Mathematical Tripos Part II: Michaelmas Term 2022

Numerical Analysis – Lecture 10

Linear systems of ODEs In all the examples of semi-discretization we have seen so far, we always reach a linear system of ODE of the form:

$$\boldsymbol{u}' = A\boldsymbol{u}, \qquad \boldsymbol{u}(0) = \boldsymbol{u}_0. \tag{2.17}$$

The solution of this linear system of ODE is given by

$$\boldsymbol{u}(t) = \mathrm{e}^{tA}\boldsymbol{u}_0 \tag{2.18}$$

where the *matrix exponential* function is defined by $\mathrm{e}^B := \sum_{k=0}^{\infty} \frac{1}{k!}B^k$. It is easily verified that $\mathrm{d}\mathrm{e}^{tA}/\mathrm{d}t = A\mathrm{e}^{tA}$, therefore (2.18) is indeed a solution of (2.17).

If A can be diagonalized $A = VDV^{-1}$, then $\mathrm{e}^{tA} = V\mathrm{e}^{tD}V^{-1}$ where e^{tD} is the diagonal matrix consisting $\mathrm{diag}\,(\mathrm{e}^{tD_{ii}})$. As such one can compute the solution of (2.17) exactly. However computing an eigenvalue decomposition can be costly, and so one would like to consider more efficient methods, based on the solution of sparse linear systems instead.

Observe that one-step methods for solving (2.17) are approximating a matrix exponential. Indeed, with $k = \Delta t$, we have:

Euler:	$\boldsymbol{u}^{n+1} = (I + kA)\boldsymbol{u}^n,$	$\mathrm{e}^z = 1 + z + \mathcal{O}(z^2);$
Implicit Euler:	$\boldsymbol{u}^{n+1} = (I - kA)^{-1}\boldsymbol{u}^n,$	$\mathrm{e}^z = (1-z)^{-1} + \mathcal{O}(z^2);$
Trapezoidal Rule:	$\boldsymbol{u}^{n+1} = \left(I - \frac{1}{2}kA\right)^{-1}\left(I + \frac{1}{2}kA\right)\boldsymbol{u}^n,$	$\mathrm{e}^z = \frac{1+\frac{1}{2}z}{1-\frac{1}{2}z} + \mathcal{O}(z^3).$

In practice the matrix A is very sparse, and this can be exploited when solving linear systems e.g., for the implicit Euler or Trapezoidal Rule.

Splitting In many cases, the matrix A is naturally expressed as a sum of two matrices, $A = B + C$. For example, when discretizing the diffusion equation in 2D with zero boundary conditions, we have $A = \frac{1}{h^2}(A_x + A_y)$ where $\frac{1}{h^2}A_x \in \mathbb{R}^{M^2\times M^2}$ corresponds to the 3-point discretization of $\frac{\partial^2}{\partial x^2}$, and $\frac{1}{h^2}A_y \in \mathbb{R}^{M^2\times M^2}$ corresponds to the 3-point discretization of $\frac{\partial^2}{\partial y^2}$. In matrix notations, if the grid points are ordered by columns, then we have:

$$A_x = \begin{bmatrix} -2I & I & & \\ I & \ddots & \ddots & \\ & \ddots & \ddots & I \\ & & I & -2I \end{bmatrix}, \quad A_y = \begin{bmatrix} G & & & \\ & G & & \\ & & \ddots & \\ & & & G \end{bmatrix}, \quad G = \begin{bmatrix} -2 & 1 & & \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & -2 \end{bmatrix} \in \mathbb{R}^{M\times M}. \tag{2.19}$$

Remark: It is convenient to note that $A_x = G \otimes I$ and $A_y = I \otimes G$, where $\otimes$ is the Kronecker product of matrices (`kron` in Matlab) defined by

$$A \otimes B = \begin{bmatrix} A_{11}B & A_{12}B & \dots & A_{1m_A}B \\ A_{21}B & A_{22}B & \dots & A_{2m_A}B \\ \vdots & & & \\ A_{n_A1}B & \dots & \dots & A_{n_Am_A}B \end{bmatrix} \in \mathbb{R}^{n_An_B\times m_Am_B}$$

where $A \in \mathbb{R}^{n_A\times m_A}$ and $B \in \mathbb{R}^{n_B\times m_B}$.

In general, $\exp(t(B+C)) \neq \exp(tB)\exp(tC)$. Equality holds however when B and C commute.

Proposition 2.31 *For any matrices B, C,*

$$\mathrm{e}^{t(B+C)} = \mathrm{e}^{tB}\mathrm{e}^{tC} + \frac{1}{2}t^2(CB - BC) + \mathcal{O}(t^3). \tag{2.20}$$

If B and C commute, then $\mathrm{e}^{B+C} = \mathrm{e}^B\mathrm{e}^C$.

Proof. We Taylor-expand both expressions $\mathrm{e}^{tB}\mathrm{e}^{tC}$ and $\mathrm{e}^{t(B+C)}$:

$$\begin{aligned}\mathrm{e}^{tB}\mathrm{e}^{tC} &= (I + tB + t^2B^2/2 + \mathcal{O}(t^3))(I + tC + t^2C^2/2 + \mathcal{O}(t^3)) \\ &= I + t(B+C) + \frac{t^2}{2}(B^2 + C^2 + 2BC) + \mathcal{O}(t^3)\end{aligned}$$

and

$$\begin{aligned}\mathrm{e}^{t(B+C)} &= I + t(B+C) + \frac{t^2}{2}(B+C)^2 + \mathcal{O}(t^3) \\ &= I + t(B+C) + \frac{t^2}{2}(B^2 + C^2 + BC + CB) + \mathcal{O}(t^3).\end{aligned}$$

Equation (**??**) follows.

When B and C commute, we can write:

$$\mathrm{e}^{B+C} = \sum_{k=0}^{\infty} \frac{1}{k!}(B+C)^k = \sum_{k=0}^{\infty} \frac{1}{k!} \sum_{i+j=k} \binom{k}{i} B^i C^j = \sum_{i,j=0}^{\infty} \frac{1}{i!j!} B^i C^j = \mathrm{e}^B \mathrm{e}^C$$

where in the second step we used the fact that B and C commute. □

Back to the 2D diffusion equation: the matrices A_x and A_y in (2.19) happen to commute: this is easy to check, and not surprising since the operators $\partial^2/\partial x^2$ and $\partial^2/\partial y^2$, which A_x/h^2 and A_y/h^2 approximate, are known to commute. So we have $\mathrm{e}^{kA} = \mathrm{e}^{kA_x/h^2}\mathrm{e}^{kA_y/h^2}$. This means that the solution of the semi-discretized diffusion equation in 2D, with zero boundary conditions, satisfies

$$\boldsymbol{u}^{n+1} = \mathrm{e}^{kA_x/h^2}\mathrm{e}^{kA_y/h^2}\boldsymbol{u}^n. \tag{2.21}$$

Split Crank-Nicolson: In the split Crank-Nicolson scheme, we approximate each exponential map in (2.21) by the rational function $r(z) = (1 + z/2)(1 - z/2)^{-1}$, which leads to

$$\boldsymbol{u}^{n+1} = (I + \tfrac{\mu}{2}A_x)(I - \tfrac{\mu}{2}A_x)^{-1}(I + \tfrac{\mu}{2}A_y)(I - \tfrac{\mu}{2}A_y)^{-1}\boldsymbol{u}^n. \tag{2.22}$$

Note that computing $\boldsymbol{u}^{n+1/2} = (I + \frac{\mu}{2}A_y)(I - \frac{\mu}{2}A_y)^{-1}\boldsymbol{u}^n$ can be done efficiently in $\mathcal{O}(M^2)$ time as A_y is block-diagonal, and the matrices G are tridiagonal (each tridiagonal solve requires $\mathcal{O}(M)$ time, and we have M of these). Computing $\boldsymbol{u}^{n+1} = (I + \frac{\mu}{2}A_x)(I - \frac{\mu}{2}A_x)^{-1}\boldsymbol{u}^{n+1/2}$ can also be done in $\mathcal{O}(M^2)$ time, since A_x is also block-diagonal provided we appropriately permute the rows and columns so that the grid ordering is by rows instead of columns. This means that the update step (2.22) of Split-Crank-Nicolson can be performed in time $\mathcal{O}(M^2)$ and only requires tridiagonal matrix solves (no FFT needed).

One can easily verify stability of the split Crank-Nicolson scheme. Indeed, we can write

$$\begin{aligned}\|\boldsymbol{u}^{n+1}\|_2 &\le \|r(\mu A_x)\|_2 \|r(\mu A_y)\boldsymbol{u}^n\|_2 \\ &\le \|r(\mu A_x)\|_2 \|r(\mu A_y)\|_2 \|\boldsymbol{u}^n\|_2 \\ &= \rho[r(\mu A_x)] \cdot \rho[r(\mu A_y)] \cdot \|\boldsymbol{u}^n\|_2\end{aligned}$$

where in the last equality we used the fact that $r(\mu A_x)$ and $r(\mu A_y)$ are symmetric to replace their operator norm with their spectral radius. The function $r(z) = (1+\frac{1}{2}z)(1-\frac{1}{2}z)^{-1}$ satisfies $|r(z)| \le 1$ for $z \in \mathbb{C}$ with $\mathrm{Re}\, z \le 0$. By the Gersgorin theorem, we see that the eigenvalues of A_x and A_y are nonpositive. This implies that $\rho[r(\mu A_x)], \rho[r(\mu A_y)] \le 1$, proving $\|\boldsymbol{u}^{n+1}\| \le \|\boldsymbol{u}^n\| \le \cdots \le \|\boldsymbol{u}^0\|$, hence stability.

In general, however, the matrices B and C in $A = B + C$ do not have to commute, as in the following example:

2D diffusion with variable diffusion coefficient The general diffusion equation with a diffusion coefficient $a(x, y) > 0$ is given by:

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(a(x,y)\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(a(x,y)\frac{\partial u}{\partial y}\right), \tag{2.23}$$

together with initial conditions on $[0,1]^2$ and Dirichlet boundary conditions along $\partial[0,1]^2\times[0,\infty)$. We 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

$$\begin{aligned} u'_{\ell,m} \;=\; \tfrac{1}{h^2}\Big[&a_{\ell-\frac12,m}u_{\ell-1,m}+a_{\ell+\frac12,m}u_{\ell+1,m}+a_{\ell,m-\frac12}u_{\ell,m-1}+a_{\ell,m+\frac12}u_{\ell,m+1}\\ &-\big(a_{\ell-\frac12,m}+a_{\ell+\frac12,m}+a_{\ell,m-\frac12}+a_{\ell,m+\frac12}\big)u_{\ell,m}\Big]. \end{aligned} \tag{2.24}$$

Assuming zero boundary conditions, we have a system $\boldsymbol{u}' = A\boldsymbol{u}$, and the matrix A can be split as $A = \frac{1}{h^2}(A_x + A_y)$. 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.

The resulting operators A_x and A_y do not necessarily commute, and so the splitting scheme

$$\boldsymbol{u}^{n+1} = \mathrm{e}^{kA_x/h^2}\mathrm{e}^{kA_y/h^2}\boldsymbol{u}^n$$

will carry an error of $\mathcal{O}(k^2)$, following (2.20).

Strang splitting: One can obtain better splitting approximations of $\mathrm{e}^{t(B+C)}$. For example it is not hard to prove that $\mathrm{e}^{\frac12 tB}\mathrm{e}^{tC}\mathrm{e}^{\frac12 tB}$ gives a $\mathcal{O}(t^3)$ approximation of $\mathrm{e}^{t(B+C)}$, i.e.,

$$\mathrm{e}^{t(B+C)} = \mathrm{e}^{\frac12 tB}\mathrm{e}^{tC}\mathrm{e}^{\frac12 tB} + \mathcal{O}(t^3). \tag{2.25}$$

Remark 2.32 (Splitting of inhomogeneous systems) Our exposition so far has been limited to the case of zero boundary conditions. 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, possibly, in a forcing term $f(x,y)$ in the original PDE (2.23)). 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

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

therefore

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

The integral on the right-hand side can be evaluated using quadrature. For example, the trapezoidal rule $\int_0^k g(\tau)\,\mathrm{d}\tau = \frac12 k[g(0)+g(k)] + \mathcal{O}(k^3)$ gives

$$\boldsymbol{u}(t_{n+1}) \;\approx\; \mathrm{e}^{kA}\boldsymbol{u}(t_n) + \tfrac12 k[\mathrm{e}^{kA}\boldsymbol{b}(t_n) + \boldsymbol{b}(t_{n+1})],$$

with a local error of $\mathcal{O}(k^3)$. We can now replace exponentials with their splittings. For example, Strang's splitting (2.25), together with the rational function approximation $r(z) = (1+z/2)/(1-z/2)$ of the exponential map, results in

$$\boldsymbol{u}^{n+1} \;=\; r\big(\tfrac12 kB\big)\,r\big(kC\big)\,r\big(\tfrac12 kB\big)\big[\boldsymbol{u}^n + \tfrac12 k\boldsymbol{b}^n\big] + \tfrac12 k\boldsymbol{b}^{n+1}.$$

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