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.

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 $A_x \in \mathbb{R}^{M^2 \times M^2}$ corresponds to the 3-point discretization of $\frac{\partial^2}{\partial x^2}$, and $A_y \in \mathbb{R}^{M^2 \times M^2}$ 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:

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

When the matrices B and C commute, we can use the following fact about the matrix exponential.

Proposition 2.31 *If B and C commute, then* $\mathrm{e}^{B+C} = \mathrm{e}^B\mathrm{e}^C$.

Proof. We have

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

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

The matrices A_x and A_y in (2.19) happen to commute (easy to check), and so $\mathrm{e}^{\Delta t A} = \mathrm{e}^{\frac{\Delta t}{h^2}A_x}\mathrm{e}^{\frac{\Delta t}{h^2}A_y}$. This means that the solution of the semi-discretized diffusion equation in 2D, with zero boundary conditions, satisfies

$$\boldsymbol{u}^{n+1} = \mathrm{e}^{\mu A_x}\mathrm{e}^{\mu A_y}\boldsymbol{u}^n. \tag{2.20}$$

Split Crank-Nicolson: In the split Crank-Nicolson scheme, we approximate each exponential map in (2.20) by the rational function $r(z) = (1 + z/2)(1 - z/2)^{-1}$, which leads to

$$\boldsymbol{u}^{n+1} = (I + \tfrac{\mu}{2}A_x)(I - \tfrac{\mu}{2}A_x)^{-1}(I + \tfrac{\mu}{2}A_y)(I - \tfrac{\mu}{2}A_y)^{-1}\boldsymbol{u}^n. \tag{2.21}$$

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 (2.21) of Split-Crank-Nicolson can be performed in time $\mathcal{O}(M^2)$ and only requires tridiagonal matrix solves (no FFT needed).

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

Example 2.32 The general diffusion equation with a diffusion coefficient $a(x, y) > 0$ is given by:

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(a(x,y)\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(a(x,y)\frac{\partial u}{\partial y}\right), \tag{2.22}$$

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

$$\begin{aligned} u'_{\ell,m} \quad = \quad & \tfrac{1}{h^2}\Big[a_{\ell-\frac{1}{2},m}u_{\ell-1,m} + a_{\ell+\frac{1}{2},m}u_{\ell+1,m} + a_{\ell,m-\frac{1}{2}}u_{\ell,m-1} + a_{\ell,m+\frac{1}{2}}u_{\ell,m+1} \\ & \qquad - \big(a_{\ell-\frac{1}{2},m} + a_{\ell+\frac{1}{2},m} + a_{\ell,m-\frac{1}{2}} + a_{\ell,m+\frac{1}{2}}\big)u_{\ell,m}\Big]. \end{aligned} \tag{2.23}$$

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 $a_{\ell\pm\frac{1}{2},m}$ terms, and A_y consists of the remaining $a_{\ell,m\pm\frac{1}{2}}$ components.

In this case the matrices A_x and A_y do not necessarily commute. The next proposition tells us that approximating $\mathrm{e}^{t(B+C)}$ by $\mathrm{e}^{tB}\mathrm{e}^{tC}$ results in an error of $\mathcal{O}(t^2)$.

Proposition 2.33 *For any matrices B, C,*

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

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

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

and

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

The result follows.

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 $r(z) = (1 + \frac{1}{2}z)/(1 - \frac{1}{2}z) = \mathrm{e}^z + O(z^3)$ results in a *split Crank–Nicolson* scheme, whose implementation reduces to a solution of tridiagonal algebraic linear systems.

Strang splitting: One can obtain better splitting approximations of $\mathrm{e}^{t(B+C)}$. For example it is not hard to prove that $\mathrm{e}^{\frac{1}{2}tB}\mathrm{e}^{tC}\mathrm{e}^{\frac{1}{2}tB}$ gives a $\mathcal{O}(t^3)$ approximation of $\mathrm{e}^{t(B+C)}$, i.e.,

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

Thus, as long as $r(z) = \mathrm{e}^z + \mathcal{O}(z^3)$, the time-stepping formula

$$\boldsymbol{u}^{n+1} = r\big(\tfrac{1}{2}\mu A_x\big)\, r\big(\mu A_y\big)\, r\big(\tfrac{1}{2}\mu A_x\big)\, \boldsymbol{u}^n$$

carries a local error of $\mathcal{O}((\Delta t)^3)$.

Stability: Consider the general diffusion equation with the splitting scheme (2.25). 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 we have

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

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 $\operatorname{Re} z \le 0$. By the Gersgorin theorem, we see that the eigenvalues of A_x and A_y are nonpositive. Then it is true that $\rho[r(\mu A_x)], \rho[r(\mu A_y)] \le 1$. This proves $\|\boldsymbol{u}^{n+1}\| \le \|\boldsymbol{u}^n\| \le \cdots \le \|\boldsymbol{u}^0\|$, hence stability.

Remark 2.34 (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, possibly, in a forcing term $f(x, y)$ in the original PDE (2.22)). 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

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

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

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

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

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

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

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

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