Dr H. Fawzi

Mathematical Tripos Part II: Michaelmas Term 2023

Numerical Analysis – Lecture 5

Stability, consistency and the Lax equivalence theorem Suppose that a numerical method for a partial differential equation of evolution can be written in the form[1]

$$\boldsymbol{u}^{n+1} = A_h \boldsymbol{u}^n,$$

where $\boldsymbol{u}^n \in \mathbb{R}^M$, $A_h \in \mathbb{R}^{M\times M}$ is a matrix, and $h = \frac{1}{M+1}$. Fix a norm $\|\cdot\|$ on $\mathbb{R}^M$, and let $\|A_h\| = \sup \frac{\|A_h \boldsymbol{x}\|}{\|\boldsymbol{x}\|}$ be the corresponding induced matrix norm. If we define *stability* as preserving the boundedness of $\boldsymbol{u}^n$ with respect to the norm $\|\cdot\|$, then since

$$\|\boldsymbol{u}^n\| \le \|A_h^n \boldsymbol{u}^0\| \le \|A_h\|^n \|\boldsymbol{u}^0\|,$$

we get:

$$\|A_h\| \le 1 \text{ as } h \to 0 \quad \Rightarrow \quad \text{the method is stable.}$$

If we denote the exact solution of the PDE by $\widehat{u}(x,t)$ and let $\widehat{\boldsymbol{u}}^n = (\widehat{u}(mk, nt))_{1\le m\le M}$, then we have $\widehat{\boldsymbol{u}}^{n+1} = A_h \widehat{\boldsymbol{u}}^n + \boldsymbol{\eta}^n$ where $\boldsymbol{\eta}^n$ is the local truncation error. The error vector $\boldsymbol{e}^n = \widehat{\boldsymbol{u}}^n - \boldsymbol{u}^n$ satisfies

$$\boldsymbol{e}^{n+1} = A_h \boldsymbol{e}^n + \boldsymbol{\eta}^n.$$

Using $\|A_h\| \le 1$ and assuming $\|\boldsymbol{e}^0\| = 0$, we get $\|\boldsymbol{e}^n\| \le \|\boldsymbol{\eta}^{n-1}\| + \cdots + \|\boldsymbol{\eta}^0\|$. If *consistency* holds, i.e., $\|\boldsymbol{\eta}^n\| = O(k^2)$, then we see that $\|\boldsymbol{e}^n\| \le nck^2$ for some constant $c > 0$. Since $n \le T/k$ we end up with $\|\boldsymbol{e}^n\| \le cTk$, and so $\|\boldsymbol{e}^n\| \to 0$ as $k \to 0$ uniformly in $n \in [1, T/k]$. This shows convergence.

We have thus arrived at the *Lax equivalence theorem*: "consistency + stability = convergence" (more precisely what we have proved here is the implication $\implies$).

Norms The discussion above involves a choice of norm on $\mathbb{R}^M$. There are two standard choices of norms:

- *Sup-norm.* Here, we choose

$$\|\boldsymbol{u}\| = \|\boldsymbol{u}\|_\infty = \max_{i=1,\dots,M} |u_i|.$$

  It can be easily shown that the corresponding induced norm for a matrix $A \in \mathbb{R}^{M\times M}$ is given by:

$$\|A\|_{\infty\to\infty} := \sup_{\boldsymbol{x}} \frac{\|A\boldsymbol{x}\|_\infty}{\|\boldsymbol{x}\|_\infty} = \max_{i=1,\dots,M} \sum_{j=1}^M |A_{ij}|.$$

  This the choice of norm we implicitly used in the convergence proof of Theorem 2.1 (Lecture 4). The matrix in this case was

$$A_h = \begin{bmatrix} 1-2\mu & \mu & & \\ \mu & \ddots & \ddots & \\ & \ddots & \ddots & \mu \\ & & \mu & 1-2\mu \end{bmatrix},$$

  for which we get $\|A_h\|_{\infty\to\infty} = |1-2\mu| + 2\mu \le 1$ if $\mu \le 1/2$.

- *Normalized Euclidean norm.* Another common of choice of norm is the normalized Euclidean length, namely,

$$\|\boldsymbol{u}\| := \sqrt{\tfrac{1}{M} \textstyle\sum_{i=1}^M |u_i|^2}.$$

[1] Assuming zero boundary conditions

The reason for the factor $\frac{1}{M}$ is to ensure that, because of the convergence of Riemann sums, we obtain

$$\|\boldsymbol{u}\| := \left[\frac{1}{M}\sum_{i=1}^{M}|u_i|^2\right]^{1/2} \to \left[\int_0^1 |u(x)|^2 \mathrm{d}x\right]^{1/2} =: \|u\|_{L_2} \qquad (h = 1/(M+1) \to 0),$$

The induced matrix norm in this case is the *spectral norm* (or the *operator norm*) and is denoted $\|A\|_2$:[2]

$$\|A\|_2 := \sup_{\boldsymbol{x}} \frac{\|A\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2}.$$

The spectral norm of A is equal to the largest singular value of A. Equivalently, we can write $\|A\|_2 = [\rho(A\bar{A}^T)]^{1/2}$ where ρ is the spectral radius:

$$\rho(M) := \max\{|\lambda| : \lambda \text{ eigenvalue of } M\}.$$

For certain matrices, such as normal matrices, one can show that $\|A\|_2 = \rho(A)$.

Definition 1.19 (Normal matrices) A complex matrix $A \in \mathbb{C}^{M\times M}$ is *normal* if it commutes with its conjugate transpose, i.e., $A\bar{A}^T = \bar{A}^T A$.

Examples of real normal matrices include symmetric matrices ($A = A^T$) and skew-symmetric matrices ($A = -A^T$). Any normal matrix A can be diagonalized in an orthonormal basis, i.e., $A = QD\bar{Q}^T$ where Q unitary, $Q\bar{Q}^T = \bar{Q}^T Q = I$, and D is diagonal. Note however that the diagonal elements D_{ii} are not necessarily real!

Proposition 1.20 *If A is normal, then $\|A\|_2 = \rho(A)$.*

Proof. Let $\boldsymbol{u}$ be any vector. We can expand it in the basis of the orthonormal eigenvectors $\boldsymbol{u} = \sum_{i=1}^{n} a_i \boldsymbol{q}_i$. Then $A\boldsymbol{u} = \sum_{i=1}^{n} \lambda_i a_i \boldsymbol{q}_i$, and since $\boldsymbol{q}_i$ are orthonormal, we obtain

$$\|A\|_2 := \sup_{\boldsymbol{u}} \frac{\|A\boldsymbol{u}\|_2}{\|\boldsymbol{u}\|_2} = \sup_{a_i} \frac{\{\sum_{i=1}^{M}|\lambda_i a_i|^2\}^{1/2}}{\{\sum_{i=1}^{M}|a_i|^2\}^{1/2}} = |\lambda_{\max}|.$$

Example 1.21 We can analyze the stability of [(2.2), Lecture 4] using the eigenvalue methods just described. The recurrence (2.2) can be written as:

$$u_m^{n+1} = u_m^n + \mu\left(u_{m-1}^n - 2u_m^n + u_{m+1}^n\right), \qquad m = 1...M,$$

in the matrix form

$$\boldsymbol{u}_h^{n+1} = A_h \boldsymbol{u}_h^n, \qquad A_h = I + \mu A_*, \qquad A_* = \begin{bmatrix} -2 & 1 & & \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & -2 \end{bmatrix}_{M\times M}.$$

Here A_* is Toeplitz, symmetric, tridiagonal (TST), with $\lambda_\ell(A_*) = -4\sin^2\frac{\pi\ell h}{2}$, hence $\lambda_\ell(A_h) = 1 - 4\mu\sin^2\frac{\pi\ell h}{2}$, so that its spectrum lies within the interval $[\lambda_M, \lambda_1] = [1-4\mu\cos^2\frac{\pi h}{2}, 1-4\mu\sin^2\frac{\pi h}{2}]$. Since A_h is symmetric, we have

$$\|A_h\|_2 = \rho(A_h) = \begin{cases} |1-4\mu\sin^2\frac{\pi h}{2}| \le 1, & \mu \le \frac{1}{2}, \\ |1-4\mu\cos^2\frac{\pi h}{2}| > 1, & \mu > \frac{1}{2} \quad (h \le h_\mu). \end{cases}$$

We recover the fact that the method is stable for $\mu \le 1/2$.

[2] Note that if $\|\cdot\|$ is the normalized Euclidean norm, then $\|A\boldsymbol{x}\|/\|\boldsymbol{x}\| = \|A\boldsymbol{x}\|_2/\|\boldsymbol{x}\|_2$ where $\|x\|_2 = (\sum_i |x_i|^2)^{1/2}$ is the usual (unnormalized) Euclidean norm