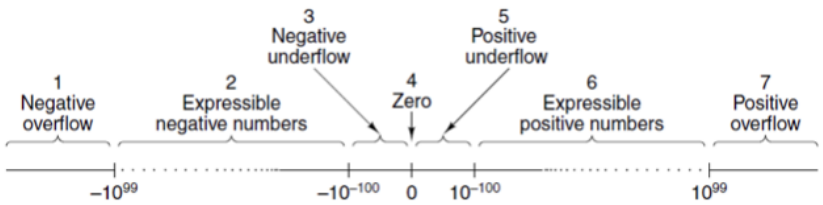
Floating point numbers

Description: Binary encoding of floating point numbers

Text reference: ***Appx B***

N = f x 10^e - N=floating point number; f=fraction (< 1.0); e=exponent

7 sections on real number line: underflow error, overflow error



IEEE Standard 754

Single & double precision
Sign bit
Exponent - Excess 127 encoding
Fraction - first bit is 2^{-1}
Normalized

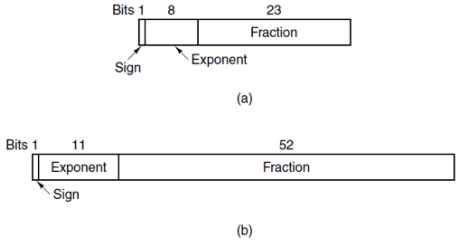


Figure B-4. IEEE floating-point formats. (a) Single precision. (b) Double precision.

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}