CA 3.5 1.3. Representation of fixed-point decimal nos. (1)

1.3.1 Binary Coded Decimal 8421: represent a digit on a tetrad

Ex: 297 = 0010 1001 0111 BCD

$i=2$ $i=1$ $i=0$

tetrad = nibble = 4 bits

weight $10^i \cdot 2^j$ — i = ith digit, j = rank in tetrad

$10^1 \cdot 2^2$

fixed weight ⇒ POSITIONAL

1.3.2. Excess of Three (E3) — excess == bias = amount that is added/subtr. from a code.

- represent digit on a tetrad.
  - E3 adds 3 units to the corresponding BCD digit ⇒ E3 is not POSITIONAL

! - facilitates addition.

Ex: 297 = 0101 1100 1010

1.3.3. Two-out-of-five

- represents a digit on 5 bits — 2 bits are 1s, 3 bits are 0s

$C_5^2 = \frac{5!}{3! \cdot 2!} = \frac{4 \cdot 5}{2} = 10$

Ex: 297 00101 10100 10001

! - error detection using parity codes

1.4. Representing floating-point (f.p.) numbers

$$X = X_M * B^{X_E}$$

X_M = mantissa = fixed-point, fractional → SM / C2

X_E = exponent = fixed-point, integer → SM / C2

B = radix: 2, power of 2

0.237×2^{17}, X_M^*

1.4.1 General considerations

sign	exponent	fractional part of mantissa
S	X_E	X_M^*
<1>	<e>	<m>

$X_M = S.X_M^*$ - sign appended X_M^*

→ mantissa.

Constraints for representing f.p. nos.

$X = X_M * 2^{X_E}$

Ⓐ Representing 0

X_E for 0 should be the smallest possible value
- fast comparison against 0
- roundings errors

X_E →
- SM: smallest value $-2^{e-1}+1$: $\overset{\langle e \rangle}{1\,1 \ldots 1}$
- C2: —"— -2^{e-1} : $1\,0 \ldots 0$

fixed-point integer

X_M for 0 — all-0s representation

X_M for 0 : $00 \ldots 00$

Putting all fields together.

0 →
- SM: $\underset{\langle 1 \rangle}{S} \mid \underset{\langle e \rangle}{1\,1 \ldots 1} \mid \underset{\langle m \rangle}{00 \ldots 00}$
- C2: $S \mid 1\,0 \ldots 0 \mid 00 \ldots 00$

fast 0 comparison → require the X_E to be all-0s

Q: how to obtain all-0s config. for X_E of 0.

A: represent X_E in a bias code. (or excess code)

bias code.
- SM: $+2^{e-1}-1$
- C2: $+2^{e-1}$

$\Rightarrow 0$: $\underset{\langle 1 \rangle}{S} \mid \underset{\langle e \rangle}{00 \ldots 00} \mid \underset{\langle m \rangle}{00 \ldots 00}$

Ⓑ Represent mantissas.

- redundant ! in decimal $0.237 = 0.0237 \times 10^1$
$= 0.000237 \times 10^3$
$= \ldots$

normalized mantissa = unique representation

- msb of the fractional part of normalized mantissa is
  - SM: $\underline{1}$
  - C2: $\overline{S}$

Ex: $0.\underbrace{1}_{msb=1}11_{SM} = \frac{7}{2^3}$ $\quad$ $1.\underbrace{1}_{msb=1}11_{SM} = \frac{-7}{2^3}$

$\underbrace{0}_{S}.\underbrace{1}_{msb=\bar{S}}11_{C2} = \frac{7}{2^3}$ $\quad$ $\underbrace{1}_{S}.\underbrace{0}_{msb=\bar{S}}01_{C2}$

Range of values for X_M

$$\frac{1}{2} \le |X_M| < 1$$

$X_M^* = 0.\cancel{0}1011 * 2^{X_E}$

normalise X_M^*:

$X_M^* = 0.\underline{1}011 * 2^{X_E-1}$

1.4.2. IEEE 754

portable formats → 32 bits (single precision)
→ 64 bits (double precision)
→ 128 bits (quadruple precision)

interchange formats → 16 bit (half-precision) → NN

Simple precision format

→ fractional part of significand.

S	X_E	X_S^*
<1>	<8>	<23>

$$X_S = \underbrace{1}_{\text{hidden bit, integer bit}}.X_S^*$$

→ fractional part of significand.

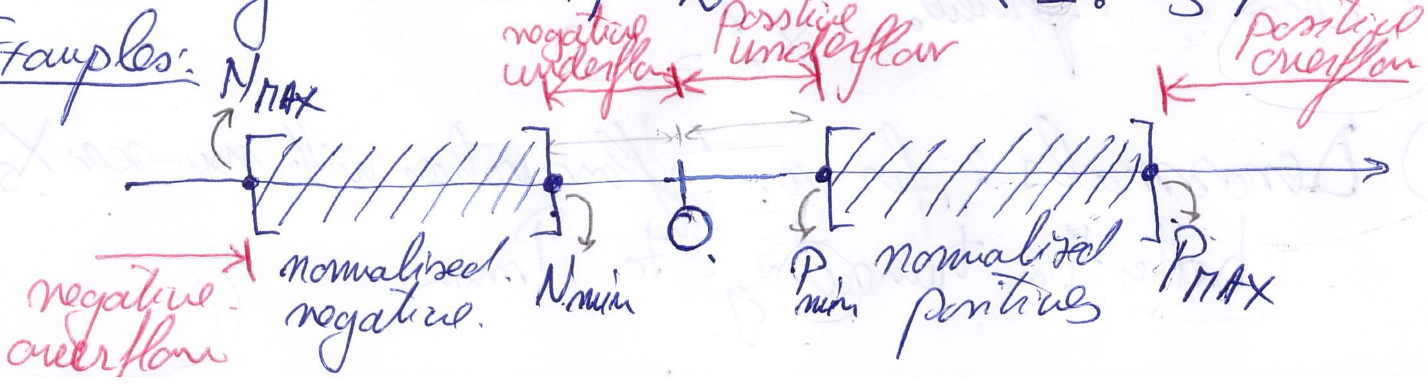
Range of values for X_S:

$$1 \le X_S < 2$$

X_E = bias of 127 $(=2^{8-1}-1)$

$$\text{Value of } X = (-1)^S * 2^{X_E-\text{bias}} * (1.X_S^*)$$

Examples:



P_{MAX}: $X_E = X_{E_{max}} - 1 = 254$

$$X_S^* = \underset{\substack{\uparrow \\ \text{binary point}}}{.}\underbrace{11\text{———}1}_{<23>} = 1 - 2^{-23}$$

$$P_{MAX} = (-1)^0 * 2^{254 - \overset{\text{bias}}{127}} * (\overset{\text{hidden bit}}{1}.\underbrace{11\text{———}1}_{<23>}) =$$

$$= 2^{127} * (1 + 1 - 2^{-23}) = 2^{127} * (2 - 2^{-23}) \simeq 3.4 * 10^{38}$$

P_{min}: $X_E = X_{E_{min}} + 1 = 1$

$$X_S^* = .\underbrace{00\text{———}0}_{<23>} = 0$$

$$P_{min} = (-1)^0 * 2^{1-127} * (1.\underbrace{00\text{———}00}_{<23>}) = 2^{-126} \simeq 1.18 * 10^{-38}$$

$x \bmod y < y$

Special quantities in IEEE 754

(A) Not a Number (NaN)

- result of: $\infty \pm \infty$, $0 * \infty$, $\frac{\infty}{\infty}$, $\frac{0}{0}$, $\sqrt{\text{negative}}$, $X \bmod 0$

Represent: $X_E = X_{E_{max}}$ (255)

$X_S^* \neq 0$ – indication of the operation that generated NaN

(B) Infinity: for overflow

- result of: $X \text{ div } 0$, $X \pm \infty$, $\sqrt{+\infty}$
- better than truncating to P_{MAX}

Represent: $X_E = X_{E_{max}}$ (255)

$X_S^* = 0$

(C) Zero: Represent: $X_E = X_{E_{min}}$ (0)

$X_S^* = 0$

(D) Denormals: for underflow values with non-zero X_S^*

- better than truncating to P_{min}

Representation $X_E = X_{E_{min}}$ (0)

$X_S^* \neq 0$

Value of the denormal no.:

$$X_{\underbrace{D}_{\text{denormal}}} = (-1)^s * 2^{1-\text{bias}} * (\underbrace{0}_{\text{hidden bit for denormals}}.X_S^*)$$

Examples:

$$612.8046875_{10} = 1001100100.1100111_2$$

$$\begin{array}{rl} 612 & - \\ 512 & 2^9 \\ \hline 100 & - \\ 64 & 2^6 \\ \hline 36 & - \\ 32 & 2^5 \\ \hline 4 & 2^2 \end{array}$$

.8046875	×2
1.6093750	×2
1.2187500	×2
0.4375000	×2
0.8750000	×2
1.7500000	×2
1.5000000	×2
1.0000000	

$$612.8046875 = 1.0011001001100111 * 2^9$$