Algorithms and Data Structures

Encoding
Peculiarities of machine arithmetic

$$a + b = b + a$$

$$a + (b + c) = (a + b) + c$$

$$a \cdot (b + c) = a \cdot b + a \cdot c$$

Learning goals

- Associative and distributive properties
- Order of addition
- Calculation of variance

PECULIARITIES OF MACHINE ARITHMETIC

Common arithmetic properties are no longer fulfilled.

For simplicity, we use decimal representation with $m = 4$ and rounding.

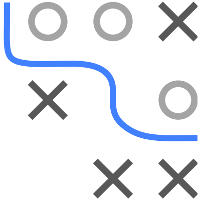
- **Associative property:**

$$a = 4, b = 5003, c = 5000 \quad \Rightarrow$$
$$a = 0.4 \cdot 10^1, b = 0.5003 \cdot 10^4, c = 0.5 \cdot 10^4$$

$$\begin{aligned}
(\tilde{a} + \tilde{b}) &= 0.4 \cdot 10^1 + 0.5003 \cdot 10^4 = 0.5007 \cdot 10^4 \\
(\tilde{a} + \tilde{b}) + \tilde{c} &= 0.5007 \cdot 10^4 + 0.5 \cdot 10^4 = 1.0007 \cdot 10^4 \\
&\approx 0.1001 \cdot 10^5 = 10010
\end{aligned}$$

$$\begin{aligned}
(\tilde{b} + \tilde{c}) &= 0.5003 \cdot 10^4 + 0.5 \cdot 10^4 = 1.0003 \cdot 10^4 \\
&\approx 0.1000 \cdot 10^5 \\
(\tilde{b} + \tilde{c}) + \tilde{a} &= 0.1000 \cdot 10^5 + 0.4 \cdot 10^1 = 0.10004 \cdot 10^5 \\
&\approx 0.1000 \cdot 10^5 = 10000
\end{aligned}$$

PECULIARITIES OF MACHINE ARITHMETIC / 2



- **Distributive property:**

$$2 \cdot (\tilde{b} - \tilde{c}) = 2 \cdot (0.5003 \cdot 10^4 - 0.5 \cdot 10^4) = 0.0006 \cdot 10^4 = 6$$

$$(2 \cdot \tilde{b} - 2 \cdot \tilde{c}) = 2 \cdot 0.5003 \cdot 10^4 - 2 \cdot 0.5 \cdot 10^4 = 1.0006 \cdot 10^4 - 1 \cdot 10^4 \approx 0.1001 \cdot 10^5 - 0.1 \cdot 10^5 = 0.0001 \cdot 10^5 = 10$$

Problem in the second example: catastrophic cancellation.

EXAMPLES

```
1e16 - 1e16
## [1] 0
```

```
(1e16 + 1) - 1e16
## [1] 0
```

```
(1e16 + 2) - 1e16
## [1] 2
```

$1e16$ cannot be represented exactly since it is larger than 2^{53}, hence the distance is greater than 1.

EXAMPLES

```
x = seq(1, 2e16, length = 100000)
s1 = sum(x)
s2 = sum(rev(x))
s1
## [1] 1e+21
```

```
s2
## [1] 1e+21
## [1] 1e+21
```

```
s1 - s2
## [1] -262144
```

ORDER OF ADDITION

General recommendation: Start with numbers having the smallest absolute values.

Assuming $0 \leq a_1 \leq a_2 \cdots \leq a_n$, there are still various ways to perform the summation, e.g.:

- $(((a_1 + a_2) + a_3) + a_4) + a_5$
- $((a_1 + a_2) + (a_3 + a_4)) + a_5$
- $((a_1 + a_2) + a_3) + (a_4 + a_5)$

Remark: Particularly bad errors can occur when calculating differences of numbers on computers (this will be discussed in another lecture).

CALCULATION OF VARIANCES

Sample: $x_1 = 356, x_2 = 357, x_3 = 358, x_4 = 359, x_5 = 360$

$$4S^2 = \sum_{i=1}^{5}(x_i - \bar{x})^2 = \sum_{i=1}^{5} x_i^2 - 5(\bar{x})^2 = 10$$

Not like that in decimal machine arithmetic with $m = 4$:

First formula OK, but second one is a disaster:

$$\tilde{x}_1^2 = .1267E6,\ \tilde{x}_2^2 = .1274E6,\ \tilde{x}_3^2 = .1282E6,$$
$$\tilde{x}_4^2 = .1289E6,\ \tilde{x}_5^2 = .1296E6,$$
$$\sum \tilde{x}_i^2 = .6408E6 \qquad 5 \cdot (\tilde{\bar{x}})^2 = 5 \cdot .1282E6 = .6410E6$$

The second formula gives a negative empirical variance!

CALCULATION OF VARIANCES / 2

Three approaches to calculate the $1/n$ normalized standard deviation of a sample:

```
sd1 = function(x) {
  s2 = mean((x - mean(x))^2)
  sqrt(s2)
}
```

```
sd2 = function(x) {
  s2 = mean(x^2) - mean(x)^2
  sqrt(s2)
}
```

```
sd3 = function(x) {
  n = length(x)
  s2 = ((n - 1) / n) * var(x)
  sqrt(s2)
}
```

CALCULATION OF VARIANCES / 3

```
options("digits" = 20)
sd1(1:9)
## [1] 2.5819888974716112
```

```
sd2(1:9)
## [1] 2.5819888974716116
```

```
sd3(1:9)
## [1] 2.5819888974716112
```

CALCULATION OF VARIANCES / 4

Algorithm Calculation of variance in R (simplified)

Input: $x \in \mathbb{R}^n$
$s1 = s2 = 0;$
for i = 1, ..., n **do**
    $s1 = s1 + x[i]$
end for
xm $\leftarrow \frac{s1}{n}$
for i = 1, ..., n **do**
    $s2 = s2 + (x[i] - xm) * (x[i] - xm)$
end for
return $\frac{s2}{n-1}$

▸ Gupta 2024