MATH50003 (2022–23)

Problem Sheet 1

This problem sheet tests the representation of numbers on the computer, using modular and floating point arithmetic.

Problem 1 With 8-bit signed integers, what are the bits for the following: $10, 120, -10$.

Problem 2 What is π to 5 binary places? Hint: recall that $\pi \approx 3.14$.

Problem 3 What are the single precision F_{32} (`Float32`) floating point representations for the following:

$$2, 31, 32, 23/4, (23/4) \times 2^{100}$$

Problem 4 Let $m(y) = \min\{x \in F_{32} : x > y\}$ be the smallest single precision number greater than y. What is $m(2) - 2$ and $m(1024) - 1024$?

Problem 5 Suppose $x = 1.25$ and consider 16-bit floating point arithmetic (F_{16}). What is the error in approximating x by the nearest float point number $\mathrm{fl}(x)$? What is the error in approximating $2x$, $x/2$, $x + 2$ and $x - 2$ by $2 \otimes x$, $x \oslash 2$, $x \oplus 2$ and $x \ominus 2$?

Problem 6 For what floating point numbers is $x \oslash 2 \neq x/2$ and $x \oplus 2 \neq x + 2$?

Problem 7 What are the exact bits for $1 \oslash 5$, $1 \oslash 5 \oplus 1$ computed using half-precision arithmetic (`Float16`) (using default rounding)?

Problem 8 Explain why the following does not return `1`. Can you compute the bits explicitly?

In [1]:
```
Float16(0.1) / (Float16(1.1) - 1)
```

Out[1]:
```
Float16(1.004)
```

Problem 9 Find a bound on the *absolute error* in terms of a constant times machine epsilon $\epsilon_{\rm m}$ for the following computations

$$\begin{gathered}(1.1 * 1.2) + 1.3 \\ (1.1 - 1)/0.1\end{gathered}$$

implemented using floating point arithmetic (with any precision). That is, each number is rounded first using fl and each operation is replaced by its floating point analogues $\oplus, \otimes, \ominus, \oslash$.