

# Understanding floating point computation in machine learning

Cerebras Systems

Rob Schreiber

“Floating-point arithmetic is considered an esoteric subject by many people.”

***What Every Computer Scientist Should Know About Floating-Point Arithmetic.*** David Goldberg. ACM TOMS 23:1 (1991)

# The Real Numbers

e.g.

$\pi = 3.14159265358979323846264338327950288419716939937510 58209749445923078164062862089986280348253421170679$   
82148086513282306647093844609550582231725359408128 48111745028410270193852110555964462294895493038196  
44288109756659334461284756482337867831652712019091 45648566923460348610454326648213393607260249141273  
72458700660631558817488152092096282925409171536436 78925903600113305305488204665213841469519415116094  
33057270365759591953092186117381932611793105118548 07446237996274956735188575272489122793818301194912  
98336733624406566430860213949463952247371907021798 60943702770539217176293176752384674818467669405132  
00056812714526356082778577134275778960917363717872 14684409012249534301465495853710507922796892589235  
42019956112129021960864034418159813629774771309960 5187072113499999837297804995105973173281609631859  
50244594553469083026425223082533446850352619311881 71010003137838752886587533208381420617177669147303  
59825349042875546873115956286388235378759375195778 18577805321712268066130019278766111959092164201989 .....

- We make do with an ugly approximation

0 1 0 0 0 1 0 1 0 1 1 1 0 0 0 0 (0x4560)

0 1 0 0 0 1 0 1 0 1 1 1 0 0 0 0 (s, e, m)

$$+ 17 - 15 = 2 \quad (1) . 0 \ 1 \ 0 \ 1 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0 = 1.34375$$

$$= (-1)^s * 2^{e-B} * f = 4 \times 1.34375 = 5.375$$

$$1 \leq f < 2$$

## Uniqueness

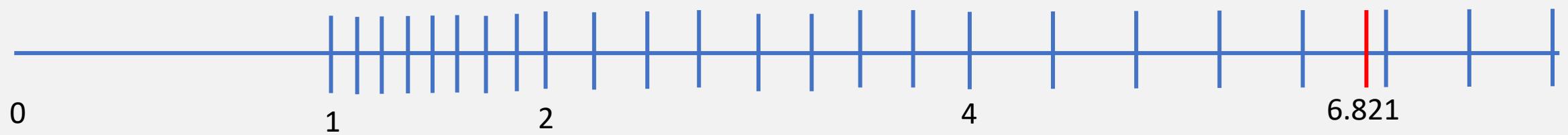
Sign, exponent, bias, fraction/significand

Normalization

Hidden bit

Rounding  $x \rightarrow \text{float}(x)$

Rounding vs truncation. Round to nearest.

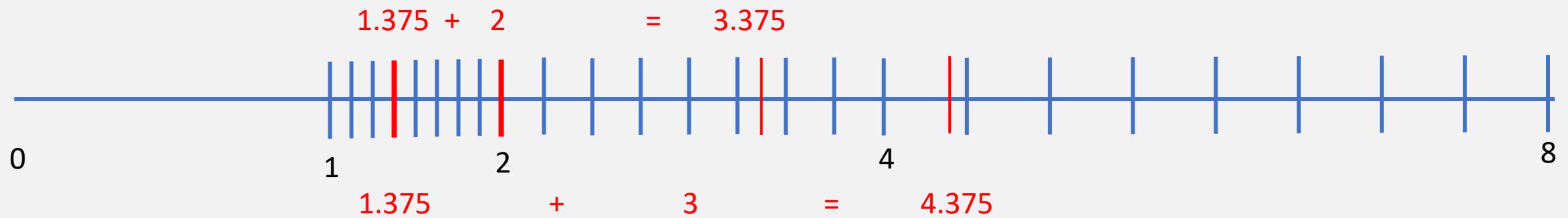


$$(6.821) = 1.101101001000101101000011100101 \dots \times 2^2$$

$$\text{float}(6.821) = ? \quad 6.5 = 1.101 \times 2^2 \quad e = 6.821 - 6.5 = .1010001 \times 2^{-1} > (1/2) \text{ ulp}$$

$$\text{float}(6.821) = ? \quad 7.0 = 1.110 \times 2^2 \quad e = 6.821 - 7.0 = .0101111 \times 2^{-1} < (1/2) \text{ ulp}$$

# Floating point *arithmetic* (+, X) -- in hardware:



$$a <+> b = \text{round}(a + b) = (a + b)(1 + \epsilon)$$

$$\epsilon \leq 2^{-\text{fraction bits}} = \frac{1}{2} \text{ ULP}$$

# How does the machine add?

$$1.234 \times 10^1 + 5.678 \times 10^{-1} \quad \text{in 4 digit decimal}$$

$$= 1.234 \times 10^1 + 0.05678 \times 10^1 \quad \text{shift the smaller for equal exponents}$$

$$= 1.29078 \times 10^1 \quad \text{add with extra precision}$$

Then round:

$$= (\text{in floating point}) 1.291 \times 10^1$$

# How much extra precision?

$$1.234 \times 10^1 + 5.678 \times 10^{-10} \quad \text{in 4 digit decimal}$$

$$= 1.234 \times 10^1 + 0.0000000005678 \times 10^1$$

$$= 1.2340000005678 \times 10^1$$

$$= (\text{in floating point}) 1.234 \times 10^1$$

**Didn't need ALL that extra precision!** Just guard digit and "sticky bit"

$$= 1.233 \times 10^1 + 1.567 \times 10^{-2} \quad (\text{shift } 1 - (-2) = 3 \text{ places to the right})$$

$$= 1.233 \times 10^1 + 0.001567 \times 10^1 = 1.233 \times 10^1 + 0.0015 b \times 10^1$$

$$= 1.2345 b = 1.235 \times 10^1 \quad (\underline{\text{round up}}, \text{ because } 5 b > \frac{1}{2} \text{ ulp})$$

# The finite set $F$

- $F \in R.$      $F \in Q.$      $0 \in F.$      $1 \in F.$      $x \in F \rightarrow -x \in F.$
- $x, y \in F \Rightarrow x + y \in F \text{ or } xy \in F.$
- Floating point arithmetic IS commutative  
 $x +_f y = rnd(x + y) = rnd(y + x) = y +_f x$
- Floating point arithmetic IS NOT associative  
 $rnd(rnd(x + y) + z) = ((x + y)(1 + \epsilon_1) + z)(1 + \epsilon_2)$   
 $rnd(x + rnd(y + z)) = ((y + z)(1 + \epsilon_3) + x)(1 + \epsilon_4)$

- A problem with the set F (thus far)
  - No zero! (due to hidden bit)
- Good answer
- Reserve  $e == \text{all zeros}$  to represent zero

- More Problems
  - Overflow ( $\text{maxval} + 1$ )
  - Underflow ( $\text{minpositive} / 3$ )
  - Incompatible machines, unportable code
- Traditional answers
  - Saturate at  $\text{maxval}$ . (So  $\text{maxval} + 2 - 1 == \text{maxval}!$ )
  - Flush to zero. Large relative error! Also, we may have  $x > y$ , but  $x - y == 0$
  - Buy my machine again...
- IEEE Standard for floating point (1985. Current: 2008)
  - Inspired by the work of Prof. W. Kahan, UCB

Directeur

**ERNEST VAUGHAN**

## ABONNEMENTS

	Un an	Six mois	Trois mois
PARIS . . . . .	20	10	5
DÉPARTEMENTS ET ALÉRIE . . . .	24	12	6
ETRANGER (UNION POSTALE) . . . .	35	18	10

## POUR LA RÉDACTION :

S'adresser à M. A. BERTHIER  
*Secrétaire de la Rédaction*

ADRESSE TÉLÉGRAPHIQUE : AURORE-PARIS

Directeur

**ERNEST VAUGHAN**

## LES ANNONCES SONT REÇUES À

142 — Rue Montmartre — 142  
AUX BUREAUX DU JOURNAL*Les manuscrits non insérés ne sont pas rendus*ADRESSER LETTRES ET MANDATS À  
A M. A. BOUIT, Administrateur

Téléphone : 102-55

# L'AURORE

Littéraire, Artistique, Sociale

# J'Accuse....!

## LETTER AU PRÉSIDENT DE LA RÉPUBLIQUE

### Par ÉMILE ZOLA

#### LETTRE

A M. FÉLIX FAURE

Président de la République

Monsieur le Président,

Me permettez-vous, dans ma gratitude pour le bienveillant accueil que vous m'avez fait un jour, d'avoir le souci de votre juste gloire et de vous dire que votre étoile, si heureuse jusqu'ici, est menacée de la plus honteuse,

lieu, des papiers disparaissaient, comme il en disparaît aujourd'hui encore; et l'auteur du bordereau était recherché, lorsqu'un *a priori* se fit peu à peu que cet auteur ne pouvait être qu'un officier de l'état-major, et un officier d'artillerie : double erreur manifeste, qui montre avec quel esprit superficiel on avait étudié ce bordereau, car un examen raisonné démontre qu'il ne pouvait s'agir que d'un officier de troupe. On cherchait donc dans la maison, on examinait les écritures, c'était comme une affaire de famille, un traître à surprendre dans les bureaux mêmes, pour l'en arrêter. Et sans une la vanille re-

Est-ce donc vrai, les choses indicibles, les choses dangereuses, capables de mettre l'Europe en flammes, qu'on a dû enterrer soigneusement derrière ce huis clos? Non! il n'y a eu, derrière, que les imaginations romanesques et démentes du commandant du Paty de Clam. Tout cela n'a été fait que pour cacher le plus saugrenu des romans-feuilletons. Et il suffit, pour s'en assurer, d'étudier attentivement l'acte d'accusation lu devant le conseil de guerre.

Ah! le néant de cet acte d'accusation! Qu'un homme ait pu être condamné sur cet acte, c'est un prodige d'iniquité. Je défie les honnêtes gens de le lire, sans que leur cœur bondisse profondément, s'inquiètent, cherchent, finissent par se convaincre de l'innocence de Dreyfus.

Je ne ferai pas l'historique des douces, puis de la conviction de M. Scheurer-Kestner. Mais, pendant qu'il foulait de son côté, il se passait des faits graves à l'état-major même. Le colonel Sandherr était mort, et le lieutenant-colonel Picquart lui avait succédé comme chef du bureau des renseignements. Et c'est à ce titre, dans l'exercice de ses fonctions, que ce dernier eut un jour entre les mains une lettre-télégramme, adressée au commandant Esterhazy, par un agent d'une puissance étrangère. Son devoir

avec lui une correspondance amicale. Seulement, il est des secrets qu'il ne fait pas bon d'avoir surpris.

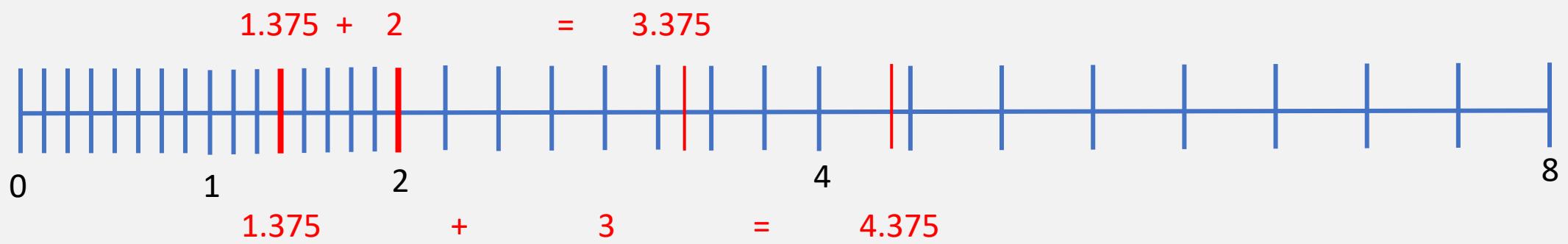
À Paris, la vérité marchait, irrésistible, et l'on sait de quelle façon l'orage attendu éclata. M. Mathieu Dreyfus dénonça le commandant Esterhazy comme le véritable auteur du bordereau, au moment où M. Scheurer-Kestner allait déposer, entre les mains du garde des sceaux, une demande en révision du procès. Et c'est ici que le commandant Esterhazy paraît. Des témoignages le montrent d'abord affolé, prêt au suicide ou à la fuite. Puis, tout d'un coup, il paye d'audace, il étonne Paris par la violence de son conseil de guerre déferrait ce qu'un conseil de guerre avait fait?

Je ne parle même pas du choix toujours possible des juges. L'idée supérieure de discipline, qui est dans le sang de ces soldats, ne suffit-elle à infirmer leur pouvoir même d'équité? Qui dit discipline dit obéissance. Lorsque le ministère de la guerre, le grand chef, a établi publiquement, aux acclamations de la représentation nationale, l'autorité absolue de la chose jugée, vous voulez qu'un conseil de guerre lui donne un formel démenti? Hélas! lorsque cela est impossible. Le général Billot a suggestionné les juges par sa déclaration, et ils ont inva-



# The IEEE Standard

- Hidden bit
- Minimum exponent reserved for 0 (if fraction == 0) and for denormalized numbers (fraction != 0)  
$$( \quad x > y \text{ now implies } x - y > 0. \quad )$$
- Maximum exponent reserved for infinity (inf) and Not a Number (NaN)
- $1/0 = \text{maxfinite} + 1 = \text{inf}$   
 $0/0 = \text{inf} - \text{inf} = \text{NaN}$
- Almost universally adopted on CPUs
- $+0$  and  $-0$



# Catastrophic cancellation

$$1.000\ 000\ 1 - 1.000\ 000\ 0 = 0.000\ 000\ 1$$

Relative error of result is  $O(1)$

The subtraction was error free, the error is in the operands

# History

- Konrad Zuse, Berlin, 1941
- Alternatives have been explored
  - Store  $\log(x)$  as a fixed-point number
  - Level-index  $x = e^e e^f$
  - Variable sized e and f
    - Tapered floating point
    - UNUMs, adds an inexact bit

# Errors. Algorithms. Error analysis

“Normal” way    **for** ( $i = 0$ ,  $\text{sum} = 0$ ;  $i < N$ ;  $i++$ ) {  $\text{sum} += X[i]$ ; }

“Compensated summation”

```
for ( $i = 0$ ,  $\text{sum} = 0$ ,  $c = 0$ ;  $i < N$ ;  $i++$ ) {
     $y = X[i] - c$ 
     $t = \text{sum} + y$  // Alas,  $\text{sum}$  is big,  $y$  small, so low-order digits of  $y$  are lost.
     $c = (t - \text{sum}) - y$  // recovers -(low part of  $y$ )
     $\text{sum} = t$  //
}
```

Analysis: relative error is  $O(N\epsilon)$  vs.  $O(\epsilon)$

(Worst case. More commonly  $O(\sqrt{N})$  growth.)

# Machine Learning, Deep Neural Networks

IEEE 32-bit (1/8/23)

(No one in ML seems to need IEEE 64-bit)

16-bit floating point

IEEE: 1 / 5 / 10

bfloat: 1 / 8 / 7 (same exponent width as IEEE 32)

# ML Problems

- Tiny gradients
  - IEEE16, smallest numbers are  $O(2^{-24})$ , gradient may underflow
  - Bfloat, low accuracy, weight + gradient = weight
- Many minibatches
  - Weight is rounded after each update. Rounding error grows
  - 32-bit weight accumulation

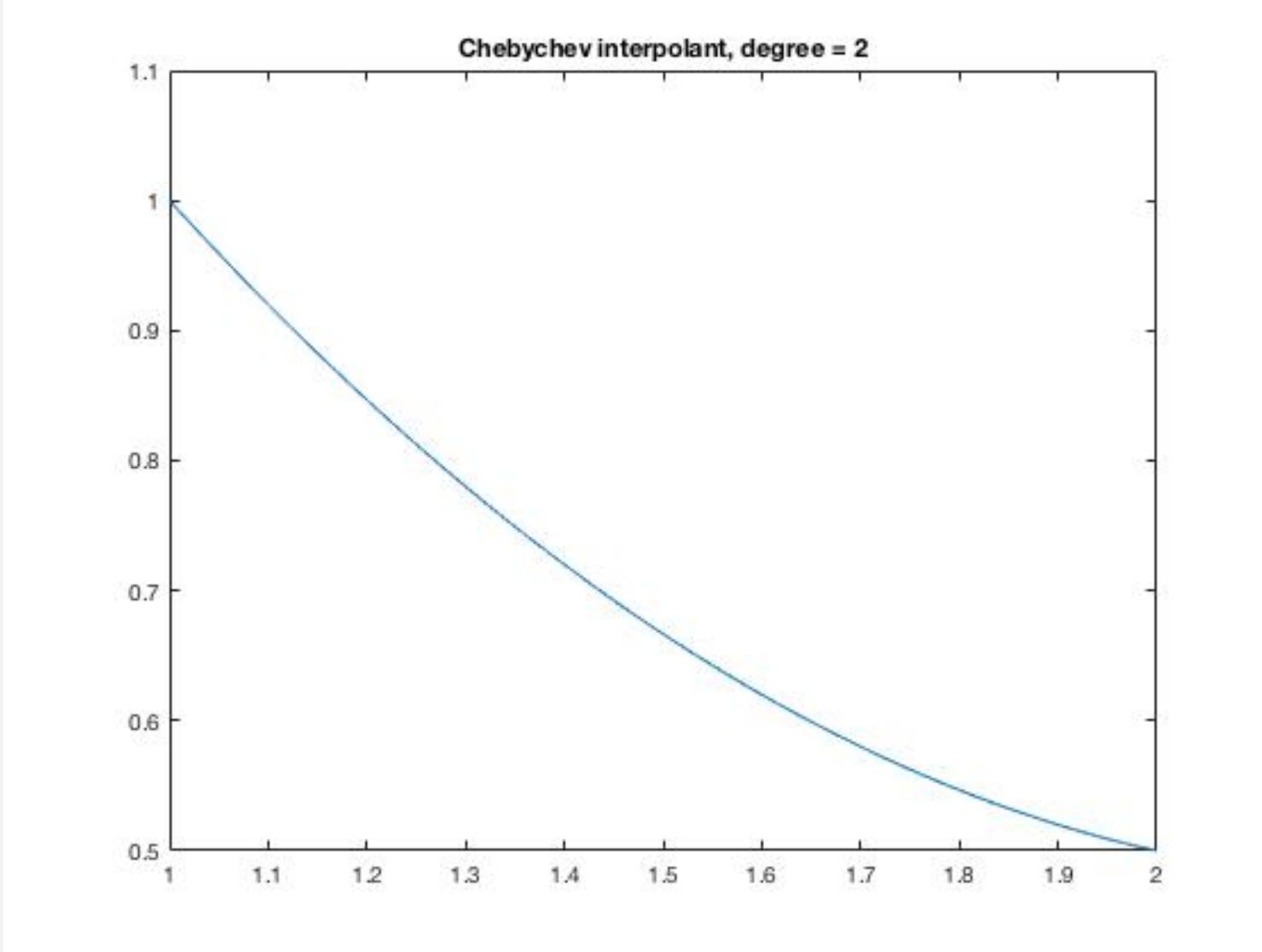
# What about division?

$$\frac{1}{x} = \frac{1}{f 2^e} = \left(\frac{1}{f}\right) 2^{-e}$$

Given an approximate reciprocal,  $y$

Suppose  $y = \left(\frac{1}{f}\right)(1 + \epsilon)$

Let  $y \leftarrow y + y(1 - fy) = \left(\frac{1}{f}\right)(1 - \epsilon^2)$



Relative error 0.0160

6-bits of precision

# Divide: IEEE wants the correctly rounded result

Possible with extra precision in the last iteration

Possible with extra careful coding with only FMA

A / B: Iterate for  $(1/B)$  and multiply by A before the last Newton iteration

# Perfectly rounded results for $f(x)$

- “The Table Maker’s Dilemma”

- Round

1.11010 10000 00000 00000 00000 00000 00000 ....

To the nearest 6 bit fraction.

1.11010 or 1.11011 ?

# The elementary functions

- #include <math.h>

sqrt

exponential, logarithm

trigonometry, hyperbolic trig, inverse trig

erf

$x^y$

$$e^x = e^{R+n(\ln 2 \div 16)} = e^R 2^{(m+j/16)} = e^R \text{tab}[j] 2^m$$

Use Taylor series for  $\exp(R)$ .

No polynomial can grow exponentially. But  $2^m$  does.

( $\exp$  is an easy case. Why?)

# tanh(x) is harder

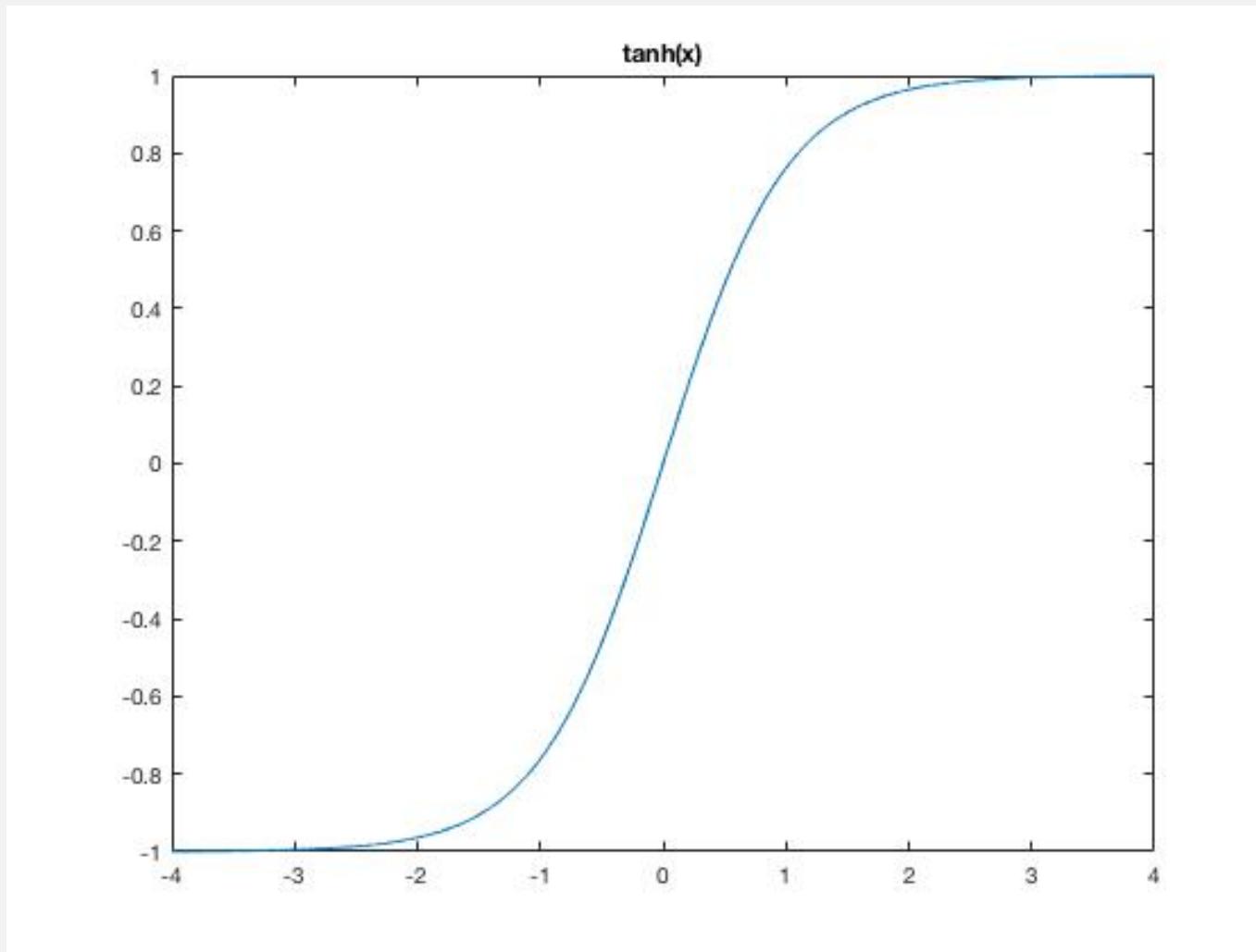
$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

$$\tanh\left(i \frac{\pi}{2}\right) = \frac{i - (-i)}{i + (-i)} = \frac{2i}{0}$$

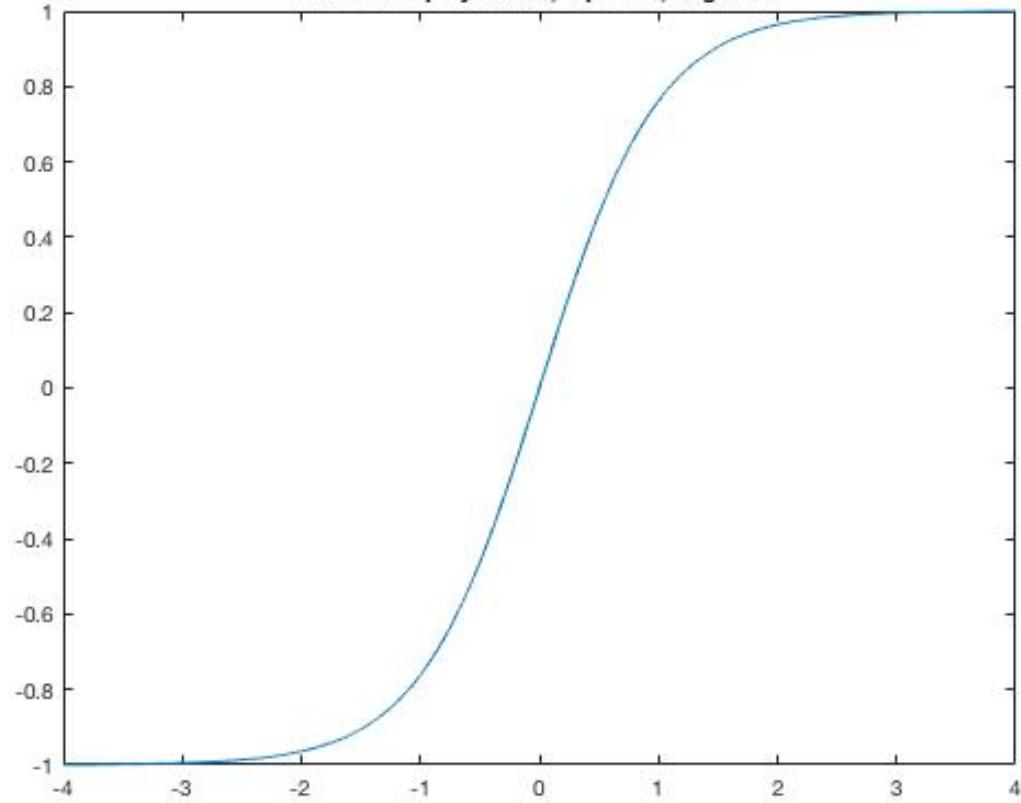
A pole near the real axis makes tanh difficult to approximate using polynomials.

Why do we care about polynomials?

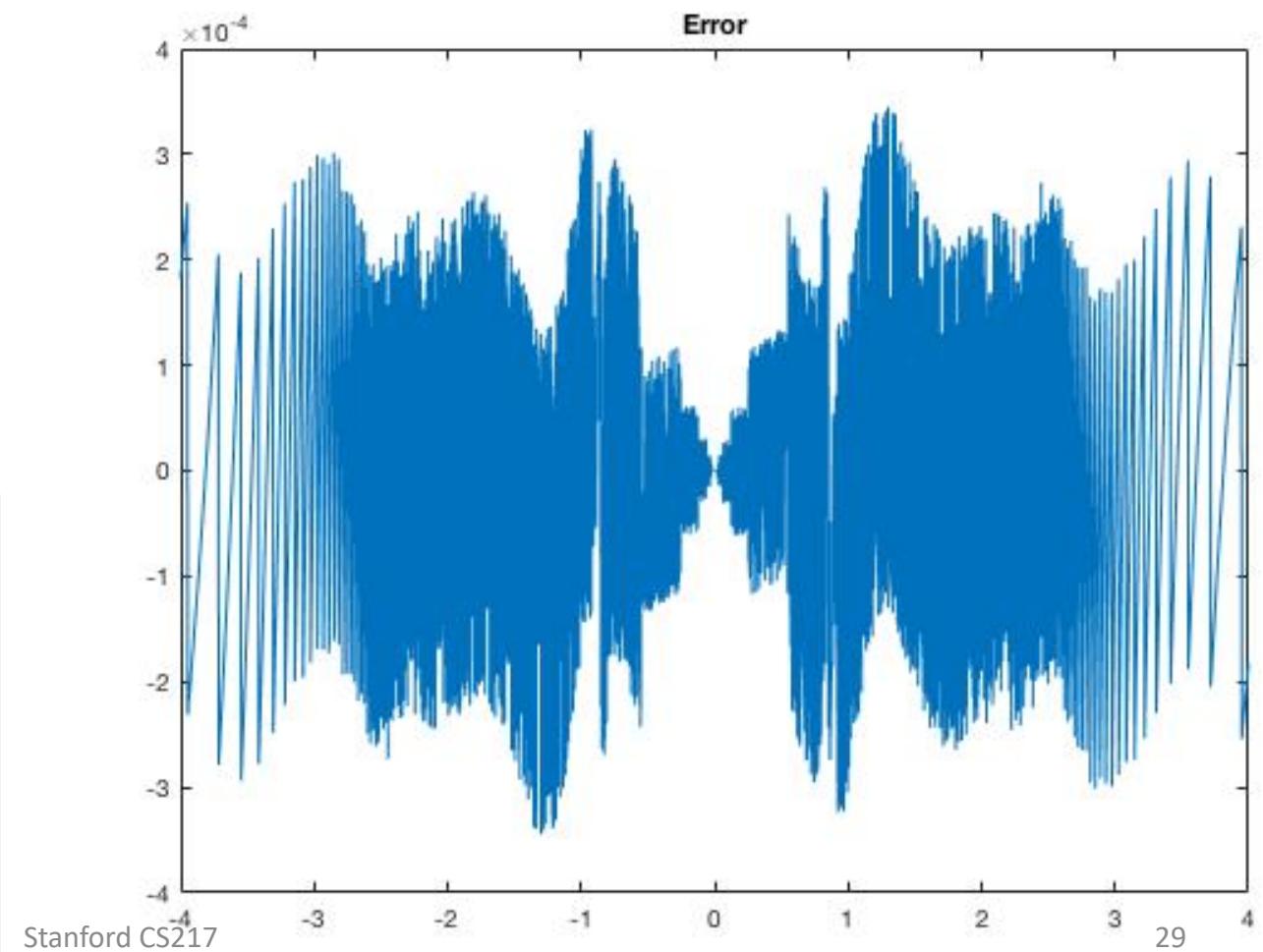
# No polynomial has asymptotes!



Piecewise polynomial, 4 pieces, degree 4



Error



# Trig functions, periodicity, range reduction

$$\sin(x) = \sin(2n\pi + \phi), \quad n = 1, 2, \dots$$
$$\phi := x - 2n\pi$$

Ping Tak Peter Tang. **Table-driven implementation of the exponential function in IEEE floating-point arithmetic.** [ACM Trans. Math. Softw. 15\(2\)](#): 144-157 (1989)

Ping Tak Peter Tang. **Table-driven implementation of the logarithm function in IEEE floating-point arithmetic.** [ACM Trans. Math. Softw. 16\(4\)](#): 378-400 (1990)

Jean-Michel Muller. **Elementary Functions: Algorithms and Implementation.** Birkhauser, 2006.

William J Cody Jr., William Waite. **Software Manual for the Elementary Functions.** Prentice Hall, 1980.

Lloyd N. Trefethen. **Approximation Theory and Approximation Practice.** SIAM, 2013.

# Summary

- The floating point numbers are a finite set that cover (well) a finite range of magnitudes
- Everything is an approximation
- The relative errors start small
- Errors propagate; whether they grow or not requires analysis
- Large numbers imply small relative, but large absolute errors



Cerebras Systems is a stealth mode startup backed by premier venture capitalists and the industry's most successful technologists. We are entrepreneurs dedicated to solving hard problems. We value integrity, passion, problem solving ability, and a sense of humor, and are always looking for extraordinary people to join our team.

---

**CONTACT**

**HELP WANTED**