Dr A. Hansen

Mathematical Tripos Part II: Michaelmas Term 2020

Numerical Analysis – Lecture 10

Technique 2.31 (Splitting) We would like to find a fast solver to the system (2.16):

$$(I - \tfrac{1}{2}\mu A_*)\,\boldsymbol{u}^{n+1} = (I + \tfrac{1}{2}\mu A_*)\,\boldsymbol{u}^n\,, \tag{2.17}$$

The matrix $B = I - \frac{1}{2}\mu A_*$ has a structure similar to that of A_* (for the 5-point scheme), so we may apply the Hockney method for solving $B\boldsymbol{u}^{n+1} = \boldsymbol{v}^n = C\boldsymbol{u}^n$. However, since the method (2.17) has a local truncation error $\mathcal{O}(k^3 + kh^2)$, we don't need the exact solution of the system: it would be enough to have one within the error.

Let us employ the notation

$$\Delta_x^2 u_{\ell,m} = u_{\ell-1,m} - 2u_{\ell,m} + u_{\ell+1,m}, \qquad \Delta_y^2 u_{\ell,m} = u_{\ell,m-1} - 2u_{\ell,m} + u_{\ell,m+1}\,.$$

Then the Crank-Nicolson method calculates $\boldsymbol{u}^{n+1}$ by solving the system

$$\left[I - \tfrac{1}{2}\mu(\Delta_x^2 + \Delta_y^2)\right] u_{\ell,m}^{n+1} = \left[I + \tfrac{1}{2}\mu(\Delta_x^2 + \Delta_y^2)\right] u_{\ell,m}^n\,, \qquad \ell, m = 1...M. \tag{2.18}$$

The local error is however preserved if we replace this formula by the difference equation

$$\left[I - \tfrac{1}{2}\mu\Delta_x^2\right]\left[I - \tfrac{1}{2}\mu\Delta_y^2\right] u_{\ell,m}^{n+1} = \left[I + \tfrac{1}{2}\mu\Delta_x^2\right]\left[I + \tfrac{1}{2}\mu\Delta_y^2\right] u_{\ell,m}^n\,, \tag{2.19}$$

which is called the *split version of Crank-Nicolson*. Indeed, the difference between two schemes is equal to

$$\begin{aligned} \tfrac{1}{4}\mu^2\Delta_x^2\Delta_y^2(u_{\ell,m}^{n+1} - u_{\ell,m}^n) &= \tfrac{k^2}{4}\tfrac{1}{h^2}\Delta_x^2\tfrac{1}{h^2}\Delta_y^2\Big(k\tfrac{\partial}{\partial t}u_{\ell,m}^n + \mathcal{O}(k^2)\Big) \\ &= \tfrac{k^3}{4}\Big(\tfrac{\partial^2}{\partial x^2}\tfrac{\partial^2}{\partial y^2}\tfrac{\partial}{\partial t}u_{\ell,m}^n + \mathcal{O}(k + h^2)\Big) = \mathcal{O}(k^3)\,, \end{aligned}$$

the same magnitude as of the local error. In the matrix form, (2.19) is equivalent to splitting the matrix A_* into the sum of two matrices A_x and A_y as

$$A_* = A_x + A_y, \qquad A_x = \begin{bmatrix} -2I & I & & \\ I & \ddots & \ddots & \\ & \ddots & \ddots & I \\ & & I & -2I \end{bmatrix}, \quad A_y = \begin{bmatrix} H & & & \\ & H & & \\ & & \ddots & \\ & & & H \end{bmatrix}, \quad H = \begin{bmatrix} -2 & 1 & & \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & -2 \end{bmatrix},$$

and solving the uncoupled system

$$\left[I - \tfrac{1}{2}\mu A_x\right]\left[I - \tfrac{1}{2}\mu A_y\right]\boldsymbol{u}^{n+1} = \left[I + \tfrac{1}{2}\mu A_x\right]\left[I + \tfrac{1}{2}\mu A_y\right]\boldsymbol{u}^n$$

as

$$B_x\boldsymbol{u}^{n+1/2} = C_xC_y\boldsymbol{u}^n, \qquad B_y\boldsymbol{u}^{n+1} = \boldsymbol{u}^{n+1/2}.$$

Matrix $B_y = I - \frac{1}{2}\mu A_y$ is block diagonal, and solving $B_y\boldsymbol{u} = \boldsymbol{v}$ is just solving one and the same tridiagonal system $B\boldsymbol{u}_i = \boldsymbol{v}_i$ with different right-hand sides. Matrix $B_x = I - \frac{1}{2}\mu A_x$ is of the same form up to a permutation (reordering of the grid), so solving $B_x\boldsymbol{v} = \boldsymbol{b}$ is again a fast procedure.

Example 2.32 Consider the general diffusion equation

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

where $a(x,y) > \alpha > 0$ 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

$$\begin{aligned} u'_{\ell,m} &= \frac{1}{h^2}\Big[a_{\ell-\frac12,m}u_{\ell-1,m} + a_{\ell+\frac12,m}u_{\ell+1,m} + a_{\ell,m-\frac12}u_{\ell,m-1} + a_{\ell,m+\frac12}u_{\ell,m+1} \\ &\qquad - \Big(a_{\ell-\frac12,m} + a_{\ell+\frac12,m} + a_{\ell,m-\frac12} + a_{\ell,m+\frac12}\Big)u_{\ell,m}\Big] + f_{\ell,m}. \end{aligned} \tag{2.21}$$

Assuming zero boundary conditions and $f \equiv 0$, we have a system $\boldsymbol{u}' = A\boldsymbol{u}$, and we may solve it again by Crank–Nicolson, and apply the split $A = A_x + A_y$. Here, A_x and A_y are again constructed from the contribution of discretizations in the x- and y-directions respectively, namely A_x includes all the $a_{\ell\pm\frac12,m}$ terms, and A_y consists of the remaining $a_{\ell,m\pm\frac12}$ components. Arguments similar to what we used in moving from (2.18) to (2.19) justify the use of the split version in this general case as well.

Intermezzo 2.33 (Linear systems of ODEs) 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 $\boldsymbol{y}(t) = \mathrm{e}^{tA}\boldsymbol{y}_0$ is a soluton.

One observes that one-step methods for solving (2.22) are approximating a matrix exponential. Thus, with $k = \Delta t$,

$$\begin{aligned} &\text{Euler:} \quad \boldsymbol{y}_n = (I + kA)^n\boldsymbol{y}_0, && 1 + z = \mathrm{e}^z + \mathcal{O}(z^2); \\ &\text{TR:} \quad \boldsymbol{y}_n = \Big[\big(I - \tfrac12 kA\big)^{-1}\big(I + \tfrac12 kA\big)\Big]^n \boldsymbol{y}_0, && \frac{1+\frac12 z}{1-\frac12 z} = \mathrm{e}^z + \mathcal{O}(z^3). \end{aligned}$$

Technique 2.34 (Splitting methods) 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 had this been true for the matrices, i.e. that $\mathrm{e}^{tA} = \mathrm{e}^{t(B+C)} = \mathrm{e}^{tB}\mathrm{e}^{tC}$, we could have approximated each component of the exponent of $A = A_x + A_y$ with the trapezoidal rule, say, to produce

$$\boldsymbol{u}^{n+1} = \Big(I - \tfrac12\mu A_x\Big)^{-1}\Big(I + \tfrac12\mu A_x\Big)\Big(I - \tfrac12\mu A_y\Big)^{-1}\Big(I + \tfrac12\mu A_y\Big)\boldsymbol{u}^n\,, \qquad \mu = k/h^2\,, \tag{2.23}$$

and since both $I - \frac12\mu A_x$ and $I - \frac12\mu A_y$ are tridiagonal, this system can be solved very cheaply.

Unfortunately, the assumption that $\mathrm{e}^{t(B+C)} = \mathrm{e}^{tB}\mathrm{e}^{tC}$ is, in general, false. 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 2.35 (Splitting) Comparing the Taylor expansions of $\mathrm{e}^{t(B+C)}$ with $\mathrm{e}^{tB}\mathrm{e}^{tC}$ we obtain

$$\mathrm{e}^{tB}\mathrm{e}^{tC} = \mathrm{e}^{t(B+C)} + \tfrac12 t^2(BC - CB) + \mathcal{O}(t^3). \tag{2.24}$$

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 $r(z) = (1 + \frac12 z)/(1 - \frac12 z)$ results in a *split Crank–Nicolson* scheme, whose implementation reduces to a solution of tridiagonal algebraic linear systems.

It is easy to prove that

$$\mathrm{e}^{t(B+C)} = \tfrac12\left(\mathrm{e}^{tB}\mathrm{e}^{tC} + \mathrm{e}^{tC}\mathrm{e}^{tB}\right) + \mathcal{O}(t^3), \qquad \mathrm{e}^{t(B+C)} = \mathrm{e}^{\frac12 tB}\mathrm{e}^{tC}\mathrm{e}^{\frac12 tB} + \mathcal{O}(t^3),$$

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 $\boldsymbol{u}^{n+1} = r\Big(\frac12\mu A_x\Big)\, r\Big(\mu A_y\Big)\, r\Big(\frac12\mu A_x\Big)\, \boldsymbol{u}^n$ 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 nonpositive, hence provided that r fulfils $|r(z)| < 1$ for $z \in \mathbb{C}$ with $\operatorname{Re} z < 0$, it is then 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 2.36 (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

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

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