B

FLOATING-POINT NUMBERS

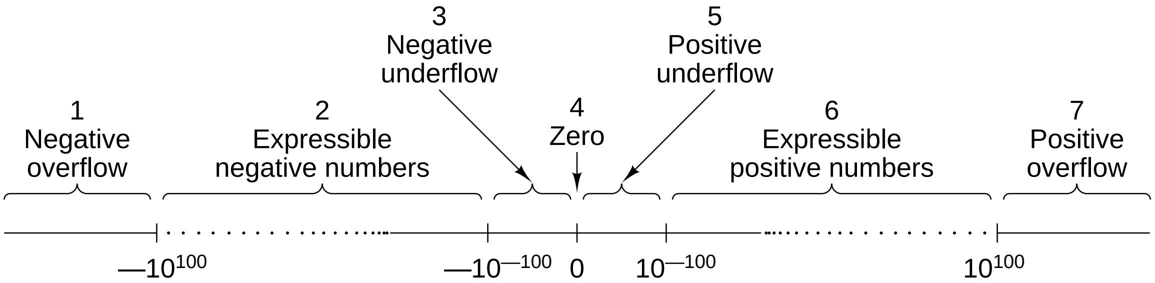
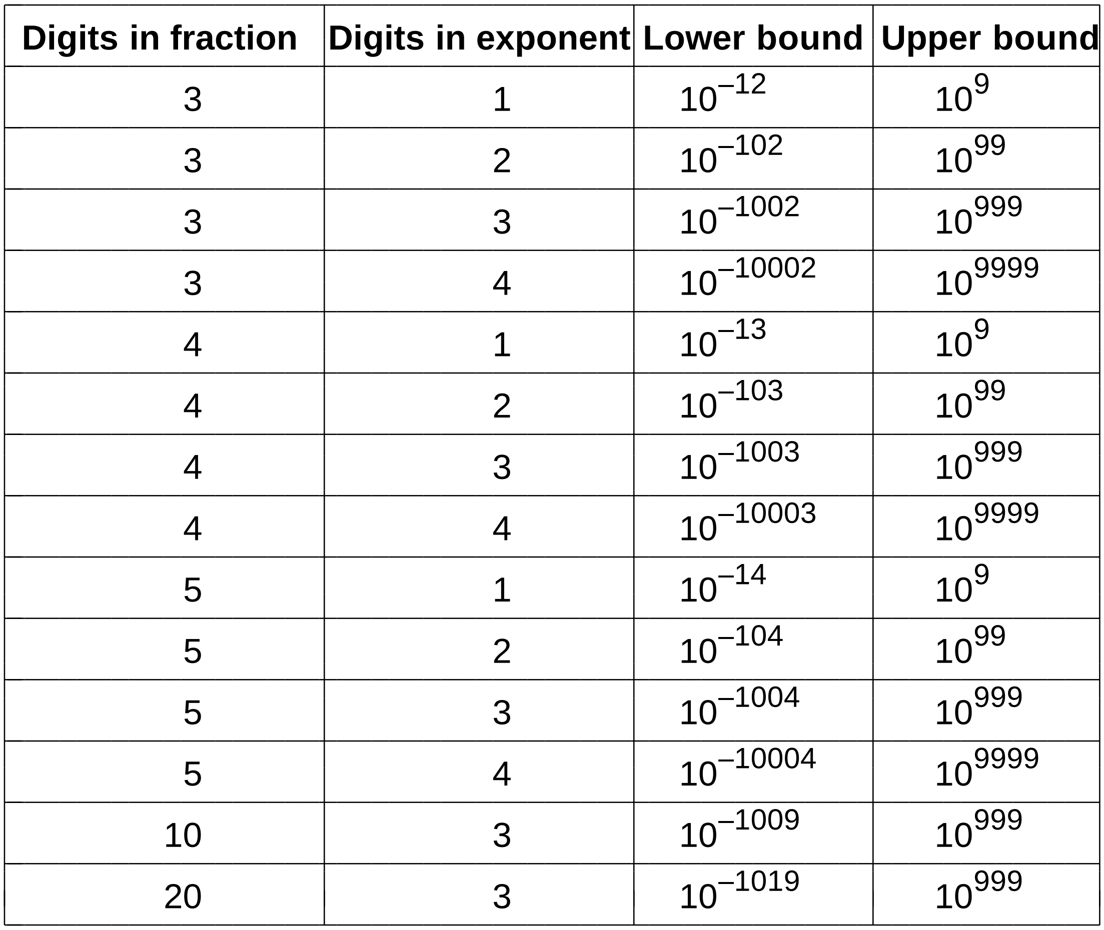


Figure B-1. The real number line can be divided into seven regions.



Digits in fraction	Digits in exponent	Lower bound	Upper bound
3	1	10^{-12}	10^{9}
3	2	10^{-102}	10^{99}
3	3	10^{-1002}	10^{999}
3	4	10^{-10002}	10^{9999}
4	1	10^{-13}	10^{9}
4	2	10^{-103}	10^{99}
4	3	10^{-1003}	10^{999}
4	4	10^{-10003}	10^{9999}
5	1	10^{-14}	10^{9}
5	2	10^{-104}	10^{99}
5	3	10^{-1004}	10^{999}
5	4	10^{-10004}	10^{9999}
10	3	10^{-1009}	10^{999}
20	3	10^{-1019}	10^{999}

Figure B-2. The approximate lower and upper bounds of expressible (unnormalized) floating-point decimal numbers.

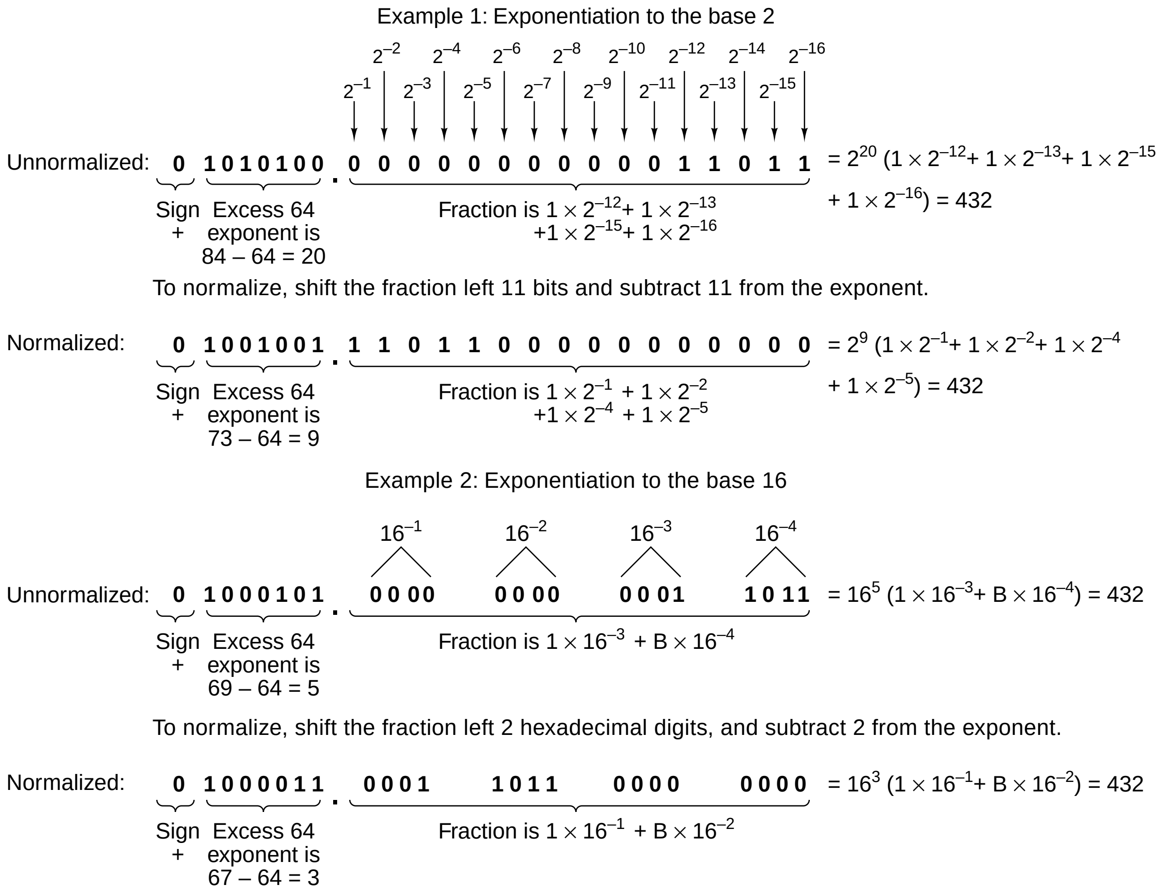

Example 1: Exponentiation to the base 2
2^–1 2^–2 2^–3 2^–4 2^–5 2^–6 2^–7 2^–8 2^–9 2^–10 2^–11 2^–12 2^–13 2^–14 2^–15 2^–16
Unnormalized: 0 1010100 . 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 1
= 2^20 (1 × 2^–12 + 1 × 2^–13 + 1 × 2^–15 + 1 × 2^–16) = 432
Sign +
Excess 64 exponent is 84 – 64 = 20
Fraction is 1 × 2^–12 + 1 × 2^–13 +1 × 2^–15 + 1 × 2^–16
To normalize, shift the fraction left 11 bits and subtract 11 from the exponent.
Normalized: 0 1001001 . 1 1 0 1 1 0 0 0 0 0 0 0 0 0 0 0
= 2^9 (1 × 2^–1 + 1 × 2^–2 + 1 × 2^–4 + 1 × 2^–5) = 432
Sign +
Excess 64 exponent is 73 – 64 = 9
Fraction is 1 × 2^–1 + 1 × 2^–2 +1 × 2^–4 + 1 × 2^–5
Example 2: Exponentiation to the base 16
16^–1 16^–2 16^–3 16^–4
Unnormalized: 0 1000101 . 0000 0000 0001 1011
= 16^5 (1 × 16^–3 + B × 16^–4) = 432
Sign +
Excess 64 exponent is 69 – 64 = 5
Fraction is 1 × 16^–3 + B × 16^–4
To normalize, shift the fraction left 2 hexadecimal digits, and subtract 2 from the exponent.
Normalized: 0 1000011 . 0001 1011 0000 0000
= 16^3 (1 × 16^–1 + B × 16^–2) = 432
Sign +
Excess 64 exponent is 67 – 64 = 3
Fraction is 1 × 16^–1 + B × 16^–2

Figure B-3. Examples of normalized floating-point numbers.

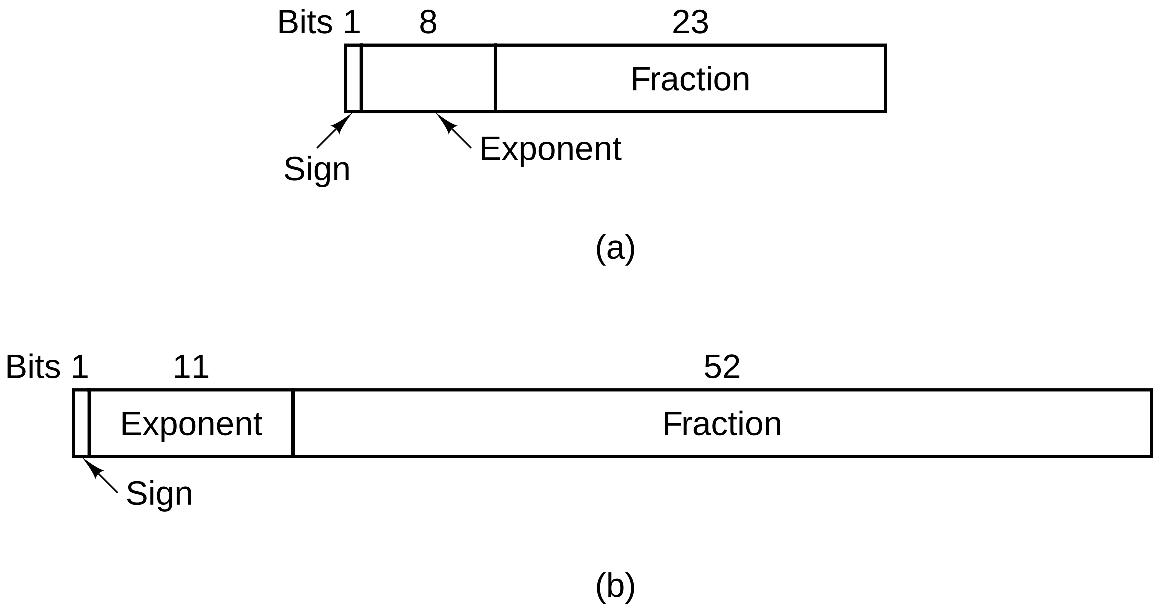
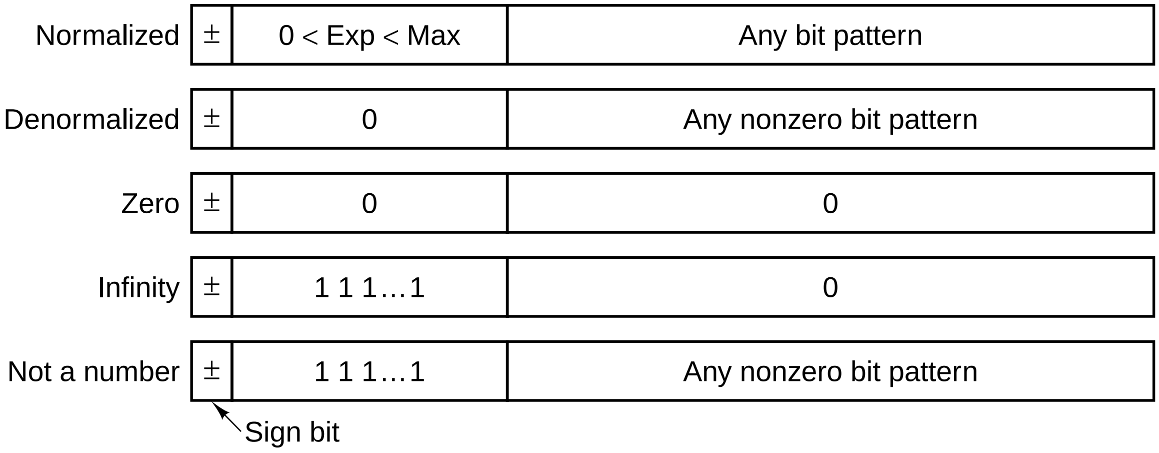


Figure B-4. IEEE floating-point formats. (a) Single precision. (b) Double precision.

Item	Single precision	Double precision
Bits in sign	1	1
Bits in exponent	8	11
Bits in fraction	23	52
Bits, total	32	64
Exponent system	Excess 127	Excess 1023
Exponent range	–126 to +127	–1022 to +1023
Smallest normalized number	2^{-126}	2^{-1022}
Largest normalized number	approx. 2^{128}	approx. 2^{1024}
Decimal range	approx. 10^{-38} to 10^{38}	approx. 10^{-308} to 10^{308}
Smallest denormalized number	approx. 10^{-45}	approx. 10^{-324}

Figure B-5. Characteristics of IEEE floating-point numbers.



Normalized	±	0 < Exp < Max	Any bit pattern
Denormalized	±	0	Any nonzero bit pattern
Zero	±	0	0
Infinity	±	1 1 1…1	0
Not a number	±	1 1 1…1	Any nonzero bit pattern

Figure B-6. IEEE numerical types.