Accuracy and Stability of Numerical Algorithms

Nick Higham
School of Mathematics
The University of Manchester

higham@ma.man.ac.uk
http://www.ma.man.ac.uk/~higham
@nhigham

SIAM Chapter Day, Cardiff University
January 21, 2013

SIAM

- $\sim 13,500$ members: ~ 9000 USA, ~ 500 UK.
- 59% in maths depts
  16% in eng depts
  11% in CS depts
- 108 student chapters:

Oxford (2007)	#53
Heriot Watt & Edinburgh	#71
Manchester	#76
Strathclyde	#93
Warwick	#93
Reading	#100
Cardiff	#104
Bath	#105

Outline

1. Rounding
2. Precision
3. Accuracy
4. Higher Precision
5. Tiny Relative Errors

Floating Point Number System

Floating point number system $F \subset \mathbb{R}$:

$$y = \pm m \times \beta^{e-t}, \qquad 0 \le m \le \beta^t - 1.$$

- *Base* β,
- *precision* t,
- *exponent range* $e_{\min} \le e \le e_{\max}$.

Floating Point Number System

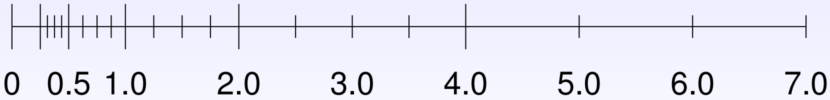
Floating point number system $F \subset \mathbb{R}$:

$$y = \pm m \times \beta^{e-t}, \qquad 0 \le m \le \beta^t - 1.$$

- *Base* β,
- *precision* t,
- *exponent range* $e_{\min} \le e \le e_{\max}$.

Floating point numbers are **not equally spaced**.

If $\beta = 2$, $t = 3$, $e_{\min} = -1$, and $e_{\max} = 3$:



Rounding

For $x \in \mathbb{R}$, $fl(x)$ is an element of F nearest to x, and the transformation $x \to fl(x)$ is called **rounding** (to nearest).

Theorem

If $x \in \mathbb{R}$ lies in the range of F then

$$fl(x) = x(1+\delta), \qquad |\delta| \le u := \frac{1}{2}\beta^{1-t}.$$

u is the **unit roundoff**, or machine precision.

Types of Rounding

- **Round to nearest**—with rule for breaking ties.

  8.12 $\rightarrow$ **8.1**, **8.17** $\rightarrow$ **8.2**.
  8.15 $\rightarrow$ **8.1** or **8.2**.

Types of Rounding

- **Round to nearest**—with rule for breaking ties.
  8.12 $\rightarrow$ **8.1**, **8.17** $\rightarrow$ **8.2**.
  8.15 $\rightarrow$ **8.1** or **8.2**.
- **Round up**: to $+\infty$.
  8.11 $\rightarrow$ **8.2**.

Types of Rounding

- **Round to nearest**—with rule for breaking ties.
  8.12 $\to$ **8.1**, **8.17** $\to$ **8.2**.
  8.15 $\to$ **8.1** or **8.2**.
- **Round up**: to $+\infty$.
  8.11 $\to$ **8.2**.
- **Round down**: to $-\infty$.
  8.68 $\to$ **8.6**.

Types of Rounding

- **Round to nearest**—with rule for breaking ties.

  8.12 $\rightarrow$ **8.1**, **8.17** $\rightarrow$ **8.2**.
  8.15 $\rightarrow$ **8.1** or **8.2**.
- **Round up**: to $+\infty$.

  8.11 $\rightarrow$ **8.2**.
- **Round down**: to $-\infty$.

  8.68 $\rightarrow$ **8.6**.

Thomas (9) homework

- *Round 17.37 to the nearest tenth*. "17.4"
- *Round 13.75 to the nearest tenth*. "13.8"

Justin Gatlin Record, 2006

BBC SPORT ATHLETICS

Low graphics | Help

Last Updated: Wednesday, 17 May 2006, 09:56 GMT 10:56 UK

E-mail this to a friend   Printable version

Gatlin denied outright 100m mark

Justin Gatlin has been denied the outright world 100m record after his time was suddenly altered almost a week after his blistering run in Qatar.

Officials have revealed the World and Olympic champion clocked 9.766 seconds, not 9.760 seconds as first thought.

Gatlin poses with what he thought was a new world record

"The American was given a time of 9.76sec at the Qatar Super grand prix but his official time was 9.766, which was rounded down instead of being rounded up to Powell's time of 9.77 set in Athens last year according to rules set out by track and field's governing body, the timekeeper Tissot admitted."

Bank of England: Inflation Rate, 2007

King faces a point of embarrassment

A TINY price movement equivalent to one hundred thousandth of one per cent on the inflation rate could make all the difference to the Bank of England this week.

With Government statisticians crunching cost of living numbers to six decimal places, the slightest rise could force Bank governor Mervyn King to explain what went wrong.

The Office for National Statistics publishes inflation numbers to one decimal place, meaning that a 3.049 rate would appear as three per cent, letting the MPC off the hook.

Vancouver Stock Exchange Index

- January 1982: Index established at 1000.
- November 1983: Index was 520.

But exchange seemed to be doing well.

Vancouver Stock Exchange Index

- January 1982: Index established at 1000.
- November 1983: Index was 520.

But exchange seemed to be doing well.

Explanation:

- Index **rounded down** to three digits at each recomputation.
- Errors always in same direction $\Rightarrow$ thousands of small errors add up to a large error.

Upon correct recalculation, the index doubled!

Virgin Media, 2007

Important information on your Virgin Media services

Dear Mr Higham

We're writing to tell you about changes to your phone charges that will be coming into effect from 1st May 2007.

The price of our monthly phone packages is coming down, so *Size: XL* (Talk Unlimited) will go from £14 a month to £9.95 and *Size: L* (Talk Evenings and Weekends) will go from £5.50 to £3.95. Your phone line will stay the same at £11 a month.

The way your call charges are calculated is also changing. Instead of charging to the nearest second, calls will be rounded up to the next minute. So, for example, a call that lasts 4 minutes 50 seconds will be rounded up to 5 minutes. If you have a phone package, any calls made outside your call plan will be rounded up to the next minute.

Virgin Media, 2007

Important information on your Virgin Media services

Dear Mr Higham

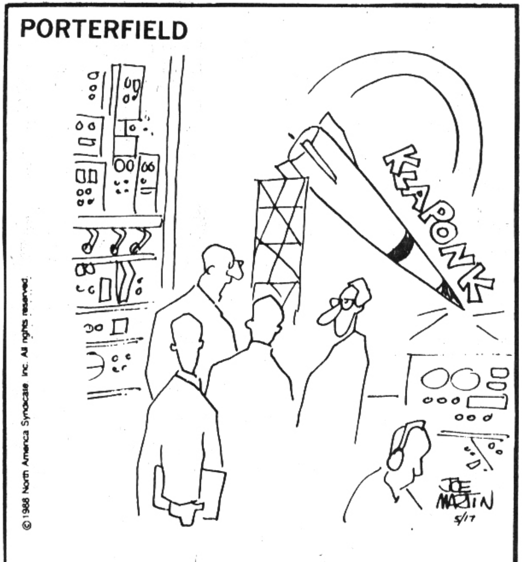
We're writing to tell you about changes to your phone charges that will be coming into effect from 1st May 2007.

The price of our monthly phone packages is coming down, so *Size: XL* (Talk Unlimited) will go from £14 a month to £9.95 and *Size: L* (Talk Evenings and Weekends) will go from £5.50 to £3.95. Your phone line will stay the same at £11 a month.

The way your call charges are calculated is also changing. Instead of charging to the nearest second, calls will be rounded up to the next minute. So, for example, a call that lasts 4 minutes 50 seconds will be rounded up to 5 minutes. If you have a phone package, any calls made outside your call plan will be rounded up to the next minute.

Remember

Rounding doesn't always mean round to nearest!



Next time, no rounding off!

Outline

Precision versus Accuracy

Unit roundoff $u = \frac{1}{2}\beta^{1-t}$.

$$\begin{aligned} fl(abc) &= ab(1+\delta_1)\cdot c(1+\delta_2) \qquad |\delta_i| \le u, \\ &= abc(1+\delta_1)(1+\delta_2) \\ &\approx abc(1+\delta_1+\delta_2). \end{aligned}$$

- Precision $= u$.
- Accuracy $\approx 2u$.

Precision versus Accuracy

Unit roundoff $u = \frac{1}{2}\beta^{1-t}$.

$$\begin{aligned} fl(abc) &= ab(1+\delta_1)\cdot c(1+\delta_2) \qquad |\delta_i| \le u, \\ &= abc(1+\delta_1)(1+\delta_2) \\ &\approx abc(1+\delta_1+\delta_2). \end{aligned}$$

- Precision $= u$.
- Accuracy $\approx 2u$.

Accuracy is not limited by precision


5
65
220

Walkers' Trouser Review

THE LOWDOWN

Fabric Nikwax Analogy Insulator (polyester microfibre outer, 100g polyester fill)
Sizes XS-XL (unisex)
Inside leg 79cm only
Waist integral belt, front flap with Velcro tabs
Pockets none

RATINGS

Comfort	80%
Fabric performance	100%
Versatility	50%
Quality/value	85%
Overall	**78.75%**

RGB to XYZ

From CIE Standard (1931):

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.49 & 0.31 & 0.20 \\ 0.17697 & 0.81240 & 0.01063 \\ 0 & 0.01 & 0.99 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}.$$

RGB to XYZ

From CIE Standard (1931):

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.49 & 0.31 & 0.20 \\ 0.17697 & 0.81240 & 0.01063 \\ 0 & 0.01 & 0.99 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}.$$

But in many books:

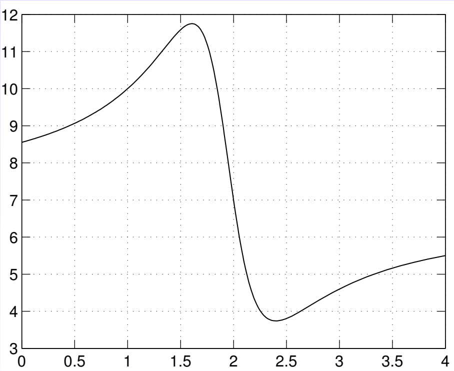
$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.49000 & 0.31000 & 0.20000 \\ 0.17697 & 0.81240 & 0.01063 \\ 0 & 0.01000 & 0.99000 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}.$$

Outline

1. Rounding
2. Precision
3. Accuracy
4. Higher Precision
5. Tiny Relative Errors

Rational Function

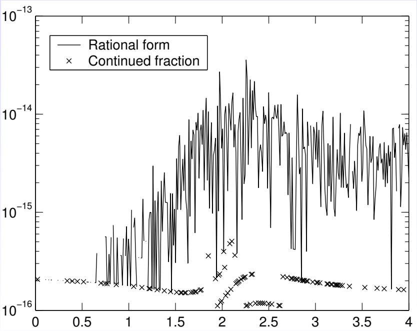
$$r(x) = \frac{(((7x - 101)x + 540)x - 1204)x + 958}{(((x - 14)x + 72)x - 151)x + 112}$$



Continued Fraction

$$r(x) = 7 - \cfrac{3}{x - 2 - \cfrac{1}{x - 7 + \cfrac{10}{x - 2 - \cfrac{2}{x - 3}}}}$$

Division by zero at $x = 1, 2, 3, 4$, but r evaluates correctly in IEEE arithmetic!



Cancellation Example

$$0 \le \frac{1-\cos x}{x^2} < 1/2, \qquad x \ne 0.$$

With $x = 1.2 \times 10^{-5}$, $\cos x$ rounded to 10 sig figs is

$$c = 0.9999\,9999\,99 \quad \Rightarrow \quad 1 - c = 0.0000\,0000\,01.$$

Then $(1-c)/x^2 = 10^{-10}/1.44 \times 10^{-10} = \boxed{0.6944\ldots!}$

To avoid cancellation, rewrite as

$$\frac{1}{2}\left(\frac{\sin(x/2)}{x/2}\right)^2.$$

Cancellation Example

$$0 \le \frac{1 - \cos x}{x^2} < 1/2, \qquad x \ne 0.$$

With $x = 1.2 \times 10^{-5}$, $\cos x$ rounded to 10 sig figs is

$$c = 0.9999\,9999\,99 \quad \Rightarrow \quad 1 - c = 0.0000\,0000\,01.$$

Then $(1 - c)/x^2 = 10^{-10}/1.44 \times 10^{-10} = \boxed{0.6944\ldots!}$

To avoid cancellation, rewrite as

$$\frac{1}{2}\left(\frac{\sin(x/2)}{x/2}\right)^2.$$

- The subtraction $1 - c$ is **exact**.

Cancellation

Theorem (Sterbenz)

Let x and y be floating point numbers with $y/2 \le x \le 2y$. Then $x - y$ is computed exactly (assuming $x - y$ does not underflow).

Cancellation

Theorem (Sterbenz)

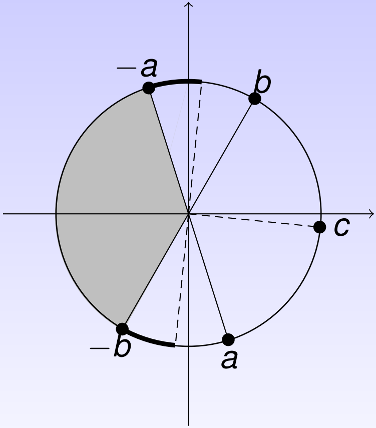
Let x and y be floating point numbers with $y/2 \le x \le 2y$. Then $x - y$ is computed exactly (assuming $x - y$ does not underflow).

Cancellation **brings earlier errors into prominence** but is not *always* a bad thing.

- Numbers being subtracted may be error free.
- Cancellation may be a symptom of intrinsic ill conditioning of problem.

Midpoint of Arc

Guo, H & Tisseur (2009):



- **Problem**: Find midpoint c of an arc (a, b).
- Obvious formula $c = (a + b)/|a + b|$ is unstable when $a \approx -b$.
- **Solution**: If $a = e^{i\theta_1}$, $b = e^{i\theta_2}$ then $c = e^{i(\theta_1+\theta_2)/2}$.

How to Compute $\log \lambda_2 - \log \lambda_1$

Define the **unwinding number**

$$\mathcal{U}(z) := \frac{z - \log e^z}{2\pi i} = \left\lceil \frac{\operatorname{Im} z - \pi}{2\pi} \right\rceil \in \mathbb{Z}.$$

Let $z = (\lambda_2 - \lambda_1)/(\lambda_2 + \lambda_1)$.

How to Compute $\log \lambda_2 - \log \lambda_1$

Define the **unwinding number**

$$\mathcal{U}(z) := \frac{z - \log e^z}{2\pi i} = \left\lceil \frac{\operatorname{Im} z - \pi}{2\pi} \right\rceil \in \mathbb{Z}.$$

Let $z = (\lambda_2 - \lambda_1)/(\lambda_2 + \lambda_1)$. Then

$$\begin{aligned}
\log \lambda_2 - \log \lambda_1 &= \log\left(\frac{\lambda_2}{\lambda_1}\right) + 2\pi i\,\mathcal{U}(\log \lambda_2 - \log \lambda_1) \\
&= \log\left(\frac{1+z}{1-z}\right) + 2\pi i\,\mathcal{U}(\log \lambda_2 - \log \lambda_1) \\
&= \operatorname{atanh}(z) + 2\pi i\,\mathcal{U}(\log \lambda_2 - \log \lambda_1).
\end{aligned}$$

H (2008): used in MATLAB `logm`.

Outline

IEEE Standard 754-2008 and 1985

Type	Size	Range	$u = 2^{-t}$
single	32 bits	$10^{\pm 38}$	$2^{-24} \approx 6.0 \times 10^{-8}$
double	64 bits	$10^{\pm 308}$	$2^{-53} \approx 1.1 \times 10^{-16}$
quadruple	128 bits	$10^{\pm 4932}$	$2^{-113} \approx 9.6 \times 10^{-35}$

- Arithmetic ops $(+, -, *, /, \sqrt{})$ performed *as if* first calculated to infinite precision, then rounded.
- Default: round to nearest, round to even in case of tie.

Need for Higher Precision

- Bailey, Simon, Barton & Fouts, **Floating Point Arithmetic in Future Supercomputers**, Internat. J. Supercomputer Appl. 3, 86–90, 1989.
- Bailey, Barrio & Borwein, **High-Precision Computation: Mathematical Physics and Dynamics**, Appl. Math. Comput. 218, 10106–10121, 2012.

- Long-time simulations.
- Large-scale simulations.
- Resolving small-scale phenomena.

Increasing the Precision

$y = e^{\pi\sqrt{163}}$ evaluated at t digit precision:

t	y
20	262537412640768744.00
25	262537412640768744.0000000
30	262537412640768743.999999999999

Is the last digit before the decimal point 4?

Increasing the Precision

$y = e^{\pi\sqrt{163}}$ evaluated at t digit precision:

t	y
20	262537412640768744.00
25	262537412640768744.0000000
30	262537412640768743.999999999999

Is the last digit before the decimal point 4?

t	y
35	262537412640768743.99999999999925007
40	262537412640768743.9999999999992500725972

So no, it's 3!

Zimbabwe resorts to the $100 trillion note

By Our Foreign Staff

ZIMBABWE'S central bank will introduce a 100 trillion Zimbabwean dollar banknote, worth about £22, on the black market, to try to ease desperate cash shortages, state-run media said yesterday. Prices are doubling every day and food and fuel are in short supply.

A cholera epidemic has killed more than 2,000 people and a deadlock between Mr Mugabe and the opposition has put hopes of ending the crisis on hold. Hyper-inflation has forced the central bank to continue to release new banknotes which quickly become almost worthless. There is an official exchange rate, but most Zimbabweans use the informal market for currency deals.

As well as the Z$100 trillion dollar note, the Reserve Bank of Zimbabwe plans to introduce Z$10 trillion, Z$20 trillion and Z$50 trillion notes, the *Herald* newspaper reported.

Zimbabweans often line up for hours outside banks to withdraw barely enough to buy a loaf of bread.

Old Mutual's new chief weighs rescue options

JUDGING by the empty state of his spacious South African office, it is quite clear that Julian Roberts has yet to settle into his role as the new chief executive of Old Mutual.

While his secretary bustles around, tidying away his few possessions – a 5p piece and a penny coin left lying on his desk – the four books on his vacant shelves stand out. The titles *Blown to Bits* and *On the Brink of Failure* could almost sum up the state of the blue-chip company Mr Roberts has just taken over. Old Mutual was the worst-performing European

PROFILE

Julian Roberts

Chief executive, Old Mutual

The economic turmoil revealed cracks in Old Mutual's model when it emerged that its $2.8bn (£1.9bn) variable annuity business in the US could not meet guarantees due to adverse movements in the Asian markets. It has been forced to inject

going to be immune. South Africa lags the rest of the world by six months to a year."

Political tensions are also playing on his mind. Old Mutual is listed not only in the UK and Johannesburg but also on the Zimbabwe Stock Exchange. Due to technical difficulties of transferring a figure with so many noughts on the end of it, Old Mutual struggled to pay shareholders an interim dividend of Z$453 trillion per share – which in November equated to just 2.45p.

"It is absolutely tragic. We have a significant business with a large

Going to Higher Precision

If we have quadruple or higher precision, what do we need to do to modify existing algorithms?

To what extent are existing algs precision-independent?

Matrix Functions

(Inverse) scaling and squaring-type algorithms for e^A, $\log(A)$, $\cos(A)$, A^t use Padé approximants.

- Padé degree chosen to achieve accuracy u.
- Padé coeffs and algorithm parameters need rederiving for a different u. Logic may change!
- MATLAB's **`expm`**, **`logm`** need changing for smaller u.

Methods based on best L_∞ approximations to e^A for Hermitian A also need higher order approximations deriving.

- Scalar elementary functions!

Outline

Tiny Relative Errors

Normwise relative errors

$$\frac{\|x-y\|_\infty}{\|x\|_\infty} = \frac{\max_i |x_i - y_i|}{\max_i |x_i|}$$

from a numerical experiment:

```
1.32e-22   3.39e-22   3.39e-21   8.67e-20
1.39e-18   4.36e-18   5.30e-18   5.83e-18
1.45e-17   3.76e-17   3.76e-17   4.27e-17
...
```

How can errors be $\ll u \approx 10^{-16}$?

Base $\beta = 2$, $u = 2^{-t}$. Dingle & H (2011):

Theorem

If $x \neq 0$ and y are distinct normalized flpt numbers then $|x - y|/|x| \geq u$ and this lower bound is attainable.

Base $\beta = 2$, $u = 2^{-t}$. Dingle & H (2011):

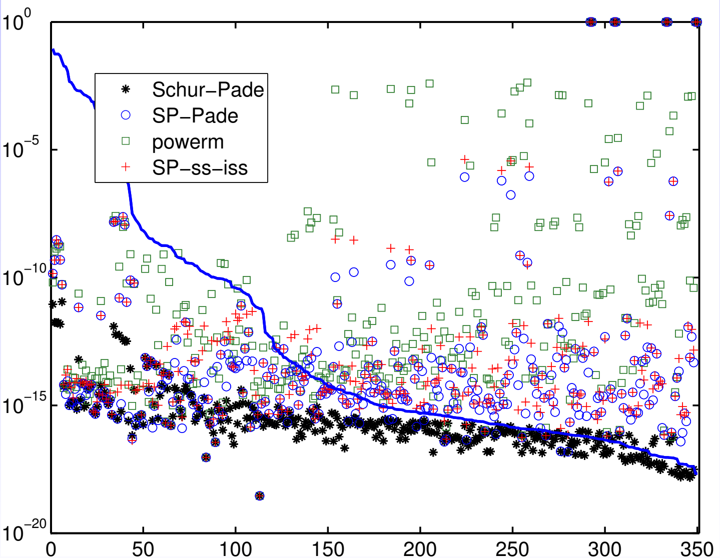
Theorem

If $x \neq 0$ and y are distinct normalized flpt numbers then $|x - y|/|x| \geq u$ and this lower bound is attainable.

But $\dfrac{\|x - y\|_\infty}{\|x\|_\infty} \ll u$ is possible.

$$x = \begin{bmatrix} 1 \\ 10^{-22} \end{bmatrix}, \quad y = \begin{bmatrix} 1 \\ 2 \times 10^{-22} \end{bmatrix}, \quad \frac{\|x - y\|_\infty}{\|x\|_\infty} = 10^{-22}.$$

Relative Errors



Performance Profiles

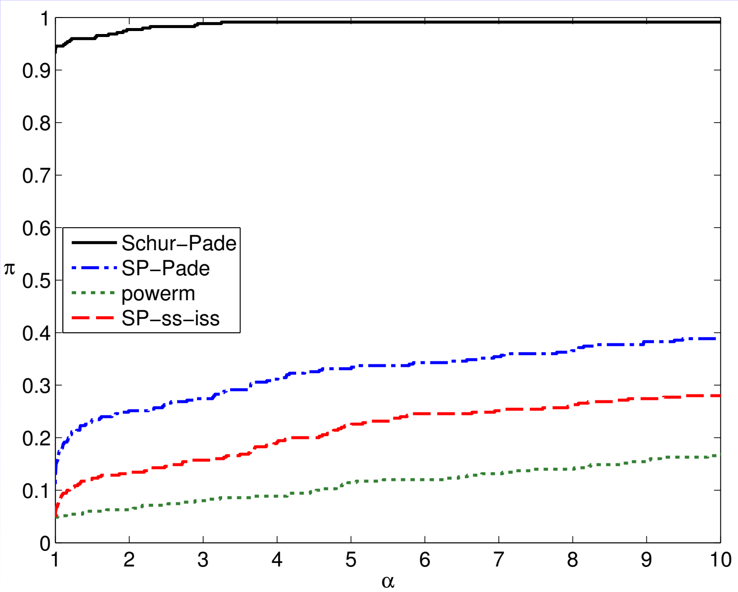
Dolan & Moré (2002).

For the given set of solvers and test problems, **plot**

x-axis: α

y-axis: probability that solver has error within factor α of smallest error over all solvers on the test set.

Performance Profile

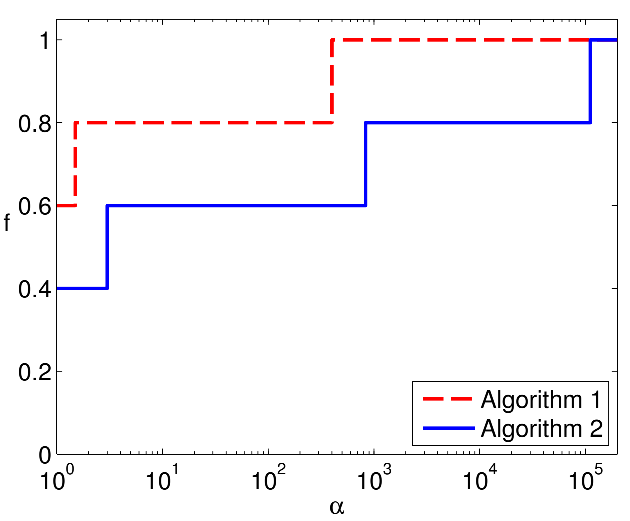


The Effect of Tiny Errors

Problem	Algorithm 1	Algorithm 2
1	4e-14	1e-16
2	6e-16	4e-16
3	1e-16	3e-16
4	9e-23	1e-17
5	6e-20	5e-17

Which algorithm is better?

Profile



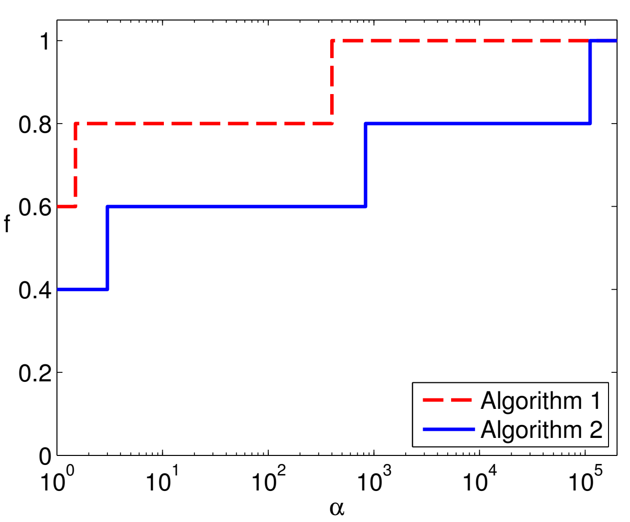
Alg 1	Alg 2
4e-14	1e-16
6e-16	4e-16
1e-16	3e-16
9e-23	1e-17
6e-20	5e-17

Transform the Data

- Map 0 to a (parameter). Typically, $a = u/20$.
- Map $[0, u]$ to $[a, u]$ linearly.
- Leave values $\geq u$ alone.

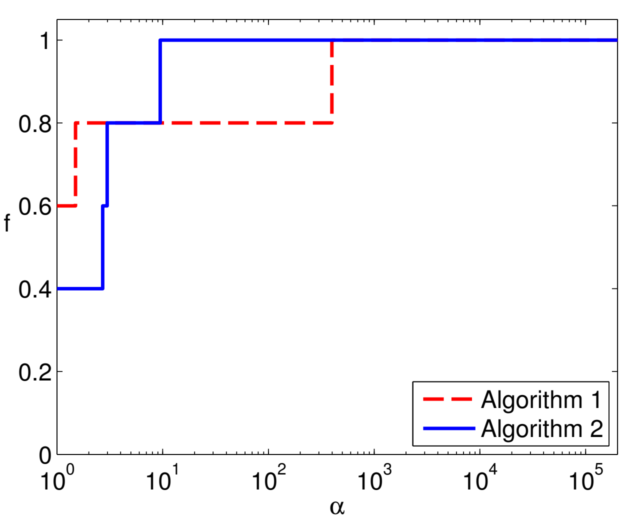
- Imposes positive minimum.
- Preserves ordering of errors.

Before



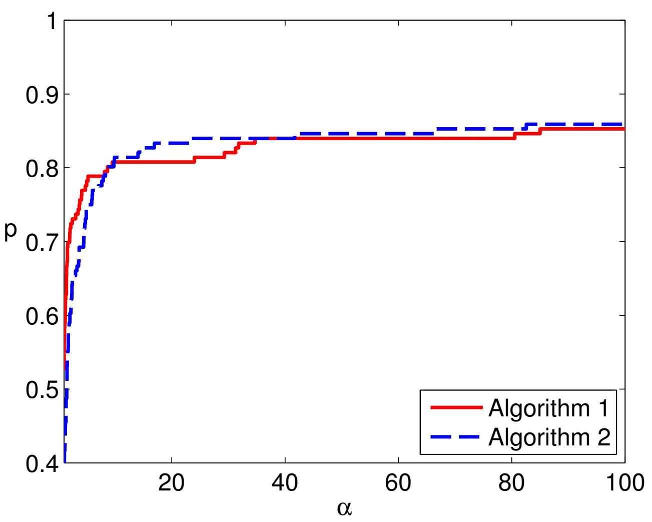
Alg 1	Alg 2
4e-14	1e-16
6e-16	4e-16
1e-16	3e-16
9e-23	1e-17
6e-20	5e-17

After

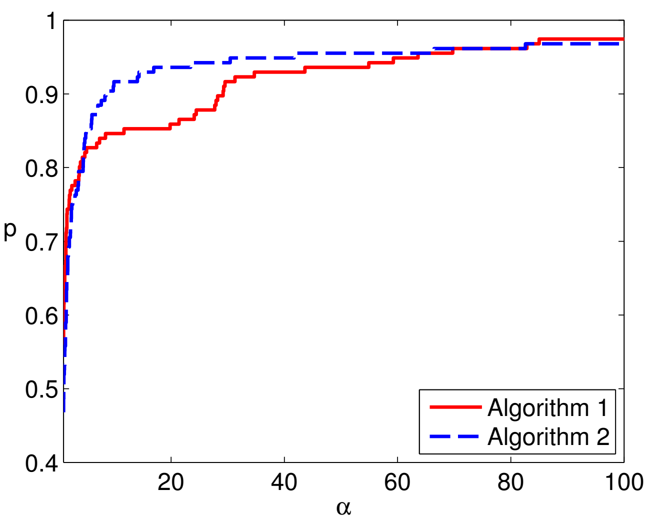


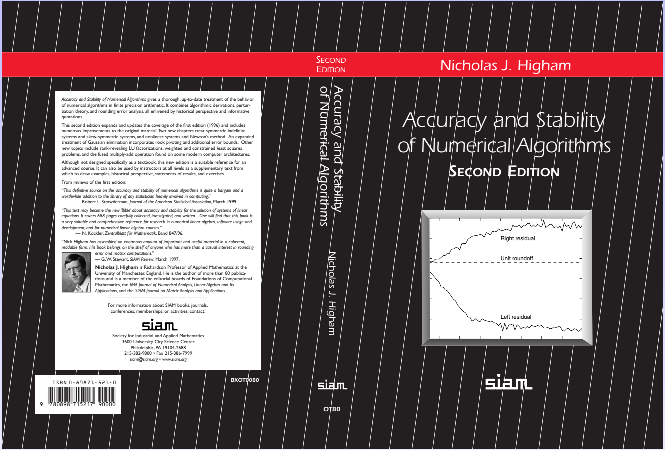
Alg 1	Alg 2
4.0e-14	**1.0e-16**
6.0e-16	4.0e-16
1.0e-16	3.0e-16
5.6e-18	**1.1e-17**
5.6e-18	**5.0e-17**

Matrix Exponential (Al-Mohy & H, 2011)



Matrix Exponential Transformed





Time to $\LaTeX$			
DX2-33	7.5 mins	Pentium 2.8Ghz	5 secs
Pentium 120Mhz	1.3 mins	Athlon X2 4400	4 secs
Pentium 500Mhz	20 secs	Core i7 @4.4Ghz	slower!
Pentium 1Ghz	10 secs		

References I

- A. H. Al-Mohy and N. J. Higham.
  Computing the action of the matrix exponential, with an application to exponential integrators.
  SIAM J. Sci. Comput., 33(2):488–511, 2011.

- N. J. Dingle and N. J. Higham.
  Reducing the influence of tiny normwise relative errors on performance profiles.
  MIMS EPrint 2011.90, Manchester Institute for Mathematical Sciences, The University of Manchester, UK, Nov. 2011.
  11 pp.

References II

- E. D. Dolan and J. J. Moré.
  Benchmarking optimization software with performance profiles.
  Math. Programming, 91:201–213, 2002.
- C.-H. Guo, N. J. Higham, and F. Tisseur.
  An improved arc algorithm for detecting definite Hermitian pairs.
  SIAM J. Matrix Anal. Appl., 31(3):1131–1151, 2009.

References III

N. J. Higham.
Accuracy and Stability of Numerical Algorithms.
Society for Industrial and Applied Mathematics,
Philadelphia, PA, USA, second edition, 2002.
ISBN 0-89871-521-0.
xxx+680 pp.