Root Finding Implementations

Dr Barry Wardell
School of Mathematics and Statistics
University College Dublin

We can get higher orders of convergence by using higher order approximations to f

Students often think it might be a good idea to develop formulas to enhance the exponent [p, the order of convergence] in this estimate to 3 or 4. However, this is an illusion. Taking two steps at a time of a quadratically convergent algorithm yields a quartically convergent one – so the difference in efficiency between quadratic and quartic is at best a constant factor. The same goes if the exponent 2, 3, or 4 is replaced by any other number greater than 1. The true distinction is between all of these algorithms that converge super-linearly, of which Newton's method is the prototype, and those that converge linearly or geometrically, where the exponent is just 1.

Lloyd N. Trefethen's essay *Numerical Analysis*,

Consider an algorithm with an order of convergence p

$$|x_{k+1} - x\ | = e_{k+1} = c\, e_k^p,$$

Error for two successive iterates

$$e_{k+2} = c\, e_{k+1}^p = c(c\, e_k^p)^p = c^{p+1}\, e_k^{p^2}.$$

Construct a new algorithm that takes two steps of the old p-convergent algorithm at each iteration

New algorithm

$$\tilde{e}_{k+1} = \tilde{c}\,\tilde{e}_k^{p^2}\,.$$

The new order of convergence is p^2

If $p = 1$ then $p^2 = 1$

If $p > 1$ then $p^2 > p$

The true distinction between algorithms is
whether they have linear or super-linear convergence

Stopping Rules

1. f-sequence convergence.

2. x-sequence convergence.

3. Cycling, no f or x sequence convergence.

4. f or x sequence divergence.

5. Cannot proceed, e.g., about to divide by zero or some other exception.

1. $f-$ Convergence

is often tested with

$$|f(x_k)-0| \leq \epsilon_f \quad \text{that is} \quad |f(x_k)| \leq \epsilon_f.$$

If ϵ_f is below the underflow threshold ($x_{\min} = 10^{-38}$ IEEE single precision) use

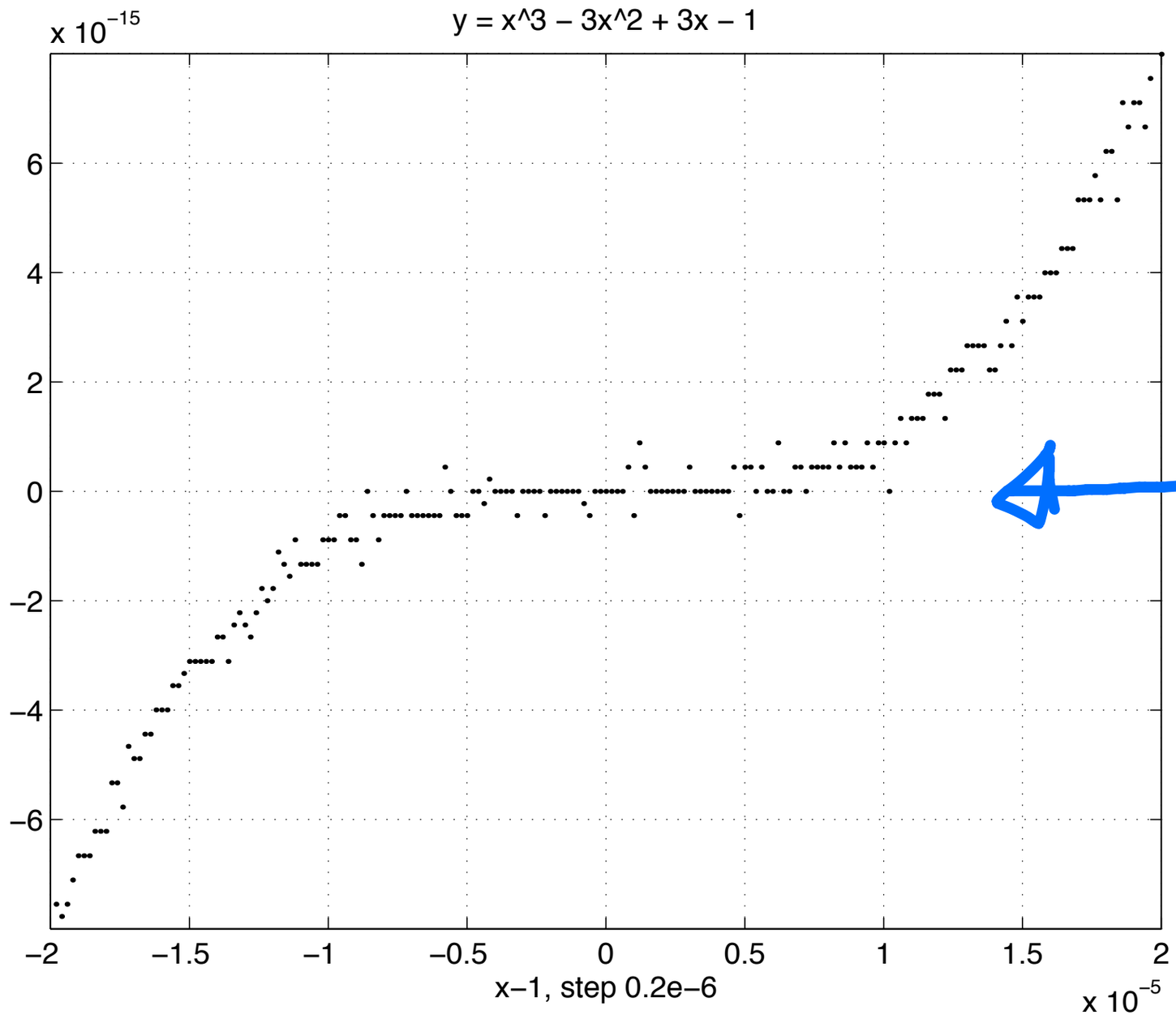
$$|f(x_k)| = 0,$$

if $\epsilon_f > x_{min}$ what should value of ϵ_f be.

Cannot answer without looking at x convergence.

Common.

The f-convergence test is chosen to be $|f(x_k)| = 0.0$



Probability of hitting $f=0$ is high.

The Function $x^3 - 3x^2 + 3x - 1$ in Machine Numbers

2. $x-$ Convergence

Expand $f(x_k)$ about the zero

$$f(x_k)=f(x)+\frac{1}{1!}f'(x)(x_k-x)+\frac{1}{2!}f''(x)(x_k-x)^2+\cdots+\frac{1}{n!}f^{(n)}(x)(x_k-x)^n+R_{n+1}.$$

then approximately

$$f(x_k)\approx f(x)+f'(x)(x_k-x) \quad \text{or} \quad f(x_k)-f(x)=f'(x)(x_k-x).$$

$$\Delta f_k = f'(x)\Delta x_k\,.$$

choose ϵ_f and ϵ_x according to the following rule :

$\Delta f_k = f'(x)\,\Delta x_k$

1. If $f'(x) \approx 1$ then choose $\epsilon_f \approx \epsilon_x$.

2. If $f'(x) \ll 1$ then choose $\epsilon_f \ll \epsilon_x$.

3. If $f'(x) \gg 1$ then choose $\epsilon_f \gg \epsilon_x$.

Problem is that in general we don't know $f'(x)$
Forced to choose ϵ_f and ϵ_x independently
Then one or other of the tests is reasonable in reasonable circumstances.

Definition: Machine Epsilon is distance between 1 and next highest f.p.n.

Spacing about a f.p.n is $\varepsilon_m |x|$

For IEEE Double Precision $\varepsilon_m = 2^{-52} \approx 10^{-16}$

$|x_k| \approx 10^{12}$

$\varepsilon_m |x_k| \approx 10^{-4}$

$|x_k| \approx 10^{-3}$

$\varepsilon |x_k| \approx 10^{-19}$

The x- sequence convergence test must be a Cauchy-type test

$|x_k| \approx 10^{12}$

$\varepsilon_m |x_k| \approx 10^{-4}$

Common but wrong

IF $|x_k - x_{k-1}| \leq 10^{-6}$ THEN STOP

$|x_k| \approx 10^{-3}$

$\varepsilon_m |x_k| \approx 10^{-19}$

Better

IF $|x_k - x_{k-1}| \leq \boldsymbol{\varepsilon}_m \times |x_k|$ THEN STOP

Full Precision.

Best

IF $|x_k - x_{k-1}| \leq \epsilon_{\text{tol}} + \boldsymbol{\varepsilon}_m \times |x_k|$ THEN STOP

Dont always need full precision.

3. Cycling

If neither f nor x sequence occurs
Stop after a set number of iterations

4. Divergence

This occurs when either x_k or $f(x_k) \longrightarrow \infty$.

Check for ∞.

5. Cannot Proceed

. Check for divide by zero or some other exception.

Exercise Does the test

$$\text{if} \quad |x_k - x_{k-1}| \leq \qquad \varepsilon_m \times |x_k| \quad \text{then stop}$$

always work. What happens if the algorithm is converging to $x = 0$? Consider finding a zero of the function $f(x) = \tan^{-1} x$.

IMPLEMENTATION NOTES

- Direct translation of an algorithm into a program rarely works.

- Remember numbers have a finite precision and the rules of arithmetic do not always work.

- Write iteration formulae in correction form. Example:

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} = x_k + \Delta x_k,$$

$$m := a + \frac{(b-a)}{2} \quad \text{rather than} \quad m := \frac{(a+b)}{2}$$

$$X_{k+1} = X_k - \frac{f(X_k)}{f'(X_k)} \qquad \text{ie.} \quad X_k + \Delta X_k$$

↑ 'good' / 'bad'
usually small

Mathematically equivalent

$$X_{k+1} = \frac{f'(X_k)X_k - f(X_k)}{f'(X_k)}$$

Entire calculation is contaminated.

- The subtraction of nearly-equal numbers may cause difficulties.

- Always check for division by zero.

Specific Rules

Newton's Algorithm

Always write iteration formulas in *correction form*. For example, the iteration formula for Newton's Algorithm is

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} = x_k + \Delta x_k,$$

Always check for division by zero.

$$\text{check for } f'(x_k) = 0.$$

Bisection Algorithm

hardest to implement correctly.

common mistakes

1. **Sign Test : if** $f(a)f(b) < 0$ **then.**
2. **Calculating the Mid-Point :** $m := (a+b)/2.0$.
3. **Convergence Test : if** $|a-b| \leq 10^{-6}$ **then.**

(1) $f(a) = 10^{-20}$, $f(b) = -10^{-20}$

in single precisim product underflows and set $= 0$. Test incorreethy fails.

Common failw as you spend time around $f = 0$.

NASTY ERROR.

(2) $m = \frac{a+b}{2}$

3 digit precision round to nearest

$a = .982, \; b = .984$

$fl(fl(1.966)/2) = fl(1.97/2) = .985$

outside interval.

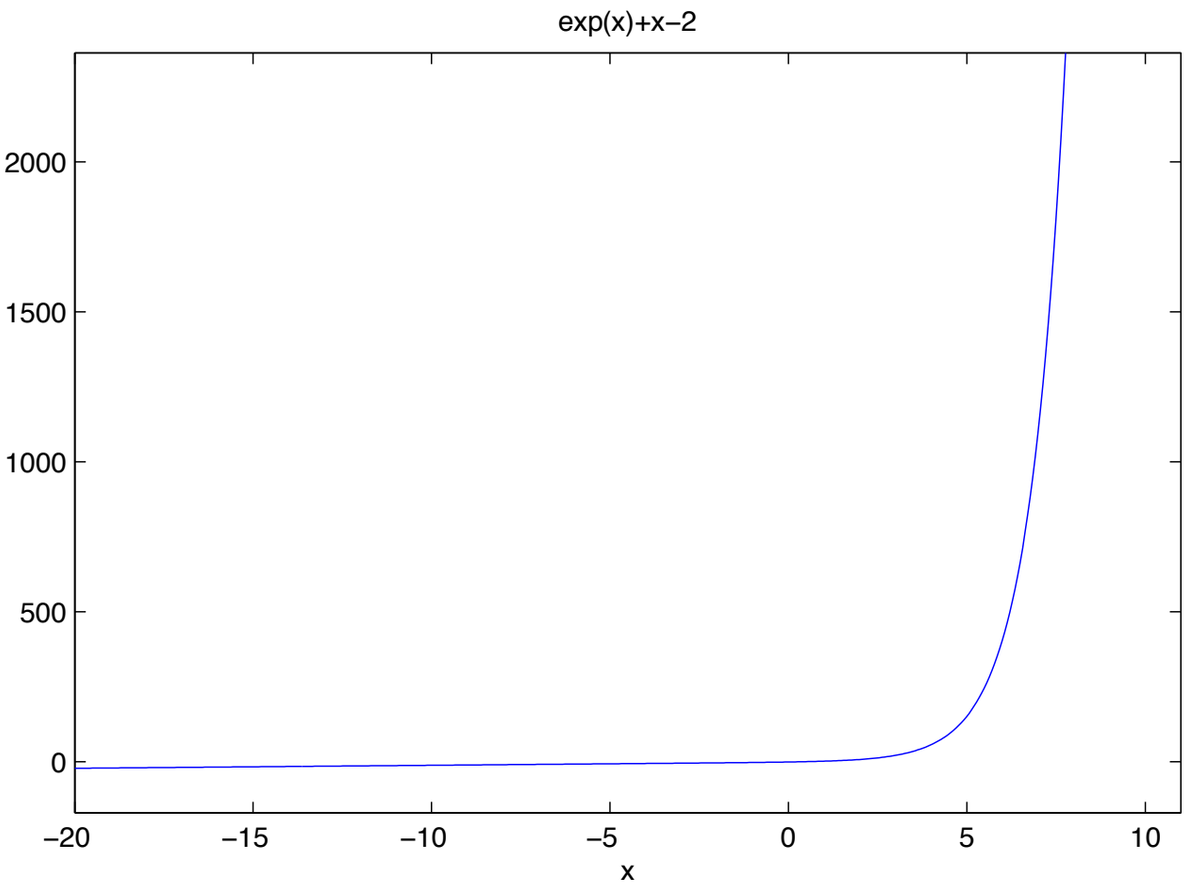
[With binary arithmetic, if there is no overflow or underflow this problem will not occur. If a and b are representable, then so is $2a, 2b$. If $a \le b$ then so is $2a \le a+b \le 2b$. The computed $\frac{a+b}{2}$ will be in $[a,b]$ (Montgomery, Microsoft). But overflow can be a problem If a, b close to overflow threshold then $a+b$ will overflow and $\frac{a+b}{2}$ will be wrong. $m = a + \frac{b-a}{2}$ will not cause overflow.

```
>> a= 2^1023; b=2^1023;
```

(3) Meaningless.

Some Examples

Example Find the zero of the function $f(x) = e^x + x - 2$ in the range $[-20, 11]$ to full precision.



: Bisection : $f(x) = e^x + x - 2$

k	a	b	E_{rel}	$f(a)$	$f(b)$
1	−20.000000	11.000000	6.888889	−22.000000	59883.141715
2	−4.500000	11.000000	4.769231	−6.488891	59883.141715
3	−4.500000	3.250000	12.400000	−6.488891	27.040340
4	−0.625000	3.250000	2.952381	−2.089739	27.040340
5	−0.625000	1.312500	5.636364	−2.089739	3.027951
6	0.343750	1.312500	1.169811	−0.246024	3.027951
7	0.343750	0.828125	0.826667	−0.246024	1.117148
⋮					
56	0.442854	0.442854	0.000000	−0.000000	0.000000
57	0.442854	0.442854	0.000000	−0.000000	0.000000
58	0.442854	0.442854	0.000000	−0.000000	0.000000

if $E_{rel} \approx 2^{-1}$ initially
expect 52,53 iterations

$$\frac{|a-b|}{|a|}$$

Transient
settling

L - bounded around

Transient and then monotone convergence

Newton : $f(x) = e^x + x - 2$

k	x_{old}	x_{new}	E_{rel}	$f(x_{\text{old}})$	$f'(x_{\text{old}})$
1	−20.000000	2.000000	11.000000	−22.000000	1.000000
2	2.000000	1.119203	0.786986	7.389056	8.389056
3	1.119203	0.582178	0.922440	2.181615	4.062412
4	0.582178	0.448802	0.297184	0.372112	2.789933
5	0.448802	0.442865	0.013405	0.015236	2.566434
6	0.442865	0.442854	0.000024	0.000028	2.557162
7	0.442854	0.442854	0.000000	0.000000	2.557146

$$\left| \frac{x_{k+1} - x_k}{x_k} \right|$$

Secant : $f(x) = e^x + x - 2$

k	x_1	x_2	E_{rel}	$f(x_1)$	$f(x_2)$
1	11.000000	−19.988615	1.550313	59883.141	−21.988615
2	−19.988615	−19.977241	0.000569	−21.988615	−21.977241
3	−19.977241	2.000000	10.988621	−21.977241	7.389056
4	2.000000	−3.529845	1.566597	7.389056	−5.500535
5	−3.529845	−1.170025	2.016896	−5.500535	−2.859666
6	−1.170025	1.385306	1.844597	−2.859666	3.381356
7	1.385306	0.000840	1648.438056	3.381356	−0.998320
8	0.000840	0.316420	0.997346	−0.998320	−0.311373
9	0.316420	0.459464	0.311326	−0.311373	0.042688
10	0.459464	0.442217	0.039000	0.042688	−0.001629
11	0.442217	0.442851	0.001431	−0.001629	−0.000008
12	0.442851	0.442854	0.000007	−0.000008	0.000000
13	0.442854	0.442854	0.000000	0.000000	−0.000000

Multiple Roots

Bisect on $(x-1)^5$

k	a	b	E_{rel}	$f(a)$	$f(b)$
1	$-2.000000e+001$	$1.100000e+001$	$6.888889e+000$	$-4.084101e+006$	$1.000000e+005$
2	$-4.500000e+000$	$1.100000e+001$	$4.769231e+000$	$-5.032844e+003$	$1.000000e+005$
3	$-4.500000e+000$	$3.250000e+000$	$1.240000e+001$	$-5.032844e+003$	$5.766504e+001$
4	$-6.250000e-001$	$3.250000e+000$	$2.952381e+000$	$-1.133096e+001$	$5.766504e+001$
5	$-6.250000e-001$	$1.312500e+000$	$5.636364e+000$	$-1.133096e+001$	$2.980232e-003$
6	$3.437500e-001$	$1.312500e+000$	$1.169811e+000$	$-1.217157e-001$	$2.980232e-003$
7	$8.281250e-001$	$1.312500e+000$	$4.525547e-001$	$-1.499904e-004$	$2.980232e-003$
8	$8.281250e-001$	$1.070313e+000$	$2.551440e-001$	$-1.499904e-004$	$1.718552e-006$
$\vdots$					
54	$1.000000e+000$	$1.000000e+000$	$3.441691e-015$	$-6.262791e-075$	$3.187232e-074$
55	$1.000000e+000$	$1.000000e+000$	$1.665335e-015$	$-6.262791e-075$	$5.397605e-079$
56	$1.000000e+000$	$1.000000e+000$	$8.881784e-016$	$-1.311618e-076$	$5.397605e-079$

flpt min

No problem.

$\sim 10^{-508}$

Newton on $(x-1)^5$

k	x_{old}	x_{new}	E_{rel}	$f(x_{\text{old}})$	$f'(x_{\text{old}})$
1	$4.000000e+000$	$3.400000e+000$	$1.764706e-001$	$2.430000e+002$	$4.050000e+002$
2	$3.400000e+000$	$2.920000e+000$	$1.643836e-001$	$7.962624e+001$	$1.658880e+002$
3	$2.920000e+000$	$2.536000e+000$	$1.514196e-001$	$2.609193e+001$	$6.794772e+001$
4	$2.536000e+000$	$2.228800e+000$	$1.378320e-001$	$8.549802e+000$	$2.783139e+001$
5	$2.228800e+000$	$1.983040e+000$	$1.239309e-001$	$2.801599e+000$	$1.139974e+001$
6	$1.983040e+000$	$1.786432e+000$	$1.100562e-001$	$9.180280e-001$	$4.669332e+000$
7	$1.786432e+000$	$1.629146e+000$	$9.654533e-002$	$3.008194e-001$	$1.912558e+000$
8	$1.629146e+000$	$1.503316e+000$	$8.370102e-002$	$9.857251e-002$	$7.833839e-001$
9	$1.503316e+000$	$1.402653e+000$	$7.176635e-002$	$3.230024e-002$	$3.208741e-001$
$\vdots$					
157	$1.000000e+000$	$1.000000e+000$	$4.440892e-016$	$8.692897e-074$	$1.779515e-058$
158	$1.000000e+000$	$1.000000e+000$	$4.440892e-016$	$3.187232e-074$	$7.974454e-059$
159	$1.000000e+000$	$1.000000e+000$	$2.220446e-016$	$9.071755e-075$	$2.918254e-059$

!! Linear Convergence.

Example (Failure of the Standard Convergence Test).

$$\text{if} \quad |x_k - x_{k-1}| \leq \epsilon_{\text{tol}} + 4.0 \times \boldsymbol{\varepsilon}_m \times |x_k| \quad \text{then stop}$$

This test has a 'flaw'. It will fail if the sequence $\{x_k\} \to 0$.

example, with $f(x) = \tan^{-1} x$,

$\tan^{-1} x = 0 \Rightarrow x = 0$.

Table Apparent Failure of Convergence Test on $\tan^{-1} x$

k	a	b	E_{rel}	$f(a)$	$f(b)$
1	$-2.000000e+001$	$1.100000e+001$	$6.888889e+000$	$-1.520838e+000$	$1.480136e+000$
2	$-4.500000e+000$	$1.100000e+001$	$4.769231e+000$	$-1.352127e+000$	$1.480136e+000$
3	$-4.500000e+000$	$3.250000e+000$	$1.240000e+001$	$-1.352127e+000$	$1.272297e+000$
4	$-6.250000e-001$	$3.250000e+000$	$2.952381e+000$	$-5.585993e-001$	$1.272297e+000$
5	$-6.250000e-001$	$1.312500e+000$	$5.636364e+000$	$-5.585993e-001$	$9.197196e-001$
6	$-6.250000e-001$	$3.437500e-001$	$6.888889e+000$	$-5.585993e-001$	$3.310961e-001$
7	$-1.406250e-001$	$3.437500e-001$	$4.769231e+000$	$-1.397089e-001$	$3.310961e-001$
8	$-1.406250e-001$	$1.015625e-001$	$1.240000e+001$	$-1.397089e-001$	$1.012154e-001$
9	$-1.953125e-002$	$1.015625e-001$	$2.952381e+000$	$-1.952877e-002$	$1.012154e-001$
⋮					
97	$-1.135960e-028$	$2.776790e-028$	$4.769231e+000$	$-1.135960e-028$	$2.776790e-028$
98	$-1.135960e-028$	$8.204153e-029$	$1.240000e+001$	$-1.135960e-028$	$8.204153e-029$
99	$-1.577722e-029$	$8.204153e-029$	$2.952381e+000$	$-1.577722e-029$	$8.204153e-029$
100	$-1.577722e-029$	$3.313216e-029$	$5.636364e+000$	$-1.577722e-029$	$3.313216e-029$

Why does Bisection take so long to converge?

No convergence

Table Convergence of Bisect on $\tan^{-1} x$

k	a	b	E_{rel}	$f(a)$	$f(b)$
1	$-2.000000e+001$	$1.100000e+001$	$6.888889e+000$	$-1.520838e+000$	$1.480136e+000$
2	$-4.500000e+000$	$1.100000e+001$	$4.769231e+000$	$-1.352127e+000$	$1.480136e+000$
3	$-4.500000e+000$	$3.250000e+000$	$1.240000e+001$	$-1.352127e+000$	$1.272297e+000$
4	$-6.250000e-001$	$3.250000e+000$	$2.952381e+000$	$-5.585993e-001$	$1.272297e+000$
5	$-6.250000e-001$	$1.312500e+000$	$5.636364e+000$	$-5.585993e-001$	$9.197196e-001$
6	$-6.250000e-001$	$3.437500e-001$	$6.888889e+000$	$-5.585993e-001$	$3.310961e-001$
7	$-1.406250e-001$	$3.437500e-001$	$4.769231e+000$	$-1.397089e-001$	$3.310961e-001$
8	$-1.406250e-001$	$1.015625e-001$	$1.240000e+001$	$-1.397089e-001$	$1.012154e-001$
9	$-1.953125e-002$	$1.015625e-001$	$2.952381e+000$	$-1.952877e-002$	$1.012154e-001$
$\vdots$					
1021	$-1.780059e-306$	$9.790325e-307$	$6.888889e+000$	$-1.780059e-306$	$9.790325e-307$
1022	$-4.005133e-307$	$9.790325e-307$	$4.769231e+000$	$-4.005133e-307$	$9.790325e-307$
1023	$-4.005133e-307$	$2.892596e-307$	$1.240000e+001$	$-4.005133e-307$	$2.892596e-307$
$\vdots$					
1075	$-4.940656e-323$	$1.037538e-322$	$5.166667e+000$	$-4.940656e-323$	$1.037538e-322$
1076	$-4.940656e-323$	$2.964394e-323$	$8.000000e+000$	$-4.940656e-323$	$2.964394e-323$
1077	$-9.881313e-324$	$2.964394e-323$	$4.000000e+000$	$-9.881313e-324$	$2.964394e-323$
1078	$-9.881313e-324$	$9.881313e-324$	Inf	$-9.881313e-324$	$9.881313e-324$

Test $|a-b| \leq \varepsilon_m |a|$

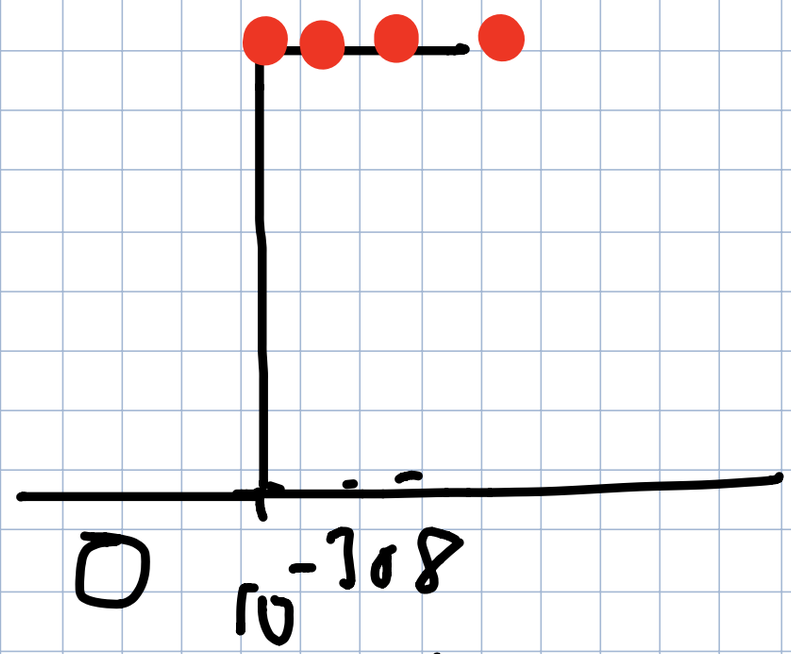
Satisfied when a, b underflow to zero.

Smallest f.p.n.

$\sim 10^{-308}$

Smallest floating pt. no- is

$2.2250738585072201 \times 10^{-308}$



gradual underflow

$4.940656458412465 \times 10^{-324}$

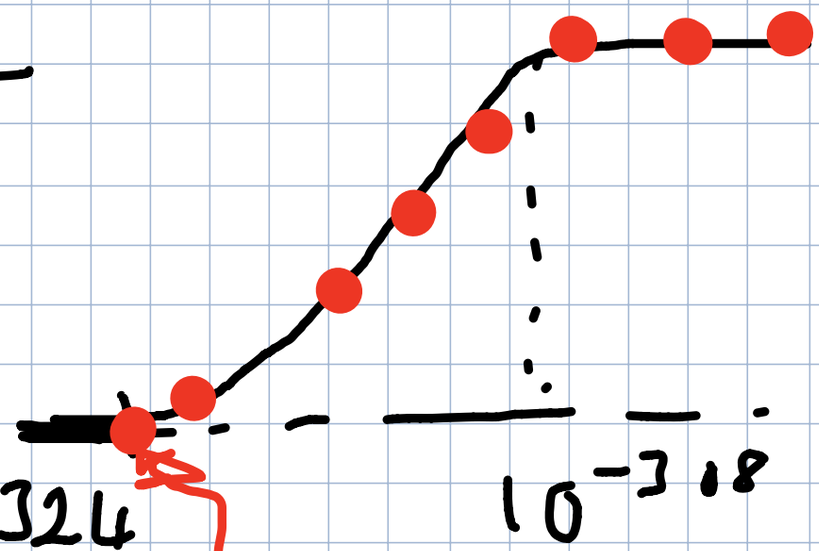


Table Convergence of Secant on $\tan^{-1} x$

k	x_1	x_2	E_{rel}	$f(x_1)$	$f(x_2)$
1	$1.100000e+001$	$-4.289777e+000$	$3.564236e+000$	$1.480136e+000$	$1.480136e+000$
2	$-4.289777e+000$	$2.980271e+000$	$2.439392e+000$	$-1.341774e+000$	$-1.341774e+000$
3	$2.980271e+000$	$-5.217652e-001$	$6.711901e+000$	$1.247061e+000$	$1.247061e+000$
4	$-5.217652e-001$	$4.528800e-001$	$2.152105e+000$	$-4.809078e-001$	$-4.809078e-001$
5	$4.528800e-001$	$-4.508330e-003$	$1.014540e+002$	$4.252463e-001$	$4.252463e-001$
6	$-4.508330e-003$	$2.898578e-004$	$1.655359e+001$	$-4.508300e-003$	$-4.508300e-003$
7	$2.898578e-004$	$-1.837521e-009$	$1.577450e+005$	$2.898578e-004$	$2.898578e-004$
8	$-1.837521e-009$	$5.146101e-017$	$3.570705e+007$	$-1.837521e-009$	$-1.837521e-009$
9	$5.146101e-017$	0	Inf	$5.146101e-017$	0