Keeping It (Approximately) Real

D. Bindel

Courant Institute for Mathematical Sciences
New York University

cSplash, 4 Apr 2009

The name of the game

- I have a problem for which the *exact* answer is x.
- I compute, but with finite precision, and get $\hat{x}$.
- I want a small *relative error* $|\hat{x} - x|/|x|$.

I have 16 digits on my computer. Do I get 16 correct digits?

A silly calculation

What is x at the end of this calculation?

```
x = 2;
for k = 1:60
  x = sqrt(x);
end
for k = 1:60
  x = x^2;
end
disp(x);
```

A quadratic equation

Compute the smaller root of

$$x^2 - x + e^{-50} = 0.$$

```
x = ( 1 - sqrt(1-4*exp(-50)) )/2;
disp(x);
```

How many digits do we get right?

Archimedes method

- Archimedes estimated π by looking at side length L_N for an inscribed N-gon. As N gets large, the semiperimeter $NL_N/2$ goes to π.
- If L_N is the length of one side of an N-gon inscribed in the unit circle, then

$$L_{2N} = \sqrt{2\left(1 - \sqrt{1 - L_N^2/4}\right)}.$$

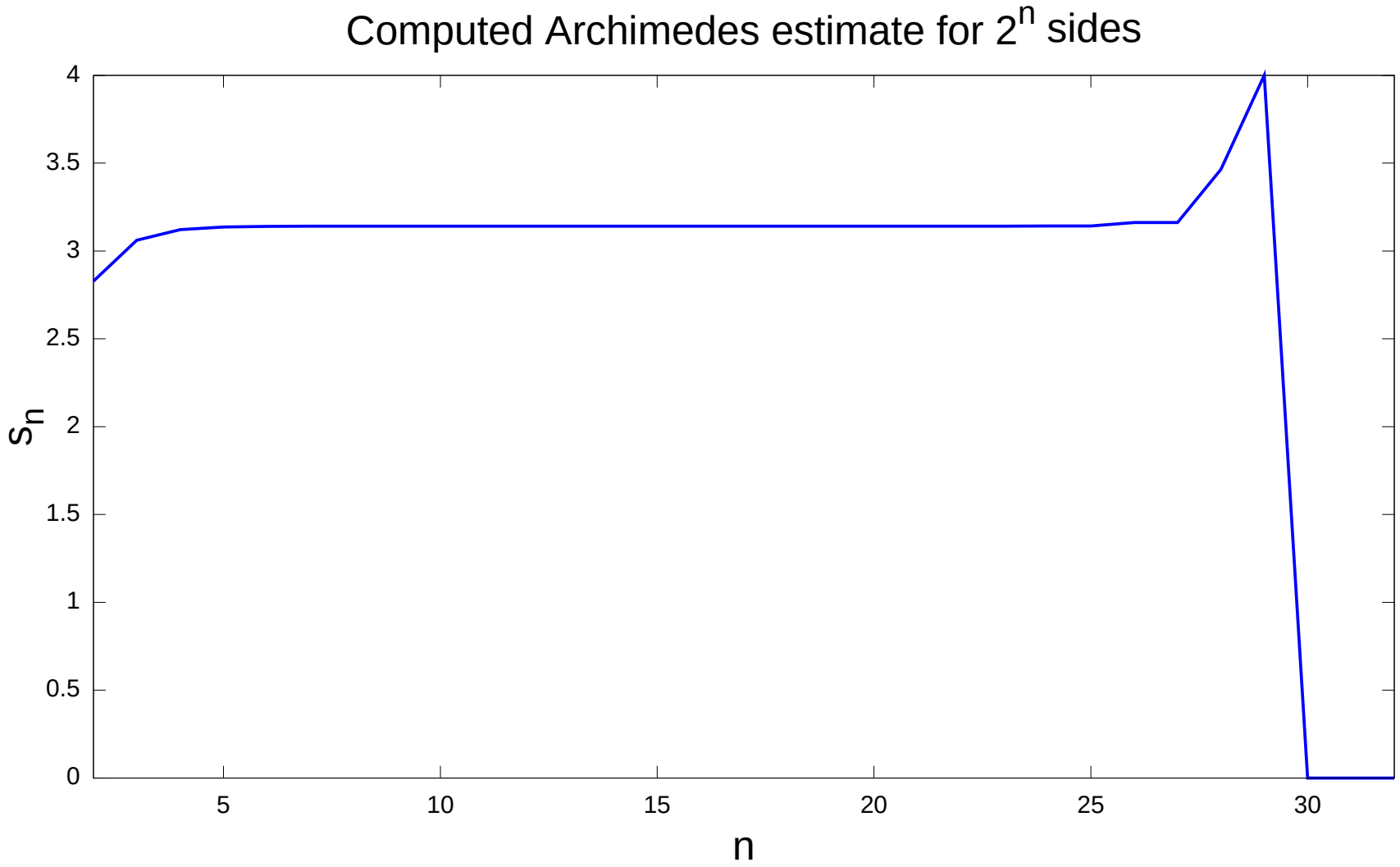
  Start from $L_4 = \sqrt{2}$.
- How well do I approximate π after 30 steps (about a billion sides?)

Archimedes method

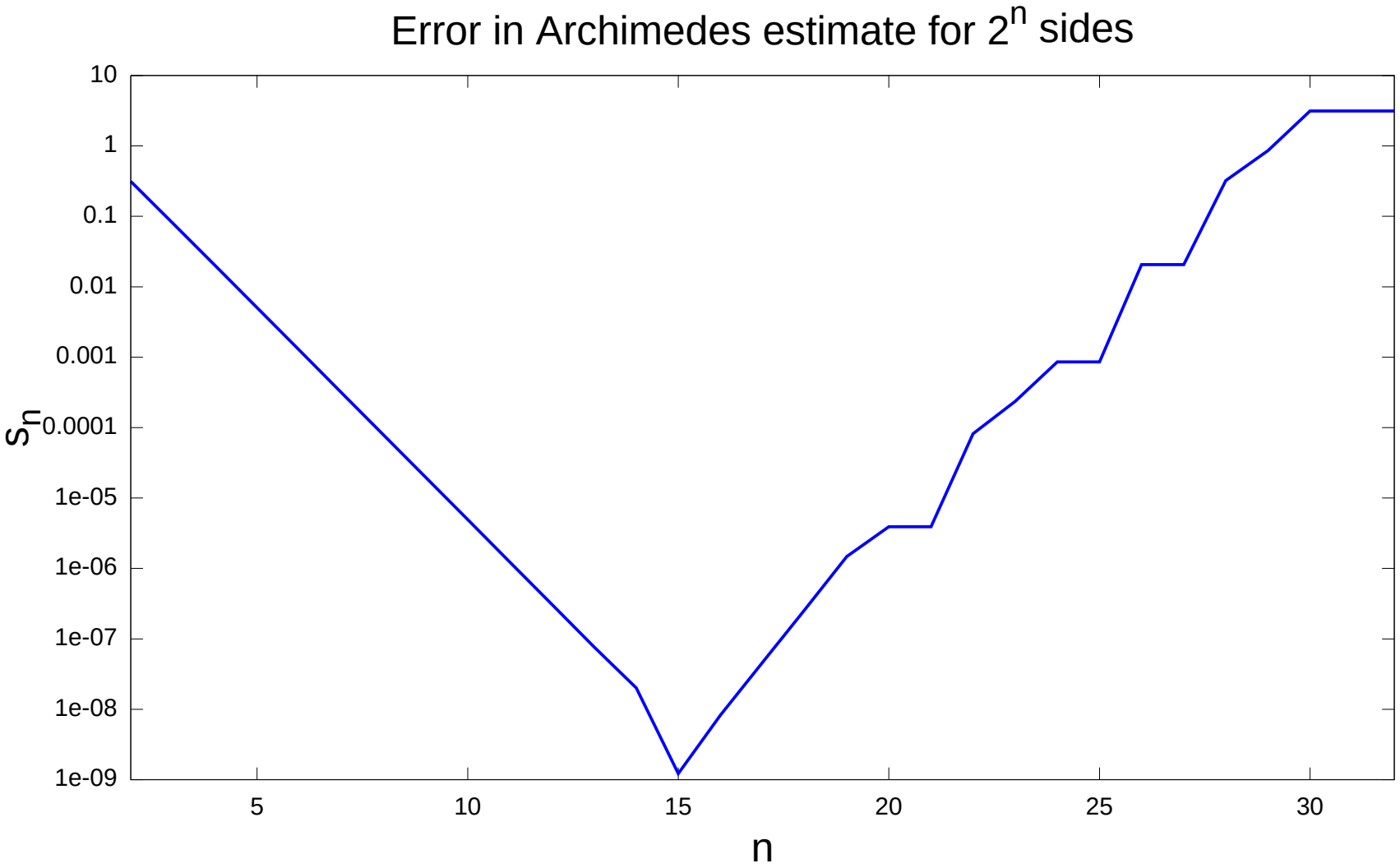
```
N = 4;
L(1) = sqrt(2);
s(1) = N*L(1)/2;

for k = 1:30
  N = N*2;
  L(k+1) = sqrt( 2*(1-sqrt(1-L(k)^2/4)) );
  s(k+1) = N*L(k+1)/2;
end
```

Archimedes, despair!



Archimedes, despair!



A little error analysis

- My computer uses IEEE floating point arithmetic (like scientific notation, but with 53 binary digits $\approx$ 16 decimal digits).
- I get the *exact result, correctly rounded* for basic floating point operations (add, subtract, multiply, divide, square root).
- This means we do basic operations to high relative accuracy. For example,

$$\mathrm{fl}(x + y) = (x + y)(1 + \delta),$$

  where $\delta < \epsilon_{\text{machine}} \approx 10^{-16}$.
- What happens when I do a sequence of steps?

Cancellation

Suppose bold digits are correct:

$$\begin{array}{rl} & \mathbf{1.09375}2543 \\ - & \mathbf{1.09374}1233 \\ \hline = & \mathbf{0.00001}1310 \end{array}$$

Inputs have six correct digits. Output has only one!

Cancellation

Suppose a and b are computed with some known error bound

$$\begin{aligned} \mathrm{fl}(a) &= a(1+\delta_a), & |\delta_a| \le \epsilon \\ \mathrm{fl}(b) &= b(1+\delta_b), & |\delta_b| \le \epsilon \end{aligned}$$

Then

$$\mathrm{fl}(a-b) = (a-b+a\delta_a - b\delta_b)(1+\delta).$$

where $|\delta| < \epsilon_{\text{machine}}$, and the relative error is

$$\frac{\mathrm{fl}(a-b)-(a-b)}{a-b} = \frac{a\delta_a - b\delta_b}{a-b}(1+\delta).$$

If $a-b \approx 0$, this might be big! Worst case:

$$\frac{\mathrm{fl}(a-b)-(a-b)}{a-b} \lesssim \frac{|a|+|b|}{|a-b|}\epsilon.$$

Paper time!

- How do we explain the behavior for our three examples? (square root and square, the quadratic equation, and Archimedes)
- How can we fix the latter two examples?

Another example

The integrals $E_n := \int_0^1 x^n e^{x-1}\, dx$ can be evaluated by

$$\begin{aligned} E_0 &= 1 - 1/e \\ E_n &= 1 - nE_{n-1}. \end{aligned}$$

These values should always be positive, and they should monotonically decay toward 0 for large n. But something goes wrong!

Disaster!

What happened?

```
 0:     6.321205588e-01
 5:     1.455329406e-01
10:     8.387707006e-02
15:     5.903379364e-02
20:    -3.019239489e+01
25:     1.927850088e+08
30:    -3.296762456e+15
35:     1.284281507e+23
40:    -1.014081007e+31
```

Paper time!

- What goes wrong?
- What happens if we reverse the recurrence?

Some more problems

1. How do we accurately evaluate $\sqrt{1+x} - \sqrt{1-x}$ when $x \ll 1$?
2. How do we accurately evaluate $\ln \sqrt{x+1} - \ln \sqrt{x}$ when $x \gg 1$?
3. How do we accurately evaluate $(1 - \cos(x))/\sin(x)$ when $x \ll 1$?
4. Here's a formula from fluid dynamics:

$$u_\theta = \frac{\Gamma_0}{2\pi r}\left(1 - \exp\left(\frac{-r^2}{4\nu t}\right)\right)$$

   Assume that we have an accurate sinh function; how can we rewrite this function for accurate evaluation when $r \ll \sqrt{4\nu t}$?
5. $f = (e^x - 1)/x$ loses most digits for small x.
   $y = e^x, f = (y-1)/\log(y)$ works fine. Why?

More problems, continued

1. For $x > 1$, the equation $x = \cosh(y)$ can be solved as

$$y = -\ln\left(x - \sqrt{x^2 - 1}\right).$$

What happens when $x = 10^8$? Can we fix it?

2. The difference equation

$$x_{k+1} = 2.25x_k - 0.5x_{k-1}$$

with starting values

$$x_1 = \frac{1}{3}, \qquad x_2 = \frac{1}{12}$$

has solution

$$x_k = \frac{4^{1-k}}{3}.$$

Is this what you actually see if you compute? What goes wrong?

Where to learn more

- Forman Acton, *Real Computing Made Real: Preventing Errors in Scientific and Engineering Calculations*. Pointed, funny, and inexpensive (Dover). Source of some of today's examples.
- Web site of W. Kahan, `http://www.cs.berkeley.edu/~wkahan`. Many notes on error analysis. Also learn how U-boats got sunk in WWII!
- G. W. Stewart, *Afternotes in Numerical Analysis*. A short introduction to numerical analysis, source of some of today's examples.
- Shampine, Allen, Pruess, *Fundamentals of Numerical Computing*. Another source of some of today's exampels.
- Nicholas Higham, *Accuracy and Stability of Numerical Algorithms*. A magisterial work – recommended mostly for advanced readers.