3X03 Derivations & Results

Neelkant Teeluckdharry

October 2024

Contents

1 IEEE 754

1.1 Special Values in Single Precision

1. +Inf: Sign Bit = 0, Exponent = 255, Mantissa = 0
2. -Inf: Sign Bit = 1, Exponent = 255, Mantissa = 0
3. NaN: Sign Bit = 0/1, Exponent = 255, Mantissa: at least 1

1.2 "Rules"

1. If $e \neq 0 \implies$ normalized
2. If $e = 0$, then 1 is added to offset.

1.3 Definitions

1. ϵ_{mach} is defined as the distance from 1 to the next largest FP number.

2 Taylor Series

2.1 e^x Series

$$f(x) = f(0) + \frac{f'(0)x}{1!} + \frac{f''(0)x^2}{2!} + \frac{f^3(0)x^3}{3!} + ..$$

$$f(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + ... = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

2.2 Truncation Error of First Order Taylor Approximation

Truncation error is defined as the error from truncating a series. Let $x \to x + h$, $c \to x$.

$$f(x+h) = f(x) + f'(x)h + \frac{f''(\zeta)h^2}{2!}$$

where $x < \zeta < x + h$.

$$f'(x) = \frac{f(x+h) - f(x)}{h} - \frac{f''(\zeta)h}{2!}$$

This implies the truncation error is $\boxed{-\frac{f''(\zeta)h}{2}}$.

3 Linear Algebra

3.1 Nullspace/Kernel of a Matrix A

Let $v = \alpha_1 v_1 + \alpha_2 v_2$ for any $v_1, v_2 \in Ker(A)$ with $\alpha_{1,2} \in R$. Then

$$Av = Av_1 + Av_2 = 0$$

which proves that $Ker(A) \subseteq R^n$.

3.2 Column Space of Matrix A

The column space of $A \in R^{m \times n}$ can be equivalently defined as all linear combinations of the columns s.t. $y = c_1 a_1 + c_2 a_2 + ... c_n a_n$, where a_i is a column vector where $x = [c_1 c_2 ... c_n]^T$. From here, it can easily be shown that for some vectors $y_1, y_2 \in col(A) \in R^m$ and reals $\alpha, \beta \in R$, the space is closed under addition and multiplication, which proves that $col(A)$ is a subspace of R^m.

4 Naive Gaussian Elimination

4.1 Algorithm

```
GaussianElimination(A):
    for k = 1 to n-1:
        for i = k+1 to n:
            m_ik = A[i,k]/A[k,k]
            for j = k+1 to n:
                A[i,j] = A[i,j] - m_ik*A[k,j]
            end
            b[i] = b[i] - m_ik*b[k]
        end
```

```
end
return A (row echelon form)
```

4.2 Relative Solution Error

$$\|x^* - x\| = \left\|A^{-1}r\right\| \leq \left\|A^{-1}\left\|r\right\|\right\|$$

And

$$\|b\| = \|Ax^*\| \leq \|A\| \|x^*\| \implies \|x^*\| \geq \frac{\|b\|}{\|A\|}$$

$$\frac{\|x^* - x\|}{\|x^*\|} \leq \frac{\|A\| \left\|A^{-1}\right\| \|r\|}{\|b\|} = \frac{\kappa(A) \|r\|}{\|b\|}$$

4.3 Time Complexity

$$\sum_{k=1}^{n-1} [2 * (n-k)^2 + (n-k)]$$

$$= \sum_{k=1}^{n-1} 2k^2 + k$$

$$= \frac{2 * (n-1) * n * (2n-1)}{6} + \frac{(n-1) * n}{2}$$

$$= \frac{2n^3}{3} - \frac{n^2}{2} - \frac{n}{6} \implies O(n^3)$$

5 LU Decomposition

5.1 LU Decomposition

Let $M_k = M_{nk} M_{n-1,k} ... M_{k+1,k}$. These correspond to the operations performed "clear" the column k of the matrix A. For naive gaussian elimination, the elementary matrix M_k, is a product of lower triangular matrices, meaning it is also lower triangular. Thus,

$$M_{n-1} ... M_2 M_1 A = U$$

$$A = M_1^{-1} M_2^{-1} ... M_{n-1}^{-1} U$$

Thus,

$$L = \boxed{M_1^{-1} M_2^{-1} ... M_{n-1}^{-1}}$$

6 Vector Norms

6.1 p-∞ norm

$$\|x\|_\infty = \lim_{p\to\infty}(\sum_{i=0}^{n}\|x_i\|^p)^{\frac{1}{p}}$$

$$= \lim_{p\to\infty}(\sum_{i=0}^{n}\frac{\|x_i\|}{\max_{i=1,..n}\|x_i\|})\max_{i=1,..n}\|x_i\|$$

$$= \max_{i=1,..n}\|x_i\|$$

7 Matrix Norms

7.1 Proof of $\|Ax\| \leq \|A\|\,\|x\|$

By definition

$$\|A\| = \max_{\|x\|\neq 0}\frac{\|Ax\|}{\|x\|} \geq \frac{\|Ax\|}{\|x\|}\forall x$$

This implies

$$\|A\|\,\|x\| \geq \|Ax\|$$

7.2 Proof of $\|AB\| \leq \|A\|\,\|B\|$

By definition

$$\|AB\| = \max_{\|x\|\neq 0}\frac{\|ABx\|}{\|x\|}$$

$$\leq \max_{\|x\|\neq 0}\frac{\|A\|\,\|Bx\|}{\|x\|}$$

$$\leq \max_{\|x\|\neq 0}\frac{\|A\|\,\|B\|\,\|x\|}{\|x\|}$$

$$= \max_{\|x\|\neq 0}\|A\|\,\|B\| = \|A\|\,\|B\|$$

7.3 Proof of 1-Norm

By definition

$$\|A\|_1 = \max_{\|x\|\neq 0}\frac{\|Ax\|_1}{\|x\|_1}$$

$$= \max_{\|x\|=1}\|Ax\|_1$$

$$= \max_{\|x\|=1}\|a_1 \cdot x_1 + a_2 \cdot x_2 + ... + a_n \cdot x_n\|$$

Max is attained when $x_i = 1$ and $x_j = 0 \; \forall i \neq j$ where i is column with largest 1-norm.

$$\|A\|_1 = \max_{j=1 \ldots n} \sum_{i=1}^{n} \|a_{ij}\|$$

7.4 Proof of 2-Norm

$$\|A\|_2 = \max_{\|x\| \neq 0} \frac{\|Ax\|_2}{\|x\|_2}$$

$$= \max_{\|x\|=1} \|Ax\|_2$$

$$= \max_{\|x\|=1} \sqrt{x^T A^T A x}$$

$$= \max_{\|x\|=1} \sqrt{\lambda^2 x^T x}$$

Since $x^T x$ is equivalent to $\|x\|_2$. The only parameter that will change across all such vectors is the eigenvalue with eigenvector x.

$$= \max_{i=1..n} \lambda_i$$

7.5 Proof of ∞-Norm

Similar proof to above except write each jth entry of final vector as linear combinations of the jth components of each of the column vectors multiplied by x_is. Max element is the row sum. Keep in mind that all entries of $x = \pm 1$ will give this since $\|x\|_\infty = 1$.

8 Eigen Results

8.1 Eigenvalue of inverse of A

Suppose A is a matrix with an eigenvalue of λ.

$$A^{-1}x = \lambda' x$$

$$\frac{1}{\lambda} x = Ax$$

$$\implies \lambda' = \frac{1}{\lambda}$$

8.2 Linear Independence of Eigenvectors with Distinct Eigenvalues

Suppose this is not the case. Then there exist $x_1, x_2...x_n \in R^n$ s.t. $0 = c_1x_1 + c_2x_2 + ...c_nx_n$, where not all $c_i = 0$. Now multiply by matrix $(A - \lambda_2 I_{n\times n})(A - \lambda_3 I_{n\times n})..(A - \lambda_n I_{n\times n})$.

$$0 = c_1(\lambda_1 - \lambda_2)...(\lambda_1 - \lambda_n)$$

Thus, $c_1 = 0$. It can also be shown that all $c_i = 0$, which means that all the eigenvectors are linearly independent and span R^n.

8.3 Eigenvectors of symmetric matrices are orthogonal

$$\begin{aligned} x_1^T A x_2 &= x_1^T \lambda_2 x_2 \\ &= \lambda_2 x_1^T x_2 \end{aligned}$$

And,

$$\begin{aligned} x_1^T A x_2 &= x_1^T A^T x_2 \\ &= (Ax_1)^T x_2 \\ &= \lambda_1 x_1^T x_2 \\ \implies (\lambda_2 - \lambda_1) x_1^T x_2 = 0 &\implies x_1^T x_2 = 0 \end{aligned}$$

9 LDL^T Transformation

9.1 Product of two lower triangular matrices

Suppose $j > i$

$$(LL')_{ij} = \sum_{k=1}^{n} l_{ik} l'_{kj}$$

Since $l_{ik} = 0$ when $k > i$ and $l'_{kj} = 0$ when $k > j > i$, this means

$$(LL')_{ij} = 0$$

9.2 Diagonal Matrix

If A is symmetric, then,

$$D = L^{-1} A L^{-T} = U L^{-T}$$

$$D^T = L^{-1} A L^{-T}$$

D is both upper triangular and symmetric $\implies$ diagonal matrix.

9.3 Algorithm

$$a_{ii} = (LDL^T)_{ii} = \sum_{k=1}^{n} l_{ik} d_k l_{ik}$$

$$= \sum_{k=1}^{i} l_{ik}^2 d_k = \sum_{k=1}^{i-1} l_{ik}^2 d_k + d_i$$

$$\implies d_i = a_{ii} - \sum_{k=1}^{i-1} d_k l_{ik}^2$$

10 Positive-Definiteness

10.1 Invertibility

$Ax = 0$ iff $x = 0$ since ($Ax = 0 \implies x^T Ax = 0$ but $x^T Ax > 0 \forall x \neq 0$) $\implies ker(A)$ has dimension 0 $\implies A$ is invertible

10.2 A has real eigenvalues

Let $\lambda \in R$ be an eigenvalue of A.

$$Av = \lambda v$$

$$v^T Av = \lambda v^T v = \lambda \|v\|^2$$

Since A is positive definite, this implies that $\lambda > 0$.

11 Iterative Methods

11.1 Cholesky Factorization

12 Power Method

For $A \in R^{n \times n}$ with n linearly independent eigenvectors spanning R^n, this means any vector $v_0 = \sum_{k=1}^{n} a_i x_i$. At each iteration,

$$\tilde{v_{k+1}} = \frac{v}{\|v\|}$$

$$v_{k+1} = A\tilde{v}$$

where

$$\lambda_k = v[1]/\tilde{v[1]}, x_k = v[1]/\tilde{v[1]}$$

Assuming $\lambda_1 \geq \lambda_i$

$$A^k v = \lambda_1^k (a_1 x_1 + \frac{\lambda_2^k}{\lambda_1^k} ...)$$

As $k \to \infty$, $A^k v \to (a_1 \lambda_1^k x_1)$

13 Linear Regression

13.1 Derivation 1

Let $\phi(a,b) = \sum_{k=0}^{n} (ax_i + b - y_i)^2$. Using the first order optimality conditions (prove another time)

$$\frac{\delta\phi}{\delta a} = \sum_{k=0}^{n} (ax_i + b - y_i) * x_i = \sum_{k=0}^{n} ax_i^2 + bx_i - x_i y_i = 0$$

$$\implies a \sum_{k=0}^{n} x_i^2 + b \sum_{k=0}^{n} x_i = \sum_{k=0}^{n} x_i y_i$$

$$\frac{\delta\phi}{\delta b} = \sum_{k=0}^{n} 2(ax_i + b - y_i) = 0$$

$$\implies a \sum_{k=0}^{n} x_i + (n+1)b = \sum_{k=0}^{n} y_i$$

This gives a linear system which can be solved.

13.2 Derivation 2

$$\begin{aligned} f(z) &= \|Az - y\| = (Az - y)^T (Az - y) \\ &= z^T A^T A z - z^T A^T y - y^T A z + \|y\|^2 \\ &= z^T A^T A z - 2 z^T A^T y + \|y\|^2 \end{aligned}$$

Since transpose of a scalar is a scalar.

$$\nabla f(z) = 2A^T A z - 2A^T y = 0$$

$$A^T A z = A^T y$$

$$z = (A^T A)^{-1} A^T y$$

This is the moore-penrose pseudoinverse.

14 Singular Value Decomposition

14.1 Decomposition

Let $A \in R^{m \times n}$, if A has a decomposition $U\Sigma V^T$, then

$$\begin{aligned} AA^T &= U\Sigma V^T V \Sigma^T U^T \\ &= U\Sigma\Sigma^T U^T \end{aligned}$$

Since AA^T is $n \times n$ and symmetric, by the spectral theorem, it is equivalent to $V\Lambda V^T$, which means that $\Lambda = \Sigma\Sigma^T$. This implies that the singular values $\sigma_i = \sqrt{\lambda_i}$.

15 Bauer-Fike Bound

$$\begin{aligned} r = A\hat{x} - \hat{\lambda}\hat{x} &= (A - \hat{\lambda}I)\hat{x} \\ \hat{x} &= (A - \hat{\lambda}I)^{-1} r \end{aligned}$$

This simplifies to

$$\begin{aligned} \|\hat{x}\| &= \left\| P(\Lambda - \hat{\lambda}I)P^{-1} \right\| \\ &\leq \|P\|_p \left\| \Lambda - \hat{\lambda}I \right\|_p \left\| P^{-1} \right\|_p \|r\|_p \\ \left\| (\Lambda - \hat{\lambda}I)^{-1} \right\|_p &= \max_{\|x\|_p = 1} \left\| (\Lambda - \hat{\lambda}I)x \right\| \\ &= \max_{\|x\|_P} (\sum_{i=1}^{n} (\frac{x_i}{\lambda_i - \hat{\lambda}})^p)^{1/p} \end{aligned}$$

If we pick λ_i s.t. $\lambda_i - \hat{\lambda}$ is minimized, then the rest evaluates to $\|x\|_p$.

$$= \frac{1}{\min_{\lambda_i \in \sigma(A)} |\lambda_i - \hat{\lambda}|}$$

This implies

$$\min_{\lambda_i \in \sigma(A)} |\lambda_i - \hat{\lambda}| \leq \frac{\kappa(P)_p \|r\|_p}{\|\hat{x}\|_p}$$

16 Newton's Method

The second order Taylor expansion of f is given by:

$$\begin{aligned} f(x) &= f(x_k) + f'(x_k)(x - x_k) + O(||x - x_k||^2) \\ &\implies 0 \approx f(x_k) + f'(x_k)(r - x_k) \end{aligned}$$

Let $x_{k+1} \triangleq r$.

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

17 First Order Optimality

Let x^* be a local minimum, such that if $\|x^* - x\| \leq \delta$, then $f(x^*) \leq f(x)$, Write the taylor series about x^*

$$f(x) = f(x^*) + \nabla f(x^*)(x - x^*) + O(\|x - x^*\|^2)$$

Choose x s.t. $x - x^* = -\alpha \nabla f(x^*)$, where $\alpha > 0$. If α is chosen s.t

$$\nabla f(x^*)^T (x - x^*) = -\alpha \left\|x - x^*\right\|^2 < 0$$

Then,

$$\left\|x - x^*\right\|^2 = \alpha^2 \left\|\nabla f(x^*)\right\|^2$$

α^2 term will be smaller in magnitude than α term.

$$f(x^*) + \nabla f(x^*)^T (x - x^*) + O(\|x - x^*\|^2) \leq f(x)$$

which means $\nabla f(x^*) = 0$.

18 Hessian is positive definite at minimum

Consider the 2nd order Taylor series about x^*

$$f(x) = f(x^*) + \nabla f(x^*)^T (x - x^*) + \frac{1}{2}(x - x^*)^T \nabla^2 f(x^*)(x - x^*) + O(\|x - x^*\|^3)$$

$$f(x) = f(x^*) + \frac{1}{2}(x - x^*)^T \nabla^2 f(x^*)(x - x^*) + O(\|x - x^*\|^3)$$

$$\approx f(x^*) + \frac{1}{2}(x - x^*)^T \nabla^2 f(x^*)(x - x^*)$$

This requires that $\nabla^2 f(x^*) > 0$.

19 Polynomial Error Bound

Assume $x \neq x_i$, define $w(t) = \prod_{i=0}^{n}(t - x_i)$ and $c = \frac{f(x) - p_n(x)}{w(x)}$. Finally let $\phi(t) = f(t) - p_n(t) - cw(t)$. Observe that $\phi(t)$ has $n + 2$ roots: $x_0....x_n$ and x. By Rolle's theorem $\phi'(t) = 0$ for some t between pairs of roots. By recursive logic, $\implies \phi^{n+1}(t)$ has at least one root. Thus

$$0 = \phi^{n+1}(t) = f^{n+1}(\zeta) - c(n + 1)!$$

since $p(t)$ are n degree polynomials and $w(t)$ has a single degree t^{n+1} term.

$$0 = f^{n+1}(\zeta) - c(n + 1)! = f^{n+1}(\zeta) - \frac{f(x) - p_n(x)}{w(x)}(n + 1)!$$

$$\implies f(x) - p_n(x) = \frac{f^{n+1}(\zeta)}{(n+1)!}\prod_{i=0}^{n}(t - x_i)$$

Let $M \triangleq \max_{a \leq t \leq b}$

$$\implies \|f(x) - p_n(x)\| \leq \frac{M}{(n+1)!}\prod_{i=0}^{n}(x - x_i)$$

By some lemma this is also:

$$\leq \frac{M}{4(n+1)!}h^{n+1}(n!) = \frac{M}{4(n+1)}h^{n+1}$$

20 Complexity of Evaluation of Polynomial

$$p(x) = c_0 + c_1 x + c_2 x^2 + \cdots + c_n x^n$$

Each power k will require $k-1$ FLOPs to compute x^k, and multiplying by coefficient gives k FLOPs for each each term of order k. Thus $\sum_{k=0}^{n} k = n(n+1)/2 \implies O(n^2)$

21 Newton Interpolation

The basis function $\phi_j(x) = (x - x_0)(x - x_1)\ldots(x - x_{j-1})$. Thus

$$p_n(x_i) = c_0 + c_1(x_i - x_0) + c_2(x_i - x_0)(x_i - x_1)\ldots c_n(x_i - x_0)..(x_i - x_{n-1}) = y_i$$

At $x = x_0$,

$$p_n(x_0) = c_0 = y_0$$

At $x = x_1$,

$$p_n(x_1) = c_0 + c_1(x_1 - x_0) = y_1$$

$$\implies c_1 = \frac{y_1 - y_0}{x_1 - x_0}$$

For further coefficients this becomes a divided difference

$$[y_i, \ldots y_j] = \frac{[y_{i+1}\ldots y_j] - [y_i \ldots y_{j-1}]}{x_j - x_i}$$

22 Integral Form of MVT

Let $F(x) \triangleq \int_0^x f(x)dx$ for continuous f on $[a,b] \implies F(x)$ is continuous. Applying MVT

$$F'(x) = f(x) = \frac{F(b) - F(a)}{b - a}$$

$$= \frac{1}{b-a}\int_a^b f(x)dx$$

23 Trapezoidal Rule

A Lagrange interpolant takes the form:

$$p_n(x) = \sum_{j=0}^{n} y_j L_j(x)$$

where $L_j(x) = \frac{(x-x_0)(x-x_1)\ldots(x-x_n)}{(x_j-x_0)\ldots(x_j-x_n)}$.

$$\int_a^b f(x)dx \approx \int_a^b p_n(x)dx = \int_a^b \sum_{j=0}^{n} f(x_j)L_j(x)dx$$

$$= \sum_{j=0}^{n} f(x_j) \int_a^b L_j(x)dx$$

Since our quadrature rule uses $n = 1$.

$$p_1(x) = f(a)\frac{x-b}{a-b} + f(b) + f(b)\frac{x-a}{b-a}$$

$$\implies \sum_{j=0}^{n} f(x_j) \approx f(a)\int_a^b \frac{x-b}{a-b}dx + f(b)\int_a^b \frac{x-a}{b-a}dx$$

$$\ldots \frac{b-a}{2}[f(a)+f(b)]$$

24 Error on Trapezoidal Rule

The error of first order interpolant is

$$f(x) - p_1(x) = \frac{f''(\eta)}{2}(x-x_0)(x-x_1)$$

Thus

$$I_f - I_{trap} = \int_a^b f''(\zeta(x))(x-x_0)(x-x_1)dx = \frac{1}{2}f''(\eta)\int_a^b (x-a)(x-b)dx$$

$$= \cdots = -\frac{f''(\eta)(b-a)^3}{12}$$

25 Composite Trapezoidal Rule

$$\int_{t_i-1}^{t_i} f(x)dx \approx \frac{t_i - t_{i-1}}{2}[f(t_{i-1}) + f(t_i)] = \frac{h}{2}[f(t_{i-1}) + f(t_i)]$$

$$\implies \int_a^b f(x)dx = \sum_{i=1}^{r} \int_{t_i-1}^{t_i} f(x)dx \approx \frac{h}{2}\sum_{i=1}^{r}[f(t_i) + f(t_{i-1})]$$

This is equal to

$$= \frac{h}{2}[f(a) + f(b)] + h\sum_{i=1}^{r-1} f(t_i)$$

Furthermore,

$$\int_{t_i-1}^{t_i} f(x)dx = \frac{h}{2}[f(t_{i-1}) + f(t_i)] - \frac{f''(\eta_i)h^3}{12}$$

Since

$$\min_{x\in[a,b]} f''(x) \leq \frac{1}{r}\sum_{i=0}^{r} f''(\eta_i) \leq \max_{x\in[a,b]} f''(x)$$

By IVT

$$f''(\mu) = \frac{1}{r}\sum_{i=1}^{r} f''(\eta_i)$$

$$\implies \text{error} = -\sum_{i=1}^{r}\frac{f''(\eta_i)h^3}{12} = \frac{-f''(\mu)(b-a)h^2}{12}$$

26 Midpoint Rule

Let $m = (a+b)/2$, write the Taylor series

$$f(x) = f(m) + f'(m)(x-m) + \frac{f''(\zeta(x))(x-m)^2}{2}$$

Notice that

$$\int_a^b (x-m)dx = 0 \implies I_f = \int_a^b f(x)dx = (b-a)f(m) + \frac{1}{2}\int_a^b f''(\zeta(x))(x-m)^2dx$$

$$I_f - I_{mid} = 1/2\int_a^b f''(\zeta(x))(x-m)^2dx$$

By integral MVT

$$= \frac{1}{2}f''(\eta)\int_a^b (x-m)^2dx$$

for some $\eta \in [a, b]$

$$= \frac{f''(\eta)(b-a)^3}{24}$$

27 Simpson's Rule

Using $n = 2$ for Lagrange basis polynomials:

$$\implies I_{simpson} = \frac{b-a}{6}[f(a) + 4f(m) + f(b)]$$

The error is:

$$\frac{-f^4(\zeta)}{90}\frac{(b-a)^5}{2^5}$$

28 Error Analysis of Adaptive Simpsons

Applying Simpson's rule on $[a, m]$ and $[m, b]$

$$\begin{aligned}
E(a,m) &= \frac{-1}{90}\frac{(h/2)^5}{2^5}f^4(\zeta) \\
&= \frac{1}{32}\left(\frac{-1}{90}\frac{h^5}{2^5}f^4(\zeta)\right) \\
&= \frac{1}{32}E_1 \\
\implies E_2 &= 2\frac{1}{32}E_1 = \frac{E_1}{16}
\end{aligned}$$

Thus

$$\begin{aligned}
I_f = S_1 + E_1 = S_2 + E_2 &\implies S_1 - S_2 = E_2 - E_1 = -15E_2 \\
&= E_2 \approx \frac{S_2 - S_1}{15} \\
&\implies I_f = S_2 + \frac{S_2 - S_1}{15}
\end{aligned}$$

29 Error on Iteration of Adaptive Simpson

Since $I_f = I_1 + I_2$

$$\begin{aligned}
|I_f - Q_1| &\leq \frac{\epsilon_{tol}}{2} \\
|I_f - Q_2| &\leq \frac{\epsilon_{tol}}{2} \\
|I - Q| = |I_1 + I_2 - Q_1 - Q_2| &\leq |I_1 - Q_1| + |I_2 - Q_2| = \epsilon_{tol}
\end{aligned}$$

30 Forward Euler's Method

Consider Taylor series centered at t_i at t_{i+1}.

$$y(t_{i+1}) = y(t_i) + y'(t_i)(t_{i+1} - t_i) + \frac{1}{2}y''(\zeta_i)(t_{i+1} - t_i)^2$$

For some $\zeta_i \in [t_i, t_{i+1}]$

$$\approx y(t_i) + y'(t_i)h$$

31 Backwards Euler's Method

Consider a Taylor Series centered at t_{i+1} at t_i.

$$y(t_i) = y(t_{i+1}) - y'(t_{i+1})h + \frac{1}{2}y''(\eta_i)(t_i - t_{i+1})^2$$

$$\begin{aligned}
y(t_i) = y(t_{i+1}) - y'(t_{i+1})h + \frac{1}{2}y''(\eta_i)h^2 &\approx y(t_{i+1}) - y'(t_{i+1})h \\
\implies y(t_{i+1}) &= y(t_i) + y'(t_{i+1})h
\end{aligned}$$

32 Forward Euler on Exponential Solution

From $y' = \lambda y$, where $\lambda < 0$

$$y_{i+1} = y_i + h \cdot f(t_i, y_i)$$

$$= y_i + h\lambda y_i = (1 + h\lambda)y_i$$

$$= (1 + h\lambda)^{i+1} y_0$$

For stability, $\|1 + h\lambda\|_2 \leq 1$

$$-1 \leq 1 + h\lambda \leq 1 \implies -2 \leq h\lambda \leq 0$$

Since $\lambda < 0$,

$$h \leq \frac{2}{|\lambda|}$$

33 Backward Euler on Exponential Solution

From $y' = \lambda y$, where $\lambda < 0$,

$$y_{i+1} = y_i + h\lambda y_{i+1}$$

$$y_{i+1}(1 - h\lambda) = y_i \implies y_{i+1} = \frac{y_i}{1 - h\lambda}$$

$$\|y_{i+1}\| = \frac{\|y_i\|}{\|1 - h\lambda\|} \leq \|y_i\|$$

Thus

$$\|1 - h\lambda\| \geq 1$$

This is true for all $h > 0$, since $\lambda < 0$.

34 LTE of Forward Euler's

LTE is error introduced in a single step.

$$d_i \triangleq \frac{y(t_{i+1}) - y(t_i)}{h} - \phi(y(t_i), t_n)/h$$

$$d_i \triangleq \frac{y(t_{i+1}) - y(t_i)}{h} - f(t_i, y_i)$$

Using taylor series about t_i.

$$= \frac{h}{2} f''(\eta_i)$$

35 LTE of Backward Euler's

$$d_i \triangleq \frac{y(t_{i+1}) - y(t_i)}{h} - f(t_{i+1}, y_{t_{i+1}})$$

Using taylor series about t_{i+1}

$$= -\frac{h}{2} f''(\eta_i)$$

36 LTE of Implicit Trapezoidal Rule

Take two second order taylor series, one about t_{i+1} and t_i. Then

$$y(t_{i+1})-y(t_i) = \frac{h}{2}(y'(t_i)+y'(t_{i+1}))+\frac{h^2}{4}(y''(t_i)-y''(t_{i+1}))+\frac{h^3}{12}(y'''(\eta_i)+y'''(\zeta_i))$$

$$\implies d_i = \frac{h}{4}[y''(t_i) - y''(t_{i+1})] + \frac{h^2}{12}[y'''(\eta_i) + y'''(\zeta_i)]$$

Using MVT

$$d_i = \frac{-h^2}{4} y'''(\gamma_i) + \frac{h^2}{12}[y'''(\eta_i) + y'''(\zeta_i)]$$

37 System of ODEs

Consider a diagonalizable matrix A in a system $y' = Ay$

$$A = P\Lambda P^{-1}$$

$$\implies y' = P\Lambda P^{-1} y \implies P^{-1} y' = \Lambda P^{-1} y$$

$$\implies z' = \Lambda z \implies z_i = \lambda_i z$$

Thus for the system to remain stable, the absolute value of the eigenvalues of A must be less than 0.

38 2-Stage RK with Trapezoidal Rule on Exponential Solution

By forward Euler's method

$$Y = (1 + h\lambda) y_i$$

Applying the trapezoidal update rule

$$y_{i+1} = y_i + \frac{h}{2}(f(t_i, y_i) + f(t_{i+1}, Y))$$

$$y_{i+1} = y_i + \frac{h}{2}(\lambda y_i + \lambda(1 + h\lambda) y_i)$$

$$y_{i+1} = y_i(1 + 2\lambda h + \frac{h^2\lambda^2}{2})$$

39 2-Stage RK with Midpoint Rule on Exponential Solution

By forward Euler's method

$$Y = (1 + \frac{h}{2}\lambda)y_i$$

$$y_{i+1} = y_i + hf(t_i + \frac{h}{2}, Y)$$
$$= y_i(1 + h\lambda + \frac{\lambda^2 h^2)}{2}$$

40 Error in Backward Difference Method

$$f(x-h) = f(x) + f'(x)(-h) + \frac{1}{2}f''(\zeta)(-h)^2$$
$$= f(x) - h \cdot f'(x) + \frac{h^2}{2}f''(\zeta)$$
$$\frac{f(x) - f(x-h)}{h} + \frac{h}{2}f''(\zeta) = f'(x)$$

Thus error is O(h), first order.

41 Error in Central Difference Method

Write two taylor series expansions at x_{i+1} and x_{i-1} about x_i.

$$f(x_{i+1}) = f(x_i) + f'(x_i)h + \frac{1}{2}f''(x_i)h^2 + \frac{h^3}{6}f'''(\zeta_+)$$

$$f(x_{i-1}) = f(x_i) - f'(x_i)h + \frac{1}{2}f''(x_i)h^2 - \frac{h^3}{6}f'''(\zeta_-)$$

where $\zeta_+ \in [x_i, x_{i+1}]$ and $\zeta_- \in [x_{i-1}, x_i]$. Adding both gives

$$f'(x_i) = \frac{f(x_{i+1}) - f(x_{i-1})}{2h} + \frac{h^2}{12}[f'''(\zeta_+) + f'''(\zeta_-)]$$

42 Second Derivative using Central Difference

Taking the third order taylors series about x_i evaluated at x_{i+1} and x_{i-1}

$$f(x_{i+1}) = f(x_i) + f'(x_i)h + \frac{h^2}{2}f''(x_i) + \frac{h^3}{6}f'''(x_i) + \frac{h^4}{24}f^{(4)}(\zeta_+)$$

$$f(x_{i-1}) = f(x_i) - f'(x_i)h + \frac{h^2}{2}f''(x_i) - \frac{h^3}{6}f'''(x_i) + \frac{h^4}{24}f^{(4)}(\zeta_+)$$

Adding both

$$f(x_{i+1}) + f(x_{i-1}) = 2f(x_i) + f''(x_i)h^2 + \frac{h^4}{24}[f^{(4)}(\eta_1) + f^{(4)}(\eta_2)]$$

By IVT

$$|f''(x_i) - \frac{f(x_{i+1}) - 2f(x_i) + f(x_{i-1})}{h^2}| \leq \frac{h^2}{12} \max_{\zeta \in [x_{i-1}, x_{i+1}]} |f^{(4)}(\zeta)|$$

43 Taylor Expansion of 2-stage Runge-Kutta Term

Consider a taylor series about t_i

$$f(y(t_{i+1}), t_{i+1}) = f(y_i + hf(y_i, t_{i+1}), t_i + h)$$

$$= f(y_i, t_i) + h\frac{df}{dt}|_{t=t_i, y=y_i} + O(h^2)$$

$$= y'(t_i) + hy''(t_i) + O(h^2)$$

44 Notes

- If slope is decreasing on interval, then forward difference underestimates slope.
- If slope is decreasing on interval, then backward difference overestimates slope.
- If slope is increasing on interval, then forward difference overestimates slope.
- if slope is increasing on interval, then backwards difference underestimates slope.

45 MT1 Problems

45.1 Problem 1

Assume x, y, z are FP numbers. Find the error bound in fl(z(x+y)).

45.2 Problem 2

Show that $V = [\alpha, \alpha] \subseteq R^2$ is a subspace.

45.3 Problem 3

Prove that the set of all eigenvectors sharing eigenvalue λ is a linear subspace.

45.4 Problem 4

Let $x, y \in R$. Find the upper bound on the relative error of fl(fl(x)fl(y)) when compared to x, y.

45.5 Problem 5

Find the bit strings for the following

- -3
- 0.25
- NaN
- Smallest positive normalized
- Unit Roundoff

45.6 Problem 6

Consider the IEEE single floating point system FP(2,24,-126,127).

- What is the smallest positive normalized number in this FP System?
- What is the largest positive denormalized number in this FP System?

45.7 Problem 7

If $x \in$ F, derive a bound on the expression

$$\frac{1}{x+1}$$

46 Answers to Problems

46.1 Problem 1

$$\begin{aligned} fl(z(x+y)) &= z(x+y)(1+\delta_z)(1+\delta_{xy}) \\ &= z(x+y)(1+\delta_{xy}+\delta_z+\delta_z\delta_{xy}) \\ &\approx z(x+y)(1+\delta_{xy}+\delta_z) \end{aligned}$$

Thus the relative error is bounded by:

$$\|\delta_{xy}+\delta_z\| \leq \|\delta_{xy}\| + \|\delta_z\| \leq \frac{2\epsilon_{mach}}{2}$$

46.2 Problem 2

Consider vectors $v_1, v_2 \in V$ and $c, d \in R$.

$$cv_1 + dv_2 =< c\alpha_1 + d\alpha_2, c\alpha_1 + d\alpha_2 >\in V$$

Since V is closed, this means it is a linear subspace of R^2.

46.3 Problem 3

Let v_1, v_2 be two such eigenvectors in R^n and $\alpha, \beta \in R$. $v = \alpha v_1 + \beta v_2$.

$$Av = \alpha Av_1 + \beta Av_2 = \lambda(\alpha v_1 + \beta v_2) = \lambda v$$

This shows that $v \in S$, where S is the set of eigenvectors sharing the eigenvalue λ. Therefore, the set S is a linear subspace of R^n.

46.4 Problem 4

$$\|RE\| = \|\delta_x + \delta_y + \delta_z\| \leq \|\delta_x\| + \|\delta_y\| + \|\delta_z\| = \frac{3\epsilon_{mach}}{2}$$

46.5 Problem 5

- 1 1000 0000 10..0
- 0 0111 1101 0..0
- 0 1111 1111 10..0
- 0 0000 0001 0..0
- 0 0110 1000 0..0

46.6 Problem 6

- $0000000010...0 = 2^{1-127} \approx 1.175 \times 10^{-38}$
- $0000000001...1 = 2^{0-127+1} \times (0.1111)_2 = 1.1..10 \times 2^{-127}$

46.7 Problem 7

Assuming $1 \in$ F

$$fl(\frac{1}{x+1}) = \frac{(1+\delta_1)}{(x+1)(1+\delta_2)}$$

$$\|RE\| = ... = \frac{\|\delta_1 - \delta_2\|}{\|1+\delta_2\|}$$

Since $\|1+\delta_2\| \geq 1 - e_{mach}/2$ and $\|\delta_1 - \delta_2\| \leq \|\delta_1\| + \|\delta_2\| \leq e_{mach}$ by the triangle inequality.

$$\|RE\| \leq \frac{e_{mach}}{1 - e_{mach}/2}$$

47 MT2 Problems

47.1 Problem 1

Find the SVD of $\begin{pmatrix} 5 & 5 \\ -1 & 7 \end{pmatrix}$ [Answer: $\begin{pmatrix} -3/\sqrt{10} & 1\sqrt{10} \\ 1/\sqrt{10} & 3/\sqrt{10} \end{pmatrix}$, $\Sigma = \begin{pmatrix} 2\sqrt{5} & 0 \\ 0 & 4\sqrt{5} \end{pmatrix}$, $\begin{pmatrix} -1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}$]

47.2 Problem 2

Write the Lagrange basis polynomials for the data set: $(1,1),(2,3),(4,3)$

48 Answers

48.1 Problem 1

$$A = U\Sigma V^T \implies AV = U\Sigma$$

Element-wise computation gives U.

48.2 Problem 2

$$L_0(x) = \frac{x^2-6x+8}{3}$$
$$L_1(x) = \frac{-(x^2-5x+4)}{2}$$
$$L_2(x) = \frac{x^2-3x+2}{6}$$