Real Numbers

- Any value on the number line (continuous, infinite).
- Computers approximate using finite bits.

Fixed Point Numbers

[sign]|integer part|.|fractional part

- Fixed number of fractional bits.
- Example: with 4 bits fractional part, $1.1011_2 = 1.6875_{10}$.

Floating Point Numbers

$$\pm(1.\text{fraction})_2 \times 2^{\text{exponent}}$$

- Normalized: leading 1 is implied.
- Mantissa: fractional bits with implicit 1.
- Biased exponent: $b = 2^{e-1} - 1$.

Special Values

- Zero: sign $\pm$, exp=0, frac=0
- $\pm\infty$: exp=all 1s, frac=0
- NaN: exp=all 1s, frac $\neq 0$

Denormalized (Subnormal) Numbers

- Exp=0, frac $\neq 0$
- No implied leading 1 $\rightarrow$ smaller precision near 0.

IEEE Standard

Single (32-bit): 1 sign, 8 exp, 23 frac
Double (64-bit): 1 sign, 11 exp, 52 frac

- realmin $= 1.0...000 \times 2^{1-b}$, realmax $= 1.11..111 \times 2^{2^e-1-b}$
- Example: Smallest positive subnormal $\approx 2^{-1074}$

Rounding

- **Round-to-nearest, ties-to-even**: avoids bias.
- **Round-towards-0**: truncates.
- **Round-to-$\pm\infty$**: ceiling/floor.

Unit Roundoff

$$\max \frac{|fl(x) - x|}{|x|} = u \quad \text{with } u = \tfrac{1}{2} \cdot 2^{-t}$$

where t = precision bits.

Floating Point Arithmetic

- Add/Sub: align exponents, add mantissas, normalize, round.
- Mult/Div: XOR signs, add/sub exponents, multiply mantissas.
- Guard + round + sticky bits $\rightarrow$ correct rounding.

Root-Finding Methods

Bisection

- Requires $f(a)f(b) < 0$.
- Update midpoint until $|b - a| < \delta$.
- Linear convergence.

Newton's Method

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$$

- Quadratic convergence if x_0 close to root.
- Example: $f(x) = x^2 - 2$, $x_0 = 1 \rightarrow x_1 = 1.5, x_2 = 1.416\ldots$

Secant Method

$$x_{k+1} = x_k - f(x_k)\frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})}$$

- Does not need f'.
- Superlinear convergence.

Taylor's Theorem

$$f(x) \approx P_n(x) = f(a) + f'(a)(x-a) + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n$$

Binary Representation Example

$13.625_{10} = 1101.101_2$
IEEE 32-bit (single):

$$\text{Sign} = 0,\ Exp = 10000010,\ Frac = 101101\ldots$$

MATLAB Example

```
% Example: smallest positive normal double
realmin

% Largest representable double
realmax

% Rounding illustration
x = 0.1 + 0.2;
disp(x == 0.3)   % returns false
```

MATLAB fzero

```
f = @(x) cos(x) - x;   % anonymous function
root = fzero(f, 0.5)   % initial guess
```

MATLAB num2bin Function

```
function binStr = num2bin(x)
% Return IEEE-754 double precision binary string
hexStr = num2hex(x);
decStr = hex2dec(hexStr');
binStrTmp = dec2bin(decStr,4);
binStr = reshape(binStrTmp.',1,[]);
end

% Example
num2bin(0.1)
```