CE 311K: Errors

Krishna Kumar

University of Texas at Austin

krishnak@utexas.edu

September 22, 2019

1 Errors

2 Bit representation

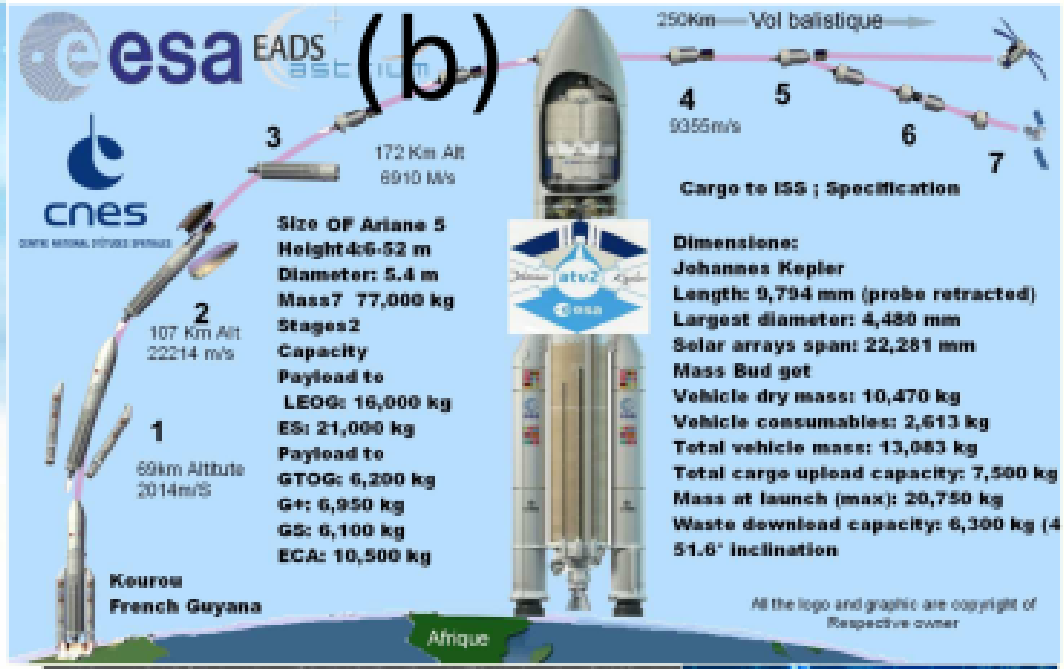
3 Numerical errors

What causes errors?

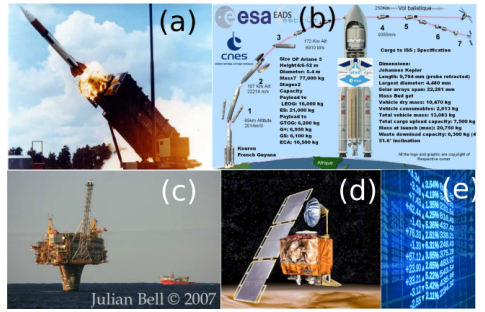
"⚠ *the computer calculated it, so it must be right*"

- the wrong mathematical model of reality (most subject areas lack models as precise and well-understood as Newtonian gravity)
- a parameterised model with parameters chosen to fit expected results ('over-fitting')
- the model being very sensitive to input or parameter values
- the discretisation of the continuous model for computation
- the build-up or propagation of inaccuracies caused by the finite precision of floating-point numbers
- plain old programming errors

The cost of errors

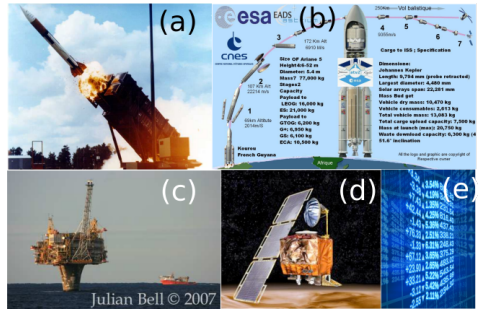


? The cost of errors



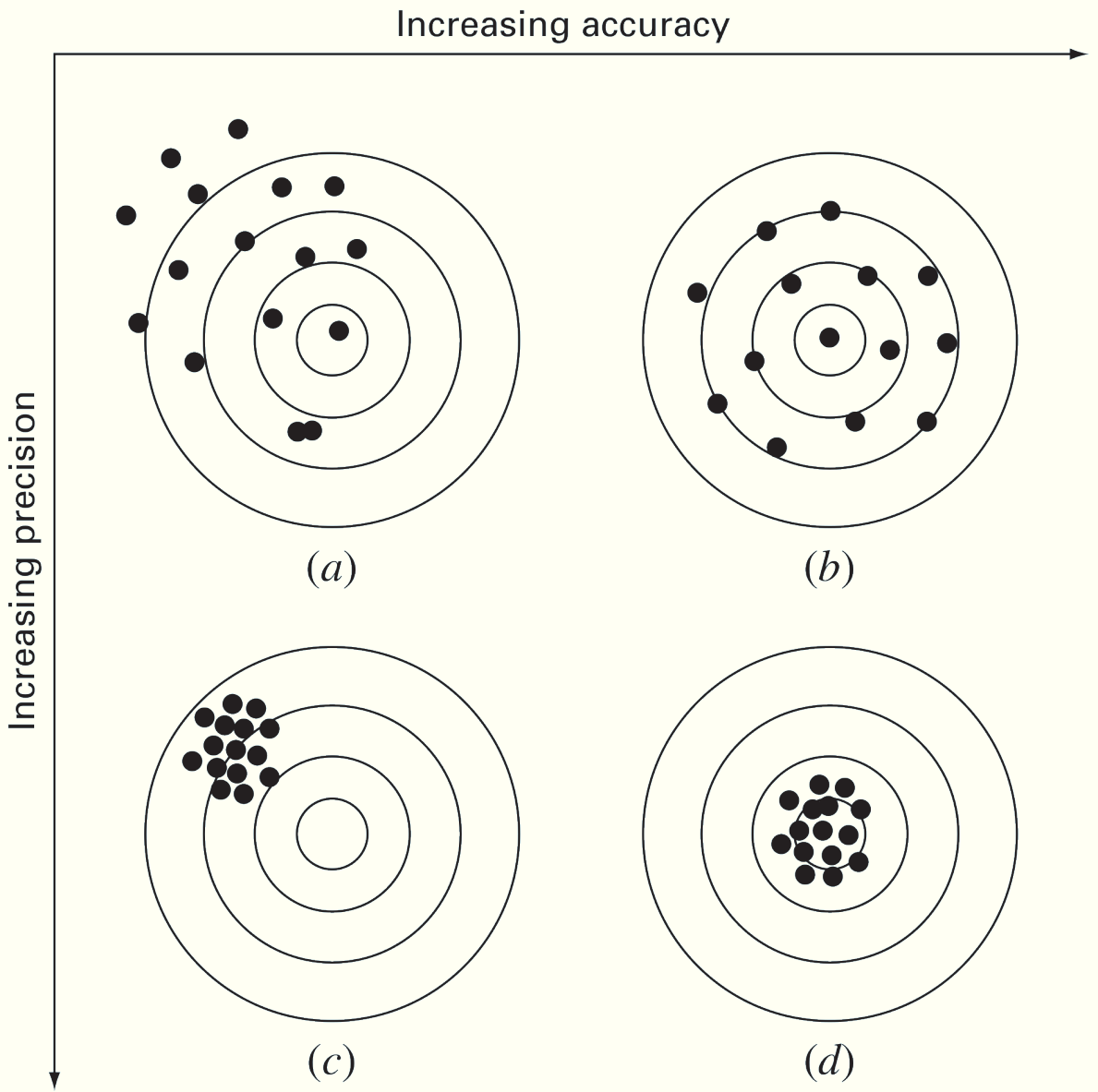
1. In 1991, a US Patriot missile failed to intercept an Iraqi Scud missile at Dhahran in Saudi Arabi, leading to a loss of life.
2. 1996 to a European Space Agency Ariane 5 unmanned rocket exploding shortly after lift-off. The rocket payload, worth US$500 Million, was destroyed.
3. Sleipner A offshore platform sprang leak and sank on 23 August 1991. Post-accident investigation traced the error to inaccurate finite element approximation of the linear elastic model of the tricell (using the popular FDTD simulator NASTRAN). The shear stresses were underestimated by 47%. he failure involved a total economic loss of about $700 million and causing a Richter scale 3.0 earthquake.

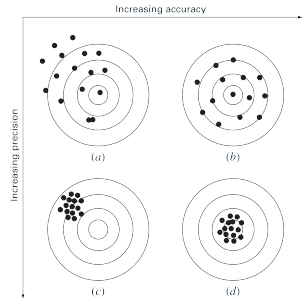
The cost of errors



1. In 1998 NASA lost its $125 million Mars Climate Orbiter spacecraft as a result of a mistake that would shame a first-year physics student failing to convert Imperial units to metric. "*Propulsion people talk in pound-seconds of thrust and navigators talk in newton-seconds*"
2. Knight Capital Group lost $440 million in 30 minutes on Aug 1, 2012 due to bug in their trading algorithms reportedly started pushing erratic trades through on nearly 150 different stocks

Quantifying errors: accuracy v precision



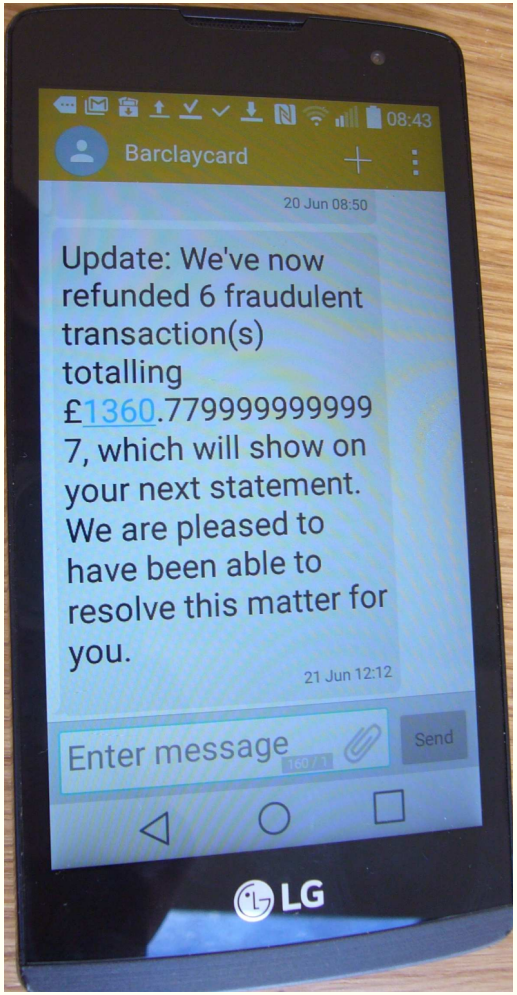


The errors associated with both calculations and measurements can be characterized with regard to their accuracy and precision.

Accuracy refers to how closely a computed or measured value agrees with the true value. Inaccuracy (also called bias) is defined as systematic deviation from the truth.

Precision refers to how closely individual computed or measured values agree with each other. Imprecision (also called uncertainty), on the other hand, refers to the magnitude of the scatter.

Significant figures

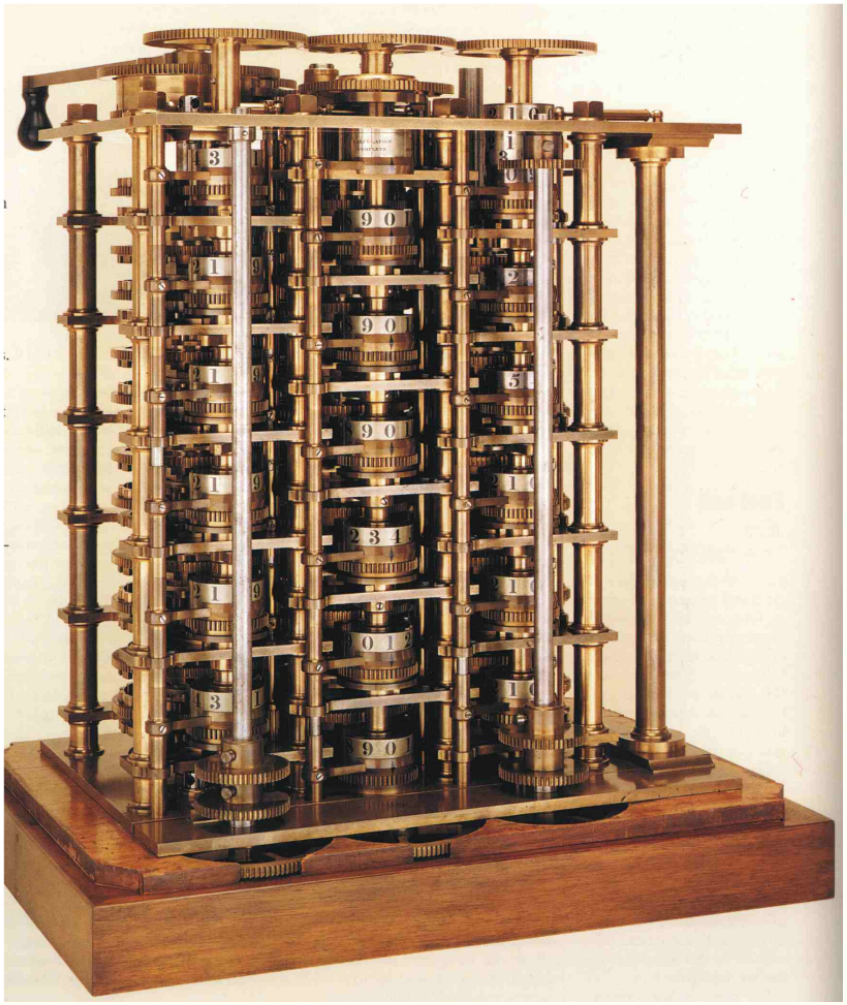


1 Errors

2 Bit representation

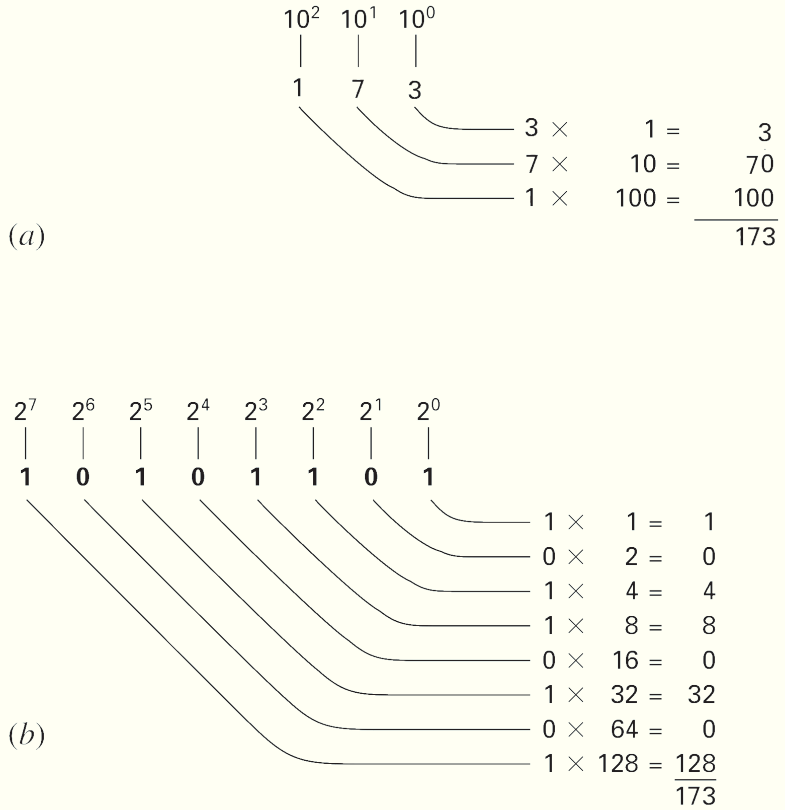
3 Numerical errors

Decimal or binary



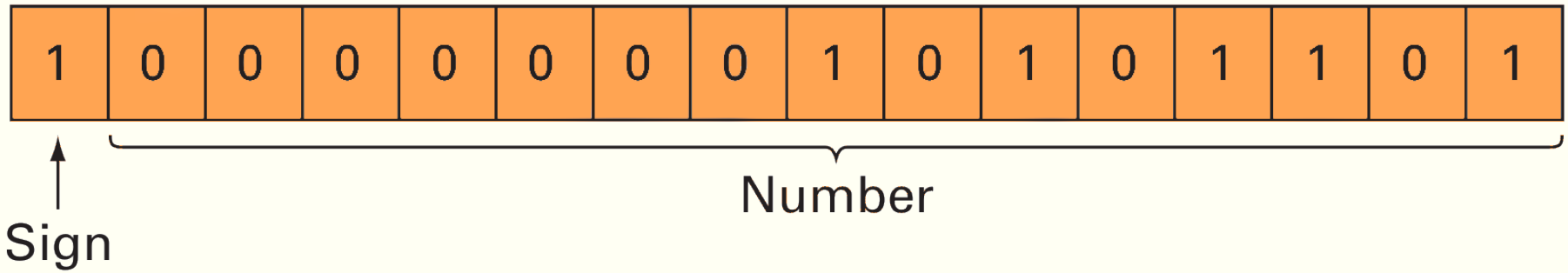
Charles Babbage's machine used base 10. CompScis decided a little later that base 2 is more funky.

Decimal or binary



Representing 173 in Decimal and Binary system

Bits and Bytes



Representing -173 on a 16-bit computer using the signed magnitude method.

Bit - **Bi**nary Digi**t**. 8 bits is a byte(?) ASCII needed 7 bits to represent all English alphabets.

The Gangam Style problem

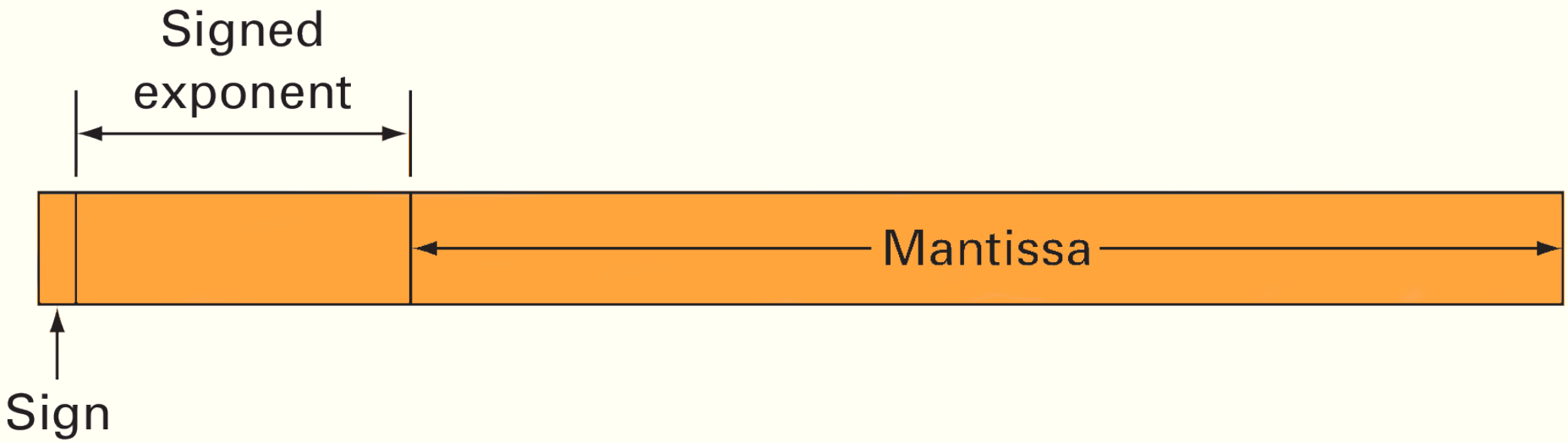
When 2,147,483,647 views of Gangam Style broke YouTube

Floating point representation

Fractional quantities are typically represented in computers using floating-point form. In this approach, the number is expressed as a fractional part, called a *mantissa* or *significand*, and an integer part, called an *exponent* or *characteristic*,

$$m.b^e$$

where m is the mantissa, b is the base of the number system being used, and e the exponent. For instance, the number 156.78 could be represented as 0.15678×10^3 in a floating-point base-10 system.

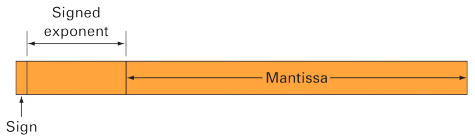


Floating point representation

Fractional quantities are typically represented in computers using floating-point form. In this approach, the number is expressed as a fractional part, called a *mantissa* or *significand*, and an integer part, called an *exponent* or *characteristic*,

$$m.b^e$$

where m is the mantissa, b is the base of the number system being used, and e the exponent. For instance, the number 156.78 could be represented as 0.15678×10^3 in a floating-point base-10 system.



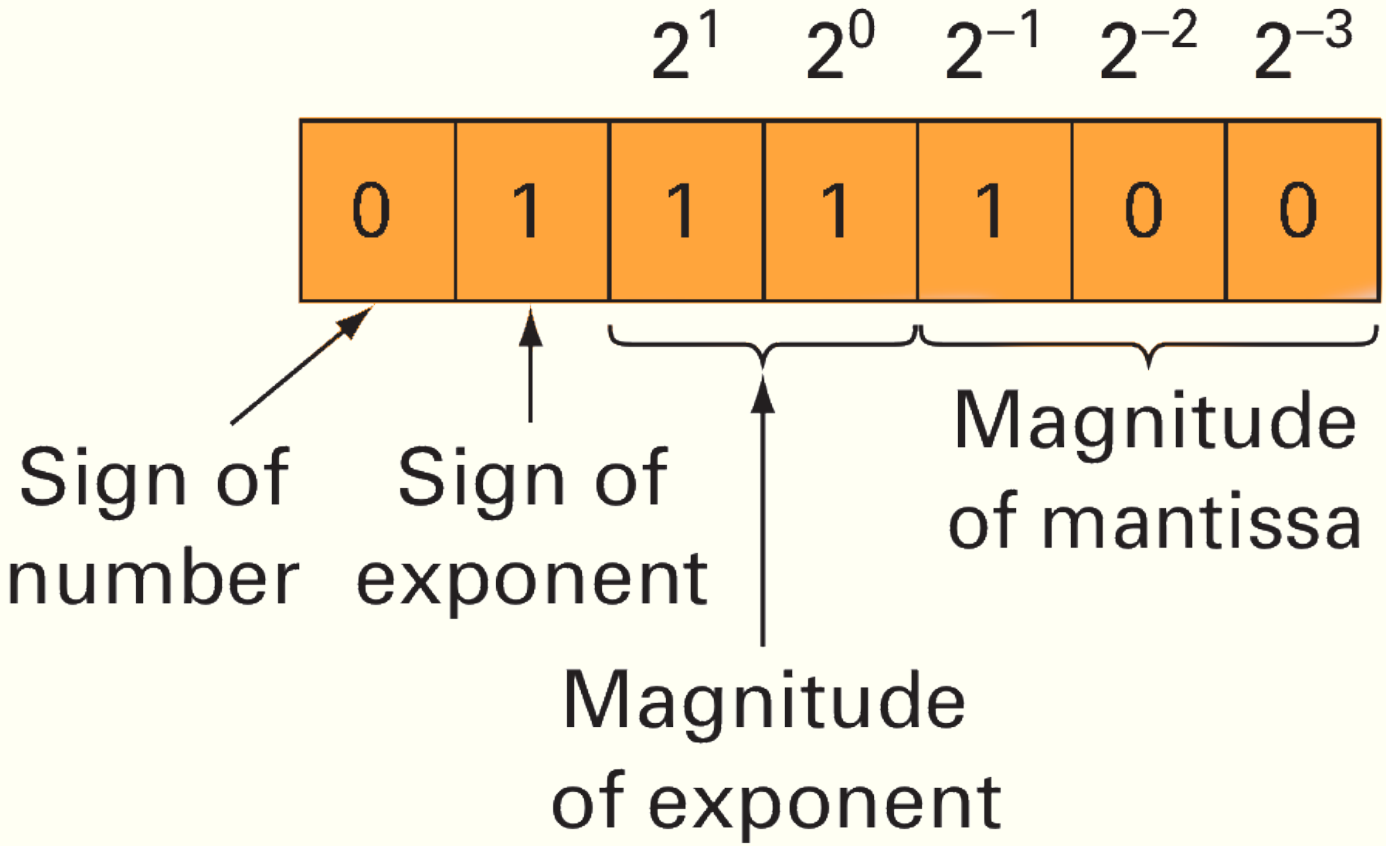
The absolute value of m is limited:

$$\frac{1}{b} \leq m < 1$$

where b the base. For example, for a base-10 system, m would range between 0.1 and 1, and for a base-2 system, between 0.5 and 1.

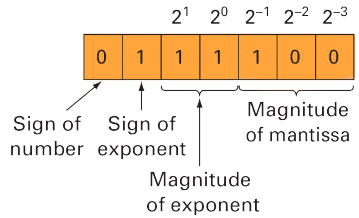
Smallest floating point for a 7-bit representation

Employ the first bit for the sign of the number, the next three for the sign and the magnitude of the exponent, and the last three for the magnitude of the mantissa:



Smallest floating point for a 7-bit representation

Employ the first bit for the sign of the number, the next three for the sign and the magnitude of the exponent, and the last three for the magnitude of the mantissa:



The initial 0 indicates that the quantity is positive. The 1 in the second place designates that the exponent has a negative sign. The 1's in the third and fourth places give a maximum value to the exponent of $1\times2^1+1\times2^0 = -3$.

Finally, the mantissa is specified by the 100 in the last three places, which conforms to: $1\times2^1+0\times2^{-2}+0\times2^{-3}=0.5$. Thus, the smallest possible positive number for this system is $+0.5\times2^-3$, which is equal to 0.0625 in the base-10 system.

1 Errors

2 Bit representation

3 Numerical errors

Numerical errors

Numerical errors arise from the use of approximations to represent exact mathematical operations and quantities.

- **round-off errors**, which result when numbers having limited significant figures are used to represent exact numbers
- **truncation errors**, which result when approximations are used to represent exact mathematical procedure.

⑦ Quantifying errors

Suppose that you have the task of measuring the lengths of a bridge and a rivet and come up with 9999 and 9 cm, respectively. If the true values are 10,000 and 10 cm, respectively. Quantify the errors.

Quantifying errors

1. The absolute error for measuring the bridge is:

$$\Delta_{bridge} = 10000 - 9999 = 1\text{cm},$$

and the rivet is:

$$\Delta_{rivet} = 10 - 9 = 1\text{cm}.$$

2. The relative error for the bridge is:

$$\delta_{bridge} = \frac{10000 - 9999}{1000} = 0.01\%,$$

and the rivet is:

$$\delta_{rivet} = \frac{10 - 9}{10} = 10\%.$$

Although both measurements have an error of 1 cm, the relative error for the rivet is much greater.

Absolute and relative errors

Let x be a real number. We will use x^* to denote its approximation. We define two ways of measuring error introduced by this approximation.

- Absolute error:
$$\varepsilon_x = \Delta_x = \|x^* - x\|$$
- Relative error:
$$\eta_x = \delta_x = \frac{\|x^* - x\|}{\|x\|}$$

The relationships (WARNING: Severe abuse of notation!)
$x^* = x \pm \varepsilon_x$ and $x^* = x(1 \pm \eta_x)$ are also commonly used to mean that x^* may take any value in the interval $[x - \varepsilon_x, x + \varepsilon_x]$.

Error accumulation

Let

$$x^* = x \pm \varepsilon_x \quad \text{and} \quad y^* = y \pm \varepsilon_y$$

Adding these two yields:

$$\begin{aligned} x^* + y^* &= (x \pm \varepsilon_x) + (y \pm \varepsilon_y) \\ &= x + y \pm \varepsilon_x \pm \varepsilon_y \\ &= x + y \pm (\varepsilon_x + \varepsilon_y) \end{aligned}$$

$$\varepsilon_{x+y} = \varepsilon_x + \varepsilon_y$$

Exercise: What about subtraction?

Error accumulation: Loss of significance

Beware: when addition or subtraction causes partial or total cancellation, the relative error of the result can be much larger than that of the operands. We call this **loss of significance**.
For example, consider we store values to 3 significant digits and we take the innocent-looking x = 9.99, y = 9.98.

- 3 significant figures means x and y are accurate to ± 0.005 absolute error. x and y thus each have a relative error of about 0.0005(0.05%), i.e. **very good**.
- However, $x - y = 0.01$, and has an absolute error of 0.01 (recall previous slide on subtraction) hence a **relative error of 100%**! We have little idea what the true value of $x - y$ is at this point.

Error accumulation

This gets even worse when the loss of significance happens in a fraction's denominator. Consider an extension of the previous example:

$$\frac{1}{x-y}$$

$$\frac{1}{x-y} = \frac{1}{0.01 \pm 0.01}$$

⚠ This can be anywhere between 50 and infinity!