Mathematical Tripos Part II: Michaelmas Term 2014

Numerical Analysis – Lecture 10

Example 2.33 Consider the general diffusion equation

$$\frac{\partial u}{\partial t} = \nabla^\top (a(x,y)\nabla u) + f(x,y) = \frac{\partial}{\partial x}\left(a(x,y)\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(a(x,y)\frac{\partial u}{\partial y}\right) + f(x,y), \tag{2.20}$$

where $a(x,y) > 0$ and $f(x,y)$ are given, together with initial conditions on $[0,1]^2$ and Dirichlet boundary conditions along $\partial[0,1]^2 \times [0,\infty)$. Replace each space derivative by *central differences* at midpoints,

$$\frac{\mathrm{d}g(\xi)}{\mathrm{d}\xi} \approx \frac{g(\xi+\frac{1}{2}h) - g(\xi-\frac{1}{2}h)}{h},$$

resulting in the ODE system

$$\begin{aligned} u'_{\ell,m} \quad = \quad & \tfrac{1}{h^2}\Big[a_{\ell-\frac{1}{2},m}u_{\ell-1,m} + a_{\ell+\frac{1}{2},m}u_{\ell+1,m} + a_{\ell,m-\frac{1}{2}}u_{\ell,m-1} + a_{\ell,m+\frac{1}{2}}u_{\ell,m+1} \\ & \quad - \big(a_{\ell-\frac{1}{2},m} + a_{\ell+\frac{1}{2},m} + a_{\ell,m-\frac{1}{2}} + a_{\ell,m+\frac{1}{2}}\big)u_{\ell,m}\Big] + f_{\ell,m}. \end{aligned} \tag{2.21}$$

The system (2.21) can be solved by an implicit ODE method, e.g. Crank–Nicolson, except that this requires a costly solution of a large algebraic system in each time step.

Intermezzo 2.34 (Linear systems of ODEs) The system (2.21) is linear and (assuming for the time being zero boundary conditions and $f \equiv 0$) homogeneous. With greater generality, let us consider the ODE system

$$\boldsymbol{y}' = A\boldsymbol{y}, \qquad \boldsymbol{y}(0) = \boldsymbol{y}_0. \tag{2.22}$$

We define formally a *matrix exponential* by Taylor series, $\mathrm{e}^B = \sum_{k=0}^\infty \frac{1}{k!}B^k$, and easily verify by formal differentiation that $\mathrm{d}\mathrm{e}^{tA}/\mathrm{d}t = A\mathrm{e}^{tA}$, therefore $\boldsymbol{y}(t) = \mathrm{e}^{tA}\boldsymbol{y}_0$.

In fact, one observes that one-step methods for ODEs, in a linear case, are approximating a matrix exponential. Thus, with $k = \Delta t$,

$$\begin{array}{lll} \text{Euler:} & \boldsymbol{y}_n = (I+kA)^n\boldsymbol{y}_0, & 1+z = \mathrm{e}^z + \mathcal{O}(z^2); \\ \text{TR:} & \boldsymbol{y}_n = \left[\left(I-\frac{1}{2}kA\right)^{-1}\left(I+\frac{1}{2}kA\right)\right]^n \boldsymbol{y}_0, & \frac{1+\frac{1}{2}z}{1-\frac{1}{2}z} = \mathrm{e}^z + \mathcal{O}(z^3). \end{array}$$

Technique 2.35 (Splitting methods) Going back to (2.21), we *split* $A = A_x + A_y$, so that A_x and A_y are constructed from the contribution of discretizations in the x and y directions respectively (similarly to Technique 2.32). In other words, A_x includes all the $a_{\ell\pm\frac{1}{2},m}$ terms and A_y consists of the remaining $a_{\ell,m\pm\frac{1}{2}}$ components. Note that, if the grid is ordered by columns, A_y is tridiagonal, and if the grid is ordered by rows, A_x is tridiagonal. Recall that, for $z_1, z_2 \in \mathbb{C}$, we have $\mathrm{e}^{z_1+z_2} = \mathrm{e}^{z_1}\mathrm{e}^{z_2}$ and suppose for a moment that this property extends to matrices, i.e. that $\mathrm{e}^{tA} = \mathrm{e}^{t(B+C)} = \mathrm{e}^{tB}\mathrm{e}^{tC}$. Had this been true, we could have approximated each component with the trapezoidal rule, say, to produce

$$\boldsymbol{u}^{n+1} = \big(I - \tfrac{1}{2}\mu A_x\big)^{-1}\big(I + \tfrac{1}{2}\mu A_x\big)\big(I - \tfrac{1}{2}\mu A_y\big)^{-1}\big(I + \tfrac{1}{2}\mu A_y\big)\boldsymbol{u}^n, \qquad \mu = k/h^2. \tag{2.23}$$

The advantage of (2.23) lies in the fact that (up to a known permutation) both $I - \frac{1}{2}\mu A_x$ and $I - \frac{1}{2}\mu A_y$ are tridiagonal, hence can be solved very cheaply.

Unfortunately, the assumption that $\mathrm{e}^{t(B+C)} = \mathrm{e}^{tB}\mathrm{e}^{tC}$ is, in general, false. [*Note*: It is true, however, for $a(x,y) \equiv$ const, for in this case A_x and A_y commute, cf. Technique 2.32.] Not all hope is lost, though, and we will demonstrate that, suitably implemented, splitting is a powerful technique to reduce drastically the expense of numerical solution.

Method 2.36 (Splitting) Comparing the Taylor expansions of $\mathrm{e}^{t(B+C)}$ with $\mathrm{e}^{tB}\mathrm{e}^{tC}$ we obtain

$$\mathrm{e}^{tB}\mathrm{e}^{tC} = \mathrm{e}^{t(B+C)} + \tfrac{1}{2}t^2(BC - CB) + \mathcal{O}(t^3). \tag{2.24}$$

In particular, $\mathrm{e}^{tB}\mathrm{e}^{tC} = \mathrm{e}^{t(B+C)}$ for all $t \geq 0$ if and only if B and C commute. The good news is, however, that approximating $\mathrm{e}^{\Delta t(B+C)}$ with $\mathrm{e}^{\Delta tB}\mathrm{e}^{\Delta tC}$ incurs an error of $\mathcal{O}((\Delta t)^2)$. So, if r is a rational function such that $r(z) = \mathrm{e}^z + \mathcal{O}(z^2)$, then

$$\boldsymbol{u}^{n+1} = r(\mu A_x) r(\mu A_y) \boldsymbol{u}^n \tag{2.25}$$

produces an error of $\mathcal{O}((\Delta t)^2)$. The choice $r(z) = (1+\frac{1}{2}z)/(1-\frac{1}{2}z)$ results in a *split Crank–Nicolson* scheme, whose implementation reduces to a solution of tridiagonal algebraic linear systems.

It is easy to prove that

$$\mathrm{e}^{t(B+C)} = \frac{1}{2}\left(\mathrm{e}^{tB}\mathrm{e}^{tC} + \mathrm{e}^{tC}\mathrm{e}^{tB}\right) + \mathcal{O}(t^3), \qquad \mathrm{e}^{t(B+C)} = \mathrm{e}^{\frac{1}{2}tB}\mathrm{e}^{tC}\mathrm{e}^{\frac{1}{2}tB} + \mathcal{O}(t^3),$$

the second formula is called the *Strang splitting*. Thus, as long as $r(z) = \mathrm{e}^z + \mathcal{O}(z^3)$, the time-stepping formula $\boldsymbol{u}^{n+1} = r\big(\frac{1}{2}\mu A_x\big)\, r\big(\mu A_y\big)\, r\big(\frac{1}{2}\mu A_x\big)\, \boldsymbol{u}^n$ carries a local error of $\mathcal{O}((\Delta t)^3)$.

As far as stability is concerned, we observe that both A_x and A_y are symmetric, hence normal, therefore so are $r(\mu A_x)$ and $r(\mu A_y)$. Then Euclidean (L_2)-norm equals the spectral radius, therefore for the splitting (2.25), we have

$$\|\boldsymbol{u}^{n+1}\| \leq \|r(\mu A_x)\| \cdot \|r(\mu A_y)\| \cdot \|\boldsymbol{u}^n\| = \rho[r(\mu A_x)] \cdot \rho[r(\mu A_y)] \cdot \|\boldsymbol{u}^n\|.$$

It is easy to verify by Gershgorin theorem that the eigenvalues of the matrices A_x and A_y are non-positive, hence provided that r fulfils $|r(z)| < 1$ for $z \in \mathbb{C}, \operatorname{Re} z < 0$, it is true that $\rho[r(\mu A_x)], \rho[r(\mu A_y)] \leq 1$. This proves $\|\boldsymbol{u}^{n+1}\| \leq \|\boldsymbol{u}^n\| \leq \cdots \leq \|\boldsymbol{u}^0\|$, hence stability.

Method 2.37 (Splitting of inhomogeneous systems) Recall our goal, namely fast methods for the two-dimensional diffusion equation. Our exposition so far has been contrived, because of the assumption that the boundary conditions are zero. In general, the linear ODE system is of the form

$$\boldsymbol{u}' = A\boldsymbol{u} + \boldsymbol{b}, \qquad \boldsymbol{u}(0) = \boldsymbol{u}^0, \tag{2.26}$$

where $\boldsymbol{b}$ originates in boundary conditions (and in a forcing term $f(x,y)$ in the original PDE (2.20)). Note that our analysis should accommodate $\boldsymbol{b} = \boldsymbol{b}(t)$, since boundary conditions might vary in time! The *exact* solution of (2.26) is provided by the *variation of constants* formula

$$\boldsymbol{u}(t) = \mathrm{e}^{tA}\boldsymbol{u}(0) + \int_0^t \mathrm{e}^{(t-s)A}\boldsymbol{b}(s)\,\mathrm{d}s, \qquad t \geq 0,$$

therefore

$$\boldsymbol{u}(t_{n+1}) = \mathrm{e}^{\Delta tA}\boldsymbol{u}(t_n) + \int_{t_n}^{t_{n+1}} \mathrm{e}^{(t_{n+1}-s)A}\boldsymbol{b}(s)\,\mathrm{d}s\,.$$

The integral can be frequently evaluated explicitly, e.g. when $\boldsymbol{b}$ is a linear combination of polynomial and exponential terms. For example, $\boldsymbol{b}(t) \equiv \boldsymbol{b} = \text{const}$ yields

$$\boldsymbol{u}(t_{n+1}) = \mathrm{e}^{\Delta tA}\boldsymbol{u}(t_n) + A^{-1}\left(\mathrm{e}^{\Delta tA} - I\right)\boldsymbol{b}.$$

This, unfortunately, is not a helpful observation, since, even if we split the exponential e^{tA}, how are we supposed to split $A^{-1} = (B+C)^{-1}$? The remedy is not to evaluate the integral explicitly but, instead, to use quadrature. For example, the trapezoidal rule $\int_0^k g(\tau)\,\mathrm{d}\tau = \frac{1}{2}k[g(0) + g(k)] + \mathcal{O}(k^3)$ gives

$$\boldsymbol{u}(t_{n+1}) \;\approx\; \mathrm{e}^{\Delta tA}\boldsymbol{u}(t_n) + \tfrac{1}{2}\Delta t[\mathrm{e}^{\Delta tA}\boldsymbol{b}(t_n) + \boldsymbol{b}(t_{n+1})],$$

with a local error of $\mathcal{O}((\Delta t)^3)$. We can now replace exponentials with their splittings. For example, Strang's splitting results in

$$\boldsymbol{u}^{n+1} \;=\; r\big(\tfrac{1}{2}\Delta tB\big)\, r\big(\Delta tC\big)\, r\big(\tfrac{1}{2}\Delta tB\big)\left[\boldsymbol{u}^n + \tfrac{1}{2}\Delta t\boldsymbol{b}^n\right] + \tfrac{1}{2}\Delta t\boldsymbol{b}^{n+1}.$$

As before, everything reduces to (inexpensive) solution of tridiagonal systems!