

Hur gör man matematik med datorer?

Warwick Tucker
Matematiska institutionen
Uppsala universitet
Box 480, 751 06 Uppsala
`warwick@math.uu.se`

24 november 2005

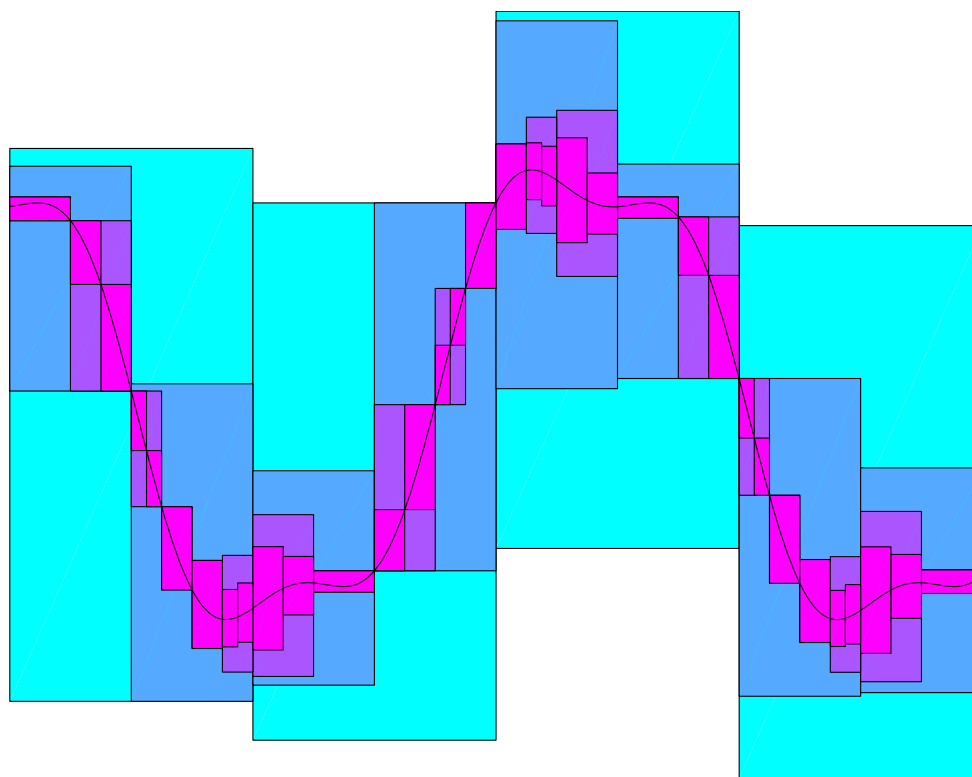
Hur gör man matematik med datorer?

Warwick Tucker

Matematiska institutionen

Uppsala universitet

warwick@math.uu.se



Kan man lita på flyttalsberäkningar?

Beräkningar med formatet `single` från C/C++

Exempel: Upprepad addition blir oförutsägbar:

$$\sum_{i=1}^{10^3} \langle 10^{-3} \rangle = 0.999990701675415,$$

$$\sum_{i=1}^{10^4} \langle 10^{-4} \rangle = 1.000053524971008.$$

Exempel: Summationsordningen är viktig:

$$1 + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{10^6} = 14.357357,$$

$$\frac{1}{10^6} + \dots + \frac{1}{3} + \frac{1}{2} + 1 = 14.392651.$$

Exempel: Betrakta funktionen

$$f(x, y) = 333.75y^6 + x^2(11x^2y^2 - y^6 - 121y^4 - 2) + 5.5y^8 + x/(2y)$$

beräknad i punkten $(x, y) = (77617, 33096)$.

IBM S/370 ($\beta = 16$) med FORTRAN:

format	p	$f(x, y)$
REAL*4	24	1.172603 ...
REAL*8	53	1.1726039400531 ...
REAL*10	64	1.172603940053178 ...

Pentium III ($\beta = 2$) med C/C++ (gcc/g++):

format	p	$f(x, y)$
float	24	178702833214061281280
double	53	178702833214061281280
long double	64	178702833214061281280

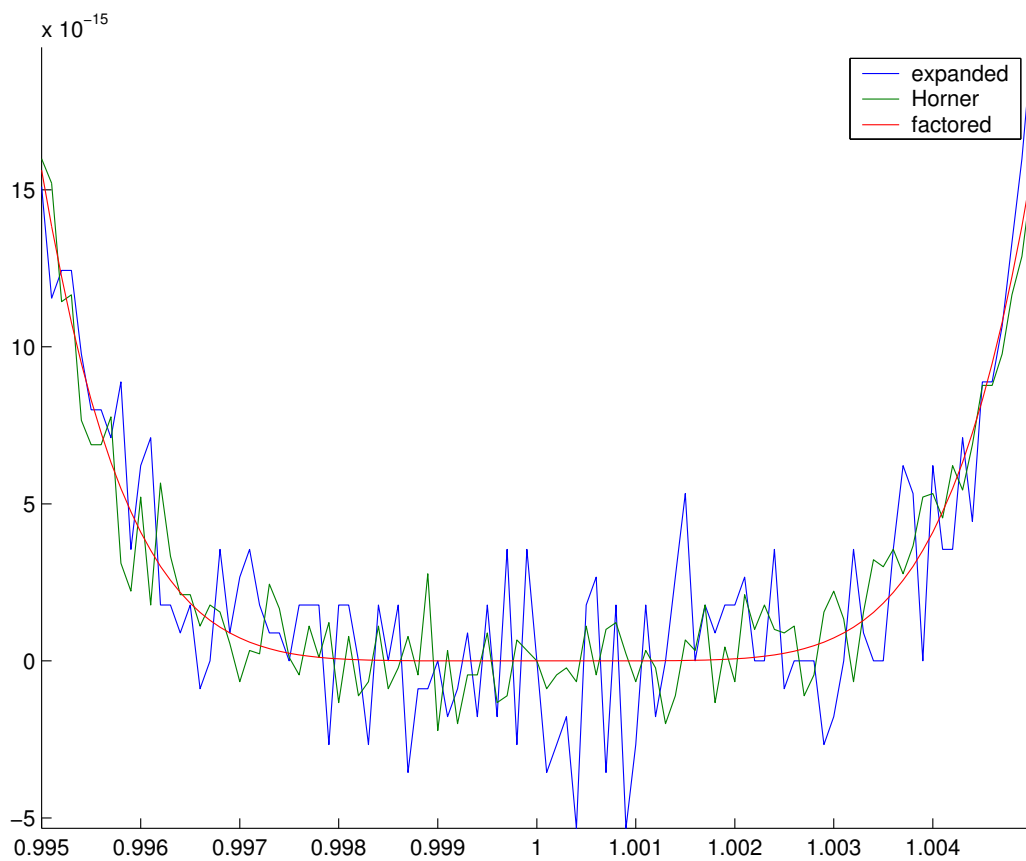
Trots all samtliga koefficienter är *exakt* representerbara i bas två, gör avrundningsfelen att resultatet blir helt missvisande:

Korrekt svar: $-0.8273960599 \dots$

Exempel: Finn samtliga rötter till polynomet

$$p(t) = t^6 - 6t^5 + 15t^4 - 20t^3 + 15t^2 - 6t + 1$$

MATLAB producerar följande graf:



Notera: $p(t) = (t - 1)^6$.

Är heltalsberäkningar tillförlitliga?

Exempel: Den harmoniska serien:

$$S_N = \sum_{k=1}^N \frac{1}{k} \quad \Rightarrow \quad \lim_{N \rightarrow \infty} S_N = +\infty.$$

Datorräkningen bekräftar detta resultat, men av helt fel anledning (heltalsfogning).

$$I_{max} + 1 = -I_{max} = I_{min}$$

Exempel: Elementära Taylorserier:

$$S_N = \sum_{k=0}^N \frac{1}{k!} \quad \Rightarrow \quad \lim_{N \rightarrow \infty} S_N = e^1.$$

Heltalsfogningen ger upphov till en talföljd som inte är växande (pga negativa termer).

N = 0	summa = 1.0000000000000000	
N = 1	summa = 2.0000000000000000	
N = 2	summa = 2.5000000000000000	
N = 3	summa = 2.6666666666666667	
N = 4	summa = 2.7083333333333333	
N = 5	summa = 2.7166666666666666	
N = 6	summa = 2.7180555555555555	
N = 7	summa = 2.718253968253968	
N = 8	summa = 2.718278769841270	
N = 9	summa = 2.718281525573192	
N = 10	summa = 2.718281801146385	
N = 11	summa = 2.718281826198493	
N = 12	summa = 2.718281828286169	
N = 13	summa = 2.718281828803753	
N = 14	summa = 2.718281829585647	
N = 15	summa = 2.718281830084572	
N = 16	summa = 2.718281830583527	
N = 17	summa = 2.718281827117590	Mindre!!!
N = 18	summa = 2.718281826004540	Mindre!!!
N = 19	summa = 2.718281835125155	
N = 20	summa = 2.718281834649448	Mindre!!!
N = 21	summa = 2.718281833812708	Mindre!!!
N = 22	summa = 2.718281831899620	Mindre!!!
N = 23	summa = 2.718281833059102	
N = 24	summa = 2.718281831770353	Mindre!!!
N = 25	summa = 2.718281832252007	
N = 26	summa = 2.718281831712599	Mindre!!!
N = 27	summa = 2.718281832386097	
N = 28	summa = 2.718281831659211	Mindre!!!
N = 29	summa = 2.718281830853743	Mindre!!!
N = 30	summa = 2.718281831563322	
N = 31	summa = 2.718281832917973	
N = 32	summa = 2.718281832452312	Mindre!!!
N = 33	summa = 2.718281831986650	Mindre!!!
N = 34	summa =	inf

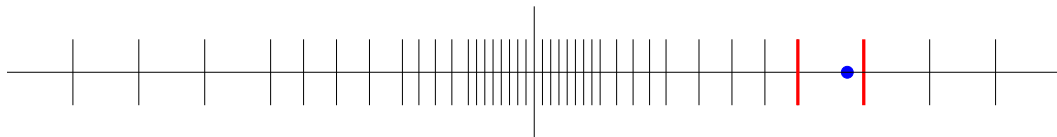
Hur kan vi kontrollera avrundningsfelen?

Avrunda alla delresultat åt båda hållen!

Om $x, y \in \mathbb{F}$ och $\star \in \{+, -, \times, \div\}$ kan vi innesluta det exakta resultatet i ett *intervall*:

$$x \star y \in [\nabla(x \star y), \Delta(x \star y)].$$

Eftersom alla (moderna) datorer avrundar med *maximal kvalitet* innesluter intervallet det exakta resultatet.



Fråga: Hur räknar man med intervall?

Intervall

Vi använder följande notation:

$$[a] = [\underline{a}, \bar{a}] = \{x \in \mathbb{R} : \underline{a} \leq x \leq \bar{a}\}.$$

Låt $\mathbb{I}\mathbb{R}$ beteckna mängden av alla kompakta intervall på \mathbb{R} :

$$\mathbb{I}\mathbb{R} = \{[a] : \underline{a}, \bar{a} \in \mathbb{R}; \underline{a} \leq \bar{a}\}.$$

Vi tillåter *tunna* intervall med $\underline{a} = \bar{a}$.

Exempel: $[1, \pi] \in \mathbb{I}\mathbb{R}$, men inte $[2, 1]$ eller $[1, \infty]$.

Vi skriver ofta x istället för $[x, x]$.

En metrik på $\mathbb{I}\mathbb{R}$

Vi kan förvandla $\mathbb{I}\mathbb{R}$ till ett metriskt rum genom att införa Hausdorffavståndet:

$$d([a], [b]) = \max\{|\underline{a} - \underline{b}|, |\bar{a} - \bar{b}|\}.$$

Då kan vi tala om konvergens:

$$\lim_{k \rightarrow \infty} [a_k] = [a] \quad \Leftrightarrow \quad \lim_{k \rightarrow \infty} d([a_k], [a]) = 0.$$

Aritmetik på \mathbb{R} :

Definition: Givet en operation $\star \in \{+, -, \times, \div\}$ och två operander $[a], [b] \in \mathbb{R}$, definierar vi

$$[a] \star [b] = \{a \star b : a \in [a], b \in [b]\},$$

med undantag för $[a] \div [b]$ då $0 \in [b]$.

Överuppräkneligt många fall att beakta!

Kontinuitet, monotonicitet, kompakthet \Rightarrow

$$[a] + [b] = [\underline{a} + \underline{b}, \bar{a} + \bar{b}]$$

$$[a] - [b] = [\underline{a} - \bar{b}, \bar{a} - \underline{b}]$$

$$[a] \times [b] = [\min\{\underline{a}\underline{b}, \underline{a}\bar{b}, \bar{a}\underline{b}, \bar{a}\bar{b}\}, \max\{\underline{a}\underline{b}, \underline{a}\bar{b}, \bar{a}\underline{b}, \bar{a}\bar{b}\}]$$

$$[a] \div [b] = [a] \times [1/\bar{b}, 1/\underline{b}], \quad \text{om } 0 \notin [b].$$

Egenskaper hos intervallaritmetiken

(1) IA är associativ och kommutativ.

(2) IA är *inte* distributiv:

$$[-1, 1]([-1, 0] + [3, 4]) = [-1, 1][2, 4] = [-4, 4],$$

$$[-1, 1][-1, 0] + [-1, 1][3, 4] = [-1, 1] + [-4, 4] = [-5, 5].$$

Dock gäller alltid

$$[a]([b] + [c]) \subseteq [a][b] + [a][c].$$

(3) IA saknar additiv och multiplikativ invers:

$$0 \in [a] - [a]; \quad 1 \in [a] \div [a].$$

(4) IA är *inklusionsmonoton*, dvs, om $[a] \subseteq [A]$ och $[b] \subseteq [B]$, så gäller

$$[a] \star [b] \subseteq [A] \star [B],$$

där vi kräver att $0 \notin [B]$ vid division.

Intervall-värda funktioner

Ett av huvudmålen är att innesluta värdemängden av en reellvärd funktion $f: D_f \rightarrow \mathbb{R}$.

$$R(f; S) = \{f(x) : x \in S\}.$$

Detta kan vi uppnå genom att konstruera en *intervallutvidgning* $F: \mathbb{R} \cap D_f \rightarrow \mathbb{R}$ av f .

Monotona funktioner

$$\begin{aligned} e^{[x]} &= [e^{\underline{x}}, e^{\bar{x}}] \\ \sqrt{[x]} &= [\sqrt{\underline{x}}, \sqrt{\bar{x}}] && \text{if } 0 \leq \underline{x} \\ \log [x] &= [\log \underline{x}, \log \bar{x}] && \text{if } 0 < \underline{x} \\ \arctan [x] &= [\arctan \underline{x}, \arctan \bar{x}]. \end{aligned}$$

Styckvis monotona funktioner

$$[x]^n = \begin{cases} [\underline{x}^n, \bar{x}^n] & : n \in \mathbb{Z}^+ \text{ är udda,} \\ [\text{mig}([x])^n, \text{mag}([x])^n] & : n \in \mathbb{Z}^+ \text{ är jämn,} \\ [1, 1] & : n = 0, \\ [1/\bar{x}, 1/\underline{x}]^{-n} & : n \in \mathbb{Z}^-; 0 \notin [x]. \end{cases}$$

Standard/elementära funktioner

Definiera klassen av *standardfunktioner* som

$$\mathfrak{S} = \{e^x, \log x, x^a, \text{abs } x, \sin x, \cos x, \tan x, \dots \\ \dots, \arccos x, \arctan x, \sinh x, \cosh x, \tanh x\}.$$

Givet ett $f \in \mathfrak{S}$, kan vi konstruera en *skarp* intervallutvidgning F , dvs

$$f \in \mathfrak{S} \quad \Rightarrow \quad R(f; [x]) = F([x]).$$

Elementära funktioner

Via ändliga konstruktioner av element från \mathfrak{S} bildar vi klassen av *elementära funktioner* \mathfrak{E} .

Sats: Om $f \in \mathfrak{E}$ och om $F([x])$ är väldefinierad, så gäller inklusionen

$$R(f; [x]) \subseteq F([x]).$$

Grafitning

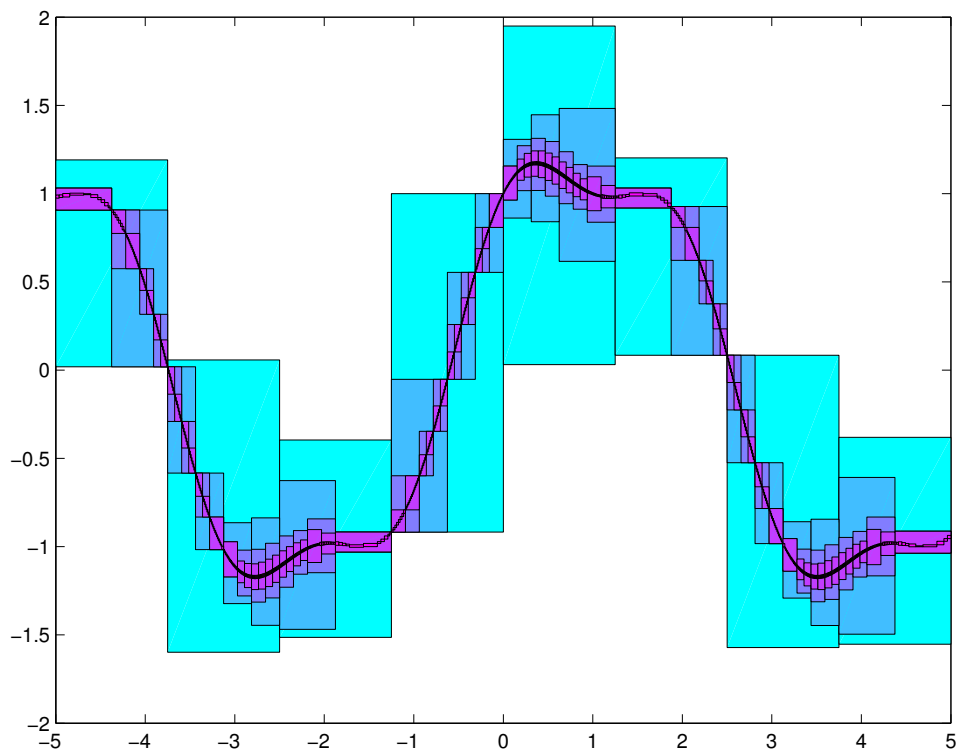
Exempel: Rita grafen till funktionen

$$f(x) = \cos^3 x + \sin x$$

på intervallet $[-5, 5]$.

Vi definierar $F([x]) = \cos^3 [x] + \sin [x]$, och partitionerar domänen tills följande gäller:

$$\max_i \{\text{width}(F([x_i]))\} \leq \text{TOL}.$$



Överuppskattningar

Utseendet är viktigt!

$$f_1(x) = 1 - x^2 = (1 - x)(1 + x) = f_2(x),$$

men

$$F_1([x]) = 1 - [x]^2 \neq (1 - [x])(1 + [x]) = F_2([x]),$$

ty

$$F_1([-1, 1]) = 1 - [-1, 1]^2 = [1, 1] - [0, 1] = [0, 1],$$

$$\begin{aligned} F_2([-1, 1]) &= (1 - [-1, 1])(1 + [-1, 1]) \\ &= [0, 2] \times [0, 2] = [0, 4]. \end{aligned}$$

Katastrofala överuppskattningar

$$\text{Om } f(x) = \frac{1}{1+x \cdot x}, \text{ så är } F([x]) = \frac{1}{1+[x] \cdot [x]}.$$

För intervallet $[x] = [-2, 2]$ får vi problem:

$$R(f; [-2, 2]) = \left[\frac{1}{5}, 1\right] \text{ men } F([-2, 2]) = \frac{1}{[-3, 5]}.$$

IA-resultatet är ej väldefinierat!

Utvidgad intervallaritmetik

Definition: Låt $\mathbb{R}^* = \{-\infty\} \cup \mathbb{R} \cup \{+\infty\}$. Givet en elementär funktion $f: D_f \rightarrow \mathbb{R}$ definierar vi $f^*: \mathcal{P}\mathbb{R}^* \rightarrow \mathcal{P}\mathbb{R}^*$ via

$$f^*(S) = R(f; S \cap D_f) \cup \left\{ \lim_{\zeta \rightarrow \zeta^*} f(\zeta) : \zeta \in D_f, \zeta^* \in S \right\}.$$

Exempel: Division med noll: Låt $f(x) = 1/x$. Då gäller

$$f^*([-3, 5]) = [-\infty, -1/3] \cup [1/5, +\infty]$$

$$f^*(0) = \{-\infty\} \cup \{+\infty\}.$$

Genom samma utvidgningsförfarande som tidigare erhålls $F^*: \mathbb{R}^* \rightarrow \mathbb{R}^*$ med egenskapen

$$f^*([x]) \subseteq F^*([x]) \quad \text{Gäller alltid!}$$

IA-resultatet är alltid väldefinierat!

Bisektionsmetoden

Kontrapositionen av inklusionsegenskapen ger

$$y \notin F^*([x]) \Rightarrow y \notin f^*([x]) \Rightarrow y \notin R(f; [x]).$$

Detta kan användas till att lokalisera nollställen.

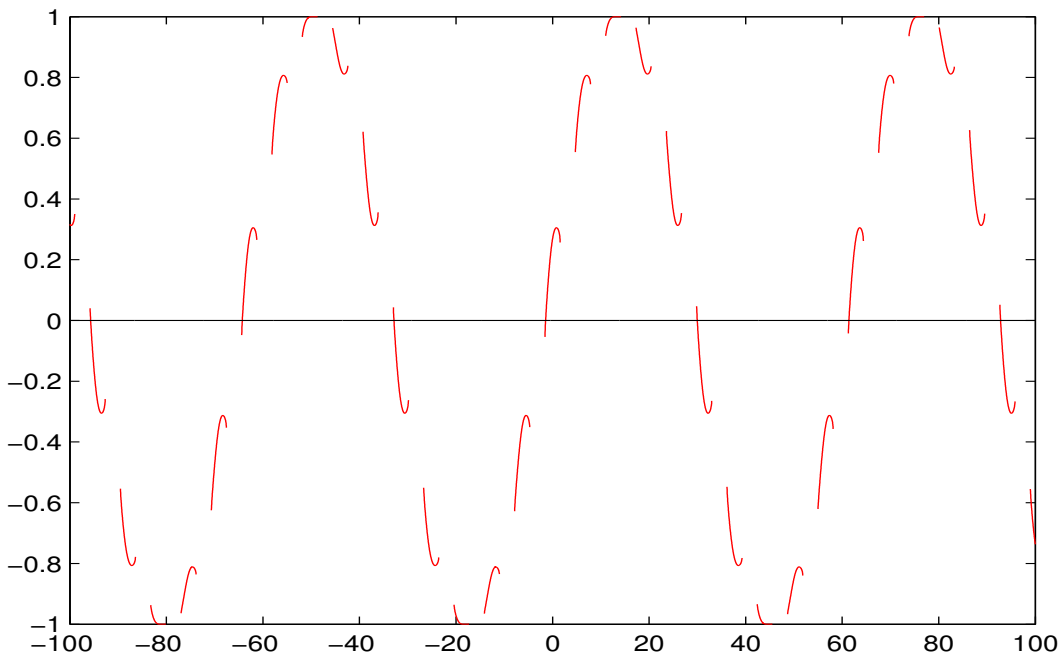
Strategi: Partitionera sökområdet och släng de delintervall där f bevisligen är nollskild.

En sådan IA-algoritm blir mycket kompakt, och fungerar även då D_f är okänd.

```
void bisect(pfcn F, interval X, double tol) {
    if ( subset(0.0, F(X)) ) // If zero is contained in F(X)
        if ( width(X) < tol ) // ... and the tolerance is met
            cout << X << endl; // ... print the subinterval.
        else { // Otherwise, divide and conquer.
            bisect(F, interval(min(X), mid(X)), tol);
            bisect(F, interval(mid(X), max(X)), tol);
        }
}
```

Bisektionsmetoden

Exempel: Låt $f(x) = \sin\left(\frac{x + e^{\sqrt{\cos x}}}{10}\right)$. Sök samtliga nollställen på intervallet $[-100, 100]$.



Test function : $f(x) = \sin\left(\frac{x + e^{\sqrt{\cos(x)}}}{10}\right)$

Search interval : $[-100, 100]$

Tolerance : $1e-14$

Ranges for zeros:

k=1 $[-95.6860825974053, -95.6860825974052]$ verified: Yes

k=2 $[-64.2701560615073, -64.2701560615073]$ verified: Yes

k=3 $[-32.8542295256094, -32.8542295256094]$ verified: Yes

k=4 $[-1.43830298971145, -1.43830298971144]$ verified: Yes

k=5 $[29.9776235461865, 29.9776235461865]$ verified: Yes

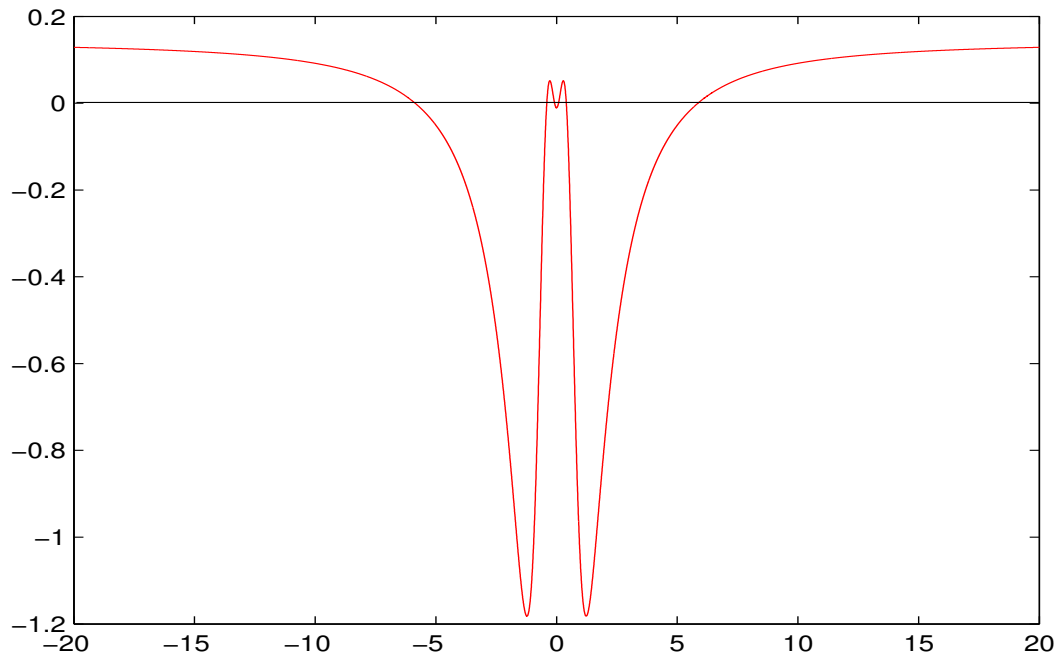
k=6 $[61.3935500820844, 61.3935500820844]$ verified: Yes

k=7 $[92.8094766179823, 92.8094766179824]$ verified: Yes

Number of bisection steps: 739

Bisektionsmetoden

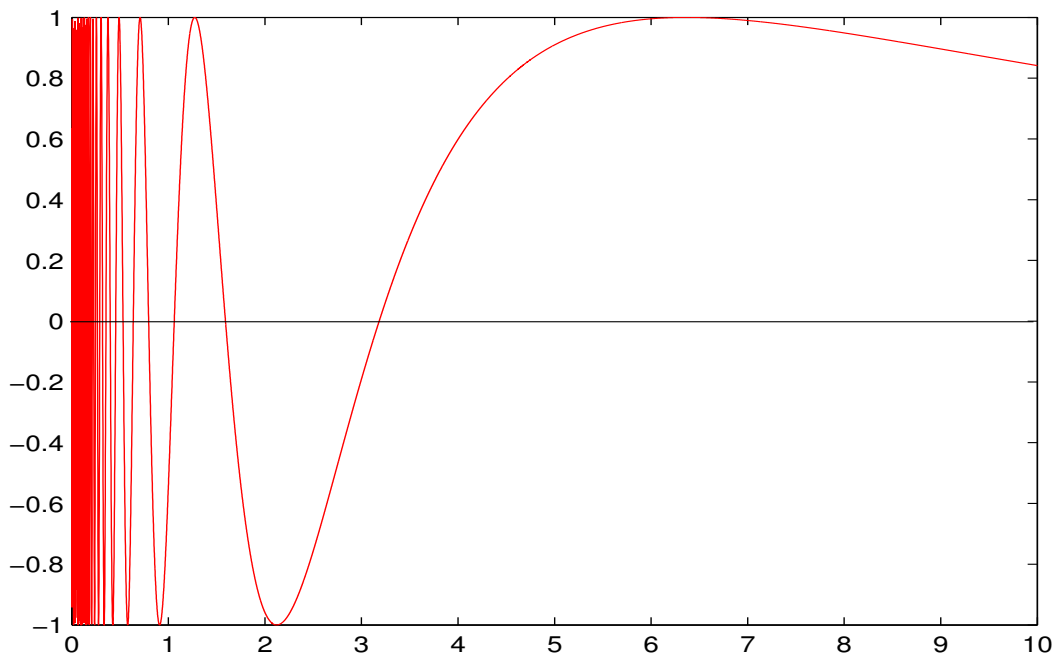
Exempel: Låt $f(x) = \sin(5/(1+x^2)+3) - e^{-x^2}$.
Sök samtliga nollställen på hela $\mathbb{R}^* = [-\infty, +\infty]$.



```
Test function      : f(x) = sin(5/(1 + x^2)+3) - exp(-x^2)
Search interval   : [ ENTIRE ]
Tolerance         : 1e-15
Ranges for zeros:
k=1 [-5.85769293576026, -5.85769293576024] verified: Yes
k=2 [-0.40422575825057, -0.40422575825055] verified: Yes
k=3 [-0.08054738301712, -0.08054738301709] verified: Yes
k=4 [ 0.08054738301709,  0.08054738301712] verified: Yes
k=5 [ 0.40422575825055,  0.40422575825057] verified: Yes
k=6 [ 5.85769293576024,  5.85769293576026] verified: Yes
Number of bisection steps: 5167
```

Bisektionsmetoden

Exempel: Låt $f(x) = \sin(10/x)$. Sök samtliga nollställen på $[0, +\infty]$.



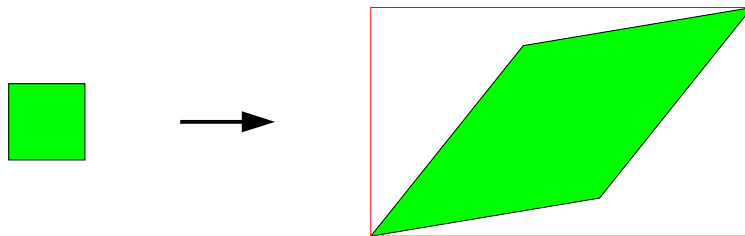
```
Test function      : f(x) = sin(10/x)
Search interval   : [0, inf]
Tolerance         : 1e-1
Ranges for zeros:
k=1 [0.000, 0.5625] verified: No
k=2 [0.625, 0.6875] verified: Yes
k=3 [0.750, 0.8125] verified: Yes
k=4 [1.000, 1.0625] verified: Yes
k=5 [1.5625, 1.625] verified: Yes
k=6 [3.125, 3.1875] verified: Yes
k=7 [ +INFTY ]     verified: No
Number of bisection steps: 2103
```

Högre dimensioner

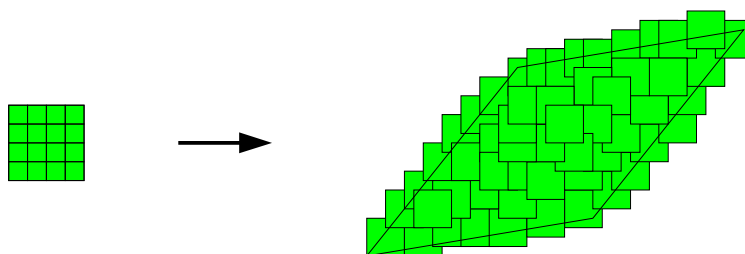
Precis som förr, fast nu arbetar vi på \mathbb{R}^n - rummet av kompakta n -dimensionella lådor:

$$[x] = ([x_1], \dots, [x_n]).$$

Packeteringseffekten: I högre dimensioner tillhör $R(f; [x])$ nästan aldrig \mathbb{R}^n .



Överuppskattningen kan reduceras...



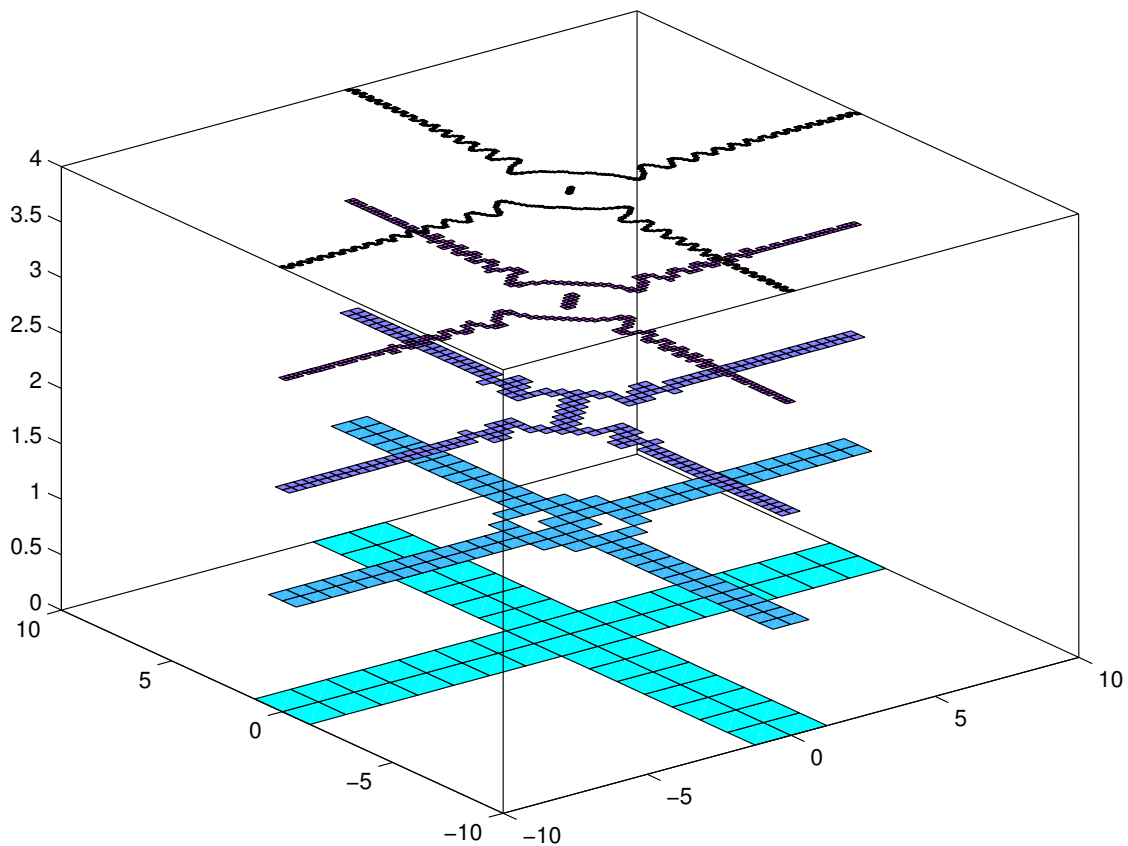
... men det kostar!

Nivåkurvor

Givet $f: \mathbb{R}^2 \rightarrow \mathbb{R}$, finn nivåkurvorna

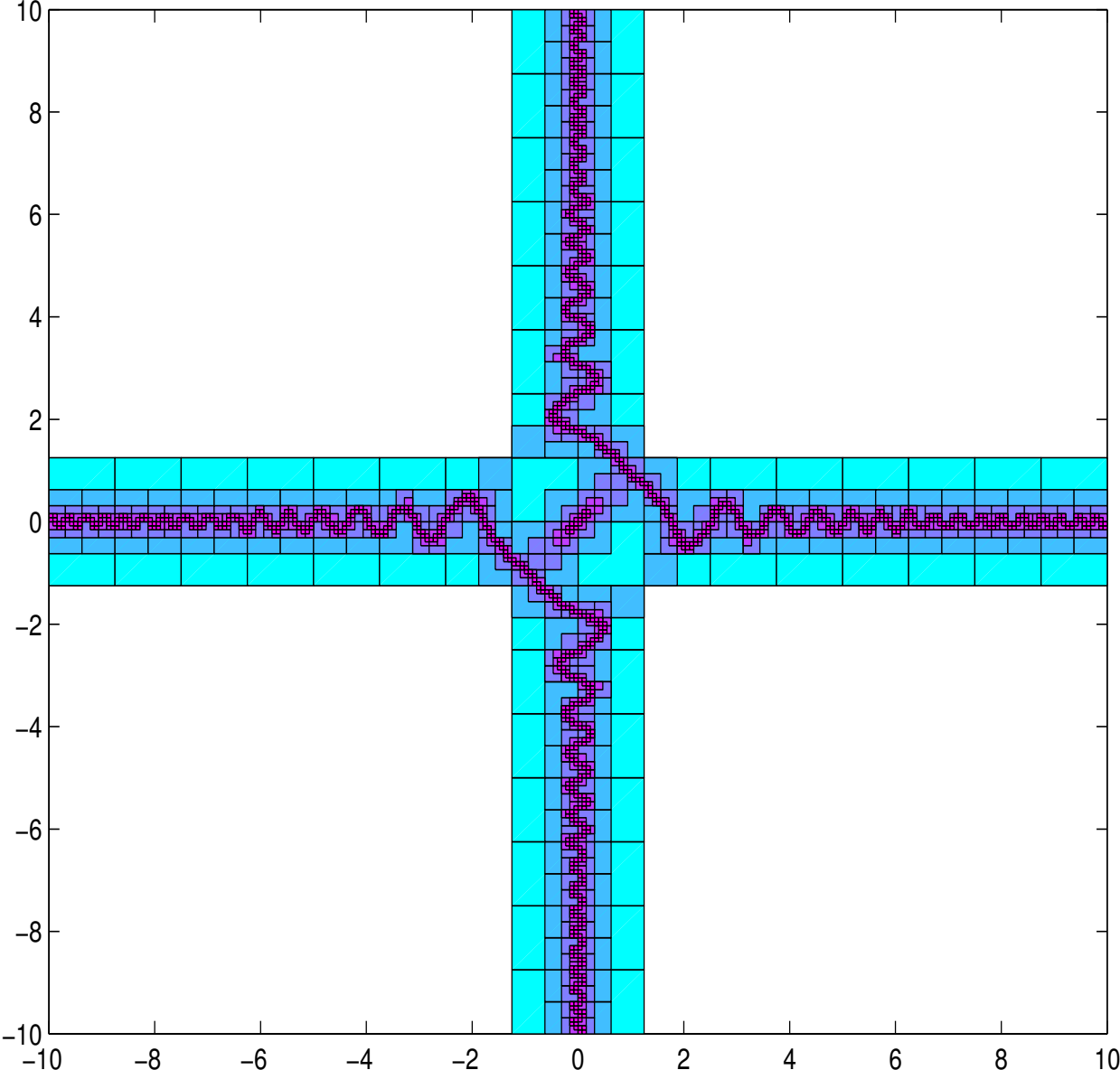
$$\Gamma(f, \alpha) = \{(x, y) \in \mathbb{R}^2 : f(x, y) = \alpha\}.$$

Exempel: Beräkna $\Gamma(f, 0)$ på $[-10, 10] \times [-10, 10]$ för funktionen $f(x, y) = \sin(x^2 + y^2) - xy$.



Nivåkurvor

Sett ovanifrån:

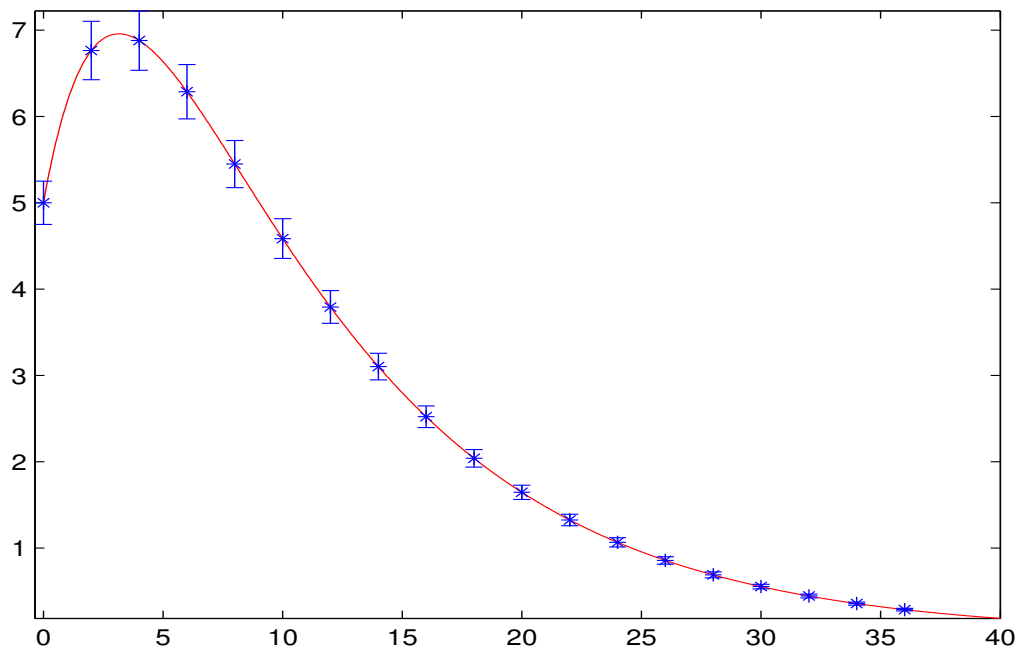


Parameterrekonstruktion

Exempel: Givet en modell

$$f(x; p) = 15e^{-p_1x} - 10e^{-p_2x}$$

och (kontaminerad) mätdata $\{(x_i, y_i)\}_{i=1}^N$, finn parametern $p = (p_1, p_2)$ som passar datamängden bäst.



Lösning: Tänk på datamängden som $\{(x_i, [y_i])\}_{i=1}^N$.

Parameterrekonstruktion

Vi säger att parametern $[p] = ([p_1], [p_2])$ är *konsistent* med datamängden om

$$C([p]) = \bigwedge_{i=1}^N \left([y_i] \cap F(x_i; [p]) \neq \emptyset \right)$$

har värdet sant. Partitionera parameterutrymmet och släng bort alla $[p]$ med $C([p]) = \text{falskt!}$

Notera: Ingen global optimering; ingen minsta-kvadratanpassning krävs.

Exempel (forts): Låt $p = (0.11, 0.32)$ och tag noderna $x_i = 4(i - 1)$ för $i = 1, \dots, 10$.

Med ett relativt fel på 5% får vi

$$p \in \left([0.108624, 0.1111432], [0.286606, 0.358746] \right).$$

Med ett relativt fel på 1% får vi

$$p \in \left([0.109721, 0.110281], [0.312931, 0.327268] \right).$$

Sammanfattning

- (1) I.A. bygger på mängdvärd matematik;
- (2) Hanterar avrundnings-/diskretiseringsfel;
- (3) Hanterar singulariteter/domäner;
- (4) Ger ett matematiskt korrekt resultat.
- (5) Ger ett kvalitetsmått på resultatet;

Artiklar som använder I.A.

D. Gabai, G. R. Mayerhoff, and N. Thurston, *Homotopy hyperbolic 3-manifolds are hyperbolic*. Annals of Mathematics, **157**, 335–431, 2003.

J. Hass, M. Hutchings, and R. Schlafly, *The Double Bubble Conjecture*. Electr. Research Announcements of the Amer. Math. Soc., **1**, 98–102, 1995.

T. C. Hales, *Some algorithms arising in the proof of the Kepler conjecture*. Discrete and computational geometry, **25**, 489–507, Algorithms Combin., 2003.

K. Makino and M. Berz, *Taylor Models and Other Validated Functional Inclusion Methods*. Int. J. of Pure and Appl. Math., **4**, 379–456, 2003.

I. Mitrea, W. Tucker, *Some Counterexamples for the Spectral Radius Conjecture*. Differential and Integral Equations, **16:12**, 1409–1439, 2003.

N. S. Nedialkov, K. R. Jackson and G. F. Corliss, *Validated Solutions of Initial Value Problems for Ordinary Differential Equations*. J. Applied Math. and Comp., **105**, 21–68, 1999.

P. Zgliczynski, *Attracting fixed points for the Kuramoto-Sivashinsky equation*, SIAM J. Applied Dynamical Systems, **1:2**, 215–235, 2002.

Referenser:

G. Alefeld and J. Herzberger, *Introduction to Interval Computations*. Academic Press, New York, 1983.

W. Kahan, *IEEE Standard 754 for Binary Floating-Point Arithmetic*. lecture notes, 1996. Available from <http://www.cs.berkeley.edu/~wkahan/>

E. N. Lorenz, *Deterministic Non-periodic Flow*. J. Atmos. Sci. **20** 130–141, 1963.

R. E. Moore, *Interval Analysis*. Prentice-Hall, Englewood Cliffs, New Jersey, 1966.

T. Sunaga, *Theory of interval algebra and its application to numerical analysis*. In: RAAG Memoirs, Ggujutsu Bunken Fukuy-kai. Tokyo, **2** 29–46, 1958.

W. Tucker, *A Rigorous ODE Solver and Smale's 14th Problem*, Found. Comput. Math. **2:1**, 53–117, 2002.

M. Warmus, *Calculus of Approximations*. Bulletin de l'Academie Polonaise de Sciences, **4:5** 253–257, 1956.

R. C. Young, *The algebra of multi-valued quantities*. Mathematische Annalen, **104** 260–290, 1931.

Interval Computations Web Page

<http://www.cs.utep.edu/interval-comp>