Error (Approximating x with $\tilde{x}$)

$$\text{absolute error} = |\tilde{x} - x|$$

$$\text{relative error} = \frac{|\tilde{x} - x|}{|x|}.$$

Floating Point Number Systems

$\mathbb{F}(\beta, k, m, M)$ is a *finite subset* of the real number system. For $f \in \mathbb{F}$:

$$f = \pm (d_1.d_2d_3 \dots d_k)_\beta \times \beta^e$$

- $\beta \in \mathbb{N}$: the base
- $d_1.d_2d_3 \dots d_k$: the significand
- $k \in \mathbb{N}$: #digits in the significand
- $e \in \mathbb{Z}$: the exponent, $m \le e \le M$

d_i are base-β digits with $d_1 \ne 0$ unless $f = 0$. For $x \in \mathbb{R}$ and $f > 0$:

$$f_{\min} = \min_{f \in \mathbb{F}} |f| = \beta^m$$

$$f_{\max} = \max_{f \in \mathbb{F}} |f| = (1 - \beta^{-k}) \beta^{M+1}.$$

Underflow: $x < f_{\min}$ (replaced by 0).
Overflow: $x > f_{\max}$ (replaced by ∞).
For $\mathbb{F}^+ = \{f \in \mathbb{F} : f > 0\}$:

$$|\mathbb{F}^+| = (M - m + 1)(\beta - 1)\beta^{k-1}.$$

Representing Real Numbers

If $x \notin \mathbb{F}$, x is rounded to the nearest representable number with $fl : \mathbb{R} \to \mathbb{F}$. To determine $fl(x)$:

1. Express x in base-β.
2. Express x in scientific form.
3. Verify that $m \le e \le M$:
   - if $e > M$, then $x = \infty$.
   - if $e < m$, then $x = 0$.
   - else, round to k digits.

$$\frac{|fl(x) - x|}{|x|} \le u = \frac{1}{2}\beta^{1-k}.$$

where u is the **unit roundoff** of $\mathbb{F}$.

Catastrophic Cancellation

The error when subtracting similar floating point numbers, where at least one is not exactly representable.

Taylor Polynomials

The nth degree **Taylor polynomial** of f approximates f for x near x_0:

$$P_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k.$$

If f is $n+1$ times differentiable on $[a, b]$ containing x_0, then for all $x \in [a, b]$, there exists a value $x_0 < c < x$ such that

$$f(x) = P_n(x) + R_n(x)$$

where

$$R_n(x) = \frac{f^{(n+1)}(c)}{(n+1)!}(x - x_0)^{n+1}$$

is the **remainder (error) term** for P_n. The maximum value of $|R_n(x)|$ on $[a, b]$ bounds the maximum absolute error of the approximation:

$$|f(x) - P_n(x)| = |R_n(x)|.$$

Ordinary Differential Equations

$\frac{\mathrm{d}y}{\mathrm{d}t} = f(t, y)$ with $y(a) = \alpha$ on $a \le t \le b$. Divide $[a, b]$ into n subintervals of width $h = (b - a)/n$. Let $t_i = a + ih$ for $i = 0, 1, \dots, n$. Then $y_i = y(t_i)$ approximates y at $t = t_i$, with $y_0 = \alpha$.

Euler's Method (First Order Taylor)

$$y(t_i + h) = y(t_i) + hy'(t_i) + \mathcal{O}(h^2).$$

where the error is proportional to h^2.

$$y_{i+1} = y_i + hf(t_i, y_i).$$

Local and Global Error

Assuming the solution was correct at the previous step:
Local: error after 1 **step** — $\mathcal{O}(h^{p+1})$.
Global: error after i **steps** — $\mathcal{O}(h^p)$.
The **order** of a method is its global error.

Second Order Taylor Method

$$y_{i+1} = y_i + hf(t_i, y_i) + \frac{h^2}{2}f'(t_i, y_i).$$

Modified Euler Method

To avoid computing $f'(t, y)$ use,

$$\frac{f(t_{i+1}, y_{i+1}) - f(t_i, y_i)}{h} + \mathcal{O}(h).$$

$$y_{i+1} = y_i + \frac{1}{2}(k_1 + k_2)$$
$$k_1 = hf(t_i, y_i)$$
$$k_2 = hf(t_i + h, y_i + k_1)$$

Runge-Kutta Method (Fourth Order)

$$y_{i+1} = y_i + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)$$
$$k_1 = hf(t_i, y_i)$$
$$k_2 = hf\left(t_i + \frac{h}{2}, y_i + \frac{k_1}{2}\right)$$
$$k_3 = hf\left(t_i + \frac{h}{2}, y_i + \frac{k_2}{2}\right)$$
$$k_4 = hf(t_i + h, y_i + k_3)$$

$i = 0, 1, \dots, n-1$ for all four methods.

Interpolation

$$P_n(x) = a_0 + a_1x + \cdots + a_nx^n.$$

Lagrange Form

Solve for a_i then factor $P_n(x_i)$ for y_i:

$$P_n(x) = \sum_{i=0}^{n} L_{n,i}(x)\, y_i$$

$$L_{n,i}(x) = \prod_{\substack{j=0 \\ j\ne i}}^{n} \frac{x - x_j}{x_i - x_j}, L_{n,i}(x_j) = \delta_{ij}$$

For distinct increasing x_i on $[a, b]$ there exists $c \in [a, b]$ such that

$$f(x) = P_n(x) + \frac{f^{(n+1)}(c)}{(n+1)!}\prod_{i=0}^{n}(x - x_i).$$

Newton's Divided Difference Form

$$\begin{aligned} P_n(x) &= a_0 + a_1(x - x_0) \\ &\quad + a_2(x - x_0)(x - x_1) + \cdots \\ &\quad + a_n(x - x_0)\cdots(x - x_{n-1}) \\ &= \sum_{k=0}^{n} f[x_0, x_1, \dots, x_k]\prod_{i=0}^{k-1}(x - x_i) \end{aligned}$$

Solve $P_n(x_i) = y_i$ for $a_0, a_1, \dots, a_n$:

$$a_0 = y_0, \quad a_1 = \frac{y_1 - y_0}{x_1 - x_0}$$

$$a_2 = \frac{\frac{y_2 - y_1}{x_2 - x_1} - \frac{y_1 - y_0}{x_1 - x_0}}{x_2 - x_0}$$

Divided Differences (Simplify a_i)

$f[x_i] = y_i$ (Zeroth divided difference)

$$f[x_i, x_{i+1}, \dots, x_{i+k}] = \frac{f[x_{i+1}, \dots, x_{i+k}] - f[x_i, \dots, x_{i+k-1}]}{x_{i+k} - x_i}$$

$$f[x_0, x_1] = \frac{f[x_1] - f[x_0]}{x_1 - x_0} = \frac{y_1 - y_0}{x_1 - x_0}$$

$$f[x_1, x_2] = \frac{f[x_2] - f[x_1]}{x_2 - x_1} = \frac{y_2 - y_1}{x_2 - x_1}$$

$$f[x_0, x_1, x_2] = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0}$$

Newton's Forward Difference Form

Equally spaced abscissas: $h = x_{i+1} - x_i$.

Forward Difference Operator

$$\Delta y_i = y_{i+1} - y_i, \quad \Delta^{k+1}y_i = \Delta(\Delta^k y_i)$$
$$\Delta^2 y_i = y_{i+2} - 2y_{i+1} + y_i$$
$$\Delta^3 y_i = y_{i+3} - 3y_{i+2} + 3y_{i+1} - y_i$$

$$f[x_0, x_1, \dots, x_k] = \frac{\Delta^k y_0}{k!h^k}$$

Substitute $x = x_0 + sh$ $(x_i = x_0 + ih)$, with $s = \frac{x - x_0}{h}$ into the divided difference form:

$$P_n(x) = \sum_{k=0}^{n} \frac{\Delta^k y_0}{k!}\prod_{i=0}^{k-1}(s - i)$$

Root Finding ($f(x) = 0$)

Intermediate Value Theorem

For continuous f on $[a, b]$ with $f(a) \le k \le f(b)$, $\exists c_1 \in [a, b] : f(c_1) = k$. If $f(a)f(b) < 0$ ($f(a)$ and $f(b)$ have opposite signs), $\exists c_2 \in [a, b] : f(c_2) = 0$.

Bisection Method

1. Find $[a, b]$ such that $f(a)f(b) < 0$.
2. For $p = \frac{a+b}{2}$, evaluate $f(p)$.
   - If $f(p) = 0$, then p is a root of f.
   - If $f(a)f(p) < 0$, then p becomes the new b and the root lies in $[a, p]$.
   - If $f(p)f(b) < 0$, then p becomes the new a and the root lies in $[p, b]$.
3. Go to step 2.

Fixed-Point Iteration

Rewrite $f(x) = 0$ as $x = g(x)$. Solve by finding a fixed-point p s.t. $g(p) = p$.

$$x_{n+1} = g(x_n) \quad (n \ge 0).$$

Newton's Method

Find the root of the tangent line at each iterate x_n using the first degree Taylor polynomial and solving for x:

$$f(x) \approx f(x_n) + f'(x_n)(x - x_n) \overset{\text{set}}{=} 0$$

$$x = x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \quad (n \ge 0)$$

Secant Method

Approximate $f'(x_n)$ with the secant between x_{n-1} and x_n:

$$f'(x_n) \approx \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}$$

with two initial values for $n \ge 1$.

Convergence of Rootfinding Methods

A convergent $\{x_n\}$ satisfies (for large n)
$$|x_{n+1}-p| \approx \lambda |x_n - p|^r$$

Fixed-point iteration ($r=1$)

p is a fixed-point and $0<\lambda<1$.

Newton's method ($r=2$)

p is a root and $\lambda>0$.

Secant method ($r=\frac{1+\sqrt{5}}{2}\approx 1.618$)

p is a root and $\lambda>0$.

Numerical Differentiation

Forward ($h=x-x_0$, $c\in[x_0,x_0+h]$)
$$f'(x_0)=\frac{f(x_0+h)-f(x_0)}{h}-\frac{h}{2}f''(c)$$

Backward ($-h=x-x_0$, $c\in[x_0-h,x_0]$)
$$f'(x_0)=\frac{f(x_0)-f(x_0-h)}{h}+\frac{h}{2}f''(c)$$

Central Difference (Second Order)

Derive using $f(x_0+h)-f(x_0-h)$:
$$f'(x_0)=\frac{f(x_0+h)-f(x_0-h)}{2h}-\frac{h^2}{6}f^{(3)}(c)$$
$f^{(3)}(c)=\frac{f^{(3)}(c_1)+f^{(3)}(c_2)}{2}$ and $c\in[c_1,c_2]$, with $c_1\in[x_0-h,x_0]$ and $c_2\in[x_0,x_0+h]$.

Second Derivative (Third Order)

Derive using $f(x_0+h)+f(x_0-h)$:
$$f''(x_0)=-\frac{h^2}{12}f^{(4)}(c)+\frac{f(x_0+h)-2f(x_0)+f(x_0-h)}{h^2}$$
$f^{(4)}(c)=\frac{f^{(4)}(c_1)+f^{(4)}(c_2)}{2}$ and $c\in[c_1,c_2]$, with $c_1\in[x_0-h,x_0]$ and $c_2\in[x_0,x_0+h]$.

Numerical Integration (Quadrature)
$$I=\int_a^b f(x)\,\mathrm{d}x \approx \sum_{i=0}^{n} w_i f(x_i)$$
for weights w_i and abscissas x_i.
Divide $[a,b]$ into n subintervals of width $h=(b-a)/n$. Let $x_i=a+ih$ for $i=0,\,1,\,\dots,\,n$, so that $x_0=a$ and $x_n=b$.

Trapezoidal Rule (Second Order)

Approximate $f(x)$ over each subinterval $[x_{i-1},x_i]$ with a degree 1 interpolant:
$$P_{1,i}(x)=y_{i-1}+s\Delta y_{i-1}=y_{i-1}+s(y_i-y_{i-1})$$
and integrate w.r.t. s: $x=x_{i-1}+sh$, $\mathrm{d}x=h\,\mathrm{d}s$, with limits $s\in[0,1]$:
$$\int_{x_{i-1}}^{x_i} f(x)\,\mathrm{d}x \approx \int_0^1 P_{1,i}(x)\,h\,\mathrm{d}s=\frac{h}{2}(y_{i-1}+y_i)\quad (i=1,2,\dots,n).$$
$$I=\sum_{i=1}^{n}\int_{x_{i-1}}^{x_i} f(x)\,\mathrm{d}x\approx\sum_{i=1}^{n}\frac{h}{2}[f(x_{i-1})+f(x_i)]$$
$$=\frac{h}{2}\left[f(x_0)+2\sum_{i=1}^{n-1}f(x_i)+f(x_n)\right]-\frac{(b-a)h^2}{12}f''(c)$$

Simpson's Rule (Fourth Order)

Approximate $f(x)$ over each subinterval $[x_{2i-2},x_{2i}]$ with a degree 2 interpolant:
$$P_{2,i}(x)=y_{2i-2}+s\Delta y_{2i-2}+\frac{s(s-1)}{2}\Delta^2 y_{2i-2}$$
$$=y_{2i-2}+s(y_{2i}-y_{2i-1})+\frac{s(s-1)}{2}(y_{2i}-2y_{2i-1}+y_{2i-2})$$
and integrate w.r.t. s: $x=x_{2i-2}+sh$, $\mathrm{d}x=h\,\mathrm{d}s$, with limits $s\in[0,2]$:
$$\int_{x_{2i-2}}^{x_{2i}} f(x)\,\mathrm{d}x\approx\int_0^2 P_{2,i}(x)\,h\,\mathrm{d}s=\frac{h}{3}(y_{2i-2}+4y_{2i-1}+y_{2i})\quad (i=1,2,\dots,n/2).$$
$$I=\sum_{i=2}^{n/2}\int_{x_{2i-2}}^{x_{2i}} f(x)\,\mathrm{d}x\approx\sum_{i=2}^{n/2}\frac{h}{3}[f(x_{2i-2})+4f(x_{2i-1})+f(x_{2i})]$$
$$=\frac{h}{3}\left[f(x_0)+4\sum_{i=1}^{n/2}f(x_{2i-1})+2\sum_{i=1}^{n/2-1}f(x_{2i})+f(x_n)\right]-\frac{(b-a)h^4}{180}f^{(4)}(c)$$

Linear Systems ($\mathbf{A}\boldsymbol{x}=\boldsymbol{b}$)

LU Decomposition ($\mathbf{A}=\mathbf{LU}\implies \mathbf{L}\boldsymbol{z}=\boldsymbol{b}$, $\mathbf{U}\boldsymbol{x}=\boldsymbol{z}$)
$$\mathbf{LU}=\begin{bmatrix}1&0&0&0\\ \frac{a_{21}}{u_{11}}&1&0&0\\ \frac{a_{31}}{u_{11}}&\frac{a_{32}-\ell_{31}u_{12}}{u_{22}}&1&0\\ \frac{a_{41}}{u_{11}}&\frac{a_{42}-\ell_{41}u_{12}}{u_{22}}&\frac{a_{43}-\ell_{41}u_{13}-\ell_{42}u_{23}}{u_{33}}&1\end{bmatrix}\begin{bmatrix}a_{11}&a_{12}&a_{13}&a_{14}\\0&a_{22}-\ell_{21}u_{12}&a_{23}-\ell_{21}u_{13}&a_{24}-\ell_{21}u_{14}\\0&0&a_{33}-\ell_{31}u_{13}-\ell_{32}u_{23}&a_{34}-\ell_{31}u_{14}-\ell_{32}u_{24}\\0&0&0&a_{44}-\ell_{41}u_{14}-\ell_{42}u_{24}-\ell_{43}u_{34}\end{bmatrix}$$
$$=\begin{bmatrix}u_{11}&u_{12}&u_{13}&u_{14}\\ \ell_{21}u_{11}&\ell_{21}u_{12}+u_{22}&\ell_{21}u_{13}+u_{23}&\ell_{21}u_{14}+u_{24}\\ \ell_{31}u_{11}&\ell_{31}u_{12}+\ell_{32}u_{22}&\ell_{31}u_{13}+\ell_{32}u_{23}+u_{33}&\ell_{31}u_{14}+\ell_{32}u_{24}+u_{34}\\ \ell_{41}u_{11}&\ell_{41}u_{12}+\ell_{42}u_{22}&\ell_{41}u_{13}+\ell_{42}u_{23}+\ell_{43}u_{33}&\ell_{41}u_{14}+\ell_{42}u_{24}+\ell_{43}u_{34}+u_{44}\end{bmatrix}$$

Symmetric Positive Definite: $\boldsymbol{x}^\top\mathbf{A}\boldsymbol{x}>0:\forall\boldsymbol{x}\in\mathbb{R}^n$.

Cholesky Decomposition ($\mathbf{A}=\mathbf{LL}^\top\implies\mathbf{L}\boldsymbol{z}=\boldsymbol{b}$, $\mathbf{L}^\top\boldsymbol{x}=\boldsymbol{z}$)
$$\mathbf{L}=\begin{bmatrix}\sqrt{a_{11}}&0&0&0\\ \frac{a_{21}}{\ell_{11}}&\sqrt{a_{22}-\ell_{21}^2}&0&0\\ \frac{a_{31}}{\ell_{11}}&\frac{a_{32}-\ell_{21}\ell_{31}}{\ell_{22}}&\sqrt{a_{33}-\ell_{31}^2-\ell_{32}^2}&0\\ \frac{a_{41}}{\ell_{11}}&\frac{a_{42}-\ell_{21}\ell_{41}}{\ell_{22}}&\frac{a_{43}-\ell_{31}\ell_{41}-\ell_{32}\ell_{42}}{\ell_{33}}&\sqrt{a_{44}-\ell_{41}^2-\ell_{42}^2-\ell_{43}^2}\end{bmatrix}$$
$$\mathbf{LL}^\top=\begin{bmatrix}\ell_{11}^2&\ell_{11}\ell_{21}&\ell_{11}\ell_{31}&\ell_{11}\ell_{41}\\ \ell_{11}\ell_{21}&\ell_{21}^2+\ell_{22}^2&\ell_{21}\ell_{31}+\ell_{22}\ell_{32}&\ell_{21}\ell_{41}+\ell_{22}\ell_{42}\\ \ell_{11}\ell_{31}&\ell_{21}\ell_{31}+\ell_{22}\ell_{32}&\ell_{31}^2+\ell_{32}^2+\ell_{33}^2&\ell_{31}\ell_{41}+\ell_{32}\ell_{42}+\ell_{33}\ell_{43}\\ \ell_{11}\ell_{41}&\ell_{21}\ell_{41}+\ell_{22}\ell_{42}&\ell_{31}\ell_{41}+\ell_{32}\ell_{42}+\ell_{33}\ell_{43}&\ell_{41}^2+\ell_{42}^2+\ell_{43}^2+\ell_{44}^2\end{bmatrix}$$

Brouwer's Fixed-Point Theorem

For g continuous on $[a,b]$, and differentiable on (a,b), with $g(x)\in[a,b]:\forall x\in[a,b]$, let a positive constant $k<1$ exist such that $|g'(x)|\le k\ \forall x\in(a,b)$. Then, g has a unique fixed-point p in $[a,b]$, and $x_{n+1}=g(x_n)$ will converge to p for all x_0 in $[a,b]$.