Many Formats for representation of binary #5:

- one's complement
- sign & mag.
- two's complement → most commonly used.

Real # in two's complement with \( \infty \) precision:

\[
x = X_m \left(-b_0 + \sum_{i=1}^{\infty} b_i 2^{-i}\right)
\]

\( X_m = \) arbitrary scale factor \( \left|X\right| < X_m \)

\( b_i = \) either zero or one
\[ b_0 = \text{sign bit} \rightarrow \begin{cases} b_0 = 0 & 0 \leq x \leq x_m \\ b_0 = 1 & -x_m \leq x < 0 \end{cases} \]

With finite \( B + 1 \) bits, we get the representation:

\[ x_B = Q_B \lceil x \rceil = x_m \left( -b_0 + \sum_{i=1}^{B} b_i 2^{-i} \right) \]

\( x \) is the quantized version of \( x \).

- Smallest difference between any 2 \( x \) in the quantized domain \( \Delta = x_m \left( 2 \right) \)

5.87923, only use 2 bits 5.87924
- quantized #s are in the range
- $X_m \leq x < X_m$

$X_B = b_0 \uparrow b_1 b_2 b_3 \ldots b_8$

Binary point.

$5.3924 \rightarrow \begin{cases} 5.4 \text{ rounding} \\ 5.3 \text{ truncation} \end{cases}$
Start a real number $x$

to get $X_B$, one can either round or truncate:

show Figs 6.37(a), 6.37(b) in O&X.

Quantization error

\[ e = Q_B[x] - x \]
\[ = X_B - x \]

2's complement:

Rounding error $-\frac{1}{2} < e < +\frac{1}{2}$

Truncation error $0 \leq e \leq a$
$|x_1| < 5000$

$|x_2| < 5000$

Overflow

$y = x_1 + x_2$

$|y| < 10000$

Saturate

Natural overflow

6.38

0.85
Interesting property of two complex numbers:

- natural overflow.

Add few #s, if the final sum doesn't overflow, then result is correct even though the intermediate results overflow.

Trade-off between overflow & rounding error.

\[ X_m \uparrow \rightarrow \text{overflow is less likely but } \Delta \uparrow, e \uparrow \]
\[ x_m \downarrow \rightarrow \text{overflow is more likely.} \]

but \[ A \uparrow \downarrow \]

\underline{But \ B \ can \ come \ to \ zero.}

- Keep \( x_m \) large to minimize chance of overflow
- But keep \( B \) large to keep \( D, e \) small.

Multiplication also introduces \( \text{overflow} \) and \( \text{rounding} \).