show Fig 6.39, 6.40, 6.41 O&S.

Binary Representation of #s.

Many Formats for representation of binary #s.

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

- Real # in two's complement with ∞ precision.

$$X = X_m \left(-b_0 + \sum_{i=1}^{\infty} b_i \, 2^{-i} \right)$$

X_m = arbitrary scale factor $|X| < X_m$

b_i = either zero or 1

b_0 = sign bit $\longrightarrow \begin{cases} b_0 = 0 & 0 \leq x \leq X_m \\ b_0 = 1 & -X_m \leq x < 0 \end{cases}$

---

with finite # of bits $(B+1)$ we get representation:

$$\hat{X}_B = Q_B[x] = X_m \left(-b_0 + \sum_{i=1}^{B} b_i 2^{-i}\right)$$

$\hat{x}$: is quantized version of x.

smallest difference between any 2 #s in quantized domain $\Delta = X_m 2^{-B}$

5.87923 , only use 2 bits 5.87924

- quantized #s are in the rang.

$$-X_m \le \hat{x} < X_m$$

$$\hat{X}_B = b_0 \, \square \, b_1 \, b_2 \, b_3 \ldots\ldots b_B$$

binary point.

$$5.3924 \longrightarrow \begin{cases} 5.4 & \text{rounding} \\ 5.3 & \text{Truncation} \end{cases}$$

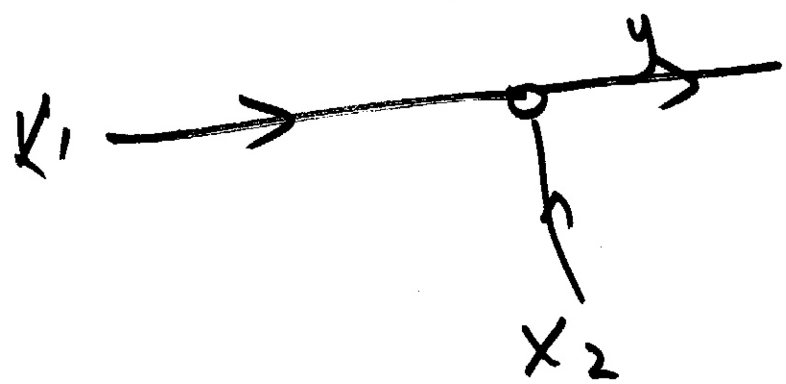
Start a real number x
to get $\hat{x}_B$, one can either
rounding or Truncation:

Show Fig 6.37 (a), 6.37 (b) in O&S.

Quantization error

$$e = Q_B[x] - x$$
$$= \hat{x}_B - x$$

2's complement: rounding error $-\Delta/2 < e < +\Delta/2$
Truncation error $0 \leq e \leq \Delta$



$|x_1 + x_2| < 10{,}000$

$y = x_1 + x_2$

$|x_1| < 5000$

$|x_2| < 5000$

overflow.

Saturate

natural overflow

6.38

O&S

- Interesting property of Two's Complement + natural overflow:

  Add few #s, if The final sum doesn't overflow, Then result is Correct even though The intermediate results overflow.

- Tradeoffs between overflow & rounding error.

  $X_m \uparrow$ $\longrightarrow$ overflow is less likely but $\Delta \uparrow$, $e \uparrow$

$X_m \downarrow \longrightarrow$ overflow is more likely.

but $\Delta \downarrow$ $e \downarrow$

But B can come to rescue.

- Keep X_m large to minimize chances of overflow
- But keep B large to keep Δ, e small.

Multiplication also introduces $\begin{cases} \text{- overflow} \\ \text{- rounding error} \end{cases}$