Numerical Analysis MATH50003 (2023–24) Problem Sheet 5

Problem 1(a) Suppose $|\epsilon_k| \leq \epsilon$ and $n\epsilon < 1$. Use induction to show that

$$\prod_{k=1}^{n}(1+\epsilon_k) = 1+\theta_n$$

for some constant θ_n satisfying

$$|\theta_n| \leq \underbrace{\frac{n\epsilon}{1-n\epsilon}}_{E_{n,\epsilon}}$$

SOLUTION

$$\prod_{k=1}^{n+1}(1+\epsilon_k) = \prod_{k=1}^{n}(1+\epsilon_k)(1+\epsilon_{n+1}) = (1+\theta_n)(1+\epsilon_{n+1}) = 1+\underbrace{\theta_n+\epsilon_{n+1}+\theta_n\epsilon_{n+1}}_{\theta_{n+1}}$$

where

$$\begin{aligned}
|\theta_{n+1}| &\leq \frac{n\epsilon}{1-n\epsilon}(1+\epsilon)+\epsilon \\
&= \frac{n\epsilon+n\epsilon^2}{1-(n+1)\epsilon}\underbrace{\frac{1-(n+1)\epsilon}{1-n\epsilon}}_{\leq 1}+\frac{\epsilon-(n+1)\epsilon^2}{1-(n+1)\epsilon} \\
&\leq \frac{(n+1)-\epsilon}{1-(n+1)\epsilon}\epsilon \leq \frac{(n+1)\epsilon}{1-(n+1)\epsilon} = E_{n+1,\epsilon}.
\end{aligned}$$

END

Problem 1(b) Show for an idealised floating point vector $\boldsymbol{x} \in F^n_{\infty,S}$ that

$$x_1 \oplus \cdots \oplus x_n = x_1 + \cdots + x_n + \sigma_n$$

where

$$|\sigma_n| \leq \|\boldsymbol{x}\|_1 E_{n-1,\epsilon_{\mathrm{m}}/2},$$

assuming $n\epsilon_{\mathrm{m}} < 2$ and where

$$\|\boldsymbol{x}\|_1 := \sum_{k=1}^{n}|x_k|.$$

Hint: use the previous part to first write

$$x_1 \oplus \cdots \oplus x_n = x_1(1+\theta_{n-1}) + \sum_{j=2}^{n} x_j(1+\theta_{n-j+1}).$$

SOLUTION

Using Problem 2.1 we write:

$$\begin{aligned}
(\cdots((x_1+x_2)(1+\delta_1)+x_3)(1+\delta_2)\cdots+x_n)(1+\delta_{n-1}) &= x_1\prod_{k=1}^{n-1}(1+\delta_k) + \sum_{j=2}^{n} x_j \prod_{k=j-1}^{n-1}(1+\delta_j) \\
&= x_1(1+\theta_{n-1}) + \sum_{j=2}^{n} x_j(1+\theta_{n-j+1})
\end{aligned}$$

where we have for $j = 2, \ldots, n$

$$|\theta_{n-j+1}| \leq E_{n-j+1,\epsilon_{\mathrm{m}}/2} \leq E_{n-1,\epsilon_{\mathrm{m}}/2}.$$

Thus we have

$$\sum_{j=1}^{n} x_j(1+\theta_{n-j+1}) = \sum_{j=1}^{n} x_j + \underbrace{\sum_{j=1}^{n} x_j\theta_{n-j+1}}_{\sigma_n}$$

where

$$|\sigma_n| \le \sum_{j=1}^{n} |x_j\theta_{n-j+1}| \le \sup_j |\theta_{n-j+1}| \sum_{j=1}^{n} |x_j| \le \|\boldsymbol{x}\|_1 E_{n-1,\epsilon_{\rm m}/2}.$$

END

Problem 1(c) For $A \in F_{\infty,S}^{n\times n}$ and $\boldsymbol{x} \in F_{\infty,S}^{n}$ consider the error in approximating matrix multiplication with idealised floating point: for

$$A\boldsymbol{x} = \begin{pmatrix} \bigoplus_{j=1}^{n} A_{1,j} \otimes x_j \\ \vdots \\ \bigoplus_{j=1}^{n} A_{1,j} \otimes x_j \end{pmatrix} + \delta$$

show that

$$\|\delta\|_\infty \le 2\|A\|_\infty \|\boldsymbol{x}\|_\infty E_{n,\epsilon_{\rm m}/2}$$

where $n\epsilon_{\rm m} < 2$ and the matrix norm is $\|A\|_\infty := \max_k \sum_{j=1}^{n} |a_{kj}|$.

SOLUTION We have for the k=th row

$$\bigoplus_{j=1}^{n} A_{k,j} \otimes x_j = \bigoplus_{j=1}^{n} A_{k,j}x_j(1+\delta_j) = \sum_{j=1}^{n} A_{k,j}x_j(1+\delta_j) + \sigma_{k,n}$$

where we know $|\sigma_n| \le M_k E_{n-1,\epsilon_{\rm m}/2}$, where from 1(b) we have

$$M_k = \Sigma_{j=1}^{n}|A_{k,j}x_j(1+\delta_j)| = \Sigma_{j=1}^{n}|A_{k,j}||x_j|(1+|\delta_j|) \le 2\max|x_j|\Sigma_{j=1}^{n}|A_{k,j}| \le 2\|\boldsymbol{x}\|_\infty\|A\|_\infty$$

Similarly, we also have

$$|\sum_{j=1}^{n} A_{k,j}x_j\delta_j| \le \|\boldsymbol{x}\|_\infty\|A\|_\infty\epsilon_{\rm m}/2$$

and so the result follows from

$$\epsilon_{\rm m}/2 + 2E_{n-1,\epsilon_{\rm m}/2} \le \frac{\epsilon_{\rm m}/2 + \epsilon_{\rm m}(n-1)}{1-(n-1)\epsilon_{\rm m}/2} \le \frac{\epsilon_{\rm m} n}{1-n\epsilon_{\rm m}/2} = 2E_{n,\epsilon_{\rm m}/2}.$$

END

Problem 2 Derive Backward Euler: use the left-sided divided difference approximation

$$u'(x) \approx \frac{u(x)-u(x-h)}{h}$$

to reduce the first order ODE

$$u(a) = c, \qquad u'(x) + \omega(x)u(x) = f(x)$$

to a lower triangular system by discretising on the grid $x_j = a + jh$ for $h = (b-a)/n$. Hint: only impose the ODE on the gridpoints $x_1, \dots, x_n$ so that the divided difference does not depend on behaviour at x_{-1}.

SOLUTION

We go through all 4 steps (this is to help you understand what to do. In an exam I will still give full credit if you get the right result, even if you don't write down all 4 steps):

(Step 1) Since we need to avoid going off the left in step 2 we start the ODE discretisation at x_1:

$$\begin{pmatrix} u(x_0) \\ u'(x_1) \\ \vdots \\ u'(x_n) \end{pmatrix} = \underbrace{\begin{pmatrix} c \\ f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix}}_{\boldsymbol{b}}$$

(Step 2) Replace with divided differences:

$$\begin{pmatrix} u(x_0) \\ (u(x_1) - u(x_0))/h \\ \vdots \\ (u(x_n) - u(x_{n-1}))/h \end{pmatrix} \approx \boldsymbol{b}$$

(Step 3) Replace with discrete system with equality:

$$\begin{pmatrix} u_0 \\ (u_1 - u_0)/h \\ \vdots \\ (u_n - u_{n-1}))/h \end{pmatrix} = \boldsymbol{b}$$

(Step 4) Write as linear system:

$$\begin{bmatrix} 1 & & & \\ -1/h & 1/h & & \\ & \ddots & \ddots & \\ & & -1/h & 1/h \end{bmatrix} \begin{pmatrix} u_0 \\ \vdots \\ u_n \end{pmatrix} = \boldsymbol{b}$$

This is the exact same matrix as Forward Euler but with a different right-hand side.

END

Problem 3 Reduce a Schrödinger equation to a tridiagonal linear system by discretising on the grid $x_j = a + jh$ for $h = (b-a)/n$:

$$u(a) = c, \qquad u''(x) + V(x)u(x) = f(x), \qquad u(b) = d.$$

SOLUTION

(Step 1)

$$\begin{pmatrix} u(x_0) \\ u''(x_1) + V(x_1)u(x_1) \\ \vdots \\ u'(x_{n-1}) + V(x_{n-1})u(x_{n-1}) \\ u(x_n) \end{pmatrix} = \underbrace{\begin{pmatrix} c \\ f(x_1) \\ \vdots \\ f(x_{n-1}) \\ d \end{pmatrix}}_{\boldsymbol{b}}$$

(Step 2) Replace with divided differences:

$$\begin{pmatrix} u(x_0) \\ (u(x_0) - 2u(x_1) + u(x_2))/h^2 + V(x_1)u(x_1) \\ \vdots \\ (u(x_{n-2} - 2u(x_{n-1}) + u(x_n))/h^2 + V(x_{n-1})u(x_{n-1}) \\ u(x_n) \end{pmatrix} \approx \boldsymbol{b}$$

(Step 3) Replace with discrete system with equality:

$$\begin{pmatrix} u_0 \\ (u_0 - 2u_1 + u_2)/h^2 + V(x_1)u_1 \\ \vdots \\ (u_{n-2} - 2u_{n-1} + u_n))/h^2 + V(x_{n-1})u_{n-1} \\ u_n \end{pmatrix} = \boldsymbol{b}$$

(Step 4) Write as a tridiagonal linear system:

$$\begin{bmatrix} 1 & & & & & \\ -1/h^2 & V(x_1) - 2/h^2 & 1/h^2 & & & \\ & \ddots & \ddots & \ddots & & \\ & & -1/h^2 & V(x_{n-1}) - 2/h^2 & 1/h^2 \\ & & & & 1 \end{bmatrix} \begin{pmatrix} u_0 \\ \vdots \\ u_n \end{pmatrix} = \boldsymbol{b}$$

END