Maths 3018/6111 - Numerical Methods

Worksheet 2 - Solutions

Theory

1. Perform the LU decomposition of
$$A = \begin{pmatrix} 1 & 3 \\ 4 & 16 \end{pmatrix}.$$
Use both standard factorisation methods.

Remember the general algorithm from Table 2.1 in the notes.
We first assume the Doolittle factorization ($\ell_{kk} = 1$). We immediately get that

$$\begin{aligned} u_{11} &= \left(a_{11} - \sum_{s=1}^{0} \ell_{1s} u_{s1} \right) / \ell_{11} \qquad \text{(using } k = 1\text{)} \\ &= 1. \end{aligned}$$

We then build the first row of U to get

$$\begin{aligned} u_{12} &= \left(a_{12} - \sum_{s=1}^{0} \ell_{1s} u_{s2} \right) / \ell_{11} \qquad \text{(using } k = 1\text{)} \\ &= 3. \end{aligned}$$

We then build the first column of L to get

$$\begin{aligned} \ell_{21} &= \left(a_{21} - \sum_{s=1}^{0} \ell_{2s} u_{s1} \right) / \ell_{11} \qquad \text{(using } k = 1\text{)} \\ &= 4. \end{aligned}$$

Given the factorization we have that $\ell_{22} = 1$ and hence that

$$\begin{aligned} u_{22} &= \left(a_{22} - \sum_{s=1}^{1} \ell_{2s} u_{s2} \right) / \ell_{22} \qquad \text{(using } k = 2\text{)} \\ &= (16 - 4 \times 3) \\ &= 4. \end{aligned}$$

Hence we have

$$L = \begin{pmatrix} 1 & 0 \\ 4 & 1 \end{pmatrix}, \qquad U = \begin{pmatrix} 1 & 3 \\ 0 & 4 \end{pmatrix}.$$

Quickly check that

$$\begin{aligned} LU &= \begin{pmatrix} 1 & 0 \\ 4 & 1 \end{pmatrix} \begin{pmatrix} 1 & 3 \\ 0 & 4 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 3 \\ 4 & 16 \end{pmatrix} \\ &= A. \end{aligned}$$

With the Crout factorization ($u_{kk} = 1$), we immediately get that

$$\begin{aligned}\ell_{11} &= \left(a_{11} - \sum_{s=1}^{0} \ell_{1s} u_{s1}\right) / u_{11} \qquad \text{(using } k = 1\text{)} \\ &= 1.\end{aligned}$$

We then build the first row of U to get

$$\begin{aligned}u_{12} &= \left(a_{12} - \sum_{s=1}^{0} \ell_{1s} u_{s2}\right) / \ell_{11} \qquad \text{(using } k = 1\text{)} \\ &= 3.\end{aligned}$$

We then build the first column of L to get

$$\begin{aligned}\ell_{21} &= \left(a_{21} - \sum_{s=1}^{0} \ell_{2s} u_{s1}\right) / \ell_{11} \qquad \text{(using } k = 1\text{)} \\ &= 4.\end{aligned}$$

Given the factorization we have that $u_{22} = 1$ and hence that

$$\begin{aligned}\ell_{22} &= \left(a_{22} - \sum_{s=1}^{1} \ell_{2s} u_{s2}\right) / u_{22} \qquad \text{(using } k = 2\text{)} \\ &= (16 - 4 \times 3) \\ &= 4.\end{aligned}$$

Hence we have

$$L = \begin{pmatrix} 1 & 0 \\ 4 & 4 \end{pmatrix}, \qquad U = \begin{pmatrix} 1 & 3 \\ 0 & 1 \end{pmatrix}.$$

Quickly check that

$$\begin{aligned}LU &= \begin{pmatrix} 1 & 0 \\ 4 & 4 \end{pmatrix} \begin{pmatrix} 1 & 3 \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 3 \\ 4 & 16 \end{pmatrix} \\ &= A.\end{aligned}$$

Note that in this case only one entry has changed; in general the different factorizations will give totally different answers.

---

2. [Additional] Write out the Thomas algorithm for a tridiagonal system.

We assume that the tridiagonal matrix A has diagonal entries $b_i, i = 1, \ldots, N$, sub-diagonal entries $a_j, j = 1, \ldots, N-1$, and super-diagonal entries $c_j, j = 1, \ldots, N-1$. We assume that we are trying to solve the problem

$$A\mathbf{x} = \mathbf{f},$$

which has size N.

The Thomas algorithm is essentially Gaussian elimination; forward elimination is used to get an upper triangular problem, and then back substitution used to find the answer. The tridiagonal nature of the problem means that it is sufficiently simple to write it out explicitly.

The first step is the forward elimination process. This yields the system

$$B\mathbf{x} = \mathbf{y},$$

where B has entries only on the diagonal, and these entries are all 1, and on the super-diagonal, which are written as $c_j/d_j, j = 1, \ldots, N-1$. The vectors $\mathbf{d}, \mathbf{y}$ are given by the procedure

(a) At the first step $d_1 = b_1$ and $y_1 = f_1/d_1$;

(b) At the k-th step $d_k = b_k - a_{k-1}c_{k-1}/d_{k-1}$ and $y_k = (f_k - y_{k-1}a_{k-1})/d_k$.

We can then solve this simplified system using back substitution to find

(a) At the first step $x_N = y_N$;

(b) In reverse order, $k = N-1, \ldots, 1$ we have $x_k = y_k - x_{k+1}c_k/d_k$.

---

3. Write down the general framework for iterative methods for linear systems. Give the convergence matrix. If the linear system uses the matrix A above, will an iterative method converge? [Hint: remember what to do with the diagonal entries]

The system to be solved is

$$A\mathbf{x} = \mathbf{b},$$

with $\det(A) \neq 0$. *We first scale the problem such that all diagonal entries of A are 1.* We split the coefficient matrix A into the matrices N, P, such that

$$A = N - P$$

where $\det(N) \neq 0$. We therefore have that

$$N\mathbf{x} = P\mathbf{x} + \mathbf{b}$$

and we use this to write an iteration scheme as

$$N\mathbf{x}^{(n)} = P\mathbf{x}^{(n-1)} + \mathbf{b}, \qquad n = 1, 2, 3, \ldots$$

where we start from some arbitrary initial guess $\mathbf{x}^{(0)}$.

The convergence matrix M is given by $M = N^{-1}P$. Convergence is guaranteed if the spectral radius of M is less than one. For the matrix given in the first part we have the *rescaled* A is

$$A = \begin{pmatrix} 1 & 3 \\ 1/4 & 1 \end{pmatrix}.$$

Clearly to talk about convergence we need to first decide how to split A into N, P matrices. For simplicity we consider Jacobi's method for which $N = I$ and hence we have

$$\begin{aligned} N &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \\ P &= \begin{pmatrix} 0 & -3 \\ -1/4 & 0 \end{pmatrix}, \\ M &= N^{-1}P \\ &= P \\ &= \begin{pmatrix} 0 & -3 \\ -1/4 & 0 \end{pmatrix}, \\ \Rightarrow \qquad \rho(M) &= \max_i |\lambda_i| \\ &= \max\left\{\pm\sqrt{-3/4}\right\} \\ &= \sqrt{3/4} \\ &< 1. \end{aligned}$$

Therefore Jacobi will converge for this matrix.

---

4. Check which of the matrices on this sheet are diagonally dominant.

For (strict) diagonal dominance we need the absolute value of the diagonal coefficient to be (strictly) greater than the sum of the absolute values of all other coefficients in the row.

For A this is not true; for the first row we have $1 \not> 3$, although dominance holds for the second row.

For B it is not true; for the second row we have $28 \not> 24 + 53$, although dominance holds for the other rows.

For C it is not true; for the first row it fails, although dominance holds for the other rows.

---

5. Briefly explain what is meant by quadrature methods based on polynomial interpolation.

Standard exam question; see, e.g., 07/08.

The function $f(x)$ to be integrated is replaced by an interpolating function, in this case a polynomial of degree n, that interpolates it at $n + 1$ *nodes* x_j, $j = 0, 1, \ldots, n$. In general, when constructing a compound quadrature formula, we may split the interval into subintervals, each subinterval containing $n+1$ nodes, and use a polynomial interpolating function on each subinterval. The integral of the function is approximated by the integral of the (piecewise polynomial) interpolating function.

---

6. [3018 only] Write down the contraction mapping theorem. Check that $g(x) = \cos(x)$ is contracting on the unit interval.

The contraction mapping theorem is

> If $g(x)$ is a contraction mapping in an interval $I = [a, b]$ then there exists one and only one fixed point of the map in $[a, b]$.

In order to state we need the definition of a contraction mapping, which is

> **Definition** - A continuous map $g(x)$ from an interval $I = [a, b] \subseteq \mathbb{R}$ into $\mathbb{R}$ is *contracting* (or a *contraction mapping*) if
>
> (a) the image of I is contained in I:
>
> $$\begin{aligned} & g(I) \subseteq I \\ \Leftrightarrow \quad & g(x) \in I \, \forall x \in I. \end{aligned}$$
>
> (b) the function $g(x)$ is Lipschitz continuous in I with Lipschitz constant $L < 1$:
>
> $$|g(x) - g(y)| \leq L|x - y| \quad \forall x, y \in I.$$

Given this we want to check $g(x) = \cos(x)$ on $I = [0, 1]$. We have that g is continuous and differentiable, so it must be Lipschitz continuous. We know that $L < \max_{x \in I} |g'(x)|$ so we look at $g' = -\sin(x)$. We note that the extrema of g' occur when $g'' = -\cos(x) = 0$ which is at $x = (n + 1/2)\pi$, of which none occur in the interval $[0, 1]$. We also note that g' is monotonic on the interval. Therefore the extreme values of g' are taken at the ends of the interval and are $[0, -\sin(1)]$, which are both less than 1 in absolute value. So $L < 1$.

We also need to check that $g(I) \subseteq I$. Again we note that g is monotonic on the interval. Hence we only need to check that $g(0) = 1 \in I$ and $g(1) = \cos(1) \in I$, both of which hold. Therefore the map is contracting.