Assignment 1

September 5, 2023 1:01 PM

1. Given $x \in \mathbb{R}$, truncation to k digits is equivalent to truncating to k-1 digits after the decimal due to the machine representation of the mantissa.

The truncation of x (denoted $\tilde{x}$) can thus be written as $\tilde{x} = \frac{\lfloor \beta^{k-1} x \rfloor}{\beta^{k-1}}$.

It should also be noted that $x \leq \frac{\lceil \beta^{k-1} x \rceil}{\beta^{k-1}}$. Therefore $x - \tilde{x} \leq \frac{\lceil \beta^{k-1} x \rceil}{\beta^{k-1}} - \frac{\lfloor \beta^{k-1} x \rfloor}{\beta^{k-1}} = \frac{1}{\beta^{k-1}}\left(\lceil \beta^{k-1} x \rceil - \lfloor \beta^{k-1} x \rfloor\right)$.

Since $\forall a \in \mathbb{R}\ \lceil a \rceil - \lfloor a \rfloor \leq 1$ and $\forall x \in \mathbb{R}\ x - \tilde{x} \leq \beta^{1-k}$, $\varepsilon_{mach} = \beta^{1-k}$. □

2. $x_1 + dx_2 = 1$ & $dx_1 + x_2 = 0 \implies \begin{pmatrix} 1 & d \\ d & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \implies \begin{pmatrix} 1 & d \\ 0 & 1-d^2 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ -d \end{pmatrix} \implies \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 + \frac{d^2}{1-d^2} \\ -d/_{1-d^2} \end{pmatrix}$.

This means the problem is well posed. Now for the condition number:

$$\text{cond} = \frac{\|f(d+\Delta d) - f(d)\|_\infty}{\|f(d)\|_\infty} \frac{\|d\|_\infty}{\|\Delta d\|_\infty}$$

$$= \left\| \begin{pmatrix} \frac{1}{1-(d+\Delta d)^2} \\ \frac{-(d+\Delta d)}{1-(d+\Delta d)^2} \end{pmatrix} - \begin{pmatrix} \frac{1}{1-d^2} \\ \frac{-d}{1-d^2} \end{pmatrix} \right\|_\infty \frac{|d|}{|\Delta d|\left(\frac{1}{1-d^2}\right)}$$

$$= \frac{|d|(1-d^2)}{|\Delta d|} \left| \frac{1}{(1-(d+\Delta d)^2)(1-d^2)} \right| \left\| \begin{pmatrix} 1-d^2-1+(d+\Delta d)^2 \\ -(1-d^2)(d+\Delta d) + d(1-(d+\Delta d)^2) \end{pmatrix} \right\|_\infty$$

$$= \frac{|d|}{|\Delta d|(1-(d+\Delta d)^2)} \left\| \begin{matrix} 2d\Delta d + \Delta d^2 \\ -(d+\Delta d - d^3 - d^2\Delta d) + (d - d^3 - 2d^2\Delta d - d^2\Delta d^2) \end{matrix} \right\|_\infty$$

$$= \frac{|d|}{|\Delta d|(1-(d+\Delta d)^2)} \left\| \begin{matrix} (2d+\Delta d)\Delta d \\ -\Delta d - d^2\Delta d - d^2\Delta d^2 \end{matrix} \right\|_\infty$$

$$= \frac{|d|}{1-(d+\Delta d)^2} \left\| \begin{matrix} 2d+\Delta d \\ 1+d^2(1+\Delta d) \end{matrix} \right\|_\infty$$

Taking $\lim_{\Delta d \to 0}$, since $|1+d^2| \geq |2d|$, $\text{cond} = \frac{|d|(1+d^2)}{1-d^2}$. □

Since cond is symmetric & monotonic on $(0,1)$, binary search can give $\boxed{|d| < 0.9125\ldots}$ (see q2.m)

3. a. See ForwardSubstitute.m

b.
```
function BackwardSubstitution(U,b)
    x_n ← b_n/U_nn
    for i=n-1 to 1 do
        s ← b_i
        for j=i+1 to n do
            s ← s - U_ij x_j
        x_i ← s/U_ii
    return x
```

```
7         x_i ← s/U_ii
8     return x
```

c. Line 2: 1 division = 1 FLOPs
  Line 6: $(n-i-1)$ multiplications & subtractions $= 2(n-i-1)$ FLOPs
  Line 7: 1 division = 1 FLOPs
  Total FLOPs:

$$T = 1 + \sum_{i=1}^{n-1} (2n-2i+1) = 1 + (2n+1)(n-1) - n(n-1)$$

$$= n^2 \in O(n^2)$$

d. See BackwardSubstitution.m

e. See LUDecomposition.m

f. See LUSolve.m

4. a. For brevity let $p_i \equiv p(t_i)$, $q_i \equiv q(t_i)$, $b_i \equiv b(t_i)$ for $i = 0, \ldots, n+1$. Using (2)&(3), (1) becomes:

$$\frac{x_{i+1} - 2x_i + x_{i-1}}{h^2} = p_i \frac{x_{i+1} - x_{i-1}}{2h} + q_i x_i + b_i, \quad x_0 = \alpha, x_{n+1} = \beta.$$

$$x_{i+1}\left(\frac{1}{h^2} - \frac{p_i}{2h}\right) - x_i\left(\frac{2}{h^2} + q_i\right) + x_{i-1}\left(\frac{1}{h^2} + \frac{p_i}{2h}\right) - b_i = 0$$

$$x_{i+1}\left(\frac{h}{2}p_i - 1\right) + x_i(h^2 q_i + 2) - x_{i-1}\left(\frac{h}{2}p_i + 1\right) = -h^2 b_i \Longrightarrow \boxed{\begin{aligned} a_{i,i-1} &= -\frac{h}{2}p_i - 1 \\ a_{i,i} &= h^2 q_i + 2 \\ a_{i,i+1} &= \frac{h}{2}p_i - 1 \end{aligned}}$$

b. $L = \|p\|_\infty = \max_{t \in [0,1]} |p(t)|$, $hL < 2$. A is strictly row diagonal dominant when

$$|a_{i,i}| > \sum_{j \neq i} |a_{i,j}| = |a_{i,i-1}| + |a_{i,i+1}|$$

$$|h^2 q_i + 2| > \left|-\frac{h}{2}p_{i+1} - 1\right| + \left|\frac{h}{2}p_{i-1} - 1\right| = 2 \quad \left(\text{since } \max_{t \in [0,1]} |p(t)| = L, \frac{hL}{2} < 1, \text{ and } |-t-1| + |t-1| = 2 \ \forall |t| < 1\right)$$

Since $h^2 > 0$ and $q_i > 0$ the above always holds, so A is strictly row diagonal dominant. □

c. $L = 20$ so $h < \frac{1}{10}$ and $\boxed{n_0 = 10}$. There does appear to be convergence as $n \to \infty$ (see q4c.png)

d. $B_J = I - D^{-1}A \Longrightarrow B_{J_{i,i}} = 0$, $B_{J_{i,i-1}} = \dfrac{-\frac{h}{2}p_i - 1}{h^2 q_i + 2}$, $B_{J_{i,i+1}} = \dfrac{\frac{h}{2}p_i - 1}{h^2 q_i + 2}$

$$\Longrightarrow \|B_J\|_\infty = \max_{i \in \{1,\ldots,n\}} \left(\frac{2}{h^2 q_i + 2}\right) \Longrightarrow \boxed{\|B_J\|_\infty < \frac{2}{h^2 q_{min} + 2}}$$

Since $q_{min} > 0$ & $h > 0$ $\frac{2}{h^2 q_i + 2} \in (0,1)$ so it always converges in this case. □

$\boxed{\lim_{h \to 0} \|B_J\|_\infty = 1}$ so it gets very slow as $h \to 0$.

e. This took $\boxed{102958 \text{ iterations } (\|B_J\|_\infty \approx 0.99992)}$, LU is way faster here. When $q = 10^8(t^2+1)$, $p = \frac{1}{20}\cos(\pi t)$ it took $\boxed{7 \text{ iterations}}$. This is due to the fact that $h^2 q_{min} \sim 10^0$ so $\boxed{\|B_J\|_\infty \approx 1/9.}$ (see q4e1.png & q4e2.png) The LU method was slower here since it had to perform $O(n^2)$ FLOPs while Jacobi effectively only needed 7 $O(n)$ operations.

5. a. First note that $f(x) = \begin{pmatrix} 1 \\ x \\ x^2 \end{pmatrix} \cdot c$. If $y_i = f(x_i) \ \forall i \in \{1, \ldots, n\}$, then $\begin{pmatrix} 1 & x_1 & x_1^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ is required.

Since M has more rows than columns, it must be overdetermined. □

$\ldots x^2) \ldots y_i \ldots \begin{pmatrix} \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix} \begin{pmatrix} \vdots \\ c_2 \end{pmatrix} = \begin{pmatrix} \vdots \\ y_n \end{pmatrix} \ldots$

Since M has more rows than columns, it must be overdetermined. $\square$

b. $A^T = (M^T M)^T = M^T (M^T)^T = M^T M = A$ $\square$

c. $Mc = y \implies M^T M c = M^T y \implies \boxed{b = Ay}$

d. See LeastSquaresQuadratic.m.

e. The best fit quadratic was $y = -19.4257 + 5.7451x - 0.2841x^2$ with $\|\tilde{y} - \tilde{y}\| = 2.347726$. (see q5e.m).