

Desperately Needed Remedies for the Undebuggability of Large Floating-Point Computations in Science and Engineering

W. Kahan, Prof. Emeritus
Math. Dept., and
Elect. Eng. & Computer Sci. Dept.
Univ. of Calif. @ Berkeley

Prepared for the

IFIP / SIAM / NIST Working Conference on
Uncertainty Quantification in Scientific Computing
3 Aug. 2011, Boulder CO.

Augmented for the

Annual Conference of the
Heilbronn Institute for Mathematical Research
8 Sept. 2011, University of Bristol, England.

And for

Computer Sci., Manchester University, 12 Sept. 2011
ICME, Stanford University, 13 Oct. 2011

This document is posted now at www.eecs.berkeley.edu/~wkahan/Boulder.pdf.

Contents

Pp. 3 - 4	Abstract and Cautionary Notice
5 - 7	Users need tools to help them investigate evidence of miscomputation
8 - 11	Why Computer Scientists haven't helped. Accuracy is Not Transitive
12 - 13	Summaries of the Stories So Far, and To Come
14	Kinds of Evidence of Miscomputation
15 - 19	EDSAC's arccos, Vancouver's Stock Index, Ranks Too Small, etc.
20 - 21	7090's Abrupt Stall; CRAYs' discordant Deflections of an Airframe
22	Clever and Knowledgeable, but Numerically Naive
23	How Suspicious Results may expose a hitherto Unsuspected Singularity
24 - 26	Pejorative Surfaces
27 - 30	The Program's Pejorative Surface may have an Extra Leaf
31 - 35	Default Quad evaluation, the humane but unlikely Prophylaxis
36	Two tools to localize roundoff-induced anomalies by rerunning the program
37 - 38	Rare counterexamples expose the tools' fallibility
39 - 42	Recomputation with three Redirected Roundings
43 - 45	Recomputation with Higher Precision
46 - 48	Floating-Point Exceptions Punished as Errors
49 - 52	Why default disruptions of control handle Floating-Point Exceptions badly
53 - 56	USS Yorktown, Ariane 5, Air France #447
57	Can losses of prestige, money and lives induce reconsideration of a policy?
58 - 60	Does this shoe leak at its toe?
61	To mitigate a Dangerous Dilemma
62 - 63	Proper Algebraic Completion for Default Presubstitution
64 - 65	Provision for individual non-default presubstitution
66 - 68	<i>flags</i> serve also as Pointers to ...
69 - 71	Retrospective Diagnostics
72 - 73	Retrospective Diagnostics' Annunciator and Interrogator
74	To be Collected: a Constellation of Competencies
76	Publications Cited
77 - 89	Responses to Questions and Comments

Desperately Needed Remedies for the Undebuggability of Large Floating-Point Computations in Science and Engineering

Abstract:

If suspicions about the accuracy of a computed result arise, how long does it take to either allay or justify them? Often diagnosis has taken longer than the computing platform's service life. Software tools to speed up diagnosis by at least an order of magnitude could be provided but almost no scientists and engineers know to ask for them, though almost all these tools have existed, albeit not all together in the same place at the same time. These tools would cope with vulnerabilities peculiar to Floating-Point, namely roundoff and arithmetic exceptions. But who would pay to develop the suite of these tools? Nobody, unless he suspects that the incidence of misleadingly anomalous Floating-Point results rather exceeds what is generally believed. Ample evidence supports that suspicion.

This document is posted now at www.eecs.berkeley.edu/~wkahan/Boulder.pdf.
More details have already been posted at .../NeeDebug.pdf and .../Mindless.pdf.

“This ... paper, by its very length, defends itself against the risk of being read.”
... attributed to Winston S. Churchill

To fit into its allotted time,
this paper’s oral presentation skips over most of the details;
but it is intended to induce you to look into those details.

“A fanatic is one who can’t change his mind and won’t change the subject.”
... Winston S. Churchill (1874 - 1965)

Am I a fanatic?

If so, you have been warned.

What is the incidence of Floating-Point computations wrong enough to mislead, but not so wrong as is obviously wrong?

Nobody knows. Nobody is keeping score.

Evidence exists implying an incidence rather greater than is generally believed.

Two Kinds of Evidence will be presented:

- **Persistence** in Software and in Programming Texts of numerically flawed formulas that have *withstood* rather than *passed* the *Test of Time*. For example, ...
Naive solutions of quadratic equations; ... of discretized differential equations
- Occasional **Revelations** of gross inaccuracies, in widely used and respected packages like MATLAB and LAPACK, caused by bugs lying hidden for years. *E.g.*, ...
Over 40 years of occasional *underestimates*, some severe, of matrices' ranks.

Evidently, providers of numerical software need help to debug it; they need abundant assistance from users.

How much debugging of numerical software is included in a chemist's job-description?

Distinctions between users and providers of numerical software are blurred by developers who incorporate, into their own software, modules developed by others. *e.g.*, LAPACK

If providers expect users to help debug numerical software,
they (and we) must find ways to reduce the costs
in time and expertise
of investigating numerical results that arouse suspicions.

Later we shall see why the earliest symptoms of hitherto unsuspected gross inaccuracies
that will befall our software at some unknown innocuous data
are highly likely to be inaccuracies, at other data, barely bad enough to arouse suspicions.

How much can investigation of a suspect Floating-Point computation's accuracy cost?
Often more than the computed result is worth.

Computers are now so cheap, most perform computations of which no one is worth very much:
Entertainment, Communications, Companionship, Embedded Controllers
are computers' most prevalent and most remunerative uses;
not our scientific and engineering computations.

A Problem of Misperception in the Marketplace:

The software tools needed to reduce by orders of magnitude the costs of debugging anomalous Floating-Point computations have almost all existed, but not all in the same package, and not in current software development systems.

Why not?

- The producers of software development systems are unaware that such tools could be produced, much less that there is a demand for them.
- The scientists and engineers who would benefit from such tools are hardly aware of them, much less that they should be requested.

Those tools will be surveyed in what follows. For more details about them see www.eecs.berkeley.edu/~wkahan/NeedDebug.pdf and .../Mindless.pdf.

Computer scientists worldwide are working hard on schemes to debug and verify software, especially in the context of parallel computation, but not Floating-Point software. Why not?

Computer Science has changed over my lifetime.

Numerical Analysis seems to have turned into a sliver under the fingernails of computer scientists.

Symptoms of Change:

- In **1983** a C.S. encyclopedia ed. by Ralston & Reilly included long articles ...
 ... on Floating-Point error-analysis (by J.H. Wilkinson) and roundoff (by Ralston)
 ... on control structures for all kinds of exception-handling (by J.L. Wagener)
 14 Years later
- In **1997** a longer C.S. encyclopedia ed. by Tucker explains a few numerical methods but mentions neither roundoff nor Floating-Point exceptions.
- In **1997** an issue of *Communications of the ACM* **40** #4 devoted pp. 26 - 74 to
 “The Debugging Scandal and What To Do About It”
 with no mention of Floating-Point arithmetic.

Cover Feature: August 2011 Issue of IEEE computer society's *Computer*, 44 #8

THE IBM PC: 30-YEAR RETROSPECTIVE

pp. 19 - 45

Four reminiscences of vignettes, design, construction, ICs, and marketing of the PC.

No mention of ...

- Embarrassingly anomalous Floating-Point arithmetic of the 1981 PC's ROM-BASIC .
An early version of my *PARANOIA* program printed out several pages of inexplicable evaluations of arithmetic expressions. Almost all these anomalies were repaired by late 1982 in the IBM PC-XT's ROM-BASIC. I wasn't told whether my printout instigated these repairs.
- Why Microsoft's software crippled the 80x87's Floating-Point in PC's, -XT and -AT.
Bill Gates predicted utterly wrongly that the PCs' sockets for the 8087 coprocessor would almost all stay empty, so he allocated at most minimal resources for its support. And today Microsoft still begrudges support for IEEE 754's arithmetic capabilities. Borland's excellent *QUATTRO* spreadsheet, programmed by Roger Schlafli, was the first to benefit from 80x87s' arithmetic, avoiding most anomalies in *VISICALC*, Lotus/IBM *123* and now Microsoft *EXCEL*. For instance see pp. 3 - 5 of www.eecs.berkeley.edu/~wkahan/Mindless.pdf.

Would *Computer*'s readers find these stories less interesting than the ones printed?

What characteristics of Floating-Point computation offend Computer Scientists?

- What you see is not exactly what you get.
 What you get is not exactly what your program asked for.
 Consequently what you get can be *Utterly Wrong* without any of the usual suspects:
i.e. no subtractive cancellation, no division, no vast number of rounded operations.

For a simple didactic example see www.eecs.berkeley.edu/~wkahan/WrongR.pdf

- Worse, unlike *Correctness* of non-numerical computer programs, *Accuracy* of Floating-Point programs is *Not Transitive*:

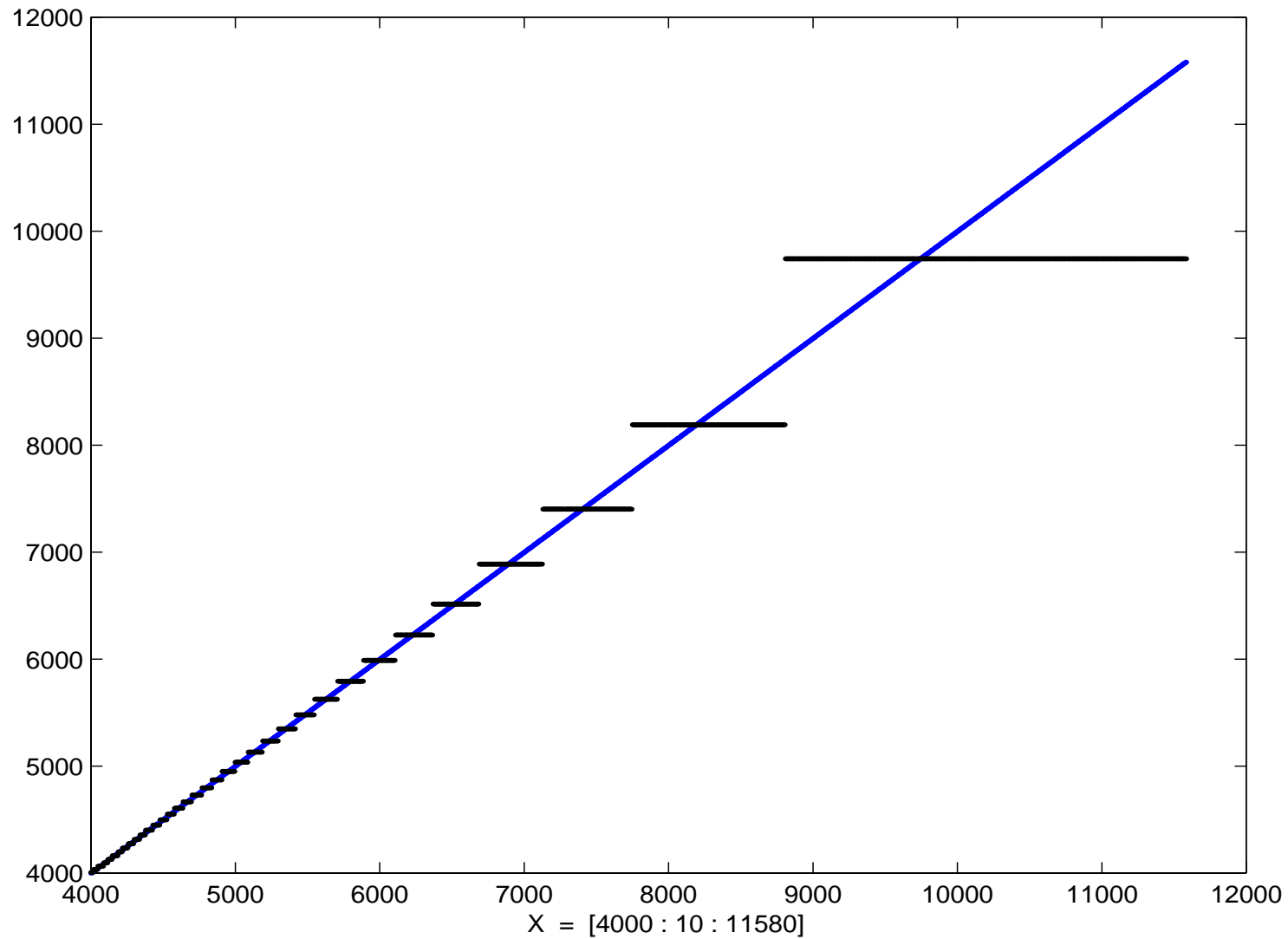
This means that ...

If program $H(X)$ approximates function $h(x)$ in all digits but its last, and if program $G(Y)$ approximates function $g(y)$ in all digits but its last, yet program $F(X) := G(H(X))$ may approximate function $f(x) := g(h(x))$ *Utterly Wrongly* over a large part of its domain.

Here is a simple didactic example, albeit contrived:

$$h(x) := \exp(-1/x^4) \quad @ \quad x > 1; \quad g(y) := 1/4\sqrt{-\log(y)} \quad @ \quad 0 < y < 1; \quad f(x) = x \quad @ \quad x > 1.$$

$$f(x) = x \quad \text{vs.} \quad F(x) = (-\log(\exp(-x^{-4})))^{-1/4}$$



This is explained in pp. 24 - 25 of my posting www.cs.berkeley.edu/~wkahan/MxMulEps.pdf.

Summary of the Story So Far:

I claim that scientists and engineers are almost all unaware ...

- ... of how high is the incidence of misleadingly inaccurate computed results.
- ... of how necessary is the investigation of every suspicious computed result as a potential harbinger of substantially worse to come.
- ... of the potential availability of software tools that would reduce those investigations' costs in expertise and time by orders of magnitude.
- ... that these tools will remain unavailable unless producers of software development systems (languages, compilers, debuggers) know these tools are in demand.

If almost nobody (but me) asks for such tools,
the demand for them will be presumed inadequate to pay for their development.

Computer scientists and programmers already have lots of other fish to fry.

Summary of the Story to Come:

- How high is the incidence of misleadingly inaccurate computed results?
What evidence suggests that it's higher than generally believed?
- How necessary is the investigation of every suspicious computed result as a possible harbinger of substantially worse to come?
What can turn almost infinitesimal rounding errors into grossly wrong results?
- Why can't arithmetic exceptions, like Over/Underflow, Division-by-Zero, etc., that may invalidate the computation simply stop it? Isn't continuation dangerous?
- What software tools would reduce those investigations' costs, in expertise and time, by *Orders of Magnitude*? How do you know?
[On a few ancient computers I implemented and enjoy some of the tools I describe.](#)

- How high is the incidence of misleadingly inaccurate computed results?

We cannot know. Nobody is keeping score.

- What evidence suggests that it's higher than generally believed?

Two kinds of evidence, **Revelation** and **Persistence** :

- **Revelation**, after long use, that a widely trusted program produces, for otherwise innocuous input data, results significantly more inaccurate than previously believed.
- **Persistence** of numerically naive and thus vulnerable formulas in the source-code of some programs, and in some published papers and textbooks.

Here is an example of naiveté too common in programming textbooks:

The zeros z of a real quadratic $\alpha \cdot z^2 - 2\beta \cdot z + \gamma$, assuming $\alpha \neq 0$ & $\gamma \neq 0$, are
 $z_1 := (\beta + \sqrt{(\beta^2 - \alpha \cdot \gamma)})/\alpha$ and $z_2 := (\beta - \sqrt{(\beta^2 - \alpha \cdot \gamma)})/\alpha$ naively.

Numerically more reliable (absent over/underflow) formulas for the zeros are

$$\delta := \beta^2 - \alpha \cdot \gamma; \text{ if } \delta < 0 \text{ then } \{ z_1 := \beta/\alpha + \mathbf{i}\sqrt{-\delta}/\alpha; z_2 := \beta/\alpha - \mathbf{i}\sqrt{-\delta}/\alpha \}$$

$$\text{else } \{ \zeta := \beta + \text{copysign}(\beta, \sqrt{\delta}); z_1 := \zeta/\alpha; z_2 := \gamma/\zeta \}.$$

Do you see why? Where are the formulas' singularities? What happens near them?

- After long use, a widely trusted program is discovered to have produced, for otherwise innocuous input data, results significantly more inaccurate than previously believed.

The earliest such instance I know befell one of the earliest electronic computers, EDSAC at Cambridge University.

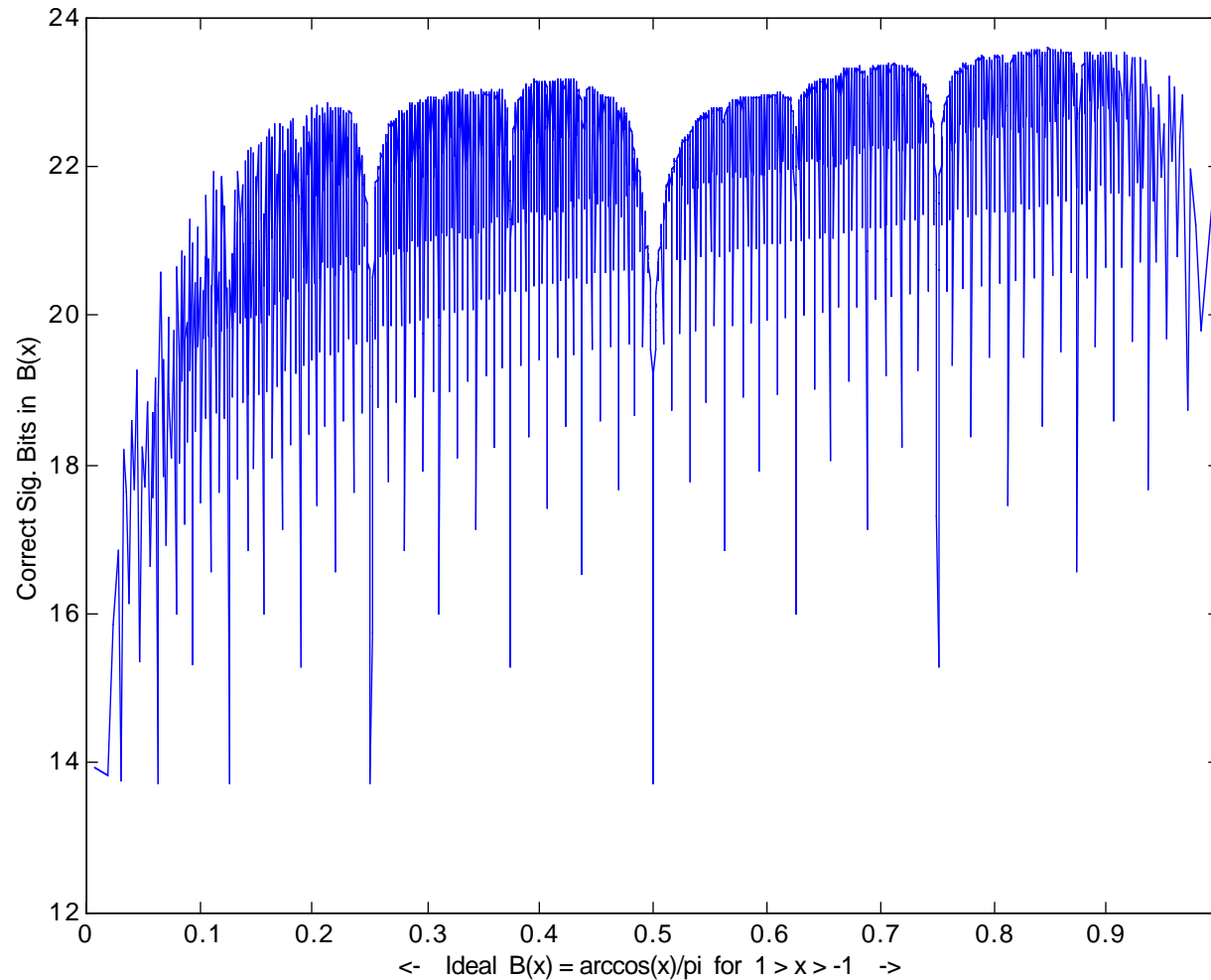
The program computed $B(x) := \arccos(x)/\pi$ from a neat algorithm (annotated here):

Set	$x_1 := x = \cos(B \cdot \pi)$;	$\beta_0 := 0$;	$B_0 := 0$;	$t_0 := 1$;	... Note $-1 \leq x \leq 1$.
While	$(B_{j-1} + t_{j-1} > B_{j-1})$	do	...	for $j := 1, 2, 3, \dots$	in turn
{	$t_j := t_{j-1}/2$;			$\dots = 1/2^j$	until it becomes negligible or zero.
	$\mu_j := \text{SignBit}(x_j)$;			$\dots = 0$ or 1	according as $x_j \geq 0$ or not.
	$\beta_j := \mu_j - \beta_{j-1} $;			$\dots = 0$ or 1	according as $\mu_j = \beta_{j-1}$ or not.
	$B_j := B_{j-1} + \beta_j \cdot t_j$;			$\dots = \sum_{1 \leq k \leq j} \beta_k / 2^k < 1$,	a binary expansion.
	$x_{j+1} := 2 \cdot x_j^2 - 1$ }			$\dots \approx \cos(2^j \cdot \arccos(x)) = \cos(2^{j+1} \cdot B \cdot \pi / 2)$.	

No subscript j appears in the actual program. With each pass around the While-loop, the program commits at most one rounding error in the last statement “ $x := 2 \cdot x^2 - 1$ ”. EDSAC ran the loop in fixed-point until $t = 0$ to get as many bits of B as the wordsize.

To get the next graph the program was run in floating-point to simulate what EDSAC would have gotten had its wordsize been 24 bits.

Of 24 Sig. Bits Carried, How Many are Correct in EDSAC's $B(x)$?



Accuracy spikes down wherever $B(x)$ comes near (but not exactly) a small odd integer multiple of a power of $1/2$. The smaller that integer, the wider and deeper the spike, down to almost half the sig. bits carried. Such arguments x are common in practice but were missed in EDSAC's tests.

Losing almost half the bits carried went unnoticed during conscientious (for that era) tests and for two years (1949 - 1951) afterwards. The testers were slightly unlucky; their probability of finding no bad errors during random testing exceeded $1/3$. For details and citations see pp. 37 - 42 of www.eecs.berkeley.edu/~wkahan/MktgMath.pdf .

- After long use, a widely trusted program is discovered to have produced, for otherwise innocuous input data, results significantly more inaccurate than previously believed.

The Vancouver Stock Exchange maintained an index of (mainly mining) stock prices.

On Fri. evening 25 Nov. 1983 the index ended at 524.811 .

On Mon. morning 28 Nov. 1983 the index began at 1098.892 .

But stock prices had not increased that much over the weekend. What had happened?

Rounding errors. The stock index was altered with each of about 3000 trades per day. The updated index was calculated to four dec. and then *chopped* (not rounded) to three. On average this lost over 20 index points/month for 22 months until *three weeks' work* by consultants from Toronto and California diagnosed and fixed the error that weekend.

Toronto *Star* 29 Nov. 1983

- After long use, a widely trusted program is discovered to have produced, for otherwise innocuous input data, results significantly more inaccurate than previously believed.

The longest running instance I know about was exposed by Zlatko Drmač & Zvonimir Bujanović [2008, 2010] in a program used heavily by LINPACK, LAPACK, MATLAB and numerous others since 1965 to estimate ranks of matrices. Given m -by- n matrix B and a small tolerance τ , we seek the least “rank” r for which

$$\begin{array}{c} n \\ \boxed{B} \\ m \end{array} \approx \begin{array}{c} r \\ \boxed{Q} \\ m \end{array} \cdot \begin{array}{c} n \\ \boxed{R} \\ r \end{array} \quad \text{within tolerance } \pm\tau .$$

Especially when $r < \min\{m, n\}/2$, this factorization reveals an important structure. The most reliable way to compute r is a *Singular Value Decomposition*, but a roughly three times faster “Pivoting QR” factorization had been preferred for over forty years despite that it could sometimes over-estimate r . Moderate over-estimates cause little damage.

Drmač & Bujanović discovered otherwise innocuous matrices B for which roundoff overlooked in the Pivoting QR program caused r to be under-estimated so severely as to violate tolerance τ when it was small enough, but not unreasonably small. This over-simplified and broke the sought structure badly. They have repaired the program’s defect.

Roundoff-Induced Anomalies Evade Expert Searches for Too Long:

- PATRIOT Anti-Missile Missiles missed a SCUD that hit a barracks in the Gulf War.
- From 1988 to 1998, MATLAB's built-in function `round(x)` that rounds x to a nearest integer-valued floating-point number malfunctioned in 386-MATLAB 3.5 and PC-MATLAB 4.2 by rounding all sufficiently big odd integers to the next bigger even integer. (Mac. MATLAB was O.K. thanks to Apple's S.A.N.E.)
- For more than a decade, MATLAB has been miscomputing $\gcd(3, 2^{80}) = 3$, $\gcd(28059810762433, 2^{15}) = 28059810762433$, $\text{lcm}(3, 2^{80}) = 2^{80}$, $\text{lcm}(28059810762433, 2^{15}) = 2^{15}$, and many others with no warning. See www.cs.berkeley.edu/~wkahan/MathH110/GCD5.pdf for corrected programs and .../HilbMats.pdf for their application to the exact construction of Hilbert matrices and their inverses to be used to test numerical linear algebra software.

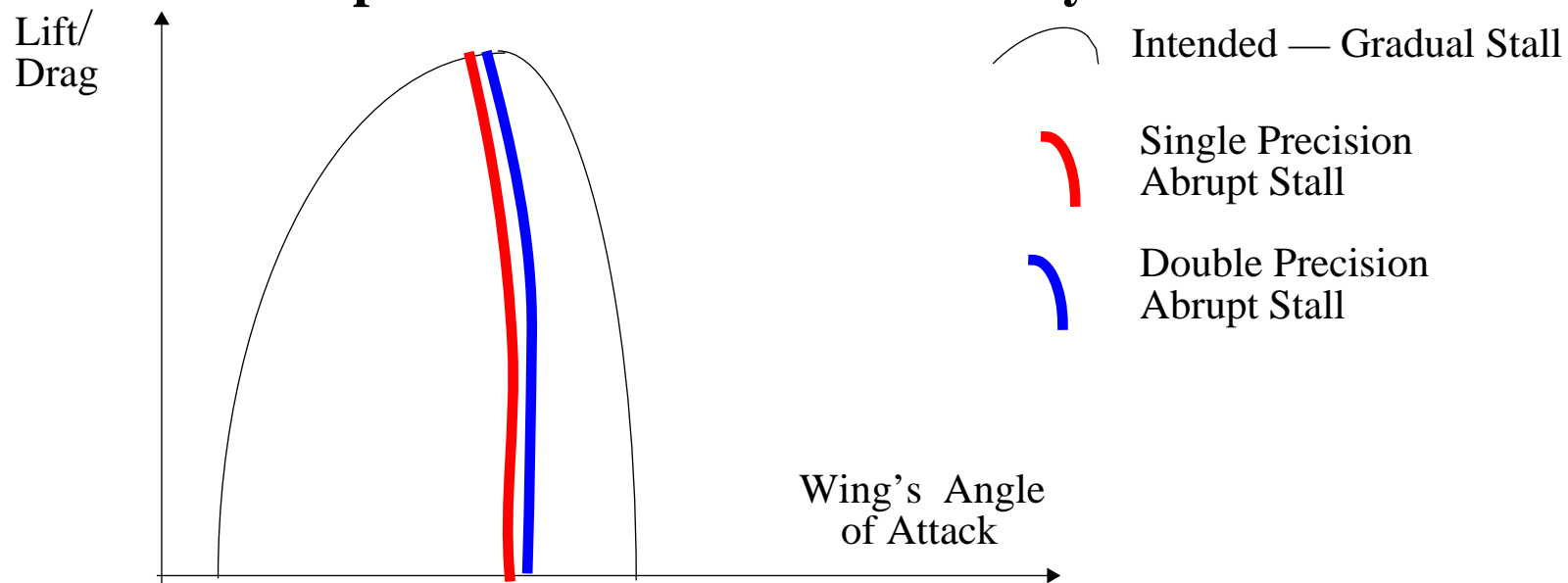
Anomalies due to Over/Underflow can evade expert searches for too long too.

In 2010, excessive inaccuracies were discovered in LAPACK's programs `_LARFP` and traced to underflows caused by the steps taken to avoid overflows. Whether the revisions to those programs promulgated subsequently are fully satisfactory remains to be seen.

- What if the user of a widely trusted program doesn't know that its results, for some otherwise innocuous input data, are significantly more inaccurate than the user believes?

This almost happened to a graduate student of aeronautical engineering in the early 1960s when his scheme to enhance lift for wings of Short-Takeoff-and-Landing aircraft seemed to suffer from abrupt onset of stall, according to his computations on an IBM 7090.

Abrupt Stall of Lift Enhanced by Blown Slots ?



Abrupt stall “caused” by inaccurate LOG in **Single**, by lack of guard digit in **Double** precision. Only after his was one of several programs chosen to test a new LOG's accuracy did he learn that the abrupt stall was entirely an artifact of roundoff. He resuscitated his research. For details see pp. 23 - 26 of www.eecs.berkeley.edu/~wkahan/NeedDebug.pdf .

I took years after the abrupt stall episode to appreciate its relevance to a question:

What exposes a misjudgment due to rounding errors ?

- A calamity severe enough to bring about an investigation, and investigators thorough and skilled enough to diagnose correctly that roundoff was the cause (if it was).
This *combination* appears to have occurred extremely rarely, if at all.
- Suspicions aroused by computed results different enough from one's expectations.
Someone would have to be exceptionally observant, experienced and diligent.
- Discordant results of recomputations using different arithmetics or different methods.
What would induce someone to go to the expense of such a recomputation?

In the mid 1990s a program written at NASA Ames predicted deflections under load of an airframe for a supersonic transport that turned out destined never to be built. Though intended for CRAY-I and CRAY-2 supercomputers, the program was developed on SGI Workstations serving as terminals. When a problem with a mesh coarse enough to fit in the workstation was run on all three machines, three results emerged disagreeing in their third sig. dec. This had ominous implications for the CRAYs' results from realistic problems with much finer meshes.

I traced the divergence to the CRAYs' idiosyncratic biased roundings. Adding iterative refinement to the program, a minor change, rendered the divergence tolerable. To rid the program of its worst errors would have required a major change; see my web page's .../Math128/FloTriK.pdf .

What exposes a misjudgment due to rounding errors ?

It's unlikely to be exposed.

Why must such misjudgments be happening?

Programs that depend upon some Floating-Point computation are being written by far more people than take a course in Numerical Analysis with enough Error-Analysis to sensitize them to the risks inherent in roundoff.

“Acquiescing to rounded arithmetic places you in a state of sin.” — D.H. Lehmer

People clever and knowledgeable in their own domains of science, engineering, statistics, finance, medicine, *etc.*, are naively using in their programs formulas mathematically correct but numerically vulnerable, instead of numerically robust but unobvious formulas.

Many such formulas are posted on my web pages; for a lengthy list see p. 22 of www.eecs.berkeley.edu/~wkahan/NeedDebug.pdf .

We may depend unwittingly upon some of these clever people's programs via the world-wide-web, the cloud, medical equipment, navigational apparatus, *etc.* How can we defend ourselves against numerical naiveté, or at least enhance the likelihood that their programs' numerical vulnerabilities will be exposed, preferably before too late?

How necessary is the investigation of every suspicious computed result as possibly a harbinger of substantially worse to come?

... if not symptomatic of a failure of some physical theory — a potential *Nobel Prize* !

“Les doutes sont fâcheux plus que toute autre chose.”

(Doubts cause more trouble than the worst truths.)

Le Misanthrope III.v (1666) by Molière (1622 - 1673)

After we have seen the most likely cause of a catastrophic numerical inaccuracy, we shall see why its possibility is most likely to be exposed by incidents that raise suspicions about computed results.

This is why suspicious computed results must be investigated.

To justify this necessity, we must understand what can turn almost infinitesimal rounding errors into grossly wrong results:

Perturbations get Amplified by Singularities Near the Data.

How Perturbations get Amplified by Singularities Near the Data.

Perturbed data $\mathbf{x} \rightarrow \mathbf{x} \pm \Delta\mathbf{x}$
 perturbs $f(\mathbf{x}) \rightarrow f(\mathbf{x} \pm \Delta\mathbf{x}) = f(\mathbf{x}) \pm \Delta f(\mathbf{x}) \approx f(\mathbf{x}) \pm f'(\mathbf{x}) \cdot \Delta\mathbf{x}.$

$\Delta f(\mathbf{x}) \approx f'(\mathbf{x}) \cdot \Delta\mathbf{x}$ can be huge when $\Delta\mathbf{x}$ is tiny only if derivative $f'(\mathbf{x})$ is gargantuan.

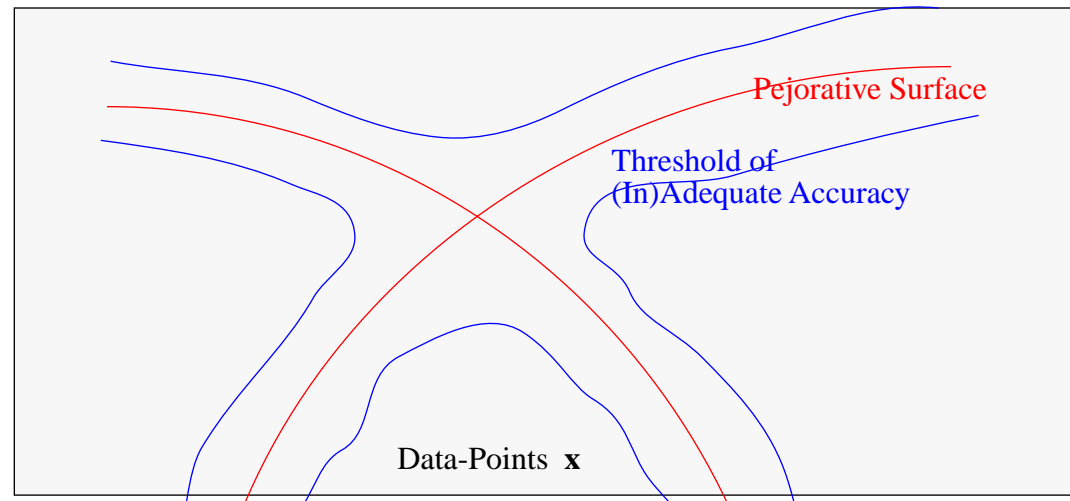
This can happen only if \mathbf{x} is near enough to a *Singularity* of f where its derivative $f' = \infty$.

Let's call the locus (point, curve, surface, hypersurface, ...) of data \mathbf{x} whereon $f'(\mathbf{x}) = \infty$ the “**Pejorative Surface**” of function f in its domain-space of data.

For example ...

Data Points	Computed Result	Data on a Pejorative Surface	Threshold Data
Matrices	Inverse	Cone of Singular Matrices	Not too “Ill-Conditioned”
Matrices	Eigensystem	... with Degenerate Eigensystems	Not too near Degenerate
Polynomials	Zeros	... with Repeated Zeros	Not too near repeated
4 Vertices	Tetrahedron's Volume	Collapsed Tetrahedra	Not too near collapse
Diff'l Equ'n	Trajectory	... with boundary-layer singularity	Not too “Stiff”

All or Most Accuracy can be Lost if Data lie on a “Pejorative” Surface



Accuracy of $f(\mathbf{x})$ is Adequate at Data \mathbf{x} Far Enough from Pejorative Surfaces.

Suppose the data’s “Precision” bounds its tiny uncertainty $\Delta\mathbf{x}$ thus: $\xi \geq \|\Delta\mathbf{x}\|$.

Then $f(\mathbf{x} \pm \Delta\mathbf{x})$ inherits uncertainty $\xi \cdot \|f'(\mathbf{x})\| \geq \|\Delta f\|$ roughly.

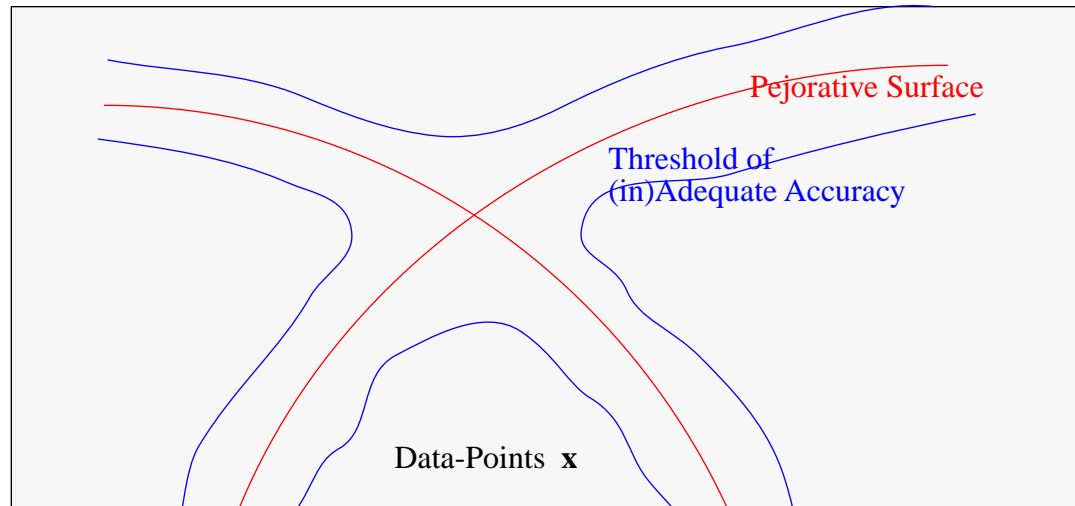
How fast does $\|f'(\mathbf{x})\| \rightarrow \infty$ as $\mathbf{x} \rightarrow$ (a Pejorative Surface) ?

Let $\pi(\mathbf{x}) :=$ (distance from \mathbf{x} to a nearest Pejorative Surface). *Typically* (not always !)

$\|f'(\mathbf{x})\|$ is roughly proportional to $1/\pi(\mathbf{x})$ while $\pi(\mathbf{x})$ is small enough;

then uncertainty $\xi \geq \|\Delta\mathbf{x}\|$ causes $f(\mathbf{x} \pm \Delta\mathbf{x})$ to “Lose” to the data’s uncertainty roughly

Const. $- \log(\pi(\mathbf{x})) + \log(\xi)$ dec. digits.



$\pi(\mathbf{x}) :=$ Distance from data \mathbf{x} to a nearest Pejorative Surface where derivative $f' = \infty$.
 $\xi \geq \|\Delta\mathbf{x}\|$ is a near-infinitesimal bound upon the uncertainty $\Delta\mathbf{x}$ in data \mathbf{x} . Typically,
 $f(\mathbf{x} \pm \Delta\mathbf{x})$ “Loses” roughly $\text{Const.} - \log(\pi(\mathbf{x})) + \log(\xi)$ dec. digits to \mathbf{x} ’s uncertainty.

How many lost digits are tolerable?

Two choices come to mind to keep the loss below a given bound Λ dec. digits:

- If data \mathbf{x} comes as close to a Pejorative Surface as $\pi(\mathbf{x}) = \Xi$ but no closer, keep the data’s “Precision” high enough that $\log(\xi) < \log(\Xi) - \text{Const} + \Lambda$.
- If given the data’s uncertainty ξ , let $\log(\Xi) > \log(\xi) - \Lambda + \text{Const.}$ constrain a Threshold Ξ , and eschew data \mathbf{x} whose $\pi(\mathbf{x}) < \Xi$, deeming such data “*Too Ill-Conditioned*” to determine f accurately enough. **Not roundoff!**

Rounding Errors are like Uncertain Data

Suppose program $F(\mathbf{X})$ is intended to compute $f(\mathbf{x})$ but actually $F(\mathbf{X}) = f(\mathbf{X}, \mathbf{r})$ in which column \mathbf{r} represents the rounding errors in F and $f(\mathbf{x}, \mathbf{0}) = f(\mathbf{x})$. The precision of the arithmetic imposes a bound like $\rho > \|\mathbf{r}\|$ analogous to the uncertainty ξ used above. To simplify exposition, assume the data \mathbf{X} we have equals the data \mathbf{x} we wish we had.

Let $f_r(\mathbf{x}) := \partial f(\mathbf{x}, \mathbf{r}) / \partial \mathbf{r} |_{\mathbf{r}=\mathbf{0}}$. Because ρ is so tiny, program $F(\mathbf{x})$ actually computes $f(\mathbf{x}, \mathbf{r}) \approx f(\mathbf{x}, \mathbf{0}) + f_r(\mathbf{x}) \cdot \mathbf{r} = f(\mathbf{x}) + f_r(\mathbf{x}) \cdot \mathbf{r}$, so $\|F(\mathbf{x}) - f(\mathbf{x})\| \approx \|f_r(\mathbf{x}) \cdot \mathbf{r}\| < \|f_r(\mathbf{x})\| \cdot \rho$.

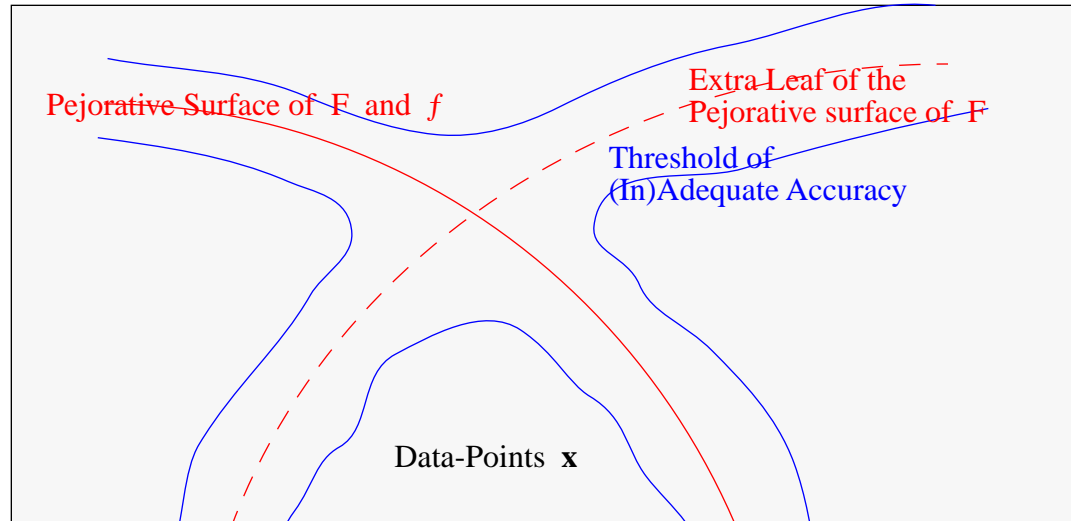
Error $F(\mathbf{x}) - f(\mathbf{x})$ can be huge when \mathbf{r} is tiny only if derivative f_r is gargantuan, which can happen only if \mathbf{x} is near enough to a *Singularity* of f where its derivative $f_r = \infty$.

Let's call the locus (point, curve, surface, hypersurface, ...) of data \mathbf{x} whereon $f_r(\mathbf{x}) = \infty$ the "*Pejorative Surface*" of program F in its domain-space of data.

Function f 's Pejorative Surface is usually contained in program F 's. Numerically bad things happen when the program's has an *Extra Leaf* extending beyond the function's.

Then at innocuous data \mathbf{x} too near that Extra Leaf of Pejorative Surface the program $F(\mathbf{x})$ produces undeservedly badly inaccurate results though $f(\mathbf{x})$ is unexceptional.

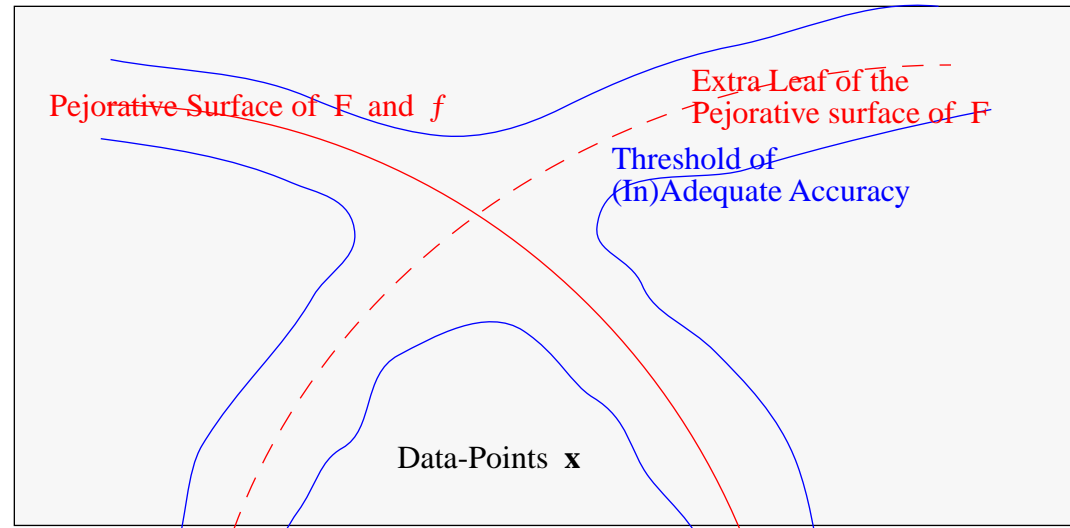
All or Most Accuracy is Lost if Data lie on a “Pejorative” Surface



Accuracy of $F(\mathbf{x})$ is Adequate at Data \mathbf{x} far enough from Pejorative Surfaces.

Let $\pi(\mathbf{x}) :=$ (distance from \mathbf{x} to a nearest Pejorative Surface). *Typically* (but not always) $\|f_r(\mathbf{x})\|$ is roughly proportional to $1/\pi(\mathbf{x})$ while $\pi(\mathbf{x})$ is small enough; then roundoff's uncertainty $\rho > \|\mathbf{r}\|$ can cause program $F(\mathbf{x})$ to lose roughly $\text{Const.} - \log(\pi(\mathbf{x}))$ dec. digits to roundoff. Since $-\log(\rho)$ is roughly the number of sig. dec. carried by the rounded arithmetic, the number of correct decimal digits left in $F(\mathbf{x})$ will be roughly $\min\{-\log(\rho), -\log(\rho) + \log(\pi(\mathbf{x})) - \text{Const.}\}$ while $\pi(\mathbf{x})$ is small enough.

Therefore some small threshold Ξ exists for which $F(\mathbf{x})$ is accurate enough only while $\pi(\mathbf{x}) > \Xi$.



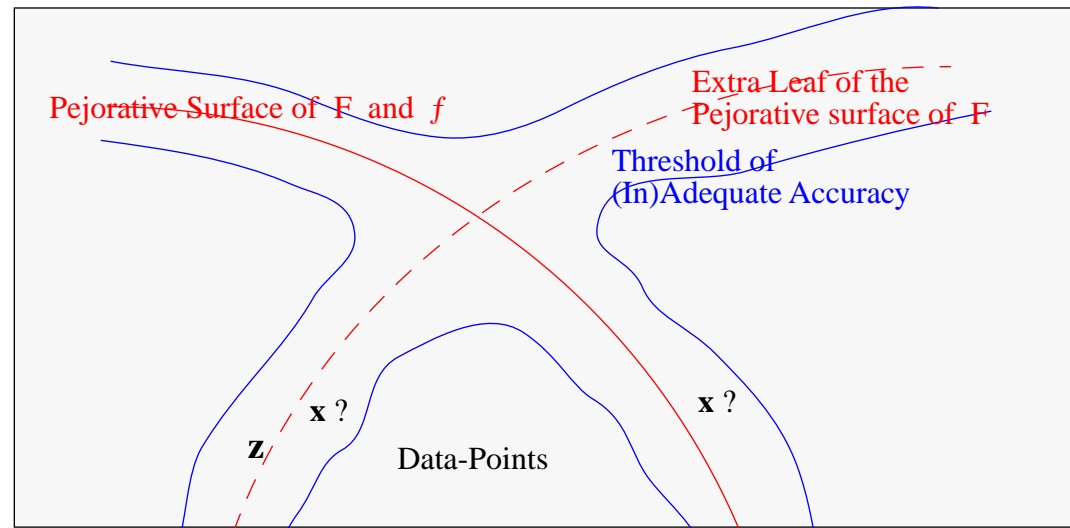
$\pi(\mathbf{x}) :=$ (distance from \mathbf{x} to a nearest Pejorative Surface of program F). Typically, the number of correct decimal digits in the result $f(\mathbf{x}, \mathbf{r})$ from program $F(\mathbf{x})$ is roughly $\min\{-\log(\rho), -\log(\rho) + \log(\pi(\mathbf{x})) - \text{Const.}\}$ while $\pi(\mathbf{x})$ is small enough.

For some small threshold Ξ the accuracy of $F(\mathbf{x})$ is adequate only while $\pi(\mathbf{x}) > \Xi$.

But Ξ and $\pi(\mathbf{x})$ are unknown, as is the location - - - - of the Extra Leaf, *if it exists*.

An opportunity to discover whether an Extra Leaf exists arises when the accuracy of $F(\mathbf{x})$ is inadequate enough to arouse suspicion. Does $F(\mathbf{x})$ deserve its inaccuracy because \mathbf{x} is “Ill-Conditioned” — too close to the Pejorative Surface of f ? Or is the inaccuracy undeserved because innocuous data \mathbf{x} is unlucky — too close to an Extra Leaf?

These important questions are difficult to resolve. Why is their resolution necessary?

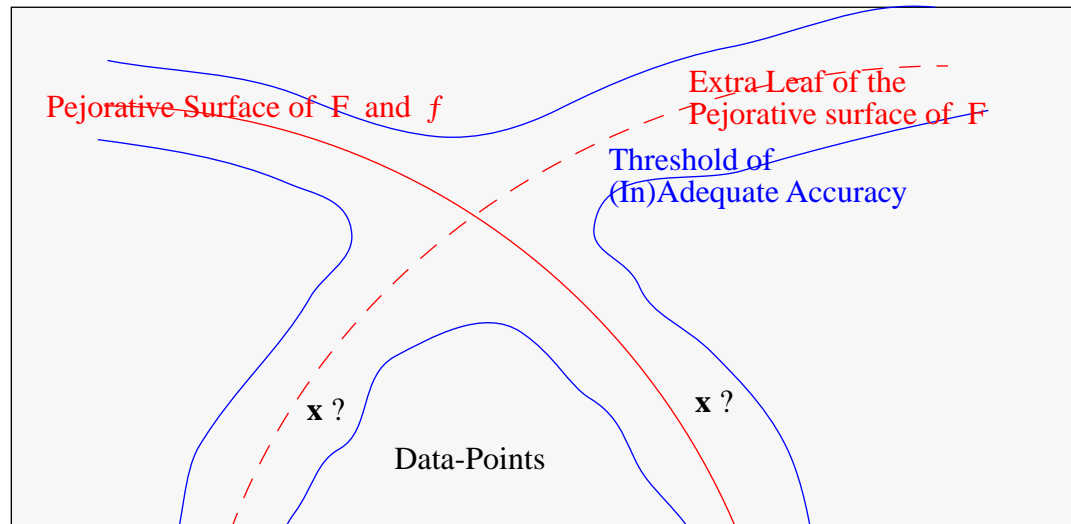


$F(\mathbf{x})$ is inaccurate enough to arouse suspicion. Does $F(\mathbf{x})$ deserve its inaccuracy because \mathbf{x} is “Ill-Conditioned” — too close to the Pejorative Surface of f ? Or is the inaccuracy undeserved because innocuous data \mathbf{x} is unlucky — too close to an Extra Leaf?

Though these important questions are difficult to resolve, their resolution is necessary lest later we accept unwittingly an utterly inaccurate $F(\mathbf{z})$ at some other innocuous data \mathbf{z} much closer to the Extra Leaf of whose existence we had chosen to remain unaware.

Two better choices present themselves:

- Enhance the likelihood of these difficult questions’ resolution by supplying tools to reduce by orders of magnitude the cost in talent and time to resolve them. OR ...
- Reduce by orders of magnitude the likelihood that these questions will arise or matter.



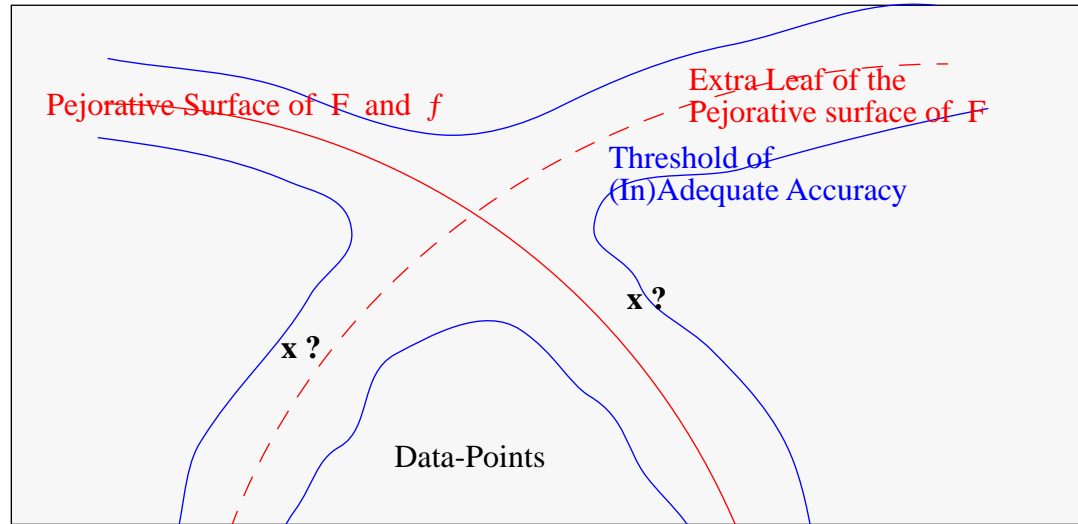
$F(\mathbf{x})$ is inaccurate enough to arouse suspicion. Where is \mathbf{x} ? Too near an Extra Leaf?

Two options present themselves:

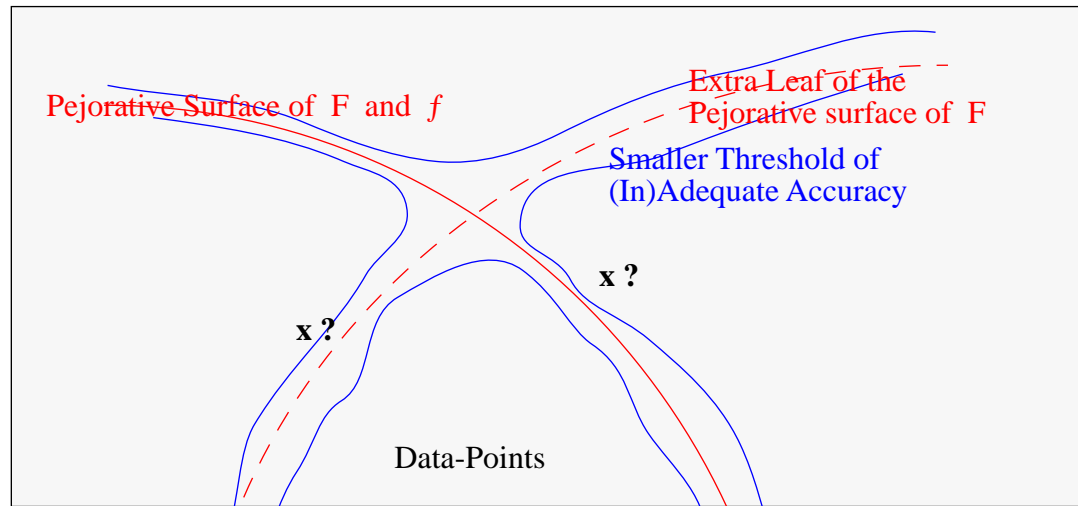
- Enhance the likelihood of these difficult questions' resolution by supplying tools to reduce by orders of magnitude the cost in talent and time to resolve them. OR ...
- Reduce by orders of magnitude the likelihood that these questions will arise or matter.

The latter option is by far the more humane and more likely to succeed. It is accomplished by changing programming languages to carry *BY DEFAULT* (except where the program specifies otherwise explicitly) extravagantly more Floating-Point precision than anyone is likely to think necessary. IEEE 754 (2008) *Quadruple* is enough; cf. COBOL's *Comp*.

Smaller $\rho \Rightarrow$ smaller threshold $\Xi \Rightarrow$ smaller volume around the Extra Leaf, if any.



Smaller $\rho \Rightarrow$ smaller threshold $\Xi \Rightarrow$ smaller volume around the Extra Leaf, if any:



Usually the threshold Ξ and **volume** around the Extra Leaf shrink in proportion with ρ .

Why is 16-byte-wide IEEE 754 (2008) *Quadruple* most likely extravagant enough?

Although the foregoing relations among arithmetic precision (ρ), distance $\pi(\mathbf{x})$ to a singularity, and consequent loss of perhaps all accuracy in $F(\mathbf{x})$ are *Typical*, the next most common relations predict a loss of about half the digits carried by the arithmetic. In other words, many programs $F(\mathbf{x})$ produce results with at least $\text{Const.} - \log(\rho)/2$ correct dec. digits no matter how near \mathbf{x} comes to a Pejorative Surface.

Some Examples:

- Nearly redundant Least-Squares problems.
- Nearly double zeros of polynomials, like the quadratic mentioned above.
- Most locations of extrema.
- Small angles between subspaces; see my web page's .../Math128/NearstQ.pdf .
- EDSAC's arccos described above. (Its Pejorative Surface looks like coarse sandpaper.)
- The financial Future Value function $FV(n, i) := ((1 + i)^n - 1)/i$ for interest rate i as a fraction, and integer n compounding periods, but *only* if FV is computed thus:

Presubstitute n for $0/0$; $FV := ((1 + i)^n - 1)/((1 + i) - 1)$. **Preserve Parentheses!**

(Because FV is the divided difference of a polynomial, it can also be computed quickly but unobviously without a division, and without losing more than a few sig. dec.)

Arithmetic precision is usually extravagant enough if it is somewhat more than twice as wide as the data's and the desired result's. Often that shrunken **volume** contains no data.

16-byte *Quad* has 113 sig. bits; 8-byte *Double* has 53; 4-byte *Float* has 24 .

What experience suggests strongly that carrying somewhat more precision in the arithmetic than twice the precision carried in the data and available for the result will vastly reduce embarrassment due to roundoff-induced anomalies?

During the 1970s, the original Kernighan-Ritchie *C* language developed for the DEC PDP-11 evaluated all Floating-Point expressions in 8-byte wide *Double* (56 sig. bits) no matter whether variables were stored as *Doubles* or as 4-byte *Floats* (24 sig. bits). They did so because of peculiarities of the PDP-11 architecture. At the time, almost all data and results on “Minicomputers” like the PDP-11 were 4-byte *Floats*.

Serendipitously, all Floating-Point computations in *C* turned out much more accurate and reliable than when programmed in FORTRAN, which must round every arithmetic operation to the precision of its one or two operand(s), or the wider operand if different.

Alas, before this serendipity could be appreciated by any but a very few error-analysts, it was ended in the early 1980s by the *C*-standards committee (ANSI X3-J11) to placate vendors of CDC 7600 & Cybers, Cray X-MP/Y-MP, and CRAY I & II supercomputers. Now most *C* compilers evaluate Floating-Point FORTRANnishly and eschew *Quad*.

Experience suggests strongly that not everyone likes *Quad* to be the default.

Why object to Default *Quad* evaluation & variables?

- 1• Languages, compilers, software and practices would have to change. This, like any other non-compatible change in the computing industry, incurs horrendous costs.
- 2• *Quad* occupies twice the memory of *Double*, especially in the cache, and takes twice as long to move through the memory system, discouraging its use in large arrays of intermediate results.
- 3• *Quad* arithmetic can take 2 to 10 times as long as *Double*, depending upon how much of a processor's area and power consumption is allocated to *Quad*. For the foreseeable future, *Quad* is likely to be microcoded, as it is on IBM mainframes, or simulated slower in software, as it is on Sun/Oracle SPARCs and Intel Processors.

Default evaluation in *Quad*, the humane option, is unlikely to be adopted widely. In consequence, at least for the foreseeable future, the other option may be our only option:

- Enhance the likelihood of these difficult questions' resolution by supplying tools to reduce by orders of magnitude the cost in talent and time to resolve them.

What tools?

What tools?

Given a program F and data \mathbf{x} at which $F(\mathbf{x})$ has aroused suspicions for some reason, we hope to find the smallest part (subprogram, block, statement, ...) of F that also arouses suspicions so that mathematical attention may be focussed upon it as a possible cause of the suspicious (mis)behavior of $F(\mathbf{x})$. Data \mathbf{x} is precious; our tools must not change data lest the change chase away the program's suspicious (mis)behavior.

Our tools will help to modify program F so as to detect hypersensitivity to roundoff by rerunning $F(\mathbf{x})$ with different roundings —

- different in Direction,
- different in Precision.

We hope a few reruns will expose a small part of F responsible for its misbehavior; this happens surprisingly often. But it does not always happen; it cannot happen in *all* cases.

Rare examples F exist that produce the same utterly wrong result $F(\mathbf{x})$ no matter how often rerun on different computer hardware, with different precisions, and with different redirected roundings, even if redirected randomly. The neatest such (counter)example I know was devised by Jean-Michel Muller in the mid-1980s and is discussed again on pp. 8 - 10 in the comprehensive handbook produced by him [2010] and his students: ...

Jean-Michel Muller's (Counter)Example

His program implements a discrete dynamical system whose state at time N is the row $[x_N, x_{N+1}]$. Starting with $x_0 := 2$ and $x_1 := -4$, the sequence $x_2, x_3, x_4, \dots, x_{N+1}, \dots$ is generated by a recurrence $x_{N+1} := 111 - (1130 - 3000/x_{N-1})/x_N$ for $N = 1, 2, 3, \dots$ in turn. An interesting challenge is the computation of, say, x_{50} using the Floating-Point hardware's arithmetic in any commercial computer or calculator, new or old.

They all get $x_{50} \approx 100$.

The correct value is $x_{50} \approx 6.0001465345614$.

Why do all those machines get *the same* utterly wrong result?

The recurrence has three fixed-points $[5, 5]$, $[6, 6]$ and $[100, 100]$. The first two are repulsive; the last is attractive. The given initial state $[2, -4]$ would generate a sequence converging to the middle fixed-point if the sequence were not perturbed by roundoff.

Computerized algebra systems can *confirm* but, so far, only a human's mathematical analysis can *discover* a numerically stable way to compute the desired sequence:

$$x_{N+1} := 11 - 30/x_N; \quad \dots \rightarrow 6.$$

Is Jean-Michel Muller's (Counter)Example Unfair?

His example's x_{50} closely approximates $x_\infty := \lim_{N \rightarrow \infty} x_N$, which is a *discontinuous* function of x_0 and x_1 wherever $x_\infty \neq 100$. This is explained in §5 of my web page's www.eecs.berkeley.edu/~wkahan/Mindless.pdf .

Floating-Point computation of a non-trivial function at its discontinuity seems foolhardy:

- If the rank of a matrix is not maximal, one rounding error will likely increase it.
- If the Jordan Normal Form of a matrix is nondiagonal, roundoff will likely undo that.
- If \mathbf{x} lies on a Pejorative Surface \mathcal{S} of f , roundoff will likely push \mathbf{x} a little off \mathcal{S} .

• • •

To counter objections to Muller's (Counter)Example, §6 of [.../Mindless.pdf](http://www.eecs.berkeley.edu/~wkahan/Mindless.pdf) has a different example $g(x) := t(q(x)^2)$ which is infinitely differentiable for all $x > 0$, as is $q(x)$; and $t(z)$ is infinitely differentiable for all z . However, when the obvious program $G(X) := T(Q(X)^2)$ is invoked to compute $G(11.)$, $G(12.)$, $G(13.)$, ..., $G(9999.)$, all but a few computed values turn out to be 0.0, which is wrong. Depending upon the precision, radix, and rounding of the arithmetic, at most a few computed values turn out to be 1.0 correctly. No mindless diagnostic tool can expose the naive part of program G unless the Math. Library's EXP has been implemented in an unlikely way.

Fortunately, this simple contrived smooth example G is extremely atypical.

A Tool for Recomputation with Redirected Rounding

IEEE 754 provides four Rounding Modes selectable (ideally) by the programmer:

The default Round-to-**Nearest** (even), Round **Up**, Round **Down**, Round-towards-**Zero**

These modes are ill-supported by programming languages; JAVA outlaws all but the first.

Given a program F and data \mathbf{x} whose result $F(\mathbf{x})$ has aroused suspicion, perhaps because \mathbf{x} is closer to the Pejorative Surface of F than ensures adequate accuracy, the user/debugger of F would use this software tool to rerun all or parts of $F(\mathbf{x})$ to find a part that seems hypersensitive to roundoff. The tool would change all the Floating-Point operations within a user-specified scope to round in a user-specified direction, and then rerun at least that scope's subprogram with *exactly* its input data that was supplied when the result of $F(\mathbf{x})$ aroused suspicion. (Of course, suspicion is insufficient for conviction.)

A crucial property of the tool is that each rerun runs about as fast as did the unaltered code. This is crucial because loops traversed a few billion times in several seconds will have to be rerun; and rerunning them too slowly will preclude that they be rerun at all.

Also crucial is that reruns must replicate intermediate results exactly up to the point where rounding is first redirected. This may take special declarations to control resources on platforms offering both resource-sharing with diverse users, and concurrency using many processors or cores. If differently many of them act in different runs, bugs flitting in and out as resources change may never be caught.

How Well does Recomputation with Redirected Rounding Work?

It works astonishingly well at exposing hypersensitivity to roundoff despite that, as we have just seen above, no mindless tool can do so infallibly. Rerunning with Redirected Roundings works on ten examples in [<.../Mindless.pdf>](#), and on all the examples appearing in the lengthy list on p. 22 of [<.../NeeDebug.pdf>](#). A typical example is summarized here; it comes from the section titled “Difficult Eigenproblems” in [<www.eecs.berkeley.edu/~wkahan/MathH110/HilbMats.pdf>](#) .

The data consist of symmetric positive definite integer matrices A and H . Sought is a column \mathbf{v} of the eigenvalues λ that satisfy $A \cdot \mathbf{b} = \lambda \cdot H \cdot \mathbf{b}$ for some $\mathbf{b} \neq \mathbf{0}$. Three such columns get computed:

- One column \mathbf{u} is computed by MATLAB’s `eig(A, H)`.
- Another column \mathbf{w} is computed by MATLAB’s `eig(X*A*X, X*H*X)` where X is obtained from the identity matrix by reversing its rows.
- A third column \mathbf{v} is obtained from the squared singular values of a bidiagonal matrix derived in an unobvious way from the given A and H because they are both Hilbert matrices. (Rarely would we have an option to compute a third column.)

In the absence of roundoff we should get $\mathbf{u} = \mathbf{v} = \mathbf{w}$, but the three computed (& sorted) columns disagree in their leading digits. ...

Columns \mathbf{u} , \mathbf{v} and \mathbf{w} were computed with arithmetic rounded the default way To Nearest. Column $\Delta\mathbf{u}_o = \mathbf{u}_o - \mathbf{u}$ shows how \mathbf{u} changed when computed with rounding directed Toward Zero. Similarly $\Delta\mathbf{u}_\uparrow$ shows how rounding Up changed \mathbf{u} , and $\Delta\mathbf{u}_\downarrow$ is for rounding Down. Likewise for $\Delta\mathbf{v}_{\dots}$ and $\Delta\mathbf{w}_{\dots}$.

\mathbf{u}	$\Delta\mathbf{u}_o$	$\Delta\mathbf{u}_\uparrow$	$\Delta\mathbf{u}_\downarrow$	\mathbf{v}	$\Delta\mathbf{v}_o$	$\Delta\mathbf{v}_\uparrow$	$\Delta\mathbf{v}_\downarrow$	\mathbf{w}	$\Delta\mathbf{w}_o$	$\Delta\mathbf{w}_\uparrow$	$\Delta\mathbf{w}_\downarrow$
0.255	-0.007	-0.004	-0.389	0.2095058938478430	-3e-16	3e-16	-3e-16	0.247	-0.029	0.002	-0.001
0.386	-0.060	-0.006	-0.136	0.3239813175038243	-9e-16	7e-16	-9e-16	0.377	-0.101	0.001	-0.000
0.512	-0.133	-0.006	-0.133	0.4391226809250292	-12e-16	12e-16	-12e-16	0.502	-0.137	0.001	0.001
0.631	-0.126	-0.006	-0.126	0.5528261852845718	-19e-16	22e-16	-19e-16	0.622	-0.129	0.002	0.002
0.740	-0.114	-0.005	-0.115	0.6612493756197405	-22e-16	26e-16	-22e-16	0.731	-0.115	0.003	0.004
0.833	-0.098	-0.004	-0.099	0.7603044306722687	-26e-16	36e-16	-26e-16	0.825	-0.098	0.003	0.005
0.908	-0.078	-0.002	-0.079	0.8461150279850096	-33e-16	36e-16	-33e-16	0.903	-0.077	0.003	0.005
0.962	-0.056	-0.001	-0.056	0.9152685078254560	-39e-16	40e-16	-39e-16	0.959	-0.055	-0.052	0.003
0.993	-0.031	-0.000	-0.032	0.9649935940457747	-40e-16	42e-16	-40e-16	0.992	-0.032	-0.031	0.001
5.724	-4.732	-3.016	-4.732	0.9932996529571477	-41e-16	44e-16	-41e-16	1.151	-0.159	-0.159	-0.005

Which column, if any, can be trusted? Rerunning each computation in three rounding modes reveals that \mathbf{v} is almost unperturbed by redirected roundoff, but it perturbs \mathbf{u} and \mathbf{w} by about as much as they differ from \mathbf{v} and each other. Afterwards an error-analysis confirms \mathbf{v} 's accuracy and explains why MATLAB's \mathbf{u} and \mathbf{w} must be inaccurate.

Redirected Rounding's Implementation Challenges

At first sight, Redirected Roundings appear to be implementable via a pre-processor that rewrites a chosen part of the text of the program being debugged and then recompiles it.

It's not always that easy.

Redirected Rounding is outlawed by JAVA and some other programming languages.

The most widespread computers redirect rounding, when they can, from a *Control Register* treated by most languages and compilers as a global variable. Some other computers redirect roundings from op-code bits that must be reloaded to change. In consequence, precompiled modules like DLLs may be affected unpredictably.

Many optimizing compilers achieve concurrency by keeping pipelines filled; to do so they interleave instructions from otherwise disjoint blocks of source-code, and "Inline" the Math. Library's functions. Then the scope of redirected rounding may be unpredictable.

For more see §14 of www.eecs.berkeley.edu/~wkahan/Mindless.pdf .

Redirected Rounding's goal may be easier to reach with a different software tool:

Recomputation with Higher Precision

It doesn't have to be much higher.

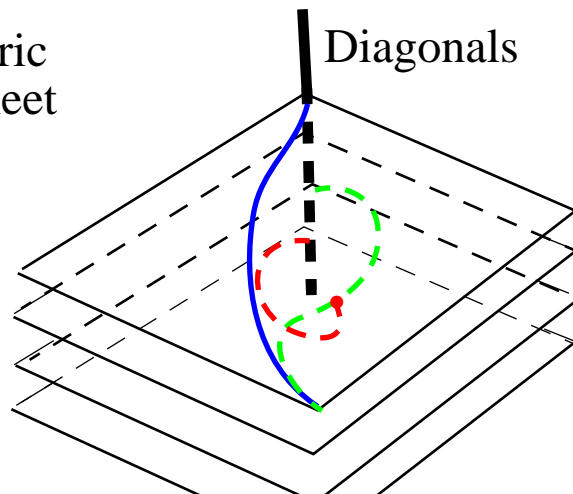
A Tool for (Slower) Recomputation with Higher Precision

This tool would ease the task of running two programs $F(\mathbf{x})$ and $\mathbb{F}(\mathbf{x})$ in lock-step. Here \mathbb{F} is derived from F by promoting all Floating-Point variables and some (probably not all) constants to a higher precision. Both programs could start with the same data \mathbf{x} .

The programs are NOT intended to be run forward in lock-step until they first diverge. That would be pointless because so many numerical processes are forward-unstable but backward-stable; this means that small perturbations like roundoff can deflect the path of a computation utterly without changing its destination significantly. For instance, the path of Gaussian Elimination with row-exchanges (“Pivoting”) can be deflected by an otherwise inconsequential rounding error if two candidates for pivots in the same column are almost equal. Deflection occurs often in eigensystem calculations; roundoff can change the order in which eigenvalues are revealed without much change to computed eigenvalues.

All the symmetric matrices in a sheet have the same eigenvalues.

Adjacent sheets differ by practically negligible roundoff.



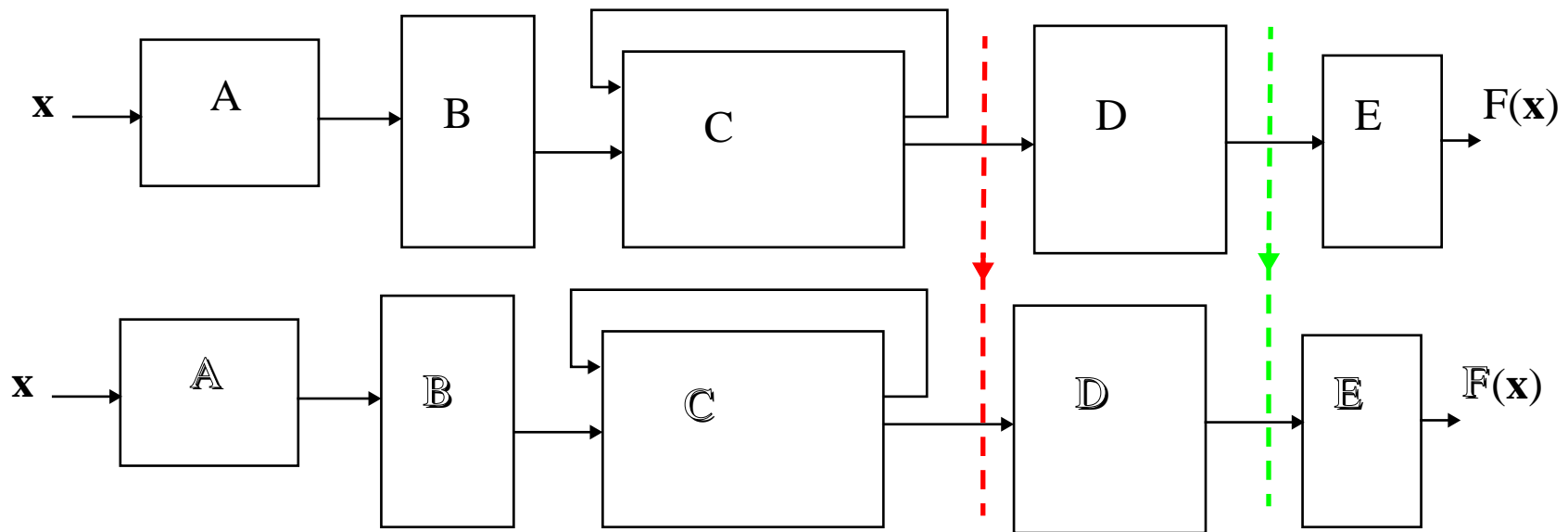
Paths followed during a program's computation of eigenvalues with ...

... no rounding errors

... the usual rounding errors

... and altered rounding errors

Instead of running F and \mathbb{F} in lock-step from their beginnings, the user of this tool will choose places in program F that I shall call “stages”. He will run $F(\mathbf{x})$ up to a chosen stage and then copy the values of all the variables alive at that stage exactly to their counterparts in \mathbb{F} ; then run \mathbb{F} to its end to see how much its result disagrees with $F(\mathbf{x})$. If they disagree too much, a later stage will be chosen; if they agree closely, an earlier stage will be chosen. With luck two adjacent stages will straddle a short section of F that causes $F(\mathbf{x})$ and $\mathbb{F}(\mathbf{x})$ to disagree too much. This section attracts focussed suspicion.



Keep in mind that *suspicion* is not yet *conviction*, which requires an error-analysis.

How Well does Recomputation with Higher Precision Work?

It almost always works, even if no short segment between stages of F can be blamed for a substantial disagreement between $F(x)$ and $\mathbb{F}(x)$, as is the case for Muller's Example. If all of program F has to be replaced by a better idea, this fact is well worth knowing.

Copying to \mathbb{F} all the values of variables in F alive at a stage can be extremely tedious without help from a software tool. And help is needed to keep track of all the technical decisions that cannot be taken out of the tool-user's hands. For instance ...

- Which functions in F from its Math Library (*log*, *cos*, ...) should not be replaced in \mathbb{F} by their higher precision counterparts ?
- Which literal constants in F should not be replaced in \mathbb{F} by their higher precision counterparts ?
- Which iterations' termination criteria in F should be changed for \mathbb{F} , and how?
- What is to be done for \mathbb{F} about software modules in F obtained from vendors pre-compiled without source-code ?

A tool to help recompute with higher precision is more interesting than first appears.

And after it works well it invites an error-analysis; learn how from N. Higham's book [2002].

And now for something entirely different ...

Floating-Point Exception-Handling

Conflicting Terminology:

Some programming languages, like *Java*, use “exception” for the policy, object or action, like a trap, that is generated by a perhaps unusual but usually anticipated event like a Time-Out, Division-by-Zero, End-of-File, or an attempt to Dereference a Null Pointer.

Here I follow IEEE 754’s slightly ambiguous use of “Floating-Point Exception” for a class of events or one of them. There are five classes:

INVALID OPERATION	like $\sqrt{-5.0}$ in a REAL arithmetic context
DIVISION-BY-ZERO	actually creation of $\pm\infty$ from finite operand(s)
OVERFLOW	an operation’s finite result is too big
UNDERFLOW	an operations nonzero result is too close to 0
INEXACT	an operation’s result has to be rounded or altered

Each exception generates, by *Default* (unless the program demands otherwise), a value *Presubstituted* for the exceptional operation’s result, continues the program’s execution and, as a side-effect, signals the event by raising a *flag* which the program can sense later, or (as happens most often) ignore.

When put forth in 1977, Presubstitution departed radically from previous practice.

When put forth in 1977, Presubstitution departed radically from previous practice which, at that time, was most often to ...

- ... Ignore Inexact, and ignore Underflow after “flushing” it to zero.
- ... Abort the program after Division-by-Zero, Overflow, and Invalid Operation as if they were Errors in a program that had failed to prevent them.

And they probably were errors if they occurred when a programmer was debugging his program by running it upon input data devised to test it.

Aborting a promulgated “Debugged” program punished its user for running ...

- ... the program upon “Invalid” input data beyond its purview, or
- ... a program that had not yet been fully debugged.

Punishment is a blunt instrument that too often befalls the innocent more than the guilty.

“ The rain it falleth on the just
And also on the unjust fella:
But chiefly on the just, because
The unjust steals the just’s umbrella.”

English jurist Lord Bowen (1835-94)

Sane computer professionals had preferred not to think about arithmetic exceptions.

Instead they acquiesced too easily to policies that punish arithmetic exceptions as errors.

Floating-Point Exceptions turn into Errors ONLY when they are Handled Badly.

Tradition has tended to conflate “Exception” with “Error” and handle both via disruptions of control, either aborting execution or jumping/trapping to a prescribed handler. ...

- FORTRAN:** Abort, showing an Error-Number and, perhaps, a traceback.
Since 1990, FORTRAN has offered a little support for IEEE 754’s defaults and flags.
- BASIC:** ON ERROR GOTO ... ; ON ERROR GOSUB to a handler.
- C :** setjmp/longjmp ... to a handler; ERRNO; abort.
Since 1999, C has let compiler writers choose to support IEEE 754’s defaults and flags.
- ADA:** *Arithmetic Error* Falls Through to a handler or the caller, or aborts.
- JAVA:** try/throw/catch/finally; abort showing error-message and traceback.
JAVA has incorporated IEEE 754’s defaults but outlawed its flags; this is *dangerous* !

These disruptions of control are appropriate when a programmer is debugging his own code into which no other provision to handle the exception has been introduced yet. Then the occurrence of the exception may well be an error; an eventuality may have been overlooked.

Otherwise IEEE Standard 754 disallows these disruptions unless a program(mer) asks for one explicitly. They must *not be the default* for any Floating-Point Exception-class.

Why *not* ?

Why must a Floating-Point Exception's default not disrupt control?

As we shall see, ...

- Disruptions of control are **Error-Prone** when they may have more than one cause.
- Disruptions of control hinder techniques for formal validations of programs.
- IEEE 754's presubstitutions and flags seem easier (although not easy) ways to cope with Floating-point Exceptions, especially by programmers who incorporate other programmers' subprograms into their own programs.
- Disruptions of control can be perilous; but so can continued execution after some exceptions. The mitigation of this dilemma requires *Retrospective Diagnostics*.

Error-Prone?

Prof. Westley Weimer's PhD. thesis, composed at U.C. Berkeley, exposed hundreds of erroneous uses of `try/throw/catch/finally` in a few million lines of non-numerical code. Mistakes were likeliest in scopes where two or more kinds of exceptions may be thrown.

See www.cs.virginia.edu/~weimer.

Floating-Point is probably more prone to error because every operation is susceptible, unless proved otherwise, to more than one kind of Exception.

Every Floating-Point operation is susceptible, unless proved otherwise, to more than one kind of exception. A program with many operations could enter a handler from any one of them, and for any of a few kinds of exception, and quite possibly unanticipatedly.

A program that handles Floating-point Exceptions by disruptions of control resembles a game ...

Snakes-and-Ladders

End	98	97	96	95	94	93	92	91	90
80	81	82	83	84	85	86	87	88	89
79	78	77	76	75	74	73	72	71	70
60	61	62	63	64	65	66	67	68	69
59	58	57	56	55	54	53	52	51	50
40	41	42	43	44	45	46	47	48	49
39	38	37	36	35	34	33	32	31	30
20	21	22	23	24	25	26	27	28	29
19	18	17	16	15	14	13	12	11	10
Start	1	2	3	4	5	6	7	8	9

... with an important difference ...

... with an important difference, for Floating-point Exceptions, ...

Invisible Snakes-and-Ladders

End	98	97	96	95	94	93	92	91	90
80	81	82	83	84	85	86	87	88	89
79	78	77	76	75	74	73	72	71	70
60	61	62	63	64	65	66	67	68	69
59	58	57	56	55	54	53	52	51	50
40	41	42	43	44	45	46	47	48	49
39	38	37	36	35	34	33	32	31	30
20	21	22	23	24	25	26	27	28	29
19	18	17	16	15	14	13	12	11	10
Start	1	2	3	4	5	6	7	8	9

None or else too many of the origins of jumps into an Exception handler are visible in the program's source-text. This hinders its formal validation.

Among programming languages, the predominant policy for handling exceptions, including Floating-Point exceptions, either disrupts control or else ignores them.

UNDERFLOW, INEXACT are usually ignored.

INVALID OPERATION, DIVIDE-BY-ZERO, OVERFLOW usually disrupt control.

A policy that predisposes every unanticipated Exception to disrupt control can have very bad consequences. *e.g.* ...

- The USS Yorktown in 1997
- The Ariane 5 in 1996
- Air France #447 in 2009
- Searches abandoned

Let's look into these examples ...

USS Yorktown (CG-48) *Aegis* Guided Missile Cruiser, 1984 — 2004



Now decommissioned, the USS Yorktown was among the first warhips extensively computerized to reduce crew (by 10% to 374) and costs (by \$2.8 million per year).

On 21 Sept. 1997, the Yorktown was maneuvering off the coast of Cape Charles, VA, when a crewman accidentally ENTERed a blank field into a data base. The blank was treated as a zero and caused a Divide-by-Zero Exception which the data-base program could not handle. It aborted to the operating system, Microsoft Windows NT 4.0, which crashed, bringing down all the ship's LAN consoles and miniature remote terminals.

The Yorktown was paralyzed for $2\frac{3}{4}$ hours, unable to control steering, engines or weapons, until the operating system had been re-booted.

Fortunately the Yorktown was not in combat nor in crowded shipping lanes.

See <www.gcn.com/Articles/1998/07/13/Software-glitches-leave-Navy-Smart-Ship-dead-in-the-water.aspx>

If IEEE 754's default had been in force, the division by zero would have insinuated into the data-base an ∞ and/or NaN, which would have been detected afterwards without a crash.

.....

The half-a-billion-dollars Ariane 5 disaster of 4 June 1996

The Ariane 5 is a French rocket that serves nowadays to lift satellites into orbit.

On its maiden flight it turned cartwheels shortly after launch and was blown up, scattering half a billion dollars worth of payload and the hopes of European scientists over a marsh in French Guiana. The disaster was traced to an Arithmetic Error,— Overflow,— in a software module monitoring acceleration (due to gravity and tidal forces) and used only while the rocket was on the launch-pad. This module’s output was destined to be ignored after rocket ignition, so it was mistakenly left enabled; but it aborted upon overflow.

A commission of inquiry blamed the disaster upon software tested inadequately. What software failure could not be blamed upon inadequate testing?

Since then the question “Who is to blame?” has spawned dozens of responses :

www.rvs.uni-bielefeld.de/publications/compendium/incidents_and_accidents/ariane5.html
...updated to 13 July 2005 by Prof. Peter B. Ladkin

Nobody else has blamed the *Fall-Through* policy of the programming language ADA.

If the overflow had not been trapped, but instead had raised a flag and generated an ∞ or any other value, both would have been ignored, and the Ariane 5 would not have crashed.

A trap too often catches creatures it was not set to catch.

.....

Air France #447 (Airbus 330) lost 1 June 2009

Modern commercial and military jet aircraft achieve their efficiencies only because they fly under control of computers that manage control surfaces (ailerons, elevators, rudder) and throttle. Only computers have the stamina to stay “on the razor’s edge” of optimal altitude, speed, and an angle of attack barely short of an *Abrupt Stall*.



35000 ft. over the Atlantic about 1000 mi. NE of Rio de Janeiro, AF#447 flew through a mild thunderstorm into one so violent that its super-cooled moisture condensed on and blocked all three *Pitot Tubes*. They could no longer sense airspeed. Bereft of consistent airspeed data, the computers relinquished command of throttles and control surfaces to the pilots with a notice that *did not explain why*. The three pilots struggled for perhaps ten seconds too long to understand why the computers had disengaged, so the aircraft stalled at too steep an angle of attack before they could institute the standard recovery procedure. Three minutes later, AF#447 pancaked into the ocean killing all 228 aboard. The computers had abandoned AF#447 too soon.

See <www.bea.aero/fr/enquetes/vol.a.point.enquete.af447.27mai2011.en.pdf>, NOVA6207 from PBS, and <www.aviationweek.com/aw/jsp_includes/articlePrint.jsp?headLine=High-Altitude%20Upset%20Recovery&storyID=news/bca0711p2.xml>

Naval embarrassment.
Half a billion dollars lost.
228 lives lost.

What more will it take to persuade the computing industry
and particularly the arbiters of taste and fashion in programming languages
to reconsider whether abortion should be the only default response
to unanticipated exceptions ?

Though a policy of continued execution after them may well pose
a difficult question for the programmer,
especially where *Embedded Systems* are concerned,
who else is better equipped to incur the obligation to answer it?

A policy that aborts execution as soon as a severe Exception occurs can also

Prematurely Abort a Search :

Suppose a program searches for an object Z that satisfies some condition upon $f(Z)$.

e.g.,

- Locate a Zero Z of $f(x)$, where $f(Z) = 0$, or
- Locate a Maximum Z of $f(x)$, where $f(Z) = \max_x f(x)$.

How can the search's trial-arguments x be restricted to the domain of f if its boundary is unknown? Is this boundary easier to find than whatever Z about f is to be sought?

Example:

$$\text{shoe}(x) := (\tan(x) - \arcsin(x)) / (x \cdot |x|^3) \quad \text{except} \quad \text{shoe}(0) := +\infty.$$

We seek a root $Z > 0$ of the equation $\text{shoe}(Z) = 0$ if such a root exists. (We don't know.) We know $x = 0.5$ lies in shoe's domain, but (pretend) we don't know its boundary.

Does your rootfinder find Z ? Or does it persuade you that Z probably does not exist?

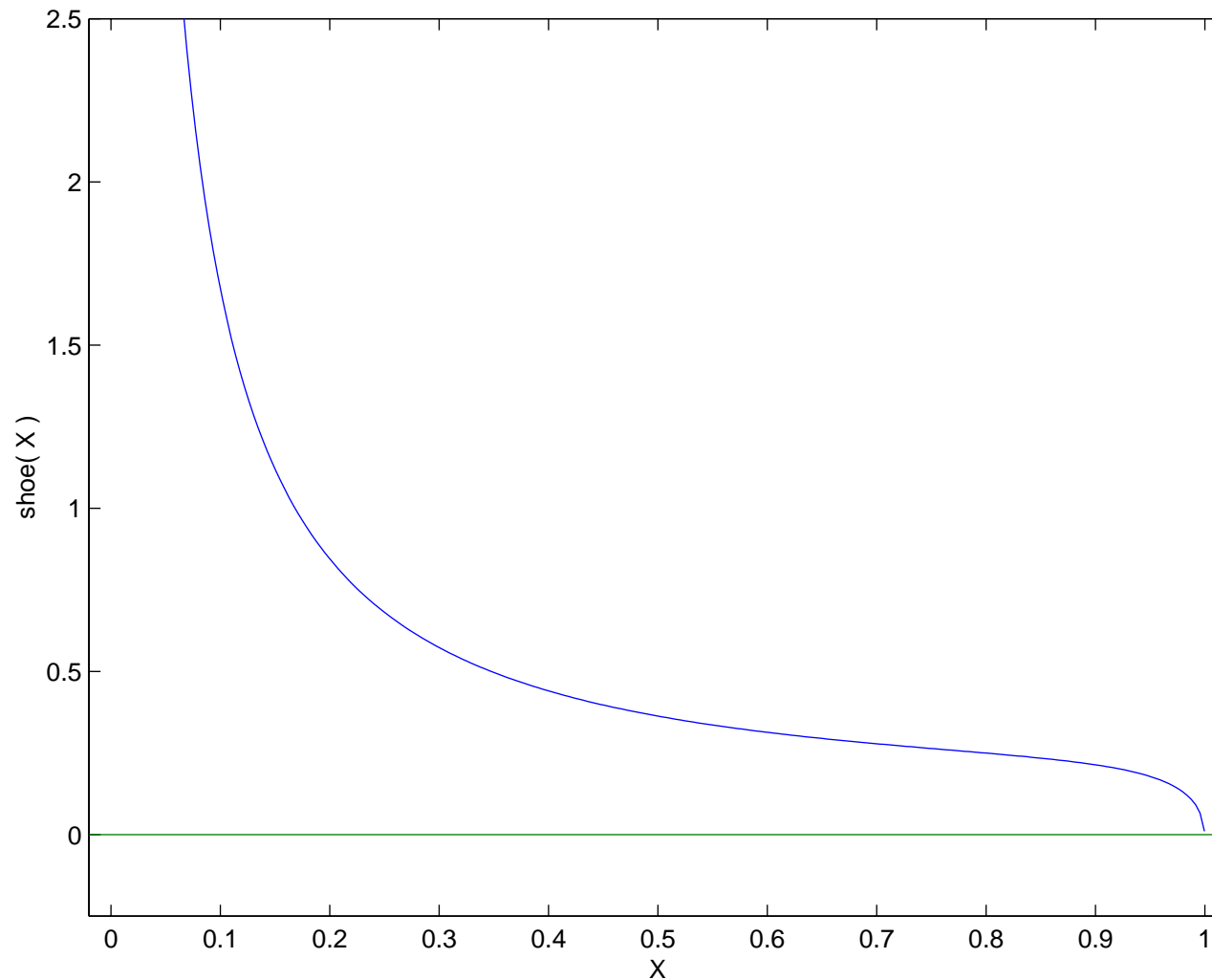
Try, say, each of 19 initial guesses $x = 0.05, 0.1, 0.15, 0.2, \dots, 0.5, \dots, 0.9, 0.95$.

`fzero` in MATLAB 6.5 on a PC said it cannot find a root near any one of them.

`root` in MathCAD 3.11 on an old Mac diverged, or converged to a huge *complex* no.

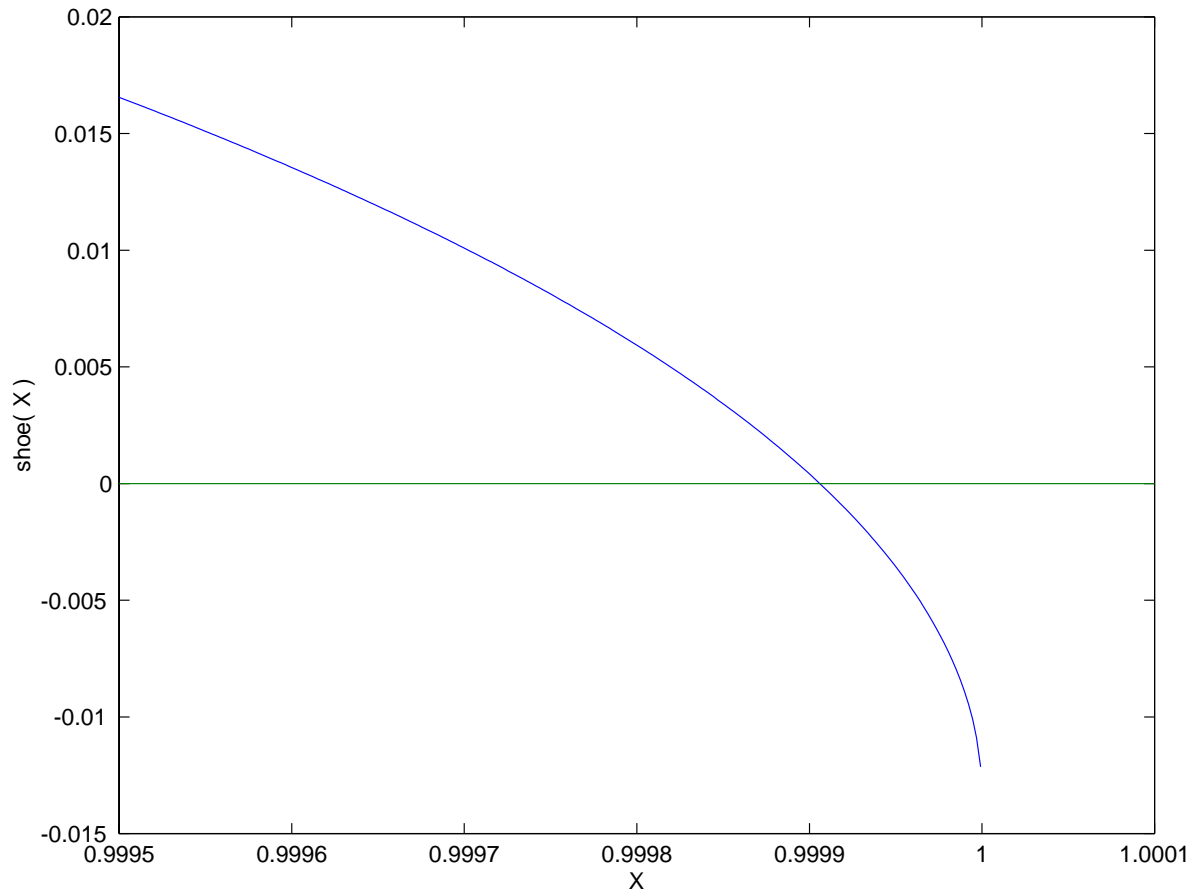
Why did [SOLV] on HP-18C, 19C and 28C handheld calculators find what they didn't?

$$\text{shoe}(x) := (\tan(x) - \arcsin(x)) / (x \cdot |x|^3)$$



If no positive Z in $\text{shoe}(x)$'s domain satisfied $\text{shoe}(Z) = 0$,
then the SHOE would leak at its toe.

$$\text{shoe}(x) := (\tan(x) - \arcsin(x)) / (x \cdot |x|^3)$$



Notice the 1000-fold change in the scale of the x - axis.

The HP-28C found the root $Z = 0.999906012413$ from each of those 19 first guesses. What did the calculator know/do that the computers didn't? ... **Defer Judgment.**

See P.J. McClellan [1987] I think some Casio calculators too may know how to do it.

Damned if you do and damned if you don't Defer Judgment

Choosing a *default* policy for handling an Exception-class runs into a ...

Dangerous Dilemma:

- Disrupting the path of a program's control can be dangerous.
- Continuing execution to a perhaps misleading result can be dangerous.

Computer systems need 3 things to mitigate the dilemma :

- 1• An *Algebraically Completed* number system for *Default Presubstitutions*.
- 2• Sticky flags to Memorialize *Leading Exceptions* in each Exception-class.
- 3• *Retrospective Diagnostics* to help the program's **User** debug it.
The program's **User** may be another program composed by maybe a different programmer.

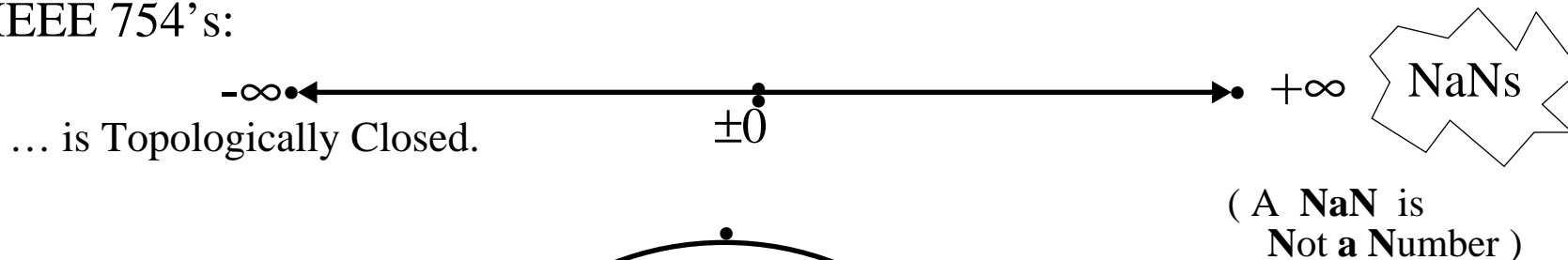
These things, to be explained hereunder, are intended for Floating-Point computations.

How well they suit other kinds of computations too is for someone else to decide.

Mathematicians do not need these 3 things for their symbolic and algebraic manipulations on paper.

Three Proper Algebraic Completions of the Real Numbers

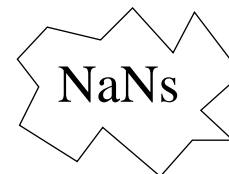
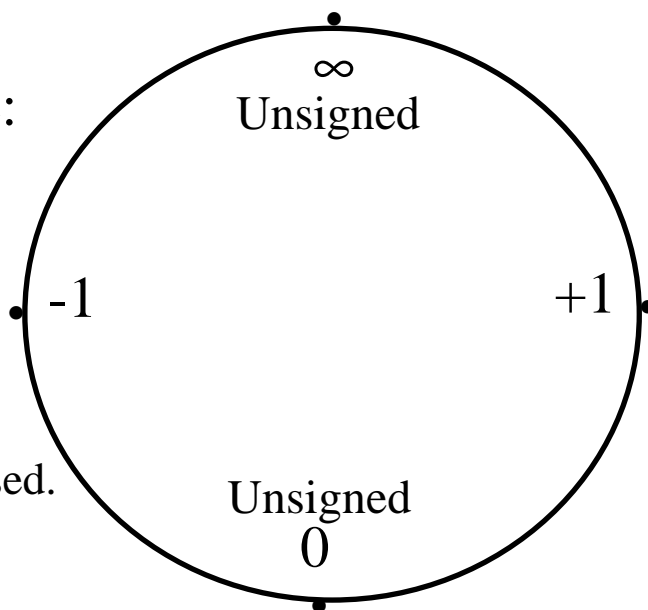
IEEE 754's:



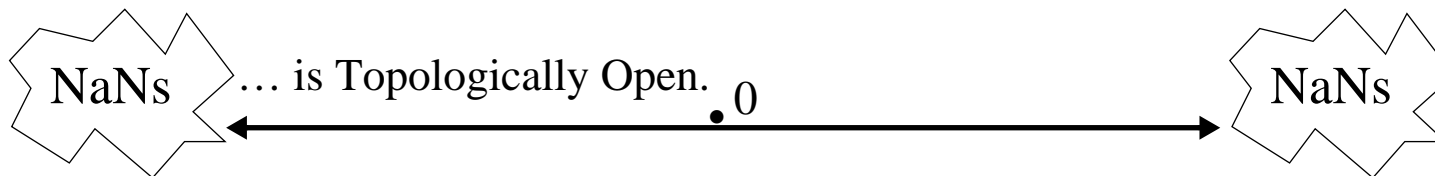
Projective Closure:

(Stereographic Projection, like the Riemann Sphere of the Complex Plane)

... is Topologically Closed.



For more about NaNs see p. 56 of <.../NeeDebug>



Proper *Algebraic Completion* maintains *Algebraic Integrity* while providing a result for *every* operation.

Algebraic Integrity: *Non-Exceptional* evaluations of algebraically equivalent expressions over the Real Numbers produce the same values.

To conserve Algebraic Integrity as much as possible, every Proper Algebraic Completion must ensure that, if Exceptions cause evaluations of algebraically equivalent expressions over the Algebraically Completed Real Numbers to produce more than one value, they can produce at most two, and if these are not $+\infty$ and $-\infty$ then at least one is NaN.

Among a few others, the Completion chosen by IEEE Standard 754 does this.

Other Completions, like APL's $0/0 := 1$ and MathCAD's $0/0 := 0$, destroy Algebraic Integrity.

For example, compare evaluations of three algebraically equivalent expressions:

x	$2/(1 + 1/x)$	$2 \cdot x/(1 + x)$	$2 + (2/x)/(-1 - 1/x)$
-1	$+\infty$!	$-\infty$!	$-\infty$!
0	0 !	0	NaN !
$\pm\infty$	2	NaN !	2

Unlike Real, Floating-Point evaluations usually conserve Algebraic Integrity at best approximately after the occurrence of roundoff and over/underflow, so some algebraically equivalent expressions evaluate more accurately than others.

For more about Algebraic Completion and Algebraic Integrity see pp. 51 - 53 of <.../NeedDebug> .

1• Presubstitution ...

... provides, within its scope, each Exception-class with a short process that supplies a value for any Floating-Point Exception that occurs, instead of aborting execution.

IEEE Standard 754 provides five presubstitutions by default for ...

INVALID OPERATION	defaults to NaN	Not-a-Number
OVERFLOW	defaults to $\pm\infty$	
DIVIDE-BY-ZERO (∞ from finite operands)	defaults to $\pm\infty$	
INEXACT RESULT	defaults to a rounded value	
UNDERFLOW is GRADUAL	and ultimately glides down to zero by default.	

These presubstitutions descend partly from the chosen Algebraic Completion of the Reals, partly from greater risks other presubstitutions may pose if their Exceptions are ignored.

Untrapped Exceptions are too likely to be overlooked and/or ignored.

- From past experience, INEXACT RESULT and UNDERFLOW are almost always ignored regardless of their presubstitutions if these are at all plausible. Ignored underflow is deemed least risky if GRADUAL.
- DIVIDE-BY-ZERO might as well be ignored because ∞ either goes away quietly ($\text{finite}/\infty = 0$) or else almost always turns into NaN during an INVALID OPERATION, which raises *its* flag.
- INVALID OPERATION should not but will be ignored inadvertently. Its NaN is harder to ignore.

Consequently, each default presubstitution has a side-effect;– it raises a *flag*. (See later.)

Ideally, a program should be allowed to choose different presubstitutions of its own.

Ideally, (on some computers today this ideal may be beyond reach)
 a program should be allowed to choose different presubstitutions of its own.

INEXACT RESULT's default presubstitution is *Round-to-Nearest* .

- IEEE 754 offers three non-default *Directed Roundings* (Up, Down, to Zero) that a program can invoke to replace or over-ride (only) the default rounding.
 ... useful for debugging as discussed previously, and for *Interval Arithmetic*.

UNDERFLOW's default presubstitution is *Gradual Underflow*, deemed most likely ignorable.

- IEEE 754 (2008) allows a kind of *Flush-to Zero* (almost), but not as the default.
 ... useful for some few iterative schemes that converge to zero very quickly, and on some hardware whose builders did not know how to make Gradual Underflow go fast.
 See www.cs.berkeley.edu/~wkahan/ARITH_17U.pdf for details.

OVERFLOW's and DIVIDE-BY-ZERO's default presubstitution is $\pm\infty$.

- Sometimes *Saturation* to $\pm(\text{Biggest finite Floating-point number})$ works better.

INVALID OPERATIONS' default presubstitutions are all NaN .

- Better presubstitutions must distinguish among $0/0$, ∞/∞ , $0\cdot\infty$, $\infty-\infty$, ...
- The scope of a presubstitution, like that of any variable, respects block structure.
- Hardware implementation is easiest with *Lightweight Traps*, each at a cost very like the cost of a rare conditional invocation of a function from the Math. library.

For examples of non-default presubstitutions see www.cs.berkeley.edu/~wkahan/Grail.pdf ,
 its pp. 1-8 explain the urgent need to implement them, and how to do it in pp. 8-10.

2• Flags

IEEE Standard 754 mandates a *Sticky flag* for each Exception-class to memorialize its every Exception that has occurred since *its flag* was last clear. Programs may raise, clear, sense, save and restore each *flag*, but not too often lest the program be slowed.

The flag of an Exception-class may be raised as a by-product of arithmetic.

The flag is a *function*, *a flag* a *variable* of data-type FLAG in memory like other variables.

The flag is not a bit in hardware's *Status Register*. Such a bit serves to update *its flag* when the program senses or saves it, perhaps after waiting for the bit to stabilize.

Any flag's data-type gets coerced to LOGICAL in conditional and LOGICAL expressions.

Any flag may also serve *Retrospective Diagnostics* by pointing to where it was raised.

An Exception that raises *its flag* need not overwrite it if it's already raised; ... faster !

Three frequent operations upon flags are ...

- Swap a saved flag with *the* current one to restore the old and sense the new.
- Merge a saved flag into *the* current *flag* (like a logical OR) to propagate one.
- Save, clear and restore all (IEEE 754's five) *flags* at once.

Reference to *the flag* is a Floating-Point operation the optimizing compiler must not swap with a prior or subsequent Floating-Point operation lest *the flag* be corrupted. This constraint upon code movement is another reason to reference *flags* sparingly.

Flags' Scopes

Variables of data-type FLAG are scoped like other variables, in so far as they respect block structure, except for *the* five Exception-classes' five *flags* which, if supported at all, have usually been treated as GLOBAL variables. Why?

By mistake; they have been conflated with bits in a status register.

The Exception-classes' five *flags* can implicitly be inherited and exported by every Floating-point operation or subprogram (or *Java* "method") unless it can specify otherwise in a language-supplied initial *Signature*.

The least annoying scheme I know for managing *flags*' inheritance and export is *APL*'s for *System Variables* []*CT* (Comparison tolerance) and []*IO* (Index Origin):

An *APL* function always inherits system variables and, if it changes one, exports the change unless this variable has been *Localized* by redeclaration at the function's start. If augmented by a command to merge a changed flag with *the flag*, this scheme works well.

Still, because they are side-effects, ...

flags are Nuisances !

flags are Nuisances. Why bother with them?

Because every known alternative can be worse :

Execution continued oblivious to Exceptions can be dangerous,
and is reckless.

Java forbids *flags*, forcing a conscientious programmer to test for an Exceptional result after every liable operation.

So many tests-and-branches are tedious and error-prone.

Recall pp. 23-4 of www.cs.berkeley.edu/~wkahan/JAVAhurt.pdf . Similarly for ...

C's single flag `ERRNO` must be sensed immediately lest another Exception overwrite it.

What can *flags* do that `try/throw/catch/finally` cannot ?

If a `throw` is hidden in a subprogram invoked more than once in the `try` clause, the `catch` clause can't know the state of variables perhaps altered between those invocations.

Recall W. Weimer's discovery that `try/throw/catch/finally` is **error-prone**.

A Floating-Point Exception *flag* costs relatively little unless the program references it.

- Apt Presubstitutions render most (not all) Exceptions and their *flags* ignorable.
- Apt non-default presubstitutions render more Exceptions and *flags* ignorable.

We should try not to burn out conscientious programmers prematurely.

Their task is difficult enough with presubstitutions and *flags*; too difficult without.

And *flags* let overlooked Exceptions be caught by *Retrospective Diagnostics*

3• Retrospective Diagnostics

We are not gods.

Sometimes some of us overlook something.

At any point in a program's execution, usually when it ends, its *Unrequited Exceptions* are those overlooked or ignored so far.

Evidence of one's existence is *its flag* still standing raised.

Retrospective Diagnostics help a program's user debug Unrequited Exceptions by facilitating interrogation of NaNs and raised *flags* now interpreted as pointers (indirectly, and perhaps only approximately) to relevant sites in the program.

Earliest Retrospective Diagnostics

See my web page's .../7094II.pdf

In the early 1960s, programs on the IBM 7090/7094 were run in batches. Each program was swept from the computer either after delivering its output, be it lines of print or card images or compile-time error-messages, or upon using up its allotment of computer time.

Often the only output was a cryptic run-time error-message and a 5-digit octal address.

I put a LOGICAL FUNCTION KICKED(...) into FORTRAN's Math. library, and altered the accounting system's summary of time used *etc.* appended to each job's output. Then ...

```
IF (KICKED(OFF)) ... executable statement ...
```

in a FORTRAN program would do nothing but record its location when executed. If later the program's execution was aborted, a few extra seconds were allotted to execute the executable statement (GO TO ..., PRINT ..., CALL ..., or REWIND ...) after the last executed invocation of KICKED . Any subsequent abortion was final.

.....

IBM's presubstitution for UNDERFLOW was 0.0 , and its other presubstitutions for ...

- DIVISION-BY-ZERO a quotient of 0.0 , or 0 for integers,
- OVERFLOW \pm (biggest floating-point number),

... were defaults a programmer could override only by a demand for abortion instead.

I added options for Gradual Underflow, and for Division-by-Zero to produce a hugest number, and for an extended exponent upon Over/Underflow. I added sticky *flags* for a program to test *etc.* any time after the Exceptions, and added Retrospective Diagnostics.

Earliest Retrospective Diagnostics continued

Each raised *flag* held the nonzero 5-digit octal address of the 7090/7094 program's site that first raised the *flag* after it had last been clear. I added tests for raised *flag* to the accounting system's summary of time used *etc.* appended to each job's output; and for each *flag* still raised at the job's end I appended a message to the job's output saying ...

“You have an unrequited ... name of Exception ... at ... octal address ... ”

This is the only change to IBM's system on the 7094 for which I was ever thanked.

... by a mathematician whose results invalidated by a DIVIDE-BY-ZERO would have embarrassed him had he announced them to the world.

My other alterations to IBM's system were taken for granted as if IBM had granted them.

Attempts over the period 1964-7 to insinuate similar facilities, all endorsed by a SHARE committee, into IBM's subsequent systems were thwarted by ...

... that's a long story for another occasion.

END OF REMINISCENCES.

.....

Note how NaNs, *flags* and Retrospective Diagnostics differ from a system's event-log:

- The system's event-log records events *chronologically*, by time of occurrence.
- NaNs and *flags* point (indirectly) to (earliest) sites (hashed) in the program.

If Exceptions were logged chronologically, they could slow the program badly, overflow the disk, and exhaust our patience even if we attempt data-mining.

Retrospective Diagnostics' Annunciator and Interrogator

How shall a program's Unrequited Exceptions be brought to the attention of its user?

- If the program's user is another program denied access to the former's *flags* by the operating system, retrospective diagnostics are thwarted.
- If the program's user is another program with access to the former's *flags*, the latter program determines their use or may pass them through to the next user.
- If the program's user is human, the program can annotate its output in a way that makes the user ...
 - *Aware* that Unrequited Exceptions exist, and then
 - *Able* to investigate them if so inclined.

“Aware” :

- Don't do it this way:

On my MS-Windows machines, some error-messages display for fractions of a second.

- Do do it this way:

On my Macs, an icon can blink or jiggle to attract my attention until I click on it.

The Math. library needs a subprogram that creates an *Annunciator*, an icon that attracts a user's attention by blinks or jiggles, which a program can invoke to annotate its output.

Clicking on an *Annunciator* should open an *Interrogator*, dropping a menu that lists unrequited Exceptions and allows displayed NaNs to be clicked-and-dragged into the list. Clicking on an item in the list should reveal (roughly) whence in the program it came.

Retrospective Diagnostics can Annoy ...

They can annoy the programmer with an implicit obligation to annotate output upon whose validity doubt may be cast deservedly by Unrequited Exceptions. This obligation is one of **Due Diligence**.

Is programming a *Profession* ? If so, one of its obligations is *Due Diligence* .

Retrospective Diagnostics can annoy a program’s user if the Annunciator resembles
The little boy who cried “Wolf !”

by calling the user’s attention to Unrequited Exceptions that seem never to matter. This may happen because the programmer decided to “Play it Safe”, actually too safe.

My IBM 7094’s retrospective diagnostics were usually torn off the end of a program’s output and discarded.

To warn or not to warn. The dilemma is intrinsic in approximate computation by one person to serve an unknown other. They share the risk. And the ***Law of Torts*** assigns to each a share of blame in proportion to his expertise, should occasion for blame arise.

.....

Retrospective Diagnostics may function better on some platforms than on others, and not at all on yet others. Debugging may be easier on some platforms than on others. Numerical software may be developed and/or run more reliably on some platforms than on others.

What Constellation of Competencies must be Collected to develop the Diagnostic Tools described herein?

Languages must be altered to support Quad by Default unless a program refuses it.

Languages must be altered to support ...

- Scopes for (re)directed roundings, and
- Scopes for non-default Presubstitutions, and for *flags*.

Compilers must be altered to augment Symbol Tables and other information attached to object modules to help debuggers (and the loaders on some architectures) implement rerunning with redirected roundings or with higher precision.

Operating Systems must be altered to support Lightweight Traps for handling non-default Presubstitutions, and *flags*' and NaNs' Retrospective Diagnostics.

Debuggers must be augmented to support users of the foregoing capabilities.

Retrospective Diagnostics may function better on some platforms than on others, and not at all on yet others. Debugging may be easier on some platforms than on others. Numerical software may be developed and/or run more reliably on some platforms than on others.

“This ... paper, by its very length, defends itself against the risk of being read.”
... attributed to Winston S. Churchill

If there be better ideas about it,
and if the reader is kind enough to pass some on to me,
this is not the subject's
Last Word.

Publications Cited

Z. Drmač & Z. Bujanović [2008] “On the failure of rank revealing QR factorization software — a case study” *ACM Trans. on Math. Software* **35** #2 Article #12, 28 pp.

Z. Drmač & Z. Bujanović [2010] “How a numerical rank revealing instability affects Computer Aided Control System Design” 12pp. <www.slicot.org/REPORTS/SLWN2010-1.pdf>

N.J. Higham [2002] *Accuracy and Stability of Numerical Algorithms* 2d. ed. ≈700 pp. (SIAM, Philadelphia).

P.J. McClellan [1987] “An Equation Solver for a Handheld Calculator” pp. 30 - 34 of the Hewlett-Packard Journal **38** #8 (Aug. 1987).

J-M. Muller et al. [2010] *Handbook of Floating-Point Arithmetic*, xxii + 572 pp. (Springer/Birkhäuser Boston, New York).

A. Ralston & E.D. Reilly Jr. [1983] eds. *Encyclopedia of Computer Science and Engineering* 2nd ed., 1694 pp. (Van Nostrand Reinhold) The articles by Wilkinson and Ralston persisted in the 4th ed. [2000] xxix+2034 pp. (Nature Publ. Group, London, England), but they disappeared from ...

E.D. Reilly [2004] ed. *Concise Encyclopedia of Computer Science* (based on the 4th ed. above) xxvi+875 pp. (Wiley, Chichester, England) which says nothing about roundoff *etc.* in Floating-Point.

A.B. Tucker Jr. [1997] ed. *The Computer Science and Engineering Handbook* 2650 pp. (CRC Press & ACM).

Responses to Questions and Comments

... from the IFIP/SIAM/NIST WoCo conference, Boulder CO, 3 Aug. 2011

from **Dr. Jeffrey Fong**, NIST Gaithersburg U.S.

"This is a sobering lecture. Critical control using computers should be duplicated online to show results of at least two independent computations agree within reasonable bounds. Do you agree that is the cost-effective way to manage a high-consequence system, where the estimated cost of a failure multiplied by the failure probability exceeds the extra cost of decision support with online verification?"

Response:

Your question and my response betray our ages. In the 1950s, when we were young, experienced engineers distrusted floating-point computation enough to follow your recommendation: Do it at least two independently different ways. This was feasible because computers did fast what had previously been done by hand using electro-mechanical calculators and slide-rules. Collecting and organizing data to put into the computer, and then presenting its output in an intelligible format, took long enough that independent recomputation's additional cost was tolerable, assuming an independent numerical method could be found. In some cases, like flutter computations for an aircraft's wings, recomputation was unavoidable because no single numerical method was likely enough to

produce reliable results. "Polyalgorithms" were proposed; these would recompute in several different ways in the hope that two or three would agree closely enough to be deemed correct.

Nowadays engineers use software packages whose numerical methods hardly ever fail despite known and unknown failure modes. Whether sufficiently different recomputation will be "cost-effective" can be decided only after assessments of too many imponderables:

- <> How likely is the chosen software package's numerical process to produce a misleading result?
- <> How likely is this misleading numerical result to cause a calamity?
- <> How much would such an imagined calamity cost, and to whom?
- <> Can a sufficiently different recomputation be found and implemented at a cost under budget and in time to meet impending deadlines?
- <> What are "reasonable bounds" for acceptance of different computed and recomputed results? Who determines these bounds, and how?
- <> What if computed and recomputed results differ excessively?

"Who shall decide, when doctors disagree, ... ?"

Epistles ... iii, l.1, by Alexander Pope (1688-1744)

"Critical control" by software in embedded systems, like a computer enhancing an aircraft's stability, would not afford their users much time to decide what to do about computational disagreements "online".

Still, like you, I would urge scientists and engineers to corroborate

(it won't be "verified") the result of a computation by a sufficiently different recomputation whenever time is available for both computations and for reconsideration if the two appear to disagree intolerably. Doing so exercises Due Diligence regardless of imponderable costs.

Enough philosophy. Now let's look at some examples:

<> Column $u = \text{sort}(\text{eig}(A, H))$ differs so much from recomputed $w = \text{sort}(\text{eig}(XAX, XHX))$ on pp. 41-2 that their errors call out for reconsideration. They are due to a shared near-null-space of columns z that both A and H nearly annihilate.

<> Single-precision's abrupt stall on p. 20 was thought to be corroborated by double-precision recomputation but, despite agreement, both were wrong, each for a different reason.

<> Mishaps that befell the Yorktown, Ariane 5 and AF447 (pp. 54-6) could not have been averted by independent recomputation so long as the default (unvoiced) policy of abortion, upon any unanticipated exception deemed "error", precluded completion of one computation. Incidentally, the rocket's computers were triply redundant.

.....

from **Dr. John Reid**, JKR Associates, U.K.

"Fortran 2003 contains facilities for controlling the modes of roundoff, although vendors are not obligated to support them."

Response:

Linguistic support for modes (rounding, precision, exception-handling) and flags is crucial for their utility. Mere compiler access to them, which is what I have had, treats them as global variables requiring too many explicit saves and restores, and does not protect these from optimizing compilers that move mode and flag references ahead or after arithmetic evaluations, thus corrupting scopes. 1980s support for modes and flags by the SANE (Standard Apple Numerics Environment) on 680x0-based Macs overburdened the programmer with the Localization of modes and flags that should have been automatic; compare p. 67 here with "Apple Numerics Manual" 2d ed. (1988) Addison-Wesley, scrapped by John Sculley in the 1990s when he put "Power" processors into Macs.

Ideally a programmer indifferent to modes and flags should not have to mention them in his program except perhaps when debugging; see the next Response. ...

from **Dr. John Reid**, JKR Associates, U.K.

"What I mostly do is write programs and check them for bugs. I therefore want the program to stop if a serious exception such as divide-by-zero occurs. My experience is that I have to request this. The default is to substitute a special value and continue, which is exactly the behavior that you want."

Response:

Like you, while debugging I have to bracket those blocks of the program where default presubstitutions would be unwelcome by statements that enable and disable traps, provided the language offers such statements. Then they must be removed before the program is put into service lest it abort prematurely. Its documentation must specify its output for inputs that precipitate an unwanted or unanticipated exception. I prefer NaNs to innocuous-looking but incorrect numerical output with a raised FLAG that is too likely to be ignored. However, NaNs can be dangerous too; for instance, among the inputs to an unwary program, NaNs can put it into an endless WHILE (Xnew .NE. Xold) DO loop.

MATLAB is different. It lacks access to FLAGS, and requires that blocks in which default presubstitutions are acceptable be bracketed by statements that disable and re-enable WARNING messages. WARNINGS are now diverse, with elaborate links to MATLAB's debugger, much more complicated than FLAGS and less helpful to the programmer who wishes to cope with all exceptions in his program without troubling its user.

Inserting/removing enabling/disabling or WRITE statements followed by recompiling has a serious flaw: the object-code debugged differs from the object-code put into service. They can differ in optimization and register allocation by the compiler, thus obscuring its bugs; see §3 of Drmač & Bujanović [2008]. If I had my way, enabling/-disabling would be unnecessary. Instead FLAGS and NaNs would point retrospectively to the places where they were raised or created in a program unaltered by recompilation.

.....

from **Sir Brian Ford**, NAG, U.K.

"Why have we been so unsuccessful in addressing and correcting these issues within the technical computing industry over the last forty years?"

Response:

Reform is inhibited by ignorance of better possibilities and by costs, the costs of changes and the costs of details, incurred initially by implementers more than beneficiaries of reform. And the talent that could institute these reforms has been fully engaged elsewhere.

Attempts to change programmers' styles and habits incur discouraging costs. Linguistic support for MODEs and FLAGS implies explicitly an obligation to attend to eventualities that previously had been ignored as uneconomical or impossible for a programmer to handle properly. How

many programmers would welcome the added burden of these obligations? That burden cannot be borne by programs intended to be widely portable so long as compiler "vendors are not obligated to support" and some choose not to support Modes and Flags, as John Reid mentioned above.

"Le bon Dieu est dans le détail." (God is in the details.)

... often attributed to Gustave Flaubert (1821-1880)

"Der Teufel steckt im Detail." (The devil is in the details.)

(First said in the 20th century, but by whom I don't know.)

Whatever resides in details, whoever pays to cope with them must wonder whether costs will ever be recovered. Apple abandoned its SANE just before finding out if it would attract developers of numerical software to prefer Macs over other platforms, and so invade a market hitherto dominated by more expensive workstations. Now this market is negligible compared with the markets from which Apple profits most today.

Here is an instance of a detail pertinent to your question: Properly to support recomputation with different roundings or precisions requires compilers to augment their symbol tables AND object-codes with marks to identify the line of source-code and subprogram from which each line of object-code came, especially if the math library is "inlined" and when aggressive optimization exploits concurrency in pipelines, cores and threads. These marks are needed to debug the object-code. Though source-code would be easier to debug if recompiled with optimization inhibited, that would obscure a deployed object-code's bug caused by overly aggressive optimization, as has often afflicted LAPACK.

Costly details multiply. Recomputation with redirected rounding can exploit those marks planted by the compiler in object-code only if it is susceptible to alteration by a sufficiently cultivated debugger without access to source-code. Further cultivation would enable the debugger to interrogate NaNs and raised flags and reveal where in the program they were created and first raised, provided the operating system provides lightweight traps to insert the necessary pointers when floating-point exceptions occur. "Lightweight" means that the trap-handlers can live and operate entirely within memory preallocated to the program without the time-consuming overhead incurred by changes to memory protection.

In short, the reforms I believe to be needed desperately entail a host of changes to programming practice and languages, compilers, operating systems and debuggers, changes that a few computer architectures may be unable to tolerate. The talent needed to implement such changes is preoccupied nowadays with the exploitation of parallelism on ever more diverse computer architectures, and with the exponentially growing disparities among the speeds of arithmetic, memory management, and communications. As programs spawned by our preoccupations proliferate, so do their bugs.

.....

from **Dr. Richard Hanson**, Rogue Wave, U.S.

~~~~~

"Occasionally the use of exceptions -- e.g. divide-by-zero -- helps performance by avoiding tests in inner loops. Example: Sturm sequences for [symmetric] tri-diagonal matrix eigenvalue problems."

### Response:

Thanks for this comment. I think your example's loop goes like this:

```
...{ Real finite x and a[1..n] and bb[1..n] > 0 are given.}
    d := -infinity ; ...{ or else d := -1 and bb[1] := +0 }
    k := 0 ;
    for j = 1 to n do
        { d := (x - a[j]) - bb[j]/d ; ...{ it's never -0 }
          k := k + signbit(d) } ;
    ...{ Now k counts the eigenvalues exceeding x .}
```

Minor restrictions upon the ranges of input data `a[...]`, `bb[...]` and `x` avert harmful over/underflows, so let's simplify discussion by ignoring them. Then the only noticeable exception is division-by-zero whenever `d` vanishes, making the next `d = -infinity`. The next pass around the loop divides by this infinite `d` to get a new finite `d` that is quite correct, or else the last `d = -infinity` with the correct sign. Here

```
signbit(z) := if ( z < 0 or z is -0 ) then 1 else 0 ;
```

but it is normally computed by a logical right-shift of leading bits rather than by a test-and-branch.

When the last  $d$  vanishes,  $x$  is eigenvalue  $\#(k+1)$  counting down. The loop is embedded in a program that uses  $k$  and  $d$  to find a sequence of values  $x$  convergent fast to a desired eigenvalue.

That pristine loop was devised in the 1950s by a physicist, Boris Davison, so far as I know. Nowadays the divide operation is so much slower than all the others that  $d$ ,  $k$  and  $x$  are arrays to exploit overlapped divisions by using more than one approximation  $x$  to one eigenvalue, and/or approximations  $x$  to more than one eigenvalue. This usage would be hindered if a test-and-branch were needed to avoid division-by-zero, so instead a tricky addition can be inserted thus:

```
{ d := ((x - a[j]) - bb[j]/d) + eta ;
```

here  $\eta$  is a tiny positive quantity, tinier than a rounding error, whose introduction noticeably restricts the admissible ranges of input data  $a[..]$  and  $bb[..]$ . This pornographic trick accommodates those few computers that must otherwise trap into the operating system to produce or consume each infinity, thereby taking at least an order of magnitude longer than an unexceptional division.

This example illustrates an important notion: The presubstitutions of infinity for  $bb[j]/0$  and zero for  $bb[j+1]/\text{infinity}$  avoid pornography incurred to prevent floating-point exceptions without thereby incurring severe performance penalties. Try to choose a value for  $\eta$  so tiny as maintains the validity and monotonicity of  $k$ , but not so tiny as risks overflows of  $bb[j]/\eta$ , to discover how greatly pornography inflates that pristine loop's capture-cross-section for programming

errors. More general presubstitutions' necessity and implementations are discussed in <[www.eecs.berkeley.edu/~wkahan/Grail.pdf](http://www.eecs.berkeley.edu/~wkahan/Grail.pdf)> .

The default presubstitutions of IEEE Standard 754 cannot be considered adequate without FLAGS. These figure in computations that almost never encounter exceptions like over/underflow that would invalidate results. Rather than test frequently for such exceptions, these computations test appropriate FLAGS occasionally at the programmer's convenience, and recompute by an alternate method when a raised FLAG requires it. The same result could be achieved by TRY-THROW-CATCH-FINALLY clauses, and faster, except that most programming languages cannot THROW when UNDERFLOW or INEXACT occurs. Besides, the scopes of THROWS and CATCHes are no easier to manage than the scopes of FLAGS.

.....

... from the Heilbronn Conference, Bristol University, 8 Sept. 2011

... from an **anonymous** member of the audience:

"Do you assert that defective software caused Air France #447's crash?"

**Response:**

Yes and no. AF#447 would not have crashed if any one of six mishaps had not befallen it. ...

<> Flying at 35000 ft., the aircraft entered a violent thunderstorm hidden from the weather RADAR by an intervening weak storm.

<> The storm's supercooled moisture froze in all three Pitot tubes, blocking them despite heaters intended to prevent this. (Since then, stronger heaters have been retrofitted to Airbus aircraft.)

<> Blocked Pitot tubes sent low or no airspeed indications to the instrument panel and to the automatic pilot's computer. It deemed these "speeds" to be "Invalid Data" inconsistent with continued flight at 35000 ft.

<> The automatic pilot's computer announced that it was relinquishing to the pilots command of the control surfaces (ailerons, elevator, rudder) and throttles, displaying only "Invalid Data" to say why. This is the software's defect. It did not say "Altitude and speed are inconsistent". It did not say "Try standard recovery procedure

(2/3 throttle, and maintain level flight)". Instead "Invalid Data" was classified implicitly as an error condition deserving abortion. Shortly afterwards, as the aircraft fell through warmer air, ice melted and the airspeed indicators recovered, but the computer did not inform the pilots that they could now trust their instruments.

<> At night, in pitch-black with no external visual references, three pilots tried to deduce which data was invalid from the instrument panel's multitudinous displays. Flying optimized "On the razor's edge" (close to stalling), the aircraft stalled before the pilots could figure out what to do. The crash came three minutes later.

<> The pilots must have been perplexed because the throttles had been reset to idle, which is no way to escape from a stall; and the younger copilot was wrongly pulling back on his joystick as if trying to climb while the older copilot was correctly pushing his forward to dive and gain speed. Because of a mistake in the design of Airbus's controls, neither copilot nor the senior pilot behind them realized until too late that the computer was averaging their cross-purposes, quietly cancelling them out.

Recently Jeff Wise's article "What Really Happened Aboard Air France 447" appeared in *Popular Mechanics*: see <[www.popularmechanics.com/technology/aviation/crashes/what-really-happened-aboard-air-france-447-6611877](http://www.popularmechanics.com/technology/aviation/crashes/what-really-happened-aboard-air-france-447-6611877)>. It is based upon extracts from the now recovered flight recorder. This posting on the internet is followed by a long list of commentators' nasty accusations about Air France's pilot training procedures, Airbus, and

especially the younger copilot, who appears likely to have to bear all the blame posthumously for the crash. But nobody objected to an implicit (accepted without debate or explanation) convention among programming languages that obliges no programmer to consider the effect his error-message (if any) would have upon users of his program after it aborts, nor to consider the states in which the program's data structures will be left after abortion caused by an unanticipated event deemed an error. (Is it the user's error, or the programmer's?) This convention amounts to a licence for irresponsibility among programmers, so it should be at least deprecated by computing professionals.

.....