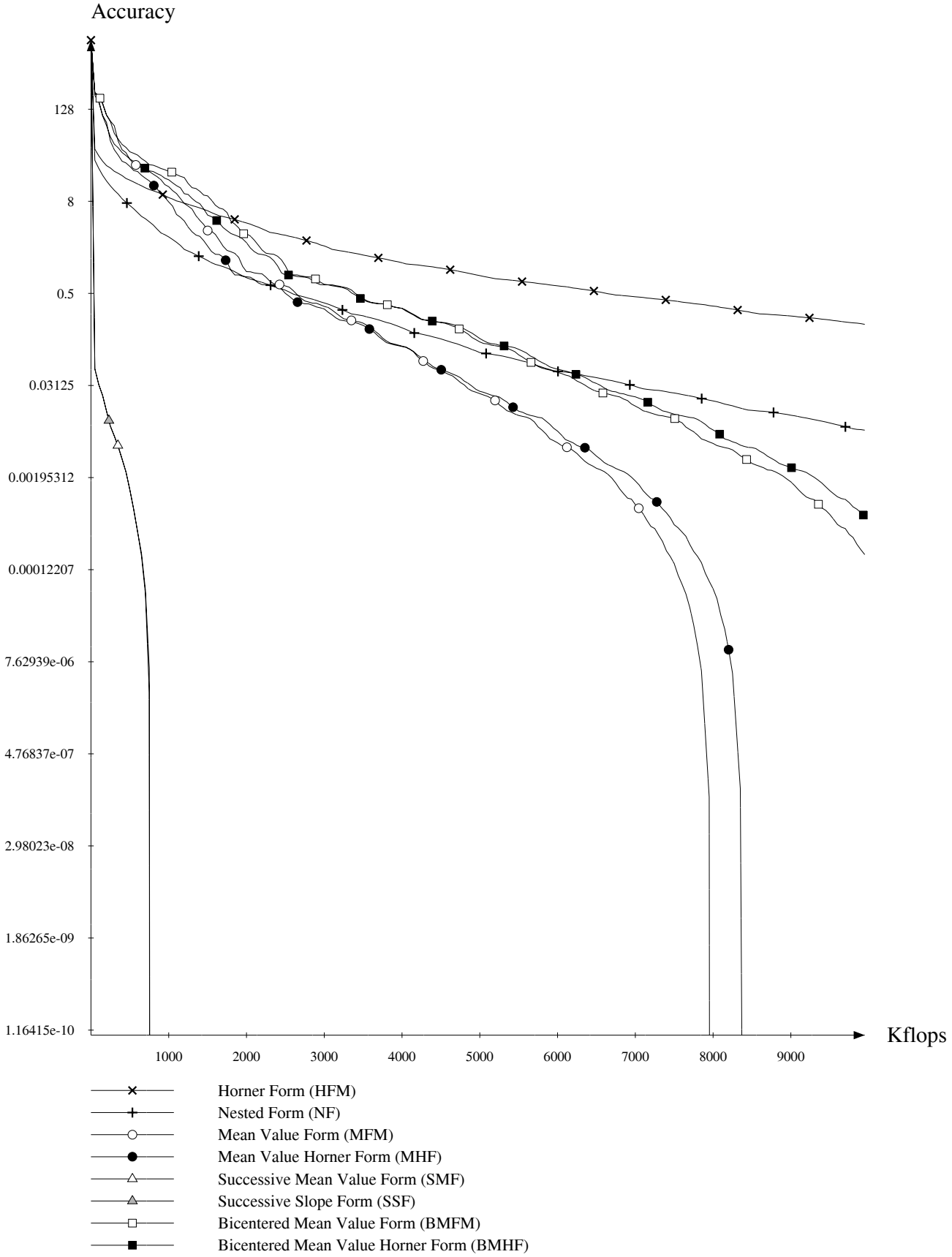


Figure 4.3.8: Brown's almost linear system, 5 variables. There are two solutions within the box. The Jacobian is ill–conditioned at these roots.

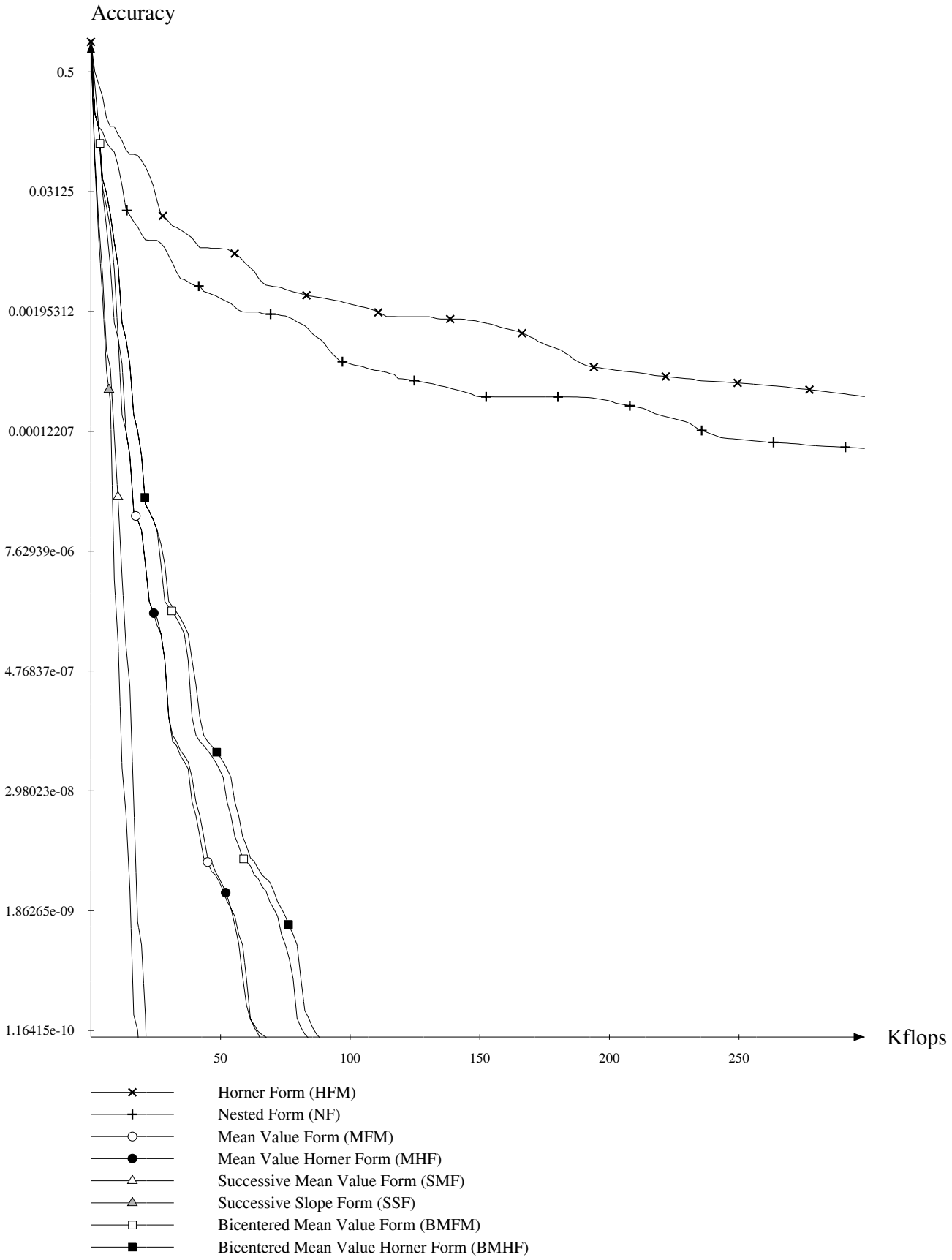


Figure 4.3.9: Combustion chemistry problem, 4 variables. There is a unique solution within the box.

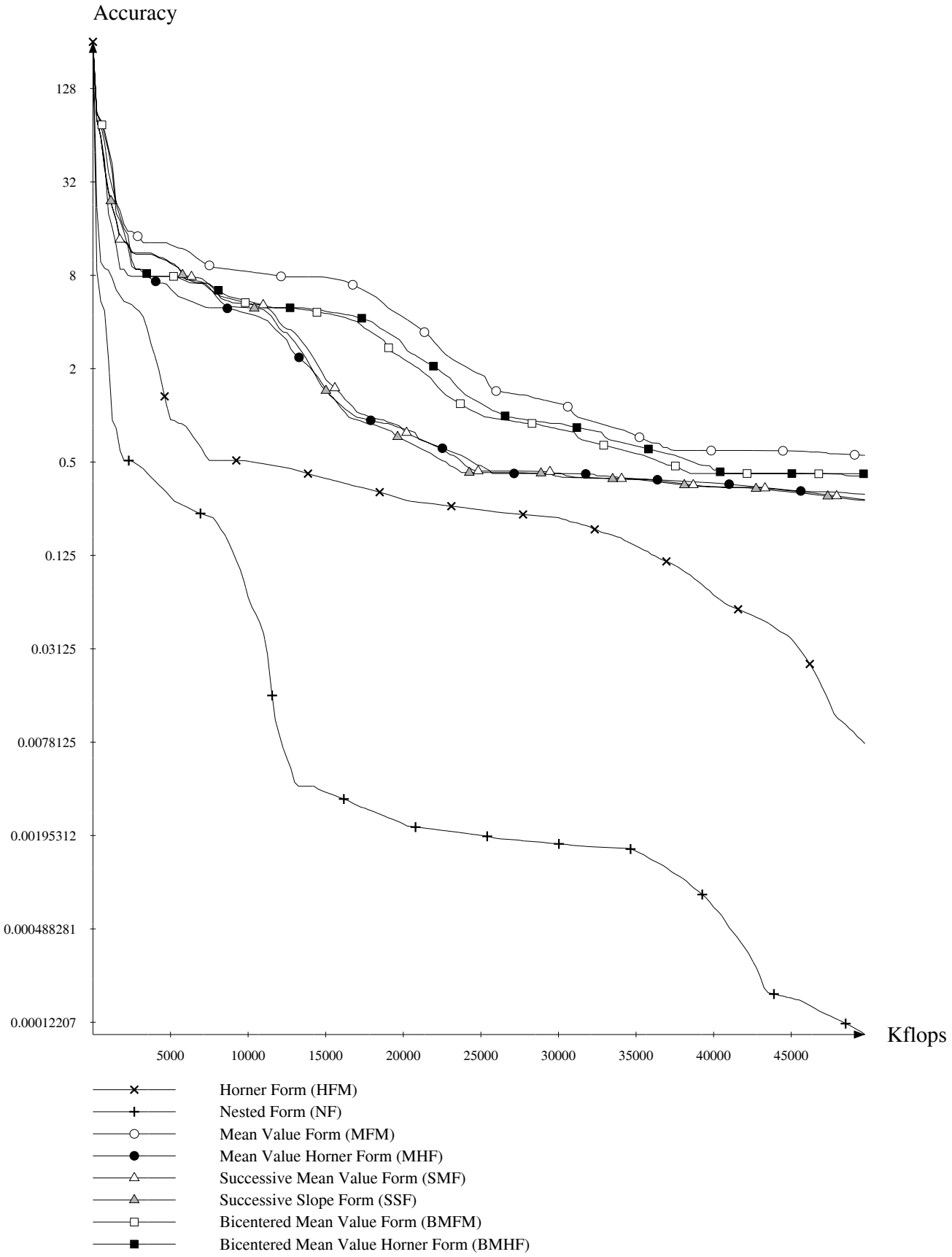


Figure 4.3.10: Robot kinematics problem, 8 variables. There are 16 solutions within the box.

Chapter 5

Isolating Boxes for Systems of Nonlinear Equations

In this chapter we study the following problem: Given a differentiable function $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^n$ and a box $\boldsymbol{B} \in \mathbb{IR}^n$, find disjoint sub–boxes $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(r)}$ of $\boldsymbol{B}$ such that

- every solution of $\boldsymbol{f}$ in $\boldsymbol{B}$ is contained in some $\boldsymbol{X}^{(i)}$,
- every $\boldsymbol{X}^{(i)}$ contains a unique solution of $\boldsymbol{f}$,
- starting from $\boldsymbol{X}^{(i)}$, an iterative method converges to the unique solution of $\boldsymbol{f}$ in $\boldsymbol{X}^{(i)}$ for every i.

The boxes $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(r)}$ are called isolating boxes for the solutions of $\boldsymbol{f}$ in $\boldsymbol{B}$.

Finding isolating boxes for the solution of systems of nonlinear equations is a very important problem in scientific computing, constraint logic programming, geometric modeling, engineering, etc. Various methods for this problem based on the Krawczyk operator [Krawczyk, 1969] and the Hansen–Sengupta operator [Hansen and Sengupta, 1981] have been described for example in [Moore, 1966] [Hansen, 1968], [Krawczyk, 1969], [Moore, 1977], [Moore and Jones, 1977], [Hansen, 1978a], [Hansen, 1978b], [Jones, 1978], [Moore, 1978], [Moore, 1979], [Jones, 1980], [Krawczyk, 1980a], [Moore, 1980a], [Moore and Kioustelidis, 1980], [Moore, 1980b], [Qi, 1980], [Wolfe, 1980], [Hansen and Sengupta, 1981], [Qi, 1981], [Moore and Qi, 1982], [Qi, 1982], [Rump, 1982], [Hansen and Greenberg, 1983], [Krawczyk, 1984], [Rump, 1984], [Krawczyk, 1985], [Shearer and Wolfe, 1985b], [Shearer and Wolfe, 1985a], [Krawczyk, 1986b], [Krawczyk, 1986c], [Krawczyk, 1986a], [Rump, 1988], [Zuhe, 1988], [Frommer and Mayer, 1989], [Kearfott, 1990a], [Kearfott, 1990b], [Neumaier, 1990], [Dimitrova, 1993], [Hong and Stahl, 1994b] and many others.

The algorithms considered in this chapter consist of three main conceptual steps:

Prune: Reduce the search space by eliminating regions which do not contain a solution.

Test: Decide whether a box contains a (unique) solution.

Bisect: Divide the search space and work on the sub–regions separately.

In Section 5.1, 5.2, 5.3 we give conditions for the existence, uniqueness and non–existence of solutions of $\boldsymbol{f}$ in a box $\boldsymbol{X}$ respectively. The key idea is to enclose $\boldsymbol{f}$ by a linear interval function i.e. a linear function with interval coefficients, and derive corresponding properties of the linearization. The non–existence condition in Section 5.3 is new. In Section 5.4 we introduce the notion of linear tightening. Linear tightening is an operator which serves for both pruning and testing. Unfortunately the conditions for the uniqueness and existence tests are too strong and are usually not satisfied. This problem is solved by preconditioning, which was introduced for systems of linear interval equations already in [Hansen, 1965]. In Section 5.5 and Section 5.6 we present two strategies for linear tightening, the linearized tightening operator, and

the Hansen–Sengupta operator [Hansen and Sengupta, 1981]. The Hansen–Sengupta operator applies preconditioning whereas the linearized tightening operator does not. Therefore the linearized tightening operator is cheaper but less powerful. The main properties of both operators are proved in an elementary geometric way.

Usually the operators are applied repeatedly until the (unique) existence test succeeds. In Section 5.7 we generalize the existence condition of the Hansen–Sengupta operator to iterations of the Hansen–Sengupta operator. A similar generalization of the uniqueness condition is not possible. In Section 5.8 we give conditions under which a sequence of boxes generated by iteration of the Hansen–Sengupta operator converges, some of them are new. Termination of a general algorithm for finding isolating boxes for the solutions of systems of nonlinear equations is studied in Section 5.9. A particular problem arises if a solution lies on the boundary of a box and we present a new method for resolving this difficulty by using results of Section 5.7 and 5.8. In Section 5.10 we review a method for finding all solutions of a system of polynomial equations when the search space is unbounded. Finally, in Section 5.11 we give an experimental comparison of two nonlinear equation system solvers, which indicates that combining the Hansen–Sengupta operator with the linearized tightening operator gives often a speed up.

5.1 Existence of Solutions in a Box

In this section we give a condition for the existence of solutions of a continuous function $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^n$ in a box $\boldsymbol{X} \in \mathbb{IR}^n$. Further, we introduce some notations which will be used in the following sections.

Definition 5.1.1 (Upper and Lower i-Face) *The upper and lower i-face $\boldsymbol{X}^{\overline{(i)}}, \boldsymbol{X}^{\underline{(i)}}$ of $\boldsymbol{X} \in \mathbb{IR}^n$ are defined as*

$$\begin{aligned} \boldsymbol{X}^{\overline{(i)}} &= (X_1, \ldots, X_{i-1}, \overline{X}_i, X_{i+1}, \ldots, X_n) \\ \boldsymbol{X}^{\underline{(i)}} &= (X_1, \ldots, X_{i-1}, \underline{X}_i, X_{i+1}, \ldots, X_n). \end{aligned}$$

Further,

$$\boldsymbol{X}^{\overline{(i)}\underline{(i)}} = \boldsymbol{X}^{\overline{(i)}} \cup \boldsymbol{X}^{\underline{(i)}}. \quad \square$$

The following Theorem is due to [Miranda, 1940].

Theorem 5.1.2 (Existence of Solutions) *If there exists a permutation*

$$\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$$

such that

$$f_i(\boldsymbol{a}) f_i(\boldsymbol{b}) \leq 0 \quad \text{for all} \quad \boldsymbol{a} \in \boldsymbol{X}^{\overline{(\pi(i))}}, \boldsymbol{b} \in \boldsymbol{X}^{\underline{(\pi(i))}}, i = 1, \ldots, n$$

then $\boldsymbol{f}$ has a solution in $\boldsymbol{X}$. $\square$

Theorem 5.1.2 can be directly applied by evaluating an interval extension of $\boldsymbol{f}$ on $\boldsymbol{X}^{\overline{(i)}}$ and on $\boldsymbol{X}^{\underline{(i)}}$ for all i, see for example [Moore and Kioustelidis, 1980]. In the following, we use a mean value form of $\boldsymbol{f}$. First, we give some basic definitions which will be used throughout the chapter.

Definition 5.1.3 (Linear Interval Function) *A function $G : \mathbb{R}^n \to \mathbb{IR}$ is called linear interval function if there exists $y \in \mathbb{R}$, $\boldsymbol{c} \in \mathbb{R}^n$ and $\boldsymbol{A} \in \mathbb{IR}^n$ such that*

$$G(\boldsymbol{x}) = y + \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{c}). \quad \square$$

Note that a linear interval function is not linear. For example, let $G(x) = [-1, 1]x$, then

$$G(1) + G(-1) = [-2, 2] \neq 0 = G(0).$$

However, a linear interval function can be considered as a set of (real) linear functions. Let $g : \mathbb{R}^n \to \mathbb{R}$ be a linear function. We write $g \in G$ if there exist $y \in \mathbb{R}$, $\boldsymbol{A} \in \mathbb{IR}^n$ and $\boldsymbol{a} \in \boldsymbol{A}$ such that

$$\begin{aligned} G(\boldsymbol{x}) &= y + \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{c}) \text{ and} \\ g(\boldsymbol{x}) &= y + \boldsymbol{a}(\boldsymbol{x} - \boldsymbol{c}) \end{aligned}$$

for all $\boldsymbol{x} \in \mathbb{R}^n$. Note that

$$G(\boldsymbol{x}) = \{g(\boldsymbol{x}) \mid g \in G\}.$$

In the following let G be a linear interval function and let $f : \mathbb{R}^n \to \mathbb{R}$.

Definition 5.1.4 (Interval Linearization) *G is called interval linearization of f in $\boldsymbol{X}$ if there exist $\boldsymbol{c} \in \mathbb{R}^n$ and $\boldsymbol{A} \in \mathbb{IR}^n$ such that*

$$G(\boldsymbol{x}) = f(\boldsymbol{c}) + \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{c})$$

and

$$f(\boldsymbol{x}) \in G(\boldsymbol{x}) \text{ for all } \boldsymbol{x} \in \boldsymbol{X}. \ \square$$

Note that in Definition 5.1.4 it is not required that $\boldsymbol{c} \in \boldsymbol{X}$. Hence, if G is an interval linearization of f in $\boldsymbol{X}$, then G is an interval linearization of f in $\boldsymbol{Y}$ for all $\boldsymbol{Y} \subseteq \boldsymbol{X}$.

Definition 5.1.5 (Variety of Interval Function) *The variety of G is defined as*

$$\mathcal{Z}(G) = \{\boldsymbol{x} \in \mathbb{R}^n \mid 0 \in G(\boldsymbol{x})\}. \ \square$$

Notation. For an interval $\boldsymbol{X} \in \mathbb{IR}^n$ we write $G(\boldsymbol{X})$ to denote the set

$$\{y \mid y \in G(\boldsymbol{x}) \text{ for some } \boldsymbol{x} \in \boldsymbol{X}\}.$$

Further, we write $\boldsymbol{X} \sim 0$ if $\boldsymbol{x} \sim 0$ for all $\boldsymbol{x} \in \boldsymbol{X}$, where $\sim \in \{>, <, \geq, \leq\}$. $\square$

Note that if $G(\boldsymbol{x}) = y + \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{c})$ then

$$G(\boldsymbol{X}) = y + \boldsymbol{A}(\boldsymbol{X} - \boldsymbol{c}).$$

Definition 5.1.6 (Face Enclosed Linear Interval Function)

- *G is weakly $\boldsymbol{X}^{\underline{(i)}}_{\overline{(i)}}$ enclosed if*

$$\left(G(\boldsymbol{X}^{\underline{(i)}}) \geq 0 \text{ and } G(\boldsymbol{X}^{\overline{(i)}}) \leq 0\right) \text{ or } \left(G(\boldsymbol{X}^{\underline{(i)}}) \leq 0 \text{ and } G(\boldsymbol{X}^{\overline{(i)}}) \geq 0\right).$$

- *G is strongly $\boldsymbol{X}^{\underline{(i)}}_{\overline{(i)}}$ enclosed if*

$$\left(G(\boldsymbol{X}^{\underline{(i)}}) > 0 \text{ and } G(\boldsymbol{X}^{\overline{(i)}}) < 0\right) \text{ or } \left(G(\boldsymbol{X}^{\underline{(i)}}) < 0 \text{ and } G(\boldsymbol{X}^{\overline{(i)}}) > 0\right). \ \square$$

Due to the natural embedding of $\mathbb{R}$ into $\mathbb{IR}$, a (real) linear function is a special case of linear interval function. Therefore, all definitions and properties of linear interval functions carry over to linear functions. In particular, if G is $\boldsymbol{X}^{\underline{(i)}}_{\overline{(i)}}$ enclosed then g is $\boldsymbol{X}^{\underline{(i)}}_{\overline{(i)}}$ enclosed for all $g \in G$.

Figure 5.1.1 illustrates Definition 5.1.6 of face enclosedness. An easy way to understand this notion is to think of a sandwich, where the bread is $\boldsymbol{X}^{\underline{(i)}}_{\overline{(i)}}$ and the ham is $\mathcal{Z}(G)$. If G is $\boldsymbol{X}^{\underline{(i)}}_{\overline{(i)}}$ enclosed, then the variety of G in $\boldsymbol{X}$ is enclosed between $\boldsymbol{X}^{\overline{(i)}}$ and $\boldsymbol{X}^{\underline{(i)}}$. In the upper two pictures, the variety (shaded area) of the linear interval function

$$G(x_1, x_2) = 0.4 + [0.4, 0.5]x_1 + [0.3, 0.4]x_2$$

is displayed. In the left graph G is strongly $\boldsymbol{X}^{\overline{(1)}}$ enclosed where $\boldsymbol{X} = ([-2.3, 1.0], [-1.5, 1.0])^{\mathrm{T}}$, in the right graph G is weakly (but not strongly) $\boldsymbol{X}^{\overline{(2)}}$ enclosed where $\boldsymbol{X} = ([-2.0, 0.5], [-2.8, 2.0])^{\mathrm{T}}$. In the lower left graph the variety of

$$G(x_1, x_2) = 0.2 + [0.2, 0.35]x_1 + [-0.15, 0.2]x_2$$

and the box $\boldsymbol{X} = ([-2.5, 1.5], [-1.5, 1])^{\mathrm{T}}$ is displayed. As one sees, G is strongly $\boldsymbol{X}^{\overline{(1)}}$ enclosed and there exists no $\boldsymbol{X}$ such that G would be $\boldsymbol{X}^{\overline{(2)}}$ enclosed. In the lower right graph, the variety of

$$G(x_1, x_2) = 0.3 + [-0.15, 0.3]x_1 + [-0.25, 0.15]x_2$$

is displayed. Apparently there is no $\boldsymbol{X}$ and i such that G would be $\boldsymbol{X}^{\overline{(i)}}$ enclosed. The following lemma gives an equivalent condition for strong $\boldsymbol{X}^{\overline{(i)}}$ enclosedness.

Lemma 5.1.7 (Strong Face Enclosedness) *G is strongly $\boldsymbol{X}^{\overline{(i)}}$ enclosed iff $0 \notin G(\boldsymbol{X}^{\overline{(i)}})$, $0 \notin G(\boldsymbol{X}^{\underline{(i)}})$ and $0 \in G(\boldsymbol{X})$.* □

Proof.

"⇒" Assume G is strongly $\boldsymbol{X}^{\overline{(i)}}$ enclosed. By Definition 5.1.6, $0 \notin G(\boldsymbol{X}^{\overline{(i)}})$, $0 \notin G(\boldsymbol{X}^{\underline{(i)}})$. It remains to show that $0 \in G(\boldsymbol{X})$. Let $g \in G$ and $x_j \in X_j$ for all $j \neq i$ arbitrary but fixed. Note that

$$g(x_1, \ldots, \overline{X}_i, \ldots, x_n)g(x_1, \ldots, \underline{X}_i, \ldots, x_n) < 0.$$

As g is continuous, there exists $x_i \in X_i$ such that $g(\boldsymbol{x}) = 0$. Hence, $0 \in G(\boldsymbol{x})$ and therefore $0 \in G(\boldsymbol{X})$.

"⇐" Assume $0 \notin G(\boldsymbol{X}^{\overline{(i)}})$, $0 \notin G(\boldsymbol{X}^{\underline{(i)}})$ and $0 \in G(\boldsymbol{X})$. From the continuity of G it follows that all elements of $G(\boldsymbol{X}^{\overline{(i)}})$ have the same sign and all elements of $G(\boldsymbol{X}^{\underline{(i)}})$ have the same sign. It remains to show that the sign of $G(\boldsymbol{X}^{\overline{(i)}})$ and $G(\boldsymbol{X}^{\underline{(i)}})$ is different. As $0 \in G(\boldsymbol{X})$ there exists $g \in G$ and $\boldsymbol{x} \in \boldsymbol{X}$ such that $g(\boldsymbol{x}) = 0$. From the linearity of g it follows that

$$g(x_1, \ldots, \overline{X}_i, \ldots, x_n)g(x_1, \ldots, \underline{X}_i, \ldots, x_n) < 0.$$

Hence, the sign of $G(\boldsymbol{X}^{\overline{(i)}})$ and $G(\boldsymbol{X}^{\underline{(i)}})$ is different and G is $\boldsymbol{X}^{\overline{(i)}}$ enclosed. □

We generalize the definitions of a linear interval function to tuples.

Definition 5.1.8 (Linear Interval Function) *A function $\boldsymbol{G} : \mathbb{R}^n \to \mathbb{IR}^n$ is called linear interval function if G_i is a linear interval function for $i = 1, \ldots, n$, i.e. if there exist*

$$\begin{aligned} y_1, \ldots, y_n &\in \mathbb{R}, \\ \boldsymbol{c}_1, \ldots, \boldsymbol{c}_n &\in \mathbb{R}^n \\ \boldsymbol{A}_1, \ldots, \boldsymbol{A}_n &\in \mathbb{IR}^{n \times n} \end{aligned}$$

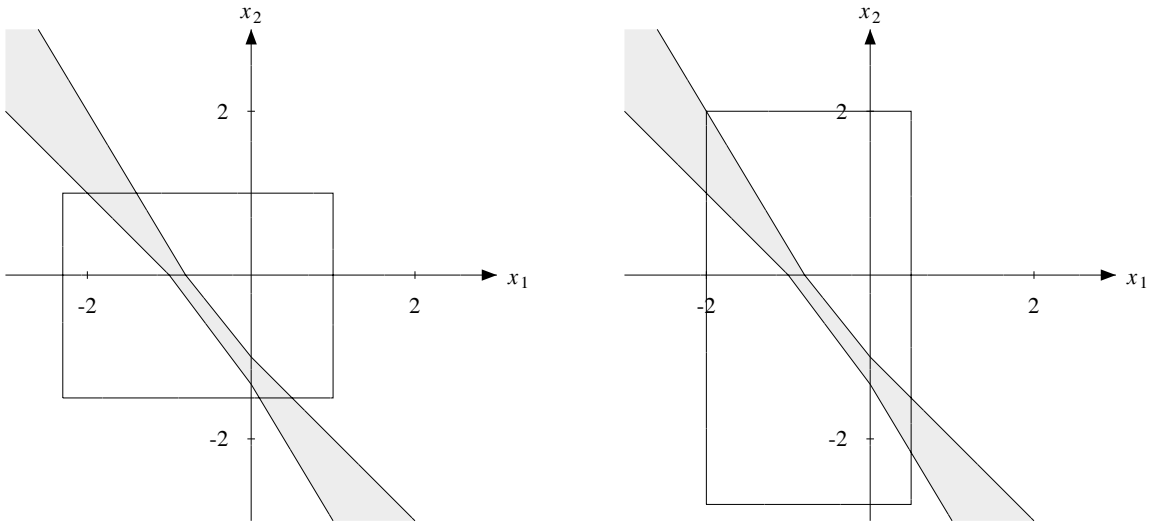
such that

$$G_i(\boldsymbol{x}) = y_i + \boldsymbol{A}_i(\boldsymbol{x} - \boldsymbol{c}_i) \text{ for } i = 1, \ldots, n.$$ □

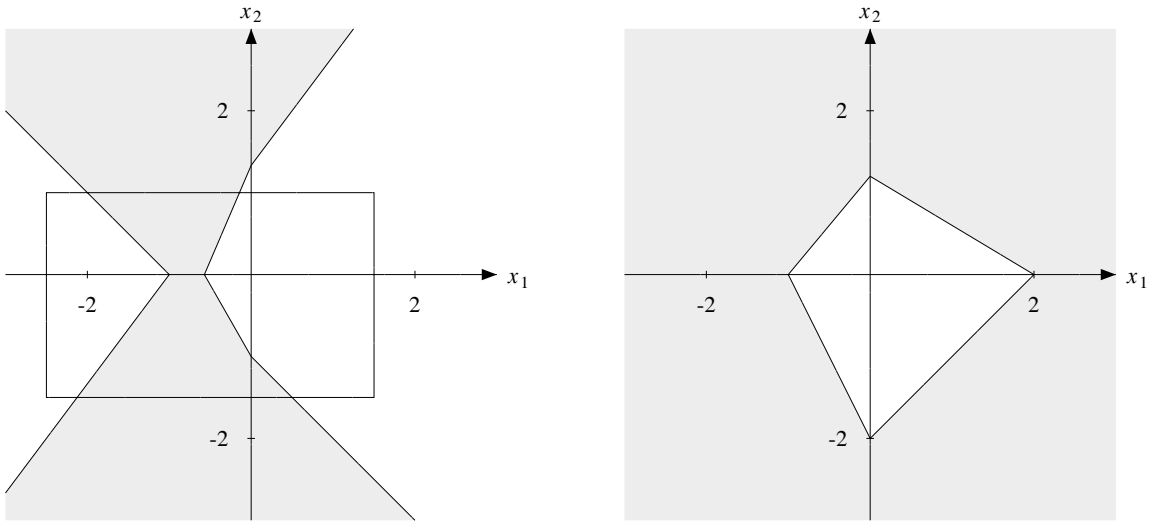
In the following let $\boldsymbol{G}$ be a linear interval function. We call the matrix

$$\mathfrak{A} = (\boldsymbol{A}_1, \ldots, \boldsymbol{A}_n)^{\mathrm{T}}$$

coefficient matrix of $\boldsymbol{G}$.



$G(x_1, x_2) = 0.4 + [0.4, 0.5]x_1 + [0.3, 0.4]x_2$



$G(x_1, x_2) = 0.2 + [0.2, 0.35]x_1 + [-0.15, 0.2]x_2$

$G(x_1, x_2) = 0.3 + [-0.15, 0.3]x_1 + [-0.25, 0.15]x_2$

Figure 5.1.1: Illustration of Definition 5.1.6.

Definition 5.1.9 (Interval Linearization) *$\boldsymbol{G}$ is called interval linearization of $\boldsymbol{f}$ in $\boldsymbol{X}$ if there exist*

$$\begin{aligned} \boldsymbol{c}_1, \ldots, \boldsymbol{c}_n &\in \mathbb{R}^n, \\ \boldsymbol{A}_1, \ldots, \boldsymbol{A}_n &\in \mathbb{IR}^{n \times n} \end{aligned}$$

such that for all i

$$G_i(\boldsymbol{x}) = f_i(\boldsymbol{c}_i) + \boldsymbol{A}_i(\boldsymbol{x} - \boldsymbol{c}_i)$$

and

$$f_i(x) \in G_i(x) \text{ for all } \boldsymbol{x} \in \boldsymbol{X}. \square$$

Definition 5.1.10 (Variety of Interval Function) *The variety of $\boldsymbol{G}$ is defined as*

$$\mathcal{Z}(\boldsymbol{G}) = \{\boldsymbol{x} \in \mathbb{R}^n \mid 0 \in G_i(\boldsymbol{x}) \text{ for all } i = 1, \ldots, n\}. \square$$

Definition 5.1.11 ($\boldsymbol{G}$ Orthogonal in $\boldsymbol{X}$) *$\boldsymbol{G}$ is weakly respectively strongly orthogonal in $\boldsymbol{X}$ if there exists a permutation*

$$\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$$

such that G_i is weakly respectively strongly $\boldsymbol{X}^{\overline{(\pi(i))}}$ enclosed for all i. $\square$

Now, Theorem 5.1.2 can be specialized as follows:

Theorem 5.1.12 (Existence of Solutions) *If there exists an interval linearization $\boldsymbol{G}$ of $\boldsymbol{f}$ in $\boldsymbol{X}$ which is weakly orthogonal in $\boldsymbol{X}$, then $\boldsymbol{f}$ has a solution in $\boldsymbol{X}$.* $\square$

Proof. Let $\boldsymbol{G}$ be an interval linearization of $\boldsymbol{f}$ in $\boldsymbol{X}$ which is weakly orthogonal in $\boldsymbol{X}$. Let $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$ be a permutation such that G_i is $\boldsymbol{X}^{\overline{(\pi(i))}}$ enclosed for all i. As $f_i(\boldsymbol{x}) \in G_i(\boldsymbol{x})$ for all $\boldsymbol{x} \in \boldsymbol{X}$, it follows that

$$f_i(\boldsymbol{a}) f_i(\boldsymbol{b}) \leq 0 \text{ for all } \boldsymbol{a} \in \boldsymbol{X}^{\overline{(\pi(i))}}, \boldsymbol{b} \in \boldsymbol{X}^{\underline{(\pi(i))}}, i = 1, \ldots, n$$

and $\boldsymbol{f}$ has a solution in $\boldsymbol{X}$ by Theorem 5.1.2. $\square$

5.2 Uniqueness of Solutions in a Box

In this section we give a condition for the uniqueness of solutions of a differentiable function $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^n$ in a box $\boldsymbol{X} \in \mathbb{IR}^n$. The main result is that if there exists a strongly orthogonal linearization $\boldsymbol{G}$ of $\boldsymbol{f}$ in $\boldsymbol{X}$ where the coefficient matrix is a Jacobian, then $\boldsymbol{f}$ has at most one solution in $\boldsymbol{X}$.

Definition 5.2.1 (Jacobian Matrix) *$\mathfrak{A} \in \mathbb{IR}^{n \times n}$ is a Jacobian of $\boldsymbol{f}$ in $\boldsymbol{X}$ if*

$$\boldsymbol{f}'(\boldsymbol{x}) \in \mathfrak{A} \text{ for all } \boldsymbol{x} \in \boldsymbol{X}. \square$$

Definition 5.2.2 (Regular Interval Matrix) *An interval matrix $\mathfrak{A}$ is regular, if every $\mathfrak{a} \in \mathfrak{A}$ is regular.* $\square$

Theorem 5.2.3 (Regular Jacobian Implies Uniqueness) *If there exists a regular Jacobian $\mathfrak{A}$ of $\boldsymbol{f}$ in $\boldsymbol{X}$, then $\boldsymbol{f}$ has at most one solution in $\boldsymbol{X}$.* $\square$

Proof. Assume $\mathfrak{A}$ is a regular Jacobian of $\boldsymbol{f}$ in $\boldsymbol{X}$ and assume $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{y}) = 0$ for some $\boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{X}$. According to the mean value theorem

$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{y}) + \mathfrak{a}(\boldsymbol{x} - \boldsymbol{y})$$

for some $\mathfrak{a} \in \mathfrak{A}$. Hence $\mathfrak{a}(\boldsymbol{x} - \boldsymbol{y}) = 0$ and from the regularity of $\mathfrak{a}$ it follows that $\boldsymbol{x} = \boldsymbol{y}$. $\square$

Theorem 5.2.4 (Linearization by Jacobian) *Let $\mathfrak{A}$ be a Jacobian of $\boldsymbol{f}$ in $\boldsymbol{X}$ and let $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_n \in \boldsymbol{X}$. Let*

$$G_i(\boldsymbol{x}) = f_i(\boldsymbol{c}_i) + \boldsymbol{A}_i(\boldsymbol{x} - \boldsymbol{c}_i) \text{ for } i = 1, \ldots, n.$$

Then $\boldsymbol{G}$ is a linearization of $\boldsymbol{f}$ in $\boldsymbol{X}$. □

Proof. Let $\boldsymbol{c}_i, \boldsymbol{A}_i, G_i$ as in Theorem 5.2.4. Let $\boldsymbol{x} \in \boldsymbol{X}$ and $i \in \{1, \ldots, n\}$ arbitrary but fixed. We have to show that $f_i(\boldsymbol{x}) \in G_i(\boldsymbol{x})$. According to the mean value theorem there exists $\boldsymbol{a}_i \in \boldsymbol{A}_i$ such that

$$f_i(\boldsymbol{x}) = f_i(\boldsymbol{c}_i) + \boldsymbol{a}_i(\boldsymbol{x} - \boldsymbol{c}_i).$$

Hence

$$f_i(\boldsymbol{x}) \in f_i(\boldsymbol{c}_i) + \boldsymbol{A}_i(\boldsymbol{x} - \boldsymbol{c}_i) = G_i(\boldsymbol{x}). \; \square$$

In the remainder of the section we derive a criterion for the regularity of an interval matrix. Let

$$G(\boldsymbol{x}) = y + \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{c})$$

be a linearization of f in $\boldsymbol{X}$.

Lemma 5.2.5 *If G is weakly $\boldsymbol{X}^{\overline{(i)}}$ enclosed then for all $g \in G$ and for all*

$$(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) \in (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$$

there exists $x_i \in X_i$ such that $g(x_1, \ldots, x_i, \ldots, x_n) = 0$. □

Proof. Lemma 5.2.5 can easily be verified at Figure 5.1.1. Assume G is weakly $\boldsymbol{X}^{\overline{(i)}}$ enclosed. Let $g \in G$ and $\hat{x}_j \in X_j$ for all $j \neq i$ arbitrary but fixed. Let $h : \mathbb{R} \to \mathbb{R}$ be defined as

$$h(x_i) = g(\hat{x}_1, \ldots, x_i, \ldots, \hat{x}_n).$$

As g is weakly $\boldsymbol{X}^{\overline{(i)}}$ enclosed, it holds that

$$\operatorname{sign}(h(\overline{X}_i)) \operatorname{sign}(h(\underline{X}_i)) \leq 0.$$

From the continuity of h it follows that there exists $\hat{x}_i \in X_i$ such that $h(\hat{x}_i) = 0$ and hence $g(\hat{\boldsymbol{x}}) = 0$. □

Corollary 5.2.6 (Enclosed Function Intersects all other Faces) *If G is weakly $\boldsymbol{X}^{\overline{(i)}}$ enclosed, then for all $g \in G$ and for all $j \neq i$*

$$0 \in g(\boldsymbol{X}^{\overline{(j)}}), \quad 0 \in g(\boldsymbol{X}^{\underline{(j)}})$$

and thus

$$0 \in G(\boldsymbol{X}^{\overline{(j)}}), \quad 0 \in G(\boldsymbol{X}^{\underline{(j)}}). \; \square$$

The following lemma states an important property of the coefficient matrix of an enclosed function.

Lemma 5.2.7 *If G is strongly $\boldsymbol{X}^{\overline{(i)}}$ enclosed then*

$$\operatorname{mig}(A_i)\operatorname{w}(X_i) > \sum_{j \neq i} \operatorname{mag}(A_j)\operatorname{w}(X_j). \; \square$$

Proof. Assume G is strongly $\boldsymbol{X}^{\overline{(i)}}$ enclosed and let $\boldsymbol{a} \in \boldsymbol{A}$ arbitrary but fixed. we have to show that

$$|a_i|\operatorname{w}(X_i) > \sum_{j \neq i} |a_j|\operatorname{w}(X_j).$$

We consider the case $a_i \geq 0$, the case $a_i \leq 0$ is similar. According to Lemma 5.2.5 there exists $\hat{x}_i \in X_i$ such that

$$\sum_{\substack{a_j \geq 0 \\ j \neq i}} a_j \underline{X}_j + \sum_{\substack{a_j < 0 \\ j \neq i}} a_j \overline{X}_j + a_i \hat{x}_i + y = 0. \tag{5.2.1}$$

Assume

$$a_i \mathrm{w}(X_i) \leq \sum_{j \neq i} |a_j| \mathrm{w}(X_j),$$

i.e.

$$a_i \mathrm{w}(X_i) \leq \sum_{\substack{a_j \geq 0 \\ j \neq i}} a_j (\overline{X}_j - \underline{X}_j) + \sum_{\substack{a_j < 0 \\ j \neq i}} a_j (\underline{X}_j - \overline{X}_j).$$

Then there exist $\hat{x}_j \in X_j$, $j \neq i$ such that

$$a_i (\hat{x}_i - \underline{X}_i) = \sum_{\substack{a_j \geq 0 \\ j \neq i}} a_j (\hat{x}_j - \underline{X}_j) + \sum_{\substack{a_j < 0 \\ j \neq i}} a_j (\hat{x}_j - \overline{X}_j). \tag{5.2.2}$$

Adding (5.2.1) and (5.2.2) we obtain

$$\sum_{j \neq i} a_j \hat{x}_j + a_i \underline{X}_i + y = 0$$

i.e. $0 \in g(\boldsymbol{X}^{\underline{(i)}})$, which contradicts the assumption that g is strongly $\boldsymbol{X}^{\overline{(i)}}$ enclosed. □

From Lemma 5.2.7 it follows immediately that if $0 \in A_i$ then there is no $\boldsymbol{X} \in \mathbb{IR}^n$ such that G is $\boldsymbol{X}^{\overline{(i)}}$ enclosed. This explains the observations in the second row of Figure 5.1.1. In the following let

$$G_i(\boldsymbol{x}) = f_i(\boldsymbol{c}_i) + \boldsymbol{A}_i(\boldsymbol{x} - \boldsymbol{c}_i), \quad i = 1, \ldots, n$$

be a linearization of f_i in $\boldsymbol{X}$ and let $\mathfrak{A} = (\boldsymbol{A}_1, \ldots, \boldsymbol{A}_n)^{\mathrm{T}} \in \mathbb{IR}^{n \times n}$.

Theorem 5.2.8 (Strong Orthogonality Implies Regularity) *If $\boldsymbol{G}$ is strongly orthogonal in $\boldsymbol{X}$, then $\mathfrak{A}$ is regular.* □

Proof. Assume $\boldsymbol{G}$ is strongly orthogonal in $\boldsymbol{X}$. For simplicity we reorder the G_i such that G_i is strongly $\boldsymbol{X}^{\overline{(i)}}$ enclosed for all i. We have to show that every $\mathfrak{a} \in \mathfrak{A}$ is regular. Hence, let $\mathfrak{a} \in \mathfrak{A}$ arbitrary but fixed. According to Lemma 5.2.7 it holds that

$$|a_{ii}| \mathrm{w}(X_i) > \sum_{j \neq i} |a_{ij}| \mathrm{w}(X_j) \quad \text{for all} \quad i.$$

Hence, the matrix

$$\begin{pmatrix} a_{1,1}\mathrm{w}(X_1) & a_{1,2}\mathrm{w}(X_2) & \ldots & a_{1,n}\mathrm{w}(X_n) \\ a_{2,1}\mathrm{w}(X_1) & a_{1,2}\mathrm{w}(X_2) & \ldots & a_{2,n}\mathrm{w}(X_n) \\ \vdots & \vdots & & \vdots \\ a_{n,1}\mathrm{w}(X_1) & a_{n,2}\mathrm{w}(X_2) & \ldots & a_{n,n}\mathrm{w}(X_n) \end{pmatrix}$$

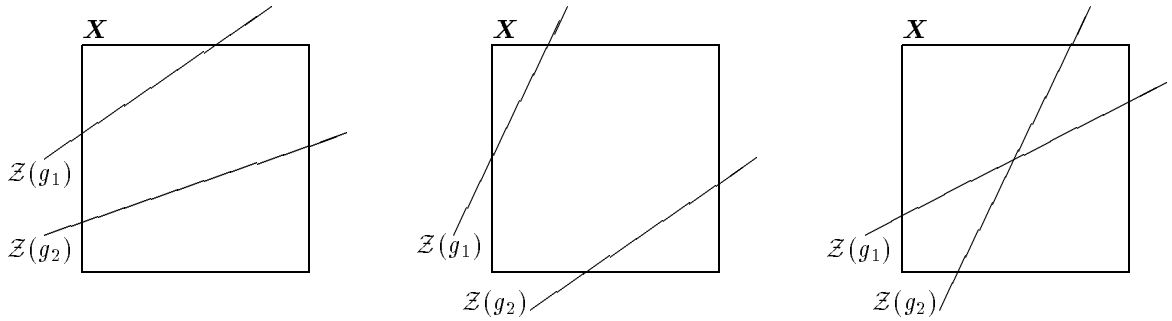
is strictly diagonally dominant and hence regular. Thus,

$$\begin{pmatrix} a_{1,1} & a_{1,2} & \ldots & a_{1,n} \\ a_{2,1} & a_{1,2} & \ldots & a_{2,n} \\ \vdots & \vdots & & \vdots \\ a_{n,1} & a_{n,2} & \ldots & a_{n,n} \end{pmatrix}$$

is also regular. □

Corollary 5.2.9 (Orthogonality Implies Uniqueness) *If $\mathfrak{A}$ is a Jacobian of $\boldsymbol{f}$ in $\boldsymbol{X}$ and $\boldsymbol{G}$ is strongly orthogonal in $\boldsymbol{X}$, then $\boldsymbol{f}$ has a unique solution in $\boldsymbol{X}$.* □

Proof. Follows from Theorem 5.1.12, Theorem 5.2.3 and Theorem 5.2.8. □



$\boldsymbol{g}$ not face disjoint from $\boldsymbol{X}$, $\boldsymbol{g}$ not orthogonal in $\boldsymbol{X}$

$\boldsymbol{g}$ face disjoint from $\boldsymbol{X}$, $\boldsymbol{g}$ not orthogonal in $\boldsymbol{X}$

$\boldsymbol{g}$ face disjoint from $\boldsymbol{X}$, $\boldsymbol{g}$ orthogonal in $\boldsymbol{X}$

Figure 5.3.1: Illustration of face disjointness and orthogonality.

5.3 Non–Existence of Solutions in a Box

In this section we give a new condition for the non–existence of solutions of a continuous function $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^n$ in a box $\boldsymbol{X} \in \mathbb{IR}^n$.

Definition 5.3.1 ($\boldsymbol{G}$ Face Disjoint from $\boldsymbol{X}$) *A linear interval function $\boldsymbol{G} : \mathbb{R}^n \to \mathbb{IR}^n$ is face disjoint from $\boldsymbol{X}$ if for every face $\boldsymbol{T}$ of $\boldsymbol{X}$ there exists $i \in \{1, \ldots, n\}$ such that*

$$\mathcal{Z}(G_i) \cap \boldsymbol{T} = \emptyset. \ \square$$

Obviously, if $\boldsymbol{G}$ is strongly orthogonal in $\boldsymbol{X}$ then $\boldsymbol{G}$ is face disjoint from $\boldsymbol{X}$. Now, the main non–existence theorem of this section is as follows:

Theorem 5.3.2 (Non–Existence of Solutions in a Box) *Let $\boldsymbol{G}$ be a linearization of $\boldsymbol{f}$ in $\boldsymbol{X}$. If $\boldsymbol{G}$ is face disjoint from $\boldsymbol{X}$ but not strongly orthogonal in $\boldsymbol{X}$ then $\boldsymbol{f}$ has no solution in $\boldsymbol{X}$.* $\square$

Theorem 5.3.2 is new. Before going into details of the proof, we give an illustration of the notions face disjointness and orthogonality in Figure 5.3.1. For simplicity, we consider only real linear functions $\boldsymbol{g}$ there. Figure 5.3.1 motivates the following Theorem, where $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^n$ is a linear function:

Theorem 5.3.3 (Non–Existence of Solutions in a Box) *Assume $\boldsymbol{g}$ is face disjoint from $\boldsymbol{X}$ but $\boldsymbol{g}$ is not strongly orthogonal in $\boldsymbol{X}$. Then $\boldsymbol{g}(\boldsymbol{x}) \neq 0$ for all $\boldsymbol{x} \in \boldsymbol{X}$.* $\square$

For the proof of Theorem 5.3.3 we need some preparation. First, we show that if some g_i is not strongly $\boldsymbol{X}^{\overline{(j)}}$ enclosed for any j, then already one half of $\boldsymbol{X}$ can not contain a solution of $\boldsymbol{g}$. Next, we show that if $\boldsymbol{g}$ is face disjoint from $\boldsymbol{X}$ but not strongly orthogonal in $\boldsymbol{X}$, then the halves corresponding to such $g_i's$ together cover $\boldsymbol{X}$ entirely.

Definition 5.3.4 (Half Box) *Let $\boldsymbol{U}, \boldsymbol{V}$ be non–opposite faces of $\boldsymbol{X}$. The set $\mathcal{H}(\boldsymbol{U}, \boldsymbol{V})$, is defined as*

$$\mathcal{H}(\boldsymbol{U}, \boldsymbol{V}) = \{\alpha \boldsymbol{u} + (1-\alpha)\boldsymbol{v} \mid 0 \leq \alpha \leq 1, \boldsymbol{u} \in \boldsymbol{U}, \boldsymbol{v} \in \boldsymbol{V}\}. \ \square$$

From the convexity of $\boldsymbol{X}$ it follows that $\mathcal{H}(\boldsymbol{U}, \boldsymbol{V}) \subseteq \boldsymbol{X}$. Geometrically, $\mathcal{H}(\boldsymbol{U}, \boldsymbol{V})$ is the half of $\boldsymbol{X}$ spanned by $\boldsymbol{U}$ and $\boldsymbol{V}$, see Figure 5.3.2. In the following let $g(\boldsymbol{x}) = y + \boldsymbol{a}\boldsymbol{x}$ be a linear function.

Lemma 5.3.5 (No Solution in Half Box) *Let $\boldsymbol{U}, \boldsymbol{V}$ be non–opposite faces of $\boldsymbol{X}$. If*

$$\mathcal{Z}(g) \cap \boldsymbol{U} = \emptyset, \quad \mathcal{Z}(g) \cap \boldsymbol{V} = \emptyset$$

then

$$\mathcal{Z}(g) \cap \mathcal{H}(\boldsymbol{U}, \boldsymbol{V}) = \emptyset. \ \square$$

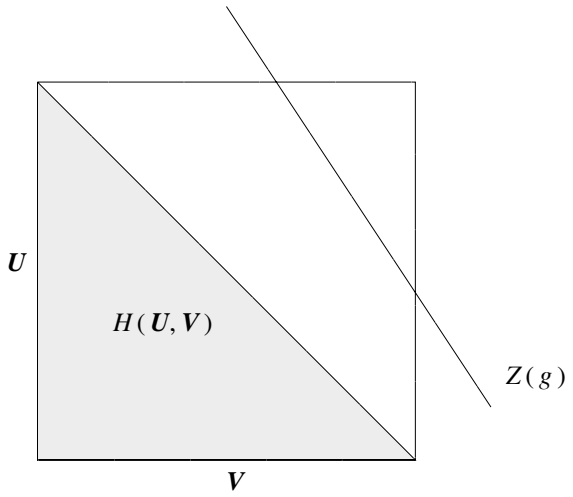


Figure 5.3.2: Illustration of Lemma 5.3.5

Proof. Let $\boldsymbol{U}, \boldsymbol{V}$ be non–opposite faces of $\boldsymbol{X}$ and assume $\mathcal{Z}(g) \cap \boldsymbol{U} = \emptyset$, $\mathcal{Z}(g) \cap \boldsymbol{V} = \emptyset$. As $\boldsymbol{U}, \boldsymbol{V}$ are not disjoint and $g(\boldsymbol{x}) \neq 0$ for all $\boldsymbol{x} \in \boldsymbol{U} \cup \boldsymbol{V}$, it follows that g has the same sign on $\boldsymbol{U}$ and $\boldsymbol{V}$. Let $\boldsymbol{x} \in \mathcal{H}(\boldsymbol{U}, \boldsymbol{V})$ arbitrary but fixed. By Definition 5.3.4 there exists $0 \leq \alpha \leq 1$ and $\boldsymbol{u} \in \boldsymbol{U}$, $\boldsymbol{v} \in \boldsymbol{V}$ such that

$$\boldsymbol{x} = \alpha \boldsymbol{u} + (1-\alpha)\boldsymbol{v}.$$

Thus,

$$\begin{aligned} g(\alpha \boldsymbol{u} + (1-\alpha)\boldsymbol{v}) &= y + \sum a_i \big(\alpha u_i + (1-\alpha) v_i\big) \\ &= \alpha \left(y + \sum a_i u_i\right) + (1-\alpha)\left(y + \sum a_i v_i\right) \\ &= \alpha g(\boldsymbol{u}) + (1-\alpha) g(\boldsymbol{v}) \\ &\neq 0. \ \square \end{aligned}$$

An illustration of Lemma 5.3.5 is given in Figure 5.3.2

Lemma 5.3.6 (Element Test for Half Box) *Let $\boldsymbol{U}, \boldsymbol{V}$ be an i- respectively j-face of $\boldsymbol{X}$, $i \neq j$ and assume $\overline{X}_i \neq \underline{X}_i$, $\overline{X}_j \neq \underline{X}_j$. Let $\boldsymbol{x} \in \boldsymbol{X}$ and let*

$$\begin{aligned} e_i &= \begin{cases} (x_i - \underline{X}_i)/(\overline{X}_i - \underline{X}_i) & \text{if} \quad \boldsymbol{U} = \boldsymbol{X}^{(i)}_{\underline{\ }} \\ (\overline{X}_i - x_i)/(\overline{X}_i - \underline{X}_i) & \text{if} \quad \boldsymbol{U} = \boldsymbol{X}^{\overline{(i)}} \end{cases} \\ e_j &= \begin{cases} (x_j - \underline{X}_j)/(\overline{X}_j - \underline{X}_j) & \text{if} \quad \boldsymbol{V} = \boldsymbol{X}^{(j)}_{\underline{\ }} \\ (\overline{X}_j - x_j)/(\overline{X}_j - \underline{X}_j) & \text{if} \quad \boldsymbol{V} = \boldsymbol{X}^{\overline{(j)}} \end{cases} \end{aligned}$$

Then $\boldsymbol{x} \in \mathcal{H}(\boldsymbol{U}, \boldsymbol{V})$ if and only if $e_i + e_j \leq 1$. $\square$

Proof. Let $i \neq j$ arbitrary but fixed and assume $\overline{X}_i \neq \underline{X}_i$, $\overline{X}_j \neq \underline{X}_j$. We give a proof for the case

$$\boldsymbol{U} = \boldsymbol{X}^{(i)}_{\underline{\ }}, \quad \boldsymbol{V} = \boldsymbol{X}^{\overline{(j)}}.$$

The other cases can be treated analogously. Assume $\boldsymbol{x} \in \boldsymbol{X}$. According to Definition 5.3.4

$$\boldsymbol{x} \in \mathcal{H}(\boldsymbol{X}^{(i)}_{\underline{\ }}, \boldsymbol{X}^{\overline{(j)}})$$

if and only if there exist $\boldsymbol{u} \in \boldsymbol{X}^{(i)}_{\underline{\ }}, \boldsymbol{v} \in \boldsymbol{X}^{\overline{(j)}}$ and $0 \leq \alpha \leq 1$ such that $\boldsymbol{x} = \alpha \boldsymbol{u} + (1-\alpha)\boldsymbol{v}$, or equivalently

$$\begin{aligned} x_i &= \alpha \underline{X}_i + (1-\alpha) v_i \\ x_j &= \alpha u_j + (1-\alpha)\overline{X}_j. \end{aligned}$$

As $v_i \in X_i$, $u_j \in X_j$ this is equivalent to

$$\begin{aligned} x_i &\leq \alpha\underline{X_i} + (1-\alpha)\overline{X_i} \\ x_j &\geq \alpha\underline{X_j} + (1-\alpha)\overline{X_j}\,. \end{aligned}$$

By some formula manipulation we obtain

$$\begin{aligned} (x_i - \overline{X_i})/(\underline{X_i} - \overline{X_i}) &\geq \alpha \\ (x_j - \overline{X_j})/(\underline{X_j} - \overline{X_j}) &\leq \alpha, \end{aligned}$$

$$\begin{aligned} (x_i - \underline{X_i})/(\overline{X_i} - \underline{X_i}) &\leq 1-\alpha \\ (\overline{X_j} - x_j)/(\overline{X_j} - \underline{X_j}) &\leq \alpha. \end{aligned}$$

and thus

$$e_i + e_j \leq 1. \ \square$$

The following lemma is trivial but will be useful later.

Lemma 5.3.7 (Condition for $\boldsymbol{X}^{\overline{(i)}} \neq \boldsymbol{X}^{\underline{(i)}}$) *If $\mathcal{Z}(g) \cap \boldsymbol{X} \neq \emptyset$ and either $\mathcal{Z}(g) \cap \boldsymbol{X}^{\overline{(i)}} = \emptyset$ or $\mathcal{Z}(g) \cap \boldsymbol{X}^{\underline{(i)}} = \emptyset$, then $\boldsymbol{X}^{\overline{(i)}} \neq \boldsymbol{X}^{\underline{(i)}}$.*

Proof. Assume $\mathcal{Z}(g) \cap \boldsymbol{X} \neq \emptyset$ and $\boldsymbol{X}^{\overline{(i)}} = \boldsymbol{X}^{\underline{(i)}}$. Then $\boldsymbol{X} = \boldsymbol{X}^{\overline{(i)}} = \boldsymbol{X}^{\underline{(i)}}$ and therefore $\mathcal{Z}(g) \cap \boldsymbol{X}^{\overline{(i)}} \neq \emptyset$, $\mathcal{Z}(g) \cap \boldsymbol{X}^{\underline{(i)}} \neq \emptyset$. $\square$

Lemma 5.3.8 (Cycle Implies Non–Existence of Solutions) *Let $2 \leq m \leq n$ and let*

$$\zeta : \{1, \ldots, m\} \to \{1, \ldots, n\}$$

be injective. Let $\boldsymbol{U}^{(i)}, \boldsymbol{V}^{(i)}$ be opposite $\zeta(i)$-faces of $\boldsymbol{X}$ for $i = 1, \ldots, m$, i.e.

$$\{\boldsymbol{U}^{(i)}, \boldsymbol{V}^{(i)}\} = \{\boldsymbol{X}^{\overline{(\zeta(i))}}, \boldsymbol{X}^{\underline{(\zeta(i))}}\}$$

and let

$$\{h_1, \ldots, h_m\} \subseteq \{g_1, \ldots, g_n\}.$$

If

- *$\mathcal{Z}(h_i)$ is disjoint from $\boldsymbol{U}^{(i)}$ and $\boldsymbol{V}^{(i+1)}$ for $i = 1, \ldots, m-1$ and*
- *$\mathcal{Z}(h_m)$ is disjoint from $\boldsymbol{U}^{(m)}$ and $\boldsymbol{V}^{(1)}$*

then $\boldsymbol{g}(\boldsymbol{x}) \neq 0$ for all $\boldsymbol{x} \in \boldsymbol{X}$. $\square$

Proof. Let $\zeta, h_i, \boldsymbol{U}^{(i)}, \boldsymbol{V}^{(i)}, i = 1, \ldots, m$ as in Lemma 5.3.8. If $\mathcal{Z}(h_i) \cap \boldsymbol{X} = \emptyset$ for some i, then obviously $\boldsymbol{g}(\boldsymbol{x}) \neq 0$ for all $\boldsymbol{x} \in \boldsymbol{X}$. Hence assume $\mathcal{Z}(h_i) \cap \boldsymbol{X} \neq \emptyset$ for all i. As every $\zeta(i)$-face of $\boldsymbol{X}$ is disjoint from some $\mathcal{Z}(h_j)$ it follows from Lemma 5.3.7 that $\overline{X}_{\zeta(i)} \neq \underline{X}_{\zeta(i)}$ for all i.

According to Lemma 5.3.5 there exists no solution of $\boldsymbol{g}$ in

$$\mathcal{X} = \bigcup_{i=1}^{m-1} \mathcal{H}(\boldsymbol{U}^{(i)}, \boldsymbol{V}^{(i+1)}) \cup \mathcal{H}(\boldsymbol{U}^{(m)}, \boldsymbol{V}^{(1)}).$$

It remains to show that $\boldsymbol{X} \subseteq \mathcal{X}$. Let $\boldsymbol{x} \in \boldsymbol{X}$ arbitrary but fixed. Let

$$e_i = \begin{cases} (x_{\zeta(i)} - \underline{X}_{\zeta(i)})/(\overline{X}_{\zeta(i)} - \underline{X}_{\zeta(i)}) & \text{if} \quad \boldsymbol{U}^{(i)} = \boldsymbol{X}^{\underline{(\zeta(i))}}, \quad \boldsymbol{V}^{(i)} = \boldsymbol{X}^{\overline{(\zeta(i))}} \\ (\overline{X}_{\zeta(i)} - x_{\zeta(i)})/(\overline{X}_{\zeta(i)} - \underline{X}_{\zeta(i)}) & \text{if} \quad \boldsymbol{U}^{(i)} = \boldsymbol{X}^{\overline{(\zeta(i))}}, \quad \boldsymbol{V}^{(i)} = \boldsymbol{X}^{\underline{(\zeta(i))}} \end{cases}$$

for $i = 1, \ldots, m$. According to Lemma 5.3.6 $\boldsymbol{x} \notin \mathcal{X}$ if and only if

$$\begin{array}{rcl} e_1 + (1 - e_2) & > & 1 \\ e_2 + (1 - e_3) & > & 1 \\ & \vdots & \\ e_m + (1 - e_1) & > & 1. \end{array} \tag{5.3.1}$$

Summing up all inequalities in (5.3.1) we obtain $m > m$, i.e. the inequalities are inconsistent and therefore $\boldsymbol{x} \in \mathcal{X}$. □

Proof of Theorem 5.3.3. Assume $\boldsymbol{g}$ is face disjoint from $\boldsymbol{X}$ but $\boldsymbol{g}$ is not strongly orthogonal in $\boldsymbol{X}$. If $\mathcal{Z}(g_i) \cap \boldsymbol{X} = \emptyset$ for some g_i then obviously $\mathcal{Z}(\boldsymbol{g}) \cap \boldsymbol{X} = \emptyset$. Thus, in the following assume $\mathcal{Z}(g_i) \cap \boldsymbol{X} \neq \emptyset$ for all i. As $\boldsymbol{g}$ is face disjoint from $\boldsymbol{X}$, there exists

$$\varphi : \{\boldsymbol{X}^{\overline{(1)}}, \boldsymbol{X}^{\underline{(1)}}, \boldsymbol{X}^{\overline{(2)}}, \boldsymbol{X}^{\underline{(2)}}, \ldots, \boldsymbol{X}^{\overline{(n)}}, \boldsymbol{X}^{\underline{(n)}}\} \to \{g_1, \ldots, g_n\}$$

such that

$$\mathcal{Z}(\varphi(\boldsymbol{T})) \cap \boldsymbol{T} = \emptyset$$

for all faces $\boldsymbol{T}$ of $\boldsymbol{X}$. Note that according to Lemma 5.3.7 no two faces of $\boldsymbol{X}$ are equal. Let $\mathcal{G}$ be an (undirected) graph whose set of vertices $\mathcal{V}$ is

$$\mathcal{V} = \{g_i \mid i = 1, \ldots, n\}$$

and whose set of edges $\mathcal{E}$ is

$$\mathcal{E} = \{\boldsymbol{X}^{\overline{(j)}} \mid j = 1, \ldots, n\}$$

where $\boldsymbol{X}^{\overline{(j)}}$ is an edge between the nodes $\varphi(\boldsymbol{X}^{\overline{(j)}})$ and $\varphi(\boldsymbol{X}^{\underline{(j)}})$. If

$$\varphi(\boldsymbol{X}^{\overline{(j)}}) = \varphi(\boldsymbol{X}^{\underline{(j)}}) = g_i$$

for some i, j then g_i is strongly $\boldsymbol{X}^{\overline{(j)}}$ enclosed and by Corollary 5.2.6 $\varphi(\boldsymbol{T}) \neq g_i$ for all $\boldsymbol{T} \neq \boldsymbol{X}^{\overline{(j)}}, \boldsymbol{X}^{\underline{(j)}}$. In this case $\boldsymbol{X}^{\overline{(j)}}$ is an edge between g_i and itself and g_i is an isolated vertex in $\mathcal{G}$, i.e. no edge except $\boldsymbol{X}^{\overline{(j)}}$ is connected to g_i. As $\boldsymbol{g}$ is not strongly orthogonal in $\boldsymbol{X}$, not all vertices are isolated. The number of non–isolated vertices and edges between non–isolated vertices is the same, hence there exists a cycle consisting of $m \geq 2$ vertices in $\mathcal{G}$. Let

$$\zeta : \{1, \ldots, m\} \to \{1, \ldots, n\}$$

be injective and $\{h_1, \ldots, h_m\} \subseteq \{g_1, \ldots, g_n\}$ such that

$$h_1 \; \boldsymbol{X}^{\overline{(\zeta(1))}} \; h_2 \; \boldsymbol{X}^{\overline{(\zeta(2))}} \; \ldots \; \boldsymbol{X}^{\overline{(\zeta(m-1))}} \; h_m \; \boldsymbol{X}^{\overline{(\zeta(m))}} \; h_1$$

is a cycle. Let $\boldsymbol{U}^{(i)}, \boldsymbol{V}^{(i)}$ be the $\zeta(i)$-faces of $\boldsymbol{X}$ such that

- h_i is disjoint from $\boldsymbol{U}^{(i)}$ and $\boldsymbol{V}^{(i+1)}$ for $i = 1, \ldots, m-1$ and
- h_m is disjoint from $\boldsymbol{U}^{(m)}$ and $\boldsymbol{V}^{(1)}$.

Now, an application of Theorem 5.3.8 completes the proof. □

In order to prove Theorem 5.3.2, we have to generalize Theorem 5.3.3 to linear interval functions. The following observation is useful therefore.

Lemma 5.3.9 *If $\boldsymbol{G}$ is face disjoint from $\boldsymbol{X}$ and not strongly orthogonal in $\boldsymbol{X}$, then $\boldsymbol{g}$ is face disjoint from $\boldsymbol{X}$ and not strongly orthogonal in $\boldsymbol{X}$ for all $\boldsymbol{g} \in \boldsymbol{G}$.* □

Proof. Assume $\boldsymbol{G}$ is face disjoint from $\boldsymbol{X}$ and not strongly orthogonal in $\boldsymbol{X}$ and let $\boldsymbol{g} \in \boldsymbol{G}$ arbitrary but fixed. Obviously $\boldsymbol{g}$ is face disjoint from $\boldsymbol{X}$. Assume $\boldsymbol{g}$ is strongly orthogonal in $\boldsymbol{X}$. Then there exists a permutation π such that g_i is strongly $\boldsymbol{X}^{\overline{(\pi(i))}}$ enclosed and by Corollary 5.2.6 $\mathcal{Z}(g_i)$ intersects $\boldsymbol{X}^{\overline{(\pi(j))}}, \boldsymbol{X}^{\underline{(\pi(j))}}$ for all $j \neq i$. Therefore $\mathcal{Z}(G_i)$ intersects $\boldsymbol{X}^{\overline{(\pi(j))}}, \boldsymbol{X}^{\underline{(\pi(j))}}$ for all $j \neq i$. As $\boldsymbol{G}$ is face disjoint from $\boldsymbol{X}$ it follows that $\mathcal{Z}(G_i)$ is disjoint from $\boldsymbol{X}^{\overline{(\pi(i))}}, \boldsymbol{X}^{\underline{(\pi(i))}}$, hence $\boldsymbol{G}$ is strongly orthogonal in $\boldsymbol{X}$ by Lemma 5.1.7, which contradicts the assumption. □

Lemma 5.3.10 *If $\boldsymbol{G}$ is face disjoint from $\boldsymbol{X}$ and not strongly orthogonal in $\boldsymbol{X}$ then $0 \notin \boldsymbol{G}(\boldsymbol{X})$.* □

Proof. Let $\boldsymbol{G}$ be face disjoint from $\boldsymbol{X}$ and not strongly orthogonal in $\boldsymbol{X}$. Assume to the contrary that $0 \in \boldsymbol{G}(\boldsymbol{X})$. Then there exists $\boldsymbol{g} \in \boldsymbol{G}$ and $\boldsymbol{x} \in \boldsymbol{X}$ such that $\boldsymbol{g}(\boldsymbol{x}) = 0$. But according to Lemma 5.3.9 $\boldsymbol{g}$ is face disjoint from $\boldsymbol{X}$ and not strongly orthogonal in $\boldsymbol{X}$ and therefore $\boldsymbol{g}(\boldsymbol{x}) \neq 0$ for all $\boldsymbol{x} \in \boldsymbol{X}$ by Theorem 5.3.3. □

Now the proof of Theorem 5.3.2 is straight forward.

Proof of Theorem 5.3.2. Let $\boldsymbol{G}$ be a linearization of $\boldsymbol{f}$ in $\boldsymbol{X}$ and assume $\boldsymbol{G}$ is face disjoint from $\boldsymbol{X}$ but not strongly orthogonal in $\boldsymbol{X}$. Then $0 \notin \boldsymbol{G}(\boldsymbol{X})$ by Lemma 5.3.10, hence $\boldsymbol{f}(\boldsymbol{x}) \neq 0$ for all $\boldsymbol{x} \in \boldsymbol{X}$. □

5.4 Linear Tightening

In this section we introduce the notion of "linear tightening". Linear tightening is an operator which takes a linear interval function G and a box $\boldsymbol{X}$ and shrinks the box in one coordinate direction *optimally* such that no solution of G in $\boldsymbol{X}$ gets lost thereby. Linear tightening is illustrated at two examples in Figure 5.4.1. From left to right, linear tightening is first applied in vertical direction (middle graph) and then horizontally (right graph). Note that in the second example horizontal tightening did not give any improvement. From this picture one can see already that after having applied linear tightening in both coordinate directions, no more reduction can be achieved by further tightening steps. One also observes that it does not matter whether we tighten first horizontally and then vertically or vice versa. Further, a reduction can be achieved only if $\mathcal{Z}(G)$ is disjoint from some face of $\boldsymbol{X}$. Note that in the second example vertical tightening gave a simultaneous reduction on both faces of $\boldsymbol{X}$. As will be shown in this section, linear tightening in some other direction will not give a further improvement in such a situation.

5.4.1 Elementary Properties of Linear Tightening

Throughout this section let $G : \mathbb{R}^n \to \mathbb{IR}$ be a linear interval function and let $\boldsymbol{X} \in \mathbb{IR}^n$.

Definition 5.4.1 (Linear Tightening) *The function*

$$\text{tight} : \mathbb{IR}^n \times (\mathbb{R}^n \to \mathbb{IR}) \times \{1, \ldots, n\} \to \mathbb{IR}^n_{\emptyset}$$

is defined as

$$\text{tight}(\boldsymbol{X}, G, j)_i = \begin{cases} X_i & \text{if } j \neq i \\ [\{x_i \in X_i \mid 0 \in G(X_1, \ldots, x_i, \ldots, X_n)\}] & \text{else.} \end{cases} \quad \square$$

In the following we implicitly extend all interval functions to empty interval arguments such that if some argument is empty, then the function value is also empty.

Theorem 5.4.2 (Linear Tightening Preserves Solutions) *Let $\boldsymbol{Y} = \text{tight}(\boldsymbol{X}, G, j)$. Then*

$$\mathcal{Z}(G) \cap \boldsymbol{X} = \mathcal{Z}(G) \cap \boldsymbol{Y}. \quad \square$$

Proof. Let $j \in \{1, \ldots, n\}$ arbitrary but fixed. Let $\boldsymbol{Y} = \text{tight}(\boldsymbol{X}, G, j)$ and let $\boldsymbol{x} \in \boldsymbol{X}$ such that $0 \in G(\boldsymbol{x})$. Assume $\boldsymbol{x} \notin \boldsymbol{Y}$. As $X_i = Y_i$ for all $i \neq j$ it holds that $x_j \notin Y_j$. From Definition 5.4.1 it follows that $0 \notin G(\boldsymbol{x})$, which is a contradiction to the assumption $0 \in G(\boldsymbol{x})$. □

The following corollary gives a condition when linear tightening does not give an improvement.

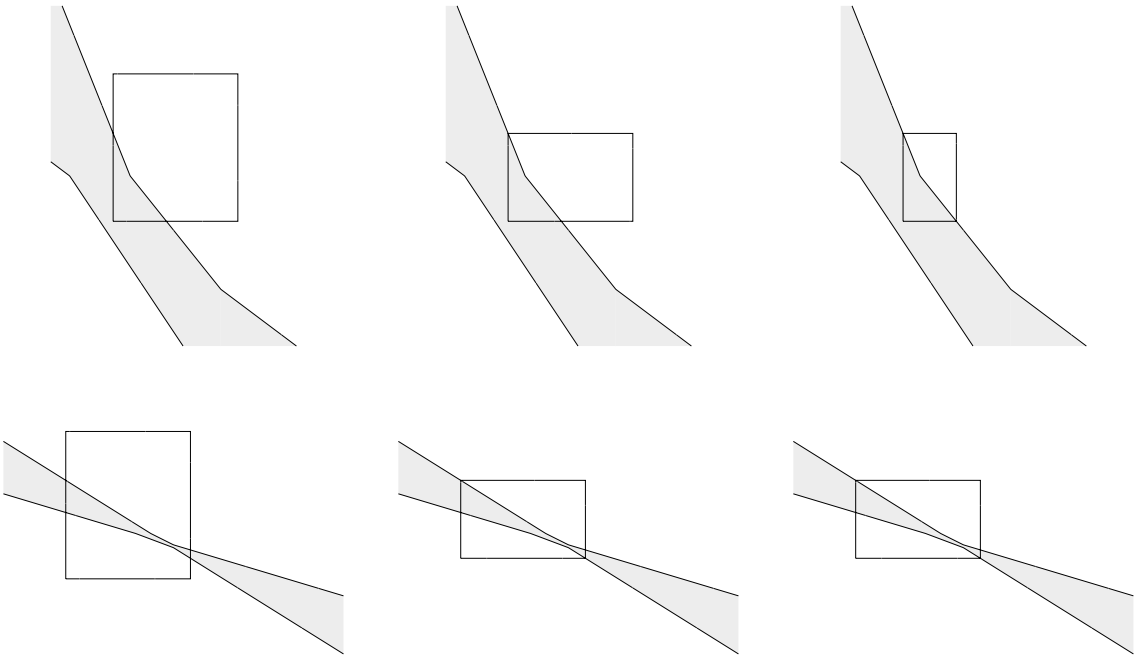


Figure 5.4.1: Illustration of linear tightening. From left to right, tightening is first applied vertically and then horizontally.

Corollary 5.4.3 (Face Condition for Linear Tightening) *Let* $\boldsymbol{X} \neq \emptyset$ *and* $\boldsymbol{Y} = \mathrm{tight}(\boldsymbol{X}, G, j)$. *Then*

$$\begin{aligned} \overline{Y}_j = \overline{X}_j \quad &\text{iff} \quad 0 \in G(\boldsymbol{X}^{\overline{(j)}}) \\ \underline{Y}_j = \underline{X}_j \quad &\text{iff} \quad 0 \in G(\boldsymbol{X}^{\underline{(j)}}). \ \square \end{aligned}$$

Next, we establish a connection between linear tightening and face enclosedness.

Corollary 5.4.4 (Linear Tightening and Strong Face Enclosedness) *Let* $\boldsymbol{Y} = \mathrm{tight}(\boldsymbol{X}, G, j) \neq \emptyset$ *such that* $Y_j \subseteq \mathrm{int}(X_j)$. *Then* G *is strongly* $\boldsymbol{X}^{\overline{(j)}}$ *enclosed.* $\square$

Proof. Follows from Corollary 5.4.3 and Lemma 5.1.7. $\square$

Definition 5.4.5 (Generalized and Hull Division) *The function* $\mathrm{gdiv} : \mathbb{IR}_{\emptyset} \times \mathbb{IR}_{\emptyset} \times \mathbb{IR}_{\emptyset} \to \mathcal{P}(\mathbb{R})$ *is defined as*

$$\mathrm{gdiv}(N, D, X) = \{x \in X \mid n = dx \ \text{ for some } \ n \in N, d \in D\}.$$

The function $\mathrm{hdiv} : \mathbb{IR}_{\emptyset} \times \mathbb{IR}_{\emptyset} \times \mathbb{IR}_{\emptyset} \to \mathbb{IR}_{\emptyset}$ *is defined as*

$$\mathrm{hdiv}(N, D, X) = [\mathrm{gdiv}(N, D, X)]. \ \square$$

Note that if $0 \notin D$ then $\mathrm{gdiv}(N, D, X) = \mathrm{hdiv}(N, D, X) = N/D \cap X$. The following theorem shows how to evaluate the linear tightening function efficiently.

Theorem 5.4.6 (Algorithm for Linear Tightening) *Let* $G(\boldsymbol{x}) = y + \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{c})$. *Then*

$$\mathrm{tight}(\boldsymbol{X}, G, j)_i = \begin{cases} X_i & \text{if } j \neq i \\ c_i + \mathrm{hdiv}(-R, A_i, X_i - c_i) & \text{else,} \end{cases}$$

where

$$R = y + \sum_{k \neq i} A_k (X_k - c_k). \ \square$$

Proof. Let G, R as in Theorem 5.4.6. We have to show that

$$[\{x_i \in X_i \mid 0 \in G(X_1, \ldots, x_i \ldots, X_n)\}] = c_i + \mathrm{hdiv}(-R, A_i, X_i - c_i).$$

$$\begin{aligned}
&[\{x_i \in X_i \mid 0 \in G(X_1, \ldots, x_i, \ldots, X_n)\}] \\
&= [\{x_i \in X_i \mid 0 \in A_i(x_i - c_i) + R\}] \\
&= [\{x_i \in X_i \mid a_i(x_i - c_i) + r = 0 \text{ for some } r \in R, a_i \in A_i\}] \\
&= c_i + [\{x_i \in X_i - c_i \mid -r = a_i x_i \text{ for some } r \in R, a_i \in A_i\}] \\
&= c_i + \mathrm{hdiv}(-R, A_i, X_i - c_i). \ \square
\end{aligned}$$

Let $\boldsymbol{Y} = \mathrm{tight}(\boldsymbol{X}, G, j) \neq \emptyset$. Geometrically it is easy to see that if $Y_j \subseteq \mathrm{int}(X_j)$, then G is strongly $\boldsymbol{X}^{\overline{(j)}}$ and weakly $\boldsymbol{Y}^{\overline{(j)}}$ enclosed. By slightly enlarging $\boldsymbol{Y}$ in its j-th coordinate direction we can achieve that G is also strongly $\boldsymbol{Y}^{\overline{(j)}}$ enclosed. Therefore, we define the following modification of linear tightening:

Definition 5.4.7 (Linear δ–Tightening) *For all $\delta > 0$ the function*

$$\mathrm{tight}_\delta : \mathbb{IR}^n \times (\mathbb{R}^n \to \mathbb{IR}) \times \{1, \ldots, n\} \to \mathbb{IR}^n$$

is defined as

$$\mathrm{tight}_\delta(\boldsymbol{X}, G, j)_i = \begin{cases} \mathrm{tight}(\boldsymbol{X}, G, j)_i & \text{if } j \neq i \\ (\mathrm{tight}(\boldsymbol{X}, G, j)_i + [-\delta, \delta]) \cap X_i & \text{else. } \square \end{cases}$$

Now, we consider iterated application of linear tightening steps.

Definition 5.4.8 (Reduction, Path) *Let $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{IR}^n$. We write*

$$\boldsymbol{X} \xrightarrow{G,j} \boldsymbol{Y} \text{ respectively } \boldsymbol{X} \xrightarrow{G,j}_\delta \boldsymbol{Y}$$

if

$$\boldsymbol{Y} = \mathrm{tight}(\boldsymbol{X}, G, j) \text{ respectively } \boldsymbol{Y} = \mathrm{tight}_\delta(\boldsymbol{X}, G, j)$$

and say $\boldsymbol{X}$ was (δ–)reduced to $\boldsymbol{Y}$ by G in direction j. Let

$$\begin{aligned} p^* &= \langle (G_{i_1}, j_1), (G_{i_2}, j_2), \ldots, (G_{i_m}, j_m) \rangle \\ \delta^* &= \langle \delta_1, \delta_2, \ldots, \delta_m \rangle \end{aligned}$$

where $i_\ell, j_\ell \in \{1, \ldots, n\}$, $\delta_\ell \geq 0$ and G_{i_ℓ} is a linear interval function for all $\ell = 1, \ldots, m$. Then p^* is called path. The ℓ-th component (G_{i_ℓ}, j_ℓ) of of p^* is denoted by p_ℓ. We write

$$\boldsymbol{X} \xrightarrow{p^*} \boldsymbol{Y} \text{ respectively } \boldsymbol{X} \xrightarrow{p^*}_{\delta^*} \boldsymbol{Y}$$

if

$$\boldsymbol{X} \xrightarrow{p_1} \ldots \xrightarrow{p_m} \boldsymbol{Y} \text{ respectively } \boldsymbol{X} \xrightarrow{p_1}_{\delta_1} \ldots \xrightarrow{p_m}_{\delta_m} \boldsymbol{Y}.$$

Further, we write

$$\boldsymbol{X} \xrightarrow{p^*} \text{ respectively } \boldsymbol{X} \xrightarrow{p^*}_{\delta^*}$$

to denote the uniquely defined $\boldsymbol{Y}$ such that $\boldsymbol{X} \xrightarrow{p^*} \boldsymbol{Y}$ respectively $\boldsymbol{X} \xrightarrow{p^*}_{\delta^*} \boldsymbol{Y}$. $\square$

Notation. Let $\delta^* = \langle \delta_1, \ldots, \delta_m \rangle$ and let $\delta \in \mathbb{R}$. For $\sim \in \{=, \leq, \geq, <, >\}$ we write $\delta^* \sim \delta$ if $\delta_i \sim \delta$ for all $i = 1, \ldots, m$.

For a single linear interval function G it does not matter in which order linear tightening steps are applied. At the end of such a tightening sequence we always obtain the smallest box in each coordinate direction where tightening was applied:

Theorem 5.4.9 (Order Independence of Tightening Steps) *Let*

$$p^* = \langle (G, j_\ell) \mid \ell = 1, \ldots, m \rangle$$

and let $\boldsymbol{X} \xrightarrow{p^*} \boldsymbol{Y}$. *Then for all* i

$$Y_i = \begin{cases} X_i & \text{if } i \notin \{j_1, \ldots, j_m\} \\ [\{x_i \in X_i \mid 0 \in G(X_1, \ldots, x_i, \ldots, X_n)\}] & \text{else. } \square \end{cases}$$

Proof. Let p^* as in Theorem 5.4.9, let $\boldsymbol{X} \xrightarrow{p^*} \boldsymbol{Y}$ and let $\tilde{\boldsymbol{Y}}$ such that for all i

$$\tilde{Y}_i = \begin{cases} X_i & \text{if } i \notin \{j_1, \ldots, j_m\} \\ [\{x_i \in X_i \mid 0 \in G(X_1, \ldots, x_i, \ldots, X_n)\}] & \text{else.} \end{cases}$$

Further, let

$$\boldsymbol{X} = {}^0\boldsymbol{Y} \xrightarrow{p_1} {}^1\boldsymbol{Y} \xrightarrow{p_2} \ldots \xrightarrow{p_m} {}^m\boldsymbol{Y} = \boldsymbol{Y}.$$

- We show that $\boldsymbol{Y} \subseteq \tilde{\boldsymbol{Y}}$, i.e. $Y_i \subseteq \tilde{Y}_i$ for all i. Let i arbitrary but fixed. If $i \notin \{j_1, \ldots, j_m\}$ then $Y_i = X_i = \tilde{Y}_i$. Otherwise, let ℓ such that $i = j_\ell$. Then

$$\begin{aligned} Y_i &\subseteq {}^\ell Y_i \\ &= [\{x_i \in {}^{\ell-1}Y \mid 0 \in G({}^{\ell-1}Y_1, \ldots, x_i, \ldots, {}^{\ell-1}Y_n)\}] \\ &\subseteq [\{x_i \in X_i \mid 0 \in G(X_1, \ldots, x_i, \ldots, X_n\}] \\ &= \tilde{Y}_i. \end{aligned}$$

- We show that $\tilde{\boldsymbol{Y}} \subseteq \boldsymbol{Y}$. Assume to the contrary there exists $\boldsymbol{y} \in \tilde{\boldsymbol{Y}}$ such that $\boldsymbol{y} \notin \boldsymbol{Y}$. Then there exists i such that $y_i \in \tilde{Y}_i$ but $y_i \notin Y_i$. From the definition of $\tilde{\boldsymbol{Y}}$ and $\boldsymbol{y} \in \tilde{\boldsymbol{Y}}$ it follows that there exists $\boldsymbol{z} \in \boldsymbol{X}$ such that $z_i = y_i$ and $0 \in G(\boldsymbol{z})$. Thus, $\boldsymbol{z}$ is a solution of G in $\boldsymbol{X}$, which is not in $\boldsymbol{Y}$. Hence, there exists $0 < \ell \leq m$ such that $\boldsymbol{z} \in {}^{\ell-1}\boldsymbol{Y}$ but $\boldsymbol{z} \notin {}^{\ell}\boldsymbol{Y}$, which is a contradiction to the solution preservation property of linear tightening (Theorem 5.4.2). $\square$

From Theorem 5.4.9 it follows immediately that for a single linear interval function G, a sequence of tightening steps may be permuted arbitrarily.

Corollary 5.4.10 (Permutation of Tightening Sequence) *Let*

$$\begin{aligned} p^* &= \langle (G, j_\ell) \mid \ell = 1, \ldots, m \rangle \\ q^* &= \langle (G, j_{\pi(\ell)}) \mid \ell = 1, \ldots, m \rangle \end{aligned}$$

where $\pi : \{1, \ldots, m\} \to \{1, \ldots, m\}$ *is a permutation. Then*

$$\boldsymbol{X} \xrightarrow{p^*} = \boldsymbol{X} \xrightarrow{q^*}$$

for all $\boldsymbol{X} \in \mathbb{IR}^n$. $\square$

Note that neither Theorem 5.4.9 nor Corollary 5.4.10 can be generalized to δ-tightening.

5.4.2 Face Disjointness by Linear Tightening

Throughout this section let $\boldsymbol{G} : \mathbb{R}^n \to \mathbb{IR}^n$ be a linear interval function and let

$$p^* = \langle (G_{i_\ell}, j_\ell) \mid \ell = 1, \ldots, m \rangle$$

for some $m \geq 0$ and $i_\ell, j_\ell \in \{1, \ldots, n\}$. We show that $\boldsymbol{X} \xrightarrow{p^*}_{\delta^*} \boldsymbol{Y} \subseteq \text{int}(\boldsymbol{X})$ for some $\delta^* > 0$ implies that $\boldsymbol{G}$ is face disjoint from $\boldsymbol{Y}$. Further, we show how to obtain for each face $\boldsymbol{T}$ of $\boldsymbol{Y}$ an index i such

that $\mathcal{Z}(G_i) \cap \boldsymbol{T} = \emptyset$ as a side product during computation of the tightening sequence. Unfortunately this works only if $\delta^* > 0$, but in the next section we show that the case $\delta^* = 0$ is in some sense equivalent to the case $\delta^* > 0$ if δ^* is "small enough".

For a given direction j it is sometimes useful to know the last tightening step in p^*, which gave a reduction in direction j or whether no reduction in direction j was achieved at all. We define functions $\overline{\Delta}$, $\underline{\Delta}$ which give us this information.

Definition 5.4.11 (Face Disjointness Functions) *For $\delta^* \geq 0$ the partial functions $\overline{\Delta}_{\delta^*}$, $\underline{\Delta}_{\delta^*}$ are defined as follows. Let*

$$\boldsymbol{X} = {}^0\boldsymbol{X} \xrightarrow{p_1}_{\delta_1} {}^1\boldsymbol{X} \xrightarrow{p_2}_{\delta_2} \ldots \xrightarrow{p_m}_{\delta_m} {}^m\boldsymbol{X} \neq \emptyset.$$

Then

$$\overline{\Delta}_{\delta^*}(\boldsymbol{X}, p^*, j) = \begin{cases} i_{\hat{\ell}} & \text{if } \hat{\ell} = \max\{\ell \mid {}^{\ell-1}\overline{X}_j > {}^{\ell}\overline{X}_j\} \text{ exists} \\ \uparrow & \text{else} \end{cases}$$

$$\underline{\Delta}_{\delta^*}(\boldsymbol{X}, p^*, j) = \begin{cases} i_{\hat{\ell}} & \text{if } \hat{\ell} = \max\{\ell \mid {}^{\ell-1}\underline{X}_j < {}^{\ell}\underline{X}_j\} \text{ exists} \\ \uparrow & \text{else. } \square \end{cases}$$

A necessary and sufficient condition for the totality of $\overline{\Delta}_{\delta^*}$, $\underline{\Delta}_{\delta^*}$ is given by the following lemma.

Lemma 5.4.12 *Assume $\boldsymbol{X} \xrightarrow{p^*}_{\delta^*} \boldsymbol{Y} \neq \emptyset$ for some $\delta^* \geq 0$. Then*

$$\overline{\Delta}_{\delta^*}(\boldsymbol{X}, p^*, j) \neq \uparrow \quad \text{and} \quad \underline{\Delta}_{\delta^*}(\boldsymbol{X}, p^*, j) \neq \uparrow \quad \text{for all} \quad j$$

if and only if

$$\boldsymbol{Y} \subseteq \operatorname{int}(\boldsymbol{X}). \ \square$$

Proof. Follows immediately from Definition 5.4.11. $\square$

The following theorem captures the essential property of the face disjointness functions $\overline{\Delta}_{\delta^*}$, $\underline{\Delta}_{\delta^*}$. It tells us which G_i are disjoint from which faces of the tightened box.

Theorem 5.4.13 (Face Disjointness Function) *Assume $\boldsymbol{X} \xrightarrow{p^*}_{\delta^*} \boldsymbol{Y} \neq \emptyset$ for some $\delta^* > 0$.*

(i) *If $\overline{\Delta}_{\delta^*}(\boldsymbol{X}, p^*, j) = i \neq \uparrow$ then $\mathcal{Z}(G_i) \cap \boldsymbol{Y}^{\overline{(j)}} = \emptyset$.*

(ii) *If $\underline{\Delta}_{\delta^*}(\boldsymbol{X}, p^*, j) = i \neq \uparrow$ then $\mathcal{Z}(G_i) \cap \boldsymbol{Y}^{\underline{(j)}} = \emptyset$. $\square$*

For the proof of Theorem 5.4.13 we need the following two lemmas. First, we show that Theorem 5.4.13 holds for paths of length one. Let $G : \mathbb{R}^n \to \mathbb{IR}$ be a linear interval function.

Lemma 5.4.14 *Let*

$$\emptyset \neq \boldsymbol{Y} = \operatorname{tight}_{\delta}(\boldsymbol{X}, G, j)$$

for some $\delta > 0$.

(i) *If $\overline{Y}_j < \overline{X}_j$ then $\mathcal{Z}(G) \cap \boldsymbol{Y}^{\overline{(j)}} = \emptyset$.*

(ii) *If $\underline{Y}_j > \underline{X}_j$ then $\mathcal{Z}(G) \cap \boldsymbol{Y}^{\underline{(j)}} = \emptyset$. $\square$*

Proof. We give a proof of (i), the proof of (ii) is analogous. Let

$$\emptyset \neq \boldsymbol{Y} = \operatorname{tight}_{\delta}(\boldsymbol{X}, G, j)$$

for some $\delta > 0$ and $\overline{Y}_j < \overline{X}_j$. Assume to the contrary that $\mathcal{Z}(G) \cap \boldsymbol{Y}^{\overline{(j)}} \neq \emptyset$. Then $0 \in G(\boldsymbol{y})$ for some $\boldsymbol{y} \in \boldsymbol{Y}^{\overline{(j)}}$. As $\delta > 0$ and $\overline{Y}_j < \overline{X}_j$, it holds that $\boldsymbol{y} \notin \operatorname{tight}(\boldsymbol{X}, G, j)$, which is a contradiction to the solution preserving property of linear tightening (Theorem 5.4.2). $\square$

Lemma 5.4.15 (Face Disjointness from Sub–Box)

(i) *Assume* $\mathcal{Z}(G) \cap \boldsymbol{X}^{\overline{(j)}} = \emptyset$. *Let* $\boldsymbol{Y} \subseteq \boldsymbol{X}$ *such that* $\overline{Y}_j = \overline{X}_j$. *Then* $\mathcal{Z}(G) \cap \boldsymbol{Y}^{\overline{(j)}} = \emptyset$.

(ii) *Assume* $\mathcal{Z}(G) \cap \boldsymbol{X}^{\underline{(j)}} = \emptyset$. *Let* $\boldsymbol{Y} \subseteq \boldsymbol{X}$ *such that* $\underline{Y}_j = \underline{X}_j$. *Then* $\mathcal{Z}(G) \cap \boldsymbol{Y}^{\underline{(j)}} = \emptyset$. □

Proof. We give a proof of (i), the proof of (ii) is analogous. Assume $\mathcal{Z}(G) \cap \boldsymbol{X}^{\overline{(j)}} = \emptyset$ and let $\boldsymbol{Y} \subseteq \boldsymbol{X}$ such that $\overline{Y}_j = \overline{X}_j$. Then $\boldsymbol{Y}^{\overline{(j)}} \subseteq \boldsymbol{X}^{\overline{(j)}}$ and hence $\mathcal{Z}(G) \cap \boldsymbol{Y}^{\overline{(j)}} = \emptyset$. □

Proof of Theorem 5.4.13. We give a proof of (i), the proof of (ii) is analogous. Let $\delta^* > 0$ and let ${}^{\ell}\boldsymbol{X}$, $\ell = 0, \ldots, m$ as in Definition 5.4.11. Assume $\overline{\Delta}_{\delta^*}(\boldsymbol{X}, p^*, j) = i \neq \uparrow$. Then there exists $\hat{\ell}$ such that

$$ {}^{\hat{\ell}-1}\overline{X}_j > {}^{\hat{\ell}}\overline{X}_j = {}^{\hat{\ell}+1}\overline{X}_j = \ldots = {}^{m}\overline{X}_j $$

and

$$ {}^{\hat{\ell}-1}\boldsymbol{X} \xrightarrow{G_i, j}_{\delta_{\hat{\ell}}} {}^{\hat{\ell}}\boldsymbol{X}. $$

From Lemma 5.4.14 it follows that $G_i \cap {}^{\hat{\ell}}\boldsymbol{X}^{\overline{(j)}} = \emptyset$ and from Lemma 5.4.15 it follows that $G_i \cap {}^{m}\boldsymbol{X}^{\overline{(j)}} = \emptyset$. □

In an algorithm for solving systems of equations one would want to use linear tightening and not linear δ-tightening because the former produces more accurate inclusions. Theorem 5.4.13 is essential because it gives conditions for face enclosedness or orthogonality of $\boldsymbol{G}$ in some box and thus ultimately for existence, uniqueness and non–existence of solutions of nonlinear systems of equations. However, the assumptions of this theorem are that $\delta^* > 0$. Now, one expects that if δ^* is small enough, then $\overline{\Delta}_{\delta^*} = \overline{\Delta}_0$ and $\underline{\Delta}_{\delta^*} = \underline{\Delta}_0$. Yet, there are counterexamples where this is not true, so we have the following slightly weaker theorem.

Theorem 5.4.16 (Face Disjointness Function for $\delta = 0$) *For all* $\delta > 0$ *there exists* $0 < \delta^* \leq \delta$ *such that for all* j

$$ \begin{aligned} \overline{\Delta}_0(\boldsymbol{X}, p^*, j) &= \overline{\Delta}_{\delta^*}(\boldsymbol{X}, p^*, j) \\ \underline{\Delta}_0(\boldsymbol{X}, p^*, j) &= \underline{\Delta}_{\delta^*}(\boldsymbol{X}, p^*, j). \ \square \end{aligned} $$

The content of the next section is to prove Theorem 5.4.16 by using a continuity property of linear tightening. As Theorem 5.4.16 is geometrically quite obvious, the reader may skip this rather lengthy proof.

5.4.3 Pseudo Continuity of Linear Tightening

In this section we prove Theorem 5.4.16. First, we show a continuity property of the hull division which is generalized to linear tightening and to linear tightening sequences.

Lemma 5.4.17 (Inclusion Monotonicity of Hull Division) *Let* $N \subseteq \tilde{N}$, $D \subseteq \tilde{D}$ *and* $X \subseteq \tilde{X}$. *Then*

$$ \mathrm{hdiv}(N, D, X) \subseteq \mathrm{hdiv}(\tilde{N}, \tilde{D}, \tilde{X}). \ \square $$

Proof. Let $N, \tilde{N}, D, \tilde{D}, X, \tilde{X}$ as in Lemma 5.4.17.

$$ \begin{aligned} \mathrm{hdiv}(N, D, X) &= [\{x \in X \mid n = dx \ \text{ for some } \ n \in N, d \in D\}] \\ &\subseteq [\{x \in \tilde{X} \mid n = dx \ \text{ for some } \ n \in \tilde{N}, d \in \tilde{D}\}] \\ &= \mathrm{hdiv}(\tilde{N}, \tilde{D}, \tilde{X}). \ \square \end{aligned} $$

Lemma 5.4.18 (Pseudo Continuity of Hull Division) *Let $N, D, X \in \mathbb{IR}$. For all $\varepsilon > 0$ there exists $\delta > 0$ such that for all*

$$\tilde{N} \subseteq N + [-\delta, \delta], \quad \tilde{D} \subseteq D + [-\delta, \delta], \quad \tilde{X} \subseteq X + [-\delta, \delta]$$

it holds that

$$\operatorname{hdiv}(\tilde{N}, \tilde{D}, \tilde{X}) \subseteq \operatorname{hdiv}(N, D, X) + [-\varepsilon, \varepsilon]. \ \square \tag{5.4.1}$$

Remark. Note that hdiv is not continuous, even if its arguments are restricted such that $\operatorname{hdiv}(N, D, X) \neq \emptyset$. For example,

$$\operatorname{hdiv}\big(1, [-1, 1], [-1, 1]\big) = [-1, 1],$$

but for every $0 < \varepsilon \leq 2$ it holds that

$$\operatorname{hdiv}\big(1, [-1, 1], [-1, 1 - \varepsilon]\big) = -1. \ \square$$

Proof. Let $N, D, X \in \mathbb{IR}$ and let $\varepsilon > 0$ arbitrary but fixed. Let $g : \mathbb{R}^3 \to \mathbb{R}$ be defined as

$$g(n, d, x) = n - dx.$$

Note that

$$\operatorname{hdiv}(N, D, X) = [\{x \in X \mid g(n, d, x) = 0 \ \text{ for some } \ n \in N, d \in D\}].$$

- Assume $\operatorname{hdiv}(N, D, X) = \emptyset$. Then

  $$g(n, d, x) \neq 0 \ \text{ for all } \ n \in N, d \in D, x \in X.$$

  From the continuity of g it follows that there exist $\delta > 0$ such that

  $$g(\tilde{n}, \tilde{d}, \tilde{x}) \neq 0 \ \text{ for all } \ \tilde{n} \in n + [-\delta, \delta], \ \ \tilde{d} \in d + [-\delta, \delta], \ \ \tilde{x} \in x + [-\delta, \delta].$$

  Hence

  $$\operatorname{hdiv}(\tilde{N}, \tilde{D}, \tilde{X}) = \emptyset \ \text{ for all } \ \tilde{N} \subseteq N + [-\delta, \delta], \ \ \tilde{D} \subseteq D + [-\delta, \delta], \ \ \tilde{X} \subseteq X + [-\delta, \delta]$$

  and in particular

  $$\operatorname{hdiv}(\tilde{N}, \tilde{D}, \tilde{X}) \subseteq \operatorname{hdiv}(N, D, X) + [-\varepsilon, \varepsilon].$$

  In the following assume $\operatorname{hdiv}(N, D, X) \neq \emptyset$.

- Assume $0 \in D$, $0 \in N$. Then

  $$\operatorname{hdiv}(\tilde{N}, \tilde{D}, \tilde{X}) = \tilde{X} \ \text{ for all } \ \tilde{N}, \tilde{D}, \tilde{X} \in \mathbb{IR}.$$

  Hence,

  $$\operatorname{hdiv}(\tilde{N}, \tilde{D}, \tilde{X}) \subseteq \operatorname{hdiv}(N, D, X) + [-\varepsilon, \varepsilon] \ \text{ for all } \ \tilde{N}, \tilde{D} \in \mathbb{IR}, \ \ \tilde{X} \subseteq X + [-\varepsilon, \varepsilon].$$

  Thus, (5.4.1) holds for $\delta = \varepsilon$.

- Assume $0 \notin D$. From the continuity of interval division restricted to denominators which do not contain zero, it follows that there exists $0 < \delta' < \operatorname{mig}(D)$ such that

  $$\operatorname{q}(\tilde{N}/\tilde{D}, N/D) < \varepsilon \ \text{ if } \ \operatorname{q}(\tilde{N}, N) < \delta', \ \ \operatorname{q}(\tilde{D}, D) < \delta'.$$

  From the inclusion monotonicity of interval division it follows that

  $$\tilde{N}/\tilde{D} \subseteq N/D + [-\varepsilon, \varepsilon] \ \text{ for all } \ \tilde{N} \subseteq N + [-\delta', \delta'], \tilde{D} \subseteq D + [-\delta', \delta'].$$

  Let $\delta = \min\{\varepsilon, \delta'\}$. Then

  $$\tilde{N}/\tilde{D} \cap \tilde{X} \subseteq N/D \cap X + [-\varepsilon, \varepsilon] \ \text{ for all } \ \tilde{N} \subseteq N + [-\delta, \delta], \tilde{D} \subseteq D + [-\delta, \delta], \tilde{X} \subseteq X + [-\delta, \delta],$$

  hence,

  $$\operatorname{hdiv}(\tilde{N}, \tilde{D}, \tilde{X}) \subseteq \operatorname{hdiv}(N, D, X) + [-\varepsilon, \varepsilon] \ \text{ for all } \ \tilde{N} \subseteq N + [-\delta, \delta], \tilde{D} \subseteq D + [-\delta, \delta], \tilde{X} \subseteq X + [-\delta, \delta].$$

- Assume $0 \in D$, $0 \notin N$. We consider the case $\underline{N} > 0$, the case $\overline{N} < 0$ is treated analogously. For arbitrary $\delta > 0$ let $N_\delta = N + [-\delta, \delta], D_\delta = D + [-\delta, \delta], X_\delta = X + [-\delta, \delta]$. Using inclusion monotonicity of hdiv it suffices to show that there exists $\delta > 0$ such that

$$\overline{\mathrm{hdiv}(N, D, X)} + \varepsilon \geq \overline{\mathrm{hdiv}(N_\delta, D_\delta, X_\delta)} \tag{5.4.2}$$
$$\underline{\mathrm{hdiv}(N, D, X)} - \varepsilon \leq \underline{\mathrm{hdiv}(N_\delta, D_\delta, X_\delta)}. \tag{5.4.3}$$

  We show (5.4.2), the proof of (5.4.3) is analogous.

  - If $\overline{\mathrm{hdiv}(N, D, X)} = \overline{X}$ then

$$\begin{aligned} \overline{\mathrm{hdiv}(N, D, X)} + \varepsilon &= \overline{X} + \varepsilon \\ &= \overline{X_\varepsilon} \\ &\geq \overline{\mathrm{hdiv}(N_\varepsilon, D_\varepsilon, X_\varepsilon)} \end{aligned}$$

    and (5.4.2) holds for $\delta = \varepsilon$.

  - Otherwise, $\overline{\mathrm{hdiv}(N, D, X)} < \overline{X}$. First, we show that in this case $\underline{D} < 0$. Assume to the contrary $\underline{D} = 0$. From $\overline{\mathrm{hdiv}(N, D, X)} < \overline{X}$ it follows that $n \neq d\overline{X}$ for all $n \in N$, $d \in D$, and as $\underline{N} > 0$, $\underline{D} = 0$ it holds that $\underline{N} > \overline{D}\,\overline{X}$. As $\mathrm{hdiv}(N, D, X) \neq \emptyset$ there exist $x \in X$ such that $n = dx$ for some $n \in N$, $d \in D$, $x \in X$ and $\overline{X} \geq 0$. But $dx \leq \overline{D}\,\overline{X} < \underline{N}$, which is a contradiction to $dx = n$ for some $n \in N$. Hence, $\underline{D} < 0$.

    Next, we show that there exists $\delta_1 > 0$ such that for all $0 \leq \delta \leq \delta_1$

$$\overline{X}_\delta \notin \mathrm{hdiv}(N_\delta, D_\delta, X_\delta). \tag{5.4.4}$$

    As $\overline{X} \notin \mathrm{hdiv}(N, D, X)$, it holds that

$$g(n, d, \overline{X}) \neq 0 \quad \text{for all} \quad n \in N, d \in D.$$

    From the continuity of g it follows that there exists $\delta_1 > 0$ such that for all $0 \leq \delta \leq \delta_1$

$$g(n, d, \overline{X}_\delta) \neq 0 \quad \text{for all} \quad n \in N_\delta, d \in D_\delta.$$

    and (5.4.4) holds.

    Finally, for all $0 \leq \delta < \min\{\delta_1, \underline{N}, -\underline{D}\}$ it holds that

$$\mathrm{gdiv}(N_\delta, D_\delta, X_\delta) = \begin{cases} X_\delta \cap \{x \in \mathbb{R} \mid x \leq \overline{N}_\delta / \underline{D}_\delta \ \text{ or } \ x \geq \underline{N}_\delta / \overline{D}_\delta\} & \text{if } \overline{D}_\delta \neq 0 \\ X_\delta \cap \{x \in \mathbb{R} \mid x \leq \overline{N}_\delta / \underline{D}_\delta\} & \text{else.} \end{cases}$$

    From $\overline{X}_\delta \notin \mathrm{hdiv}(N_\delta, D_\delta, X_\delta)$ it follows that $\overline{X}_\delta \notin \mathrm{gdiv}(N_\delta, D_\delta, X_\delta)$, hence

$$\overline{\mathrm{hdiv}(N_\delta, D_\delta, X_\delta)} = \overline{N}_\delta / \underline{D}_\delta.$$

    Thus, there exists $\delta > 0$ such that

$$\overline{\mathrm{hdiv}(N_\delta, D_\delta, X_\delta)} - \overline{\mathrm{hdiv}(N, D, X)} = \frac{\overline{N} + \delta}{\underline{D} - \delta} - \frac{\overline{N}}{\underline{D}} \leq \varepsilon$$

    and (5.4.2) holds. □

Lemma 5.4.19 (Inclusion Monotonicity of Linear Tightening) *Let* $\boldsymbol{X} \subseteq \tilde{\boldsymbol{X}}$. *Then*

$$\mathrm{tight}(\boldsymbol{X}, G, j) \subseteq \mathrm{tight}(\tilde{\boldsymbol{X}}, G, j). \ \square$$

Proof. Follows immediately from Lemma 5.4.17. □

We apply the pseudo continuity of the hull division to sequences of linear tightening steps. In the following we write $[-1, 1]^n$ to denote the n-dimensional interval vector with each component $[-1, 1]$.

Lemma 5.4.20 (Pseudo Continuity of Linear Tightening) *For all $\varepsilon > 0$ there exists $\delta > 0$ such that for all $\tilde{\boldsymbol{X}} \subseteq \boldsymbol{X} + \delta[-1,1]^n$*

$$\mathrm{tight}(\tilde{\boldsymbol{X}}, G, j) \subseteq \mathrm{tight}(\boldsymbol{X}, G, j) + \varepsilon[-1,1]^n. \ \square$$

Proof. Follows immediately from Lemma 5.4.18. $\square$

Theorem 5.4.21 (Pseudo Continuity of Linear Tightening Sequence) *Let*

$$\boldsymbol{X} \xrightarrow{p^*} \boldsymbol{Y}.$$

For all $\varepsilon > 0$ there exists $\hat{\delta} > 0$ such that for all $0 \leq \delta^* \leq \hat{\delta}$

$$\boldsymbol{X} \xrightarrow{p^*}_{\delta^*} \subseteq \boldsymbol{Y} + \varepsilon[-1,1]^n. \ \square$$

Proof. Assume $\boldsymbol{X} \xrightarrow{p^*} \boldsymbol{Y}$ and let $\varepsilon > 0$ arbitrary but fixed. We prove Theorem 5.4.21 by induction on the length of p^*. If the length of p^* is zero, then

$$\boldsymbol{X} \xrightarrow{p^*}_{\delta^*} \boldsymbol{X} = \boldsymbol{Y} \subseteq \boldsymbol{Y} + \varepsilon[-1,1]^n.$$

for all δ^*. Now, assume Theorem 5.4.21 is proved for all paths of length $m-1$ and let p^* be a path of length m. Let

$$\boldsymbol{X} = {}^0\boldsymbol{X} \xrightarrow{p_1} \ldots \xrightarrow{p_{m-1}} {}^{m-1}\boldsymbol{X} \xrightarrow{p_m} {}^m\boldsymbol{X} = \boldsymbol{Y}.$$

Let $0 < \delta' < \varepsilon$ arbitrary but fixed. According to Lemma 5.4.20 there exists $\varepsilon' > 0$ such that

$${}^{m-1}\boldsymbol{X} + \varepsilon'[-1,1]^n \xrightarrow{p_m} \subseteq {}^m\boldsymbol{X} + (\varepsilon - \delta')[-1,1]^n.$$

Hence,

$${}^{m-1}\boldsymbol{X} + \varepsilon'[-1,1]^n \xrightarrow{p_m}_{\delta'} \subseteq {}^m\boldsymbol{X} + \varepsilon[-1,1]^n.$$

By induction hypothesis, there exists $\delta'' > 0$ such that

$$\boldsymbol{X} \xrightarrow{p_1}_{\delta''} \ldots \xrightarrow{p_{m-1}}_{\delta''} \subseteq {}^{m-1}\boldsymbol{X} + \varepsilon'[-1,1]^n.$$

Now, for

$$\delta = \min\{\delta', \delta''\}$$

it follows from the inclusion monotonicity of linear tightening (Lemma 5.4.19) that

$$\boldsymbol{X} \xrightarrow{p^*}_{\delta^*} \subseteq \boldsymbol{Y} + \varepsilon[-1,1]^n$$

for all $\delta^* \leq \delta$. $\square$

Proof of Theorem 5.4.16. Let $\boldsymbol{X}$ arbitrary but fixed. We give a proof by induction on the length of p^*. If the length of p^* is zero, then obviously

$$\begin{aligned} \overline{\Delta}_0(\boldsymbol{X}, p^*, j) &= \overline{\Delta}_{\delta^*}(\boldsymbol{X}, p^*, j) \\ \underline{\Delta}_0(\boldsymbol{X}, p^*, j) &= \underline{\Delta}_{\delta^*}(\boldsymbol{X}, p^*, j) \end{aligned} \tag{5.4.5}$$

for all δ^* and for all j. Now, let p^* be an arbitrary but fixed path of length m, let $p^{*'} = \langle p_1, \ldots, p_{m-1} \rangle$ and let $p_m = (G, k)$ for some G, k. The induction hypothesis is as follows: For all $\delta > 0$ there exists $0 < \delta^{*'} \leq \delta$ such that for all j

$$\begin{aligned} \overline{\Delta}_0(\boldsymbol{X}, p^{*'}, j) &= \overline{\Delta}_{\delta^{*'}}(\boldsymbol{X}, p^{*'}, j) \\ \underline{\Delta}_0(\boldsymbol{X}, p^{*'}, j) &= \underline{\Delta}_{\delta^{*'}}(\boldsymbol{X}, p^{*'}, j) \end{aligned} \tag{5.4.6}$$

Let $\delta > 0$ arbitrary but fixed. If $j \neq k$ then (5.4.5) holds for $\delta^* = \delta^{*'} \circ \delta_m$ where $\delta^{*'}$ satisfies (5.4.6) and δ_m is arbitrary. Hence, it remains to show that there exists $\delta^* \leq \delta$ such that (5.4.5) holds for $j = k$. Let

$$\boldsymbol{X} \xrightarrow{p^{*'}} \boldsymbol{Y} \xrightarrow{G,k} \boldsymbol{Z}.$$

and

$$\boldsymbol{X} \xrightarrow{p^{*'}}_{\delta^{*'}} \tilde{\boldsymbol{Y}}.$$

We distinguish 4 cases.

- Assume $Z_k = Y_k$. According to the induction hypothesis and Theorem 5.4.21 there exists $\delta^{*'} \leq \delta$ such that 5.4.6 holds and $\tilde{Y}_k \subseteq Y_k + [-\delta, \delta]$. Let

$$\tilde{\boldsymbol{Y}} \xrightarrow{G,k}_{\delta} \tilde{\boldsymbol{Z}}.$$

  Then $\tilde{Z}_k = \tilde{Y}_k$ and (5.4.5) holds for $\delta^* = \delta^{*'} \circ \delta$.

- Assume $Z_k \subseteq \mathrm{int}(Y_k)$. Then $\mathcal{Z}(G)$ is disjoint from both k-faces of $\boldsymbol{Y}$. From the continuity of G it follows that there exists $\varepsilon > 0$ such that $\mathcal{Z}(G)$ is disjoint from both k-faces of $\boldsymbol{Y}'$ for all $\boldsymbol{Y} \subseteq \boldsymbol{Y}' \subseteq \boldsymbol{Y} + \varepsilon[-1,1]^n$. According to the induction hypothesis and Theorem 5.4.21 there exists $\delta^{*'} \leq \delta$ such that 5.4.6 holds and $\boldsymbol{Y} \subseteq \tilde{\boldsymbol{Y}} \subseteq \boldsymbol{Y} + \varepsilon[-1,1]^n$, i.e. $\mathcal{Z}(G)$ is disjoint from both k-faces of $\tilde{\boldsymbol{Y}}$. Therefore, there exists $0 < \delta_m \leq \delta$ such that

$$\tilde{\boldsymbol{Y}} \xrightarrow{G,k}_{\delta_m} \tilde{\boldsymbol{Z}},$$

  $\tilde{Z}_k \subseteq \mathrm{int}(\tilde{Y}_k)$ and (5.4.5) holds for $\delta^* = \delta^{*'} \circ \delta_m$.

- Assume $\overline{Z}_k < \overline{Y}_k$ and $\underline{Z}_k = \underline{Y}_k$. According to the pseudo continuity (Lemma 5.4.20) and the inclusion monotonicity (Lemma 5.4.19) of linear tightening there exists $\varepsilon > 0$ such that for all $\boldsymbol{Y} \subseteq \boldsymbol{Y}' \subseteq \boldsymbol{Y} + \varepsilon[-1,1]^n$ it holds that

$$\overline{Y}'_k - \overline{Z}'_k > \underline{Z}'_k - \underline{Y}'_k$$

  and

$$\underline{Z}'_k - \underline{Y}'_k \leq \delta,$$

  where

$$\boldsymbol{Y}' \xrightarrow{G,k} \boldsymbol{Z}'.$$

  According to the induction hypothesis and Theorem 5.4.21 there exists $\delta^{*'} \leq \delta$ such that 5.4.6 holds and $\boldsymbol{Y} \subseteq \tilde{\boldsymbol{Y}} \subseteq \boldsymbol{Y} + \varepsilon[-1,1]^n$. Let $\delta_m \leq \delta$ such that

$$\overline{\tilde{Y}}_k - \overline{\tilde{Z}}_k > \delta_m > \underline{\tilde{Z}}'_k - \underline{\tilde{Y}}_k,$$

  where

$$\tilde{\boldsymbol{Y}} \xrightarrow{G,k} \tilde{\boldsymbol{Z}}'.$$

  Now, let

$$\tilde{\boldsymbol{Y}} \xrightarrow{G,k}_{\delta_m} \tilde{\boldsymbol{Z}}.$$

  Then $\overline{\tilde{Z}}_k < \overline{\tilde{Y}}_k$ and $\underline{\tilde{Z}}_k = \underline{\tilde{Y}}_k$ and (5.4.5) holds for $\delta^* = \delta^{*'} \circ \delta_m$.

- Assume $\overline{Z}_k = \overline{Y}_k$ and $\underline{Z}_k > \underline{Y}_k$. This case is analogous to the previous case. □

In the following we write $\overline{\Delta}$, $\underline{\Delta}$ instead of $\overline{\Delta}_0$, $\underline{\Delta}_0$.

5.4.4 Existence, Uniqueness and Non–Existence by Linear Tightening

In this section we derive existence, uniqueness and non–existence tests for solutions of a nonlinear system of equations $\boldsymbol{f}$ in a box $\boldsymbol{X}$. The tests are based on the results of Section 5.1, 5.2, 5.3. Hence, we consider a linearization $\boldsymbol{G}$ of $\boldsymbol{f}$ in $\boldsymbol{X}$ and derive conditions for orthogonality of $\boldsymbol{G}$ in $\boldsymbol{X}$, for the regularity of the coefficient matrix of $\boldsymbol{G}$ and for $0 \notin \boldsymbol{G}(\boldsymbol{X})$. The conditions are given by properties of the boxes in a linear tightening sequence of $\boldsymbol{G}$ starting at $\boldsymbol{X}$. In the following let $\boldsymbol{G} : \mathbb{R}^n \to \mathbb{IR}^n$ and let

$$p^* = \langle (G_{i_1}, j_1), (G_{i_2}, j_2), \ldots, (G_{i_m}, j_m) \rangle.$$

Theorem 5.4.22 *Assume* $\boldsymbol{X} \xrightarrow{p^*} \boldsymbol{Y} \neq \emptyset$ *and*

$$\overline{\Delta}(\boldsymbol{X}, p^*, j) = \underline{\Delta}(\boldsymbol{X}, p^*, j) = i \neq \uparrow$$

for some $i, j \in \{1, \ldots, n\}$. *Then the following holds.*

(i) *For all* $\delta > 0$ *there exists* $0 < \delta^* \leq \delta$ *such that the linear interval function* G_i *is strongly* $\tilde{\boldsymbol{Y}}^{\overline{(j)}}$ *enclosed, where* $\boldsymbol{X} \overset{p^*}{\longrightarrow}_{\delta^*} \tilde{\boldsymbol{Y}}$.

(ii) G_i *is weakly* $\boldsymbol{Y}^{\overline{(j)}}$ *enclosed.*

(iii) $\overline{\Delta}(\boldsymbol{X}, p^*, j') \neq i, \ \underline{\Delta}(\boldsymbol{X}, p^*, j') \neq i$ *for all* $j' \neq j$.

(iv) $\boldsymbol{Y} \overset{G_i, j'}{\longrightarrow} \boldsymbol{Y}$ *for all* $j' \neq j$.

Proof. Assume $\boldsymbol{X} \overset{p^*}{\longrightarrow} \boldsymbol{Y} \neq \emptyset$ and $\overline{\Delta}(\boldsymbol{X}, p^*, j) = \underline{\Delta}(\boldsymbol{X}, p^*, j) = i \neq \uparrow$.

(i) Let $\delta > 0$ arbitrary but fixed. According to Theorem 5.4.16 there exists $0 < \delta^* \leq \delta$ such that $\overline{\Delta}_{\delta^*} = \overline{\Delta}$ and $\underline{\Delta}_{\delta^*} = \underline{\Delta}$. Let

$$\boldsymbol{X} = {}^0\boldsymbol{X} \overset{p_1}{\longrightarrow}_{\delta_1} {}^1\boldsymbol{X} \overset{p_2}{\longrightarrow}_{\delta_2} \ldots \overset{p_m}{\longrightarrow}_{\delta_m} {}^m\boldsymbol{X} = \tilde{\boldsymbol{Y}}.$$

We have to show that G_i is strongly $\tilde{\boldsymbol{Y}}^{\overline{(j)}}$ enclosed. Let $\ell_1, \ell_2 < m$ such that

$$\begin{aligned} {}^{\ell_1}\overline{X}_j &> {}^{\ell_1+1}\overline{X}_j = \ldots = {}^m\overline{X}_j \\ {}^{\ell_2}\underline{X}_j &< {}^{\ell_2+1}\underline{X}_j = \ldots = {}^m\underline{X}_j. \end{aligned}$$

Assume $\ell_1 \leq \ell_2$, the case $\ell_1 \geq \ell_2$ is analogous. According to Lemma 5.4.14

$$G_i({}^{\ell_1}\boldsymbol{X}^{\overline{(j)}}) > 0 \quad \text{or} \quad G_i({}^{\ell_1}\boldsymbol{X}^{\overline{(j)}}) < 0.$$

Assume $G_i({}^{\ell_1}\boldsymbol{X}^{\overline{(j)}}) > 0$, the case $G_i({}^{\ell_1}\boldsymbol{X}^{\overline{(j)}}) < 0$ is analogous. From Lemma 5.4.15 it follows that

$$G_i({}^{\ell}\boldsymbol{X}^{\overline{(j)}}) > 0 \quad \text{for all} \quad \ell \geq \ell_1$$

and in particular,

$$G_i({}^{\ell_2}\boldsymbol{X}^{\overline{(j)}}) > 0. \tag{5.4.7}$$

Now, it suffices to show that $G_i({}^{\ell_2}\boldsymbol{X}^{\underline{(j)}}) < 0$ because then $G_i({}^{\ell}\boldsymbol{X}^{\underline{(j)}}) < 0$ for all $\ell \geq \ell_2$, hence

$$G_i(\boldsymbol{Y}^{\overline{(j)}}) > 0, \ \ G_i(\boldsymbol{Y}^{\underline{(j)}}) < 0$$

Assume to the contrary $\overline{G_i({}^{\ell_2}\boldsymbol{X}^{\underline{(j)}})} \geq 0$. According to Lemma 5.4.14, $0 \notin G_i({}^{\ell_2}\boldsymbol{X}^{\underline{(j)}})$, hence

$$G_i({}^{\ell_2}\boldsymbol{X}^{\underline{(j)}}) > 0. \tag{5.4.8}$$

As G_i is a linear interval function, (5.4.7) and (5.4.8) imply $G_i({}^{\ell_2}\boldsymbol{X}) > 0$ and thus ${}^{\ell_2+1}\boldsymbol{X} = \emptyset$. Hence $\tilde{\boldsymbol{Y}} = \emptyset$ and as $\tilde{\boldsymbol{Y}} \supseteq \boldsymbol{Y}$, it follows that $\boldsymbol{Y} = \emptyset$, which is a contradiction to the assumption.

(ii) Assume G_i is not weakly $\boldsymbol{Y}^{\overline{(j)}}$ enclosed. From the continuity of G_i, it follows that G_i is not strongly $\tilde{\boldsymbol{Y}}^{\overline{(j)}}$ enclosed for all $\tilde{\boldsymbol{Y}}$ in a neighborhood of $\boldsymbol{Y}$, which contradicts (i) and Theorem 5.4.21.

(iii) Let δ^* as in (i), and let $\boldsymbol{X} \overset{p^*}{\longrightarrow}_{\delta^*} \tilde{\boldsymbol{Y}}$. Then G_i is strongly $\tilde{\boldsymbol{Y}}^{\overline{(j)}}$ enclosed and by Corollary 5.2.6

$$\mathcal{Z}(G_i) \cap \tilde{\boldsymbol{Y}}^{\overline{(j')}} \neq \emptyset, \ \ \mathcal{Z}(G_i) \cap \tilde{\boldsymbol{Y}}^{\underline{(j')}} \neq \emptyset$$

for all $j' \neq j$. From Theorem 5.4.13 it follows that $\overline{\Delta}_{\delta^*}(\boldsymbol{X}, p^*, j') \neq i$, $\underline{\Delta}_{\delta^*}(\boldsymbol{X}, p^*, j') \neq i$ and thus $\overline{\Delta}(\boldsymbol{X}, p^*, j') \neq i$, $\underline{\Delta}(\boldsymbol{X}, p^*, j') \neq i$. □

(iv) According to (ii) G_i is weakly $\boldsymbol{Y}^{\overline{(j)}}$ enclosed. By Theorem 5.2.6 $\mathcal{Z}(G_i)$ intersects $\boldsymbol{Y}^{\overline{(j')}}$, $\boldsymbol{Y}^{\underline{(j')}}$ for all $j \neq j'$. Hence, by Corollary 5.4.3 $\boldsymbol{Y} \overset{G_i, j'}{\longrightarrow} \boldsymbol{Y}$ for all $j' \neq j$. □

In the following let $\mathfrak{A}$ be the coefficient matrix of $\boldsymbol{G}$.

Theorem 5.4.23 (Orthogonality, Regularity and Non–Existence) *Assume* $\boldsymbol{X} \overset{p^*}{\longrightarrow} \boldsymbol{Y} \subseteq \mathrm{int}(\boldsymbol{X})$ *and* $\boldsymbol{Y} \neq \emptyset$.

(i) *Assume* $\overline{\Delta}(\boldsymbol{X}, p^*, j) = \underline{\Delta}(\boldsymbol{X}, p^*, j)$ *for all* j. *Then for all* $\delta > 0$ *there exists* $0 < \delta^* \leq \delta$ *such that* $\boldsymbol{G}$ *is strongly orthogonal in* $\tilde{\boldsymbol{Y}}$ *where* $\boldsymbol{X} \overset{p^*}{\longrightarrow}_{\delta^*} \tilde{\boldsymbol{Y}}$.

(ii) *Assume* $\overline{\Delta}(\boldsymbol{X}, p^*, j) = \underline{\Delta}(\boldsymbol{X}, p^*, j)$ *for all* j. *Then* $\mathfrak{A}$ *is regular and* $\boldsymbol{G}$ *is weakly orthogonal in* $\boldsymbol{Y}$.

(iii) *Assume* $\overline{\Delta}(\boldsymbol{X}, p^*, j) \neq \underline{\Delta}(\boldsymbol{X}, p^*, j)$ *for some* j. *Then* $0 \notin \boldsymbol{G}(\boldsymbol{x})$ *for all* $\boldsymbol{x} \in \boldsymbol{X}$. □

Proof. Assume $\boldsymbol{X} \overset{p^*}{\longrightarrow} \boldsymbol{Y} \subseteq \mathrm{int}(\boldsymbol{X})$ and $\boldsymbol{Y} \neq \emptyset$. From Theorem 5.4.12 it follows that

$$\overline{\Delta}(\boldsymbol{X}, p^*, j) \neq \uparrow, \quad \underline{\Delta}(\boldsymbol{X}, p^*, j) \neq \uparrow \quad \text{for all} \quad j.$$

Let $\delta > 0$ arbitrary but fixed and let $0 < \delta^* \leq \delta$ such that

$$\overline{\Delta}_{\delta^*}(\boldsymbol{X}, p^*, j) = \overline{\Delta}(\boldsymbol{X}, p^*, j), \quad \underline{\Delta}_{\delta^*}(\boldsymbol{X}, p^*, j) = \underline{\Delta}(\boldsymbol{X}, p^*, j)$$

and

$$\boldsymbol{Y} \subseteq \tilde{\boldsymbol{Y}} \subseteq \mathrm{int}(\boldsymbol{X})$$

where

$$\boldsymbol{X} \overset{p^*}{\longrightarrow}_{\delta^*} \tilde{\boldsymbol{Y}}.$$

(i) Assume $\overline{\Delta}(\boldsymbol{X}, p^*, j) = \underline{\Delta}(\boldsymbol{X}, p^*, j)$ for all j. From Theorem 5.4.22 (i) it follows that for all j there exists i such that G_i is strongly $\tilde{\boldsymbol{Y}}^{\overline{(j)}}$ enclosed. According to Theorem 5.4.22 (iii) no G_i is strongly $\tilde{\boldsymbol{Y}}^{\overline{(j')}}$ and $\tilde{\boldsymbol{Y}}^{\overline{(j)}}$ enclosed for $j' \neq j$. Hence, $\boldsymbol{G}$ is strongly orthogonal in $\tilde{\boldsymbol{Y}}$.

(ii) From (i) and Theorem 5.2.8 it follows that $\mathfrak{A}$ is regular. Weak orthogonality of $\boldsymbol{G}$ in $\boldsymbol{Y}$ follows from Theorem 5.4.22, (ii) and (iii).

(iii) Assume $\overline{\Delta}(\boldsymbol{X}, p^*, j) = i' \neq i'' = \underline{\Delta}(\boldsymbol{X}, p^*, j)$ for some j. First, we show that $\boldsymbol{G}$ is not strongly orthogonal in $\tilde{\boldsymbol{Y}}$. Assume to the contrary that $\boldsymbol{G}$ is strongly orthogonal in $\tilde{\boldsymbol{Y}}$. Then there exists a permutation π such that G_i is strongly $\tilde{\boldsymbol{Y}}^{\overline{(\pi(i))}}$ enclosed for all i. According to Corollary 5.2.6, $\mathcal{Z}(G_i)$ intersects $\tilde{\boldsymbol{Y}}^{\overline{(j)}}$ and $\tilde{\boldsymbol{Y}}^{\underline{(j)}}$ for all $i \neq \pi^{-1}(j)$. From Theorem 5.4.13 it follows that $\mathcal{Z}(G_{i'})$ is disjoint from $\tilde{\boldsymbol{Y}}^{\overline{(j)}}$ and $\mathcal{Z}(G_{i''})$ is disjoint from $\tilde{\boldsymbol{Y}}^{\underline{(j)}}$, hence $i' = i'' = \pi^{-1}(j)$ which is a contradiction to the assumption.

As $\overline{\Delta}_{\delta^*}$ and $\underline{\Delta}_{\delta^*}$ are total functions, it follows from Theorem 5.4.13 that $\boldsymbol{G}$ is face disjoint from $\tilde{\boldsymbol{Y}}$. Applying Lemma 5.3.10 we get $0 \notin \boldsymbol{G}(\boldsymbol{x})$ for all $\boldsymbol{x} \in \tilde{\boldsymbol{Y}}$ and according to Theorem 5.4.2 it holds that $0 \notin \boldsymbol{G}(\boldsymbol{x})$ for all $\boldsymbol{x} \in \boldsymbol{X}$. □

Theorem 5.4.24 (Non–Existence) *If* $\boldsymbol{X} \overset{p^*}{\longrightarrow} \emptyset$ *then* $0 \notin \boldsymbol{G}(\boldsymbol{x})$ *for all* $\boldsymbol{x} \in \boldsymbol{X}$. □

Proof. Follows from Theorem 5.4.2. □

Remark. It is not obvious whether there are cases where Theorem 5.4.23 (iii) is algorithmically advantageous over Theorem 5.4.24 for proving non–existence of solutions. Figure 5.4.2 displays such a situation for the two–dimensional case. In order to apply Theorem 5.4.24 more tightening steps are necessary as compared to Theorem 5.4.23. □

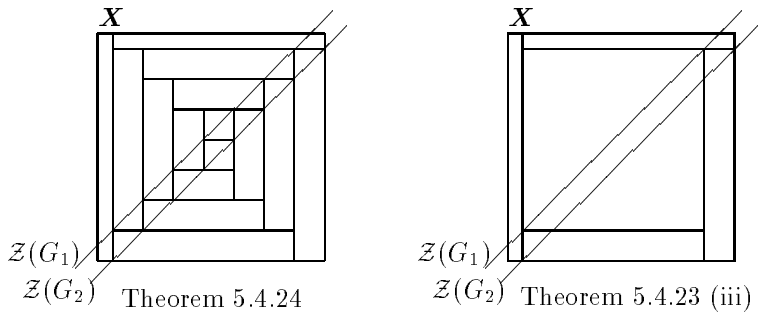


Figure 5.4.2: In order to apply Theorem 5.4.24 for proving non–existence, tightening steps have to be applied until an empty box is obtained. For Theorem 5.4.23 (iii) the tightening sequence can be stopped as soon as a box is obtained which is in the interior of the original box.

5.4.5 The Unique Existence Condition of Linearized Tightening is Too Strong

Theorem 5.4.23 can be used for testing whether a system of nonlinear equations $\boldsymbol{f}$ has a unique solution in $\boldsymbol{X}$ in the following straight forward way: Choose a linearization $\boldsymbol{G}$ of $\boldsymbol{f}$ in $\boldsymbol{X}$, where the coefficient matrix of $\boldsymbol{G}$ is a Jacobian of $\boldsymbol{f}$ in $\boldsymbol{X}$. Choose a path p^* consisting of pairs (G_i, j) and check whether the conditions $\boldsymbol{X} \xrightarrow{p^*} \boldsymbol{X}' \subseteq \mathrm{int}(\boldsymbol{X})$, $\boldsymbol{X}' \neq \emptyset$ and $\overline{\Delta}(\boldsymbol{X}, p^*, j) = \underline{\Delta}(\boldsymbol{X}, p^*, j)$ for $j = 1, \ldots, n$ of Theorem 5.4.23 are satisfied. Unfortunately, even if $\boldsymbol{f}$ has a unique solution in $\boldsymbol{X}$, these conditions will usually not be satisfied. The reason is as follows. Assume the conditions are satisfied. Then there exists $\delta^* > 0$ such that $\boldsymbol{X} \xrightarrow{p^*}_{\delta^*} \boldsymbol{Y}$ and $\boldsymbol{G}$ is strongly orthogonal in $\boldsymbol{Y}$. Hence, there exists a permutation π such that $G_{\pi(i)}$ is strongly $\boldsymbol{Y}^{\overline{(i)}}$ enclosed. According to Lemma 5.2.7 this means that

$$\mathrm{mig}(A_{\pi(i)})\mathrm{w}(Y_i) > \sum_{j \neq i} \mathrm{mag}(A_{\pi(j)})\mathrm{w}(Y_j) \tag{5.4.9}$$

for all i. Let

$$\mathfrak{a} = \boldsymbol{f}'(\boldsymbol{x})$$

where $\boldsymbol{x}$ is the unique solution of $\boldsymbol{f}$ in $\boldsymbol{X}$. It follows that

$$|a_{\pi(i)}|\mathrm{w}(Y_i) > \sum_{j \neq i} |a_{\pi(j)}|\mathrm{w}(Y_j)$$

for all i. This means that the matrix

$$\begin{pmatrix} a_{1,1}\mathrm{w}(Y_1) & a_{1,2}\mathrm{w}(Y_2) & \ldots & a_{1,n}\mathrm{w}(Y_n) \\ a_{2,1}\mathrm{w}(Y_1) & a_{2,2}\mathrm{w}(Y_2) & \ldots & a_{2,n}\mathrm{w}(Y_n) \\ \vdots & \vdots & & \vdots \\ a_{n,1}\mathrm{w}(Y_1) & a_{n,2}\mathrm{w}(Y_2) & \ldots & a_{n,n}\mathrm{w}(Y_n) \end{pmatrix}$$

is strictly diagonally dominant after some row permutation. Hence, a necessary condition for the assumptions of Theorem 5.4.23 is that there exists a permutation matrix $\mathfrak{p}$ and a diagonal matrix $\mathfrak{d}$ such that $\mathfrak{p}\mathfrak{a}\mathfrak{d}$ is strictly diagonally dominant. In general such matrices $\mathfrak{p}$, $\mathfrak{d}$ do not exist if $n > 2$. As an example consider the matrix

$$\mathfrak{a} = \begin{pmatrix} 3 & 2 & 2 \\ 2 & 3 & 2 \\ 2 & 2 & 3 \end{pmatrix}.$$

5.5 Linearized Tightening Operator

In this section we present the linearized tightening operator, which defines a strategy for choosing a path p^* and a linearization $\boldsymbol{G}$ of $\boldsymbol{f}$ and returns $\boldsymbol{X} \xrightarrow{p^*}$. The main objective of this strategy is that $\boldsymbol{X} \xrightarrow{p^*}$ can

be computed efficiently and that each G_i is tightened once in every direction j in p^*. As pointed out in Section 5.4.5, the linearized tightening operator will usually not detect whether $\boldsymbol{f}$ has a unique solution in $\boldsymbol{X}$. However, as will be shown in Section 5.11, it can give significant speedups in combination with a more powerful method.

Informally, the linearized tightening operator can be described as follows: Let ${}^0\boldsymbol{X} = \boldsymbol{X}$. Find a linearization G_1 of f_1 in ${}^0\boldsymbol{X}$. Tighten G_1 in every coordinate direction with starting box ${}^0\boldsymbol{X}$ obtaining ${}^1\boldsymbol{X}$. Find a linearization G_2 of f_2 in ${}^1\boldsymbol{X}$. Tighten G_2 in every coordinate direction with starting box ${}^1\boldsymbol{X}$ obtaining ${}^2\boldsymbol{X}$, etc. Finally, return ${}^n\boldsymbol{X}$. Note that G_i is a linearization of f_i in ${}^{i-1}\boldsymbol{X}$, but not necessarily in $\boldsymbol{X}$. This results usually in tighter inclusions as if we would require that $\boldsymbol{G}$ is a linearization of $\boldsymbol{f}$ in $\boldsymbol{X}$. The following algorithm describes this process more formally.

Algorithm 5.5.1 [Linearized Tightening Operator]

In: $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^n$,
$\boldsymbol{X} \in \mathbb{IR}^n$.

Out: $\boldsymbol{Y} \subseteq \boldsymbol{X}$ such that $\boldsymbol{Y}$ contains all solutions of $\boldsymbol{f}$ in $\boldsymbol{X}$.

(1) [Initialize.]
${}^0\boldsymbol{X} \longleftarrow \boldsymbol{X}$.

(2) [Iterate over equations.]
for $i = 1, \ldots, n$

(2.1) [Linearize.]
Choose $\boldsymbol{c}_i \in \mathbb{R}^n$ and $\boldsymbol{A}_i \in \mathbb{IR}^n$ such that $G_i(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{c}_i) + \boldsymbol{A}_i(\boldsymbol{x} - \boldsymbol{c}_i)$ is a linearization of f_i in ${}^{i-1}\boldsymbol{X}$.

(2.2) [Path.]
${}^ip^* \longleftarrow \langle (G_i, 1), (G_i, 2), \ldots, (G_i, n) \rangle$.

(2.3) [Tighten.]
Let ${}^i\boldsymbol{X}$ such that ${}^{i-1}\boldsymbol{X} \xrightarrow{{}^ip^*} {}^i\boldsymbol{X}$.

(3) [Return.]
$\boldsymbol{Y} \longleftarrow {}^n\boldsymbol{X}$.
return $\boldsymbol{Y}$.

Definition 5.5.2 (Linearized Tightening Operator) *Let $\boldsymbol{c}_i$, $\boldsymbol{A}_i$, $i = 1, \ldots, n$ and $\boldsymbol{Y}$ as in Algorithm 5.5.1. Let $\mathfrak{c} = (\boldsymbol{c}_1, \ldots, \boldsymbol{c}_n)^{\mathrm{T}}$ and $\mathfrak{A} = (\boldsymbol{A}_1, \ldots, \boldsymbol{A}_n)^{\mathrm{T}}$. Then the linearized tightening operator $\mathcal{LT}$ is defined as*

$$\mathcal{LT}(\boldsymbol{f}, \boldsymbol{X}, \mathfrak{c}, \mathfrak{A}) = \boldsymbol{Y}. \ \square$$

In the following let $\boldsymbol{c}_i$, $\boldsymbol{A}_i$, ${}^i\boldsymbol{X}$, ${}^ip^*$ and $\boldsymbol{Y}$ as in Algorithm 5.5.1.

Theorem 5.5.3 (Properties of Linearized Tightening Operator)

(i) *If $\boldsymbol{f}(\boldsymbol{x}) = 0$ for some $\boldsymbol{x} \in \boldsymbol{X}$ then $\boldsymbol{x} \in \mathcal{LT}(\boldsymbol{f}, \boldsymbol{X}, \mathfrak{c}, \mathfrak{A})$.*

(ii) *If ${}^i\boldsymbol{X} = \emptyset$ for some $0 \leq i \leq n$ then $\boldsymbol{f}$ has no solution in $\boldsymbol{X}$.*

(iii) *If ${}^i\boldsymbol{X} \subseteq \mathrm{int}(\boldsymbol{X})$ for some $i < n$ then $\boldsymbol{f}$ has no solution in $\boldsymbol{X}$.*

(iv) *If $\boldsymbol{Y} \subseteq \mathrm{int}(\boldsymbol{X})$ and there exists $1 \leq i \leq n$ such that for all j it holds that ${}^iX_j \not\subseteq \mathrm{int}({}^{i-1}X_j)$ then $\boldsymbol{f}$ has no solution in $\boldsymbol{X}$.*

(v) *If $\emptyset \neq \boldsymbol{Y} \subseteq \mathrm{int}(\boldsymbol{X})$ and for all i there exists j such that ${}^iX_j \subseteq \mathrm{int}({}^{i-1}X_j)$ then $\boldsymbol{f}$ has a solution in $\boldsymbol{X}$.*

(vi) *If* $\emptyset \neq \boldsymbol{Y} \subseteq \mathrm{int}(\boldsymbol{X})$ *and for all* i *there exists* j *such that* ${}^{i}X_j \subseteq \mathrm{int}({}^{i-1}X_j)$ *and* $\mathfrak{A}$ *is a Jacobian of* $\boldsymbol{f}$ *in* $\boldsymbol{Y}$*, then* $\boldsymbol{f}$ *has a unique solution in* $\boldsymbol{X}$. □

Proof.

(i) Follows from the solution preserving property of linear tightening (Theorem 5.4.2).

(ii) Follows from (i).

(iii) Assume ${}^{i}\boldsymbol{X} \subseteq \mathrm{int}(\boldsymbol{X})$ for some $i < n$. Let $p^* = {}^{1}p^* \circ \ldots \circ {}^{i}p^*$. From Lemma 5.4.12 it follows that

$$\overline{\Delta}(\boldsymbol{X}, p^*, j) \neq \uparrow, \quad \underline{\Delta}(\boldsymbol{X}, p^*, j) \neq \uparrow \quad \text{for all} \quad j.$$

As $i < n$ it follows from Theorem 5.4.22 (iii) that

$$\overline{\Delta}(\boldsymbol{X}, p^*, j) \neq \underline{\Delta}(\boldsymbol{X}, p^*, j) \quad \text{for some} \quad j.$$

According to Theorem 5.4.23 (iii) it holds that $0 \notin \boldsymbol{G}(\boldsymbol{x})$ for all $\boldsymbol{x} \in \boldsymbol{X}$ and hence for all $\boldsymbol{x} \in \boldsymbol{Y}$. As $\boldsymbol{G}$ is a linearization of $\boldsymbol{f}$ in $\boldsymbol{Y}$ it follows that $\boldsymbol{f}(\boldsymbol{x}) \neq 0$ for all $\boldsymbol{x} \in \boldsymbol{Y}$ and by (i), $\boldsymbol{f}(\boldsymbol{x}) \neq 0$ for all $\boldsymbol{x} \in \boldsymbol{X}$.

(iv) Assume $\boldsymbol{Y} \subseteq \mathrm{int}(\boldsymbol{X})$ and there exists i such that

$${}^{i}X_j \not\subseteq \mathrm{int}({}^{i-1}X_j) \quad \text{for all} \quad j.$$

Then there exists no j such that $\overline{\Delta}(\boldsymbol{X}, p^*, j) = \underline{\Delta}(\boldsymbol{X}, p^*, j) = i$, where $p^* = {}^{1}p^* \circ \ldots \circ {}^{n}p^*$. From Theorem 5.4.22 (iii) it follows that

$$\overline{\Delta}(\boldsymbol{X}, p^*, j') \neq \underline{\Delta}(\boldsymbol{X}, p^*, j') \quad \text{for some} \quad j'.$$

The rest of the proof is the same as for (iii).

(v) Assume $\emptyset \neq \boldsymbol{Y} \subseteq \mathrm{int}(\boldsymbol{X})$ and for all i there exists j_i such that ${}^{i}X_{j_i} \subseteq \mathrm{int}({}^{i-1}X_{j_i})$. Then G_i is strongly ${}^{i-1}\boldsymbol{X}^{\overline{(j_i)}}$ enclosed by Corollary 5.4.4. Hence, by Corollary 5.2.6 ${}^{i}X_j = {}^{i-1}X_j$ for all $j \neq j_i$. As $\boldsymbol{Y} \subseteq \mathrm{int}(\boldsymbol{X})$ it follows that $i \neq i'$ implies $j_i \neq j_{i'}$. Hence,

$$\overline{\Delta}(\boldsymbol{X}, p^*, j) = \underline{\Delta}(\boldsymbol{X}, p^*, j) \quad \text{for all} \quad j$$

where $p^* = {}^{1}p^* \circ \ldots \circ {}^{n}p^*$. Thus, $\boldsymbol{G}$ is weakly orthogonal in $\boldsymbol{Y}$ by Theorem 5.4.23 (ii). As $\boldsymbol{G}$ is a linearization of $\boldsymbol{f}$ in $\boldsymbol{Y}$, it follows from Theorem 5.1.12 that $\boldsymbol{f}$ has a solution in $\boldsymbol{Y}$ and hence in $\boldsymbol{X}$.

(vi) Assume $\emptyset \neq \boldsymbol{Y} \subseteq \mathrm{int}(\boldsymbol{X})$ and for all i there exists j such that ${}^{i}X_j \subseteq \mathrm{int}({}^{i-1}X_j)$ and $\mathfrak{A}$ is a Jacobian of $\boldsymbol{f}$ in $\boldsymbol{Y}$. According to (v) $\boldsymbol{f}$ has a solution in $\boldsymbol{X}$. From Theorem 5.4.23 (ii) it follows that $\mathfrak{A}$ is regular. As $\mathfrak{A}$ is a Jacobian of $\boldsymbol{f}$ in $\boldsymbol{Y}$, it follows from Theorem 5.2.3 that $\boldsymbol{f}$ has at most one solution in $\boldsymbol{Y}$. Hence, by (i) there exists a unique solution of $\boldsymbol{f}$ in $\boldsymbol{X}$. □

Below, we give an optimized algorithm for evaluating the linearized tightening operator. During computation of the reduction sequence

$${}^{i-1}\boldsymbol{X} \overset{G_i,1}{\longrightarrow} \ldots \overset{G_i,n}{\longrightarrow} {}^{i}\boldsymbol{X}$$

some intermediate results can be precomputed and reused. This is possible because the order of the tightening steps does not matter as long as G_i is fixed, see Theorem 5.4.9. Further, Theorem 5.4.22 (iv) is used in Step 2.5 to cancel a tightening sequence ${}^{i}p^*$ as soon as it is clear that no further reduction will be achieved. We assume that $\mathfrak{A}$ is sparse, i.e. each row of $\mathfrak{A}$ has at most $n' \leq n$ non–zero elements.

Algorithm 5.5.4 (LTO) [Linearized Tightening Operator]

In: $\boldsymbol{f} \in \mathbb{R}^n \to \mathbb{R}^n$,
$\boldsymbol{X} \in \mathbb{IF}^n$.

Out: $\boldsymbol{X}' \supseteq \mathcal{LT}(\boldsymbol{f}, \boldsymbol{X}, \mathfrak{c}, \mathfrak{A})$, for some suitably chosen $\mathfrak{c}$, $\mathfrak{A}$,
$exist \in \{\texttt{true}, \texttt{false}\}$ such that if $exist = \texttt{true}$ then $\mathfrak{A}$ is regular and $\boldsymbol{f}$ has a solution in $\boldsymbol{X}$,
$nonexist \in \{\texttt{true}, \texttt{false}\}$ such that if $nonexist = \texttt{true}$ then $\boldsymbol{f}$ has no solution in $\boldsymbol{X}$.

(1) [Initialize.]
${}^0\boldsymbol{X} \longleftarrow \boldsymbol{X}$, $exist \longleftarrow \texttt{false}$, $nonexist \longleftarrow \texttt{false}$, $orthogonal \longleftarrow \texttt{true}$.

(2) [Iterate over equations.]
for $i = 1, \ldots, n$

(2.1) [Linearize.]
Choose $\boldsymbol{c}_i \in \mathbb{F}^n$ and $\boldsymbol{A}_i \in \mathbb{IF}^n$ such that $G_i(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{c}_i) + \boldsymbol{A}_i(\boldsymbol{x} - \boldsymbol{c}_i)$ is a linearization of f_i in ${}^{i-1}\boldsymbol{X}$.

(2.2) [Translate.]
for $j = 1, \ldots, n$ where $A_{ij} \neq 0$ do $Y_j \longleftarrow {}^{i-1}X_j - c_{ij}$.

(2.3) [Precompute products.]
for $j = 1, \ldots, n$ where $A_{ij} \neq 0$ do $P_j \longleftarrow A_{ij} Y_j$.

(2.4) [Accumulate sums right to left.]
$S_{n+1} \longleftarrow f_i(\boldsymbol{c}_i)$ (overestimate).
for $j = n, \ldots, 2$ where $A_{ij} \neq 0$ do $S_{j-1} \longleftarrow S_j + P_j$.

(2.5) [Tighten along ${}^ip^*$.]
$R \longleftarrow 0$.
for $j = 1, \ldots, n$ where $A_{ij} \neq 0$
$\tilde{Y}_j \longleftarrow \mathrm{HDIV}(-R - S_{j+1}, A_{ij}, Y_j)$.
if $\tilde{Y}_j \neq Y_j$
if $\tilde{Y}_j = \emptyset$ then return $(\emptyset, \texttt{false}, \texttt{true})$.
if $\tilde{Y}_j \subseteq \mathrm{int}(Y_j)$ then $Y_j \longleftarrow \tilde{Y}_j$, goto Step (2.6).
$Y_j \longleftarrow \tilde{Y}_j$, $orthogonal \longleftarrow \texttt{false}$.
$R \longleftarrow R + P_j$.

(2.6) [Translate Back.]
for $j = 1, \ldots, n$ where $A_{ij} \neq 0$
if $A_{ij} \neq 0$ then ${}^iX_j \longleftarrow Y_j + c_{ij}$, else ${}^iX_j \longleftarrow {}^{i-1}X_j$.

(3) [Test if in interior.]
if ${}^n\boldsymbol{X} \subseteq \mathrm{int}(\boldsymbol{X})$
if $orthogonal = \texttt{true}$ then $exist \longleftarrow true$, else $nonexist \longleftarrow \texttt{true}$.

(4) [Return.]
return $({}^n\boldsymbol{X}, exist, nonexist)$.

Theorem 5.5.5 (Complexity) *The costs of Algorithm 5.5.4 (*LTO*) except for Step 2.1 and the overestimation of $f_i(\boldsymbol{c}_i)$ in Step 2.4 are*

$$\begin{array}{rl} n'n & \textit{interval multiplication,} \\ 2n'n & \textit{number divisions,} \\ 10n'n & \textit{number additions.} \ \square \end{array}$$

Proof.

- Step 2.2 costs $2n'$ number additions.
- Step 2.3 costs n' interval multiplications.
- Step 2.4 costs $2n'$ number additions.
- The j-th iteration in Step 2.5 costs 2 number divisions and 4 number additions. Hence, Step 2.5 costs $2n'$ number divisions and $4n'$ number additions.

- Step 2.6 costs $2n'$ number additions.

As Step 2 is iterated n times, Theorem 5.5.5 follows. □

Remark. If $\boldsymbol{c}_i$ is chosen to be the midpoint of ${}^{i-1}\boldsymbol{X}$ in Step 2.2, then all interval multiplications can be turned into number multiplications using the formula

$$A(X - \mathrm{mid}(X)) = \mathrm{mag}(A)\mathrm{rad}(X)[-1, 1].$$

5.6 Hansen–Sengupta Operator

As pointed out in Section 5.4.5, the unique existence condition of linearized tightening (Theorem 5.4.23) is too strong. Unless the Jacobian of $\boldsymbol{f}$ at some solution $\boldsymbol{x}$ has certain diagonal dominance properties, then for all $\boldsymbol{X}$ containing $\boldsymbol{x}$ and for all linearizations $\boldsymbol{G}$ of $\boldsymbol{f}$ in $\boldsymbol{X}$ the condition of Theorem 5.4.23 is not satisfied. In this section we present the Hansen–Sengupta operator [Hansen and Sengupta, 1981], which solves this problem. Instead of weakening the conditions of Theorem 5.4.23, we modify the given equations $\boldsymbol{f}$ by premultiplying a matrix $\mathfrak{m} \in \mathbb{R}^{n \times n}$. The matrix $\mathfrak{m}$ is called preconditioning matrix. If $\mathfrak{m}$ is regular, then $\boldsymbol{f}$ and $\mathfrak{m}\boldsymbol{f}$ have the same solution set. Now, we choose $\mathfrak{m}$ such that the interval Jacobian of $\mathfrak{m}\boldsymbol{f}$ in $\boldsymbol{X}$ is approximately the identity matrix. If this approximation is close enough, then the necessary condition (5.4.9) holds and the problem pointed out in Section 5.4.5 is solved.

In order to obtain a linearization of $\mathfrak{m}\boldsymbol{f}$ in $\boldsymbol{X}$, it is not necessary to multiply $\mathfrak{m}$ and $\boldsymbol{f}$, as the following lemma shows:

Lemma 5.6.1 (Linearization of Preconditioned System) *If $\boldsymbol{G} = \boldsymbol{f}(c) + \mathfrak{A}(\boldsymbol{x} - \boldsymbol{c})$ is a linearization of $\boldsymbol{f}$ in $\boldsymbol{X}$, then $\mathfrak{m}\boldsymbol{G}$ is a linearization of $\mathfrak{m}\boldsymbol{f}$ in $\boldsymbol{X}$. Further, $\mathfrak{m}\boldsymbol{G}(\boldsymbol{x}) = \mathfrak{m}\boldsymbol{f}(\boldsymbol{c}) + (\mathfrak{m}\mathfrak{A})(\boldsymbol{x} - \boldsymbol{c})$.* □

Proof. Assume $\boldsymbol{G}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{c}) + \mathfrak{A}(\boldsymbol{x} - \boldsymbol{c})$ is a linearization of $\boldsymbol{f}$ in $\boldsymbol{X}$ and let $i \in \{1, \ldots, n\}$ arbitrary but fixed. Then

$$\begin{aligned}
(\mathfrak{m}\boldsymbol{G}(\boldsymbol{x}))_i &= (\mathfrak{m}(\boldsymbol{f}(\boldsymbol{c}) + \mathfrak{A}(\boldsymbol{x} - \boldsymbol{c})))_i \\
&= \boldsymbol{m}_i(\boldsymbol{f}(\boldsymbol{c}) + \mathfrak{A}(\boldsymbol{x} - \boldsymbol{c})) \\
&= \sum_{j=1}^{n} m_{ij}(f_j(\boldsymbol{c}) + \boldsymbol{A}_j(\boldsymbol{x} - \boldsymbol{c})) \\
&= \sum_{j=1}^{n} m_{ij} f_j(\boldsymbol{c}) + \sum_{j=1}^{n} m_{ij} \boldsymbol{A}_j(\boldsymbol{x} - \boldsymbol{c}) \\
&= \sum_{j=1}^{n} m_{ij} f_j(\boldsymbol{c}) + \sum_{j=1}^{n} m_{ij} \sum_{k=1}^{n} A_{jk}(x_k - c_k) \\
&= \sum_{j=1}^{n} m_{ij} f_j(\boldsymbol{c}) + \sum_{j=1}^{n} \sum_{k=1}^{n} m_{ij} A_{jk}(x_k - c_k) \\
&= \sum_{j=1}^{n} m_{ij} f_j(\boldsymbol{c}) + \sum_{j=1}^{n} (m_{ij} \boldsymbol{A}_j)(\boldsymbol{x} - \boldsymbol{c}) \\
&= \boldsymbol{m}_i \boldsymbol{f}(\boldsymbol{c}) + (\boldsymbol{m}_i \mathfrak{A})(\boldsymbol{x} - \boldsymbol{c}) \\
&= (\mathfrak{m}\boldsymbol{f}(\boldsymbol{c}) + (\mathfrak{m}\mathfrak{A})(\boldsymbol{x} - \boldsymbol{c}))_i.
\end{aligned}$$

Hence $\mathfrak{m}\boldsymbol{G}$ is a linear interval function and $\mathfrak{m}\boldsymbol{f}(\boldsymbol{x}) \in \mathfrak{m}\boldsymbol{G}(\boldsymbol{x})$ for all $\boldsymbol{x} \in \boldsymbol{X}$ follows immediately from $\boldsymbol{f}(\boldsymbol{x}) \in \boldsymbol{G}(\boldsymbol{x})$. □

The choice of $\mathfrak{m}$ has been studied thoroughly in the literature. If $\mathfrak{m} = \mathfrak{a}^{-1}$ for some $\mathfrak{a} \in \mathfrak{A}$, then $\mathfrak{i} \in \mathfrak{m}\mathfrak{A}$, where $\mathfrak{i}$ denotes the identity matrix. In this case we may view $\mathfrak{m}\mathfrak{A}$ as an approximation of $\mathfrak{i}$. In [Xiaojun

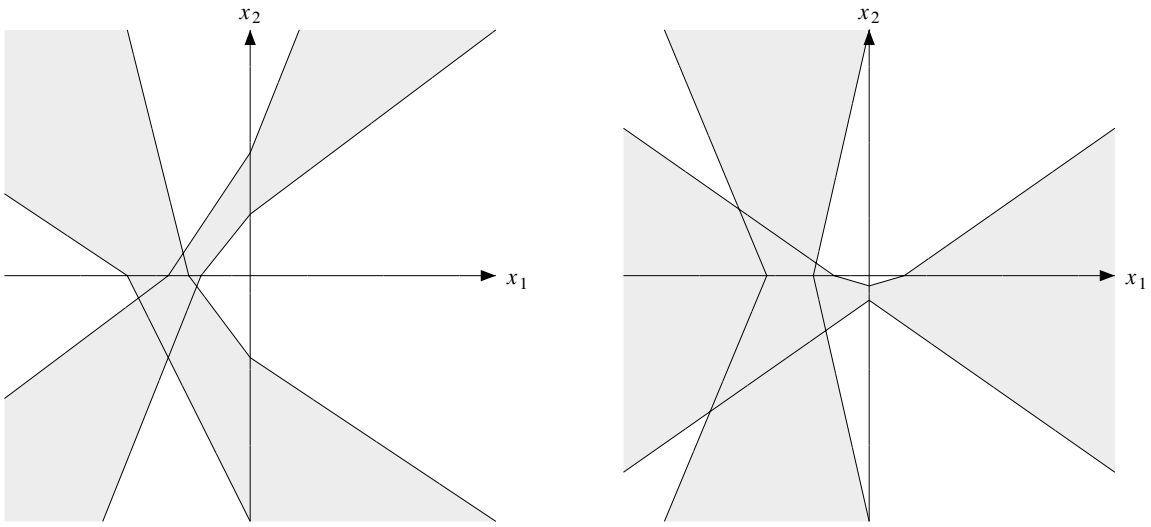


Figure 5.6.1: Preconditioning of the linearized system.

and Deren, 1987], [Neumaier, 1990] optimality properties of the choice

$$\mathfrak{m} = (\mathrm{mid}(\mathfrak{A}))^{-1}$$

are given. The effect of preconditioning is illustrated in Figure 5.6.1. The left graph shows a system of linear interval equations $\boldsymbol{G}$, the right graph shows $\mathfrak{m}\,\boldsymbol{G}$, where $\mathfrak{m} = (\mathrm{mid}(\mathfrak{A}))^{-1}$ and $\mathfrak{A}$ is the coefficient matrix of $\boldsymbol{G}$. The varieties of the equations in the right graph are approximately parallel to the coordinate axes, which reflects the fact that $\mathfrak{m}\,\mathfrak{A}$ contains the identity matrix. Note that $\mathcal{Z}(\boldsymbol{G}) \subseteq \mathcal{Z}(\mathfrak{m}\,\boldsymbol{G})$. It is easy to see that this is always the case.

In the following let $\boldsymbol{G}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{c}) + \mathfrak{A}(\boldsymbol{x} - \boldsymbol{c})$ be a linearization of $\boldsymbol{f}$ in $\boldsymbol{X}$ and let $\mathfrak{m} \in \mathbb{R}^{n\times n}$. We use the notation

$$\begin{array}{lcll} \mathfrak{m}\,\boldsymbol{f} & = & \boldsymbol{f}^{\mathfrak{m}} & = (f_1^{\mathfrak{m}}, f_2^{\mathfrak{m}}, \ldots, f_n^{\mathfrak{m}})^{\mathrm{T}} \\ \mathfrak{m}\,\mathfrak{A} & = & \mathfrak{A}^{\mathfrak{m}} & = (\boldsymbol{A}_1^{\mathfrak{m}}, \boldsymbol{A}_2^{\mathfrak{m}}, \ldots, \boldsymbol{A}_n^{\mathfrak{m}})^{\mathrm{T}} \\ \mathfrak{m}\,\boldsymbol{G} & = & \boldsymbol{G}^{\mathfrak{m}} & = (G_1^{\mathfrak{m}}, G_2^{\mathfrak{m}}, \ldots, G_n^{\mathfrak{m}})^{\mathrm{T}}. \end{array}$$

If $\mathfrak{i} \in \mathfrak{m}\,\mathfrak{A}$, then we know a priori that certain tightening steps will not give a simultaneous reduction on opposite faces.

Lemma 5.6.2 (Failing Tightening Steps) *Assume* $0 \in A_{ij}^{\mathfrak{m}}$ *and* $\boldsymbol{Y} = \mathrm{tight}(\boldsymbol{X}, G_i^{\mathfrak{m}}, j) \neq \emptyset$. *Then*

$$Y_j \notin \mathrm{int}(X_j). \ \square$$

Proof. Assume $0 \in A_{ij}^{\mathfrak{m}}$, $\boldsymbol{Y} = \mathrm{tight}(\boldsymbol{X}, G_i^{\mathfrak{m}}, j) \neq \emptyset$ and $Y_j \in \mathrm{int}(X_j)$. Then $G_i^{\mathfrak{m}}$ is strongly $\boldsymbol{X}^{\overline{(j)}}$ enclosed by Corollary 5.4.4 and thus

$$\mathrm{mig}(A_{ij}^{\mathfrak{m}})\mathrm{w}(X_j) > \sum_{k\neq j} \mathrm{mag}(A_{ik}^{\mathfrak{m}})\mathrm{w}(X_k)$$

by Lemma 5.2.7. But this is a contradiction because $\mathrm{mig}(A_{ij}) = 0$ as $\mathfrak{i} \in \mathfrak{m}\,\mathfrak{A}$ and $i \neq j$. $\square$

Lemma 5.6.2 can be verified at the right graph of Figure 5.6.1.

In order to apply Theorem 5.4.23 we are only interested in tightening steps $(G_i^{\mathfrak{m}}, j)$ which reduce the given box simultaneously on both j-faces. This justifies why the Hansen–Sengupta operator uses only tightening steps $(G_i^{\mathfrak{m}}, j)$ where $i = j$.

Definition 5.6.3 (Hansen–Sengupta operator) *Let* $\boldsymbol{G}(\boldsymbol{x}) = \boldsymbol{f}(c) + \mathfrak{A}(\boldsymbol{x} - \boldsymbol{c})$ *be a linearization of* $\boldsymbol{f}$ *in* $\boldsymbol{X}$ *and let* $\mathfrak{m} \in \mathbb{R}^{n\times n}$. *Then the Hansen–Sengupta operator* $\mathcal{HS}$ *is defined as*

$$\mathcal{HS}(\boldsymbol{f}, \boldsymbol{X}, \boldsymbol{c}, \mathfrak{A}, \mathfrak{m}) = \boldsymbol{Y}$$

where $p^* = \langle (G_1^{\mathfrak{m}}, 1), (G_2^{\mathfrak{m}}, 2), \ldots, (G_n^{\mathfrak{m}}, n) \rangle$ *and* $\boldsymbol{X} \xrightarrow{p^*} \boldsymbol{Y}$. □

Algorithm 5.6.4 gives the main steps in the computation of the Hansen–Sengupta operator.

Algorithm 5.6.4 [Hansen–Sengupta Operator]

In: $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^n$.
$\boldsymbol{X} \in \mathbb{IR}^n$.

Out: $\boldsymbol{Y} \subseteq \boldsymbol{X}$ such that $\boldsymbol{Y}$ contains all solutions of $\boldsymbol{f}$ in $\boldsymbol{X}$.

(1) [Initialize.]
${}^0\boldsymbol{X} \longleftarrow \boldsymbol{X}$.

(2) [Linearize.]
Choose $\boldsymbol{c} \in \mathbb{R}^n$, $\mathfrak{A} \in \mathbb{IR}^n$ such that $\boldsymbol{G}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{c}) + \mathfrak{A}(\boldsymbol{x} - \boldsymbol{c})$ is a linearization of $\boldsymbol{f}$ in $\boldsymbol{X}$.

(3) [Precondition.]
$\mathfrak{m} \longleftarrow \mathrm{mid}(\mathfrak{A})^{-1}$.
Compute $\mathfrak{m}\,\boldsymbol{f}(\boldsymbol{c})$ and $\mathfrak{m}\,\mathfrak{A}$ and let $\boldsymbol{G}^{\mathfrak{m}}(\boldsymbol{x}) = \mathfrak{m}\,\boldsymbol{f}(\boldsymbol{c}) + \mathfrak{m}\,\mathfrak{A}(\boldsymbol{x} - \boldsymbol{c})$.

(4) [Tighten.]
for $i = 1, \ldots, n$
${}^i\boldsymbol{X} \longleftarrow \mathrm{tight}({}^{i-1}\boldsymbol{X}, G_i^{\mathfrak{m}}, i)$.

(5) [Return.]
$\boldsymbol{Y} \longleftarrow {}^n\boldsymbol{X}$.
return $\boldsymbol{Y}$.

Comparing Algorithm 5.6.4 for the Hansen–Sengupta operator and Algorithm 5.5.1 for the linearized tightening operator one makes the following observations:

- The matrix $\mathfrak{A}$ is computed at the beginning of the Hansen–Sengupta algorithm, but is computed successively in the linearized tightening algorithm. Successive computation is preferable because it gives usually tighter inclusions. However, as already for the computation of $\mathfrak{m}$ and for each $\boldsymbol{A}_i^{\mathfrak{m}}$ the entire matrix $\mathfrak{A}$ is needed, this is not possible for the Hansen–Sengupta operator. Before the i-th tightening step of the Hansen–Sengupta operator one could recompute $\mathfrak{A}$ using ${}^i\boldsymbol{X}$, but this seems too costly.
- The linearization of $\boldsymbol{f}$ chosen for the Hansen–Sengupta operator is such that for every f_i the same $\boldsymbol{c}$ is used, whereas for the linearized tightening operator we allow different $\boldsymbol{c}_i$. We could have allowed the same generality for the Hansen–Sengupta operator but there is no necessity for it. As the computation of $f_i^{\mathfrak{m}}(\boldsymbol{c}_i)$ requires evaluation of $f_j(\boldsymbol{c}_i)$ for all j, the choice of different $\boldsymbol{c}_i$ would cost n times as many evaluations of $\boldsymbol{f}$ as needed if the same $\boldsymbol{c}$ is used for the linearization of all f_i.

The following theorem states the most important properties of the Hansen–Sengupta operator.

Theorem 5.6.5 (Properties of Hansen–Sengupta Operator) *Let* $\boldsymbol{c}$, $\mathfrak{A}$ *as in Definition 5.6.3, let* $\mathfrak{m} \in \mathbb{R}^{n \times n}$ *and let* $\boldsymbol{Y} = \mathcal{HS}(\boldsymbol{f}, \boldsymbol{X}, \boldsymbol{c}, \mathfrak{A}, \mathfrak{m})$.

(i) *If* $\boldsymbol{f}(\boldsymbol{x}) = 0$ *for some* $\boldsymbol{x} \in \boldsymbol{X}$ *then* $\boldsymbol{x} \in \mathcal{HS}(\boldsymbol{f}, \boldsymbol{X}, \boldsymbol{c}, \mathfrak{A}, \mathfrak{m})$.

(ii) *If* $\boldsymbol{Y} = \emptyset$ *then* $\boldsymbol{f}$ *has no solution in* $\boldsymbol{X}$.

(iii) *If* $\emptyset \neq \boldsymbol{Y} \subseteq \mathrm{int}(\boldsymbol{X})$ *then* $\boldsymbol{f}$ *has a solution in* $\boldsymbol{X}$.

(iv) *If* $\emptyset \neq \boldsymbol{Y} \subseteq \mathrm{int}(\boldsymbol{X})$ *and* $\mathfrak{A}$ *is a Jacobian of* $\boldsymbol{f}$ *in* $\boldsymbol{X}$, *then* $\boldsymbol{f}$ *has a unique solution in* $\boldsymbol{X}$. □

Proof. Let $\boldsymbol{c}$, $\mathfrak{A}$, $\boldsymbol{G}$, $\mathfrak{m}$, $\boldsymbol{Y}$ as in Theorem 5.6.5 and let p^* as in Definition 5.6.3.

(i) Follows from Theorem 5.4.2.

(ii) Follows from (ii).

(iii) Assume $\emptyset \neq \boldsymbol{Y} \subseteq \mathrm{int}(\boldsymbol{X})$. Then

$$\overline{\Delta}(\boldsymbol{X}, p^*, i) = \underline{\Delta}(\boldsymbol{X}, p^*, i) = i \quad \text{for all} \quad i.$$

From Theorem 5.4.23 (ii) it follows that $\boldsymbol{G}^{\mathfrak{m}}$ is weakly orthogonal in $\boldsymbol{Y}$ and $\boldsymbol{f}^{\mathfrak{m}}$ has a solution in $\boldsymbol{Y}$ and hence in $\boldsymbol{X}$ by Theorem 5.1.12. In order to prove that $\boldsymbol{f}$ has a solution in $\boldsymbol{X}$, we have to show that $\mathfrak{m}$ is regular. According to Theorem 5.4.23 (i) there exists $\delta^* > 0$ such that $\boldsymbol{G}^{\mathfrak{m}}$ is strongly orthogonal in $\boldsymbol{X} \xrightarrow{p^*}_{\delta^*}$. From Theorem 5.2.8 it follows that $\mathfrak{m}\mathfrak{A}$ is regular and therefore $\mathfrak{m}$ is regular.

(iv) Assume $\emptyset \neq \boldsymbol{Y} \subseteq \mathrm{int}(\boldsymbol{X})$ and $\mathfrak{A}$ is a Jacobian of $\boldsymbol{f}$ in $\boldsymbol{X}$. In the proof of (iii) we have already shown that $\mathfrak{A}$ is regular and $\boldsymbol{f}$ has a solution in $\boldsymbol{X}$. According to Theorem 5.2.3 this solution is unique. □

Below, we give an optimized algorithm for evaluating the Hansen–Sengupta operator. As in the case of the linearized tightening operator, we assume that $\mathfrak{A}$ is sparse, i.e. each row of $\mathfrak{A}$ has at most $n' < n$ non–zero elements. Note however that if $\mathfrak{m} = \mathrm{mid}(\mathfrak{A})^{-1}$, then $\mathfrak{m}$ and $\mathfrak{m}\mathfrak{A}$ are usually dense.

Algorithm 5.6.6 (HSO) [Hansen–Sengupta Operator]

In: $\boldsymbol{f} \in \mathbb{R}^n \to \mathbb{R}^n$,
$\boldsymbol{X} \in \mathbb{IF}^n$.

Out: $\boldsymbol{X}' \supseteq \mathcal{HS}(\boldsymbol{f}, \boldsymbol{X}, \boldsymbol{c}, \mathfrak{A}, \mathfrak{m})$, for some suitably chosen $\boldsymbol{c}$, $\mathfrak{A}$, $\mathfrak{m}$,
$exist \in \{\texttt{true}, \texttt{false}\}$ such that if $exist = \texttt{true}$ then $\mathfrak{A}$ is regular and $\boldsymbol{f}$ has a solution in $\boldsymbol{X}$,
$nonexist \in \{\texttt{true}, \texttt{false}\}$ such that if $nonexist = \texttt{true}$ then $\boldsymbol{f}$ has no solution in $\boldsymbol{X}$.

(1) [Initialize.]
$exist \longleftarrow \texttt{false}$, $nonexist \longleftarrow \texttt{false}$.

(2) [Linearize.]
Choose $\boldsymbol{c} \in \mathbb{F}^n$, $\mathfrak{A} \in \mathbb{IF}^{n \times n}$ such that $\boldsymbol{G}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{c}) + \mathfrak{A}(\boldsymbol{x} - \boldsymbol{c})$ is a linearization of $\boldsymbol{f}$ in $\boldsymbol{X}$.

(3) [Precondition.]
$\mathfrak{m} \longleftarrow \mathrm{mid}(\mathfrak{A})^{-1}$ (approximate, perturb if singular).
$\mathfrak{A}^{\mathfrak{m}} \longleftarrow \mathfrak{m}\mathfrak{A}$.
$\boldsymbol{B}^{\mathfrak{m}} \longleftarrow \mathfrak{m}\boldsymbol{f}(\boldsymbol{c})$ (use interval arithmetic).

(4) [Translate.]
$\boldsymbol{Y} \longleftarrow \boldsymbol{X} - \boldsymbol{c}$.

(5) [Tighten.]
for $i = 1, \ldots, n$
$R \longleftarrow \sum_{j \neq i} A^{\mathfrak{m}}_{ij} Y_j + B^{\mathfrak{m}}_i$.
$Y_i \longleftarrow \mathrm{HDIV}(-R, A^{\mathfrak{m}}_{ii}, Y_i)$.
if $Y_i = \emptyset$ then return $(\emptyset, \texttt{false}, \texttt{true})$.

(6) [Translate Back.]
$\boldsymbol{X}' \longleftarrow \boldsymbol{Y} + \boldsymbol{c}$.

(7) [Test if in Interior.]
if $\boldsymbol{X}' \subseteq \mathrm{int}(\boldsymbol{X})$ then $exist \longleftarrow \texttt{true}$, else $exist \longleftarrow \texttt{false}$.

(8) [Return.]
return $(\boldsymbol{X}', exist, \texttt{false})$.

Theorem 5.6.7 (Complexity) *The costs of Algorithm 5.6.6* (HSO) *except for Step 2, the overestima-*

tion of $\boldsymbol{f}(\boldsymbol{c})$ *and the approximation of* $\mathrm{mid}(\mathfrak{A})^{-1}$ *in Step 3 are*

$$\begin{array}{rl} n'n^2 + 2n^2 - n & \textit{interval multiplications,} \\ 2n & \textit{number divisions,} \\ 2n'n^2 - 2n'n + 4n^2 & \textit{number additions.} \ \square \end{array}$$

Proof.

- The computation of $\mathfrak{A}^{\mathfrak{m}}$ in Step 3 costs $n'n^2$ interval multiplications and $2n'n^2 - 2n'n$ number additions. The computation of $\boldsymbol{B}^{\mathfrak{m}}$ costs n^2 interval multiplications, and $2n^2 - 2n$ number additions.
- Step 4 costs $2n$ number additions.
- The i-th iteration in Step 5 costs $n-1$ interval multiplications, 2 number divisions, and $2n-2$ number additions. Hence, Step 5 costs $n^2 - n$ interval multiplications, $2n$ number divisions, and $2n^2 - 2n$ number additions.
- Step 6 costs $2n$ number additions. $\square$

Note that the linearized tightening operator is significantly cheaper than the Hansen–Sengupta operator, even if the costs for the inversion of $\mathfrak{m}\mathfrak{A}$ are ignored.

5.7 Iterated Linear Tightening

If the Hansen–Sengupta operator is applied to a box $\boldsymbol{X}$, it is often the case that the resulting box $\boldsymbol{Y}$ is a proper subset of $\boldsymbol{X}$ but not in the interior of $\boldsymbol{X}$. This means that some reduction was achieved but we do not know whether $\boldsymbol{f}$ has a (unique) solution in $\boldsymbol{X}$. Hence, one would apply the Hansen–Sengupta operator again on $\boldsymbol{Y}$ and so on. Let us assume after several iterations we obtain a box $\emptyset \neq \boldsymbol{Z} \subseteq \mathrm{int}(\boldsymbol{X})$. Can we conclude from this that $\boldsymbol{f}$ has a solution in $\boldsymbol{X}$ and that this solution is unique if in each iteration the linearization had a Jacobian coefficient matrix? In both cases the answer is no, counterexamples exist already for $n = 1$. However, for the existence property the examples show why this is not the case and how the problem can be solved. There seems to be no useful way for preserving the uniqueness property though. As in the previous sections we give more general proofs by using arbitrary linear tightening sequences. The results in this section are new and can be applied directly to the Hansen–Sengupta operator and to the linearized tightening operator.

We begin by giving counterexamples for $n = 1$, where the Hansen–Sengupta operator is applied twice and the resulting box is in the interior of the starting box but in one case there is no solution in the box and in the other case there are three solutions. Note that these are also counterexamples for the linearized tightening operator, applied to the preconditioned system.

Counterexample for Existence. Let $f(x) \equiv 1$ and let $X = [0, 2]$. Then $A = [-2, 4]$ is a (widely overestimated) Jacobian of f in X. For $c^{(1)} = 0.2$ we obtain a linearization $G^{(1)}(x) = 1 + [-2, 4](x - 0.2)$ of f in X. The inverse midpoint preconditioner is in this case $m = 1$ and we obtain $mA = [-2, 4]$, $mf(c^{(1)}) = 1$ and thus $Y = \mathcal{HS}(f, X, c^{(1)}, A, m) = [0.7, 2]$. We apply the Hansen–Sengupta operator again with the same A and m but this time we choose $c^{(2)} = 1.8$. We obtain $Z = \mathcal{HS}(f, Y, c^{(2)}, A, m) = [0.7, 1.55]$. Now, $\emptyset \neq Z \subseteq \mathrm{int}(X)$ but obviously f has no solution in X. This example is shown in Figure 5.7.1, left graph.

Counterexample for Uniqueness. Let $f(x) = x^3 - x$ and let $X = [-1.5, 1.5]$. Then $A = [-2, 6]$ is a Jacobian of f in X. For $c^{(1)} = -1.4$ we obtain a linearization $G^{(1)}(x) = -1.344 + [-2, 6](x + 1.4)$ of f in X. The inverse midpoint preconditioner is in this case $m = 0.5$ and we obtain $mA = [-1, 3]$, $mf(c) = -0.672$ and thus $Y = \mathcal{HS}(f, X, c^{(1)}, A, m) = [-1.176, 1.5]$. We apply the Hansen–Sengupta operator again with the same A and m but this time we choose $c^{(2)} = 1.4$. We obtain $Z = \mathcal{HS}(f, Y, c^{(2)}, A, m) = [-0.176, 1.176]$. Now, $\emptyset \neq Z \subseteq \mathrm{int}(X)$ but f has 3 solutions in X. This example is shown in Figure 5.7.1, right graph.

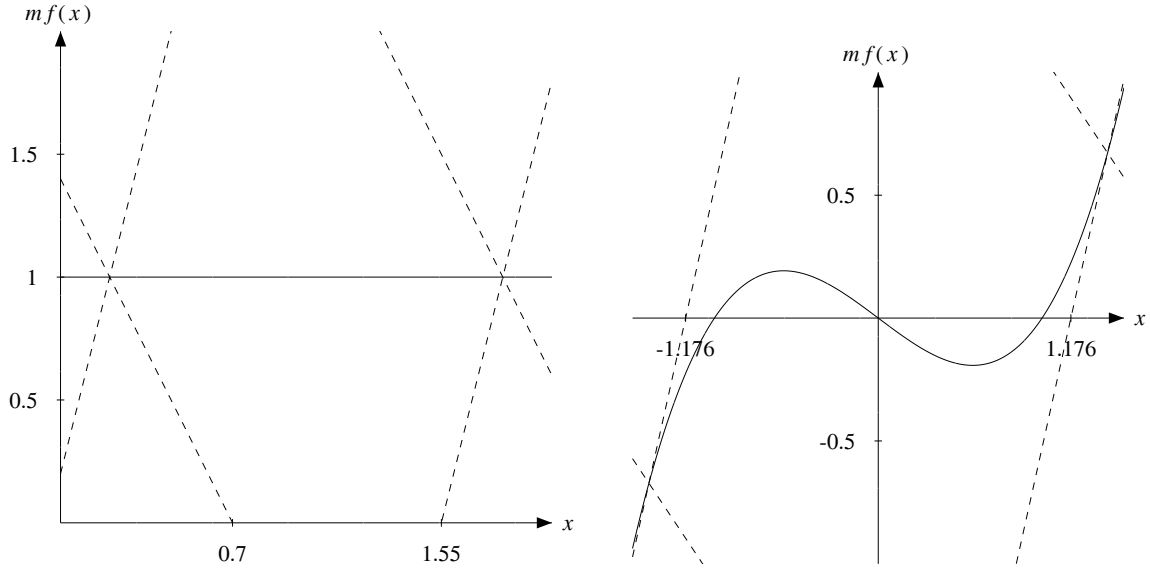


Figure 5.7.1: Counterexamples for iterated Hansen–Sengupta operator.

Let us analyze what happened in the first example. Here, we have two linearizations $mG^{(1)}$, $mG^{(2)}$ of mf in Z, where

$$\begin{aligned} mG^{(1)}(x) &= 1 + [-2,4](x - 0.2) \\ mG^{(2)}(x) &= 1 + [-2,4](x - 1.8). \end{aligned}$$

Let $G^{\cap}(x) = mG^{(1)}(x) \cap mG^{(2)}(x)$. Then $mf(x) \in G^{\cap}(x)$ for all $x \in X$. Extending the definition of face enclosedness to intersections of linear interval functions in the natural way, one would expect that $G^{\cap}$ is $X^{\overline{(1)}}$ enclosed. However, this is not the case. Instead, $G^{\cap}(X^{\overline{(1)}}) \geq 0$, $G^{\cap}(X^{\underline{(1)}}) \geq 0$ but neither $G^{\cap}(X^{\overline{(1)}}) \leq 0$ nor $G^{\cap}(X^{\underline{(1)}}) \leq 0$. Therefore Theorem 5.1.12, which is the basis of the existence proof, can not be applied in this case.

In the counterexample for uniqueness we have $Z \subseteq \mathrm{int}(X)$ but A is not regular and therefore Theorem 5.2.3 does not apply. One might conjecture that this situation will not arise if we apply only linear tightening steps where the denominator in the hull division of Theorem 5.4.6 does not contain zero. At least in the one–dimensional case this would trivially hold, because if the derivative does not contain zero, then f has at most one solution. But as Figure 5.7.2 shows, it is not true for $n = 2$. Here,

$$\begin{aligned} f_1(x_1, x_2) &= x_1^3 + 0.5x_1 - x_2 \\ f_2(x_1, x_2) &= x_2^3 + 0.5x_2 - x_1 \end{aligned}$$

and $\boldsymbol{X} = ([-2,2],[-2,2])^{\mathrm{T}}$. Then

$$\mathfrak{A} = \begin{pmatrix} [0.5, 12.5] & -1 \\ -1 & [0.5, 12.5] \end{pmatrix}$$

is a Jacobian of $\boldsymbol{f}$ in $\boldsymbol{X}$. The inverse midpoint preconditioner is

$$\mathfrak{m} \approx \begin{pmatrix} 0.157576 & 0.024242 \\ 0.024242 & 0.157576 \end{pmatrix}$$

and thus

$$\mathfrak{m}\,\mathfrak{A} \approx \begin{pmatrix} [0.054545, 1.945455] & [-0.145455, 0.145455] \\ [-0.145455, 0.145455] & [0.054545, 1.945455] \end{pmatrix}.$$

In the first graph of Figure 5.7.2 one sees that $\mathfrak{m}\,\boldsymbol{f}$ has 3 solutions in $\boldsymbol{X}$. In the second row we show the first application of the Hansen–Sengupta operator, where $\boldsymbol{c}^{(1)} = (-1.5, 1.5)^{\mathrm{T}}$. Here, we have the

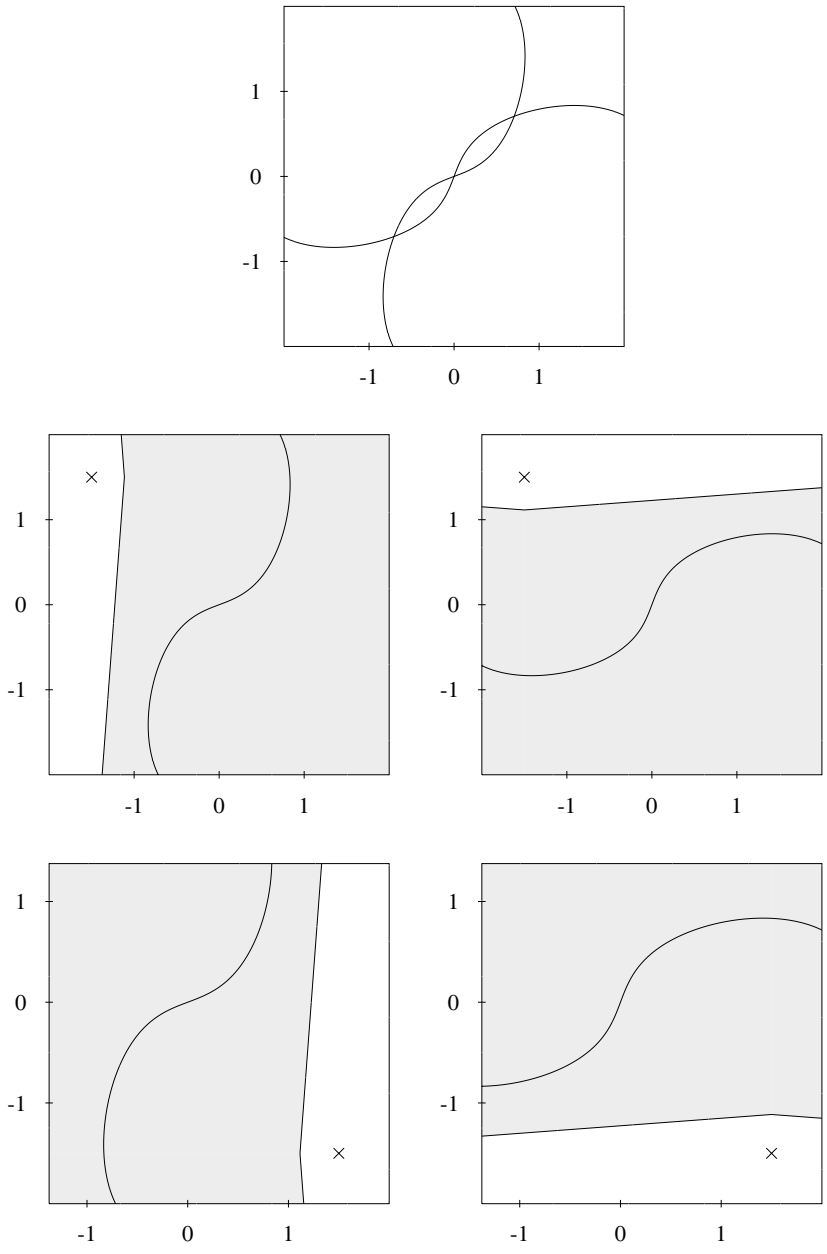


Figure 5.7.2: Counterexample for the uniqueness property of the iterated Hansen–Sengupta operator. The denominator in the hull division does not contain zero.

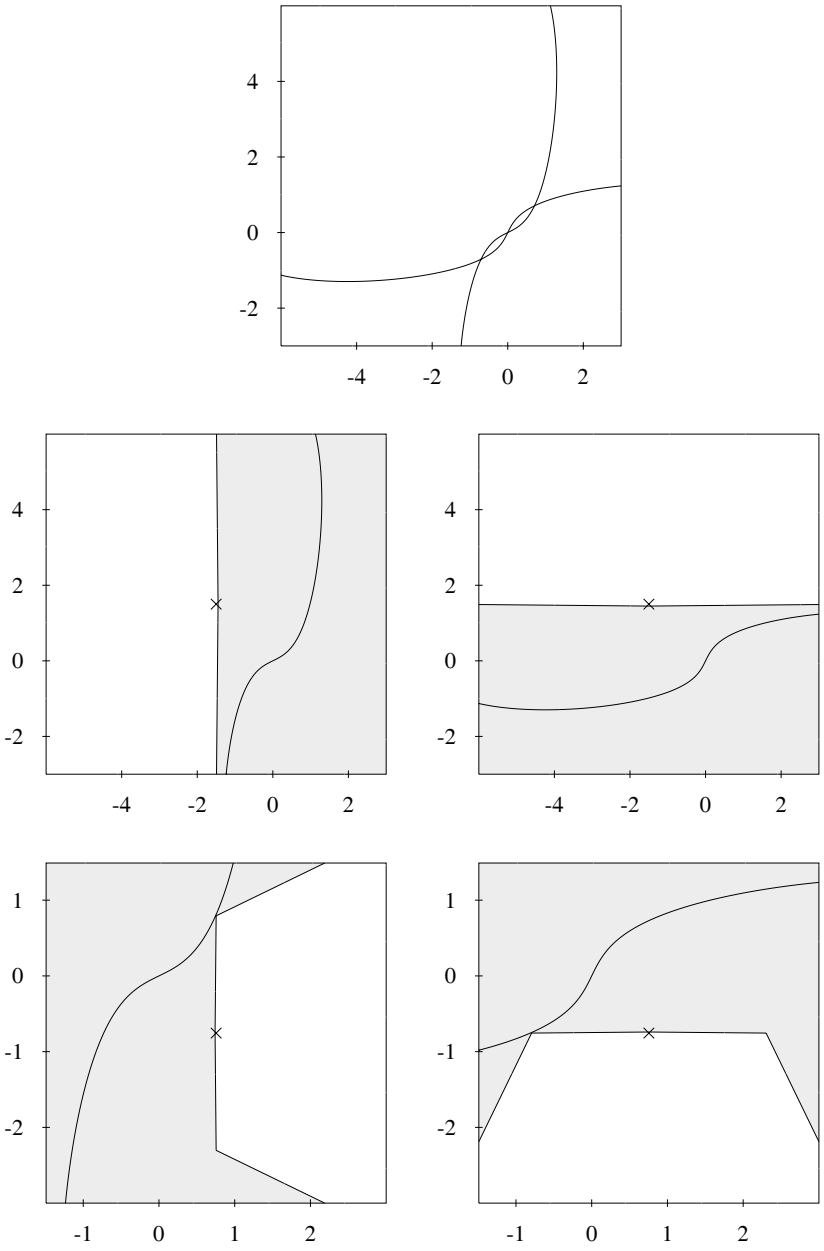


Figure 5.7.3: Counterexample for the uniqueness property of the iterated Hansen–Sengupta operator. The denominator in the hull division does not contain zero and the center is always chosen as the midpoint.

linearizations

$$\begin{aligned} G_1^{\mathfrak{m}(1)}(\boldsymbol{x}) &= -0.75 + [0.054545, 1.945455](x_1 + 1.5) + [-0.145455, 0.145455](x_2 - 1.5) \\ G_2^{\mathfrak{m}(1)}(\boldsymbol{x}) &= 0.75 + [-0.145455, 0.145455](x_1 + 1.5) + [0.054545, 1.945455](x_2 - 1.5) \end{aligned}$$

and thus

$$\mathcal{HS}(\boldsymbol{f}, \boldsymbol{X}, \boldsymbol{c}^{(1)}, \mathfrak{A}, \mathfrak{m}) = \boldsymbol{Y} \approx ([-1.376168, 2], [-2, 1.376168])^{\mathrm{T}}.$$

In the second row of Figure 5.7.2 we apply the Hansen–Sengupta operator again on $\boldsymbol{Y}$ with the same $\mathfrak{A}$, $\mathfrak{m}$, but this time we choose $\boldsymbol{c}^{(2)} = (1.5, -1.5)^{\mathrm{T}}$. We obtain the linearizations

$$\begin{aligned} G_1^{\mathfrak{m}(2)}(\boldsymbol{x}) &= 0.75 + [0.054545, 1.945455](x_1 - 1.5) + [-0.145455, 0.145455](x_2 + 1.5) \\ G_2^{\mathfrak{m}(2)}(\boldsymbol{x}) &= -0.75 + [-0.145455, 0.145455](x_1 - 1.5) + [0.054545, 1.945455](x_2 + 1.5) \end{aligned}$$

and

$$\mathcal{HS}(\boldsymbol{f}, \boldsymbol{X}, \boldsymbol{c}^{(2)}, \mathfrak{A}, \mathfrak{m}) = \boldsymbol{Z} \approx ([-1.376168, 1.329527], [-1.329527, 1.376168])^{\mathrm{T}}.$$

Hence, $\boldsymbol{Z} \subseteq \mathrm{int}(\boldsymbol{X})$ but in none of the tightening steps a hull division with a denominator containing zero occured.

Looking at Figure 5.7.2, one sees that the choice of $\boldsymbol{c}^{(1)}$, $\boldsymbol{c}^{(2)}$ close to the corners of the box played an essential role. If, as an additional restriction we would require that the chosen $\boldsymbol{c}$ is always the midpoint of the current box, would then the uniqueness property hold for iterated application of the Hansen–Sengupta operator? Again the answer is no, as Figure 5.7.3 shows. We use the same $\boldsymbol{f}$ as above but $\boldsymbol{X} = ([-6, 3], [-3, 6])^{\mathrm{T}}$, the Jacobian

$$\mathfrak{A} = \begin{pmatrix} [0.5, 108.5], -1 \\ -1, [0.5, 108.5] \end{pmatrix}$$

and the inverse midpoint preconditioner $\mathfrak{m}$. For $\boldsymbol{c}^{(1)} = \mathrm{mid}(\boldsymbol{X})$ we obtain $\boldsymbol{Y} = ([-1.4902, 3], [-3, 1.4902])^{\mathrm{T}}$, and for $\boldsymbol{c}^{(2)} = \mathrm{mid}(\boldsymbol{Y})$ we get $\boldsymbol{Z} = ([-1.4902, 2.188772], [-2.188772, 1.4902])^{\mathrm{T}}$. Again, $\boldsymbol{Z} \in \mathrm{int}(\boldsymbol{X})$, in none of the tightening steps a hull division with a denominator containing zero occured and $\boldsymbol{c}$ was in both iterations the midpoint of the current box. One possibility to restrict the Hansen–Sengupta operator such the uniqueness property holds also for iterated application would be to allow only simultaneous reductions on opposite faces. However, instead of pursuing such strong restrictions, we give a different uniqueness criterion in Section 5.8.

In the remainder of this section we modify the existence conditions of the Hansen–Sengupta or linearized tightening operator such that they still hold if the operators are applied iteratively.

5.7.1 Existence by Intersected Linearizations

Iterative application of operators like the Hansen–Sengupta operator or the linearized tightening operator is a special case of a tightening sequence, where different linearizations of the same nonlinear function occur. In this section we use such tightening sequences in order to prove existence of solutions of systems of nonlinear equations. As in the case of linear interval functions, the existence condition is based on the notion of weak face enclosedness and orthogonality. Given two linearizations $G_i^{(1)}$, $G_i^{(2)}$ of a nonlinear function f_i in $\boldsymbol{X}$ and a direction j, it can happen that neither $G_i^{(1)}$ nor $G_i^{(2)}$ is weakly $\boldsymbol{X}^{\underline{(j)}}$ enclosed, but $G_i^{(1)}(\boldsymbol{X}^{\overline{(j)}}) > 0$, $G_i^{(2)}(\boldsymbol{X}^{\underline{(j)}}) < 0$, and thus $G_i^{(1)} \cap G_i^{(2)}$ is $\boldsymbol{X}^{\overline{(j)}}$ enclosed. However, the intersection of two linear interval functions is usually not a linear interval function. Therefore, we first extend some basic properties shown in Section 5.1 for linear interval functions to intersections of linear interval functions.

Definition 5.7.1 (Intersected Linear Interval Function) *A function $G^{\cap} : \mathbb{R}^n \to \mathbb{IR}_{\emptyset}$ is called intersected linear interval function if there exist linear interval functions $G^{(1)}$, $G^{(2)}$ such that*

$$G^{\cap}(\boldsymbol{x}) = G^{(1)}(\boldsymbol{x}) \cap G^{(2)}(\boldsymbol{x})$$

for all $\boldsymbol{x} \in \mathbb{R}^n$. □

In the following let $G^\cap$ be an intersected linear interval function and let $f : \mathbb{R}^n \to \mathbb{R}$.

Definition 5.7.2 (Intersected Interval Linearization) *$G^\cap$ is called intersected interval linearization of f in $\boldsymbol{X}$ if there exist interval linearizations $G^{(1)}$, $G^{(2)}$ of f in $\boldsymbol{X}$ such that $G^\cap = G^{(1)} \cap G^{(2)}$.* □

If $G^\cap$ is an intersected interval linearization of f in $\boldsymbol{X}$ then $f(\boldsymbol{x}) \in G^\cap(\boldsymbol{x})$ for all $\boldsymbol{x} \in \boldsymbol{X}$, hence $G^\cap(\boldsymbol{x}) \neq \emptyset$ for all $\boldsymbol{x} \in \boldsymbol{X}$. Further, $G^\cap$ is an intersected interval linearization of f in every $\boldsymbol{Y} \subseteq \boldsymbol{X}$.

Let $G^\cap = G^{(1)} \cap G^{(2)}$ where $G^{(1)}$, $G^{(2)}$ are linear interval functions. The variety $\mathcal{Z}(G^\cap)$ of $G^\cap$ is defined as

$$\mathcal{Z}(G^\cap) = \{\boldsymbol{x} \in \mathbb{R}^n \mid 0 \in G^\cap(\boldsymbol{x})\}$$

and obviously

$$\mathcal{Z}(G^\cap) = \mathcal{Z}(G^{(1)}) \cap \mathcal{Z}(G^{(2)}).$$

As in the case of linear interval functions we use the following notation.

Notation. For an interval $\boldsymbol{X} \in \mathbb{IR}^n$ we write $G^\cap(\boldsymbol{X})$ to denote the set

$$\{y \mid y \in G^\cap(\boldsymbol{x}) \text{ for some } \boldsymbol{x} \in \boldsymbol{X}\}.$$

Further, we write $\boldsymbol{X} \sim 0$ if $\boldsymbol{x} \sim 0$ for all $\boldsymbol{x} \in \boldsymbol{X}$, where $\sim \in \{>, <, \geq, \leq\}$. □

Next, we extend the notion of face enclosedness to intersected linear interval functions in a straight forward way.

Definition 5.7.3 (Face Enclosed Intersected Linear Interval Function)

- *$G^\cap$ is weakly $\boldsymbol{X}^{\underline{(i)}}$ enclosed if $G^\cap(\boldsymbol{x}) \neq \emptyset$ for all $\boldsymbol{x} \in \boldsymbol{X}$ and*

$$\left(G^\cap(\boldsymbol{X}^{\underline{(i)}}) \geq 0 \text{ and } G^\cap(\boldsymbol{X}^{\overline{(i)}}) \leq 0\right) \text{ or } \left(G^\cap(\boldsymbol{X}^{\underline{(i)}}) \leq 0 \text{ and } G^\cap(\boldsymbol{X}^{\overline{(i)}}) \geq 0\right).$$

- *$G^\cap$ is strongly $\boldsymbol{X}^{\overline{(i)}}$ enclosed if $G^\cap(\boldsymbol{x}) \neq \emptyset$ for all $\boldsymbol{x} \in \boldsymbol{X}$ and*

$$\left(G^\cap(\boldsymbol{X}^{\underline{(i)}}) > 0 \text{ and } G^\cap(\boldsymbol{X}^{\overline{(i)}}) < 0\right) \text{ or } \left(G^\cap(\boldsymbol{X}^{\underline{(i)}}) < 0 \text{ and } G^\cap(\boldsymbol{X}^{\overline{(i)}}) > 0\right).$$ □

As Figure 5.7.1 shows, there exist intersected linear interval functions $G^\cap$ such that $G^\cap(\boldsymbol{x}) \neq \emptyset$ for all $\boldsymbol{x} \in \boldsymbol{X}$, $0 \notin G^\cap(\boldsymbol{X}^{\overline{(i)}})$, $0 \notin G^\cap(\boldsymbol{X}^{\underline{(i)}})$, $0 \in G^\cap(\boldsymbol{X})$ but still $G^\cap$ is not strongly $\boldsymbol{X}^{\overline{(i)}}$ enclosed. This is the crucial difference to linear interval functions, see Lemma 5.1.7. The following lemma gives a useful criterion when $G^\cap$ is $\boldsymbol{X}^{\overline{(i)}}$ enclosed.

Lemma 5.7.4 (Face Enclosed Intersected Linear Interval Function) *Let $G^\cap = G^{(1)} \cap G^{(2)}$ such that $G^\cap(\boldsymbol{x}) \neq \emptyset$ for all $\boldsymbol{x} \in \boldsymbol{X}$.*

- *If*

$$\left(G^{(1)}(\boldsymbol{X}^{\underline{(i)}}) \geq 0 \text{ and } G^{(2)}(\boldsymbol{X}^{\overline{(i)}}) \leq 0\right) \text{ or } \left(G^{(1)}(\boldsymbol{X}^{\underline{(i)}}) \leq 0 \text{ and } G^{(2)}(\boldsymbol{X}^{\overline{(i)}}) \geq 0\right)$$

  then $G^\cap$ is weakly $\boldsymbol{X}^{\overline{(i)}}$ enclosed.

- *If*

$$\left(G^{(1)}(\boldsymbol{X}^{\underline{(i)}}) > 0 \text{ and } G^{(2)}(\boldsymbol{X}^{\overline{(i)}}) < 0\right) \text{ or } \left(G^{(1)}(\boldsymbol{X}^{\underline{(i)}}) < 0 \text{ and } G^{(2)}(\boldsymbol{X}^{\overline{(i)}}) > 0\right)$$

  then $G^\cap$ is strongly $\boldsymbol{X}^{\overline{(i)}}$ enclosed. □

Proof. Obvious. □

The definitions of intersected linear interval function and intersected interval linearization extend to tuples component wise. In the following let $\boldsymbol{G}^\cap : \mathbb{R}^n \to \mathbb{IR}^n_\emptyset$ be an intersected linear interval function.

Definition 5.7.5 ($G^\cap$ Orthogonal in X) *$G^\cap$ is weakly respectively strongly orthogonal in $\boldsymbol{X}$ if there exists a permutation*

$$\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$$

such that $G_i^\cap$ is weakly respectively strongly $\boldsymbol{X}^{\overline{(\pi(i))}}$ enclosed for all i. □

Now, Theorem 5.1.2 can be specialized as follows.

Theorem 5.7.6 (Existence of Solutions) *If there exists an intersected interval linearization $\boldsymbol{G}^\cap$ of $\boldsymbol{f}$ in $\boldsymbol{X}$ which is weakly orthogonal in $\boldsymbol{X}$, then $\boldsymbol{f}$ has a solution in $\boldsymbol{X}$.* □

Proof. Let $\boldsymbol{G}^\cap$ be an interval linearization of $\boldsymbol{f}$ in $\boldsymbol{X}$ which is weakly orthogonal in $\boldsymbol{X}$. Let $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$ be a permutation such that $G_i^\cap$ is $\boldsymbol{X}^{\overline{(\pi(i))}}$ enclosed for all i. As $f_i(\boldsymbol{x}) \in G_i^\cap(\boldsymbol{x})$ for all $\boldsymbol{x} \in \boldsymbol{X}$, it follows that

$$f_i(\boldsymbol{a}) f_i(\boldsymbol{b}) \leq 0 \quad \text{for all} \quad \boldsymbol{a} \in \boldsymbol{X}^{\overline{(\pi(i))}}, \boldsymbol{b} \in \boldsymbol{X}^{\underline{(\pi(i))}}, i = 1, \ldots, n$$

and $\boldsymbol{f}$ has a solution in $\boldsymbol{X}$ by Theorem 5.1.2. □

5.7.2 Tightening with Multiple Linearizations

In Theorem 5.4.22 it was shown that if $\overline{\Delta}(\boldsymbol{X}, p^*, j) = \underline{\Delta}(\boldsymbol{X}, p^*, j) = i$, then G_i is weakly $\boldsymbol{Y}^{\overline{(j)}}$ enclosed, where $\boldsymbol{X} \xrightarrow{p^*} \boldsymbol{Y}$. From Figure 5.7.1, left graph, we see that this Theorem does not generalize to intersected linear interval functions: If $\mathcal{Z}(G^{(1)}) \cap \boldsymbol{Y}^{\overline{(j)}} = \emptyset$, $\mathcal{Z}(G^{(2)}) \cap \boldsymbol{Y}^{\underline{(j)}} = \emptyset$ then $G^{(1)} \cap G^{(2)}$ need not be $\boldsymbol{Y}^{\overline{(j)}}$ enclosed, even if $G^{(1)}$, $G^{(2)}$ are linearizations of the same f in $\boldsymbol{X}$. However, Lemma 5.7.4 shows that if the signs of $G^{(1)}$, $G^{(2)}$ on $\boldsymbol{X}^{\overline{(j)}}$, $\boldsymbol{X}^{\underline{(j)}}$ are opposite, then $G^{(1)} \cap G^{(2)}$ is in fact $\boldsymbol{X}^{\overline{(j)}}$ enclosed. Thus, after having computed the tightening sequence, we have to check this sign condition. Actually it is possible to compute the signs already during tightening and memorize them. Let us now formalize this method.

In the following let

$$p^* = \langle (G_{i_\ell}^{(\ell)}, j_\ell) \mid i_\ell, j_\ell \in \{1, \ldots, n\}, \ell = 1, \ldots, m \rangle$$

and

$$^{\ell-1}\boldsymbol{X} \xrightarrow{G_{i_\ell}^{(\ell)}, j_\ell} {}^{\ell}\boldsymbol{X},$$

where $G_{i_\ell}^{(\ell)}$ is a linearization of f_i over $^{\ell-1}\boldsymbol{X}$ and $^0\boldsymbol{X} = \boldsymbol{X}$. We modify the definition of the face disjointness functions $\overline{\Delta}$, $\underline{\Delta}$ as follows:

Definition 5.7.7 (Face Disjointness Functions)

$$\overline{\Delta}(\boldsymbol{X}, p^*, j) = \begin{cases} (i_{\hat{\ell}}, \hat{\ell}, \operatorname{sign}(G_{i_{\hat{\ell}}}^{(\hat{\ell})}(^{\hat{\ell}-1}\boldsymbol{X}^{\overline{(j)}}))) & \text{if } \hat{\ell} = \max\{\ell \mid {}^{\ell-1}\overline{X}_j > {}^{\ell}\overline{X}_j\} \text{ exists} \\ \uparrow & \text{else} \end{cases}$$

$$\underline{\Delta}(\boldsymbol{X}, p^*, j) = \begin{cases} (i_{\hat{\ell}}, \hat{\ell}, \operatorname{sign}(G_{i_{\hat{\ell}}}^{(\hat{\ell})}(^{\hat{\ell}-1}\boldsymbol{X}^{\underline{(j)}}))) & \text{if } \hat{\ell} = \max\{\ell \mid {}^{\ell-1}\underline{X}_j < {}^{\ell}\underline{X}_j\} \text{ exists} \\ \uparrow & \text{else.} \end{cases}$$ □

Note that

$${}^{\ell-1}\overline{X}_j > {}^{\ell}\overline{X}_j \text{ implies } 0 \notin G_{i_\ell}^{(\ell)}(^{\ell-1}\boldsymbol{X}^{\overline{(j)}})$$

and

$${}^{\ell-1}\underline{X}_j < {}^{\ell}\underline{X}_j \text{ implies } 0 \notin G_{i_\ell}^{(\ell)}(^{\ell-1}\boldsymbol{X}^{\underline{(j)}}),$$

hence

$$\operatorname{sign}(G_{i_\ell}^{(\ell)}(^{\ell-1}\boldsymbol{X}^{\overline{(j)}})), \quad \operatorname{sign}(G_{i_\ell}^{(\ell)}(^{\ell-1}\boldsymbol{X}^{\underline{(j)}}))$$

are well defined.

The essential property of the face disjointness functions is captured by the following theorem.

Theorem 5.7.8 (Face Disjointness Functions) *Let* $\boldsymbol{X} \xrightarrow{p^*} \boldsymbol{Y} \neq \emptyset$.

(i) *If* $\overline{\Delta}(\boldsymbol{X}, p^*, j) = (i, \ell, s) \neq \uparrow$ *then* $sG_i^{(\ell)}(\boldsymbol{Y}^{\overline{(j)}}) \geq 0$.

(ii) *If* $\underline{\Delta}(\boldsymbol{X}, p^*, j) = (i, \ell, s) \neq \uparrow$ *then* $sG_i^{(\ell)}(\boldsymbol{Y}^{\underline{(j)}}) \geq 0$. □

Proof. We give a proof of (i), the proof of (ii) is similar. Assume $\boldsymbol{X} \xrightarrow{p^*} \boldsymbol{Y} \neq \emptyset$. Let

$$\overline{\Delta}(\boldsymbol{X}, p^*, j) = (i, \ell, s) \neq \uparrow .$$

Then

$$sG_i^{(\ell)}({}^{\ell-1}\boldsymbol{X}^{\overline{(j)}}) > 0,$$

hence

$$sG_i^{(\ell)}({}^{\ell}\boldsymbol{X}^{\overline{(j)}}) \geq 0.$$

As

$${}^{\ell}\overline{X}_j = {}^{\ell+1}\overline{X}_j = \ldots = \overline{Y}_j ,$$

it holds that

$$\boldsymbol{Y}^{\overline{(j)}} \subseteq {}^{\ell}\boldsymbol{X}^{\overline{(j)}},$$

hence

$$sG_i^{(\ell)}(\boldsymbol{Y}^{\overline{(j)}}) \geq 0. \; \square$$

Theorem 5.7.9 (Weak Enclosedness) *Assume* $\boldsymbol{X} \xrightarrow{p^*} \boldsymbol{Y} \neq \emptyset$ *and*

$$\overline{\Delta}(\boldsymbol{X}, p^*, j) = (i, \ell, s), \quad \underline{\Delta}(\boldsymbol{X}, p^*, j) = (i, \ell', -s)$$

for some $i \in \{1, \ldots, n\}$, $\ell, \ell' \in \{1, \ldots, m\}$ *and* $s \in \{-1, 1\}$. *Then* $G_i^{(\ell)} \cap G_i^{(\ell')}$ *is weakly* $\boldsymbol{Y}^{\overline{(j)}}$ *enclosed.* □

Proof. Assume $\boldsymbol{X} \xrightarrow{p^*} \boldsymbol{Y} \neq \emptyset$ and

$$\overline{\Delta}(\boldsymbol{X}, p^*, j) = (i, \ell, s), \quad \underline{\Delta}(\boldsymbol{X}, p^*, j) = (i, \ell', -s).$$

As $G_i^{(\ell)}$, $G_i^{(\ell')}$ are linearizations of f_i in $\boldsymbol{Y}$, it holds that $G_i^{(\ell)}(\boldsymbol{x}) \neq \emptyset$, $G_i^{(\ell')}(\boldsymbol{x}) \neq \emptyset$ for all $\boldsymbol{x} \in \boldsymbol{Y}$. From Theorem 5.7.8 it follows that $sG_i^{(\ell)}(\boldsymbol{Y}^{\overline{(j)}}) \geq 0$, $sG_i^{(\ell')}(\boldsymbol{Y}^{\underline{(j)}}) \leq 0$ and thus $G_i^{(\ell)} \cap G_i^{(\ell')}$ is weakly $\boldsymbol{Y}^{\overline{(j)}}$ enclosed by Lemma 5.7.4. □

Theorem 5.7.10 (Existence of Solutions) *Assume* $\boldsymbol{X} \xrightarrow{p^*} \boldsymbol{Y} \neq \emptyset$ *and there exists a permutation* $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$ *such that*

$$\overline{\Delta}(\boldsymbol{X}, p^*, j) = (\pi(j), \ell_j, s_j), \quad \underline{\Delta}(\boldsymbol{X}, p^*, j) = (\pi(j), \ell'_j, -s_j)$$

for all j. *Then* $\boldsymbol{f}$ *has a solution in* $\boldsymbol{X}$. □

Proof. Follows from Theorem 5.7.9 and Theorem 5.7.6. □

In the following let

$$\boldsymbol{G}^{(\ell)}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{c}^{(\ell)}) + \mathfrak{A}^{(\ell)}(\boldsymbol{x} - \boldsymbol{c}^{(\ell)})$$

be a linearization of $\boldsymbol{f}$ in ${}^{\ell-1}\boldsymbol{X}$ for $\ell = 1, \ldots, m$.

Theorem 5.7.11 (Weak Enclosedness) *Assume* $\boldsymbol{X} \xrightarrow{p^*} \boldsymbol{Y} \neq \emptyset$ *and*

$$\overline{\Delta}(\boldsymbol{X}, p^*, j) = (i, \ell, s), \quad \underline{\Delta}(\boldsymbol{X}, p^*, j) = (i, \ell', s').$$

(i) *If* $A_{ij}^{(\ell)} \geq 0$ *and* $A_{ij}^{(\ell')} \geq 0$ *then* $s = 1$, $s' = -1$.

(ii) *If* $A_{ij}^{(\ell)} \leq 0$ *and* $A_{ij}^{(\ell')} \leq 0$ *then* $s = -1$, $s' = 1$.

In both cases $G_i^{(\ell)} \cap G_i^{(\ell')}$ *is weakly* $\boldsymbol{Y}^{\overline{(j)}}$ *enclosed.* □

Proof. We give a proof of (i), the proof of (ii) is analogous. Assume $\boldsymbol{X} \xrightarrow{p^*} \boldsymbol{Y} \neq \emptyset$,

$$\overline{\Delta}(\boldsymbol{X}, p^*, j) = (i, \ell, s), \quad \underline{\Delta}(\boldsymbol{X}, p^*, j) = (i, \ell', s').$$

and $A_{ij}^{(\ell)} \geq 0$, $A_{ij}^{(\ell')} \geq 0$. Using Theorem 5.7.8 it suffices to show that

$$G_i^{(\ell)}({}^{\ell-1}\boldsymbol{X}^{\overline{(j)}}) > 0, \quad G_i^{(\ell')}({}^{\ell'-1}\boldsymbol{X}^{\underline{(j)}}) < 0.$$

As

$${}^{\ell}\boldsymbol{X} = \mathrm{tight}({}^{\ell-1}\boldsymbol{X}, G_i^{(\ell)}, j)$$

and ${}^{\ell-1}\overline{X}_j > {}^{\ell}\overline{X}_j$, it holds that

$$0 \notin G_i^{(\ell)}({}^{\ell-1}\boldsymbol{X}^{\overline{(j)}}).$$

Assume

$$G_i^{(\ell)}({}^{\ell-1}\boldsymbol{X}^{\overline{(j)}}) < 0.$$

As $G_i^{(\ell)}$ is a set of linear functions and $A_{ij}^{(\ell)} \geq 0$, it holds that

$$G_i^{(\ell)}({}^{\ell-1}\boldsymbol{X}^{\underline{(j)}}) < 0,$$

and hence $0 \notin G_i^{(\ell)}({}^{\ell-1}\boldsymbol{X})$. But then ${}^{\ell}\boldsymbol{X} = \emptyset$, which contradicts the assumption $\boldsymbol{Y} \neq \emptyset$. The proof of

$$G_i^{(\ell')}({}^{\ell'-1}\boldsymbol{X}^{\underline{(j)}}) < 0$$

is analogous. Hence $G_i^{(\ell)} \cap G_i^{(\ell')}$ is weakly $\boldsymbol{Y}^{\overline{(j)}}$ enclosed by Theorem 5.7.9. □

Theorem 5.7.12 (Existence of Solutions) *Assume* $\boldsymbol{X} \xrightarrow{p^*} \boldsymbol{Y} \neq \emptyset$ *and there exists a permutation* $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$ *such that for all* j

$$\overline{\Delta}(\boldsymbol{X}, p^*, j) = (\pi(j), \ell_j, s_j), \quad \underline{\Delta}(\boldsymbol{X}, p^*, j) = (\pi(j), \ell'_j, s'_j)$$

and either

$$A_{\pi(j),j}^{(\ell_j)} \geq 0, \quad A_{\pi(j),j}^{(\ell'_j)} \geq 0$$

or

$$A_{\pi(j),j}^{(\ell_j)} \leq 0, \quad A_{\pi(j),j}^{(\ell'_j)} \leq 0.$$

Then $s_j = -s'_j$ *and* $\boldsymbol{f}$ *has a solution in* $\boldsymbol{X}$. □

Proof. Follows from Theorem 5.7.11 and Theorem 5.7.6. □

5.7.3 Iteration of the Hansen–Sengupta Operator

In this section we use Theorem 5.7.12 for proving existence of solutions through iterated application of the Hansen–Sengupta operator. Let $\boldsymbol{X}^{(1)} = \boldsymbol{X}$, $\mathfrak{m} \in \mathbb{R}^{n \times n}$,

$$\boldsymbol{X}^{(k+1)} = \mathcal{HS}(\boldsymbol{f}, \boldsymbol{X}^{(k)}, \boldsymbol{c}^{(k)}, \mathfrak{A}^{(k)}, \mathfrak{m})$$

such that

$$\boldsymbol{G}^{(k)} = \boldsymbol{f}(\boldsymbol{c}^{(k)}) + \mathfrak{A}^{(k)}(\boldsymbol{x} - \boldsymbol{c}^{(k)})$$

is a linearization of $\boldsymbol{f}$ in $\boldsymbol{X}^{(k)}$, $k = 1, \ldots, r$ for some $r \geq 1$ and let $\boldsymbol{Y} = \boldsymbol{X}^{(r+1)}$. Note that throughout the iteration the preconditioning matrix $\mathfrak{m}$ is fixed.

Theorem 5.7.13 (Existence of Solution by Iterated Hansen–Sengupta Operator) *If* $A_{ii}^{\mathfrak{m}(k)} \geq 0$ *for all* $k = 1, \ldots, r$, $i \in \{1, \ldots, n\}$ *and* $\boldsymbol{Y} \subseteq \mathrm{int}(\boldsymbol{X})$, *then* $\boldsymbol{f}$ *has a solution in* $\boldsymbol{Y}$. □

Proof. Assume $A_{ii}^{\mathfrak{m}(k)} \geq 0$ for all k, i and $\boldsymbol{Y} \subseteq \mathrm{int}(\boldsymbol{X})$. Let

$$\begin{aligned} p^* \quad = \langle \quad & (G_1^{\mathfrak{m}(1)}, 1), (G_2^{\mathfrak{m}(1)}, 2), \ldots, (G_n^{\mathfrak{m}(1)}, n), \\ & (G_1^{\mathfrak{m}(2)}, 1), (G_2^{\mathfrak{m}(2)}, 2), \ldots, (G_n^{\mathfrak{m}(2)}, n), \ldots \\ & (G_1^{\mathfrak{m}(r)}, 1), (G_2^{\mathfrak{m}(r)}, 2), \ldots, (G_n^{\mathfrak{m}(r)}, n) \quad \rangle. \end{aligned}$$

As $\boldsymbol{Y} \subseteq \mathrm{int}(\boldsymbol{X})$ it holds that

$$\overline{\Delta}(\boldsymbol{X}, p^*, i) = (i, k_i, s_i), \quad \underline{\Delta}(\boldsymbol{X}, p^*, i) = (i, k_i', s_i')$$

for all i and for some $k_i, k_i' \in \{1, \ldots, r\}$, $s_i, s_i' \in \{-1, 1\}$ and an application of Theorem 5.7.12 completes the proof. □

5.8 Convergence of the Hansen–Sengupta Operator

In this section we give criteria when iterated application of the Hansen–Sengupta operator converges. First, let us define what we mean by a convergent sequence of interval vectors. In order to simplify the presentation, we define $\hat{\mathrm{w}}(\emptyset) = 0$.

Definition 5.8.1 (Convergent Sequence of Interval Vectors) *A sequence of interval vectors*

$$\langle \boldsymbol{X}^{(k)} \in \mathbb{IR}_\emptyset \mid k = 1, 2, \ldots \rangle$$

is called convergent if for every $\varepsilon > 0$ *there exists* $\hat{k} \in \mathbb{N}$ *such that for all* $k \geq \hat{k}$ *it holds that* $\hat{\mathrm{w}}(\boldsymbol{X}^{(k)}) < \varepsilon$. □

Note that this is the usual definition for convergence of the sequence $\langle \hat{\mathrm{w}}(\boldsymbol{X}^{(k)}) \in \mathbb{R} \mid k = 1, 2, \ldots \rangle$ to 0, but not for convergence of the sequence $\langle \boldsymbol{X}^{(k)} \in \mathbb{IR}_\emptyset \mid k = 1, 2, \ldots \rangle$. We will use the following well known criterion for convergence.

Theorem 5.8.2 (Criterion for Convergence) *If there exists* $\alpha < 1$ *such that*

$$\hat{\mathrm{w}}(\boldsymbol{X}^{(k+1)}) \leq \alpha \hat{\mathrm{w}}(\boldsymbol{X}^{(k)})$$

for all k, *then* $\langle \boldsymbol{X}^{(k)} \mid k = 1, 2, \ldots \rangle$ *is convergent.* □

Throughout this section let

$$\boldsymbol{X}^{(k+1)} = \mathcal{HS}(\boldsymbol{f}, \boldsymbol{X}^{(k)}, \boldsymbol{c}^{(k)}, \mathfrak{A}^{(k)}, \mathfrak{m})$$

where

$$\boldsymbol{G}^{(k)} = \boldsymbol{f}(\boldsymbol{c}^{(k)}) + \mathfrak{A}^{(k)}(\boldsymbol{x} - \boldsymbol{c}^{(k)})$$

is a linearization of $\boldsymbol{f}$ in $\boldsymbol{X}^{(k)}$, $\boldsymbol{c}^{(k)} \in \boldsymbol{X}^{(k)}$ and $\mathfrak{A}^{(k)} \subseteq \mathfrak{A}^{(1)}$ for $k = 1, 2, \ldots$. Instead of $\mathfrak{A}^{(1)}$ we will simply write $\mathfrak{A}$. The following theorem is a slight generalization of [Hansen and Sengupta, 1981] to the case when $\boldsymbol{c}^{(k)}$ is arbitrarily chosen from $\boldsymbol{X}^{(k)}$ and $\boldsymbol{X}$ need not contain a solution of $\boldsymbol{f}$.

Theorem 5.8.3 (Convergence of Hansen–Sengupta Operator) *Let*

$$\begin{aligned} d_i &= \mathrm{q}(A_{ii}^{\mathfrak{m}}, 1) \\ r_i &= \sum_{\substack{j=1 \\ j \neq i}}^{n} \mathrm{mag}(A_{ij}^{\mathfrak{m}}) \end{aligned}$$

and let

$$\alpha_i = \frac{2(2d_i r_i + d_i^2 + d_i + r_i)}{1 - d_i^2}.$$

If $d_i < 1$ *and* $\alpha_i < 1$ *for all* $i = 1, \ldots, n$ *then* $\langle \boldsymbol{X}^{(k)} \mid k = 1, 2, \ldots \rangle$ *is convergent and*

$$\hat{\mathrm{w}}(\boldsymbol{X}^{(k+1)}) \leq \alpha \hat{\mathrm{w}}(\boldsymbol{X}^{(k)})$$

where $\alpha = \max_i \alpha_i$. □

Proof. Let $d_i, r_i, \alpha_i, \alpha$ as in Theorem 5.8.3 and assume $d_i < 1$, $\alpha_i < 1$ for all i. We show that $\hat{\mathrm{w}}(\boldsymbol{X}^{(k+1)}) \leq \alpha \hat{\mathrm{w}}(\boldsymbol{X}^{(k)})$ for all k, and apply Theorem 5.8.2. Hence, let $k \geq 1$ arbitrary but fixed. We distinguish two cases.

- Assume
$$0 \notin f_i^{\mathfrak{m}}(\boldsymbol{c}^{(k)}) + \boldsymbol{A}_i^{\mathfrak{m}(k)}(\boldsymbol{X}^{(k)} - \boldsymbol{c}^{(k)}) \quad \text{for some} \quad i. \tag{5.8.1}$$
We show that $\boldsymbol{X}^{(k+1)} = \emptyset$. From (5.8.1) it follows that
$$A_{ii}^{\mathfrak{m}(k)}(X_i^{(k)} - c_i^{(k)}) \cap \left(-f_i^{\mathfrak{m}}(\boldsymbol{c}^{(k)}) - \sum_{j \neq i} A_{ij}^{\mathfrak{m}(k)}(X_j^{(k)} - c_j^{(k)}) \right) = \emptyset.$$
As $d_i < 1$ it holds that $0 \notin A_{ii}^{\mathfrak{m}(k)}$ and
$$X_i^{(k)} \cap c_i^{(k)} - \frac{-f_i^{\mathfrak{m}}(\boldsymbol{c}^{(k)}) - \sum_{j \neq i} A_{ij}^{\mathfrak{m}(k)}(X_j^{(k)} - c_j^{(k)})}{A_{ii}^{\mathfrak{m}(k)}} = \emptyset.$$
Hence, $\boldsymbol{X}^{(k+1)} = \mathcal{HS}(\boldsymbol{f}, \boldsymbol{X}^{(k)}, \boldsymbol{c}^{(k)}, \mathfrak{A}, \mathfrak{m}) = \emptyset$.

- Assume
$$0 \in f_i^{\mathfrak{m}}(\boldsymbol{c}^{(k)}) + \boldsymbol{A}_i^{\mathfrak{m}(k)}(\boldsymbol{X}^{(k)} - \boldsymbol{c}^{(k)}) \quad \text{for all} \quad i. \tag{5.8.2}$$
As $d_1 < 1$ it holds that $0 \notin A_{ii}^{\mathfrak{m}}$ for all i and
$$\begin{aligned} X_i^{(k+1)} &\subseteq c_i^{(k)} - \frac{f_i^{\mathfrak{m}}(\boldsymbol{c}^{(k)}) + \sum_{j \neq i} A_{ij}^{\mathfrak{m}}(X_j^{(k)} - c_j^{(k)})}{A_{ii}^{\mathfrak{m}}} \\ &\subseteq c_i^{(k)} - \frac{f_i^{\mathfrak{m}}(\boldsymbol{c}^{(k)}) + \sum_{j \neq i} \mathrm{mag}(A_{ij}^{\mathfrak{m}})\mathrm{w}(X^{(k)})[-1,1]}{A_{ii}^{\mathfrak{m}}} \\ &\subseteq c_i^{(k)} - \frac{f_i^{\mathfrak{m}}(\boldsymbol{c}^{(k)}) + r_i \hat{\mathrm{w}}(\boldsymbol{X}^{(k)})[-1,1]}{[1 - d_i, 1 + d_i]} \\ &\subseteq c_i^{(k)} - \frac{f_i^{\mathfrak{m}}(\boldsymbol{c}^{(k)})}{[1 - d_i, 1 + d_i]} - \frac{r_i \hat{\mathrm{w}}(\boldsymbol{X}^{(k)})[-1,1]}{1 - d_i}. \end{aligned}$$
Hence,
$$\mathrm{w}(X_i^{(k+1)}) \leq \mathrm{w}\left(\frac{f_i^{\mathfrak{m}}(\boldsymbol{c}^{(k)})}{[1 - d_i, 1 + d_i]} \right) + \frac{2 r_i \hat{\mathrm{w}}(\boldsymbol{X}^{(k)})}{1 - d_i}.$$
Now,
$$\begin{aligned} \mathrm{w}\left(\frac{f_i^{\mathfrak{m}}(\boldsymbol{c}^{(k)})}{[1 - d_i, 1 + d_i]} \right) &= \frac{|f_i^{\mathfrak{m}}(\boldsymbol{c}^{(k)})|}{1 - d_i} - \frac{|f_i^{\mathfrak{m}}(\boldsymbol{c}^{(k)})|}{1 + d_i} \\ &= |f_i^{\mathfrak{m}}(\boldsymbol{c}^{(k)})| \left(\frac{2 d_i}{1 - d_i^2} \right). \end{aligned}$$

From 5.8.2 it follows that

$$\begin{aligned} f_i^{\mathrm{m}}(\boldsymbol{c}^{(k)}) &\in -\sum_{j\neq i} A_{ij}^{\mathrm{m}}(X_j^{(k)} - c_j^{(k)}) - A_{ii}^{\mathrm{m}}(X_i^{(k)} - c_i^{(k)}) \\ &\subseteq \sum_{j\neq i} \mathrm{mag}(A_{ij}^{\mathrm{m}})\mathrm{w}(X_j^{(k)})[-1,1] + \mathrm{mag}(A_{ii}^{\mathrm{m}})\mathrm{w}(X_i^{(k)})[-1,1] \\ &\subseteq r_i \hat{\mathrm{w}}(\boldsymbol{X}^{(k)})[-1,1] + (1+d_i)\hat{\mathrm{w}}(\boldsymbol{X}^{(k)})[-1,1]. \end{aligned}$$

Hence

$$|f_i^{\mathrm{m}}(\boldsymbol{c}^{(k)})| \leq (r_i + d_i + 1)\hat{\mathrm{w}}(\boldsymbol{X}^{(k)}).$$

Summarizing, we obtain

$$\begin{aligned} \mathrm{w}(X_i^{(k+1)}) &\leq \left(\frac{2d_i(r_i+d_i+1)}{1-d_i^2} + \frac{2r_i}{1-d_i}\right)\hat{\mathrm{w}}(\boldsymbol{X}^{(k)}) \\ &= \frac{2d_i(r_i+d_i+1) + 2r_i(1+d_i)}{1-d_i^2}\hat{\mathrm{w}}(\boldsymbol{X}^{(k)}) \\ &= \frac{2(2d_ir_i + d_i^2 + d_i + r_i)}{1-d_i^2}\hat{\mathrm{w}}(\boldsymbol{X}^{(k)}) \\ &= \alpha_i \hat{\mathrm{w}}(\boldsymbol{X}^{(k)}). \end{aligned}$$

Thus,

$$\hat{\mathrm{w}}(\boldsymbol{X}^{(k+1)}) \leq \alpha\, \hat{\mathrm{w}}(\boldsymbol{X}^{(k)})$$

and $\langle \boldsymbol{X}^{(k)} \mid k = 1,2,\ldots\rangle$ is convergent by Theorem 5.8.2. □

Theorem 5.8.4 (Convergence of Hansen–Sengupta Operator) *Let*

$$\alpha_i = \sum_{\substack{j=1 \\ j\neq i}}^{n} \mathrm{mag}(A_{ij}^{\mathrm{m}}) + \mathrm{mag}(1 - A_{ii}^{\mathrm{m}})$$

and let $\alpha = \max_i \alpha_i$.

(i) *If* $\alpha < 1/2$ *then* $\langle \boldsymbol{X}^{(k)} \mid k = 1,2,\ldots\rangle$ *is convergent and*

$$\hat{\mathrm{w}}(\boldsymbol{X}^{(k+1)}) \leq 2\alpha\, \hat{\mathrm{w}}(\boldsymbol{X}^{(k)})$$

(ii) *If* $\alpha < 1$ *and* $\boldsymbol{c}^{(k)} = \mathrm{mid}(\boldsymbol{X}^{(k)})$ *for all* k, *then* $\langle \boldsymbol{X}^{(k)} \mid k = 1,2,\ldots\rangle$ *is convergent and*

$$\hat{\mathrm{w}}(\boldsymbol{X}^{(k+1)}) \leq \alpha\, \hat{\mathrm{w}}(\boldsymbol{X}^{(k)}).$$ □

Proof. See [Neumaier, 1990], Corollary 5.2.4 and Theorem 4.3.5. □

Before we give another convergence theorem for the Hansen–Sengupta operator, we state the following two technical lemmas about the width of an interval quotient and an interval product.

Lemma 5.8.5 (Width of Quotient) *Let* $N, D \in \mathbb{IR}$ *and* $y \in \mathbb{R}$ *such that* $0 \in N$, $0 \notin D$. *Then*

(i) $\mathrm{w}(N/D) = \mathrm{w}(N)/\mathrm{mig}(D)$.

(ii) $\mathrm{w}(N/D) \leq \mathrm{w}((N+y)/D)$. □

Proof. Let N, D, y as in Lemma 5.8.5. As $\mathrm{w}(N/D) = \mathrm{w}(-N/D)$ we may assume $D > 0$.

(i)

$$\begin{aligned} \mathrm{w}(N/D) &= \overline{N/D} - \underline{N/D} \\ &= \overline{N}/\mathrm{mig}(D) - \underline{N}/\mathrm{mig}(D) \\ &= \mathrm{w}(N)/\mathrm{mig}(D). \end{aligned}$$

(ii) If $0 \in N + y$ then according to (i)

$$\begin{aligned} \mathrm{w}(N/D) &= \mathrm{w}(N)/\mathrm{mig}(D) \\ &= \mathrm{w}(N+y)/\mathrm{mig}(D) \\ &= \mathrm{w}((N+y)/D). \end{aligned}$$

Otherwise, as $\mathrm{w}((N+y)/D) = \mathrm{w}(-(N+y)/D)$ we may assume $N + y > 0$. Then

$$\begin{aligned} \mathrm{w}(N/D) &= \mathrm{w}(N)/\mathrm{mig}(D) \\ &= \mathrm{w}(N+y)/\mathrm{mig}(D) \\ &= (\overline{N}+y)/\mathrm{mig}(D) - (\underline{N}+y)/\mathrm{mig}(D) \\ &\le (\overline{N}+y)/\mathrm{mig}(D) - (\underline{N}+y)/\mathrm{mag}(D) \\ &= \overline{(N+y)/D} - \underline{(N+y)/D} \\ &= \mathrm{w}((N+y)/D). \ \square \end{aligned}$$

Lemma 5.8.6 (Width of Product) *Let $A, X \in \mathbb{IR}$ and let $c \in X$. Then*

$$\mathrm{w}(A(X-c)) \le 2\mathrm{mag}(A)\mathrm{q}(X,c). \ \square$$

Proof. Let A, X, c as in Lemma 5.8.6. Then

$$\begin{aligned} \mathrm{w}(A(X-c)) &= \overline{A(X-c)} - \underline{A(X-c)} \\ &\le \max\{|a(x-c)| \mid a \in A, x \in X\} - \min\{-|a(x-c)| \mid a \in A, x \in X\} \\ &= \mathrm{mag}(A)\mathrm{q}(X,c) + \mathrm{mag}(A)\mathrm{q}(X,c) \\ &= 2\mathrm{mag}(A)\mathrm{q}(X,c). \ \square \end{aligned}$$

The following theorem is new. It gives a convergence condition for the Hansen–Sengupta operator which is even weaker than Theorem 5.8.4.

Theorem 5.8.7 (Convergence of Hansen–Sengupta Operator) *Assume $0 \notin A_{ii}^{\mathfrak{m}}$ for $i = 1, \ldots, n$, let*

$$\alpha_i = \sum_{\substack{j=1 \\ j \ne i}}^{n} \mathrm{mag}(A_{ij}^{\mathfrak{m}})/\mathrm{mig}(A_{ii}^{\mathfrak{m}})$$

and let $\alpha = \max_i \alpha_i$.

(i) *If $\alpha < 1/2$ and $\boldsymbol{X}^{(k)} \ne \emptyset$ then*

$$\hat{\mathrm{w}}(\boldsymbol{X}^{(k+1)}) < \hat{\mathrm{w}}(\boldsymbol{X}^{(k)})$$

for all k.

(ii) *If $\boldsymbol{X}^{(k)} \ne \emptyset$ and there exists $\beta < 1$ such that*

$$\mathrm{q}(c_i^{(k)}, X_i^{(k)}) \le \beta \mathrm{w}(X_i^{(k)})$$

for all k, i and $\alpha\beta < 1/2$ then $\langle \boldsymbol{X}^{(k)} \mid k = 1, 2, \ldots \rangle$ is convergent and

$$\hat{\mathrm{w}}(\boldsymbol{X}^{(k+1)}) \le \beta \max(2\alpha, 1)\hat{\mathrm{w}}(\boldsymbol{X}^{(k)}). \ \square$$

Proof. Let α_i, α as in Theorem 5.8.7 and assume $0 \notin A_{ii}^{\mathfrak{m}}$ for all i. Let k arbitrary but fixed and assume $\boldsymbol{X}^{(k)} \neq \emptyset$. Further, let i such that $\mathrm{w}(X_i^{(k+1)}) = \hat{\mathrm{w}}(\boldsymbol{X}^{(k+1)})$.

- Assume $c_i^{(k)} \notin X_i^{(k+1)}$.

  (i) As $c_i^{(k)} \in X_i^{(k)}$ and $X_i^{(k+1)} \subseteq X_i^{(k)}$, it holds that $X_i^{(k+1)} \subset X_i^{(k)}$. Hence $\hat{\mathrm{w}}(\boldsymbol{X}^{(k+1)}) < \hat{\mathrm{w}}(\boldsymbol{X}^{(k)})$.

  (ii) Assume $\mathrm{q}(c_i^{(k)}, X_i^{(k)}) \leq \beta \mathrm{w}(X_i^{(k)})$. Then $\mathrm{w}(X_i^{(k+1)}) \leq \mathrm{q}(c_i^{(k)}, X_i^{(k)}) \leq \beta \mathrm{w}(X_i^{(k)})$. Hence $\hat{\mathrm{w}}(\boldsymbol{X}^{(k+1)}) \leq \beta \hat{\mathrm{w}}(\boldsymbol{X}^{(k)})$.

- Assume $c_i^{(k)} \in X_i^{(k+1)}$. Let

$$\boldsymbol{Y} = \mathrm{tight}(\boldsymbol{X}^{(k)}, G_i^{(k)}, i).$$

  According to Lemma 5.4.19 it holds that $X_i^{(k+1)} \subseteq Y_i$. As $0 \notin A_{ii}^{\mathfrak{m}}$ it holds that

$$Y_i - c_i^{(k)} \subseteq -\frac{f_i(\boldsymbol{c}^{(k)}) + \sum_{j \neq i} A_{ij}^{\mathfrak{m}} (X_j^{(k)} - c_j^{(k)})}{A_{ii}^{\mathfrak{m}}}.$$

  As $c_i^{(k)} \in Y_i$ it follows that

$$0 \in f_i(\boldsymbol{c}^{(k)}) + \sum_{j \neq i} A_{ij}^{\mathfrak{m}} (X_j^{(k)} - c_j^{(k)}).$$

  Further, as $c_j^{(k)} \in X_j^{(k)}$ for all j, obviously

$$0 \in \sum_{j \neq i} A_{ij}^{\mathfrak{m}} (X_j^{(k)} - c_j^{(k)}).$$

  Hence, by Lemma 5.8.5

$$\begin{aligned}
\mathrm{w}(Y_i) &\leq \mathrm{w}\left(\frac{f_i(\boldsymbol{c}^{(k)}) + \sum_{j \neq i} A_{ij}^{\mathfrak{m}} (X_j^{(k)} - c_j^{(k)})}{A_{ii}^{\mathfrak{m}}} \right) \\
&= \mathrm{w}\left(\frac{\sum_{j \neq i} A_{ij}^{\mathfrak{m}} (X_j^{(k)} - c_j^{(k)})}{A_{ii}^{\mathfrak{m}}} \right) \\
&= \mathrm{w}\left(\frac{\sum_{j \neq i} A_{ij}^{\mathfrak{m}} (X_j^{(k)} - c_j^{(k)})}{\mathrm{mig}(A_{ii}^{\mathfrak{m}})} \right) \\
&= \sum_{j \neq i} \mathrm{w}(A_{ij}^{\mathfrak{m}} (X_j^{(k)} - c_j^{(k)}))/\mathrm{mig}(A_{ii}^{\mathfrak{m}}). \qquad (5.8.3)
\end{aligned}$$

  (i) Assume $\alpha < 1/2$. As $c_j^{(k)} \in X_j^{(k)}$ for all j it holds that

$$\begin{aligned}
\mathrm{w}(Y_i) &\leq \sum_{j \neq i} 2\mathrm{mag}(A_{ij}^{\mathfrak{m}}) \mathrm{w}(X_j^{(k)})/\mathrm{mig}(A_{ii}^{\mathfrak{m}}) \\
&\leq \sum_{j \neq i} 2\mathrm{mag}(A_{ij}^{\mathfrak{m}})/\mathrm{mig}(A_{ii}^{\mathfrak{m}})\ \hat{\mathrm{w}}(\boldsymbol{X}^{(k)}) \\
&= 2\alpha_i \hat{\mathrm{w}}(\boldsymbol{X}^{(k)}) \\
&< \hat{\mathrm{w}}(\boldsymbol{X}^{(k)}).
\end{aligned}$$

  Hence, $\hat{\mathrm{w}}(\boldsymbol{X}^{(k+1)}) < \hat{\mathrm{w}}(\boldsymbol{X}^{(k)})$.

  (ii) Assume $\mathrm{q}(c_j^{(k)}, X_j^{(k)}) \leq \beta \mathrm{w}(X_j^{(k)})$ for all j and $\alpha\beta < 1/2$. According to Lemma 5.8.6

$$\begin{aligned}
\mathrm{w}(A_{ij}^{\mathfrak{m}} (X_j^{(k)} - c_j^{(k)})) &\leq 2\mathrm{mag}(A_{ij}^{\mathfrak{m}}) \mathrm{q}(X_j^{(k)}, c_j^{(k)}) \\
&\leq 2\beta \mathrm{mag}(A_{ij}^{\mathfrak{m}}) \mathrm{w}(X_j^{(k)})
\end{aligned}$$

for all j, and thus

$$\begin{aligned} \mathrm{w}(Y_i) &\leq \sum_{j \neq i} 2\beta \mathrm{mag}(A_{ij}^{\mathfrak{m}})\mathrm{w}(X_j^{(k)})/\mathrm{mig}(A_{ii}^{\mathfrak{m}}) \\ &\leq \sum_{j \neq i} 2\beta \mathrm{mag}(A_{ij}^{\mathfrak{m}})/\mathrm{mig}(A_{ii}^{\mathfrak{m}})\ \hat{\mathrm{w}}(\boldsymbol{X}^{(k)}) \\ &= 2\alpha_i\beta\hat{\mathrm{w}}(\boldsymbol{X}^{(k)}). \end{aligned}$$

Hence, $\hat{\mathrm{w}}(\boldsymbol{X}^{(k+1)}) \leq 2\alpha\beta\hat{\mathrm{w}}(\boldsymbol{X}^{(k)})$. □

Corollary 5.8.8 *Let α as in Theorem 5.8.7. If $\boldsymbol{c}^{(k)} = \mathrm{mid}(\boldsymbol{X}^{(k)})$ for all k and $\alpha < 1$ then $\langle \boldsymbol{X}^{(k)} \mid k = 1, 2, \ldots \rangle$ is convergent and*

$$\hat{\mathrm{w}}(\boldsymbol{X}^{(k+1)}) \leq \max(\alpha, 1/2)\hat{\mathrm{w}}(\boldsymbol{X}^{(k)}). \ \square$$

The following Theorem is known for the Krawczyk operator [Qi, 1982], but is new for the Hansen–Sengupta operator. It states that if a box is mapped into its interior by the Hansen–Sengupta operator, then further application of the Hansen–Sengupta operator converges.

Theorem 5.8.9 (Convergence of Hansen–Sengupta Operator) *Assume $\emptyset \neq \boldsymbol{X}^{(2)} \subseteq \mathrm{int}(\boldsymbol{X}^{(1)})$ and $\boldsymbol{c}^{(k)} = \mathrm{mid}(\boldsymbol{X}^{(k)})$ for all $k = 1, 2, \ldots$. Then $\langle \boldsymbol{X}^{(k)} \mid k = 1, 2, \ldots \rangle$ is convergent and*

$$\mathrm{w}(X_i^{(k)}) \leq \max(\alpha, 1/2)^{k-1}\mathrm{w}(X_i^{(1)})$$

for all i, where

$$\alpha = \max_i \mathrm{w}(X_i^{(2)})/\mathrm{w}(X_i^{(1)}). \ \square$$

Proof. Assume $\emptyset \neq \boldsymbol{X}^{(2)} \subseteq \mathrm{int}(\boldsymbol{X}^{(1)})$ and $\boldsymbol{c}^{(k)} = \mathrm{mid}(\boldsymbol{X}^{(k)})$ for all $k = 1, 2, \ldots$. According to Theorem 5.6.5 (i) and (iii), $\boldsymbol{X}^{(k)} \neq \emptyset$ for all k. From Lemma 5.6.2 it follows that $0 \notin A_{ii}^{\mathfrak{m}}$ for all i. Let α_i such that

$$\mathrm{w}(X_i^{(2)}) = \alpha_i \mathrm{w}(X_i^{(1)}).$$

As $\boldsymbol{X}^{(2)} \subseteq \mathrm{int}(\boldsymbol{X}^{(1)})$ it holds that $\alpha_i < 1$. By induction on k we show that

$$\mathrm{w}(X_i^{(k)}) \leq \max(\alpha, 1/2)^{k-1}\mathrm{w}(X_i^{(1)}) \quad \text{for all} \quad i. \tag{5.8.4}$$

Obviously (5.8.4) holds for $k = 2$. Assume (5.8.4) holds for some $k \geq 2$ and let i arbitrary but fixed.

- If $c_i^{(k)} \notin X_i^{(k+1)}$ then $\mathrm{w}(X_i^{(k+1)}) \leq 1/2\mathrm{w}(X_i^{(k)})$ and by induction $\mathrm{w}(X_i^{(k+1)}) \leq \max(\alpha, 1/2)^k \mathrm{w}(X_i^{(1)})$.
- Assume $c_i^{(k)} \in X_i^{(k+1)}$. From (5.8.3) it follows that

$$\mathrm{mig}(A_{ii}^{\mathfrak{m}})\mathrm{w}(X_i^{(k+1)}) \leq \sum_{j \neq i} \mathrm{w}(A_{ij}^{\mathfrak{m}}(X_j^{(k)} - c_j^{(k)})).$$

By induction hypothesis and the assumption $\boldsymbol{c}^{(k)} = \mathrm{mid}(\boldsymbol{X}^{(k)})$ it holds that

$$(X_j^{(k)} - c_j^{(k)}) \subseteq \max(\alpha, 1/2)^{k-1}(X_j^{(1)} - c_j^{(1)}).$$

Hence,

$$\mathrm{mig}(A_{ii}^{\mathfrak{m}})\mathrm{w}(X_i^{(k+1)}) \leq \max(\alpha, 1/2)^{k-1} \sum_{j \neq i} \mathrm{w}(A_{ij}^{\mathfrak{m}}(X_j^{(1)} - c_j^{(1)})).$$

Now, as $0 \in \sum_{j \neq i} A_{ij}^{\mathrm{m}}(X_j^{(1)} - c_j^{(1)})$, it follows from Lemma 5.8.5 that

$$\begin{aligned}
&\sum_{j \neq i} \mathrm{w}(A_{ij}^{\mathrm{m}}(X_j^{(1)} - c_j^{(1)}))/\mathrm{mig}(A_{ii}^{\mathrm{m}}) \\
&= \mathrm{w}\left(\frac{\sum_{j \neq i} A_{ij}^{\mathrm{m}}(X_j^{(1)} - c_j^{(1)})}{A_{ii}^{\mathrm{m}}}\right) \\
&\leq \mathrm{w}\left(\frac{f_i(\boldsymbol{c}^{(1)}) + \sum_{j \neq i} A_{ij}^{\mathrm{m}}(X_j^{(1)} - c_j^{(1)})}{A_{ii}^{\mathrm{m}}}\right) \\
&= \mathrm{w}(X_i^{(2)}) \\
&\leq \alpha \mathrm{w}(X_i^{(1)}).
\end{aligned}$$

Thus,

$$\mathrm{w}(X_i^{(k+1)}) \leq \max(\alpha, 1/2)^k \mathrm{w}(X_i^{(1)}). \ \square$$

5.9 Termination

Let $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^n$ be a continuous functions with finitely many solutions in $\mathbb{R}^n$ and let $\boldsymbol{B} \in \mathbb{IR}^n$. We want to find isolating boxes for all solutions of $\boldsymbol{f}$ in $\boldsymbol{B}$. In the previous sections we studied algorithms for testing existence, uniqueness and non–existence of solutions of $\boldsymbol{f}$ in a box. As the tests may fail, it is not clear whether an algorithm based on these tests will terminate. In this section we give necessary conditions for termination and a terminating algorithm. The content of this section is new.

In the following let $\mathcal{E}$, $\mathcal{N}$ be predicates on $\mathbb{IR}^n$ such that if $\mathcal{E}(\boldsymbol{X}) = \texttt{true}$ then $\boldsymbol{f}$ has a unique solution in $\boldsymbol{X}$ and if $\mathcal{N}(\boldsymbol{X}) = \texttt{true}$ then $\boldsymbol{f}$ has no solution in $\boldsymbol{X}$. Note that the Hansen–Sengupta operator, the linearized tightening operator or simply a function which returns always `false` can be used to implement $\mathcal{E}$ and $\mathcal{N}$. The following algorithm is a skeleton for many interval methods for solving systems of nonlinear equations. From now on we will not take care of problems arising from finite precision arithmetic any more. However, the results still hold in the case of finite precision arithmetic, if the accuracy is "high enough".

Algorithm 5.9.1 [Isolating Boxes]

In: $\boldsymbol{f} \in \mathbb{R}^n \to \mathbb{R}^n$ continuous with finitely many solutions,
$\boldsymbol{B} \in \mathbb{IR}^n$.

Out: *Solutions*, a list of sub–boxes of $\boldsymbol{B}$ such that each $\boldsymbol{X}$ in *Solutions* contains a unique solution of $\boldsymbol{f}$ in $\boldsymbol{B}$ and each solution of $\boldsymbol{f}$ in $\boldsymbol{B}$ is contained in some $\boldsymbol{X}$ in *Solutions*.

(1) [Initialize.]
$Work \longleftarrow \{\boldsymbol{B}\}$, $Solutions \longleftarrow$ empty list.

(2) [Terminate.]
if $Work =$ empty list, then return *Solutions*.

(3) [Get box to work on.]
choose and remove an element $\boldsymbol{X}$ from $Work$.

(4) [Test unique existence.]
if $\mathcal{E}(\boldsymbol{X}) = \texttt{true}$ then add $\boldsymbol{X}$ to *Solutions*, goto Step 2.

(5) [Test non–existence.]
if $\mathcal{N}(\boldsymbol{X}) = \texttt{true}$ then goto Step 2.

(6) [Bisect.]
bisect $\boldsymbol{X}$ in the midpoint of its component with largest width, add both halves to $Work$ and goto Step 2.

Algorithm 5.9.1 is correct but need not terminate. For example, if $\mathcal{E}(\boldsymbol{X}) = \mathcal{N}(\boldsymbol{X}) = \texttt{false}$ for all $\boldsymbol{X} \in \mathbb{IR}^n$, then Algorithm 5.9.1 will not terminate. So, we impose further conditions on $\mathcal{E}$ and $\mathcal{N}$ which guarantee termination.

Theorem 5.9.2 (Termination)

- *Assume there exists $\varepsilon > 0$ such that if $\boldsymbol{X} \subseteq \boldsymbol{B}$, $\hat{\mathrm{w}}(\boldsymbol{X}) \leq \varepsilon$ and $\boldsymbol{X}$ contains a unique solution of $\boldsymbol{f}$, then $\mathcal{E}(\boldsymbol{X}) = \texttt{true}$.*
- *Assume there exists $\lambda > 0$ such that for all $\boldsymbol{X} \subseteq \boldsymbol{B}$ it holds that*

$$\min\{|f_i(\boldsymbol{x})| \mid \boldsymbol{x} \in \boldsymbol{X}\} > \lambda \hat{\mathrm{w}}(\boldsymbol{X}) \quad \text{for some} \quad i \text{ implies } \mathcal{N}(\boldsymbol{X}) = \texttt{true}.$$

Then Algorithm 5.9.1 terminates. □

Proof. Assume $\mathcal{E}$ and $\mathcal{N}$ satisfy the conditions of Theorem 5.9.2 for some $\varepsilon > 0$, $\lambda > 0$ but Algorithm 5.9.1 does not terminate. According to König's Lemma there exists a sequence $\langle \boldsymbol{X}^{(k)} \in \mathbb{IR}^n \mid k = 1, 2, \ldots \rangle$, such that $\boldsymbol{X}^{(k+1)} \subseteq \boldsymbol{X}^{(k)}$, $\lim_{k\to\infty} \hat{\mathrm{w}}(\boldsymbol{X}^{(k)}) = 0$ and $\mathcal{E}(\boldsymbol{X}^{(k)}) = \mathcal{N}(\boldsymbol{X}^{(k)}) = \texttt{false}$. Let $\hat{k} > 0$ such that $\hat{\mathrm{w}}(\boldsymbol{X}^{(\hat{k})}) \leq \varepsilon$. From the condition on $\mathcal{E}$ it follows that $\boldsymbol{X}^{(k)}$ does not contain a solution of $\boldsymbol{f}$ for all $k \geq \hat{k}$. According to Theorem 1.2.8 $\boldsymbol{x}^* = \lim_{k\to\infty} \boldsymbol{X}^{(k)}$ exists. As $\boldsymbol{x}^* \in \boldsymbol{X}^{(k)}$ for all k, it holds that $\boldsymbol{f}(\boldsymbol{x}^*) \neq 0$. Hence, there exists i such that $f_i(\boldsymbol{x}^*) \neq 0$ and from the continuity of f_i it follows that $f_i(\boldsymbol{x}) \neq 0$ for all $\boldsymbol{x}$ in a neighborhood $\mathcal{X}$ of $\boldsymbol{x}^*$. Let $\tilde{k} > 0$ such that $\boldsymbol{X}^{(\tilde{k})} \subseteq \mathcal{X}$. As $\boldsymbol{X}^{(\tilde{k})}$ is a closed subset of $\mathbb{R}^n$ and f_i is continuous, there exists $\delta > 0$ such that $|f_i(\boldsymbol{x})| \geq \delta$ for all $\boldsymbol{x} \in \boldsymbol{X}^{(\tilde{k})}$. Now, let $\tilde{k}' \geq \tilde{k}$ such that $\hat{\mathrm{w}}(\boldsymbol{X}^{(\tilde{k}')}) < \delta/\lambda$. As $\boldsymbol{X}^{(\tilde{k}')} \subseteq \boldsymbol{X}^{(\tilde{k})}$ it holds that

$$\begin{aligned} \min\{|f_i(\boldsymbol{x})| \mid \boldsymbol{x} \in \boldsymbol{X}^{(\tilde{k}')}\} &\geq \min\{|f_i(\boldsymbol{x})| \mid \boldsymbol{x} \in \boldsymbol{X}^{(\tilde{k})}\} \\ &\geq \delta \\ &> \lambda \hat{\mathrm{w}}(\boldsymbol{X}^{(\tilde{k}')}). \end{aligned}$$

By the condition on $\mathcal{N}$, it holds that $\mathcal{N}(\boldsymbol{X}^{(\tilde{k}')}) = \texttt{true}$, which is a contradiction to the assumption $\mathcal{N}(\boldsymbol{X}^{(k)}) = \texttt{false}$ for all k. □

As a next step towards an algorithm for finding isolating boxes for the solutions of systems of nonlinear equations, we try to use interval methods for computing $\mathcal{E}$, $\mathcal{N}$. For $\mathcal{N}$ this is easy, all we need is an interval extension $\boldsymbol{F}$ of $\boldsymbol{f}$, which is Lipschitz in $\boldsymbol{B}$.

Theorem 5.9.3 *Let $\boldsymbol{F}$ be an interval extension of $\boldsymbol{f}$, which is Lipschitz in $\boldsymbol{B}$, i.e. there exists $\lambda > 0$ such that*

$$\hat{\mathrm{w}}(\boldsymbol{F}(\boldsymbol{X})) \leq \lambda \hat{\mathrm{w}}(\boldsymbol{X}) \quad \text{for all} \quad \boldsymbol{X} \subseteq \boldsymbol{B}.$$

Let $\mathcal{N}$ be defined as

$$\mathcal{N}(\boldsymbol{X}) = \begin{cases} \texttt{true} & \text{if } 0 \notin \boldsymbol{F}(\boldsymbol{X}) \\ \texttt{false} & \text{else.} \end{cases}$$

Then the following holds:

(i) *Let $\boldsymbol{X} \in \mathbb{IR}^n$. If $\mathcal{N}(\boldsymbol{X}) = \texttt{true}$ then $\boldsymbol{f}$ has no solution in $\boldsymbol{X}$.*

(ii) *If $\boldsymbol{X} \subseteq \boldsymbol{B}$ and*

$$\min\{|f_i(\boldsymbol{x})| \mid \boldsymbol{x} \in \boldsymbol{X}\} > \lambda \hat{\mathrm{w}}(\boldsymbol{X})\} \quad \text{for some} \quad i$$

then $\mathcal{N}(\boldsymbol{X}) = \texttt{true}$. □

Proof. Let $\boldsymbol{F}$, λ and $\mathcal{N}$ as in Theorem 5.9.3.

(i) Let $\boldsymbol{X} \in \mathbb{IR}^n$ arbitrary but fixed. If $\mathcal{N}(\boldsymbol{X}) = \texttt{true}$ then $0 \notin \boldsymbol{F}(\boldsymbol{X})$, hence $\boldsymbol{f}(\boldsymbol{x}) \neq 0$ for all $\boldsymbol{x} \in \boldsymbol{X}$.

(ii) Let $\boldsymbol{X} \subseteq \boldsymbol{B}$ and assume $y = \min\{|f_i(\boldsymbol{x})| \mid \boldsymbol{x} \in \boldsymbol{X}\} > \lambda \hat{\mathrm{w}}(\boldsymbol{X})$ for some i. From the Lipschitz property of $\boldsymbol{F}$ it follows that $y > \hat{\mathrm{w}}(\boldsymbol{F}(\boldsymbol{X}))$, hence $y > \mathrm{w}(F_i(\boldsymbol{X}))$. As $y \leq \mathrm{mag}(F_i(\boldsymbol{X}))$ it follows that $\mathrm{mag}(F_i(\boldsymbol{X})) > \mathrm{w}(F_i(\boldsymbol{X}))$. Therefore $0 \notin F_i(\boldsymbol{X})$ and $0 \notin \boldsymbol{F}(\boldsymbol{X})$. Hence, $\mathcal{N}(\boldsymbol{X}) = \texttt{true}$. □

Using Theorem 5.6.5 (iv) one could define $\mathcal{E}$ for example as

$$\mathcal{E}(\boldsymbol{X}) = \begin{cases} \texttt{true} & \text{if } \mathcal{HS}(\boldsymbol{f}, \boldsymbol{X}, \boldsymbol{c}, \mathfrak{A}, \mathfrak{m}) \subseteq \mathrm{int}(\boldsymbol{X}) \\ \texttt{false} & \text{else,} \end{cases}$$

where $\boldsymbol{f}(\boldsymbol{c}) + \mathfrak{A}(\boldsymbol{X} - \boldsymbol{c})$ is a linearization of $\boldsymbol{f}$ in $\boldsymbol{X}$ and $\mathfrak{A}$ is a Jacobian of $\boldsymbol{f}$ in $\boldsymbol{X}$. But no matter what restrictions we impose on the choice of $\boldsymbol{c}$, $\mathfrak{A}$, $\mathfrak{m}$, the condition of Theorem 5.9.2 on $\mathcal{E}$ will not be satisfied for the following two reasons:

- If $\boldsymbol{f}$ has a multiple solution in $\boldsymbol{X}$, i.e. there exists $\hat{x} \in \boldsymbol{X}$ such that $\boldsymbol{f}(\hat{\boldsymbol{x}}) = 0$ and $\boldsymbol{f}'(\hat{\boldsymbol{x}})$ is singular, then $\mathfrak{A}$ and $\mathfrak{A}^{\mathfrak{m}}$ are singular and $\mathcal{HS}(\boldsymbol{f}, \boldsymbol{X}, \boldsymbol{c}, \mathfrak{A}, \mathfrak{m}) \not\subseteq \mathrm{int}(\boldsymbol{X})$ for any choice of $\boldsymbol{c}$ and $\mathfrak{m}$.
- If $\boldsymbol{f}(\boldsymbol{x}) = 0$ for some $\boldsymbol{x} \in \partial(\boldsymbol{X})$ then by Theorem 5.6.5 (i), $\mathcal{HS}(\boldsymbol{f}, \boldsymbol{X}, \boldsymbol{c}, \mathfrak{A}, \mathfrak{m}) \not\subseteq \mathrm{int}(\boldsymbol{X})$ for any choice of $\boldsymbol{c}, \mathfrak{A}, \mathfrak{m}$.

From now on we assume that $\boldsymbol{f}$ has no multiple solutions in $\boldsymbol{B}$. In the next section we give a terminating algorithm for finding isolating boxes of the solutions of $\boldsymbol{f}$, which solves the problem of solutions at the boundary of boxes.

5.9.1 Solutions at the Boundary of Boxes

The key idea to solve the problem with solutions of $\boldsymbol{f}$ at the boundary of $\boldsymbol{X}$ is as follows. First, enlarge $\boldsymbol{X}$ in each direction obtaining a box $\tilde{\boldsymbol{X}}$ such that $\boldsymbol{X} \subseteq \mathrm{int}(\tilde{\boldsymbol{X}})$. Let $\mathfrak{A}$ be a Jacobian of $\boldsymbol{f}$ in $\tilde{\boldsymbol{X}}$ and assume $\langle \boldsymbol{X}^{(k)} \mid k = 1, 2, \ldots \rangle$ converges, where $\boldsymbol{X}^{(1)} = \tilde{\boldsymbol{X}}$, $\boldsymbol{X}^{(k+1)} = \mathcal{HS}(\boldsymbol{f}, \boldsymbol{X}^{(k)}, \boldsymbol{c}^{(k)}, \mathfrak{A}, \mathfrak{m})$ for some $\mathfrak{m} \in \mathbb{R}^{n \times n}$ and $\boldsymbol{c}^{(k)} \in \boldsymbol{X}^{(k)}$. Then there exists $\hat{k}$ such that $\boldsymbol{X}^{(\hat{k})}$ satisfies one of the following two conditions:

- $\boldsymbol{X}^{(\hat{k})} \cap \boldsymbol{X} = \emptyset$. In this case $\boldsymbol{f}$ has no solution in $\boldsymbol{X}$.
- $\boldsymbol{X}^{(\hat{k})} \subseteq \mathrm{int}(\tilde{\boldsymbol{X}})$. According to Theorem 5.7.13, $\boldsymbol{f}$ has a unique solution in $\boldsymbol{X}^{(\hat{k})}$ if $A^{\mathfrak{m}}_{ii} > 0$ for all i. Note that $\boldsymbol{X}^{(\hat{k})}$ need not be a subset of $\boldsymbol{X}$, hence we do not know whether $\boldsymbol{f}$ has a solution in $\boldsymbol{X}$ or not.

In the following let $\boldsymbol{F}'$ be an interval extension of the Jacobian of $\boldsymbol{f}$ and let $\mathcal{C}$ be a predicate on $\mathbb{IR}^n$ such that if $\mathcal{C}(\boldsymbol{X}) = \texttt{true}$ then

- the sequence $\langle \boldsymbol{X}^{(k)} \mid k = 1, 2, \ldots \rangle$, where $\boldsymbol{X}^{(1)} = \boldsymbol{X}$, $\boldsymbol{X}^{(k+1)} = \mathcal{HS}(\boldsymbol{f}, \boldsymbol{X}^{(k)}, \boldsymbol{c}^{(k)}, \mathfrak{A}, \mathfrak{m})$ converges,
- $A^{\mathfrak{m}}_{ii} \geq 0$ for all i, and
- $\mathrm{mid}(\mathfrak{A})$ is regular

where $\boldsymbol{c}^{(k)} = \mathrm{mid}(\boldsymbol{X}^{(k)})$, $\mathfrak{A} = \boldsymbol{F}'(\boldsymbol{X})$ and $\mathfrak{m} = \mathrm{mid}(\mathfrak{A})^{-1}$. Further, let $\gamma \in \mathbb{R}$, $\gamma > 0$ be a constant, which determines the amount by which $\tilde{\boldsymbol{X}}$ is larger than $\boldsymbol{X}$. Now, we modify Algorithm 5.9.1 as follows:

Algorithm 5.9.4 [Isolating Boxes]

In: $\boldsymbol{f} \in \mathbb{R}^n \to \mathbb{R}^n$ continuous with finitely many solutions, all of them are simple,
$\boldsymbol{B} \in \mathbb{IR}^n$, $\hat{\mathrm{w}}(\boldsymbol{B}) > 0$.

Out: *Solutions*, a list of sub–boxes of $\boldsymbol{B}$ such that every $\boldsymbol{X}$ in *Solutions* contains a unique solution of $\boldsymbol{f}$ and every solution of $\boldsymbol{f}$ in $\boldsymbol{B}$ is contained in some $\boldsymbol{X}$ in *Solutions*.

(1) [Initialize.]
$Work \longleftarrow \{\boldsymbol{B}\}$, $Solutions \longleftarrow$ empty list.

(2) [Terminate.]
if $Work =$ empty list, then return $Solutions$.

(3) [Get box to work on.]
chose and remove an element $\boldsymbol{X}$ from $Work$.

(4) [Test unique existence.]
choose $\tilde{\boldsymbol{X}}$ such that $\boldsymbol{X} \subseteq \mathrm{int}(\tilde{\boldsymbol{X}})$ and $\tilde{\boldsymbol{X}} \subseteq \boldsymbol{X} + \gamma\hat{\mathrm{w}}(\boldsymbol{X})[-1,1]^n$.
$\mathfrak{A} \longleftarrow \boldsymbol{F}'(\tilde{\boldsymbol{X}})$, $\mathfrak{m} \longleftarrow \mathrm{mid}(\mathfrak{A})^{-1}$.
if $\mathcal{C}(\tilde{\boldsymbol{X}}) =$ `true`

(4.1) [Initialize Hansen–Sengupta iteration.]
$k \longleftarrow 1$, $\boldsymbol{X}^{(1)} \longleftarrow \tilde{\boldsymbol{X}}$.

(4.2) [Test.]
if $\boldsymbol{X}^{(k)} \cap \boldsymbol{X} = \emptyset$ goto Step 2.
if $\boldsymbol{X}^{(k)} \subseteq \mathrm{int}(\tilde{\boldsymbol{X}})$ then add $\boldsymbol{X}^{(k)}$ to $Solutions$ and goto Step 2.

(4.3) [Hansen–Sengupta operator.]
$\boldsymbol{X}^{(k+1)} \longleftarrow \mathcal{HS}(\boldsymbol{f}, \boldsymbol{X}^{(k)}, \mathrm{mid}(\boldsymbol{X}^{(k)}), \mathfrak{A}, \mathfrak{m})$.

(4.4) [Iterate.]
$k \longleftarrow k+1$, goto Step 4.2.

(5) [Test non–existence.]
if $\mathcal{N}(\boldsymbol{X}) =$ `true` then goto Step 2.

(6) [Bisect.]
bisect $\boldsymbol{X}$ in the midpoint of the component with largest width, add both halves to $Work$ and goto Step 2.

We modify Theorem 5.9.2 such that it gives sufficient conditions for the termination of Algorithm 5.9.4. Let $\tilde{\boldsymbol{B}} = \boldsymbol{B} + \gamma\hat{\mathrm{w}}(\boldsymbol{B})[-1,1]^n$.

Theorem 5.9.5 (Termination)

- *Assume there exists $\varepsilon > 0$ such that if $\boldsymbol{X} \subseteq \tilde{\boldsymbol{B}}$, $\hat{\mathrm{w}}(\boldsymbol{X}) \leq \varepsilon$ and $\boldsymbol{X}$ contains a unique solution of $\boldsymbol{f}$, then $\mathcal{C}(\boldsymbol{X}) =$* `true`.
- *Assume there exists $\lambda > 0$ such that for all $\boldsymbol{X} \subseteq \boldsymbol{B}$ it holds that*

$$\min\{|f_i(\boldsymbol{x})| \mid \boldsymbol{x} \in \boldsymbol{X}\} > \lambda\hat{\mathrm{w}}(\boldsymbol{X}) \quad \text{for some} \quad i \text{ implies } \mathcal{N}(\boldsymbol{X}) = \texttt{true}.$$

Then Algorithm 5.9.4 terminates. □

Proof. The proof is similar to the proof of Theorem 5.9.2. First, we show that the loop in Step 4 terminates, i.e. there exists k such that either $\boldsymbol{X}^{(k)} \cap \boldsymbol{X} = \emptyset$ or $\boldsymbol{X}^{(k)} \subseteq \mathrm{int}(\tilde{\boldsymbol{X}})$. As $\mathcal{C}(\tilde{\boldsymbol{X}}) =$ `true`, it holds that $\boldsymbol{X}^{(k)}$ converges and there exists $\tilde{k}$ such that

$$\hat{\mathrm{w}}(\boldsymbol{X}^{(\tilde{k})}) < \min_i \min\{|\overline{X}_i - \overline{\tilde{X}}_i|, |\underline{X}_i - \underline{\tilde{X}}_i|\}.$$

Assume $\boldsymbol{X}^{(\tilde{k})} \not\subseteq \mathrm{int}(\tilde{\boldsymbol{X}})$. As $\boldsymbol{X}^{(\tilde{k})} \subseteq \tilde{\boldsymbol{X}}$, there exists i such that $\overline{X}_i^{(\tilde{k})} = \overline{\tilde{X}}_i$ or $\underline{X}_i^{(\tilde{k})} = \underline{\tilde{X}}_i$. Assume $\overline{X}_i^{(\tilde{k})} = \overline{\tilde{X}}_i$. Then $\underline{X}_i^{(\tilde{k})} > \overline{X}_i$, hence $\boldsymbol{X}^{(\tilde{k})} \cap \boldsymbol{X} = \emptyset$ and the loop in Step 4 terminates. The case $\underline{X}_i^{(\tilde{k})} = \underline{\tilde{X}}_i$ is analogous.

Assume $\mathcal{C}$ and $\mathcal{N}$ satisfy the conditions of Theorem 5.9.5 for some $\varepsilon > 0$, $\lambda > 0$ but Algorithm 5.9.4 does not terminate. According to König's Lemma there exists a sequence $\langle \boldsymbol{X}^{(k)} \in \mathbb{IR}^n \mid k = 1, 2, \ldots \rangle$, such

that $\boldsymbol{X}^{(k+1)} \subseteq \boldsymbol{X}^{(k)}$, $\lim_{k\to\infty} \hat{\mathrm{w}}(\boldsymbol{X}^{(k)}) = 0$ and $\mathcal{C}(\tilde{\boldsymbol{X}}^{(k)}) = \mathcal{N}(\boldsymbol{X}^{(k)}) = \texttt{false}$ where $\boldsymbol{X}^{(k)} \subseteq \mathrm{int}(\tilde{\boldsymbol{X}}^{(k)})$ and $\tilde{\boldsymbol{X}}^{(k)} \subseteq \boldsymbol{X}^{(k)} + \gamma\hat{\mathrm{w}}(\boldsymbol{X}^{(k)})[-1,1]^n$. Hence, there exists $\hat{k} > 0$ such that $\hat{\mathrm{w}}(\tilde{\boldsymbol{X}}^{(\hat{k})}) \leq \varepsilon$. From the condition on $\mathcal{C}$ it follows that $\boldsymbol{X}^{(\hat{k})}$ and hence $\boldsymbol{X}^{(k)}$ does not contain a solution of $\boldsymbol{f}$ for all $k \geq \hat{k}$. The rest is the same as in the proof of Theorem 5.9.2. □

Now, we can use any of the convergence theorems of Section 5.8 to obtain an algorithm for $\mathcal{C}$, which satisfies the conditions of Theorem 5.9.5. For simplicity, we use Theorem 5.8.4.

Theorem 5.9.6 *Let $\mathcal{C}$ be defined as*

$$\mathcal{C}(\boldsymbol{X}) = \begin{cases} \texttt{true} & \text{if } \alpha < 1/2, A_{ii}^{\mathfrak{m}} \geq 0 \text{ for all } i \text{ and } \mathrm{mid}(\mathfrak{A}) \text{ is regular} \\ \texttt{false} & \text{else,} \end{cases}$$

for all $\boldsymbol{X} \subseteq \tilde{\boldsymbol{B}}$, where α is as in Theorem 5.8.4, $\mathfrak{A} = \boldsymbol{F}'(\boldsymbol{X})$ for some continuous Lipschitz interval extension $\boldsymbol{F}'$ of $\boldsymbol{f}'$ in $\tilde{\boldsymbol{B}}$ and $\mathfrak{m} = \mathrm{mid}(\mathfrak{A})^{-1}$ if $\mathrm{mid}(\mathfrak{A})$ is regular. Then there exists $\varepsilon > 0$ such that $\boldsymbol{X} \subseteq \tilde{\boldsymbol{B}}$, $\hat{\mathrm{w}}(\boldsymbol{X}) \leq \varepsilon$ and $\boldsymbol{X}$ contains a unique solution of $\boldsymbol{f}$ implies $\mathcal{C}(\boldsymbol{X}) = \texttt{true}$. □

Proof. Let $\alpha, \mathfrak{A}, \boldsymbol{F}', \mathfrak{m}$ as in Theorem 5.9.6. As $\boldsymbol{f}$ has no multiple solutions, there exists $\varepsilon' > 0$ such that $\boldsymbol{F}'(\boldsymbol{X})$ is regular for all $\boldsymbol{X} \in \mathcal{X}$ where

$$\mathcal{X} = \{\boldsymbol{X} \subseteq \tilde{\boldsymbol{B}} \mid \hat{\mathrm{w}}(\boldsymbol{X}) \leq \varepsilon', \boldsymbol{f}(\boldsymbol{x}) = 0 \text{ for some } \boldsymbol{x} \in \boldsymbol{X}\}.$$

As $\mathcal{X}$ is closed and $\boldsymbol{F}'$ is continuous, the set $\{\mathrm{mid}(\boldsymbol{F}'(\boldsymbol{X}))^{-1} \mid \boldsymbol{X} \in \mathcal{X}\}$ is component wise bounded, i.e. there exists $\hat{m}$ such that

$$\hat{m} \geq |m_{ij}| \text{ for all } i,j \text{ and } \mathfrak{m} \in \{\mathrm{mid}(\boldsymbol{F}'(\boldsymbol{X}))^{-1} \mid \boldsymbol{X} \in \mathcal{X}\}.$$

As $\boldsymbol{F}'$ is Lipschitz in $\tilde{\boldsymbol{B}}$, there exists $\lambda > 0$ such that $\mathrm{w}(\boldsymbol{F}'_{ij}(\boldsymbol{X})) \leq \lambda\hat{\mathrm{w}}(\boldsymbol{X})$ for all i,j and $\boldsymbol{X} \subseteq \tilde{\boldsymbol{B}}$. Hence, for all $\boldsymbol{X} \in \mathcal{X}$

$$\hat{\mathrm{w}}(\mathrm{mid}(\boldsymbol{F}'(\boldsymbol{X}))^{-1}\boldsymbol{F}'(\boldsymbol{X})) \leq n\hat{m}\lambda\hat{\mathrm{w}}(\boldsymbol{X}).$$

From $\mathrm{i} \in \mathrm{mid}(\boldsymbol{F}'(\boldsymbol{X}))^{-1}\boldsymbol{F}'(\boldsymbol{X})$ it follows that

$$\alpha \leq n^2\hat{m}\lambda\hat{\mathrm{w}}(\boldsymbol{X}).$$

Thus, let

$$\varepsilon < \min\left\{\frac{1}{2n^2\hat{m}\lambda}, \varepsilon'\right\}.$$

Now, if $\boldsymbol{X} \subseteq \tilde{\boldsymbol{B}}$, $\hat{\mathrm{w}}(\boldsymbol{X}) \leq \varepsilon$ and $\boldsymbol{X}$ contains a solution of $\boldsymbol{f}$, then $\alpha < 1/2$ and $\mathrm{mid}(\mathfrak{A})$ is regular. Further, $\alpha < 1/2$ implies $A_{ii}^{\mathfrak{m}} \geq 0$ for all i, hence $\mathcal{C}(\boldsymbol{X}) = \texttt{true}$. □

The output of Algorithm 5.9.4 is not precisely what one would expect from an algorithm for finding root isolating boxes. It is true that every $\boldsymbol{X}$ in *Solutions* contains a unique solution of $\boldsymbol{f}$ and that every solution of $\boldsymbol{f}$ in $\boldsymbol{B}$ is contained in some $\boldsymbol{X} \in$ *Solutions*. However, the elements of *Solution* need not be disjoint and it is even possible that there are two different elements $\boldsymbol{X}, \boldsymbol{Y}$ in *Solution* which contain the same unique solution of $\boldsymbol{f}$. The reason is that in Step 4 of Algorithm 5.9.4 the box was enlarged and it can happen that the enlarged box contains a solution but the original box does not or has a solution on its boundary. However, this deficiency can easily be repaired:

Let $\boldsymbol{X}, \boldsymbol{Y} \in$ *Solutions* such that $\boldsymbol{X} \cap \boldsymbol{Y} \neq \emptyset$. Let $\langle \boldsymbol{X}^{(k)} \mid k = 1,2,\ldots\rangle$, $\langle \boldsymbol{Y}^{(k)} \mid k = 1,2,\ldots\rangle$ be sequences which are generated by repeated application of the Hansen–Sengupta operator starting with $\boldsymbol{X}$ respectively $\boldsymbol{Y}$ using the respective $\mathfrak{A}$, $\mathfrak{m}$ as in Algorithm 5.9.4, Step 4. Then both sequences are convergent and there exists $\hat{k}$ such that either

$$\boldsymbol{X}^{(\hat{k})} \cap \boldsymbol{Y}^{(\hat{k})} = \emptyset$$

or

$$\mathcal{C}([\boldsymbol{X}^{(\hat{k})} \cup \boldsymbol{Y}^{(\hat{k})}]) = \texttt{true}.$$

In the first case, $\boldsymbol{X}^{(\hat{k})}$ and $\boldsymbol{Y}^{(\hat{k})}$ each contain a unique solution of $\boldsymbol{f}$. In the second case, $\boldsymbol{X}^{(\hat{k})} \cup \boldsymbol{Y}^{(\hat{k})}$ contains a unique solution. By simultaneous computation of $\boldsymbol{X}^{(k)}$, $\boldsymbol{Y}^{(k)}$ for $k = 1,2,\ldots$ it is therefore possible to refine the list *Solutions* such that it has the same properties as stated in Algorithm 5.9.4 and all its elements are mutually disjoint.

5.10 Unbounded Start Regions for Polynomial Systems

So far we were searching for the solutions of a nonlinear system of equations $\boldsymbol{f}$ within a given search interval $\boldsymbol{B}$. In this section we tackle the problem of finding all solutions of $\boldsymbol{f}$ in $\mathbb{R}^n$ where $\boldsymbol{f}$ is a *polynomial* system. The method was invented by [Neumaier, 1990]. We give a more comprehensive description, point out the problem of solutions at infinity and generalize it to polynomial inequalities.

The idea of the algorithm is to decompose $\mathbb{R}^n$ into $n+1$ subsets $\mathcal{D}^{(\ell)}$, $\ell = 1, \ldots, n+1$ and map each subset to the interval $[-1,1]^n$. For each subset $\mathcal{D}^{(\ell)}$ the polynomial system $\boldsymbol{f}$ is transformed into a system $\boldsymbol{g}^{(\ell)}$ such that the roots of $\boldsymbol{f}$ in $\mathcal{D}^{(\ell)}$ correspond to the roots of $\boldsymbol{g}$ in $[-1,1]^n$. After having found the roots of $\boldsymbol{g}$ in $[-1,1]^n$ using a method for bounded start regions, we have to map them back to $\mathcal{D}^{(\ell)}$. First, we give a simple but more general theorem which forms the theoretical basis of the algorithm.

Let $\mathcal{D}$ be a set, let $\boldsymbol{p}$ be a predicate on $\mathcal{D}$ and let $\mathcal{D}^{(\ell)}$, $\ell = 1, \ldots, m$ be subsets of $\mathcal{D}$ such that

$$\bigcup_{\ell=1}^{m} \mathcal{D}^{(\ell)} \supseteq \{\boldsymbol{x} \in \mathcal{D} \mid \boldsymbol{p}(\boldsymbol{x})\}.$$

Let $\mathcal{E}^{(\ell)}$ be a set and let

$$\boldsymbol{t}^{(\ell)} : \mathcal{E}^{(\ell)} \to \mathcal{D}^{(\ell)}$$

be surjective for all ℓ. Further, let $\boldsymbol{q}^{(\ell)}$ be a predicate on $\mathcal{E}^{(\ell)}$, which is defined by

$$\boldsymbol{q}^{(\ell)}(\boldsymbol{y}) \text{ iff } \boldsymbol{p}(\boldsymbol{t}^{(\ell)}(\boldsymbol{y}))$$

for all $\boldsymbol{y} \in \mathcal{E}^{(\ell)}$.

Theorem 5.10.1 *For all $\boldsymbol{x} \in \mathcal{D}$ it holds that $\boldsymbol{p}(\boldsymbol{x})$ if and only if there exists $\ell \in \{1, \ldots, m\}$ and $\boldsymbol{y} \in \mathcal{E}^{(\ell)}$ such that $\boldsymbol{q}^{(\ell)}(\boldsymbol{y})$ and $\boldsymbol{t}^{(\ell)}(\boldsymbol{y}) = \boldsymbol{x}$.* □

Proof. Let $\boldsymbol{x} \in \mathcal{D}$ arbitrary but fixed.

"⇒" Assume $\boldsymbol{p}(\boldsymbol{x})$. Then $\boldsymbol{x} \in \mathcal{D}^{(\ell)}$ for some $\ell \in \{1, \ldots, m\}$. As $\boldsymbol{t}^{(\ell)}$ is surjective, there exists $\boldsymbol{y} \in \mathcal{E}^{(\ell)}$ such that $\boldsymbol{t}^{(\ell)}(\boldsymbol{y}) = \boldsymbol{x}$. Hence $\boldsymbol{q}^{(\ell)}(\boldsymbol{y})$.

"⇐" Let $\ell \in \{1, \ldots, m\}$ and $\boldsymbol{y} \in \mathcal{E}^{(\ell)}$ arbitrary but fixed such that $\boldsymbol{q}^{(\ell)}(\boldsymbol{y})$ and $\boldsymbol{t}^{(\ell)}(\boldsymbol{y}) = \boldsymbol{x}$. Then $\boldsymbol{p}(\boldsymbol{t}^{(\ell)}(\boldsymbol{y}))$ and hence $\boldsymbol{p}(\boldsymbol{x})$. □

In the following we apply Theorem 5.10.1 to the special case when $\mathcal{D} = \mathbb{R}^n$ and $\boldsymbol{p}(\boldsymbol{x}) \equiv \boldsymbol{f}(\boldsymbol{x}) = 0$, where $\boldsymbol{f}$ is a polynomial system $\mathbb{R}^n \to \mathbb{R}^n$ with finitely many solutions in $\mathbb{R}^n$. We make the following choices for $\mathcal{D}^{(\ell)}$, $\mathcal{E}^{(\ell)}$, $\boldsymbol{t}^{(\ell)}$ which are justified later: Let $m = n+1$, and for $\ell = 1, \ldots, n$ let

$$\begin{aligned}
\mathcal{D}^{(\ell)} &= \{\boldsymbol{x} \in \mathbb{R}^n \mid |x_\ell| \geq 1, |x_i| \leq |x_\ell| \text{ for } i \neq \ell\}, \\
\mathcal{E}^{(\ell)} &= \{\boldsymbol{y} \in [-1,1]^n \mid y_\ell \neq 0\}, \\
t_i^{(\ell)}(\boldsymbol{y}) &= \begin{cases} y_i/y_\ell & i \neq \ell \\ 1/y_\ell & i = \ell. \end{cases} \quad \text{for } i = 1, \ldots, n
\end{aligned}$$

and let

$$\begin{aligned}
\mathcal{D}^{(n+1)} &= [-1,1]^n \\
\mathcal{E}^{(n+1)} &= [-1,1]^n \\
\boldsymbol{t}^{(n+1)}(\boldsymbol{y}) &= \boldsymbol{y}.
\end{aligned}$$

One easily checks that $\boldsymbol{t}^{(\ell)} : \mathcal{E}^{(\ell)} \to \mathcal{D}^{(\ell)}$ is surjective and $\bigcup_{\ell=1}^{n+1} \mathcal{D}^{(\ell)} = \mathcal{D}$.

Let $\boldsymbol{g}^{(\ell)} : \mathcal{E}^{(\ell)} \to \mathbb{R}^n$ be defined as

$$\boldsymbol{g}^{(\ell)}(\boldsymbol{y}) = \boldsymbol{f}(\boldsymbol{t}^{(\ell)}(\boldsymbol{y})).$$

From Theorem 5.10.1 it follows that $\hat{\boldsymbol{x}}$ is a solution of $\boldsymbol{f}$ in $\mathbb{R}^n$ if and only if there exists $\ell \in \{1, \ldots, n\}$ and $\hat{\boldsymbol{y}} \in \mathcal{E}^{(\ell)}$ such that $\boldsymbol{g}^{(\ell)}(\hat{\boldsymbol{y}}) = 0$ and $\boldsymbol{t}^{(\ell)}(\hat{\boldsymbol{y}}) = \hat{\boldsymbol{x}}$. Thus, it suffices to find the solutions of $\boldsymbol{g}^{(\ell)}$ in the bounded set $\mathcal{E}^{(\ell)}$ for all ℓ. Before we go into details of solving this problem let us consider a simple example.

Example. Let $n = 2$ and let

$$\begin{aligned} f_1(\boldsymbol{x}) &= x_1^2 + x_2 - 1 \\ f_2(\boldsymbol{x}) &= x_1 x_2 - 1. \end{aligned}$$

The decomposition of $\mathbb{R}^2$ into $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3$ is displayed in Figure 5.10.1. We obtain

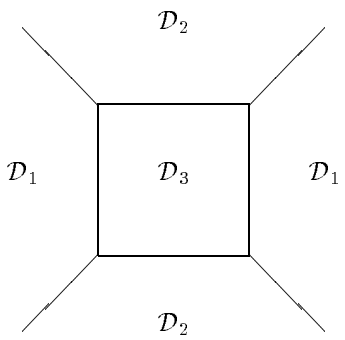


Figure 5.10.1: Decomposition of $\mathbb{R}^2$.

$$\begin{aligned} \boldsymbol{t}^{(1)}(\boldsymbol{y}) &= (1/y_1, y_2/y_1) \\ \boldsymbol{t}^{(2)}(\boldsymbol{y}) &= (y_1/y_2, 1/y_2) \\ \boldsymbol{t}^{(3)}(\boldsymbol{y}) &= (y_1, y_2) \end{aligned}$$

and thus

$$\begin{aligned} g_1^{(1)}(\boldsymbol{y}) &= f_1(\boldsymbol{t}^{(1)}(\boldsymbol{y})) = 1/y_1^2 + y_2/y_1 - 1 = 1/y_1^2(1 + y_1 y_2 - y_1^2) \\ g_2^{(1)}(\boldsymbol{y}) &= f_2(\boldsymbol{t}^{(1)}(\boldsymbol{y})) = 1/y_1 \; y_2/y_1 - 1 = 1/y_1^2(y_2 - y_1^2) \\ \\ g_1^{(2)}(\boldsymbol{y}) &= f_1(\boldsymbol{t}^{(2)}(\boldsymbol{y})) = y_1^2/y_2^2 + 1/y_2 - 1 = 1/y_2^2(y_1^2 + y_2 - y_2^2) \\ g_2^{(2)}(\boldsymbol{y}) &= f_2(\boldsymbol{t}^{(2)}(\boldsymbol{y})) = y_1/y_2 \; 1/y_2 - 1 = 1/y_2^2(y_1 - y_2^2). \\ \\ g_1^{(3)}(\boldsymbol{y}) &= f_1(\boldsymbol{t}^{(3)}(\boldsymbol{y})) = y_1^2 + y_2 - 1 \\ g_2^{(3)}(\boldsymbol{y}) &= f_2(\boldsymbol{t}^{(3)}(\boldsymbol{y})) = y_1 y_2 - 1 \end{aligned}$$

Applying interval methods directly for finding the solutions of $\boldsymbol{g}^{(\ell)}$ in $\mathcal{E}^{(\ell)}$ is not possible because of the singularity of $g_i^{(\ell)}$ at $y_\ell = 0$ and because $\mathcal{E}^{(\ell)}$ is not an interval for $\ell \neq n+1$. However, the special choice of $\boldsymbol{t}^{(\ell)}$ and the fact that $\boldsymbol{f}$ is a polynomial system allow us to write $g_i^{(\ell)}$ in the form

$$\begin{aligned} g_i^{(\ell)}(\boldsymbol{y}) &= 1/y_\ell^{d_i} \; \tilde{g}_i^{(\ell)}(\boldsymbol{y}) \\ g_i^{(n+1)}(\boldsymbol{y}) &= \tilde{g}_i^{(n+1)}(\boldsymbol{y}) \end{aligned}$$

where d_i is the total degree of f_i and $\tilde{g}_i^{(\ell)}$ is a polynomial. Obviously

$$g_i^{(\ell)}(\boldsymbol{y}) = 0 \text{ iff } \tilde{g}_i^{(\ell}(\boldsymbol{y}) = 0$$

for all $\boldsymbol{y} \in \mathcal{E}^{(\ell)}$. Note that $\tilde{g}_i^{\ell}$ has the same number of terms as f_i, which means that sparsity of the input system is preserved.

In the next step we solve the $\tilde{\boldsymbol{g}}^{(\ell)}$ with starting box $[-1,1]^n$. For each ℓ we obtain a set $\mathcal{A}^{(\ell)}$ of sub–intervals of $[-1,1]^n$ such that each solution of $\tilde{\boldsymbol{g}}^{(\ell)}$ is contained in some $\boldsymbol{A}^{(\ell)} \in \mathcal{A}^{(\ell)}$. In order to obtain a set $\mathcal{B}$ of interval vectors which enclose all solutions of $\boldsymbol{f}$ in $\mathbb{R}^n$ we have to transform the $\boldsymbol{A}^{(\ell)}$ back using $\boldsymbol{t}^{(\ell)}$. Note however, that $\boldsymbol{t}^{(\ell)}(\boldsymbol{A}^{(\ell)})$ is defined only if $\boldsymbol{A}^{(\ell)} \subseteq \mathcal{E}^{(\ell)}$, i.e.

$$0 \notin A_\ell^{(\ell)}.$$

If this is not the case, then extended interval arithmetic has to be used during the back transformation and for the corresponding solutions of $\boldsymbol{f}$ enclosures by unbounded intervals are obtained. This situation occurs whenever $\boldsymbol{f}$ has solutions at infinity.

In the following we extend the method described above to systems of polynomial inequalities

$$f_i(\boldsymbol{x}) \sim_i 0$$

where $\sim_i \in \{=, \neq, <, >, \leq, \geq\}$ for $i = 1, \ldots, r$

A straight forward generalization does not work because

$$g_i^{(l)}(\boldsymbol{y}) \sim_i 0 \text{ iff } \tilde{g}_i^{(l)}(\boldsymbol{y}) \sim_i 0 \text{ for all } \boldsymbol{y} \in \mathcal{E}^{(\ell)}$$

holds only if $\sim_i \in \{=, \neq\}$ or if d_i is even. The problem can be solved if $\mathbb{R}^n$ is decomposed into $2n+1$ subsets instead of only $n+1$ subsets as in the case of polynomial equations. Let

$$\begin{aligned}
\mathcal{D}^{(\ell)} &= \begin{cases} \{\boldsymbol{x} \in \mathbb{R}^n \mid x_\ell \geq 1, |x_i| \leq |x_\ell| \text{ for } i \neq \ell\} & \ell = 1, \ldots, n \\ \{\boldsymbol{x} \in \mathbb{R}^n \mid x_\ell \leq -1, |x_i| \leq |x_\ell| \text{ for } i \neq \ell\} & \ell = n+1, \ldots, 2n \end{cases} \\
\mathcal{E}^{(\ell)} &= \begin{cases} \{\boldsymbol{y} \in [-1,1]^n \mid y_\ell < 0\} & \ell = 1, \ldots, n \\ \{\boldsymbol{y} \in [-1,1]^n \mid y_\ell > 0\} & \ell = n+1, \ldots, 2n. \end{cases} \\
t_i^{(\ell)}(y_1, \ldots, y_n) &= \begin{cases} y_i/y_\ell & i \neq \ell \\ 1/y_\ell & i = \ell \end{cases} \quad \ell = 1, \ldots, 2n
\end{aligned}$$

and let

$$\begin{aligned}
\mathcal{D}^{(2n+1)} &= [-1,1]^n \\
\mathcal{E}^{(2n+1)} &= [-1,1]^n \\
\boldsymbol{t}^{(2n+1)}(\boldsymbol{y}) &= \boldsymbol{y}.
\end{aligned}$$

As before let $\boldsymbol{g}^{(\ell)}(\boldsymbol{y}) = \boldsymbol{f}\big(\boldsymbol{t}^{(\ell)}(\boldsymbol{y})\big)$. For $\ell = 1, \ldots, 2n$ let $\tilde{\boldsymbol{g}}^{(\ell)} : \mathbb{R}^n \to \mathbb{R}^n$ be polynomial systems such that

$$g_i^{(\ell)}(\boldsymbol{y}) = \begin{cases} \tilde{g}_i^{(\ell)}(\boldsymbol{y})/y_\ell^{d_i} & \text{if } \ell > n \text{ or } d_i \text{ is even} \\ -\tilde{g}_i^{(\ell)}(\boldsymbol{y})/y_\ell^{d_i} & \text{if } \ell \leq n \text{ and } d_i \text{ is odd} \end{cases}$$

and $\boldsymbol{g}^{(2n+1)}(\boldsymbol{y}) = \tilde{\boldsymbol{g}}^{(2n+1)}(\boldsymbol{y})$. Then $\tilde{g}_i^{(\ell)}(\boldsymbol{y}) \sim_i 0$ iff $g_i^{(\ell)}(\boldsymbol{y}) \sim_i 0$ for all $\boldsymbol{y} \in \mathcal{E}^{(\ell)}$. Back transformation of the solution intervals is done in the same way as for polynomial systems of equations.

5.11 Experimental Results

In this section we compare two algorithms for finding isolating boxes for all solutions of a given system of equations. The first algorithm, RHB (Range test, Hansen–Sengupta operator, Bisection) repeats the following operations until isolating boxes for all solutions of $\boldsymbol{f}$ are found, see [Hansen and Sengupta, 1981].

- **Range test:**
  Overestimate $\boldsymbol{f}(\boldsymbol{X})$ and test if it contains zero.
- **Hansen–Sengupta operator:**
  Replace a box $\boldsymbol{X}$ by the result of the Hansen–Sengupta operator applied to $\boldsymbol{X}$ using Algorithm 5.6.6. Check the existence, uniqueness and non–existence conditions of Theorem 5.6.5.

- **Bisection:**
  Bisect a box $\boldsymbol{X}$ in the midpoint of the largest width direction.

In the second algorithm, RLHB (Range test, Linearized tightening, Hansen–Sengupta operator, Bisection), we use in addition the linearized tightening operator:

- **Linearized tightening operator:**
  Replace a box $\boldsymbol{X}$ by the result of the linearized tightening operator applied to $\boldsymbol{X}$ using Algorithm 5.5.1. Check the existence, uniqueness and non–existence conditions of Theorem 5.5.3.

A few remarks are needed here:

- In the examples below we consider only polynomial equations. For the range test and the evaluation of the derivatives we use the linearly convergent Horner form.
- After application of the Hansen–Sengupta operator, a Jacobian of $\mathfrak{m}\,\boldsymbol{f}$ in some box $\boldsymbol{X}$ is available and we use Theorem 5.8.4 for testing convergence of the Hansen–Sengupta operator starting from $\boldsymbol{X}$. If the conditions of this Theorem are satisfied, we use the technique described in Algorithm 5.9.4 for avoiding non–termination if a solution of $\boldsymbol{f}$ is on the boundary of $\boldsymbol{X}$.
- If the hull division in a linear tightening step is replaced by a generalized division, then the result is sometimes the disjoint union of two intervals. In such a situation we memorize the gap between the intervals and instead of a bisection we split the box at the widest gap.
- If slopes are used instead of derivatives for the Hansen–Sengupta or for the linearized tightening operator, then the computed boxes are usually smaller but the uniqueness property as stated by Theorem 5.6.5 and Theorem 5.5.3 does no longer hold. Thus, one might use slopes until existence of a solution in a box is proved and from then on derivatives. However, in order to keep the algorithms simple, we use derivatives from the beginning.

Based on experimental observations, we found that the following strategy seems to work well:

(1) Range test.

(2) (Only for RLHB.) Apply successive linearized tightening. If the box is reduced "significantly" thereby, go back to (2).

(3) Next, apply Hansen–Sengupta operator. If the box is reduced "significantly" thereby, go back to (2).

(4) If during (2) or (3) some gaps were detected, split the box at the widest gap. Otherwise bisect in the midpoint of the largest width direction.

(5) Goto (1).

In order to apply this heuristics, one should decide how much change is "significant". In our implementation, we consider a change significant if it is greater than 10% in size, where the size of a box is the sum of the widths of its components. We compared the algorithms on the following examples.

A: This example is taken from [Moore, 1977].

$$\begin{aligned} x_1^2 + x_2^2 &= 1 \\ x_1^2 - x_2 &= 0 \end{aligned}$$

Starting box: $[-1.5, 1.5]^2$.

B: This example is taken from [Morgan, 1987].

$$\begin{aligned} x_1 + x_2 + x_3 + x_4 - 1 &= 0 \\ x_1 + x_2 - x_3 + x_4 - 3 &= 0 \\ x_1^2 + x_2^2 + x_3^2 + x_4^2 - 4 &= 0 \\ x_1^2 + x_2^2 + x_3^2 + x_4^2 - 2x_1 - 3 &= 0 \end{aligned}$$

Starting box: $[-10, 10]^4$.

C1: This example is taken from [Moore and Jones, 1977], where the specific values of the coefficients a_i, b_i and the indices i_1, i_2, i_3 are given.

$$x_i - a_i - b_i x_{i_1} x_{i_2} x_{i_3} = 0, \quad i = 1, \ldots, 10$$

Starting box: $[-2, 2]^{10}$.

C2: This example is also taken from [Moore and Jones, 1977]. The difference from **C1** is that it has 20 variables.

$$x_i - a_i - b_i x_{i_1} x_{i_2} x_{i_3} = 0, \quad i = 1, \ldots, 20$$

Starting box: $[-1, 2]^{20}$.

C3: This example is a modification of **C1** in that each variable x_i is replaced by x_i^2. Note that if $(z_1, \ldots, z_{10})$ is a solution of **C4** then $(\pm z_1, \ldots \pm z_{10})$ is also a solution, hence the number of solutions is a multiple of 1024. In fact, both methods found exactly 1024 isolating boxes.

$$x_i^2 - a_i - b_i x_{i_1}^2 x_{i_2}^2 x_{i_3}^2 = 0, \quad i = 1, \ldots, 10$$

Starting box: $[-1, 1]^{10}$.

C4: This example is another modification of **C1**, where we use the same coefficients but increase the degrees of some variables and add one more term.

$$x_i - a_i - b_i x_{i_1}^3 x_{i_2}^3 x_{i_3}^3 + x_{i_2}^4 x_{i_3}^7 = 0, \quad i = 1, \ldots, 10.$$

Starting box: $[-1, 1]^{10}$.

D1: This is a sparse system with 12 variables and low degree.

$$\begin{aligned}
-x_3x_{10}x_{11} - x_5x_{10}x_{11} - x_7x_{10}x_{11} + x_4x_{12} + x_6x_{12} + x_8x_{12} &= 0.4077 \\
x_2x_4x_9 + x_2x_6x_9 + x_2x_8x_9 + x_1x_{10} &= 1.9115 \\
x_3x_9 + x_5x_9 + x_7x_9 &= 1.9791 \\
3x_2x_4 + 2x_2x_6 + x_2x_8 &= 4.0616 \\
3x_1x_4 + 2x_1x_6 + x_1x_8 &= 1.7172 \\
3x_3 + 2x_5 + x_7 &= 3.9701 \\
x_i^2 + x_{i+1}^2 &= 1, \quad i \text{ odd}
\end{aligned}$$

Starting box: $[0.38, 0.40] \times [0.92, 0.93] \times [0.56, 0.57] \times [0.82, 0.83] \times [-1, 1]^8$.

D2: This example is the same as **D1**. The only difference is that its starting box is larger. Starting box: $[0.38, 0.40] \times [0.92, 0.93] \times [-1, 1]^{10}$.

D3: This example is again the same as **D1**. The only difference is that its starting box is even larger. Starting box: $[-1, 1]^{12}$.

In Table 5.11.1, we report various statistics. At top, we report the number of arithmetic floating point operations. It seems that the RLHB method is usually faster.

The second table shows that the number of bisections is much smaller for RLHB, which explains why RLHB is usually faster.

For the readers who might be interested in more details, we provide further statistics such as the number of Hansen–Sengupta operator, linearized tightening and range test calls. The reason why a Hansen–Sengupta operator call is faster for RLHB than for RHB is that for RLHB the Jacobian is already evaluated from the previous linearized tightening iteration when the Hansen–Sengupta operator is called.

	A	B	C1	C2	C3	C4	D1	D2	D3
					Total (Kflop)				
RLHB	3.19	165	5405	54904	63728	12170	1830	19360	41177
RHB	4.10	169	5511	959077	167244	14185	22254	181307	433538
					Bisections / Splits				
RLHB	5/0	44/1	480/1	656/0	5119/0	867/1	36/0	413/4	877/20
RHB	5/2	71/4	495/2	10232/169	9215/0	1191/0	794/48	6671/337	16899/874
RLHB									
					Range Test				
calls	11	130	979	1337	10239	1737	113	1185	2545
avg (Kflop)	0.03	0.18	0.46	2.03	0.47	0.74	1.01	1.02	1.01
percent	10.3	14.0	8.40	4.96	7.55	10.5	6.26	6.23	6.23
					Linearized Tightening				
calls	25	163	630	1759	12287	1079	199	2001	4107
avg (Kflop)	0.08	0.45	1.74	6.85	1.76	3.09	3.18	3.16	3.16
percent	64.4	44.2	20.3	22.0	33.9	27.4	34.6	32.6	31.6
					Hansen–Sengupta Operator				
calls	5	86	497	680	5119	868	84	887	1943
avg (Kflop)	0.12	0.75	7.62	58.7	7.14	8.57	12.7	13.2	13.0
percent	18.2	39.1	70.1	72.7	57.3	61.1	58.4	32.6	61.5
RHB									
					Range Test				
calls	25	214	1084	24849	25599	2403	2056	16697	41659
avg (Kflop)	0.03	0.16	0.43	1.75	0.51	0.69	0.90	0.90	0.87
percent	18.7	20.6	8.49	4.53	7.86	11.8	8.31	8.28	8.41
					Hansen–Sengupta Operator				
calls	21	149	605	14765	17407	1212	1393	11309	27381
avg (Kflop)	0.15	0.87	8.25	61.8	8.78	10.23	14.6	14.6	14.4
percent	75.2	76.9	90.6	95.2	91.41	87.4	91.2	91.2	91.1

Table 5.11.1: Experimental comparison of RLHB and RHB

Chapter 6

Nonlinear Tightening

In the previous chapter we introduced the notion of linear tightening. The idea was to iterate the following process:

(1) Choose an equation f_i and a direction j.

(2) Find a linearization G_i of f_i in $\boldsymbol{X}$.

(3) Shrink $\boldsymbol{X}$ in direction j optimally obtaining $\boldsymbol{X}'$, such that all solutions of G_i in $\boldsymbol{X}$ are in $\boldsymbol{X}'$.

In this chapter we do not linearize f_i but apply tightening directly to the nonlinear equation. This operation was introduced in [Hong and Stahl, 1994b]. Nonlinear tightening does not have uniqueness, existence and non–existence properties corresponding to linear tightening, but gives in many cases better reductions as the linearized tightening operator or the Hansen–Sengupta operator, especially for "large" boxes. Thus, nonlinear tightening is a pure pruning operation and can not be used for testing existence or uniqueness of solutions.

Roughly put, nonlinear tightening works as follows: Choose an equation f_i from the system $\boldsymbol{f}$ and a direction j. Evaluate f_i in all variables except for x_j on the given box $\boldsymbol{X}$, obtaining a univariate function $F : \mathbb{R} \to \mathbb{IR}$ in x_j. Next, find enclosures $Z_1, \ldots, Z_r$ for the solutions of F in X_i. Now, every solution of f in $\boldsymbol{X}$ and hence every solution of $\boldsymbol{f}$ in $\boldsymbol{X}$ is still contained in one of $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(r)}$ where

$$\boldsymbol{X}^{(k)} = (X_1, \ldots, X_{i-1}, Z_k, X_{i+1}, \ldots, X_n).$$

Nonlinear tightening has been investigated in AI community [Mackworth, 1977, Cleary, 1987, Older and Vellino, 1990], for solving simple equations and inequalities such as $xy = z$ and $x > 0$.

In Section 6.1, we give a precise definition of nonlinear tightening. In Section 6.2, we describe an algorithm for finding isolating boxes, which uses the tightening operation. In Section 6.3, we illustrate the algorithm described in the previous section on a simple example. In Section 6.4, we give a general description of a tightening procedure, and a more detailed one for multivariate polynomials. In Section 6.5, we report some experimental results.

6.1 Definition of Nonlinear Tightening

In this section we define the notion "nonlinear tightening". Throughout this section, let $f : \mathbb{R}^n \to \mathbb{R}$ and $\boldsymbol{X} \in \mathbb{IR}^n$.

Definition 6.1.1 (Variety) *The variety $\mathcal{Z}(f)$ of f, is defined by*

$$\mathcal{Z}(f) = \{\boldsymbol{x} \in \mathbb{R}^n \mid f(\boldsymbol{x}) = 0\}. \ \square$$

Definition 6.1.2 (Projection) *Let $\boldsymbol{x} \in \mathbb{R}^n$. The i-th projection of $\boldsymbol{x}$, written as $\pi_i(\boldsymbol{x})$, is defined by*

$$\pi_i(\boldsymbol{x}) = x_i.$$

Let $\mathcal{S} \subseteq \mathbb{R}^n$. Then the i-th projection of $\mathcal{S}$, written also as $\pi_i(\mathcal{S})$, is defined by

$$\pi_i(\mathcal{S}) = \{\pi_i(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathcal{S}\}. \ \square$$

Definition 6.1.3 (Optimal Tightening) *The optimal tightening of $\boldsymbol{X}$ on x_i by f is defined as the set*

$$\pi_i(\mathcal{Z}(f) \cap \boldsymbol{X}). \ \square$$

Theorem 6.1.4 (Solution Preservation) *Let X_i' be the optimal tightening of $\boldsymbol{X}$ on x_i by f, and let $\boldsymbol{X}' = (X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n)$. Then we have*

$$\mathcal{Z}(f) \cap \boldsymbol{X} = \mathcal{Z}(f) \cap \boldsymbol{X}'. \ \square$$

Thus, we can always safely replace $\boldsymbol{X}$ by $\boldsymbol{X}'$ when we are interested only in the solution of f in $\boldsymbol{X}$. Unlike in the case of linear tightening, it is expensive to compute the optimal tightening, since in general it requires exact computations with real algebraic numbers. Thus we relax the definition as follows:

Definition 6.1.5 (Tightening) *A tightening of $\boldsymbol{X}$ on x_i by f is a finite set of disjoint sub–intervals of X_i whose union contains the optimal tightening. More precisely, it is a set*

$$\{X_i^{(1)}, \ldots, X_i^{(\ell)}\}$$

such that $X_i^{(k)} \in \mathbb{IR}$, $X_i^{(k)} \subseteq X_i$, $X_i^{(k)} \cap X_i^{(j)} = \emptyset$ for $k \neq j$, and $\biguplus_k X_i^{(k)} \supseteq \pi_i(\mathcal{Z}(f) \cap \boldsymbol{X})$. $\square$

Definition 6.1.6 (Tightening Operator) *A tightening operator is a procedure that, given f, $\boldsymbol{X}$ and i, produces a finite set of boxes $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(\ell)}$ such that for every $k = 1, \ldots, \ell$*

$$\boldsymbol{X}^{(k)} = (X_1, \ldots, X_{i-1}, X_i^{(k)}, X_{i+1}, \ldots, X_n)$$

where the set $\{X_i^{(1)}, \ldots, X_i^{(\ell)}\}$ forms a tightening of $\boldsymbol{X}$ on x_i by f. $\square$

One extreme tightening operator is the most expensive one which computes the optimal tightening and the other extreme is the cheapest one which trivially returns $\{\boldsymbol{X}\}$. Obviously, we are interested in one between these two extremes, which strikes a "good" compromise between accuracy and computational cost. One such tightening operator for multivariate polynomial functions will be described in Section 6.4.

6.2 Combined Algorithm for Finding Isolating Boxes

In this section, we use the Hansen–Sengupta operator together with a tightening operator for finding isolating boxes for all zeros of a function within a box.

Algorithm 6.2.1 [Isolating Boxes by Hansen–Sengupta Operator and Tightening]

In: $\boldsymbol{f} : \mathbb{R}^n \longrightarrow \mathbb{R}^n$,
$\boldsymbol{X} \in \mathbb{IR}^n$.

Out: *Solutions*, a finite set of isolating boxes for the solutions of $\boldsymbol{f}$ in $\boldsymbol{X}$,
TooSmall, a finite set of sub–boxes of $\boldsymbol{X}$ that are too small to work on.

(1) [Initialize.]
$Work \longleftarrow \{\boldsymbol{X}\}$. $Solutions \longleftarrow \{\}$. $TooSmall \longleftarrow \{\}$.

(2) [Choose a box and an operation.]
Choose and remove a box from $Work$.
Choose one of the following operations:

- range test,
- tightening,
- Hansen–Sengupta operator,
- bisection.

(3) [Process the box.]
Apply the operation on the chosen box, obtaining possibly one ore more sub–boxes.
Insert the resulting boxes into the proper sets: Solutions, TooSmall, or Work.

(4) [Loop.]
If there is a box in Work, then go to Step (2). Otherwise, we are done. □

Algorithm 6.2.1 is correct no matter which boxes and operations are chosen in Step (2). Since we are interested in finding all isolating boxes (not just one), the efficiency of the algorithm does not depend on which box we choose, because we need to analyze every box eventually. Thus, it is fine to choose the first one in the data structure of $Work$. But the efficiency of the algorithm depends heavily on which operation is chosen. Based on experimental results, we found that the following strategy seems to work well.

(1) First, apply tightening with respect to every equation and variable. During tightening carry out range test, since it can be done cheaply using the intermediate results of tightening.

(2) Next, apply Hansen–Sengupta operator.

(3) If the state has been "significantly" changed during (1) and (2), then goto (1).

(4) If during tightening or during the Hansen–Sengupta operator a gap was detected, split in the direction of the largest gap. Otherwise bisect in the midpoint of the largest width direction.

(5) Goto (1).

In order to apply this heuristics, one should decide how much change is "significant" to avoid bisection. In our implementation, we consider a change significant if it is greater than 10% in size, where the size of a box is the sum of the widths of its components. For comparison purpose, we have implemented the method described in [Hansen and Sengupta, 1981], and we applied a similar strategy.

6.3 Illustration

Before jumping into the details of tightening operation, we illustrate the main algorithm described in the previous section on a simple example, taken from [Moore, 1977]:

$$x^2 + y^2 - 1$$
$$x^2 - y$$

The graph on the left hand side of Figure 6.3.1 traces the boxes produced during the execution of the algorithm of the last section. The one on the right hand side provides the same information produced by the algorithm of [Hansen and Sengupta, 1981] which is basically run by Hansen–Sengupta operator, range test, and bisection.

From the picture on the left hand side, we see that the initial box $[-1.5, 1.5]^2$ is first tightened by the circle and then by the parabola. The white patch is the portion that has been tightened out. The remaining box can neither be reduced by the Hansen–Sengupta operator nor by tightening, so it is bisected vertically.

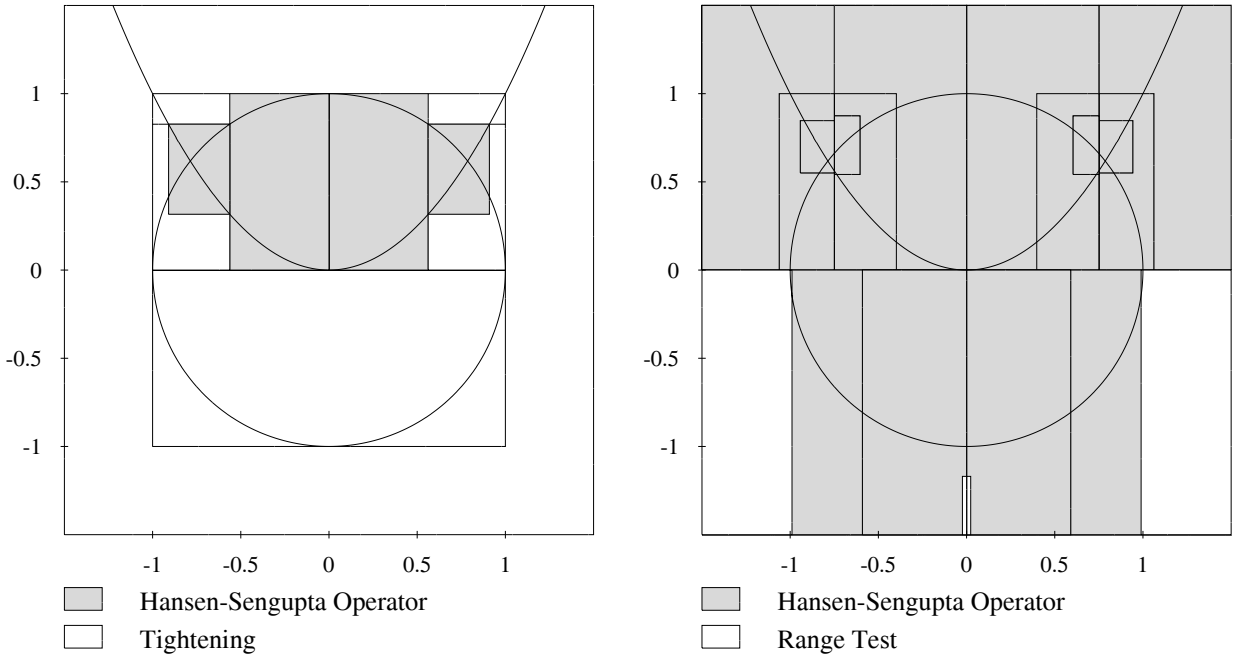


Figure 6.3.1: Illustration of the main algorithm at a simple example

In each sub–box tightening is not successful, but the Hansen–Sengupta operator leads to some reduction, pruning out the gray strip. The remaining box is tightened by the circle and then by the parabola. Now, the Hansen–Sengupta operator detects that the remaining box is an isolating box.

For the picture on the right hand side we do not go into details, but note that it produced more intermediate boxes. This is partly due to the increased number of bisections (the left picture has 1 bisection, while the right one has 7 bisections). The experimental results in Section 6.5 show that this effect seems to become more dramatic for higher dimensional problems.

6.4 Algorithm for Tightening

In this section, we describe a general scheme for tightening continuous functions, and in particular give a detailed procedure for the case of multivariate polynomials. Let us begin by recalling the problem of tightening.

In: $f : \mathbb{R}^n \to \mathbb{R}$, $\boldsymbol{X} \in \mathbb{IR}^n$, and $i \in \{1, \ldots, n\}$.

Out: a tightening of $\boldsymbol{X}$ on x_i by f.

In the following, we will reduce this problem into two sub–problems. Note that

$$\pi_i(\mathcal{Z}(f) \cap \boldsymbol{X}) = \{x_i \in X_i \mid 0 \in f(X_1, \ldots, X_{i-1}, x_i, X_{i+1}, \ldots, X_n)\}.$$

Let $F : \mathbb{R} \longrightarrow \mathbb{IR}$ be defined as

$$F(x_i) = [f(X_1, \ldots, X_{i-1}, x_i, X_{i+1}, \ldots, X_n)].$$

Using this function, we can rewrite the above formulas as

$$\pi_i(\mathcal{Z}(f) \cap \boldsymbol{X}) \subseteq \{x_i \in X_i \mid 0 \in F(x_i)\}. \tag{6.4.1}$$

Note that if f is continuous then equality holds in (6.4.1). Now, let $\underline{F}, \overline{F} : \mathbb{R} \to \mathbb{R}$ be the bounding functions of F, i.e.

$$F(x_i) = [\,\underline{F}(x_i), \overline{F}(x_i)\,].$$

Continuing using these two functions, we have

$$\begin{aligned}\pi_i(\mathcal{Z}(f) \cap \boldsymbol{X}) &\supseteq \{x_i \in X_i \mid 0 \in F(x_i)\} \\ &= \{x_i \in X_i \mid 0 \in [\,\underline{F}(x_i), \overline{F}(x_i)\,]\} \\ &= \{x_i \in X_i \mid \underline{F}(x_i) \leq 0 \ \wedge \ \overline{F}(x_i) \geq 0\} \\ &= \{x_i \in X_i \mid \underline{F}(x_i) \leq 0\} \ \cap \ \{x_i \in X_i \mid \overline{F}(x_i) \geq 0\}.\end{aligned}$$

Thus, we have reduced the problem of tightening into two sub–problems:

1. finding the functions $\underline{F}$ and $\overline{F}$,
2. solving the inequalities $\underline{F}(x_i) \leq 0$ and $\overline{F}(x_i) \geq 0$.

For the purpose of just tightening (not necessary optimal), it is sufficient to overestimate. This gives us the following main algorithm:

Algorithm 6.4.1 (Tightening)

(1) Compute the functions $\underline{F}$ and $\overline{F}$ (suitably overestimate).

(2) Solve the two inequalities $\underline{F} \leq 0$ and $\overline{F} \geq 0$ (overestimate again).

(3) Return the intersection of the two solution sets. □

In the following two sections, we show the details of the first two steps. For the first step, we present an algorithm for polynomials only, while for the second step, we give a general algorithm. Thus we present the second step first in the order of generality.

6.4.1 Solving Inequalities

We begin with the second step, namely that of solving the inequalities. The two inequalities are solved in the same way, and thus we describe how to solve only one of them, $\overline{F} \geq 0$. So, here is the sub–problem statement:

In: $\overline{F} : \mathbb{R} \to \mathbb{R}$ and $X \in \mathbb{IR}$.

Out: disjoint intervals $P_1, \ldots, P_k$ such that $\biguplus_j P_j \supseteq \{x \in X \mid \overline{F}(x) \geq 0\}$.

The basic idea is to compute all real roots of $\overline{F}$ within X, which induces a finite set of intervals on which the sign of $\overline{F}$ is constant. From these, we only need to select the ones with non–negative signs. In doing this, we face the following technical problems: due to finite precision arithmetic and multiple roots, we get not the exact roots, but intervals which contain them. Moreover such an interval may contain more than one root. The following algorithm handles such difficulties.

Algorithm 6.4.2 (Solving Inequality)

(1) Compute disjoint (root) intervals $R_1, \ldots, R_r$ such that $\biguplus_j R_j \supseteq \{x \in X \mid \overline{F}(x) = 0\}$. (This can be done for instance by the Interval Newton Method.)

(2) The following code extracts intervals where $\overline{F}$ is non–negative. It essentially scans the root intervals from left to right while picking up solution intervals. In the code, R_i is the current root interval being checked, k keeps track of the number of the solution intervals extracted so far, b holds the upper end point of the previous root interval, and Y is the interval evaluation of $\overline{F}$ on the middle of the gap between two consecutive root intervals.

Initialize $b \longleftarrow \underline{X}$ and $k \longleftarrow 0$.
for $i = 1, \ldots, r$
  $Y \longleftarrow \overline{F}(1/2(b + \underline{R_i}))$ using interval arithmetic.
  case: $\overline{Y} < 0$: $k \longleftarrow k + 1$. $P_k \longleftarrow R_i$.

case: $\overline{Y} \geq 0$ and $k > 0$: $P_k = \text{hull}(P_k, [b, \overline{R}_i])$.
case: $\overline{Y} \geq 0$ and $k = 0$: $k \longleftarrow k+1$. $P_k \longleftarrow [b, \overline{R}_i]$.
$b \longleftarrow \overline{R}_i$.
if $b \neq \overline{X}$ or ($\underline{X} = \overline{X}$ and $r = 0$)
$Y \longleftarrow \overline{F}(1/2(b + \overline{X}))$ using interval arithmetic.
case: $\underline{Y} \geq 0$ and $k > 0$: $P_k = \text{hull}(P_k, [b, \overline{X}])$.
case: $\underline{Y} \geq 0$ and $k = 0$: $k \longleftarrow 1$. $P_k \longleftarrow X$. □

6.4.2 Finding Bounding Functions

Now we tackle the problem of finding the bounding functions for the case of polynomial functions. Here is the problem statement.

In: $f \in \mathbb{R}[x_1, \ldots, x_n]$, $\boldsymbol{X} \in \mathbb{IR}^n$, and $i \in \{1, \ldots, n\}$.

Out: $\underline{F}$ and $\overline{F}$ as defined above (overestimated).

Recall that $\underline{F}$ and $\overline{F}$ are defined by

$$\begin{aligned} F(x_i) &= [\{f(X_1, \ldots, X_{i-1}, x_i, X_{i+1}, \ldots, X_n\}] \\ &= [\,\underline{F}(x_i), \overline{F}(x_i)\,]. \end{aligned}$$

The following algorithm solves the problem.

Algorithm 6.4.3 (Bounding Functions for Polynomials)

(1) Obtain the coefficients $A_j \in \mathbb{IR}$ of an interval polynomial $F(x_i) = \sum_{j=0}^{d} A_j x^j$ by evaluating f on $x_j = X_j$ for every $j \neq i$, using interval arithmetic.

(2) Compute the bounding functions of $F(x)$: (We drop the subscript i for simplicity.)

$$\underline{F}(x) = \begin{cases} \sum_{j=0}^{d} \underline{A}_j x^j & \text{if} \quad x \geq 0 \\ \sum_{j=0}^{d} A_j^{\inf} x^j & \text{else} \end{cases}$$

$$\overline{F}(x) = \begin{cases} \sum_{j=0}^{d} \overline{A}_j x^j & \text{if} \quad x \geq 0 \\ \sum_{j=0}^{d} A_j^{\sup} x^j & \text{else.} \end{cases}$$

where

$$A_j^{\inf} = \begin{cases} \underline{A}_j & \text{if } j \text{ is even} \\ \overline{A}_j & \text{if } j \text{ is odd} \end{cases} \qquad A_j^{\sup} = \begin{cases} \overline{A}_j & \text{if } j \text{ is even} \\ \underline{A}_j & \text{if } j \text{ is odd.} \end{cases} \quad \square$$

The proof of Step (2) is straightforward. One remark is needed here. The resulting bounding functions are piecewise differentiable but not differentiable at 0. This does not cause problems in solving the corresponding inequalities since one only needs to apply the method of the previous section on each piece separately and merge the resulting intervals. While merging, it might be necessary to concatenate two intervals containing 0.

6.5 Experimental Results

In the sequel THB denotes the method described above (tightening, Hansen–Sengupta operator, bisection) and RHB stands for the method described in [Hansen and Sengupta, 1981] (range test, Hansen–Sengupta operator, bisection). We have tested the algorithms on the same examples as described in Section 5.11.

In Table 6.5.1, we report various statistics. At top, we report the number of arithmetic floating point operations. It seems that the THB is faster, in particular for the problems with many variables.

The second table shows that the number of bisections is much smaller for THB, which explains why THB is usually faster. The reduction in the number of bisections is mainly due to the tightening operation.

	A	**B**	**C1**	**C2**	**C3**	**C4**	**D1**	**D2**	**D3**
					Total (Kflop)				
THB	1.87	74.9	24.4	211	30557	57.1	573	7655	17340
RHB	4.10	169	5511	959077	167244	14185	22254	181307	433538
					Bisections / Splits				
THB	1/0	8/0	0/0	0/0	0/1023	1/0	5/1	100/3	240/3
RHB	5/2	71/4	495/2	10232/169	9215/0	1191/0	794/48	6671/337	16899/874
THB									
					Tightening				
calls	6	42	2	3	3072	4	34	427	987
avg (Kflop)	0.17	0.93	3.03	8.81	2.83	6.40	3.59	3.52	3.48
percent	55.5	52.0	24.8	12.5	28.5	44.8	21.3	19.6	19.8
					Hansen–Sengupta Operator				
calls	6	40	2	3	3072	3	33	421	951
avg (Kflop)	0.13	0.87	9.13	61.7	7.06	10.4	13.6	14.6	14.6
percent	41.3	46.8	74.8	87.4	71.0	54.8	78.4	80.1	78.8
RHB									
					Range Test				
calls	25	214	1084	24849	25599	2403	2056	16697	41659
avg (Kflop)	0.03	0.16	0.43	1.75	0.51	0.69	0.90	0.90	0.87
percent	18.7	20.6	8.49	4.53	7.86	11.8	8.31	8.28	8.41
					Hansen–Sengupta Operator				
calls	21	149	605	14765	17407	1212	1393	11309	27381
avg (Kflop)	0.15	0.87	8.25	61.8	8.78	10.23	14.6	14.6	14.4
percent	75.2	76.9	90.6	95.2	91.41	87.4	91.2	91.2	91.1

Table 6.5.1: Experimental comparison of THB and RHB

Chapter 7

Solution of the Robot Inverse Kinematics Problem

In this chapter we apply the interval methods presented in the previous chapters for solving the robot inverse kinematics problem. The inverse kinematics problem is stated as follows: Given a parametric description of a robot and a desired position and orientation of the end effector, find all possible sets of joint values such that the end effector is in the desired position and orientation.

There are important classes of robots where this problem has a closed form solution. In general, however, one has to rely on iterative methods. In the following we consider robots with 6 joints corresponding to the 3 degrees of freedom of the end effector in both position and orientation. The joints are either revolute or prismatic. For an introduction to robot kinematics we refer to the excellent monographs [Paul, 1981] and [Craig, 1986].

7.1 Kinematic Equations

The kinematic equations of a robot are given by

$$\mathfrak{a}_1\mathfrak{a}_2\cdots\mathfrak{a}_6 = \mathfrak{e}, \tag{7.1.1}$$

where $\mathfrak{a}_1,\ldots,\mathfrak{a}_6,\mathfrak{e} \in \mathbb{R}^{4\times 4}$. Each $\mathfrak{a}_i$ corresponds to a joint and a connected link of the robot and $\mathfrak{e}$ describes the end effector position and orientation. More precisely,

$$\mathfrak{a}_i = \mathrm{rot}_z(\theta_i)\,\mathrm{trans}_z(d_i)\,\mathrm{trans}_x(a_i)\,\mathrm{rot}_x(\alpha_i)$$

where

$$\mathrm{rot}_z(\theta_i) = \begin{pmatrix} \cos(\theta_i) & -\sin(\theta_i) & 0 & 0 \\ \sin(\theta_i) & \cos(\theta_i) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$\mathrm{rot}_x(\alpha_i) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos(\alpha_i) & -\sin(\alpha_i) & 0 \\ 0 & \sin(\alpha_i) & \cos(\alpha_i) & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$\mathrm{trans}_z(d_i) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & d_i \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$\mathrm{trans}_x(a_i) = \begin{pmatrix} 1 & 0 & 0 & a_i \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

If the i-th joint of the robot is revolute, then d_i, a_i, α_i are constants given by the structure of the robot and θ_i is a variable describing the joint angle. If the i-th joint is prismatic, then θ_i, a_i, α_i are constants and d_i is a variable describing the joint distance.

The matrices $\mathrm{rot}_z(\theta)$, $\mathrm{rot}_x(\alpha)$ and $\mathrm{trans}_z(d)$, $\mathrm{trans}_x(a)$ are called elementary rotation and translation frames respectively. Frames are a class of 4×4 matrices with special properties.

Definition 7.1.1 (Frame) *A matrix $\mathfrak{a} \in \mathbb{R}^{4\times 4}$ is called frame if*

$$\mathfrak{a} = \begin{pmatrix} & \mathfrak{o} & & \boldsymbol{p} \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

for some orthogonal $\mathfrak{o} \in \mathbb{R}^{3\times 3}$ and $\boldsymbol{p} \in \mathbb{R}^3$. The matrix $\mathfrak{o}$ is called orientation of $\mathfrak{a}$, the vector $\boldsymbol{p}$ is called position of $\mathfrak{a}$. □

Obviously $\mathrm{rot}_z(\theta)$, $\mathrm{rot}_x(\alpha)$, $\mathrm{trans}_z(d)$, $\mathrm{trans}_x(a)$ are frames, and as the product of two frames is a frame, it holds that $\mathfrak{a}_1, \ldots, \mathfrak{a}_6$ and $\mathfrak{e}$ are frames.

7.2 Forward Kinematics

Before we come to the inverse kinematics problem, we discuss the forward kinematics problem, which is much simpler. The results of this section will be used in an algorithm for computing the inverse kinematics. The forward kinematics problem can be stated as follows: Given $\theta_i, d_i, \alpha_i, a_i$, $i = 1, \ldots, 6$, find $\mathfrak{e}$ such that

$$\mathfrak{e} = \mathfrak{a}_1\mathfrak{a}_2 \cdots \mathfrak{a}_6.$$

In the following let

$$x_i = \begin{cases} d_i & \text{if the } i\text{-th joint is prismatic} \\ \theta_i & \text{if the } i\text{-th joint is revolute} \end{cases}$$

and let $\boldsymbol{x} = (x_1, \ldots, x_6)^{\mathrm{T}}$. Note that $\mathfrak{a}_i$ depends only on x_i but not on x_j for $j \neq i$. If the i-th joint is prismatic, then we call x_i distance variable, otherwise x_i is called angle variable.

Computing the forward kinematics for a given robot means simply evaluating the function $\mathfrak{f} : \mathbb{R}^6 \to \mathbb{R}^{4\times 4}$,

$$\mathfrak{f}(\boldsymbol{x}) = \mathfrak{a}_1\mathfrak{a}_2 \cdots \mathfrak{a}_6. \tag{7.2.1}$$

Now, we want to compute the forward kinematics where the joint values are given by intervals $X_i \in \mathbb{IR}$. Therefore, we define an interval extension $\mathfrak{F}$ of $\mathfrak{f}$. As the associative law does not hold for interval matrix multiplication, we have to explicitly specify the order in which the matrices are multiplied. It seems that the smallest number of arithmetic operations are needed if the elementary matrices are multiplied successively from left to right. Hence, let

$$\mathfrak{F}(\boldsymbol{X}) = \mathfrak{F}_6^{(4)}$$

where

$$\begin{aligned}
\mathfrak{F}_i^{(0)} &= \mathfrak{F}_{i-1}^{(4)} \\
\mathfrak{F}_i^{(1)} &= \begin{cases} \mathfrak{F}_i^{(0)}\mathrm{rot}_z(\theta_i) & \text{if the } i\text{-th joint is prismatic} \\ \mathfrak{F}_i^{(0)}\mathrm{rot}_z(X_i) & \text{if the } i\text{-th joint is revolute} \end{cases} \\
\mathfrak{F}_i^{(2)} &= \begin{cases} \mathfrak{F}_i^{(1)}\mathrm{trans}_z(d_i) & \text{if the } i\text{-th joint is revolute} \\ \mathfrak{F}_i^{(1)}\mathrm{trans}_z(X_i) & \text{if the } i\text{-th joint is prismatic} \end{cases} \\
\mathfrak{F}_i^{(3)} &= \mathfrak{F}_i^{(2)}\mathrm{trans}_x(a_i) \\
\mathfrak{F}_i^{(4)} &= \mathfrak{F}_i^{(3)}\mathrm{rot}_x(\alpha_i)
\end{aligned}$$

for $i = 1, \ldots, 6$ and let

$$\mathfrak{F}_0^{(4)} = \mathrm{i}.$$

Obviously $\boldsymbol{x} \in \boldsymbol{X}$ implies $\mathfrak{f}(\boldsymbol{x}) \in \mathfrak{F}(\boldsymbol{X})$ but due to interval dependencies usually

$$\mathfrak{F}(\boldsymbol{X}) \supset \mathfrak{f}(\boldsymbol{X}).$$

In order to obtain closer inclusions of $\mathfrak{f}$ one can apply standard techniques like centered forms as described in Chapter 4 for polynomials. However, the special structure of (7.2.1) gives us further possibilities as described below.

Frame Property. Note that each $\mathfrak{F}_i^{(j)}$, $i = 1, \ldots, 6$, $j = 1, \ldots, 4$ bounds a set of frames. Eliminating all matrices which are not frames from $\mathfrak{F}_i^{(j)}$ would in general not give an interval matrix. Hence, we reduce $\mathfrak{F}_i^{(j)}$ in a sub–optimal way without leaving the class of interval matrices. So, the problem can be stated as follows: Given $\mathfrak{F} \in \mathbb{IR}^{4\times 4}$, find $\hat{\mathfrak{F}} \subseteq \mathfrak{F}$ such that each frame in $\mathfrak{F}$ is also in $\hat{\mathfrak{F}}$. This problem can be solved by tightening using the orthogonality properties of the orientation of frames.

Dependency between $\sin(X)$ and $\cos(X)$. For revolute joints, overestimation occurs already during the multiplication of a matrix $\mathfrak{F}$ by $\mathrm{rot}_z(X)$ because of the dependency between $\sin(X)$ and $\cos(X)$.

Let

$$\mathfrak{F} = \begin{pmatrix} O_{11} & O_{12} & O_{13} & P_1 \\ O_{21} & O_{22} & O_{23} & P_1 \\ O_{31} & O_{32} & O_{33} & P_1 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Then

$$\begin{aligned} \mathfrak{F}\,\mathrm{rot}_z(X) &= \begin{pmatrix} O_{11} & O_{12} & O_{13} & P_1 \\ O_{21} & O_{22} & O_{23} & P_1 \\ O_{31} & O_{32} & O_{33} & P_1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \cos(X) & -\sin(X) & 0 & 0 \\ \sin(X) & \cos(X) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} O_{11}\cos(X) + O_{12}\sin(X) & -O_{11}\sin(X) + O_{12}\cos(X) & O_{13} & P_1 \\ O_{21}\cos(X) + O_{22}\sin(X) & -O_{21}\sin(X) + O_{22}\cos(X) & O_{23} & P_2 \\ O_{31}\cos(X) + O_{32}\sin(X) & -O_{31}\sin(X) + O_{32}\cos(X) & O_{33} & P_3 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \end{aligned}$$

In order to avoid this kind of overestimation, we have to solve the following problem: Given $A, B, X \in \mathbb{IR}$, find

$$\{A\sin(x) + B\cos(x) \mid x \in X\}.$$

When solving the inverse kinematics problem, we are are usually given the intervals $\sin(X)$, $\cos(X)$ but not X. Therefore, the problem is slightly different: Given $A, B, S, C \in \mathbb{IR}$, find

$$\{As + Bc \mid s^2 + c^2 = 1, s \in S, c \in C\}.$$

In order to simplify the problem, we assume $0 \notin \mathrm{int}(S)$, $0 \notin \mathrm{int}(C)$. This can always be achieved through bisections. As

$$\begin{aligned} \{As + Bc \mid s^2 + c^2 = 1, s \in S, c \in C\} &= \{(-A)s + Bc \mid s^2 + c^2 = 1, s \in -S, c \in C\} \\ &= \{As + (-B)c \mid s^2 + c^2 = 1, s \in S, c \in -C\} \end{aligned}$$

we can reduce the problem to $\underline{S} \geq 0$, $\underline{C} \geq 0$. Thus,

$$\begin{aligned} \overline{\{As + Bc \mid s^2 + c^2 = 1, s \in S, c \in C\}} &= \max\{\overline{A}s + \overline{B}c \mid s^2 + c^2 = 1, s \in S, c \in C\} \\ \underline{\{As + Bc \mid s^2 + c^2 = 1, s \in S, c \in C\}} &= \min\{\underline{A}s + \underline{B}c \mid s^2 + c^2 = 1, s \in S, c \in C\} \end{aligned}$$

and the problem can be solved for example by the Lagrange multiplier method.

7.3 Inverse Kinematics

In this section we give an algorithm for solving the inverse kinematics problem

$$\mathfrak{a}_1\mathfrak{a}_2\cdots\mathfrak{a}_6 = \mathfrak{e} \tag{7.3.1}$$

based on tightening and the Hansen–Sengupta operator. The algorithm is in principle the same as the one presented in Chapter 6, but exploits the structure of the kinematic equations (7.3.1).

7.3.1 Tightening

The idea of tightening is to evaluate an equation in all variables except for one and solve the univariate interval equation. For the kinematic equations (7.3.1) this is particularly straight forward, because (7.3.1) can be solved explicitly for each variable:

For revolute joints we can rewrite (7.3.1) as

$$\begin{array}{lcccccccccccl}
\mathrm{rot}_z(x_1) & = & & & & & & \mathfrak{e} & \mathfrak{a}_6^{-1} & \mathfrak{a}_5^{-1} & \mathfrak{a}_4^{-1} & \mathfrak{a}_3^{-1} & \mathfrak{a}_2^{-1} & \mathrm{rot}_x^{-1}(\alpha_1)\mathrm{trans}_x^{-1}(a_1)\mathrm{trans}_z^{-1}(d_1) \\
\mathrm{rot}_z(x_2) & = & & & & & \mathfrak{a}_1^{-1} & \mathfrak{e} & \mathfrak{a}_6^{-1} & \mathfrak{a}_5^{-1} & \mathfrak{a}_4^{-1} & \mathfrak{a}_3^{-1} & & \mathrm{rot}_x^{-1}(\alpha_2)\mathrm{trans}_x^{-1}(a_2)\mathrm{trans}_z^{-1}(d_2) \\
\mathrm{rot}_z(x_3) & = & & & & \mathfrak{a}_2^{-1} & \mathfrak{a}_1^{-1} & \mathfrak{e} & \mathfrak{a}_6^{-1} & \mathfrak{a}_5^{-1} & \mathfrak{a}_4^{-1} & & & \mathrm{rot}_x^{-1}(\alpha_3)\mathrm{trans}_x^{-1}(a_3)\mathrm{trans}_z^{-1}(d_3) \\
\mathrm{rot}_z(x_4) & = & & & \mathfrak{a}_3^{-1} & \mathfrak{a}_2^{-1} & \mathfrak{a}_1^{-1} & \mathfrak{e} & \mathfrak{a}_6^{-1} & \mathfrak{a}_5^{-1} & & & & \mathrm{rot}_x^{-1}(\alpha_4)\mathrm{trans}_x^{-1}(a_4)\mathrm{trans}_z^{-1}(d_4) \\
\mathrm{rot}_z(x_5) & = & & \mathfrak{a}_4^{-1} & \mathfrak{a}_3^{-1} & \mathfrak{a}_2^{-1} & \mathfrak{a}_1^{-1} & \mathfrak{e} & \mathfrak{a}_6^{-1} & & & & & \mathrm{rot}_x^{-1}(\alpha_5)\mathrm{trans}_x^{-1}(a_5)\mathrm{trans}_z^{-1}(d_5) \\
\mathrm{rot}_z(x_6) & = & \mathfrak{a}_5^{-1} & \mathfrak{a}_4^{-1} & \mathfrak{a}_3^{-1} & \mathfrak{a}_2^{-1} & \mathfrak{a}_1^{-1} & \mathfrak{e} & & & & & & \mathrm{rot}_x^{-1}(\alpha_6)\mathrm{trans}_x^{-1}(a_6)\mathrm{trans}_z^{-1}(d_6)
\end{array}$$

For prismatic joints we have

$$\begin{array}{lcccccccccccl}
\mathrm{trans}_z(x_1) & = & \mathrm{rot}_z^{-1}(\theta_1) & & & & & & \mathfrak{e} & \mathfrak{a}_6^{-1} & \mathfrak{a}_5^{-1} & \mathfrak{a}_4^{-1} & \mathfrak{a}_3^{-1} & \mathfrak{a}_2^{-1} & \mathrm{rot}_x^{-1}(\alpha_1)\mathrm{trans}_x^{-1}(a_1) \\
\mathrm{trans}_z(x_2) & = & \mathrm{rot}_z^{-1}(\theta_2) & & & & & \mathfrak{a}_1^{-1} & \mathfrak{e} & \mathfrak{a}_6^{-1} & \mathfrak{a}_5^{-1} & \mathfrak{a}_4^{-1} & \mathfrak{a}_3^{-1} & & \mathrm{rot}_x^{-1}(\alpha_2)\mathrm{trans}_x^{-1}(a_2) \\
\mathrm{trans}_z(x_3) & = & \mathrm{rot}_z^{-1}(\theta_3) & & & & \mathfrak{a}_2^{-1} & \mathfrak{a}_1^{-1} & \mathfrak{e} & \mathfrak{a}_6^{-1} & \mathfrak{a}_5^{-1} & \mathfrak{a}_4^{-1} & & & \mathrm{rot}_x^{-1}(\alpha_3)\mathrm{trans}_x^{-1}(a_3) \\
\mathrm{trans}_z(x_4) & = & \mathrm{rot}_z^{-1}(\theta_4) & & & \mathfrak{a}_3^{-1} & \mathfrak{a}_2^{-1} & \mathfrak{a}_1^{-1} & \mathfrak{e} & \mathfrak{a}_6^{-1} & \mathfrak{a}_5^{-1} & & & & \mathrm{rot}_x^{-1}(\alpha_4)\mathrm{trans}_x^{-1}(a_4) \\
\mathrm{trans}_z(x_5) & = & \mathrm{rot}_z^{-1}(\theta_5) & & \mathfrak{a}_4^{-1} & \mathfrak{a}_3^{-1} & \mathfrak{a}_2^{-1} & \mathfrak{a}_1^{-1} & \mathfrak{e} & \mathfrak{a}_6^{-1} & & & & & \mathrm{rot}_x^{-1}(\alpha_5)\mathrm{trans}_x^{-1}(a_5) \\
\mathrm{trans}_z(x_6) & = & \mathrm{rot}_z^{-1}(\theta_6) & \mathfrak{a}_5^{-1} & \mathfrak{a}_4^{-1} & \mathfrak{a}_3^{-1} & \mathfrak{a}_2^{-1} & \mathfrak{a}_1^{-1} & \mathfrak{e} & & & & & & \mathrm{rot}_x^{-1}(\alpha_6)\mathrm{trans}_x^{-1}(a_6)
\end{array}$$

Thus, tightening means simply evaluating the right hand sides of the above equations by interval arithmetic and intersect with the matrix on the left hand side. If tightening is done successively in each variable, then intermediate results of the right hand side evaluation can be reused and overestimation is reduced because most recently updated values are used in the computation. Further, as all matrix products on the right hand side are frames, we can use the techniques described in the previous section for reducing overestimation during frame multiplication. Inverse frames can be handled in the same way as ordinary frames because

$$\begin{aligned}
\mathfrak{a}_i^{-1} &= \mathrm{rot}_x^{-1}(\alpha_i)\mathrm{trans}_x^{-1}(a_i)\mathrm{trans}_z^{-1}(d_i)\mathrm{rot}_z^{-1}(\theta_i) \\
&= \mathrm{rot}_x(-\alpha_i)\mathrm{trans}_x(-a_i)\mathrm{trans}_z(-d_i)\mathrm{rot}_z(-\theta_i).
\end{aligned}$$

In order to avoid trigonometric function calls during tightening, we replace each angle variable x_i by two new variables $s_i = \sin(x_i)$, $c_i = \cos(x_i)$. Whenever a new interval S_i for s_i or C_i for c_i is computed, we tighten the other interval optimally using $s_i^2 + c_i^2 = 1$.

7.3.2 Hansen–Sengupta Operator

Note that the matrix kinematic equation

$$\mathfrak{f}(\boldsymbol{x}) = \mathfrak{e}$$

consists of 16 equations in 6 variables. Still, it is not overdetermined as due to the frame property there are only 6 independent equations. In order to apply the Hansen–Sengupta operator, the number of variables must be the same as the number of equations. Hence we choose 6 independent equations, for example

$$\begin{aligned}
f_{11}(\boldsymbol{x}) &= e_{11} \\
f_{12}(\boldsymbol{x}) &= e_{12}
\end{aligned}$$

$$\begin{aligned} f_{21}(\boldsymbol{x}) &= e_{21} \\ f_{14}(\boldsymbol{x}) &= e_{14} \\ f_{24}(\boldsymbol{x}) &= e_{24} \\ f_{34}(\boldsymbol{x}) &= e_{34}. \end{aligned} \tag{7.3.2}$$

In the following let

$$\begin{aligned} \boldsymbol{f}(\boldsymbol{x}) &= (f_{11}, f_{12}, f_{21}, f_{14}, f_{24}, f_{34})^{\mathrm{T}} \\ \boldsymbol{e} &= (e_{11}, e_{12}, e_{21}, e_{14}, e_{24}, e_{34})^{\mathrm{T}}. \end{aligned}$$

In order to compute the partial derivatives of $\boldsymbol{f}$ efficiently, we use

$$\frac{\partial}{\partial x_i}\mathfrak{f}(\boldsymbol{x}) = \mathfrak{a}_1 \mathfrak{a}_2 \cdots \mathfrak{a}_i' \cdots \mathfrak{a}_6$$

where

$$\mathfrak{a}_i' = \begin{pmatrix} -\sin(x_i) & -\cos(x_i) & 0 & 0 \\ \cos(x_i) & -\sin(x_i) & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \mathrm{trans}_z(d_i)\,\mathrm{trans}_x(a_i)\,\mathrm{rot}_x(\alpha_i)$$

if the i-th joint is revolute and

$$\mathfrak{a}_i' = \mathrm{rot}_z(\theta_i) \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \mathrm{trans}_x(a_i)\,\mathrm{rot}_x(\alpha_i)$$

if the i-th joint is prismatic. This allows to reuse intermediate results for the computation of all partial derivatives as well as a reduction of the overestimation error in the same way as described in Section 7.2 for the forward kinematics.

As in Chapter 6 we apply tightening and the Hansen–Sengupta operator alternately. During tightening we compute intervals for the sine and the cosine of each angle variable, hence no trigonometric function calls are needed for the computation of the Jacobian. But in order to obtain a starting box for the Hansen–Sengupta operator we need an interval version of the arctan function, and for the evaluation of $\boldsymbol{f}(\mathrm{mid}(\boldsymbol{X}))$ interval versions of the sine and the cosine function are necessary. Finally, in order to get the sines and cosines of the result of the Hansen–Sengupta operator for the next tightening step, further interval sine and cosine function calls are needed. The interval versions of trigonometric and inverse trigonometric functions which we use in our experiments are based on direction rounded, truncated Taylor series, see for example [Rothmaier, 1971], [Braune and Krämer, 1987], [Braune, 1987].

7.4 Experimental Results

In this section we compare three algorithms for solving the robot inverse kinematics problem.

- The first algorithm IKG (Inverse Kinematics General algorithm) is the same as in Chapter 6. For the Hansen–Sengupta operator we use the system of equations (7.3.2). For tightening we use all 12 non–trivial equations of (7.3.1). Each angle variable x_i is replaced by two variables s_i, c_i for its sine and cosine and an equation $s_i^2 + c_i^2 = 1$ is added.
- The second algorithm IKS (Inverse Kinematics Specialized algorithm) is as described in this chapter and exploits the special structure of the kinematic equations.
- The third algorithm IKO (Inverse Kinematics Optimized algorithm) is the same as IKS, except that for the Hansen–Sengupta operator we are using slopes instead of derivatives. The slopes are defined as for the successive mean value form, see Section 4.2.1. Only after the Hansen–Sengupta operator with a slope maps a box into its interior, we use a Jacobian for testing uniqueness of the solution in the box.

In our experiments we consider the following four examples. The costs for the computation are reported in Table 7.4.

- Elbow manipulator, see for example [Paul, 1981]. This robot has only revolute joints.

i	θ_i	d_i	α_i	a_i
1	x_1	0	$\pi/2$	0
2	x_2	0	0	2
3	x_3	0	0	3
4	x_4	0	$-\pi/2$	4
5	x_5	0	$\pi/2$	0
6	x_6	0	0	0

  Starting Box: $[-\pi,\pi]\times[-\pi,\pi]\times[-\pi,\pi]\times[-\pi,\pi]\times[-\pi,\pi]\times[-\pi,\pi]$.
  End Effector Frame:

$$\mathfrak{e} = \begin{pmatrix} -0.611 & -0.767 & 0.194 & 0.571 \\ 0.018 & -0.258 & -0.966 & 0.177 \\ 0.791 & -0.587 & 0.171 & 7.537 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

  The problem has 4 solutions.

- Stanford manipulator, see for example [Paul, 1981]. This robot has 5 revolute joints and 1 prismatic joint.

i	θ_i	d_i	α_i	a_i
1	x_1	0	$-\pi/2$	0
2	x_2	2	$\pi/2$	0
3	0	x_3	0	0
4	x_4	0	$-\pi/2$	0
5	x_5	0	$\pi/2$	0
6	x_6	0	0	0

  Starting Box: $[-\pi,\pi]\times[-\pi,\pi]\times[0,1]\times[-\pi,\pi]\times[-\pi,\pi]\times[-\pi,\pi]$.
  End Effector Frame:

$$\mathfrak{e} = \begin{pmatrix} 0.020 & -0.762 & 0.647 & -0.345 \\ 0.871 & 0.332 & 0.363 & 1.987 \\ -0.491 & 0.556 & 0.670 & 0.306 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

  The problem has 4 solutions.

- Modified Stanford manipulator. All parameters are the same as for the Stanford manipulator except for α_5 and a_5. Note that neither $\sin(\alpha_5) = 0$ nor $\cos(\alpha_5) = 0$, hence the corresponding algebraic system of equations is more dense than in the previous example.

i	θ_i	d_i	α_i	a_i
1	x_1	0	$-\pi/2$	0
2	x_2	2	$\pi/2$	0
3	0	x_3	0	0
4	x_4	0	$-\pi/2$	0
5	x_5	0	1.3	2
6	x_6	0	0	0

  Starting Box: $[-\pi,\pi]\times[-\pi,\pi]\times[0,1]\times[-\pi,\pi]\times[-\pi,\pi]\times[-\pi,\pi]$.
  End Effector Frame:

$$\mathfrak{e} = \begin{pmatrix} 0.006 & -0.907 & 0.422 & -0.153 \\ 0.860 & 0.220 & 0.461 & 3.653 \\ -0.511 & 0.360 & 0.781 & -0.783 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

  The problem has 3 solutions.

- This example is the same as the previous one except that the last joint is replaced by a prismatic joint.

i	θ_i	d_i	α_i	a_i
1	x_1	0	$-\pi/2$	0
2	x_2	2	$\pi/2$	0
3	0	x_3	0	0
4	x_4	0	$-\pi/2$	0
5	x_5	0	1.3	2
6	0	x_6	0	0

Starting Box: $[-\pi,\pi]\times[-\pi,\pi]\times[0,1]\times[-\pi,\pi]\times[-\pi,\pi]\times[0,1]$.
End Effector Frame:

$$\mathfrak{e} = \begin{pmatrix} 0.096 & -0.902 & 0.422 & -0.111 \\ 0.833 & 0.304 & 0.461 & 3.700 \\ -0.544 & 0.307 & 0.781 & -0.705 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

The problem has 1 solution.

In all examples except for the Stanford Manipulator IKS is faster than IKG and in all cases IKO is faster than IKS. While the difference between IKO and IKS is roughly the same in all cases, the performance of IKG deviates significantly. In the Elbow Manipulator example IKG is more than 200 times more expensive than IKO whereas in the case of the Stanford Manipulator the difference is negligible.

7.5 Parallelization

We parallelize algorithm IKO on a workstation network and on the super computer CS-2HA using PVM [Geist *et al.*, 1994] for the parallelization primitives. In both cases we apply a manager–worker scheme. For the super computer implementation this works well up to about 30 processors. If more processors are used, the manager is a bottle neck. We solve this problem by using a more sophisticated load balancing method and report experimental results with up to 65 processors.

7.5.1 Parallelization on a Workstation Network

In this section we describe the parallelization of algorithm IKO on a workstation network and report experimental results with up to 9 machines. For the parallelization we use the following simple manager–worker scheme.

Manager. The manager holds the list of boxes which have to be worked on and a list of idle workers. If there is a box in the list and an idle worker, he sends the box to the worker. Otherwise he waits for a message from a worker and updates the list of boxes, solutions and idle workers. If there are no more boxes and all workers are idle, he sends a termination message to each worker.

Worker. Each worker performs the following loop.

(1) Receive a message from the manager which tells the worker either to terminate or contains a box to work on.

(2) Perform tightening and the Hansen–Sengupta operator as described in the previous section as long as the box is reduced significantly. If the box is canceled thereby then send a message to the manager indicating that the worker is idle and go to (1). If the box is too small to continue working on it or if it is detected that it contains a unique solution, then send the box to the manager and indicate that the worker is idle.

Elbow Manipulator						
	Total (MFlop)	Hansen–Sengupta		Tightening		Bisections
		calls	avg. (Flop)	calls	avg. (Flop)	
IKG	12966.0	377717	19641	495439	11139	161451
IKS	66.4	7191	4760	12251	2606	5703
IKO	56.8	6702	3858	11745	2581	5446

Stanford Manipulator						
	Total (MFlop)	Hansen–Sengupta		Tightening		Bisections
		calls	avg. (Flop)	calls	avg. (Flop)	
IKG	8.1	408	11601	482	6940	113
IKS	8.9	1177	3840	1744	2476	745
IKO	6.5	994	2738	1527	2432	647

Modified Stanford Manipulator 1						
	Total (MFlop)	Hansen–Sengupta		Tightening		Bisections
		calls	avg. (Flop)	calls	avg. (Flop)	
IKG	662.0	25358	14624	31147	9296	10077
IKS	121.4	13503	4623	22342	2589	10450
IKO	95.1	11517	3905	19074	2574	8918

Modified Stanford Manipulator 2						
	Total (MFlop)	Hansen–Sengupta		Tightening		Bisections
		calls	avg. (Flop)	calls	avg. (Flop)	
IKG	85.3	4274	11282	4860	7585	1550
IKS	4.4	507	4470	858	2485	389
IKO	3.7	465	3608	760	2512	342

Table 7.4.1: Experimental comparison of 3 algorithms for solving the inverse kinematics problem.

(3) Bisect the box, send one half to the manager and continue with the other half at (1).

The algorithm is implemented in the C++ language using PVM for the parallelization primitives. The network consists of Silicon Graphics workstations with MIPS R2000A/R3000 processors and 33 MHz clock frequency. The manager and each worker runs on a separate machine. Experimental results with up to 8 workers are reported in Table 7.5.1. The utilization is obtained by dividing the speedup with n workers by n, i.e. the manager is not taken into account. For the experiments we are using the same examples as in the previous section. All computing times are measured in seconds using wall clock.

7.5.2 Parallelization on the Super Computer CS-2HA

In this section we parallelize algorithm IKO on the MIMD super computer CS-2HA at the European Center for Parallel Computing in Vienna (VCPC).

In its current configuration the CS-2HA has 128 processing elements, each consisting of 2 Super SPARC shared memory 50 MHz processors, 1 MB cache and 64 MB RAM. The CS-2HA is divided into several partitions where the largest partition comprises 68 processors. The communication is supported by proprietary communication and switching chips. Each processing element interfaces the network through an Elan communications chip, which relives the CPU from communications processing. The Elan is a RISC processor which has a shared memory interface to the CPU and has two 2-way 70 MHz byte wide data links to connect to the network providing 50 MB/s user bandwidth. The network itself is a multi–stage packet switch with 100 MB/sec/link and is built of Elite 8×8 crosspoint switches. As well as supporting point–to–point connectivity, the data network provides hardware broadcast at full bandwidth.

Manager–Worker Scheme. First, we tested the same simple manager–worker scheme as for the workstation implementation described in the previous section on the CS-2HA. Both implementations are based on PVM, hence only minor modifications of the program were necessary. Experimental results are reported in Table 7.5.2. As loading and starting the processes involves negotiation with the resource manager, which can cause arbitrary delays, we do not take the startup time into account.

For the larger two examples (Elbow and Modified Stanford 1) the worker utilization is higher than 90% up to 20 processors! However, in all cases the maximum speedup is already achieved with roughly 25–30 processors and using more than 30 processors increases the computing time. There might be two reasons for this:

- Limited number of boxes which can be worked on simultaneously.
- The manager is a bottle neck.

Modified Manager–Worker Scheme. Let us first assume that there are not enough boxes available which can be processed at the same time. Note that at the beginning of the program only one box is processed, after that there are two boxes which can be worked on in parallel and so on. This means that although a large number of boxes has to be processed, it is possible that there are only few boxes which are available at the same time. The low parallelization degree especially at the beginning of the program execution can be resolved if the strategy of the manager is modified slightly as follows: If there is more than one idle worker but only one available box, then the manager splits the box and assigns one half to a worker. This means that immediately after the program starts, all workers are busy. Experimental results obtained with this modification are reported in Table 7.5.3.

Apparently the modification in the manager strategy does not resolve the problem that using more than 30 processors gives no performance improvement. Qualitatively the results in Table 7.5.3 are the same as in Table 7.5.2. Hence, it seems that the manger is the bottleneck.

Explicit Load Balancing Scheme. The idea to overcome the manager bottleneck is to employ several managers and assign a fixed number of workers to each manager. Further, there is a meta manager which is in charge of load balancing between the managers.

Elbow Manipulator									
Workers	seq	1	2	3	4	5	6	7	8
Time (sec)	485.2	552.5	292.9	201.4	151.7	124.7	105.7	91.6	81.8
Speedup	—	0.88	1.66	2.41	3.20	3.89	4.59	5.30	5.93
Utilization	—	0.88	0.83	0.80	0.80	0.78	0.77	0.76	0.74

Stanford Manipulator									
Workers	seq	1	2	3	4	5	6	7	8
Time (sec)	59.3	66.6	35.3	24.4	19.8	15.9	13.7	12.0	11.1
Speedup	—	0.89	1.68	2.43	2.99	3.73	4.33	4.94	5.34
Utilization	—	0.89	0.84	0.81	0.75	0.75	0.72	0.71	0.67

Modified Stanford Manipulator 1									
Workers	seq	1	2	3	4	5	6	7	8
Time (sec)	781.7	876.7	466.5	318.5	239.6	195.8	169.1	149.4	133.7
Speedup	—	0.89	1.68	2.45	3.26	3.99	4.62	5.23	5.85
Utilization	—	0.89	0.84	0.82	0.82	0.80	0.77	0.75	0.73

Modified Stanford Manipulator 2									
Workers	seq	1	2	3	4	5	6	7	8
Time (sec)	32.1	35.2	19.2	13.0	10.2	8.4	7.3	6.3	5.9
Speedup	—	0.91	1.67	2.46	3.15	3.82	4.40	5.10	5.44
Utilization	—	0.91	0.84	0.82	0.79	0.76	0.73	0.72	0.68

Table 7.5.1: Workstation network implementation of algorithm IKO for the inverse kinematics problem.

Elbow Manipulator											
Workers	seq	2	4	8	16	24	32	40	48	56	64
Time (sec)	144.3	76.6	39.3	18.9	9.8	7.3	6.8	7.3	6.7	6.9	7.1
Speedup	—	1.88	3.67	7.63	14.72	19.77	21.22	19.77	21.54	20.91	20.32
Utilization	—	0.94	0.92	0.95	0.92	0.82	0.66	0.49	0.45	0.37	0.32

Stanford Manipulator											
Workers	seq	2	4	8	16	24	32	40	48	56	64
Time (sec)	16.4	8.5	4.6	2.3	1.4	1.2	1.2	1.3	1.4	1.4	1.5
Speedup	—	1.93	3.57	7.13	11.71	13.67	13.67	12.62	11.71	11.71	10.93
Utilization	—	0.96	0.89	0.89	0.73	0.57	0.43	0.32	0.24	0.21	0.17

Modified Stanford Manipulator 1											
Workers	seq	2	4	8	16	24	32	40	48	56	64
Time (sec)	220.7	115.3	57.2	29.8	15.1	11.6	10.5	10.3	10.4	10.5	14.5
Speedup	—	1.91	3.86	7.41	14.62	19.03	21.02	21.43	21.22	21.02	15.22
Utilization	—	0.96	0.96	0.93	0.91	0.79	0.66	0.54	0.44	0.38	0.24

Modified Stanford Manipulator 2											
Workers	seq	2	4	8	16	24	32	40	48	56	64
Time (sec)	8.40	4.34	2.23	1.18	0.80	0.73	0.75	0.83	0.91	0.93	1.02
Speedup	—	1.94	3.77	7.12	10.50	11.51	11.20	10.12	9.23	9.03	8.24
Utilization	—	0.97	0.94	0.89	0.66	0.48	0.35	0.25	0.19	0.16	0.13

Table 7.5.2: Manager–worker implementation of algorithm IKO for the inverse kinematics problem on a super computer.

Elbow Manipulator											
Workers	seq	2	4	8	16	24	32	40	48	56	64
Time (sec)	144.3	75.3	37.6	19.1	9.9	7.2	6.7	6.7	7.1	7.0	7.5
Speedup	—	1.92	3.84	7.55	14.58	20.04	21.54	21.54	20.32	20.61	19.24
Utilization	—	0.96	0.96	0.94	0.91	0.84	0.67	0.54	0.42	0.37	0.30

Stanford Manipulator											
Workers	seq	2	4	8	16	24	32	40	48	56	64
Time (sec)	16.4	8.4	4.6	2.2	1.3	1.2	1.3	1.3	1.5	1.6	1.7
Speedup	—	1.95	3.57	7.45	12.62	13.67	12.62	12.62	10.93	10.25	9.65
Utilization	—	0.98	0.89	0.93	0.79	0.57	0.39	0.32	0.23	0.18	0.15

Modified Stanford Manipulator 1											
Workers	seq	2	4	8	16	24	32	40	48	56	64
Time (sec)	220.7	114.1	57.3	29.2	15.1	11.1	10.4	10.6	10.5	10.3	10.7
Speedup	—	1.93	3.85	7.56	14.62	19.88	21.22	20.82	21.02	21.43	20.63
Utilization	—	0.97	0.96	0.94	0.91	0.83	0.66	0.52	0.44	0.38	0.32

Modified Stanford Manipulator 2											
Workers	seq	2	4	8	16	24	32	40	48	56	64
Time (sec)	8.40	4.38	2.20	1.22	0.77	0.75	0.97	0.92	1.03	1.19	1.43
Speedup	—	1.92	3.82	6.89	10.91	11.20	8.66	9.13	8.16	7.06	5.87
Utilization	—	0.96	0.95	0.86	0.68	0.47	0.27	0.23	0.17	0.13	0.09

Table 7.5.3: Modified manager–worker implementation of algorithm IKO for the inverse kinematics problem on a super computer. The manager bisects a box if there are more than one idle workers and only one available box.

The task of the managers and workers is basically the same as in the previous approach except that from time to time each manager reports his current load to the meta manager. The load of a manager is the difference between the number of boxes on his stack and the number of his idle workers. If the meta manager detects a significant load imbalance, he sends a message to the most loaded manager indicating that he should give work to the least loaded manager.

This approach requires that some parameters are fixed. Experimentally the following values turned out to be reasonable.

- The number of managers is determined such that each manager has approximately 10 workers.
- A manager reports his load to the meta manger if the load information at the meta manager differs at least by 3.
- If according to the information available at the meta manager the most loaded manager has a load of at least 2 and the least loaded manager has a load of less than 0, then the meta manager issues a work shift between these two managers.
- If a manager is told to send work to another manager, he attempts to send as many boxes such that both managers are equally loaded. The information about the load of the recipient of the work is provided by the meta manager.

A problem with this approach is how to detect termination. Each manager sends a message to the meta manager when he has no more work and all his workers are idle. But this information is not sufficient for the meta manager to decide whether all work is done because there might still be some load balance messages under way. Hence, in addition each manager counts the number of sent and received messages between himself, other managers and the meta manger. If a manager has no more work and all his workers are idle, then he sends these numbers to the meta manager. After the meta manager received such a message from every manager, he checks whether the total number of sent and received messages is consistent with his own message count and if so, he sends a termination message to all managers and workers.

Experimental results of this implementation are reported in Table 7.5.4. For the larger examples (Elbow and Modified Stanford 1) we obtain a worker utilization of at least 75% even with the maximum number of processors. In this case the computing time is less than half of the manager–worker approach. For the Stanford Manipulator example we get speedups until the maximum number of processors which are significantly higher than in the previous approach. Only for the Modified Stanford Manipulator 2 example no improvement over the simple manager–worker scheme is achieved. However, as shown in Table 7.4, in this example the total number of boxes which are processed is only 343 and it seems that this is not enough for an efficient parallelization.

Elbow Manipulator											
Workers	seq	1	3	7	15	22	29	36	44	51	58
Managers	—	1	1	1	1	2	3	4	4	5	6
Time (sec)	144.3	149.9	49.5	22.0	10.5	7.3	5.8	4.9	4.1	3.7	3.3
Speedup	—	0.96	2.92	6.56	13.74	19.77	24.88	29.45	35.20	39.00	43.73
Utilization	—	0.96	0.97	0.94	0.92	0.90	0.86	0.82	0.80	0.76	0.75

Stanford Manipulator											
Workers	seq	1	3	7	15	22	29	36	44	51	58
Managers	—	1	1	1	1	2	3	4	4	5	6
Time (sec)	16.4	16.9	5.9	2.6	1.4	1.1	1.0	1.0	1.0	0.9	0.9
Speedup	—	0.97	2.78	6.31	11.71	14.91	16.40	16.40	16.40	18.22	18.22
Utilization	—	0.97	0.93	0.90	0.78	0.68	0.57	0.46	0.37	0.36	0.31

Modified Stanford Manipulator 1											
Workers	seq	1	3	7	15	22	29	36	44	51	58
Managers	—	1	1	1	1	2	3	4	4	5	6
Time (sec)	220.7	238.1	77.3	34.0	16.2	11.2	8.4	7.1	5.9	5.2	4.8
Speedup	—	0.93	2.86	6.49	13.62	19.71	26.27	31.08	37.41	42.44	45.98
Utilization	—	0.93	0.95	0.93	0.91	0.90	0.91	0.86	0.85	0.83	0.79

Modified Stanford Manipulator 2											
Workers	seq	1	3	7	15	22	29	36	44	51	58
Managers	—	1	1	1	1	2	3	4	4	5	6
Time (sec)	8.40	8.55	2.90	1.41	0.88	0.75	0.72	0.78	0.80	0.85	0.90
Speedup	—	0.98	2.90	5.96	9.55	11.20	11.67	10.77	10.50	9.88	9.33
Utilization	—	0.98	0.97	0.85	0.64	0.51	0.40	0.30	0.24	0.19	0.16

Table 7.5.4: Explicit load balancing implementation of algorithm IKO for the inverse kinematics problem on a super computer. The work is distributed dynamically between the managers.

Bibliography

[Aho *et al.*, 1977]
A. V. Aho, S. C. Johnson, and J. D. Ullman.
Code Generation for Expressions with Common Subexpression.
J. of the Association for Computing Machinery, 24(1):146–160, 1977.

[Alefeld and Herzberger, 1974]
G. Alefeld and J. Herzberger.
Einführung in die Intervallrechnung.
Bibliographisches Institut, Mannheim–Wien–Zürich, 1974.

[Alefeld and Herzberger, 1983]
G. Alefeld and J. Herzberger.
Introduction to Interval Computations.
Academic Press, N.Y., 1983.

[Alefeld and Lohner, 1985]
G. Alefeld and R. Lohner.
On Higher Order Centered Forms.
Computing, 35:177–184, 1985.

[Alefeld and Rokne, 1981]
G. Alefeld and J. Rokne.
On the Evaluation of Rational Functions in Interval Arithmetic.
SIAM J. Numerical Analysis, 18:862–870, 1981.

[Alefeld, 1970]
G. Alefeld.
Eine Modifikation des Newtonverfahrens zur Bestimmung der reellen Nullstellen einer reellen Funktion.
ZAMM, 50:T32–T33, 1970.

[Alefeld, 1981]
G. Alefeld.
Bounding the Slope of Polynomial Operators and some Applications.
Computing, 26:227–237, 1981.

[Alefeld, 1990]
G. Alefeld.
On the Approximation of the Range of Values by Interval Expressions.
Computing, 44:273–278, 1990.

[Apostolatos and Kulisch, 1967]
N. Apostolatos and U. Kulisch.
Grundlagen einer Maschinenintervallarithmetik.
Computing, 2:89–101, 1967.

[Armstrong, 1983]
M. A. Armstrong.
Basic Topology.
Springer, 1983.

[Baumann, 1988]
E. Baumann.
Optimal Centered Forms.
BIT, 28:80–87, 1988.

[Braune and Krämer, 1987]
K. Braune and W. Krämer.
High Accuracy Standard Functions for Real and Complex Intervals.
In E. Kaucher, U. Kulisch, and C. Ullrich, editors, *Scientific Computation and Programming Languages*, pages 81–114, Teubner, Stuttgart, 1987.

[Braune, 1987]
K. D. Braune.
Hochgenaue Standardfunktionen für reelle und komplexe Punkte und Intervalle in beliebigen Gleitpunktrastern.
PhD thesis, Universität Karlsruhe, 1987.

[Caprani and Madsen, 1980]
O. Caprani and K. Madsen.
Mean Value Forms in Interval Analysis.
Computing, 25:147–154, 1980.

[Cargo and Shisha, 1966]
G. T. Cargo and O. Shisha.
The Bernstein Form of a Polynomial.
Journal of Research of Nat. Bur. Standards, 70B:79–81, 1966.

[Chuba and Miller, 1972]
W. Chuba and W. Miller.
Quadratic Convergence in Interval Arithmetic, Part 1.
BIT, 12:284–290, 1972.

[Cleary, 1987]
J. G. Cleary.
Logical Arithmetic.
Future Computing Systems, 2(2):125–149, 1987.

[Cornelius and Lohner, 1984]
H. Cornelius and R. Lohner.
Computing the Range of Values of Real Functions with Accuracy Higher than Second Order.
Computing, 33:331–347, 1984.

[Craig, 1986]
J. J. Craig.
Introduction to Robotics, Mechanics and Control.
Addison-Wesley, 1986.

[Dimitrova, 1993]
N. Dimitrova.
On Some Properties of an Interval Newton Type Method and its Modification.
In R. Albrecht, G. Alefeld, and H. J. Stetter, editors, *Validation Numerics — Theory and Applications*, pages 21–32, Springer Verlag, 1993.

[Frommer and Mayer, 1989]
A. Frommer and G. Mayer.
Safe Bounds for the Solution of Nonlinear Problems using a Parallel Multisplitting Method.
Computing, 42:171–186, 1989.

[Garloff, 1985]
J. Garloff.
Convergent Bounds for the Range of Multivariate Polynomials.
In K. Nickel, editor, *Interval Mathematics*, pages 37–56. Springer, 1985.

[Geist *et al.*, 1994]
A. Geist, A. Beguelin, J. Dongarra, W. Jiang, R. Manchek, and V. Sunderam.
PVM 3 User's Guide and Reference Manual.
Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, September 1994.

[Grassmann and Rokne, 1979]
E. Grassmann and J. Rokne.
The Range of Values of a Circular Complex Polynnomial over a Circular Complex Interval.
Computing, 23:139–169, 1979.

[Hammer *et al.*, 1993]
R. Hammer, M. Hocks, U. Kulisch, and D. Ratz.
Numerical Toolbox for Verified Computing I.
Springer, 1993.

[Hansen and Greenberg, 1983]
E. Hansen and R. I. Greenberg.
An Interval Newton Method.
Appl. Math. Comp, 12:89–98, 1983.

[Hansen and Sengupta, 1980]
E. R. Hansen and S. Sengupta.
Global Constrained Optimization using Interval Analysis.
In K. Nickel, editor, *Interval Mathematics*, pages 25–48. Springer, 1980.

[Hansen and Sengupta, 1981]
E. Hansen and S. Sengupta.
Bounding Solutions of Systems of Equations using Interval Analysis.
BIT, 21:203–211, 1981.

[Hansen, 1965]
E. Hansen.
Interval Arithmetic in Matrix Computations, Part 1.
SIAM J. Numerical Analysis, 2(2):308–320, 1965.

[Hansen, 1968]
E. Hansen.
On Solving Systems of Equations using Interval Arithmetic.
Math. Comp., 22:374–384, 1968.

[Hansen, 1969]
E. Hansen.
On the Centered Form.
In E. Hansen, editor, *Topics in Interval Analysis*, pages 102–106, Clarendon Press, Oxford, 1969.

[Hansen, 1978a]
E. Hansen.
A Globally Convergent Interval Method for Computing and Bounding Real Roots.
BIT, 18:415–424, 1978.

[Hansen, 1978b]
E. Hansen.
Interval Forms of Newtons Method.
BIT, 20:153–163, 1978.

[Hansen, 1979]
E. Hansen.
Global Optimization using Interval Analysis: The One-Dimensional Case.
Optimization Theory and Applications, 29:331–344, 1979.

[Hansen, 1980]
E. Hansen.
Global Optimization using Interval Analysis: The Multidimensional Case.
Numerische Mathematik, 34:247–270, 1980.

[Hansen, 1988]
E. Hansen.
An Overview of Global Optimization using Interval Analysis.
In R. E. Moore, editor, *Reliability in Computing*, pages 289–307, Academic Press, London, 1988.

[Hansen, 1992]
E. Hansen.
Global Optimization Using Interval Analysis.
Marcel Dekker, Inc., 1992.

[Hong and Stahl, 1994a]
H. Hong and V. Stahl.
Bernstein Form is Inclusion Monotone.
Technical Report 94-52, Research Institute for Symbolic Computation, Johannes Kepler University, Linz, Austria, 1994.

[Hong and Stahl, 1994b]
H. Hong and V. Stahl.
Safe Starting Regions by Fixed Points and Tightening.
Computing, 53:323–335, 1994.

[Hong and Stahl, 1995]
H. Hong and V. Stahl.
Bernstein Form is Inclusion Monotone.
Computing, 55:43–53, 1995.

[Householder, 1970]
A. S. Householder.
The Numerical Treatment of a Single Non-Linear Equation.
Mc Graw-Hill, 1970.

[Ichida and Fujii, 1979]
K. Ichida and Y. Fujii.
An Interval Arithmetic Method for Global Optimization.
Computing, 23:85–97, 1979.

[IEEE, 1985]
IEEE.
IEEE Standard 754–1985 for Binary Floating Point Arithmetic.
Reprinted in SIGPLAN, 22(2):9–25, 1985.

[Jones, 1978]
S. T. Jones.
Searchning for Solutions of Finite Nonlinear Systems – An Interval Approach.
PhD thesis, University of Wisconsin–Madison, 1978.

[Jones, 1980]
S. T. Jones.
Locating Safe Starting Regions for Iterative Methods: A Heuristic Algorithm.
In K. Nickel, editor, *Interval Mathematics*, pages 377–386. Springer, 1980.

[Kearfott, 1987]
R. Baker Kearfott.
Some Tests of Generalized Bisection.
ACM Trans. Math. Software, 13(3), 1987.

[Kearfott, 1990a]
R. Baker Kearfott.
Interval Arithmetic Techniques in the Computational Solution of Nonlinear Systems of Equations: Introduction, Examples and Comparison.
Lectures in Applied Mathematics, 26:337–357, 1990.

[Kearfott, 1990b]
R. Baker Kearfott.

Interval Newton / Generalized Bisection When There are Singularities Near Roots.
Annals of Operations Research, 25:181–196, 1990.

[Knuth, 1981]
D. E. Knuth.
The Art of Computer Programming, Seminumerical Algorithms.
Addison–Wesley, 1981.

[Krawczyk and Neumaier, 1985]
R. Krawczyk and A. Neumaier.
Interval Slopes for Rational Functions and Associated Centered Forms.
SIAM J. Numerical Analysis, 22(3):604–615, 1985.

[Krawczyk and Nickel, 1982]
R. Krawczyk and K. Nickel.
Die zentrische Form in der Intervallarithmetik, ihre quadratische Konvergenz und ihre Inklusionsisotonie.
Computing, 20:117–137, 1982.

[Krawczyk, 1969]
R. Krawczyk.
Newton-Algorithmen zur Bestimmung von Nullstellen mit Fehlerschranken.
Computing, 4:187–201, 1969.

[Krawczyk, 1980a]
R. Krawczyk.
Interval Extensions and Interval Iterations.
Computing, 24:119–129, 1980.

[Krawczyk, 1980b]
R. Krawczyk.
Zur äusseren und inneren Einschließung des Wertebereichs einer Funktion.
Freiburger Intervall–Berichte, 7:1–19, 1980.

[Krawczyk, 1982]
R. Krawczyk.
Zentrische Formen und Intervalloperatoren.
Freiburger Intervall–Berichte, 1:1–30, 1982.

[Krawczyk, 1983]
R. Krawczyk.
Intervallsteigungen für rationale Funktionen und zugeordnete zentrische Formen.
Freiburger Intervall–Berichte, 2:1–30, 1983.

[Krawczyk, 1984]
R. Krawczyk.
Interval Iterations for Including a Set of Solutions.
Computing, 32:13–31, 1984.

[Krawczyk, 1985]
R. Krawczyk.
Interval Operators and Fixed Intervals.
In K. Nickel, editor, *Interval Mathematics*, pages 81–94. Springer, 1985.

[Krawczyk, 1986a]
R. Krawczyk.
A Class of Interval Newton Operators.
Computing, 37:179–183, 1986.

[Krawczyk, 1986b]
R. Krawczyk.
An Improved Interval Newton Operator.
J. Mathematical Analysis and Applications, 118:194–207, 1986.

[Krawczyk, 1986c]
R. Krawczyk.
Properties of Interval Operators.
Computing, 37:227–245, 1986.

[Lane and Riesenfeld, 1981]
J. M. Lane and R. F. Riesenfeld.
Bounds on a Polynomial.
BIT, 21:112–117, 1981.

[Lipson, 1981]
J. D. Lipson.
Elements of Algebra and Algebraic Computing.
Addison–Weseley, 1981.

[Mackworth, 1977]
A. K. Mackworth.
Consistency in Networks of Relations.
Artificial Intelligence, 8:99–118, 1977.

[Miller, 1972]
W. Miller.
Quadratic Convergence in Interval Arithmetic, Part 2.
BIT, 12:291–298, 1972.

[Miller, 1973]
W. Miller.
More on Quadratic Convergence in Interval Arithmetic.
BIT, 13:76–83, 1973.

[Miller, 1975]
W. Miller.
The Error in Interval Arithmetic.
In K. Nickel, editor, *Interval Mathematics*, pages 246–250. Springer, 1975.

[Miranda, 1940]
C. Miranda.
Un 'Osservatione su un Teorema di Brouwer.
Bolletino Unione Math. Ital., 2(3):5–7, 1940.

[Moore and Jones, 1977]
R. E. Moore and S. T. Jones.
Safe Starting Regions for Iterative Methods.
SIAM J. Numerical Analysis, 14:1051–1065, 1977.

[Moore and Kioustelidis, 1980]
R. E. Moore and J. B. Kioustelidis.
A Simple Test for Accuracy of Approximate Solutions to Nonlinear (or Linear) Systems.
SIAM J. Numerical Analysis, 17:521–529, 1980.

[Moore and Qi, 1982]
R. E. Moore and L. Qi.
A successive interval test for nonlinear systems.
SIAM J. Numerical Analysis, 19:845–850, 1982.

[Moore, 1966]
R. E. Moore.
Interval Anaysis.
Prentice-Hall, Englewood Cliffs, N.J., 1966.

[Moore, 1976]
R. E. Moore.
On Computing the Range of a Rational Function of n Variables over a Bounded Region.
Computing, 16:1–15, 1976.

[Moore, 1977]
R. E. Moore.
A test for Existence of Solution to Nonlinear Systems.
SIAM J. Numerical Analysis, 14:611–615, 1977.

[Moore, 1978]
R. E. Moore.
A computational test for convergence of iterative methods for nonlinear systems.
SIAM J. Numerical Analysis, 15:1194–1196, 1978.

[Moore, 1979]
R. E. Moore.
Methods and Applications of Interval Anaysis.
SIAM, Philadelphia, 1979.

[Moore, 1980a]
R. E. Moore.
Interval Methods for Nonlinear Systems.
Computing, Suppl., 2:113–120, 1980.

[Moore, 1980b]
R. E. Moore.
New Results on Nonlinear Systems.
In K. Nickel, editor, *Interval Mathematics*, pages 165–180. Springer, 1980.

[Morgan and Shapiro, 1987]
A. P. Morgan and V. Shapiro.
Box–Bisection for Solving Second–Degree Systems and the Problem of Clustering.
ACM Trans. Math. Software, 13(2):152–167, 1987.

[Morgan, 1983]
A. P. Morgan.
A Method for Computing All Solutions to Systems of Polynomial Equations.
ACM Trans. Math. Software., 9(1):1–17, 1983.

[Morgan, 1987]
A. P. Morgan.
Solving Polynomial Systems Using Continuation for Engineering and Scientific Problems.
Prentice Hall, 1987.

[Neumaier, 1990]
A. Neumaier.
Interval Methods for Systems of Equations.
Cambridge University Press, 1990.

[Older and Vellino, 1990]
W. Older and A. Vellino.
Extending Prolog with Constraint Arithmetic on Real Intervals.
In *Proceedings of the Eight Biennial Conference of the Canadian Society for Computational Studies of Intelligence*, 1990.

[Paul, 1981]
R. P. Paul.
Robot Manipulators: Mathematics, Programming and Control.
MIT Press, 1981.

[Qi, 1980]
L. Qi.
A Generalization of the Krawczyk–Moore Algorithm.
In K. Nickel, editor, *Interval Mathematics*, pages 481–488. Springer, 1980.

[Qi, 1981]
L. Qi.

Interval Boxes of Solutions of Nonlinear Systems.
Computing, 27:137–144, 1981.

[Qi, 1982]
L. Qi.
A Note on the Moore test for nonlinear system.
SIAM J. Numerical Analysis, 19:851–857, 1982.

[Rall, 1983]
L. B. Rall.
Mean Value and Taylor Forms in Interval Analysis.
SIAM, J. Math. An., 14:223–238, 1983.

[Ratschek and Rokne, 1980a]
H. Ratschek and J. Rokne.
About the Centered Form.
SIAM J. Numerical Analysis, 17(3):333–337, 1980.

[Ratschek and Rokne, 1980b]
H. Ratschek and J. Rokne.
Optimality of the Centered Form.
In K. Nickel, editor, *Interval Mathematics*, pages 499–508. Springer, 1980.

[Ratschek and Rokne, 1984]
H. Ratschek and J. Rokne.
Computer Methods for the Range of Functions.
Ellis Horwood Limited, 1984.

[Ratschek and Rokne, 1988]
H. Ratschek and J. Rokne.
New Computer Methods for Global Optimization.
Ellis Horwood Limited, 1988.

[Ratschek and Schröder, 1981]
H. Ratschek and G. Schröder.
Centered Forms for Functions in Several Variables.
Mathematical Analysis and Applications, 82:543–552, 1981.

[Ratschek, 1977]
H. Ratschek.
Mittelwertsätze der Intervallarithmetik.
Beiträge Numer. Mathematik, 6:133–144, 1977.

[Ratschek, 1978]
H. Ratschek.
Zentrische Formen.
ZAMM, 58:434–436, 1978.

[Ratschek, 1980a]
H. Ratschek.
Centered Forms.
SIAM J. Numerical Analysis, 17(5):656–662, 1980.

[Ratschek, 1980b]
H. Ratschek.
Optimal Approximations in Interval Analysis.
In K. Nickel, editor, *Interval Mathematics*, pages 181–202. Springer, 1980.

[Rivlin, 1970]
T. J. Rivlin.
Bounds on a Polynomial.
Journal of Research of Nat. Bur. Standards, 74B:47–54, 1970.

[Rokne and Wu, 1982]
J. G. Rokne and T. Wu.
The Circular Complex Centered Form.
Computing, 28:17–30, 1982.

[Rokne and Wu, 1983]
J. G. Rokne and T. Wu.
A Note on the Circular Complex Centered Form.
Computing, 30:201–211, 1983.

[Rokne, 1977]
J. G. Rokne.
Bounds for an Interval Polynomial.
Computing, 18:225–240, 1977.

[Rokne, 1979a]
J. G. Rokne.
A note on the Bernstein Algorithm for Bounds for Interval Polynomials.
Computing, 21:159–170, 1979.

[Rokne, 1979b]
J. G. Rokne.
The Range of Values of a Complex Polynomial over a Complex Interval.
Computing, 22:153–169, 1979.

[Rokne, 1981]
J. G. Rokne.
The Centered Form for Interval Polynomials.
Computing, 27:339–348, 1981.

[Rokne, 1982]
J. G. Rokne.
Optimal Computation of the Bernstein Algorithm for the Bound of an Interval Polynomial.
Computing, 28:239–246, 1982.

[Rokne, 1985]
J. G. Rokne.
A Low Complexity Explicit Rational Centered Form.
Computing, 34:261–263, 1985.

[Rokne, 1986]
J. G. Rokne.
Low Complexity $k-$dimensional Centered Forms.
Computing, 37:247–253, 1986.

[Rothmaier, 1971]
B. Rothmaier.
Die Berechnung der elementaren Funktionen mit beliebiger Genauigkeit.
PhD thesis, Universität Karlsruhe, 1971.

[Rump, 1982]
S. M. Rump.
Solving Nonlinear Systems with Least Significant Bit Accuracy.
Computing, 29:183–200, 1982.

[Rump, 1984]
S. M. Rump.
Solution of Linear and Nonlinear Algebraic Problems with Sharp, Guaranteed Bounds.
Computing Supplement, 5:147–168, 1984.

[Rump, 1988]
S. M. Rump.
Algorithms for Verified Inclusions: Theory and Practice.
In R. E. Moore, editor, *Reliability in Computing*, pages 109–126, Academic Press, London, 1988.

[Sethi and Ullman, 1970]
R. Sethi and J. D. Ullman.
The Generation of Optimal Code for Arithmetic Expressions.
J. of the Association for Computing Machinery, 17:715–728, 1970.

[Shearer and Wolfe, 1985a]
J. M. Shearer and M. A. Wolfe.
Some Algorithms for the Solution of a class of Nonlinear Algebraic Equations.
Computing, 35:63–72, 1985.

[Shearer and Wolfe, 1985b]
J. M. Shearer and M. A. Wolfe.
Some Computable Existence, Uniqueness, and Convergence Tests for Nonlinear Systems.
SIAM J. Numerical Analysis, 22(6):1200–1207, 1985.

[Skelboe, 1974]
S. Skelboe.
Computation of Rational Interval Functions.
BIT, 14:87–95, 1974.

[Tsai and Morgan, 1984]
L. W. Tsai and A. P. Morgan.
Solving the Kinematics of the most general Six– and Five– Degree–of–Freedom Manipulators by Continuation Methods.
Technical Report Rep. GMR-4631, General Motors Research Labs., Warren, Mich., 1984.

[Ullman, 1973]
J. D. Ullman.
Fast Algorithms for the Elimination of Common Subexpressions.
Acta Informatica, 2:191–213, 1973.

[Wolfe, 1980]
M. A. Wolfe.
A Modification of Krawczyk's Algorithm.
SIAM J. Numerical Analysis, 17(3):376–379, 1980.

[Xiaojun and Deren, 1987]
C. Xiaojun and W. Deren.
On the Optimal Properties of the Krawczyk–Type Interval Operator.
Freiburger Intervall–Berichte, 5, 1987.

[Zuhe, 1988]
Shen Zuhe.
A Note on Moore's Interval Test for Zeros of Nonlinear Systems.
Computing, 40:85–90, 1988.

Vita

Address:

Volker Stahl
Research Institute for Symbolic Computation
Johannes Kepler Universität
A-4040 Linz, Austria
vstahl@risc.uni-linz.ac.at

Personal Data:

Born March 19, 1966 in Albstadt, Germany
German citizen, not married.

Education:

1976 - 1985	Primary and high school.
May 85	Graduation from high school, emphasis on mathematics and physics.
Oct. 86 - Aug. 91	Student at the University Erlangen in Computer Science.
Aug. 91	Graduation as Diplom Informatiker (equivalent to M.S. in Computer Science).
Oct. 91 - Dec. 95	Graduate student at the Research Institute for Symbolic Computation, University Linz (Prof. Buchberger).

Diploma Thesis:

Formale Untersuchung gebräuchlicher und unkonventioneller Datenkompressionsverfahren unter Berücksichtigung der Einsatzmöglichkeit hochparalleler Spezialhardware.

Awards:

1985	High school final exam with honors.
1991	University diploma exam with honors.
1992	DAAD scholarship.

Professional Experience:

Nov. 90 - Aug. 91	Working student at Siemens AG Erlangen, Medical Technology Department.
Jan. 93	Implementation of a symbolic algorithm for solving linear equation systems on a Sun workstation network at the University Tübingen, using the distributed thread system developed there.
Jul. 93 - Dec. 95	Development of a library for parallel constraint solving in a group of 12 graduate students.

Research Interests:

- Interval methods for constraint solving and optimization.
- Parallel computing.

Publications:

- V. Stahl: *Solving a System of Linear Equations with Modular Arithmetic on a MIMD Computer.* Technical Report No. 92-62, Research Institute for Symbolic Computation, Linz, 1992
- V. Stahl: *Exact Real Root Isolation with Sturm Sequences on a Shared Memory Multiprocessor.* Technical Report No. 93-27, Research Institute for Symbolic Computation, Linz, 1993
- H. Hong and V. Stahl: *Safe Start Regions by Fixed Points and Tightening.* Proc. of the SCAN (International Symposium on Scientific Computing, Computer Arithmetic and Validated Numerics), Wien, 1993
- H. Hong and V. Stahl: *Safe Start Regions by Fixed Points and Tightening.* Computing 53, 323-335, 1994
- V. Stahl: *The Overestimation Error of the Centered Form for Univariate Polynomials can be Reduced by Half.* Technical Report No. 94-03, Research Institute for Symbolic Computation, Linz, 1994
- V. Stahl: *Interval Horner Evaluation of Univariate Polynomials gives Sometimes the Range.* Technical Report No. 94-04, Research Institute for Symbolic Computation, Linz, 1994
- H. Hong and V. Stahl: *Bernstein Form is Inclusion Monotone.* Computing 55, 43-53, 1995