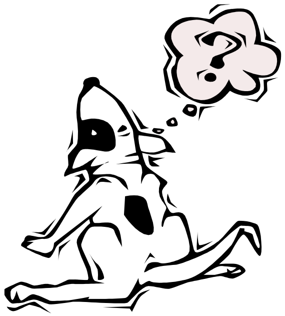
Lecture 3 Floating Point Representations

Recitation

In comparison

- The smallest and largest possible 32-bit integers in two's complement are only -2^{31} and $2^{31} - 1$
- How can we represent so many more values in the IEEE 754 format, even though we use the same number of bits as regular integers?



what's the next representable FP number?

$+2^{-126}(1+2^{-23})$ **differ from the smallest number by 2^{-149}**

How is this derived?

Exercise:

- Convert 17.5 into IEEE 754 single precision format

$17.5 = 10001.1_2 = 1.00011 \times 2^4$

S = 0

e = 4 + 127 = 131 = 1000 0011

f = 000110…0

The final format: 0 1000 0011 000110…0

Exercise:

- Compare the following single-precision IEEE 754 FP numbers:

0 1101 0110 101000111100100..0	0 1101 1110 10…0
0 0111 1111 010..0	1 0111 1111 010..0
0 1111 0111 11100001101010..0	0 1111 0110 000…0
1 0000 0010 000…0	1 0000 0001 000…0
1 1010 0101 110…0	1 1010 0101 1010…0

Exercise:

- Compare the following single-precision IEEE 754 FP numbers:

0 1101 0110 101000111100100..0	<	0 1101 1110 10…0
0 0111 1111 010..0	>	1 0111 1111 010..0
0 1111 0111 11100001101010..0	>	0 1111 0110 000…0
1 0000 0010 000…0	<	1 0000 0001 000…0
1 1010 0101 110…0	<	1 1010 0101 1010…0

Quiz 2

1. Convert the following FP number into IEEE 754 single-precision format

   -21.75

2. Compare the following pairs of single-precision IEEE 754 FP numbers

0 1011 0100 0010… 0	0 1011 0100 010…0
0 0000 0011 110…0	0 0000 0001 110…0
1 1111 0010 1010…0	1 1111 0001 1110…0