UNIVERSITY OF CAMBRIDGE

Department of Applied Mathematics and Theoretical Physics

Numerical Solution of Differential Equations

A. Iserles

Part III

Michaelmas 2007

Numerical Analysis Group
Centre for Mathematical Sciences
Wilberforce Rd
Cambridge CB3 0WA

version 2.7182818284590
September 2006

Numerical Solution of Differential Equations

Contents

These notes are based in the main on parts of

A. Iserles, *A First Course in the Numerical Analysis of Differential Equations,* Cambridge University Press, Cambridge (1996)

with the addition of some material.

1 Introduction

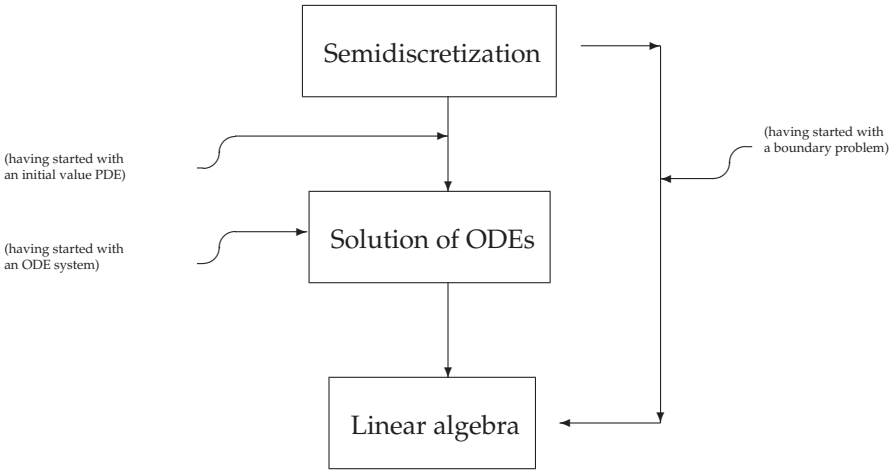
Preamble. Mathematical subjects are typically studied according to what might be called the *tree of knowledge* paradigm, in a natural progression from axioms to theorems to applications. However, there is a natural temptation to teach applicable subjects as a *phone directory*, whereby imparting maximum of implementable knowledge without bothering with mathematical niceties. In this lecture course we steer a middle course, although sympathetic to the tree of knowledge model.

Like the human brain, computers are algebraic machines. They can perform a finite number of elementary arithmetical and logical operations – and that's just about all! Conceptual understanding and numerical solution of differential equations (an analytic construct!) can be accomplished only by translating 'analysis' into 'algebra'.

Computation is not an alternative to rigourous analysis. The two go hand-in-hand and the dichotomy between 'qualitative' and 'quantitative' mathematics is a false one.

The purpose of this course is to help you to think about numerical calculations in a more professional manner, whether as a preparation for career in numerical maths/scientific computing or as useful background material in computationally-heavy branches of applied maths. Don't worry, you will not be a numerical analyst by the end of the course! But you might be able to read a book or a paper on a numerical subject and understand it.

Structure of the subject. Typically, given a differential equation with an initial value (and perhaps boundary values), the computation can be separated into three conceptual stages:



Example: The diffusion equation $u_t = u_{xx}$, zero boundary conditions. Herewith we sketch our reasoning without any deep analysis, which will be fleshed out in the sequel.

Stage 1 Semidiscretization:

$$u_k' = \frac{1}{(\Delta x)^2}(u_{k-1} - 2u_k + u_{k+1}), \qquad k = 1, 2, \ldots, m-1, \tag{1.1}$$

an ordinary differential equation (ODE). Here $\Delta x = 1/m$.
In matrix form $\boldsymbol{u}' = A\boldsymbol{u}$, $\boldsymbol{u}(0) = \boldsymbol{u}_0$, hence the *exact* solution of (1.1) is $\boldsymbol{u}(t) = \mathrm{e}^{tA}\boldsymbol{u}_0$.

Stage 2 Ordinary Differential Equations:
Familiar methods for $\boldsymbol{y}' = \boldsymbol{f}(t, \boldsymbol{y})$:

Forward Euler (FE): $\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + \Delta t \boldsymbol{f}(t_n, \boldsymbol{y}_n)$;

Backward Euler (BE): $\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + \Delta t \boldsymbol{f}(t_{n+1}, \boldsymbol{y}_{n+1})$;

Trapezoidal Rule (TR): $\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + \frac{1}{2}\Delta t(\boldsymbol{f}(t_n, \boldsymbol{y}_n) + \boldsymbol{f}(t_{n+1}, \boldsymbol{y}_{n+1}))$.

In our case

FE: $\boldsymbol{u}^n = (I + \Delta t A)\boldsymbol{u}^{n-1} = \cdots = (I + \Delta t A)^n \boldsymbol{u}^0$;

BE: $\boldsymbol{u}^n = (I - \Delta t A)^{-1}\boldsymbol{u}^{n-1} = \cdots = (I - \Delta t A)^{-n}\boldsymbol{u}^0$;

TR: $\boldsymbol{u}^n = (I - \frac{1}{2}\Delta t A)^{-1}(I + \frac{1}{2}\Delta t A)\boldsymbol{u}^{n-1} = \cdots = \left((I - \frac{1}{2}\Delta t A)^{-1}(I + \frac{1}{2}\Delta t A)\right)^n \boldsymbol{u}^0$.

The matrix A is symmetric $\Rightarrow A = QDQ^\top$, where Q is orthogonal and D diagonal, $D = \text{diag}\,\{d_1, d_2, \ldots, d_{m-1}\}$. Moreover,

$$d_k = \frac{2}{(\Delta x)^2}\left(-1 + \cos\frac{k\pi}{m}\right) = -4m^2 \sin^2\frac{k\pi}{2m}, \quad k = 1, 2, \ldots, m-1.$$

In the FE case

$$\boldsymbol{u}^n = Q(I + \Delta t D)^n Q^\top \boldsymbol{u}^0.$$

The exact solution of $u_t = u_{xx}$: uniformly bounded, dissipates to 0 at ∞. To mimic this, we require $\rho(I + \Delta t D) \le 1$. Let $\mu := \Delta t/(\Delta x)^2$. Then

$$\begin{aligned}\sigma(I + \Delta t D) = \left\{1 - 4\mu \sin^2\frac{k\pi}{2m}\right\} \;\Rightarrow\; \rho(I + \Delta t D) &= \max\left\{\left|1 - 4\mu\sin^2\frac{(m-1)\pi}{2m}\right|, 1\right\} \\ &\approx \max\{|1 - 4\mu|, 1\}.\end{aligned}$$

Thus,

$$\rho(I + \Delta t D) < 1 \quad \Longrightarrow \quad \mu \le \frac{1}{2} \quad \Longrightarrow \quad \Delta t \le \frac{1}{2}(\Delta x)^2.$$

We can do better with either BE or TR: in each case, uniform boundedness holds for all Δt and Δx.

Stage 3 Linear algebra:
'Good' ODE methods (BE and TR) entail the solution of a linear system of equations. There are several options:

Option 1: LU factorization (i.e. Gaussian elimination): LU-factorizing the matrix A costs $\mathcal{O}(m^3)$ flops and each subsequent 'backsolve' costs $\mathcal{O}(m^2)$ flops. This is bearable in a 1D problem but worse to come in, say, 2 space dimensions ($\mathcal{O}(m^6)$ instead of $\mathcal{O}(m^3)$ etc.).

Option 2: Sparse LU factorization (the Thomas algorithm): Can be performed in $\approx 5m$ flops per time step – substantial saving but costs significantly mount in several space dimensions.

Option 3: Iterative methods (Jacobi, Gauss–Seidel…): quite expensive, although the cost less sensitive to the number of space dimensions. Converge in the present case.

Option 4: Splitting methods (to come!): retain the advantages of sparse LU for any number of space dimensions.

2 Ordinary Differential Equations

Formulation of the problem. We solve

$$\boldsymbol{y}' = \boldsymbol{f}(t, \boldsymbol{y}), \quad \boldsymbol{y}(0) = \boldsymbol{y}_0 \in \mathbb{R}^d. \tag{2.1}$$

Without loss of generality, (1) The system is *autonomous*, i.e. $\boldsymbol{f} = \boldsymbol{f}(\boldsymbol{y})$; and (2) $\boldsymbol{f}$ is analytic (and hence so is $\boldsymbol{y}$). Both assumptions may be lifted when they breach generality.

Denote $h = \Delta t > 0$ and consider numerical methods that approximate the exact solution $\boldsymbol{y}(nh)$ by $\boldsymbol{y}_n$ for all $n \in \mathbb{Z}_+$ (or in a smaller range of interest). Of course (and unless we are perverse) $\boldsymbol{y}_0$ coincides with the initial value.

Order of a method. We say that a method is of *order* p if for all $n \in \mathbb{Z}_+$ it is true that $\boldsymbol{y}_{n+1} = \tilde{\boldsymbol{y}}((n+1)h) + \mathcal{O}\big(h^{p+1}\big)$, where $\tilde{\boldsymbol{y}}$ is the exact solution of (2.1) with the initial value $\tilde{\boldsymbol{y}}(nh) = \boldsymbol{y}_n$.

High derivatives. These can be obtained by repeatedly differentiating (2.1). This gives equations of the form $\boldsymbol{y}^{(k)} = \boldsymbol{f}_k(\boldsymbol{y})$, where

$$\boldsymbol{f}_0(\boldsymbol{y}) = \boldsymbol{y}, \qquad \boldsymbol{f}_1(\boldsymbol{y}) = \boldsymbol{f}(\boldsymbol{y}), \qquad \boldsymbol{f}_2(\boldsymbol{y}) = \frac{\partial \boldsymbol{f}(\boldsymbol{y})}{\partial \boldsymbol{y}} \boldsymbol{f}(\boldsymbol{y}), \qquad \ldots. \tag{2.2}$$

2.1 Taylor methods

From the Taylor theorem and (2.2)

$$\boldsymbol{y}((n+1)h) = \sum_{k=0}^{\infty} \frac{1}{k!} h^k \boldsymbol{y}^{(k)}(nh) = \sum_{k=0}^{\infty} \frac{1}{k!} h^k \boldsymbol{f}_k(\boldsymbol{y}(nh)), \quad n \in \mathbb{Z}_+.$$

This leads us to consider the *Taylor method*

$$\boldsymbol{y}_{n+1} = \sum_{k=0}^{p} \frac{1}{k!} h^k \boldsymbol{f}_k(\boldsymbol{y}_n), \quad n \in \mathbb{Z}_+. \tag{2.3}$$

Examples:

$p = 1$: $\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + h\boldsymbol{f}(\boldsymbol{y}_n)$ (forward Euler)

$p = 2$: $\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + h\boldsymbol{f}(\boldsymbol{y}_n) + \frac{1}{2} h^2 \frac{\partial \boldsymbol{f}(\boldsymbol{y}_n)}{\partial \boldsymbol{y}} \boldsymbol{f}(\boldsymbol{y}_n)$

Theorem 1 *The Taylor method (2.3) is of order p.*

Proof By induction. Assuming $\boldsymbol{y}_n = \tilde{\boldsymbol{y}}(nh)$, it follows that $\boldsymbol{f}_k(\boldsymbol{y}_n) = \boldsymbol{f}_k(\tilde{\boldsymbol{y}}(nh))$, hence $\boldsymbol{y}_{n+1} = \tilde{\boldsymbol{y}}((n+1)h) + \mathcal{O}\big(h^{p+1}\big)$. □

Connection with e^z. The *differential operator*: $\mathrm{D}\boldsymbol{g}(t) = \boldsymbol{g}'(t)$; the *shift operator*: $\mathrm{E}\boldsymbol{g}(t) = \boldsymbol{g}(t+h)$ ($\mathrm{E} = \mathrm{E}_h$). Thus, the ODE (2.1) is $\mathrm{D}\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{y})$, whereas $\mathrm{D}^k\boldsymbol{y} = \boldsymbol{f}_k(\boldsymbol{y})$. Numerical solution of the ODE – equivalent to approximating the action of the shift operator and its powers. But the Taylor theorem implies that $\forall$ analytic function $\boldsymbol{g}$

$$\mathrm{E}\boldsymbol{g}(t) = \sum_{k=0}^{\infty} \frac{1}{k!} h^k \mathrm{D}^k \boldsymbol{g}(t) = \mathrm{e}^{h\mathrm{D}} \boldsymbol{g}(t).$$

Let $R(z) = \sum_{k=0}^{\infty} r_k z^k$ be an analytic function s.t. $R(z) = \mathrm{e}^z + \mathcal{O}\big(z^{p+1}\big)$ (i.e., $r_k = \frac{1}{k!}$, $k = 0, 1, \ldots, p$). Identically to Theorem 1 we can prove that the formal 'method'

$$\boldsymbol{y}_{n+1} = R(h\mathrm{D})y_n = \sum_{k=0}^{\infty} r_k h^k \boldsymbol{f}_k(\boldsymbol{y}_n), \quad n \in \mathbb{Z}_+, \tag{2.4}$$

is of order p. Indeed, the Taylor method (2.3) follows from (2.4) by letting $R(z) = \sum_{k=0}^{p} \frac{1}{k!} z^k$, the pth section of the Taylor expansion of e^z.

Stability. The test equation: $y' = \lambda y$, $y(0) = 1$, $h = 1$. The exact *stability set*: $\lambda \in \mathbb{C}^- = \{z \in \mathbb{C} : \operatorname{Re} z < 0\}$. The *stability domain* $\mathcal{D}$ of a method: the set of all $\lambda \in \mathbb{C}$ s.t. $\lim_{n\to\infty} y_n = 0$.

A-stability: We say that a method is *A-stable* if $\mathbb{C}^- \subseteq \mathcal{D}$.

Why is A-stability important? Consider the equation

$$\boldsymbol{y}' = \begin{bmatrix} -1 & 1 \\ 0 & -10^5 \end{bmatrix} \boldsymbol{y}, \qquad \boldsymbol{y}(0) = \boldsymbol{y}_0.$$

There are two solution components, e^{-t} (which decays gently) and $\mathrm{e}^{-10^5 t}$ (which decays almost at once, $\mathrm{e}^{-10^5} \approx 3.56 \times 10^{-43430}$). Inasmuch as the second component dies out fast, we require $-10^5 h \in \mathcal{D}$ – otherwise the solution gets out of hand. This requirement to depress the step length (for non-A-stable methods) is characteristic of *stiff equations*.

In the case of Taylor's method, $f_k(y) = \lambda^k y$, $k \in \mathbb{Z}_+$, hence

$$y_n = \left(\sum_{k=0}^{p} \frac{1}{k!} \lambda^k \right)^n, \quad n \in \mathbb{Z}_+.$$

Hence

$$\mathcal{D} = \left\{ z \in \mathbb{C} : \left| \sum_{k=0}^{p} \frac{1}{k!} z^k \right| < 1 \right\}$$

and it follows that $\mathcal{D}$ must be bounded – in particular, the method can't be A-stable.

2.2 Rational methods

Choose R as a rational function,

$$R(z) = \frac{\sum_{k=0}^{M} p_k z^k}{\sum_{k=0}^{N} q_k z^k}.$$

This corresponds to the numerical method

$$\sum_{k=0}^{N} q_k h^k \boldsymbol{f}_k(\boldsymbol{y}_{n+1}) = \sum_{k=0}^{M} p_k h^k \boldsymbol{f}_k(\boldsymbol{y}_n). \tag{2.5}$$

If $N \geq 1$ then (2.5) is an algebraic system of equations – more about the solution of such (nonlinear) systems later.

Padé approximations. Given a function f, analytic at the origin, the $[M/N]$ *Padé approximation* is the quotient of an Mth degree polynomial over an Nth degree polynomial that matches the Taylor

series of f to the highest order of accuracy. For $f(z) = \exp z$ we have $R_{M/N} = P_{M/N}/Q_{M/N}$, where

$$P_{M/N}(z) = \sum_{k=0}^{M} \binom{M}{k} \frac{(M+N-k)!}{(M+N)!} z^k,$$
$$Q_{M/N}(z) = \sum_{k=0}^{N} \binom{N}{k} \frac{(M+N-k)!}{(M+N)!} (-z)^k = P_{N/M}(-z).$$

Lemma 2 *$R_{M/N}(z) = \mathrm{e}^z + \mathcal{O}\big(z^{M+N+1}\big)$ and no $[M/N]$ function can do better.*

Corollary 1 *The* Padé method

$$\sum_{k=0}^{N} (-1)^k \binom{N}{k} \frac{(M+N-k)!}{(M+N)!} h^k \boldsymbol{f}_k(\boldsymbol{y}_{n+1}) = \sum_{k=0}^{M} \binom{M}{k} \frac{(M+N-k)!}{(M+N)!} h^k \boldsymbol{f}_k(\boldsymbol{y}_n) \tag{2.6}$$

is of order $M+N$.

Examples:

[0/1]: $\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + h\boldsymbol{f}(\boldsymbol{y}_{n+1})$ (backward Euler);

[1/1]: $\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + \frac{1}{2}h[\boldsymbol{f}(\boldsymbol{y}_n) + \boldsymbol{f}(\boldsymbol{y}_{n+1})]$ (trapezoidal rule);

[0/2]: $\boldsymbol{y}_{n+1} - h\boldsymbol{f}(\boldsymbol{y}_{n+1}) + \frac{1}{2}h^2 \dfrac{\partial \boldsymbol{f}(\boldsymbol{y}_{n+1})}{\partial \boldsymbol{y}} \boldsymbol{f}(\boldsymbol{y}_{n+1}) = \boldsymbol{y}_n.$

A-stability. Solving $y' = \lambda y$, $y(0) = 1$, $h = 1$, with (2.5) we obtain $y_n = R^n(\lambda)$, hence

$$\mathcal{D} = \{z \in \mathbb{C} : |R(z)| < 1\}.$$

Lemma 3 *The method (2.5) is A-stable if and only if (a) all the poles of R reside in $\mathbb{C}^+ := \{z \in \mathbb{C} : \mathrm{Re}\, z > 0\}$; and (b) $|R(\mathrm{i}y)| \le 1$ for all $y \in \mathbb{R}$.*

Proof By the maximum modulus principle. □

It easy to verify that all three methods in the last example are A-stable. In general, according to a theorem of Wanner, Hairer & Nørsett, the Padé method (2.6) is A-stable iff $M \le N \le M+2$, $M \in \mathbb{Z}_+$. Figure 2.1 displays linear stability domains of four popular methods.

2.3 Multistep methods

These exploit past values of the solution. Thus (with a single derivative – the whole theory can be generalized to use higher derivatives)

$$\sum_{l=0}^{m} \alpha_l \boldsymbol{y}_{n+l} = h \sum_{l=0}^{m} \beta_l \boldsymbol{f}(\boldsymbol{y}_{n+l}), \qquad \alpha_m = 1. \tag{2.7}$$

An operatorial interpretation. Let $\rho(w) := \sum_{l=0}^{m} \alpha_l w^l$, $\sigma(w) = \sum_{l=0}^{m} \beta_l w^l$. Since $hD = \log E$, substituting the exact solution in (2.7) yields

$$[\rho(E) - \log E\, \sigma(E)]\boldsymbol{y}(nh) = \text{small perturbation.} \tag{2.8}$$

Suppose that

$$\rho(w) = \log w\sigma(w) + \mathcal{O}\big(|1-w|^{p+1}\big)\,.$$

Then

$$\{\rho(E) - \log E\sigma(E)\}\,\boldsymbol{y}(nh) = \mathcal{O}\big(h^{p+1}\big)\,. \tag{2.9}$$

Subtracting (2.9) from (2.7) and using the implicit function theorem we 'deduce'

"Lemma" *The method (2.7) is of order p.*

The snag in the "lemma" is that not always are we allowed to use the implicit function theorem in this manner.

The root condition. We say that ρ obeys the root condition if all its zeros are in $|w| \le 1$ and the zeros on $|w| = 1$ are simple. The root condition suffices for the application of the implicit function theorem. Moreover, we say that a method is *convergent* if, as $h \downarrow 0$, the numerical error is uniformly bounded throughout a compact interval by a constant multiple of the errors in the choice of the starting values and in the solution of algebraic equations.

Theorem 4 (The Dahlquist Equivalence Theorem) *The method (2.7) is convergent iff $p \ge 1$ and ρ obeys the root condition.*

Proof in the easy direction: Let $y' \equiv 0$, $y(0) = 1$. Then $\sum_{l=0}^{m} \alpha_l y_{n+l} = 0$, a linear recurrence with the solution $y_n = \sum_{i=1}^{r} \sum_{j=0}^{\mu_i - 1} \alpha_{i,j} n^j \omega_i^n$, where ω_i is a zero of ρ of multiplicity μ_i, $\sum_{i=1}^{r} \mu_i = m$. The $\alpha_{i,j}$s are determined by the starting values. Hence, whether $|\omega_i| > 1$ or $|\omega_j| = 1$ and $\mu_i \ge 2$ for some i, some starting values imply that $\lim_{n\to\infty} |y_n| = \infty$. This can't converge to $y(t) \equiv 1$ on any bounded interval as $h \downarrow 0$. □

Examples: (Here and elsewhere $\boldsymbol{f}_m = \boldsymbol{f}(\boldsymbol{y}_m)$.)

1. $\boldsymbol{y}_{n+2} = \boldsymbol{y}_n + 2h\boldsymbol{f}_{n+1}$ (explicit midpoint rule, a.k.a. leapfrog), order 2, convergent.
2. $\boldsymbol{y}_{n+2} - (1+a)\boldsymbol{y}_{n+1} + a\boldsymbol{y}_n = \frac{1}{12}h[(5+a)\boldsymbol{f}_{n+2} + 8(1-a)\boldsymbol{f}_{n+1} - (1+5a)\boldsymbol{f}_n]$, convergent for $-1 \le a < 1$, of order 3 for $a \ne -1$ and order 4 for $a = -1$.
3. $\boldsymbol{y}_{n+3} + \frac{27}{11}\boldsymbol{y}_{n+2} - \frac{27}{11}\boldsymbol{y}_{n+1} - \boldsymbol{y}_n = \frac{3}{11}h\left(\boldsymbol{f}_n + 9\boldsymbol{f}_{n+1} + 9\boldsymbol{f}_{n+2} + \boldsymbol{f}_{n+3}\right)$, order 6. But

$$\rho(w) = (w-1)\left(w + \frac{19 + 4\sqrt{15}}{11}\right)\left(w + \frac{19 - 4\sqrt{15}}{11}\right)$$

   and the root condition is violated.

Highest order of a multistep method. Let

$$\rho(w) - \log w\sigma(w) = c(w-1)^{p+1} + \mathcal{O}\big(|w-1|^{p+2}\big)\,, \qquad c \ne 0, \tag{2.10}$$

and define

$$R(\zeta) := \left(\frac{\zeta - 1}{2}\right)^m \rho\left(\frac{\zeta+1}{\zeta-1}\right) = \sum_{l=0}^{m} r_l \zeta^l, \qquad S(\zeta) := \left(\frac{\zeta - 1}{2}\right)^m \sigma\left(\frac{\zeta+1}{\zeta-1}\right) = \sum_{l=0}^{m} s_l \zeta^l.$$

Proposition 5 *The following is true:*

(a) $p \geq 1 \Rightarrow r_m = 0$, *hence* $\deg R = m - 1$;

(b) *Order* $p \Leftrightarrow R(\zeta) - \log \frac{\zeta+1}{\zeta-1} S(\zeta) = c \left(\frac{2}{\zeta}\right)^{p+1-m} + \cdots$ *as* $\zeta \to \infty$;

(c) *The root condition* $\Rightarrow r_{m-1} \neq 0$ *and all the nonzero* r_l*s have the same sign.*

Proof **(a)** $r_m = 2^{-m}\rho(1)$. But $p \geq 1$ implies $\rho(1) = 0$. **(b)** Follows at once from (2.10); **(c)** The root condition $\Leftrightarrow$ all the zeros of R are in $\operatorname{cl} \mathbb{C}^-$, no multiple zeros reside on $\mathrm{i}\mathbb{R}$ and $r_{m-1} \neq 0$ (since $r_{m-1} = 2^{-m}(2\rho'(1) - m\rho(1))$, the latter corresponds to ρ having no multiple zero at 1). Denote the zeros of R by $\xi_1, \ldots, \xi_M, \xi_{M+1} \pm \mathrm{i}\nu_{M+1}, \ldots, \xi_N \pm \mathrm{i}\nu_N$. Thus,

$$R(\zeta) = r_{m-1} \prod_{j=1}^{M} (\zeta - \xi_j) \prod_{j=M+1}^{N} [(\zeta - \xi_j)^2 + \nu_j^2].$$

Since $-\xi_j, \nu_j^2 \geq 0$ and the r_ls are convex linear combinations of products of these quantities, the lemma follows. □

Theorem 6 (Dahlquist's first barrier) *Convergence implies* $p \leq 2\,[(m+2)/2]$.

Proof Let $G(\zeta) := \left(\log \frac{\zeta+1}{\zeta-1}\right)^{-1}$. Thus, **(b)** $\Rightarrow$

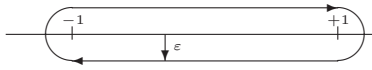
$$R(\zeta)G(\zeta) - S(\zeta) = c \left(\frac{2}{\zeta}\right)^{p-m} + \mathcal{O}\big(\zeta^{-p+m-1}\big)$$

As $|\zeta| \to \infty$, we have $G(\zeta) \to \frac{1}{2}\zeta$, hence

$$G(\zeta) = \tfrac{1}{2}\zeta + \sum_{l=0}^{\infty} g_l \zeta^{-l}.$$

However, $G(\zeta) = -G(-\zeta)$, hence $g_{2l} = 0$, $l \in \mathbb{Z}_+$. By the *Cauchy integral formula*,

$$g_{2l+1} = \frac{1}{2\pi\mathrm{i}} \int_{\Gamma_\varepsilon} v^{2l} G(v)\,\mathrm{d}v,$$

where Γ_ε: 

. Letting $\varepsilon \downarrow 0$, for all $l \in \mathbb{Z}_+$

$$g_{2l+1} = \frac{1}{2\pi\mathrm{i}} \int_{-1}^{1} v^{2l} \left\{ \frac{1}{\log \frac{1+v}{1-v} + \mathrm{i}\pi} - \frac{1}{\log \frac{1+v}{1-v} - \mathrm{i}\pi} \right\} \mathrm{d}v = -\int_{-1}^{1} \frac{v^{2l}\,\mathrm{d}v}{\left(\log \frac{1+v}{1-v}\right)^2 + \pi^2} < 0.$$

But, for general polynomials R and S of degrees $m-1$ and m, respectively, $R(\zeta)G(\zeta) - S(\zeta) = \sum_{l=-m}^{\infty} e_l \zeta^{-l}$. Order conditions $\Rightarrow e_{-m} = \cdots = e_{p-m-1} = 0$. **(c)** and $g_{2l+1} < 0$, $l \in \mathbb{Z}_+$, imply

$m = 2s$**:** $|e_2| = \left|\sum_{j=1}^{s} r_{2j-1} g_{2j+1}\right| \geq |r_{2s-1} g_{2s+1}| > 0 \Rightarrow p \leq m + 2$;

$m = 2s + 1$**:** $|e_1| = \left|\sum_{j=0}^{s} r_{2j} g_{2j+1}\right| \geq |r_{2s} g_{2s+1}| > 0 \Rightarrow p \leq m + 1$.

This proves the theorem. □

Attaining the first Dahlquist barrier. When m is even, order $m+2$ attainable but all the zeros of ρ live on $|w|=1$ and this is unhealthy. Better choice: $p=m+1$ for all m. The 'stablest' method with $\rho(w)=w^m-w^{m-1}$, since all the zeros of ρ (except for the one at 1) are at 0. This gives the *Adams methods*.

Adams–Moulton: Implicit ($\beta_m \neq 0$), order $m+1$:

$m=1$: $\rho(w)=w-1$, $\sigma(w)=\frac{1}{2}(w+1)$,
$m=2$: $\rho(w)=w^2-w$, $\sigma(w)=\frac{1}{12}(5w^2+8w-1)$.

Adams–Bashforth: Explicit ($\beta_m=0$), order m:

$m=1$: $\rho(w)=w-1$, $\sigma(w)\equiv 1$ (forward Euler),
$m=2$: $\rho(w)=w^2-w$, $\sigma(w)=\frac{1}{2}(3w-1)$.

To derive Adams–Moulton, say, choose an mth degree polynomial σ that matches the Taylor expansion of $\rho(w)/\log w$ about $w=1$.

A-stability. Let $T(z,w)=\rho(w)-z\sigma(w)$. When applied to $y'=\lambda y$, $y(0)=1$, $h=1$, the method reads

$$\sum_{l=0}^{m}(\alpha_l-\lambda\beta_l)y_{n+l}=T(\lambda,E)y_n=0.$$

This is a difference equation whose *characteristic polynomial* is the function $T(\lambda,\cdot)$. Let its zeros be $\omega_1(\lambda),\dots,\omega_{N(\lambda)}(\lambda)$, of multiplicities $\mu_1(\lambda),\dots,\mu_{N(\lambda)}(\lambda)$ resp., $\sum_1^{N(\lambda)}\mu_k(\lambda)\le m$. Then

$$y_n=\sum_{j=1}^{N(\lambda)}\sum_{i=0}^{\mu_j(\lambda)-1} n^i\omega_j^n(\lambda)\xi_{i,j}(\lambda), \qquad \xi_{i,j}(\lambda) \text{ independent of } n.$$

Hence, the *linear stability domain* is the set of all $\lambda\in\mathbb{C}$ such that all the zeros of $T(\lambda,w)=0$ reside in $|w|<1$ (cf. Figure 2.1). We have

Lemma 7 *A-stability $\Leftrightarrow$ for every $\lambda\in\mathbb{C}^-$ all the zeros of $T(\lambda,w)=0$ are in $|w|<1$.*

Theorem 8 (Dahlquist's second barrier) *A-stability implies that $p\le 2$. Moreover, the 2nd order A-stable method with the least truncation error is the trapezoidal rule.*

Multistep–multiderivative methods. Motivated by the above, we consider methods that employ both information across a range of steps and higher derivatives. We already know that there exists a 1-step, N-derivative method ($[N/N]$ Padé) of order $2N$.

Theorem 9 (Wanner–Hairer–Nørsett) *A-stability implies that $p\le 2N$ for any multistep N-derivative method. Moreover, the $(2N)$-order A-stable method with the least truncation error is the 1-step $[N/N]$ Padé.*

Checking for A-stability of 2-step methods. We again employ the maximum principle, checking for (i) absence of poles in $\mathrm{cl}\,\mathbb{C}^-$; and (ii) the root condition of $T(\mathrm{i}t,\cdot)$, $t\in\mathbb{R}$. We can use the

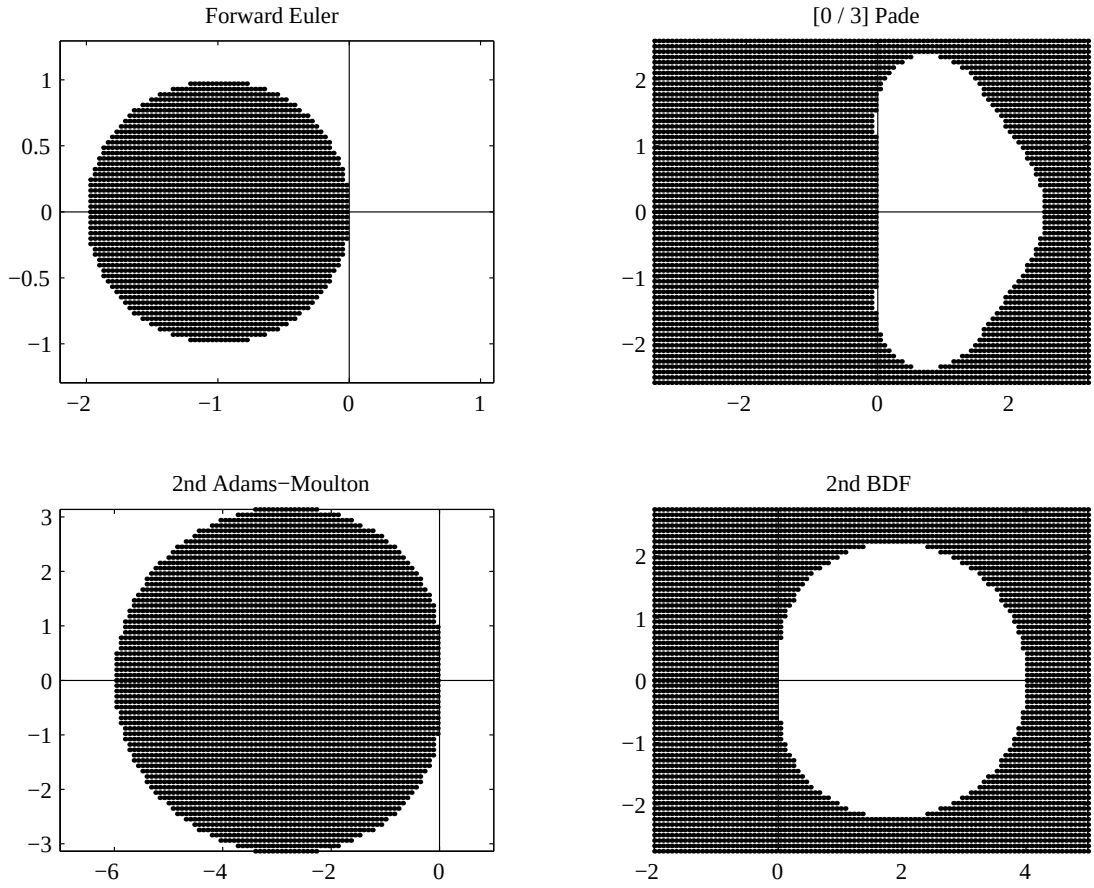


Figure 2.1 Linear stability domains of four numerical methods.

Cohn–Schur criterion: The quadratic $aw^2 + bw + c$, $a, b, c \in \mathbb{C}$, $a \neq 0$, obeys the root condition iff **(a)** $|a| \geq |c|$; **(b)** $\left(|a|^2 - |c|^2\right)^2 \geq \left|a\bar{b} - b\bar{c}\right|^2$; **(c)** If (b) is obeyed as an equality then $|b| < 2|a|$.

Relaxed stability concepts. Requiring stability only across a wedge in $\mathbb{C}^-$ of angle α results in $A(\alpha)$*-stability*. Thus, A-stability $\Leftrightarrow$ $A(90°)$-stability. This is sufficient for most purposes. High-order $A(\alpha)$-stable methods exist for $\alpha < 90°$.

Backward differentiation formulae (BDF). We want 'stability' when $|\lambda| \gg 1$ ($\mathrm{Re}\,\lambda < 0$). Since $T(\lambda, w) \approx -\lambda\sigma(w)$, the 'best' choice is $\sigma(w) = \beta_m w^m$. Stipulating order m, we have

$m = 1$**:** $\rho(w) = w - 1$, $\sigma(w) = w$ (backward Euler, A-stable);

$m = 2$**:** $\rho(w) = w^2 - \frac{4}{3}w + \frac{1}{3}$, $\sigma(w) = \frac{2}{3}w^2$ (A-stable);

$m = 3$**:** $\rho(w) = w^3 - \frac{18}{11}w^2 + \frac{9}{11}w - \frac{2}{11}$, $\sigma(w) = \frac{6}{11}w^3$ ($A(86°2')$-stable).

BDF methods are the standard workhorse for stiff equations.

Are BDF convergent? It is possible to prove that BDF is convergent iff $m \leq 6$. That's enough for all intents and purposes.

2.4 Implementation of multistep methods

Implicit methods need be solved by iteration, which must be employed in every time step. Typically, starting values for the iterative scheme are provided by an explicit method of comparable order. This leads to *predictor–corrector (PC)* pairs, where "P" is explicit and "C" is implicit – e.g. Adams–Bashforth and Adams–Moulton.

Modes of iteration. Either iterating until the error is beneath tolerance (*iterating to convergence*, PC$^\infty$) or executing a fixed (small) number of iterations and abandoning the process unless the error is beneath tolerance (PC^m, where m is the number of iterations). The choice between the two is dictated by the interplay between the *set-up cost* $\mathbf{C_S}$ and the *iteration cost* $\mathbf{C_I}$. Thus, $\mathbf{C_S} \gg \mathbf{C_I} \Rightarrow$ PC$^\infty$, otherwise PC^m with $m \in \{1, 2, 3\}$. The cost is influenced by several factors:

1. Function evaluations;
2. Solution of nonlinear algebraic equations;
3. Coping with stiffness;
4. Error and stepsize control.

Solving nonlinear algebraic equations. The algebraic system is

$$\boldsymbol{y} - \beta h \boldsymbol{f}(\boldsymbol{y}) = \boldsymbol{v},$$

where $\boldsymbol{v}$ is known. *Direct iteration:*

$$\boldsymbol{y}^{[j+1]} = \boldsymbol{v} + \beta h \boldsymbol{f}(\boldsymbol{y}^{[j]}). \tag{2.11}$$

This is a special case of the *functional iteration* $\boldsymbol{x}^{[j+1]} = \boldsymbol{g}(\boldsymbol{x}^{[j]})$ to approach a fixed point of $\boldsymbol{g}$.

Theorem 10 (Banach's contraction mapping theorem) *Let $\|\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{y})\| \le L\|\boldsymbol{x} - \boldsymbol{y}\|$, $0 < L < 1$, for all $\|\boldsymbol{x} - \boldsymbol{x}^{[0]}\|, \|\boldsymbol{y} - \boldsymbol{x}^{[0]}\| \le r$ for some $r > 0$. Provided that $\|\boldsymbol{g}(\boldsymbol{x}^{[0]}) - \boldsymbol{x}^{[0]}\| \le (1 - L)r$, it is true that*

(a) $\|\boldsymbol{x}^{[j]} - \boldsymbol{x}^{[0]}\| \le r \; \forall j \in \mathbb{Z}_+$;

(b) $\boldsymbol{x}^\star = \lim_{j\to\infty} \boldsymbol{x}^{[j]}$ *exists and is a fixed point of* $\boldsymbol{g}$;

(c) $\boldsymbol{x}^\star$ *is the unique fixed point of* $\boldsymbol{g}$ *in* $S_r := \{\|\boldsymbol{x} - \boldsymbol{x}^{[0]}\| \le r\}$.

Proof We prove that $\|\boldsymbol{x}^{[j+1]} - \boldsymbol{x}^{[j]}\| \le L^j(1 - L)r$. It is true for $k = 0$ and, by induction,

$$\|\boldsymbol{x}^{[j+1]} - \boldsymbol{x}^{[j]}\| = \|\boldsymbol{g}(\boldsymbol{x}^{[j]}) - \boldsymbol{g}(\boldsymbol{x}^{[j-1]})\| \le L\|\boldsymbol{x}^{[j]} - x^{[j-1]}\| \le L^j(1 - L)r.$$

Therefore, by the triangle inequality,

$$\|\boldsymbol{x}^{[j+1]} - \boldsymbol{x}^{[0]}\| = \left\|\sum_{i=0}^{j} (\boldsymbol{x}^{[i+1]} - \boldsymbol{x}^{[i]})\right\| \le \sum_{i=0}^{j} L^i(1 - L)r = (1 - L^{j+1})r \le r.$$

This proves **(a)**.
$\{\boldsymbol{x}^{[j]}\}_{j=0}^\infty$ is a *Cauchy sequence*, since

$$\|\boldsymbol{x}^{[k+j]} - \boldsymbol{x}^{[j]}\| = \left\|\sum_{i=0}^{k-1} (\boldsymbol{x}^{[j+i+1]} - \boldsymbol{x}^{[j+i]})\right\| \le L^j r \overset{j\to\infty}{\to} 0.$$

Therefore a limit $\boldsymbol{x}^\star$ exists in the compact set S_r. It is a fixed point and **(b)** follows.
Finally, suppose that $\boldsymbol{x}^\circ \in S_r$, $\boldsymbol{x}^\circ \neq \boldsymbol{x}^\star$, is a fixed point. Then

$$\|\boldsymbol{x}^\star - \boldsymbol{x}^\circ\| = \|\mathbf{g}(\boldsymbol{x}^\star) - \mathbf{g}(\boldsymbol{x}^\circ)\| \leq L\|\boldsymbol{x}^\star - \boldsymbol{x}^\circ\| < \|\boldsymbol{x}^\star - \boldsymbol{x}^\circ\|,$$

a contradiction. □

For the iteration (2.11) $L \approx h|\beta|\rho(\partial \boldsymbol{f}/\partial \boldsymbol{y})$, hence, for *stiff equations*, attaining $L < 1$ may radically depress $h > 0$. In that case we may consider *Newton–Raphson (NR)*, namely functional iteration on

$$\hat{\boldsymbol{g}}(\boldsymbol{x}) := \boldsymbol{x} - \left(I - \frac{\partial \boldsymbol{g}(\boldsymbol{x})}{\partial \boldsymbol{x}}\right)^{-1} (\boldsymbol{x} - \boldsymbol{g}(\boldsymbol{x})).$$

This gives the scheme

$$\boldsymbol{x}^{[j+1]} = \boldsymbol{x}^{[j]} - \left(I - \frac{\partial \boldsymbol{g}(\boldsymbol{x}^{[j]})}{\partial \boldsymbol{x}}\right)^{-1} (\boldsymbol{x}^{[j]} - \boldsymbol{g}(\boldsymbol{x}^{[j]})). \tag{2.12}$$

The scheme (2.12) is *very* expensive, since (i) the Jacobian must be re-evaluated in each iteration; (ii) a new linear system must be solved for every j. Instead, we use *modified NR (MNR)*, keeping the Jacobian constant:

$$\boldsymbol{x}^{[j+1]} = \boldsymbol{x}^{[j]} - \left(I - \frac{\partial \boldsymbol{g}(\boldsymbol{x}^\circ)}{\partial \boldsymbol{x}}\right)^{-1} (\boldsymbol{x}^{[j]} - \boldsymbol{g}(\boldsymbol{x}^{[j]})), \tag{2.13}$$

with, for example, $\boldsymbol{x}^\circ = \boldsymbol{x}^{[0]}$.

Conclusion. (2.11) for nonstiff, (2.13) for stiff.

But... $\mathbf{C_S}$ negligible for (2.11), whereas $\mathbf{C_S} \gg \mathbf{C_I}$ for MNR (as long as we reuse the same LU factorization in every step). Hence

Conclusion. Solve nonstiff ODE with PC^m, solve stiff ODE by iterating to convergence.

2.5 Strategies for error control

The following are some of the most popular devices to control local error and to choose the step length so that the error estimate does not exceed given tolerance.

The Milne device. Let c_{P} and c_{C} be the *error constants* of "P" and "C" resp., hence $\boldsymbol{y}_{n+1}^{(P)} = \boldsymbol{y}(t_{n+1}) + c_{\mathrm{P}} h^{p+1}\, \mathrm{d}^{p+1}\boldsymbol{y}(t_n)/\,\mathrm{d}t^{p+1} + \cdots$ etc. Hence

$$\begin{aligned} & \boldsymbol{y}_{n+1}^{(P)} - \boldsymbol{y}_{n+1}^{(C)} \approx (c_{\mathrm{P}} - c_{\mathrm{C}}) h^{p+1} \frac{\mathrm{d}^{p+1}\boldsymbol{y}(t_n)}{\mathrm{d}t^{p+1}} \\ \Rightarrow \qquad & \left\| \boldsymbol{y}_{n+1}^{(C)} - \boldsymbol{y}(t_{n+1}) \right\| \approx \left| \frac{c_{\mathrm{C}}}{c_{\mathrm{P}} - c_{\mathrm{C}}} \right| \left\| \boldsymbol{y}_{n+1}^{(P)} - \boldsymbol{y}_{n+1}^{(C)} \right\|. \end{aligned}$$

This provides an estimate of the local error.

Deferred correction. As an example, consider the trapezoidal rule $\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + \frac{1}{2}h[\boldsymbol{f}(\boldsymbol{y}_n) + \boldsymbol{f}(\boldsymbol{y}_{n+1})]$. The error is $-\frac{1}{12}h^3\boldsymbol{y}'''(t_n) + \mathcal{O}(h^4)$. Let

$$\boldsymbol{s}(\boldsymbol{w}_{n-1}, \boldsymbol{w}_n, \boldsymbol{w}_{n+1}) := -\frac{1}{12}h(\boldsymbol{f}(\boldsymbol{w}_{n+1}) - 2\boldsymbol{f}(\boldsymbol{w}_n) + \boldsymbol{f}(\boldsymbol{w}_{n-1})).$$

Then $\boldsymbol{s}(\boldsymbol{y}_{n-1}, \boldsymbol{y}_n, \boldsymbol{y}_{n+1}) = -\frac{1}{12}h^3\boldsymbol{y}'''(t_n) + \mathcal{O}(h^4)$. We retain an extra value $\boldsymbol{y}_{n-1}$ and use $\boldsymbol{s}$ to estimate the local error.

The Zadunaisky device. Given a p-order solution sequence $\{\boldsymbol{y}_j\}_{j=0}^n$, choose a polynomial $\boldsymbol{q}$, $\deg \boldsymbol{q} = p$, that interpolates $\boldsymbol{y}$ at the last $p+1$ grid points. Let $\boldsymbol{d}(t) := \boldsymbol{q}'(t) - \boldsymbol{f}(\boldsymbol{q}(t))$ and consider the *auxiliary system* $\boldsymbol{z}' = \boldsymbol{f}(\boldsymbol{z}) + \boldsymbol{d}(t)$ (with the same past values). Then **(a)** Since $\boldsymbol{q}(t) = \boldsymbol{y}(t) + \mathcal{O}(h^{p+1})$ and $\boldsymbol{y}$ obeys $\boldsymbol{y}' - \boldsymbol{f}(\boldsymbol{y}) = \boldsymbol{0}$, $\boldsymbol{d}(t) = \mathcal{O}(h^p)$ and the system is very near (within $\mathcal{O}(h^p)$) of the original ODE; and **(b)** The function $\boldsymbol{q}$ solves exactly the auxiliary equation; it makes sense to use $\boldsymbol{z}_{n+1} - \boldsymbol{q}(t_{n+1})$ as an estimate of $\boldsymbol{y}_{n+1} - \boldsymbol{y}(t_{n+1})$.

Gear's automatic integration. This is not just an error control device but an integrated approach to the implementation of multistep methods. We estimate the local error of an order-p multistep method by repeatedly differentiating an interpolating polynomial. Moreover: suppose that we have a whole family of m-step methods for $m = 1, 2, \ldots, m^*$, say, each of order $p_m = m + K$ (thus, $K = 1$ for Adams–Moulton and $K = 0$, $m^* \le 6$, for BDF) and with the error constant c_m.

1. Commence the iteration with $m = 1$.
2. At the nth step, working with the m-step method, evaluate error estimates
$$E_j \approx c_j h^{j+K+1}\boldsymbol{y}^{(j+K+1)}(t_n), \qquad j \in I_m := \{m-1, m, m+1\} \cap \{1, 2, \ldots, m^*\}.$$
3. Use E_m to check whether $\|\boldsymbol{y}_{n+1} - \boldsymbol{y}(t_{n+1})\|$ is beneath the error tolerance.
4. Using E_j find the method in I_m that is likely to produce a result within the error tolerance in the next step *with the longest step size*.
5. Change to that method and step-size, using interpolation to re-grid the values.

Remarks on Gear's method:

- No starting values are required beyond the initial value.
- We must retain enough past values for error control and step-size management – this is well in excess of what is required by the multistep method.
- The set I_m of 'locally allowed' methods is likely to be further restricted: we are not allowed to increase step-size too soon after a previous increase.
- In practical packages it is usual to use the *Nordsieck representation*, whereby, instead of past values of $\boldsymbol{y}_j$, we store (use, interpolate…) finite differences.
- Many popular all-purpose ODE solvers – `DIFFSYS`, `EPISODE`, `FACSIMILE`, `DASSLE` etc. – are all based on Gear's method. (However, other popular solvers, e.g. `STRIDE` and `SIMPLE`, use Runge–Kutta methods.)

2.6 Runge–Kutta methods

Quadrature. Consider $y'(t) = f(t)$. Thus
$$y(t_0 + h) = y_0 + \int_0^h f(t_0 + \tau)\,\mathrm{d}\tau \approx y_0 + h\sum_{l=1}^{s} b_l f(t_0 + c_l h),$$

where the latter is a *quadrature formula*. Following this logic, we can try for $\boldsymbol{y}' = \boldsymbol{f}(t, \boldsymbol{y})$ the 'scheme'

$$\boldsymbol{y}(t_0 + h) \approx \boldsymbol{y}_0 + h \sum_{l=1}^{s} b_l \boldsymbol{f}(t_0 + c_l h, \boldsymbol{y}(t_0 + c_l h)).$$

Although impossible in an exact form, this provides a useful paradigm. Runge–Kutta schemes can be seen as an attempt to flesh it out. . . .

An RK scheme. Let A be an $s \times s$ *RK matrix* and $\boldsymbol{b} \in \mathbb{R}^s$ a vector of *RK weights*. $\boldsymbol{c} := A\mathbf{1}$ (where $\mathbf{1} \in \mathbb{R}^s$ is a vector of 1s) is the vector of *RK nodes*. The corresponding s-stage *RK method* reads

$$\begin{aligned}
\boldsymbol{k}_1 &= \boldsymbol{f}\left(t_n + c_1 h, \boldsymbol{y}_n + h \sum_{j=1}^{s} a_{1,j} \boldsymbol{k}_j\right), \\
\boldsymbol{k}_2 &= \boldsymbol{f}\left(t_n + c_2 h, \boldsymbol{y}_n + h \sum_{j=1}^{s} a_{2,j} \boldsymbol{k}_j\right), \\
&\vdots \\
\boldsymbol{k}_s &= \boldsymbol{f}\left(t_n + c_s h, \boldsymbol{y}_n + h \sum_{j=1}^{s} a_{s,j} \boldsymbol{k}_j\right), \\
\boldsymbol{y}_{n+1} &= \boldsymbol{y}_n + h \sum_{l=1}^{s} b_l \boldsymbol{k}_l.
\end{aligned} \tag{2.14}$$

Butcher's notation. Denote an RK method by the tableau

$$\begin{array}{c|c} \boldsymbol{c} & A \\ \hline & \boldsymbol{b}^\top \end{array} \quad = \quad \begin{array}{c|cccc} c_1 & a_{1,1} & a_{1,2} & \cdots & a_{1,s} \\ c_2 & a_{2,1} & a_{2,2} & \cdots & a_{2,s} \\ \vdots & \vdots & \vdots & & \vdots \\ c_s & a_{s,1} & a_{s,2} & \cdots & a_{s,s} \\ \hline & b_1 & b_2 & \cdots & b_s \end{array} .$$

Different categories of RK methods:
Explicit RK (ERK): A strictly lower triangular;
Diagonally-implicit RK (DIRK): A lower triangular;
Singly-diagonally-implicit RK (SDIRK): A lower triangular, $a_{l,l} \equiv \text{const} \neq 0$.
Implicit RK (IRK): Otherwise.

An example: ERK, $s = 3$**:** Unless otherwise stated, all quantities derived at (t_n, y_n). We assume that the ODE is scalar and autonomous. (For order ≥ 5 *this represents loss of generality,* but this is not the case at present.) Hence

$$\begin{aligned}
k_1 &= f, \\
k_2 &= f(y + h a_{2,1} k_1) = f + c_2 h f_y f + \tfrac{1}{2} h^2 c_2^2 f_{yy} f^2 + \cdots, \\
k_3 &= f(y + h(a_{3,1} k_1 + a_{3,2} k_2)) = f + h c_3 f_y f + h^2 \left(c_2 a_{3,2} f_y^2 f + \tfrac{1}{2} c_3^2 f_{yy} f^2\right) + \cdots
\end{aligned}$$

and

$$\begin{aligned}
y_{n+1} = y &+ h(b_1 + b_2 + b_3) f + h^2 (b_2 c_2 + b_3 c_3) f_y f \\
&+ h^3 \left(\frac{b_2 c_2^2 + b_3 c_3^2}{2} f_{yy} f^2 + b_3 c_2 a_{3,2} f_y^2 f\right) + \mathcal{O}\left(h^4\right).
\end{aligned} \tag{2.15}$$

However, $\frac{\mathrm{d}}{\mathrm{d}t}f = f_y f$, $\frac{\mathrm{d}^2}{\mathrm{d}^2 t}f = f_{yy}f^2 + f_y^2 f \Rightarrow$

$$y(t_n + h) = y + hf + \tfrac{1}{2}h^2 f_y f + \tfrac{1}{6}h^3(f_{yy}f^2 + f_y^2 f) + \mathcal{O}\big(h^4\big)\,. \tag{2.16}$$

Comparison of (2.15) and (2.16) yields the third-order conditions

$$b_1 + b_2 + b_3 = 1, \quad b_2 c_2 + b_3 c_3 = \tfrac{1}{2}, \quad b_2 c_2^2 + b_3 c_3^2 = \tfrac{1}{3}, \quad b_3 c_2 a_{3,2} = \tfrac{1}{6}.$$

Examples of 3-stage ERK of order 3:

Kutta:
$$\begin{array}{c|ccc} 0 & & & \\ \frac{1}{2} & \frac{1}{2} & & \\ 1 & -1 & 2 & \\ \hline & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \end{array}\,, \qquad \text{Nystrom:} \quad \begin{array}{c|ccc} 0 & & & \\ \frac{2}{3} & \frac{2}{3} & & \\ \frac{2}{3} & 0 & \frac{2}{3} & \\ \hline & \frac{1}{4} & \frac{3}{8} & \frac{3}{8} \end{array}\,.$$

Highest order attainable by ERK.

stages	1	2	3	4	5	6	7	8	9	10	11
order	1	2	3	4	4	5	6	6	7	7	?

Elementary differentials. Error expansions of RK can be arranged into a well-behaved mathematical framework by using *elementary differentials* of Butcher and graph theory. The idea is to establish a recursive relationship between f, $f_y f$, $f_y^2 f$, $f_{yy}f^2$ etc. (elementary differentials). Note that each kth derivative of y can be expressed as a linear combination (with positive integer coefficients) of elementary differentials of 'order' $k-1$. A handy way of expressing the recurrence is by associating elementary differentials with *rooted trees* and the expansion coefficients with certain combinatorial attributes of these trees. Likewise, the RK method can be expanded in elementary differentials and comparison of the two expansions allows to ascertain the order of any given (explicit or otherwise) method. However, this approach is nonconstructive – given a method, we can check its order, but the technique provides only partial clues how to design high-order methods with suitable properties.

Embedded RK. An error-control device specific to RK. We *embed* a method in a larger method. For example, let

$$\tilde{A} = \begin{bmatrix} A & \mathbf{0} \\ \boldsymbol{a}^\top & \tilde{a} \end{bmatrix}, \qquad \tilde{\boldsymbol{c}} = \begin{bmatrix} \boldsymbol{c} \\ \tilde{c} \end{bmatrix},$$

such that $\begin{array}{c|c} \tilde{\boldsymbol{c}} & \tilde{A} \\ \hline & \tilde{\boldsymbol{b}}^\top \end{array}$ is of higher order than $\begin{array}{c|c} \boldsymbol{c} & A \\ \hline & \boldsymbol{b}^\top \end{array}$. Comparison of the two yields an estimate of the error in the latter method.

Collocation methods. Assuming that $c_1, \ldots, c_s$ are distinct, find an s-degree polynomial $\boldsymbol{u}$ s.t. $\boldsymbol{u}(t_n) = \boldsymbol{y}_n$ and

$$\boldsymbol{u}'(t_n + c_l h) = \boldsymbol{f}(t_n + c_l h, \boldsymbol{u}(t_n + c_l h)), \qquad l = 1, 2, \ldots, s. \tag{2.17}$$

We let $\boldsymbol{y}_{n+1} := \boldsymbol{u}(t_n + h)$ be the approximation at t_{n+1}.
Let $\omega(t) := \prod_{l=1}^{s}(t - c_l)$ and $\omega_l(t) := \omega(t)/(t - c_l)$, $l = 1, 2, \ldots, s$.

Lemma 11 *Let the c_ls be distinct. The RK method*

$$a_{k,l} = \frac{1}{\omega_l(c_l)} \int_0^{c_k} \omega_l(\tau)\,\mathrm{d}\tau,\ k = 1, 2, \ldots, s, \quad b_l = \frac{1}{\omega_l(c_l)} \int_0^1 \omega_l(\tau)\,\mathrm{d}\tau, \quad l = 1, 2, \ldots, s$$

is identical *to the collocation method (2.17).*

Proof The polynomial $\boldsymbol{u}'$ coincides with its $(s-1)$st degree Lagrange interpolation polynomial. Thus, denoting by

$$L_j(t) := \frac{\omega_j(t)}{\omega_j(c_j)}$$

the *jth Lagrange cardinal polynomial* at $c_1, c_2, \ldots, c_s$ of degree $s-1$ (thus $L_j(c_j) = 1$, $L_j(c_i) = 0$ for all $i \neq j$), we have

$$\begin{aligned}\boldsymbol{u}'(t) &= \sum_{j=1}^{s} L_j\left(\tfrac{t-t_n}{h}\right) \boldsymbol{u}'(t_n + c_i h) = \sum_{j=1}^{s} L_j\left(\tfrac{t-t_n}{h}\right) \boldsymbol{f}(t_n + c_j h, \boldsymbol{u}(t_n + c_j h)) \\ &= \sum_{j=1}^{s} \frac{\omega_j((t-t_n)/h)}{\omega_j(c_j)} \boldsymbol{f}(t_n + c_j h, \boldsymbol{u}(t_n + c_j h))\end{aligned}$$

and integration yields

$$\boldsymbol{u}(t) = \boldsymbol{y}_n + h \sum_{j=1}^{s} \int_0^{(t-t_n)/h} \frac{\omega_j(\tau)}{\omega_j(c_j)} \,\mathrm{d}\tau \boldsymbol{f}(t_n + c_j h, \boldsymbol{u}(t_n + c_j h)).$$

Letting

$$\boldsymbol{k}_j := \boldsymbol{f}(t_n + c_j h, \boldsymbol{u}(t_n + c_j h)), \quad j = 1, \ldots, s,$$

we have

$$\boldsymbol{u}(t_n + c_l h) = \boldsymbol{y}_n + h \sum_{j=1}^{s} a_{l,j} \boldsymbol{k}_j, \qquad j = 1, \ldots, s,$$

and

$$\boldsymbol{y}_{n+1} = \boldsymbol{u}(t_n + h) = \boldsymbol{y}_n + h \sum_{l=1}^{s} b_l \boldsymbol{k}_l.$$

This and the definition (2.14) of an RK method prove the lemma. □

An intermezzo: numerical quadrature. Let w be a positive weight function in (a, b). We say that the *quadrature*

$$\int_a^b g(\tau) w(\tau) \,\mathrm{d}\tau \approx \sum_{l=1}^{s} b_l g(c_l) \tag{2.18}$$

is of order p if it is correct for all $g \in \mathbb{P}_{p-1}$. (For connoisseurs of mathematical analysis: instead of a terminology of weight functions, we may use, with greater generality, Borel measures $\mathrm{d}\mu(t) = \omega(t)\,\mathrm{d}t$.)

We denote by $p_s \in \mathbb{P}_s$ an sth orthogonal polynomial, i.e. $p_s \not\equiv 0$, $\int_a^b q(\tau) p_s(\tau) w(\tau) \,\mathrm{d}\tau = 0 \ \forall q \in \mathbb{P}_{s-1}$.

Theorem 12 *Let $c_1, \ldots, c_s$ be the zeros of p_s and let $b_1, \ldots, b_s$ be the solution of the (nonsingular) Vandermonde linear system $\sum_{l=1}^{s} b_l c_l^j = \int_a^b \tau^j w(\tau) \,\mathrm{d}\tau$, $j = 0, \ldots, s-1$. Then*

(a) *(2.18) is of order exactly $2s$,*

(b) *Every other quadrature must be of order $\leq 2s-1$.*

Proof Let $\mu_j := \int_a^b \tau^j w(\tau)\,\mathrm{d}\tau$, $j \in \mathbb{Z}_+$, be the *moments* of w. Then order $2s$ is equivalent to $\sum_{l=1}^s b_l c_l^j = \mu_j$, $j = 0, \ldots, 2s-1$. In other words, it is equivalent to

$$\sum_{j=0}^{2s-1} \alpha_j \mu_j = \sum_{l=1}^{s} b_l \sum_{j=0}^{2s-1} \alpha_j c_l^j$$

for any $\alpha_0, \ldots, \alpha_{2s-1} \in \mathbb{R}$. Choose $\sum_{j=0}^{2s-1} \alpha_j t^j = p_s(t)q(t)$, where $q \in P_{s-1}$. Then

$$\sum_{j=0}^{2s-1} \alpha_j \mu_j = \int_a^b p_s(\tau)q(\tau)w(\tau)\,\mathrm{d}\tau = 0 \tag{2.19}$$

and

$$\sum_{l=0}^{s} b_l \sum_{j=0}^{2s-1} \alpha_j c_l^j = \sum_{l=0}^{s} b_l p_s(c_l)q(c_l) = 0. \tag{2.20}$$

We prove first that (2.18) is of order $2s$. Expressing $v \in P_{2s-1}$ as $v = p_s q + \tilde{v}$, where $q, \tilde{v} \in P_{s-1}$, the definition of $b_1, \ldots, b_s$ means that

$$\sum_{l=1}^{s} b_l \tilde{v}(c_l) = \sum_{j=0}^{s-1} \tilde{v}_j \sum_{l=1}^{s} b_l c_l^j = \sum_{j=0}^{s-1} \tilde{v}_j \mu_j = \int_a^b \tilde{v}(\tau)\,\mathrm{d}\tau.$$

This, in tandem with (2.19) and (2.20), proves that (2.18) is of order $2s$.

It is of order *exactly* $2s$, since $\int_a^b [p_s(\tau)]^2 w(\tau)\,\mathrm{d}\tau > 0$, whereas $\sum_{l=1}^s b_l [p_s(c_l)]^2 = 0$.

To prove that no other method can match or exceed this order, we choose $q = L_m$ (the m Lagrange interpolation polynomial), $m \in \{0, 1, \ldots, s\}$. It follows from (2.20) that $b_m p_s(c_m) = 0$. It is impossible that $b_m = 0$, otherwise the $(s-1)$-point method omitting (b_m, c_m) will be of order $> 2s-2$, and this leads to a contradiction, identically to the last paragraph. Hence $p_s(c_m) = 0$, $m = 1, \ldots, s$. □

Corollary 2 *Quadrature (2.18) is of order $s + r$ for $r \in \{0, 1, \ldots, s\}$ iff $b_1, \ldots, b_s$ are chosen as in the theorem, whereas*

$$\int_0^1 \tau^j \omega(\tau) w(\tau)\,\mathrm{d}\tau = 0, \quad j = 0, \ldots, r-1 \qquad \textit{where} \qquad \omega(t) = \prod_{k=1}^{s} (t - c_k). \tag{2.21}$$

Corollary 3 *Letting $(a, b) = (0, 1)$ and $w \equiv 1$, the highest order of quadrature is obtained when $c_1, \ldots, c_s$ are the zeros of a Legendre polynomial P_s, shifted from $[-1, 1]$ to $[0, 1]$.*

Back to collocation…
Frequently – and this is the case with collocation – we have a numerical solution which is a smooth (i.e. C^1) function $\boldsymbol{u}$, say (rather than merely having the solution at grid points). In that case we can evaluate the *defect*, i.e. the "departure" $\boldsymbol{f}(t, \boldsymbol{u}(t)) - \boldsymbol{u}'(t)$ from the solution of the exact ODE. How much does its magnitude tell us about the numerical error?

Theorem 13 (The Alekseev–Gröbner Lemma) *Let $\boldsymbol{u}$ be a smooth function s.t. $\boldsymbol{u}(t_0) = \boldsymbol{y}(t_0)$, where $\boldsymbol{y}$ solves $\boldsymbol{y}' = \boldsymbol{f}(t, \boldsymbol{y})$. Then*

$$\boldsymbol{u}(t) - \boldsymbol{y}(t) = \int_{t_0}^{t} \Phi(t, \tau, \boldsymbol{u}(\tau))[\boldsymbol{f}(\tau, \boldsymbol{u}(\tau)) - \boldsymbol{u}'(\tau)]\,\mathrm{d}\tau,$$

where Φ is the matrix of partial derivatives of the solution of $\boldsymbol{v}' = \boldsymbol{f}(t, \boldsymbol{v})$, $\boldsymbol{v}(\tau) = \boldsymbol{u}(\tau)$, w.r.t. $\boldsymbol{u}(\tau)$.

Theorem 14 *Provided that ω obeys (2.21) with $(a,b)=(0,1)$, $w\equiv 1$ and $r\in\{0,1,\dots,s\}$, the collocation method is of order $s+r$.*

Proof By estimating $\boldsymbol{u}-\boldsymbol{y}$ with the Alekseev–Gröbner lemma and approximating the underlying integral with the corresponding quadrature rule. □

Corollary 4 *The highest-order s-stage RK method corresponds to collocation at shifted Legendre points (Gauss–Legendre RK, of order $2s$).*

Examples:

$$\begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array}, \qquad \begin{array}{c|cc} \frac{1}{2}-\frac{\sqrt{3}}{6} & \frac{1}{4} & \frac{1}{4}-\frac{\sqrt{3}}{6} \\ \frac{1}{2}+\frac{\sqrt{3}}{6} & \frac{1}{4}+\frac{\sqrt{3}}{6} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}.$$

2.7 Stability of RK methods

A-Stability Solving $y'=\lambda y$, $y(0)=1$, $h=1$, and denoting the vector of stages by $\boldsymbol{k}$, we have $\boldsymbol{k}=\lambda(\mathbf{1}+A\boldsymbol{k})\in\mathbb{R}^s$, thus $\boldsymbol{k}=\lambda(I-\lambda A)^{-1}\mathbf{1}$. We obtain $y_{n+1}=R(\lambda)y_n$, where

$$\begin{aligned} R(\lambda) &= 1+\lambda\boldsymbol{b}^\top(I-\lambda A)^{-1}\mathbf{1}=\boldsymbol{b}^\top(I+\lambda(I-\lambda A)^{-1})\mathbf{1} \\ &= \boldsymbol{b}^\top(I-\lambda A)^{-1}(I-\lambda(A-I))\mathbf{1}=\frac{1}{\det(I-\lambda A)}\boldsymbol{b}^\top\operatorname{adj}(I-\lambda A)(I-\lambda(A-I))\mathbf{1}. \end{aligned}$$

It follows that R is a rational function in $\mathbb{P}_{s/s}$.

Lemma 15 *The Gauss–Legendre RK is A-stable.*

Proof $R\in\mathbb{P}_{s/s}$ and it approximates $\exp z$ of order $2s$, hence it necessarily is the s/s Padé approximation. Thus A-stability. □

Nonlinear stability analysis. Suppose that it is known that

$$\langle\boldsymbol{u}-\boldsymbol{v},\boldsymbol{f}(\boldsymbol{u})-\boldsymbol{f}(\boldsymbol{v})\rangle\le 0,\qquad \boldsymbol{u},\boldsymbol{v}\in\mathbb{R}^d, \tag{2.22}$$

where $\langle\,\cdot\,,\,\cdot\,\rangle$ is an inner product. Let $\|\boldsymbol{u}\|:=\sqrt{\langle\boldsymbol{u},\boldsymbol{u}\rangle}$.

Lemma 16 *The solution of $\boldsymbol{y}'=\boldsymbol{f}(\boldsymbol{y})$ is* dissipative, *i.e. $\|\boldsymbol{u}(t)-\boldsymbol{v}(t)\|$ is monotonically nonincreasing for any two solutions $\boldsymbol{u}(t)$ and $\boldsymbol{v}(t)$ and $t\ge 0$.*

Proof Let $\phi(t):=\|\boldsymbol{u}(t)-\boldsymbol{v}(t)\|^2$. Then, by (2.22),

$$\tfrac{1}{2}\phi'(t)=\langle\boldsymbol{u}(t)-\boldsymbol{v}(t),\boldsymbol{u}'(t)-\boldsymbol{v}'(t)\rangle=\langle\boldsymbol{u}(t)-\boldsymbol{v}(t),\boldsymbol{f}(\boldsymbol{u}(t))-\boldsymbol{f}(\boldsymbol{v}(t))\rangle\le 0.$$

Hence ϕ is monotonically nonincreasing. □

Do RK methods share this feature?

Herewith, $\langle \cdot, \cdot \rangle$ is the standard *Euclidean* inner product. Denote the stages in the nth step by $\boldsymbol{k}_1, \ldots, \boldsymbol{k}_s$ (for $\boldsymbol{u}$) and by $\boldsymbol{l}_1, \ldots, \boldsymbol{l}_s$ (for $\boldsymbol{v}$). Then

$$\|\boldsymbol{u}_{n+1} - \boldsymbol{v}_{n+1}\|^2 = \|\boldsymbol{u}_n - \boldsymbol{v}_n\|^2 + 2h \left\langle \boldsymbol{u}_n - \boldsymbol{v}_n, \sum_j b_j(\boldsymbol{k}_j - \boldsymbol{l}_j) \right\rangle + h^2 \left\| \sum_j b_j(\boldsymbol{k}_j - \boldsymbol{l}_j) \right\|^2 .$$

Thus, for $\|\boldsymbol{u}_{n+1} - \boldsymbol{v}_{n+1}\| \leq \|\boldsymbol{u}_n - \boldsymbol{v}_n\|$ we require

$$2 \left\langle \boldsymbol{u}_n - \boldsymbol{v}_n, \sum_j b_j(\boldsymbol{k}_j - \boldsymbol{l}_j) \right\rangle + h \left\| \sum_j b_j(\boldsymbol{k}_j - \boldsymbol{l}_j) \right\|^2 \leq 0. \tag{2.23}$$

Let $\boldsymbol{d}_j := \boldsymbol{k}_j - \boldsymbol{l}_j$ and set

$$\boldsymbol{p}_j := \boldsymbol{u}_n + h \sum_{i=1}^{s} a_{j,i} \boldsymbol{k}_i, \quad \boldsymbol{q}_j := \boldsymbol{v}_n + h \sum_{i=1}^{s} a_{j,i} \boldsymbol{l}_i, \qquad j = 1, \ldots, s.$$

Then $\boldsymbol{k}_j = \boldsymbol{f}(\boldsymbol{p}_j)$, $\boldsymbol{l}_j = \boldsymbol{f}(\boldsymbol{q}_j)$, $j = 1, 2, \ldots, s$, and, provided that $b_1, \ldots, b_s \geq 0$,

$$\begin{aligned}
\left\langle \boldsymbol{u}_n - \boldsymbol{v}_n, \sum_j b_j \boldsymbol{d}_j \right\rangle &= \sum_j b_j \left\langle \boldsymbol{p}_j - h \sum_i a_{j,i} \boldsymbol{k}_i - \boldsymbol{q}_j + h \sum_i a_{j,i} \boldsymbol{l}_i, \boldsymbol{d}_j \right\rangle \\
&= \sum_j b_j \left\{ \langle \boldsymbol{p}_j - \boldsymbol{q}_j, \boldsymbol{f}(\boldsymbol{p}_j) - \boldsymbol{f}(\boldsymbol{q}_j) \rangle - h \sum_i a_{j,i} \langle \boldsymbol{d}_i, \boldsymbol{d}_j \rangle \right\} \\
&\leq -h \sum_{i,j} b_j a_{j,i} \boldsymbol{d}_j^\top \boldsymbol{d}_i.
\end{aligned}$$

Thus

$$\frac{2}{h} \left\langle \boldsymbol{u}_n - \boldsymbol{v}_n, \sum_j b_j \boldsymbol{d}_j \right\rangle + \left\| \sum_j b_j \boldsymbol{d}_j \right\|^2 = \sum_{i,j} \boldsymbol{d}_j^\top (b_i b_j - b_j a_{j,i} - b_i a_{i,j}) \boldsymbol{d}_i = - \sum_{i,j} \boldsymbol{d}_i^\top m_{i,j} \boldsymbol{d}_j,$$

where $m_{i,j} := b_i a_{i,j} + b_j a_{j,i} - b_i b_j$. Suppose, though, that the symmetric matrix $M = (m_{i,j})$ is positive semidefinite and denote the matrix with the columns $\boldsymbol{d}_1, \ldots, \boldsymbol{d}_s$ by D. Let $\boldsymbol{\delta}_1^\top, \ldots, \boldsymbol{\delta}_s^\top$ be the rows of D. Then

$$\sum_{i,j} \boldsymbol{d}_i^\top m_{i,j} \boldsymbol{d}_j = \sum_{i,j,k} d_{i,k} m_{i,j} d_{j,k} = \sum_k \sum_{i,j} d_{i,k} m_{i,j} d_{j,k} = \sum_k \boldsymbol{\delta}_k^\top M \boldsymbol{\delta}_k \geq 0.$$

□

Theorem 17 (Butcher) *A RK method is* algebraically stable *(i.e., mimics the dissipation) iff* $\boldsymbol{b} \geq \boldsymbol{0}$ *and* M *is negative semidefinite.*

2.8 Additional ODE problems and themes

- **ODEs as nonlinear dynamical systems** – Increasing insight into the asymptotics of ODE solvers has been gleaned in recent years by treating a numerical method as a map that approximates the ODE flow, analysing its dependence on parameters like $\boldsymbol{y}_0$, h etc.

1. It can be shown that certain methods may display, in a fixed-step implementation, spurious modes of behaviour, inclusive of fixed points, oscillations, Hopf bifurcations or chaos. Other methods are more immune to this phenomena.
2. Certain methods are better than others in displaying correct asymptotic behaviour (omega sets, bifurcations etc.). Likewise, some error control techniques are safer than others.

- **Geometric integration** – Most of advance in numerical ODEs in the last decade occurred in *geometric integration:* computation of initial-value problems for ODEs (and PDEs) while retaining *exactly* their known invariants of mathematical and physical significance. For example

  1. *Symplectic methods:* Main qualitative features of *Hamiltonian* ODEs are conserved by methods that conserve the symplectic invariant, e.g. by Gauss–Legendre RK (but only as long as the step size remains fixed!). This is important because Hamiltonian ODEs are of crucial importance in many subjects, their solution is typically desired along extended time intervals and they exhibit very 'sensitive' behaviour, that can be easily corrupted by a numerical method.
  2. *Differential equations on manifolds:* An example: $Y' = A(Y)Y$, $Y(0) = Y_0$, where Y_0 is a $d \times d$ *orthogonal* matrix and the function A maps orthogonal to skew-symmetric matrices. (Equations of this form widely occur in robotics and in the engineering of mechanical systems.) It is easy to prove that $Y(t)$ remains orthogonal for all $t \geq 0$. Yet, most numerical methods destroy orthogonality! In greater generality, it is often known that the solution of an ODE system possesses an *invariant* (equivalently, evolves on a manifold) and a new generation of numerical methods attempts to discretize while retaining this qualitative feature.

- **High-order equations** – An example: $y''(t) = f(t, y, y')$, $y(0) = y_0$, $y'(0) = y'_0$. In principle, they can be always converted into an ODE system by letting $y_1 = y$, $y_2 = y'$. However, because of their ubiquity, there are special variants of multistep (*Numerov's method*) and RK–Nystrom methods for second-order equations.

- **Two-point boundary value problems** – An example: $y'' = f(t, y, y')$, $y(0) = a$, $y'(1) = b$. Typical methods: *shooting*, *finite differences* and *finite elements*.

  1. **Shooting:** The idea is to treat $c = y'(0)$ as a parameter. Thus, $y(t) = y(t; y(0), y'(0))$ and we try to find c so that $y(1; a, c) = b$. In reality, $y(1; a, c)$ is evaluated by an initial-value ODE solver and nonlinear iterative techniques (recall Newton–Raphson) are employed to find the right value of c.
  2. **Finite differences:** Discretize the derivative locally, e.g.

     $$\tfrac{1}{h^2}(y_{n+1} - 2y_n + y_{n-1}) = f\left(nh, y_n, \tfrac{1}{2h}(y_{n+1} - y_{n-1})\right), \quad n = 1, 2, \ldots, N-1,$$

     where $h = \frac{1}{N}$. This, together with $y(0) = a$, $(y_N - y_{N-1})/h = b$, say, yields a (typically) nonlinear algebraic system.
  3. **Finite elements:** Discussion deferred till later.

- **Differential delay equations (DDE)** – An example: $y'(t) = f(y(t - \tau))$, $y(t) = \phi(t)$ for $-\tau < t \leq 0$. Can be solved by a 'continuous' extension of ODE methods. However, the main source of problems is analytic: the DDEs are not ODEs. Thus, the solution of the equation is, in general, of low smoothness at $\{m\tau\}_{m=0}^{\infty}$, even if f is analytic – unlike ODEs, where analytic f implies an analytic solution. Even more curious is the solution of $y'(t) = -y(qt)$, $y(0) = 1$, where $q \in (0, 1)$. Even if q is arbitrarily near to 1, $|y|$ cannot be uniformly bounded!

- **Differential algebraic equations (DAE)** – An example: $x' = f(t,x,y)$, $0 = g(t,x,y)$. In other words, the solution is forced (at the price of tying down some degrees of freedom) to live on a nonlinear, multivariate manifold. Again, it is misleading to treat DAEs as ODEs, disregarding their special nature. There exists an extensive modern theory, inclusive of DAE extensions of RK, BDF and other standard methods.

Exercises

2.1 Let

$$R(z) = \frac{1 + (1-a)z + \left(b - a + \frac{1}{2}\right) z^2}{1 - az + bz^2}.$$

1. Determine p such that $R(z) = \mathrm{e}^z + \mathcal{O}\left(z^{p+1}\right)$.
2. Write the one-step two-derivative order-p method that 'corresponds' to the rational function R.
3. Determine conditions on a and b so that the method is A-stable.

2.2 Prove that Padé approximations to $\exp z$ are unique: Let $R_k(z) = P_k(z)/Q_k(z)$, $\deg P_k = m$, $\deg Q_k = n$, $Q_k(0) = 1$, $R_k(z) = \mathrm{e}^z + \mathcal{O}\left(z^{m+n+1}\right)$, $k = 1, 2$. Then necessarily $R_1 \equiv R_2$.

2.3 Let integer $m, n \geq 0$ be given and

$$P_{m/n}(z) := \sum_{k=0}^{m} \binom{m}{k} \frac{(m+n-k)!}{(m+n)!} z^k;$$

$$Q_{m/n}(z) := \sum_{k=0}^{n} \binom{n}{k} \frac{(m+n-k)!}{(m+n)!} (-z)^k = P_{n/m}(-z).$$

Set

$$\psi_{m/n}(z) := P_{m/n}(z) - \mathrm{e}^z Q_{m/n}(z).$$

1. Prove that the $\psi_{m/n}$'s obey the recurrence relation

$$\psi_{m/n}(z) = \psi_{m/(n-1)}(z) - \frac{mz}{(m+n-1)(m+n)} \psi_{(m-1)/(n-1)}(z), \quad m, n \geq 1.$$

2. Prove by induction or otherwise that

$$\psi_{m/n}(z) = (-1)^{n-1} \sum_{k=0}^{\infty} \frac{m!(k+n)!}{(n+m)!k!(k+n+m+1)!} z^{k+n+m+1}.$$

Deduce the explicit form of Padé approximations to e^z.

2.4 1. The equation $\boldsymbol{y}' = \boldsymbol{f}(t, \boldsymbol{y})$, $\boldsymbol{y}(t_0) = \boldsymbol{y}_0$, is solved by consecutive steps of forward and backward Euler,

$$\begin{aligned} \boldsymbol{y}_{2n+1} &= \boldsymbol{y}_{2n} + h\boldsymbol{f}(t_{2n}, \boldsymbol{y}_{2n}), \\ \boldsymbol{y}_{2n+2} &= \boldsymbol{y}_{2n+1} + h\boldsymbol{f}(t_{2n+2}, \boldsymbol{y}_{2n+2}). \end{aligned}$$

Prove that the sequence $\{\boldsymbol{y}_{2n}\}_{n\geq 0}$ approximates $\{\boldsymbol{y}(t_{2n})\}_{n\geq 0}$ to second order.
Is the method A-stable?

2. The same equation is solved by the combination

$$\begin{aligned} \boldsymbol{y}_{3n+1} &= \boldsymbol{y}_{3n} + h_0 \boldsymbol{f}(t_{3n}, \boldsymbol{y}_{3n}), \\ \boldsymbol{y}_{3n+2} &= \boldsymbol{y}_{3n+1} + \tfrac{1}{2} h_1 \left(\boldsymbol{f}(t_{3n+1}, \boldsymbol{y}_{3n+1}) + \boldsymbol{f}(t_{3n+2}, \boldsymbol{y}_{3n+2}) \right), \\ \boldsymbol{y}_{3n+3} &= \boldsymbol{y}_{3n+2} + h_0 \boldsymbol{f}(t_{3n+3}, \boldsymbol{y}_{3n+3}), \end{aligned}$$

of forward Euler, the trapezoidal rule and backward Euler. Prove that there exist no $h_0, h_1 > 0$ such that $\{\boldsymbol{y}_{3n}\}_{n\geq 0}$ approximates $\{\boldsymbol{y}((2h_0 + h_1)n)\}_{n\geq 0}$ to third order.

2.5 Determine the range of the real parameter α such that the multistep method

$$\begin{aligned}&\boldsymbol{y}_{n+3} - (1+2\alpha)\boldsymbol{y}_{n+2} + (1+2\alpha)\boldsymbol{y}_{n+1} - \boldsymbol{y}_n \\ &= \tfrac{1}{6}h[(5+\alpha)\boldsymbol{f}(\boldsymbol{y}_{n+3}) - (4+8\alpha)\boldsymbol{f}(\boldsymbol{y}_{n+2}) + (11-5\alpha)\boldsymbol{f}(\boldsymbol{y}_{n+1})]\end{aligned}$$

is convergent.

What is the order of the method for different values of α?

For which values of α is the method A-stable?

2.6 Derive the coefficients of the BDF methods for $m = 2, 3, 4$. Are these methods A-stable?

2.7 Consider the two-step (one-derivative) methods of order $p \geq 2$.

1. Show that they form a two-parameter family.
2. Characterise all the A-stable methods of this kind.
3. Find the A-stable method with the least magnitude of the error constant.

2.8 We say that a method is $R^{[1]}$ if, for any ODE system $\boldsymbol{y}' = \boldsymbol{f}(\boldsymbol{y})$ with continuous $\boldsymbol{f}$, the existence and boundedness of the limit $\hat{\boldsymbol{y}} = \lim_{n\to\infty} \boldsymbol{y}_n$ (with any constant step-size $h > 0$) implies that $\hat{\boldsymbol{y}}$ is a fixed point of the ODE (i.e. $\boldsymbol{f}(\hat{\boldsymbol{y}}) = \boldsymbol{0}$).

1. Prove that every convergent multistep method (iterated to convergence, if implicit) is $R^{[1]}$.
2. Show that the second-order Runge–Kutta method

$$\begin{array}{c|cc} 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \hline & 0 & 1 \end{array}$$

is not $R^{[1]}$. *[Hint: Consider the logistic equation $y' = \kappa y(1-y)$.]*

2.9 A method is $R^{[2]}$ if, for all equations $\boldsymbol{y}' = \boldsymbol{f}(\boldsymbol{y})$, there exists no solution sequence (with any constant step-size $h > 0$) such that both $\hat{\boldsymbol{y}}_o := \lim_{n\to\infty} \boldsymbol{y}_{2n+1}$ and $\hat{\boldsymbol{y}}_e := \lim_{n\to\infty} \boldsymbol{y}_{2n}$ exist, are bounded and $\hat{\boldsymbol{y}}_o \neq \hat{\boldsymbol{y}}_e$ (such solution sequence is necessarily false!). Prove that, for any convergent multistep method determined by the polynomials (ρ, σ) (that are relatively prime, i.e. have no zeros in common), $R^{[2]}$ is equivalent to $\sigma(-1) = 0$.

2.10 A multistep *one-leg* method for $\boldsymbol{y}' = \boldsymbol{f}(t, \boldsymbol{y})$ is defined as

$$\sum_{l=0}^{k} \rho_l \boldsymbol{y}_{n-k+l} = h\boldsymbol{f}\left(\sum_{l=0}^{k} \sigma_l t_{n-k+l}, \sum_{l=0}^{k} \sigma_l \boldsymbol{y}_{n-k+l}\right).$$

Letting $\rho(z) := \sum_0^k \rho_l z^l$, $\sigma(z) := \sum_0^k \sigma_l z^l$, derive necessary and sufficient conditions on $\{\rho, \sigma\}$ for *(a)* order 2; and *(b)* A-stability.

2.11 Derive the order of the one-leg method (*implicit midpoint rule*)

$$\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + h\boldsymbol{f}\left(t_n + \tfrac{1}{2}h, \tfrac{1}{2}(\boldsymbol{y}_n + \boldsymbol{y}_{n+1})\right).$$

Is it A-stable? It is $R^{[1]}$? $R^{[2]}$?

2.12 We say that an ODE method is *conservative* if, given that the exact solution of $\boldsymbol{y}' = \boldsymbol{f}(\boldsymbol{y})$ obeys a quadratic conservation law of the form $\boldsymbol{y}(t)^\top S\boldsymbol{y}(t) \equiv c$, $t \geq 0$, where S is a symmetric, positive-definite matrix and c is a (positive) constant, it is also true that $\boldsymbol{y}_n^\top S\boldsymbol{y}_n \equiv c$, $n = 0, 1, \ldots$. Methods like this are important in the solution of Hamiltonian systems. Prove that the one-leg method from Exercise 11 is conservative.

2.13 Find the order of the explicit Runge-Kutta method

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array}.$$

2.14 Determine conditions on $\boldsymbol{b}$, $\boldsymbol{c}$, and A such that the method

$$\begin{array}{c|cc} c_1 & a_{1,1} & a_{1,2} \\ c_2 & a_{2,1} & a_{2,2} \\ \hline & b_1 & b_2 \end{array}$$

is of order $p \geq 3$.

2.15 Prove that the implicit Runge–Kutta scheme

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{5}{24} & \frac{1}{3} & -\frac{1}{24} \\ 1 & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \\ \hline & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \end{array}$$

can be expressed as a collocation method with a cubic collocation polynomial. Determine the order of this scheme.

2.16 Let a ν-stage Runge–Kutta method be defined by collocation with the collocation points $c_1, c_2, \ldots, c_\nu$, which are distinct. Suppose that the polynomial $\omega(t) := \prod_{l=1}^{\nu}(t - c_l)$ can be expressed in the form

$$\omega(t) = \alpha \tilde{P}_\nu(t) + \beta \tilde{P}_{\nu-1}(t),$$

where $\tilde{P}_n$ is the nth Legendre polynomial, shifted to the interval $[0, 1]$. (Hence $\int_0^1 \tau^j \tilde{P}_n(\tau)\,\mathrm{d}\tau = 0$, $j = 0, 1, \ldots, n-1$.)

1. Prove that the method is at least of order $2\nu - 1$.
2. The constants α and β are chosen so that the matrix A is invertible and $\boldsymbol{b}^\top A^{-1}\mathbf{1} = 1$. Prove that the stability function is a $(\nu - 1)/\nu$ rational function, hence deduce that the method is A-stable.

2.17 The function R is a rational fourth-order approximation to $\exp z$. We solve $\boldsymbol{y}' = \boldsymbol{f}(\boldsymbol{y})$, $\boldsymbol{y}(t_0) = \boldsymbol{y}_0$ by the numerical scheme

$$\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + A^{-1}\left(R(hA) - I\right)\boldsymbol{f}(\boldsymbol{y}_n),$$

where A is a nonsingular matrix that may depend on n.

1. Prove that

$$A = \frac{\partial \boldsymbol{f}(\boldsymbol{y}(t_n))}{\partial \boldsymbol{y}} + \mathcal{O}(h)$$

gives a second-order method.

2. Discuss the stability properties of the above method.

Bibliography

[1] E. Hairer, S.P. Nørsett and G. Wanner, *Solving Ordinary Differential Equations I: Nonstiff Problems*, Springer–Verlag, Berlin, 1987.

[2] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations II: Stiff Problems and Differenial Algebraic Equations*, Springer–Verlag, Berlin, 1991.

[3] P. Henrici, *Discrete Variable Methods in Ordinary Differential Equations*, Wiley, New York, 1962.

[4] J.D. Lambert, *Numerical Methods for Ordinary Differential Equations*, Wiley, London, 1991.

3 Finite difference methods for PDEs

3.1 Calculus of finite differences

Given $\{y_n\}_{n=-\infty}^{\infty}$, we define

$\mathrm{E}y_n = y_{n+1}$	The shift operator
$\Delta_+ y_n = y_{n+1} - y_n$	The forward difference operator
$\Delta_- y_n = y_n - y_{n-1}$	The backward difference operator
$\Delta_0 y_n = y_{n+\frac{1}{2}} - y_{n-\frac{1}{2}}$	The central difference operator
$\mu_0 y_n = \frac{1}{2}(y_{n+\frac{1}{2}} + y_{n-\frac{1}{2}})$	The averaging operator.

Note that Δ_0 and μ_0 are ill-defined – but watch this space!

Assume further that $y_n = y(nh)$, where y is analytic in $\mathbb{R}$ with radius of convergence $> h$, and define

$\mathrm{D}y_n = y'(nh)$	The differential operator.

All operators can be conveniently expressed in terms of each other. For example,

$$\mu_0 = \tfrac{1}{2}(\mathrm{E}^{\frac{1}{2}} + \mathrm{E}^{-\frac{1}{2}}) \qquad \Leftrightarrow \qquad \mathrm{E} = 2\mu_0^2 - \mathrm{I} + 2\mu_0\sqrt{\mu_0^2 - \mathrm{I}}.$$

We conclude that all the above operators commute.

Approximating D^s. Using 'slanted' (i.e. $\Delta_\pm$) differences, we have

$$\begin{aligned}\mathrm{D}^s &= \frac{1}{h^s}\left(\log(\mathrm{I}+\Delta_+)\right)^s = \frac{1}{h^s}\left\{\Delta_+^s - \tfrac{1}{2}s\Delta_+^{s+1} + \tfrac{1}{24}s(3s+5)\Delta_+^{s+2} - \cdots\right\}\\ &= \frac{(-1)^s}{h^s}\left(\log(\mathrm{I}-\Delta_-)\right)^s = \frac{1}{h^s}\left\{\Delta_-^s + \tfrac{1}{2}s\Delta_-^{s+1} + \tfrac{1}{24}s(3s+5)\Delta_-^{s+2} + \cdots\right\}.\end{aligned}$$

For example,

$$\mathrm{D}^s y_n \approx \frac{1}{h^s}\left(\Delta_+^s - \tfrac{1}{2}s\Delta_+^{s+1} + \tfrac{1}{24}s(3s+5)\Delta_+^{s+2}\right) y_n \qquad \text{(error } \mathcal{O}\big(h^3\big)\text{, bandwidth } s+2\text{)}.$$

Central differences. Although Δ_0 and μ_0 aren't well-defined on a grid, $\Delta_0^2 y_n = y_{n+1} - 2y_n + y_{n-1}$ and $\Delta_0\mu_0 y_n = \frac{1}{2}(y_{n+1} - y_{n-1})$ are!

We have $\mathrm{D} = \frac{2}{h}\log\left(\frac{1}{2}\Delta_0 + \sqrt{\mathrm{I} + \frac{1}{4}\Delta_0^2}\right)$ and we let $g(z) := \log(z + \sqrt{1+z^2})$. By the (generalized) binomial theorem

$$g'(z) = (1+z^2)^{-\frac{1}{2}} = \sum_{j=0}^{\infty}(-1)^j\binom{2j}{j}\left(\frac{z}{2}\right)^{2j}.$$

Since $g(0) = 0$, integration yields

$$g(z) = 2\sum_{j=0}^{\infty}\frac{(-1)^j}{2j+1}\binom{2j}{j}\left(\frac{z}{2}\right)^{2j+1}.$$

Hence

$$\mathrm{D} = \frac{2}{h}g\left(\tfrac{1}{2}\Delta_0\right) = \frac{4}{h}\sum_{j=0}^{\infty}\frac{(-1)^j}{2j+1}\binom{2j}{j}\left(\tfrac{1}{4}\Delta_0\right)^{2j+1} = \frac{1}{h}(\Delta_0 - \tfrac{1}{24}\Delta_0^3 + \tfrac{3}{640}\Delta_0^5 - \cdots).$$

We have

$$\mathrm{D}^s = \frac{1}{h^s}\left(\Delta_0^s - \tfrac{1}{24}s\Delta_0^{s+2} + \tfrac{1}{5760}s(5s+22)\Delta_0^{s+4} - \cdots\right). \tag{3.1}$$

This works beautifully for even s, e.g.

$$\mathrm{D}^2 y_n \approx \frac{1}{h^2}\left(\Delta_0^2 - \tfrac{1}{12}\Delta_0^4\right) y_n \qquad \text{(error } \mathcal{O}\big(h^4\big)\text{, bandwidth 4)}$$

For odd s we exploit $\mu_0 = \left(\mathrm{I} + \frac{1}{4}\Delta_0^2\right)^{\frac{1}{2}}$ to multiply (3.1) by

$$\mathrm{I} = \mu_0\left(\mathrm{I} + \tfrac{1}{4}\Delta_0^2\right)^{-\frac{1}{2}} = \mu_0 \sum_{j=0}^{\infty}(-1)^j \frac{(2j)!}{(j!)^2}\left(\frac{\Delta_0}{4}\right)^{2j}.$$

This gives

$$\mathrm{D}^s = \frac{1}{h^s}\mu_0\Delta_0\left(\Delta_0^{s-1} - \tfrac{1}{24}(s+3)\Delta_0^{s+1} + \tfrac{1}{5760}(5s^2+52s+135)\Delta_0^{s+3} - \cdots\right).$$

For example,

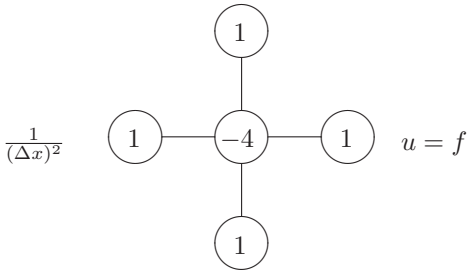
$$\mathrm{D}y_n \approx \frac{1}{h}\left(\tfrac{1}{12}y_{n-2} - \tfrac{2}{3}y_{n-1} + \tfrac{2}{3}y_{n+1} - \tfrac{1}{12}y_{n+2}\right) \qquad \text{(error } \mathcal{O}\big(h^4\big)\text{, bandwidth 4.)}$$

3.2 Synthesis of finite difference methods

An example – the Poisson equation $\nabla^2 u = f$ with Dirichlet conditions on the boundary of $[0,1]^2$. Let $u_{k,l} \approx u(k\Delta x, l\Delta x)$. We solve the Poisson equation with the *five point formula*

$$(\Delta x)^2(\Delta_{0,x}^2 + \Delta_{0,y}^2)u_{k,l} = u_{k-1,l} + u_{k+1,l} + u_{k,l-1} + u_{k,l+1} - 4u_{k,l} = (\Delta x)^2 f_{k,l}. \tag{3.2}$$

A compact notation is given via *computational stencils* (a.k.a. computational molecules). Thus, (3.2) can be written as



Computational stencils can be formally 'added', 'multiplied' etc.

Curved boundaries. It is often impossible to fit a grid into a domain so that all the intersections of the grid with the boundary are themselves grid points. The easiest quick fix is to use, when necessary, finite-difference formulae with non-equidistant points. This, in practice, means using larger stencils near (curved) boundaries.

Initial value problems. Again, we can use finite differences and computational stencils. Two approaches: *full discretization (FD)*, whereby both time and space are discretized in unison, and *semidiscretization (SD)* – only space is discretized, and this gives an ODE system. Schemes can be *explicit* or *implicit*.

Example: $u_t = u_{xx}$, $x \in [0,1]$, $t \geq 0$, with 0 b.c. at $x = 0, 1$ and initial conditions for $t = 0$ (the *diffusion equation*, a.k.a. the *heat equation*). Let $u_m(t) \approx u(m\Delta x, t)$, $u_m^n \approx u(m\Delta x, n\Delta t)$. Then

$$\text{(Explicit) SD:} \quad u_m' = \frac{1}{(\Delta x)^2}(u_{m-1} - 2u_m + u_{m+1}),$$

$$\text{(Explicit) FD:} \quad u_m^{n+1} = u_m^n + \frac{\Delta t}{(\Delta x)^2}(u_{m-1}^n - 2u_m^n + u_{m+1}^n),$$

$$\text{(Implicit) FD:} \quad u_m^{n+1} = u_m^n + \frac{\Delta t}{2(\Delta x)^2}(u_{m-1}^n - 2u_m^n + u_{m+1}^n + u_{m-1}^{n+1} - 2u_m^{n+1} + u_{m+1}^{n+1}).$$

Note that the explicit FD (*Euler's method*) is the result of SD followed by forward Euler, whereas the implicit FD (*Crank–Nicolson*) is obtained by solving the SD equations with the trapezoidal rule.

An analytic approach. Works for linear equations with constant coefficients. It is obvious in a hand-waiving manner, but more rigourous justification requires tools like Fourier analysis.

We consider first the *Poisson equation* $\nabla^2 u = f$ in $\Omega \subset \mathbb{R}^2$, with Dirichlet b.c. $u = g$, $(x,y) \in \partial\Omega$ or Neumann b.c. $\frac{\partial}{\partial n} u = g$, $(x,y) \in \partial\Omega$. It is approximated by the linear combination

$$\mathcal{L}_{\Delta x} u_{k,l} := \sum_{(i,j)\in I} a_{i,j} u_{k+i,l+j} = (\Delta x)^2 f_{k,l}. \tag{3.3}$$

Recalling that $\nabla^2 = \mathrm{D}_x^2 + \mathrm{D}_y^2 = (\Delta x)^{-2}[(\log \mathrm{E}_x)^2 + (\log \mathrm{E}_y)^2]$, we set

$$L(x,y) := (\log x)^2 + (\log y)^2,$$
$$L_{\Delta x}(x,y) := \sum_{(i,j)\in I} a_{i,j} x^i y^j.$$

Suppose that

$$L_{\Delta x}(x,y) = L(x,y) + \mathcal{O}((\Delta x)^{p+3}), \qquad x, y = 1 + \mathcal{O}(\Delta x)\,.$$

Let $\tilde{u}_{k,l} = u(k\Delta x, l\Delta x)$ (the exact solution). Then

$$\mathcal{L}_\Delta \tilde{u}_{k,l} - (\Delta x)^2 f_{k,l} = (\Delta x)^2 \left(\nabla^2 \tilde{u}_{k,l} - f_{k,l}\right) + \mathcal{O}\big((\Delta x)^{p+3}\big) = \mathcal{O}\big((\Delta x)^{p+3}\big)\,.$$

Subtracting $\mathcal{L}_\Delta u_{k,l} - (\Delta x)^2 f_{k,l} = 0$ gives

$$\mathcal{L}_{\Delta x}(\boldsymbol{u} - \tilde{\boldsymbol{u}}) = \mathcal{O}\big((\Delta x)^{p+3}\big)\,.$$

But $L(x,y) = \mathcal{O}\big((\Delta x)^2\big)$ (since $L(x,y) = (1-x)^2 + (1-y)^2 + \text{h.o.t.}$) and the implicit function theorem implies (in simple geometries, with 'nice' boundary conditions) that the error in (3.3) is $\mathcal{O}\big((\Delta x)^{p+1}\big)$.

Example: The 5-point formula (3.2):

$$L_{\Delta x}(\mathrm{e}^{\mathrm{i}\theta}, \mathrm{e}^{\mathrm{i}\psi}) = -4\left(\sin^2\tfrac{\theta}{2} + \sin^2\tfrac{\psi}{2}\right) = -(\theta^2 + \psi^2) + \mathcal{O}\big((\Delta x)^4\big) = L(\mathrm{e}^{\mathrm{i}\theta}, \mathrm{e}^{\mathrm{i}\psi}) + \mathcal{O}\big((\Delta x)^4\big)\,,$$

hence error $\mathcal{O}\big((\Delta x)^2\big)$.

Example: The 9-point formula

$$\frac{1}{(\Delta x)^2} \begin{array}{ccc} \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \\ \frac{2}{3} & -\frac{10}{3} & \frac{2}{3} \\ \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \end{array} \qquad u = f$$

We can prove, by proceeding as before, that

$$L_{\Delta x}(\mathrm{e}^{\mathrm{i}\theta}, \mathrm{e}^{\mathrm{i}\psi}) = -(\theta^2+\psi^2) + \tfrac{1}{12}(\theta^2+\psi^2)^2 + \mathcal{O}\big((\Delta x)^6\big)\,.$$

Hence, as before, the error is $\mathcal{O}\big((\Delta x)^2\big)$. However, when $f \equiv 0$, we are solving the equation

$$\left(1 + \frac{(\Delta x)^2}{12}\nabla^2\right)\nabla^2 u = 0 \tag{3.4}$$

to $\mathcal{O}\big((\Delta x)^4\big)$ – and, disregarding the highly nontrivial matter of boundary conditions, for small Δx, the differential operator on the left is invertible and the solutions of (3.4) and of $\nabla^2 u = 0$ coincide. Consequently, the 9-point formula carries local error of $\mathcal{O}\big((\Delta x)^2\big)$ for Poisson, but $\mathcal{O}\big((\Delta x)^4\big)$ for Laplace!

Mehrstellenverfahren. How to extend the benefits of the 9-point formula to the Poisson equation? We have seen that $L_\Delta = L - \frac{1}{12}L^2 + \mathcal{O}\big((\Delta x)^6\big)$. Let

$$\mathcal{M}_{\Delta x} = \sum_{(i,j)\in J} b_{i,j}\mathrm{E}_x^i\mathrm{E}_y^j := M_{\Delta x}(\mathrm{E}_x, \mathrm{E}_y)$$

be a finite-difference operator such that $M_{\Delta x}(x,y) = 1 + \frac{1}{12}L(x,y) + \mathcal{O}\big((\Delta x)^4\big)$. We apply $\mathcal{M}_{\Delta x}$ to the *right-hand side of the equation*, i.e. solve

$$\mathcal{L}_{\Delta x}u_{k,l} = (\Delta x)^2\mathcal{M}_{\Delta x}f_{k,l}.$$

This means that we are solving

$$\left[1 + \tfrac{1}{12}(\Delta x)^2\nabla^2\right](\nabla^2 u - f) = 0 \tag{3.5}$$

and the local error is $\mathcal{O}\big((\Delta x)^4\big)$. As before, for small Δx, the solution of (3.5) and of Poisson's equation coincide.

Another interpretation: $L(x,y)$ is being approximated by $L_{\Delta x}(x,y)/M_{\Delta x}(x,y)$, a rational function.

The d-dimensional case. Let $u = u(x_1, x_2, \ldots, x_d)$ and consider $\nabla^2 u = f$, hence we have $L(\boldsymbol{x}) = \sum_1^d (\log x_k)^2$.

Theorem 18 *Let*

$$L_{\Delta x}(\boldsymbol{x}) = -\tfrac{2}{3}(2d+1) + \tfrac{2}{3}\sum_1^d\left(x_k + \frac{1}{x_k}\right) + \tfrac{2}{3}\cdot\frac{1}{2^d}\prod_1^d\left(x_k + \frac{1}{x_k}\right),$$

$$M_{\Delta x}(\boldsymbol{x}) = 1 - \tfrac{1}{6}d + \tfrac{1}{12}\sum_1^d\left(x_k + \frac{1}{x_k}\right).$$

Then the solution of $\mathcal{L}_{\Delta x}u_{\mathbf{k}} = (\Delta x)^2\mathcal{M}_{\Delta x}f_{\mathbf{k}}$ approximates the solution of $\nabla^2 u = f$ to $\mathcal{O}\big((\Delta x)^4\big)$.

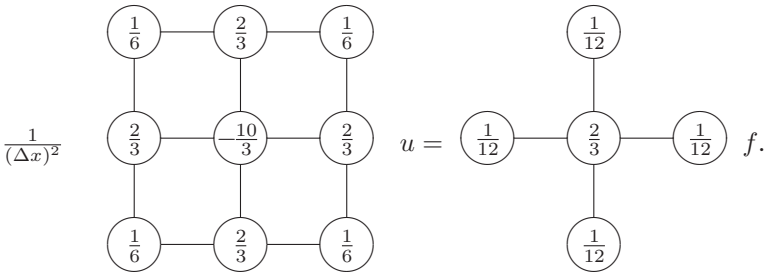
Proof Follows at once from

$$L_{\Delta x}(\mathrm{e}^{\mathrm{i}\theta_1}, \ldots, \mathrm{e}^{\mathrm{i}\theta_d}) = -\tfrac{2}{3}(2d+1) + \tfrac{4}{3}\sum_1^d\cos\theta_k + \tfrac{2}{3}\prod_1^d\cos\theta_k = L - \tfrac{1}{12}L^2 + \mathcal{O}\big((\Delta x)^6\big)\,,$$

$$M_{\Delta x}(\mathrm{e}^{\mathrm{i}\theta_1}, \ldots, \mathrm{e}^{\mathrm{i}\theta_d}) = 1 - \tfrac{1}{6}d + \tfrac{1}{6}\sum_1^d\cos\theta_k = 1 - \tfrac{1}{12}L + \mathcal{O}\big((\Delta x)^4\big)\,.$$

□

In the special case $d = 2$ we obtain



3.3 Equations of evolution

An analytic approach. Consider $u_t = \frac{\partial^L}{\partial x^L} u$ with given boundary and initial conditions. Set $\mu = \frac{\Delta t}{(\Delta x)^L}$, the *Courant number*. For example, $L = 1$ is the *advection equation*, whereas $L = 2$ yields the *diffusion equation*.

Semidiscretizations. We consider the ODEs

$$u'_m - \tfrac{1}{(\Delta x)^L} \sum_{-r}^{s} \alpha_k u_{m+k} = 0 \tag{3.6}$$

and denote the exact solution of the original PDE by $\tilde{u}$. Since $\mathrm{D}_t \tilde{u} = \mathrm{D}_x^L \tilde{u}$,

$$\begin{aligned} \tilde{u}'_m - \tfrac{1}{(\Delta x)^L} \sum_{k=-r}^{s} \alpha_k \tilde{u}_{m+k} &= \left(\mathrm{D}_t - \tfrac{1}{(\Delta x)^L} \sum_{k=-r}^{s} \alpha_k \mathrm{E}_x^k \right) \tilde{u}_m \\ &= \frac{1}{(\Delta x)^L} \left((\log \mathrm{E}_x)^L - \sum_{k=-r}^{s} \alpha_k \mathrm{E}_x^k \right) \tilde{u}_m. \end{aligned}$$

Suppose that

$$h(z) = \sum_{k=-r}^{s} \alpha_k z^k = (\log z)^L + \mathcal{O}\big(|z-1|^{p+1}\big)$$

and denote $e_m = u_m - \tilde{u}_m$. Subtracting (3.6) from

$$\tilde{u}'_m - \tfrac{1}{(\Delta x)^L} \sum_{-r}^{s} \alpha_k \tilde{u}_{m+k} = \mathcal{O}\big((\Delta x)^{p-L+1}\big)$$

yields

$$e'_m = \tfrac{1}{(\Delta x)^L} \sum_{-r}^{s} \alpha_k e_{m+k} + \mathcal{O}\big((\Delta x)^{p-L+1}\big)$$

with *zero initial and boundary conditions*. Therefore, as long as we solve the SD equations with constant μ (which takes care of $(\Delta x)^{-L}$), $e_m = \mathcal{O}\big((\Delta x)^{p+1}\big)$ and the method is of order p.

Example $u_t = u_x$, $L = 1$, $r = s = 1 \Rightarrow u'_m = \frac{1}{2\Delta x}(u_{m+1} - u_{m-1})$.

Full discretizations. Let

$$\sum_{k=-r}^{s} \gamma_k u_{m+k}^{n+1} = \sum_{k=-r}^{s} \delta_k u_{m+k}^{n}, \tag{3.7}$$

where $\gamma_k = \gamma_k(\mu)$, $\delta_k = \delta_k(\mu)$ and $\sum_{-r}^{s} \gamma_k(0) \neq 0$. Proceeding like before, we let

$$H(z;\mu) := \frac{\sum_{k=-r}^{s} \delta_k z^k}{\sum_{k=-r}^{s} \gamma_k z^k}.$$

Provided that the rational function H is irreducible (i.e., it has no common factors), the method (3.7) is of order p if

$$H(z;\mu) = \mathrm{e}^{\mu(\log z)^L} + \mathcal{O}\big(|z-1|^{p+1}\big).$$

Example $u_t = u_x$, $L = 1$, $r = s = 1 \Rightarrow u_m^{n+1} = u_m^n + \frac{\mu}{2}(u_{m+1}^n - u_{m-1}^n)$.

Well posedness. Let $u_t = \mathcal{L}u + f$, where $\mathcal{L}$ is a spatial linear differential operator. The solution is $u = \mathcal{E}(t)u_0$, where u_0 is the initial condition and $\mathcal{E}$ is the *evolution operator*. Note that $\mathcal{E}(0) = I$ and that $\mathcal{E}$ is a *semigroup*: $\mathcal{E}(t+s) = \mathcal{E}(t)\mathcal{E}(s)$. We say that the equation is *well posed* in a Banach space H (with the norm $|\cdot|$) if $|\mathcal{E}(t)| \leq C$ uniformly for all $0 \leq t \leq T$. Examples:

1. The advection equation $u_t = u_x$: Here $u(x,t) = u(x+t,0)$, hence $\mathcal{E}(t) = E_x^t$ and (provided the initial values are given on all of $\mathbb{R}$), $|\mathcal{E}| \equiv 1$.

2. The diffusion equation $u_t = u_{xx}$ with zero boundary conditions: By Fourier analysis for $x \in [-\pi, \pi]$,

$$u(x,0) = \sum_{m=-\infty}^{\infty} \alpha_m \mathrm{e}^{\mathrm{i}mx} \qquad \Longrightarrow \qquad u(x,t) = \sum_{m=-\infty}^{\infty} \alpha_m \mathrm{e}^{\mathrm{i}mx - m^2 t}. \tag{3.8}$$

   Therefore $|\mathcal{E}(t)u_0| \leq |u_0|$, hence $|\mathcal{E}| \leq 1$.

3. The 'reversed' diffusion equation $u_t = -u_{xx}$: $\mathrm{e}^{\mathrm{i}mx - m^2 t}$ is replaced by $\mathrm{e}^{\mathrm{i}mx + m^2 t}$ in (3.8) and we have a blow-up, hence no well-posedness.

Convergence. The FD scheme

$$\boldsymbol{u}_{\Delta x}^{n+1} = \mathcal{A}_{\Delta x} \boldsymbol{u}_{\Delta x}^{n} + \boldsymbol{f}_{\Delta x}^{n}, \tag{3.9}$$

where all coefficients are allowed to depend on μ, is said to be *convergent* if, given $T > 0$, for all $\Delta x \to 0$, $n, m \to \infty$, s.t. $m\Delta x \to x$, $n\Delta t \to t$ (x in the spatial domain of definition, $t \in (0, T]$) and fixed μ, $(u_m^n)_{\Delta x}$ tends to $u(x,t)$ and the progression to the limit is uniform in $t \in (0, T]$ and x. Trivial generalization to several space variables.

We let $\|\boldsymbol{u}\|_{\Delta x} = \big[(\Delta x)\sum |u_m|^2\big]^{\frac{1}{2}}$, where the sum is carried out over the grid points. Note that if $u_m = g(m\Delta x)$, where g is suitably smooth, Riemann sums imply that $\lim_{\Delta x \downarrow 0} \|\boldsymbol{u}\|_{\Delta x} = \|g\|$, where $\|g\|$ is the standard Euclidean norm acting on functions.

Stability. We say that (3.9) is *stable* (in the sense of Lax) if it is true that $\|\mathcal{A}_{\Delta x}^n\|_{\Delta x}$ is uniformly bounded when $\Delta x \to 0$ (μ being constant) and for all $n \in \mathbb{Z}_+$, $n\Delta t \in [0, T]$. **Health warning:** This concept is *different* from A-stability!

Theorem 19 (The Lax equivalence theorem) *For linear well posed PDEs of evolution, convergence is equivalent to consistency (i.e. order ≥ 1) and stability.*

SD schemes. We now consider

$$\boldsymbol{u}'_{\Delta x} = \frac{1}{(\Delta x)^L}\mathcal{P}_{\Delta x}\boldsymbol{u}_{\Delta x} + \boldsymbol{f}_{\Delta x}(t). \tag{3.10}$$

Convergence means that the solution of the ODE system (3.10) tends to the solution of the PDE when $\Delta x \to 0$, uniformly in Δx and $t \in [0, T]$.

Stability: $\|\exp(t\mathcal{P}_{\Delta x})\|$ is uniformly bounded for all $t \in [0, T]$, $\Delta x \to 0$. The Lax equivalence theorem remains valid.

3.4 Stability analysis

Von Neumann's theory I: Eigenvalue analysis. A matrix A is *normal* if $AA^* = A^*A$. Examples: Hermitian and skew-Hermitian matrices.

Lemma 20 *A matrix is normal iff it has a full set of unitary eigenvectors.*

Thus, $A = Q^*DQ$, where Q is unitary and D is diagonal.

For general matrices it is true that $\rho(A) \le \|A\|$ (in every norm). Moreover, in Euclidean norm, $\|A\| = \sqrt{\rho(A^*A)}$.

Corollary 5 *Suppose that A is normal. Then $\|A\| = \rho(A)$.*

Proof Since $\|A\| = \sqrt{\rho(A^*A)}$, $A^* = Q^*\bar{D}Q$ and because multiplication by unitary matrices is an isometry of the Euclidean norm. □

Theorem 21 *Suppose that $\mathcal{A}_{\Delta x}$ is normal for all $\Delta x \to 0$ and that there exists $\alpha \ge 0$ s.t. $\rho(\mathcal{A}_{\Delta x}) \le \mathrm{e}^{\alpha\Delta t}$. Then the method (3.9) is stable.*

Proof The Euclidean norm of a normal matrix coincides with its spectral radius. For every vector $\boldsymbol{v}_{\Delta x}$, $\|\boldsymbol{v}_{\Delta x}\|_{\Delta x} = 1$ and n s.t. $n\Delta t \le T$ it is true that

$$\begin{aligned}\|\mathcal{A}^n_{\Delta x}\boldsymbol{v}_{\Delta x}\|^2_{\Delta x} &= \langle \mathcal{A}^n_{\Delta x}\boldsymbol{v}_{\Delta x}, \mathcal{A}^n_{\Delta x}\boldsymbol{v}_{\Delta x}\rangle_{\Delta x} = \langle \boldsymbol{v}_{\Delta x}, (\mathcal{A}^n_{\Delta x})^*\mathcal{A}^n_{\Delta x}\boldsymbol{v}_{\Delta x}\rangle_{\Delta x}\\ &= \langle \boldsymbol{v}_{\Delta x}, (\mathcal{A}^*_{\Delta x}\mathcal{A}_{\Delta x})^n\boldsymbol{v}_{\Delta x}\rangle_{\Delta x} \le \|\boldsymbol{v}_{\Delta x}\|^2_{\Delta x}\|\mathcal{A}^*_{\Delta x}\mathcal{A}_{\Delta x}\|^n_{\Delta x} = [\rho(\mathcal{A}_{\Delta x})]^{2n}.\end{aligned}$$

Therefore

$$\|\mathcal{A}^n_{\Delta x}\|_{\Delta x} \le [\rho(\mathcal{A}_{\Delta x})]^n \le \mathrm{e}^{\alpha n\Delta t} \le \mathrm{e}^{\alpha T},$$

uniform boundedness. □

An alternative interpretation. The factorization $\mathcal{A} = VDV^{-1}$ implies that $\|\mathcal{A}^n\| \le \kappa(V)\|D\|^n$, where $\kappa(V) = \|V\| \cdot \|V^{-1}\|$ is the *spectral condition number*. As long as $\mathcal{A}$ is normal, V is unitary and $\kappa(V) \equiv 1$ (irrespective of Δx). However, in general it is possible that $\lim_{\Delta x\to 0}\kappa(\mathcal{V}_{\Delta x}) = \infty$. Therefore, uniformly bounded eigenvalues are necessary but *not* sufficient for stability!

Example: $u_t = u_{xx} + f$, $0 \le x \le 1$, is being solved with Euler's scheme

$$u^{n+1}_m = u^n_m + \mu(u^n_{m-1} - 2u^n_m + u^n_{m+1}) + f^n_m.$$

Hence $\mathcal{A}_{\Delta x}$ is tridiagonal and symmetric. Moreover, it is a *Toeplitz matrix*, i.e. constant along the diagonals. We denote matrices like this by *TST*. The dependence on Δx is, by the way, expressed via the matrix dimension.

Lemma 22 *(with a straightforward proof) Let A be a $d \times d$ TST matrix, $a_{k,k} = \alpha$, $a_{k,k\pm1} = \beta$. Then the eigenvalues of A are $\lambda_k = \alpha + 2\beta\cos\frac{k\pi}{d+1}$, with the corresponding (orthogonal) eigenvectors $v_{k,l} = \sin\frac{\pi kl}{2d+2}$, $k, l = 1, 2, \ldots, d$.*

In our case $\alpha = 1 - 2\mu$, $\beta = \mu$, hence $\lambda_k = 1 - 4\mu\sin^2\frac{k\pi}{2d+2}$. Consequently $\max|\lambda_k| \le 1$ means that $\mu \le \frac{1}{2}$. Since the matrix is symmetric, this is necessary and sufficient for stability.

Example We solve the diffusion equation with the Crank–Nicolson method,

$$u_m^{n+1} = u_m^n + \tfrac{1}{2}\mu(u_{m-1}^n - 2u_m^n + u_{m+1}^n + u_{m-1}^{n+1} - 2u_m^{n+1} + u_{m+1}^{n+1}) + f_m^n.$$

Now $\mathcal{A}_{\Delta x} = B^{-1}A$, where both A and B are TST, with $\alpha = 1 - \mu$, $\beta = \frac{1}{2}\mu$ and $\alpha = 1 + \mu$, $\beta = -\frac{1}{2}\mu$ respectively. Since all TST matrices commute,

$$\lambda_k = \frac{1 - 2\mu\sin^2\frac{k\pi}{2d+2}}{1 + 2\mu\sin^2\frac{k\pi}{2d+2}} \in (-1, 1), \quad k = 1, 2, \ldots, d,$$

hence stability for all $\mu > 0$.

Example The advection equation $u_t = u_x$, $x \in [0, 1]$, with 0 b.c. at $x = 1$, is solved by Euler's method

$$u_m^{n+1} = (1 - \mu)u_m^n + \mu u_{m+1}^n.$$

Hence $\mathcal{A}$ is bidiagonal with $1 - \mu$ along the main diagonal and $\rho(\mathcal{A}) = |1 - \mu|$. It follows that $0 < \mu < 2$ is *necessary* for stability. To convince ourselves that it is not sufficient, consider a general $d \times d$ matrix

$$A_d = \begin{bmatrix} a & b & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & & \ddots & 0 \\ \vdots & & 0 & a & b \\ 0 & \cdots & \cdots & 0 & a \end{bmatrix} \quad \Rightarrow \quad A_d^\top A_d = \begin{bmatrix} a^2 & ab & 0 & \cdots & 0 \\ ab & a^2+b^2 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & ab & a^2+b^2 & ab \\ 0 & \cdots & 0 & ab & a^2+b^2 \end{bmatrix}.$$

Thus, by the *Gerschgorin theorem*, $\|A_d\|^2 = \rho(A_d^\top A_d) \le (|a| + |b|)^2$. On the other hand, let $v_{d,k} = \left(\operatorname{sgn}\frac{a}{b}\right)^{k-1}$, $k = 1, \ldots, d$. It is easy to see that $\|A_d\boldsymbol{v}_d\|/\|\boldsymbol{v}_d\| \overset{d\to\infty}{\longrightarrow} |a| + |b|$. Consequently, $\|A_d\| \overset{d\to\infty}{\longrightarrow} |a| + |b|$. In our special example $a = 1 - \mu$, $b = \mu$, hence the norm is at least $\mu + |1 - \mu|$ and stability is equivalent to $0 < \mu \le 1$.

Eigenvalue analysis for SD schemes.

Theorem 23 *Let $\mathcal{P}_{\Delta x}$ be normal and suppose that $\exists\beta \in \mathbb{R}$ s.t. $\operatorname{Re}\lambda \le \beta$ for all $\lambda \in \sigma(\mathcal{P}_{\Delta x})$ and $\Delta x \to 0$. Then the SD method is stable.*

Proof Let $\|\boldsymbol{v}\| = 1$. Then

$$\begin{aligned} \|\mathrm{e}^{t\mathcal{P}}\boldsymbol{v}\|^2 &= \left\langle \boldsymbol{v}, \left(\mathrm{e}^{t\mathcal{P}}\right)^* \mathrm{e}^{t\mathcal{P}}\boldsymbol{v} \right\rangle = \left\langle \boldsymbol{v}, \mathrm{e}^{t\mathcal{P}^*}\mathrm{e}^{t\mathcal{P}}\boldsymbol{v} \right\rangle = \left\langle \boldsymbol{v}, \mathrm{e}^{t(\mathcal{P}+\mathcal{P}^*)}\boldsymbol{v} \right\rangle \\ &\le \|\boldsymbol{v}\|^2 \left\|\mathrm{e}^{t(\mathcal{P}+\mathcal{P}^*)}\right\| = \rho\left(\mathrm{e}^{t(\mathcal{P}+\mathcal{P}^*)}\right) = \max\left\{\mathrm{e}^{2t\operatorname{Re}\lambda} : \lambda \in \sigma(\mathcal{P})\right\} \le \mathrm{e}^{2\beta T}. \end{aligned}$$

This completes the proof. □

Von Neumann theory II: Fourier analysis. We restrict attention to linear PDEs with *constant coefficients* and to the *Cauchy problem*: the initial value is given on all of $\mathbb{R}$ (all this can be trivially generalized to several space dimensions), with no boundary conditions.
FD schemes. Let

$$\sum_{k=-r}^{s} \gamma_k u_{m+k}^{n+1} = \sum_{k=-r}^{s} \delta_k u_{m+k}^{n}, \tag{3.11}$$

and set, as before,

$$H(z;\mu) = \frac{\sum_{k=-r}^{s} \delta_k z^k}{\sum_{k=-r}^{s} \gamma_k z^k}.$$

Recall that the *Fourier transform* of $\{v_m\}_{m\in\mathbb{Z}}$ is $\hat{v}(\theta) = \sum_{m=-\infty}^{\infty} v_m \mathrm{e}^{-\mathrm{i}m\theta}$ and that

$$\left(\sum_{-\infty}^{\infty} |v_m|^2\right)^{\frac{1}{2}} = \|\boldsymbol{v}\| = \||\hat{v}|\| = \left[\frac{1}{2\pi}\int_{-\pi}^{\pi} |\hat{v}(\theta)|^2 \,\mathrm{d}\theta\right]^{\frac{1}{2}}.$$

In other words the Fourier transform is an $\ell_2 \to L_2$ *isomorphic isometry*.

Multiplying (3.11) by $\mathrm{e}^{-\mathrm{i}m\theta}$ and summing up for all $m \in \mathbb{Z}_+$ we obtain $\hat{u}^{n+1} = H(\mathrm{e}^{\mathrm{i}\theta};\mu)\hat{u}^n$, hence $\hat{u}^n = \big(H(\mathrm{e}^{\mathrm{i}\theta};\mu)\big)^n \hat{u}^0$. Thus, $\|\mathcal{A}_{\Delta x}^n\| = \||\hat{u}^n|\|$ means that $|H(\mathrm{e}^{\mathrm{i}\theta};\mu)| \le 1$ for all $|\theta| \le \pi \Rightarrow$ stability.

As a matter of fact, $|H(\mathrm{e}^{\mathrm{i}\theta};\mu)| \le 1$ for all $|\theta| \le \pi \Leftrightarrow$ stability. To prove in the $\Leftarrow$ direction, take a function u s.t.

$$\hat{u}(\theta) = \begin{cases} 1, & \theta \in (\alpha,\beta), \\ 0, & \theta \notin (\alpha,\beta), \end{cases}, \quad \text{i.e.} \quad u_m = \frac{1}{2\pi}\int_{\alpha}^{\beta} \mathrm{e}^{\mathrm{i}m\theta}\mathrm{d}\theta = \begin{cases} (\beta-\alpha)/(2\pi), & m = 0, \\ (\mathrm{e}^{\mathrm{i}m\beta} - \mathrm{e}^{\mathrm{i}m\alpha})/(2\pi m), & m \ne 0, \end{cases}$$

where $|H(\mathrm{e}^{\mathrm{i}\theta},\mu)| \ge 1+\varepsilon$ for $\alpha \le \theta \le \beta$ and $\varepsilon > 0$.

Back to the last example. Now $H(\mathrm{e}^{\mathrm{i}\theta};\mu) = 1 - \mu + \mu\mathrm{e}^{\mathrm{i}\theta}$ and $|H(\mathrm{e}^{\mathrm{i}\theta};\mu)| \le 1$ iff $0 \le \mu \le 1$, as required.

Semidiscretizations. Let, as before,

$$u_m' = \frac{1}{(\Delta x)^L}\sum_{k=-r}^{s} \alpha_k u_{m+k}, \qquad h(z) := \sum_{k=-r}^{s} \alpha_k z^k.$$

Similar analysis establishes that stability is equivalent to $\operatorname{Re} h(\mathrm{e}^{\mathrm{i}\theta}) \le 0$ for all $|\theta| \le \pi$.

Example: Suppose that $u_t = u_x$ is solved with $u_m' = \frac{1}{(\Delta x)}\left(-\frac{3}{2}u_m + 2u_{m+1} - \frac{1}{2}u_{m+2}\right)$. Hence $h(\mathrm{e}^{\mathrm{i}\theta}) = -\frac{1}{2}(3 - 4\mathrm{e}^{\mathrm{i}\theta} + \mathrm{e}^{2\mathrm{i}\theta})$, thus $\operatorname{Re} h(\mathrm{e}^{\mathrm{i}\theta}) = -(1-\cos\theta)^2 \le 0$ and stability follows.

A Toeplitz matrix interpretation. Equation (3.11) can be written in a matrix form as $B\boldsymbol{u}^{n+1} = A\boldsymbol{u}^n$ and we wish to bound the norm of $B^{-1}A$. Both A and B are *bi-infinite* Toeplitz matrices. A spectrum of an operator C is the set of all $\lambda \in \mathbb{C}$ such that $C - \lambda I$ has no inverse – when the number of dimensions is infinite this encompasses both eigenvalues and more exotic creatures. In fact, bi-infinite Toeplitz matrices have no eigenvalues, just a *continuous spectrum*.[1] A general Toeplitz matrix F reads $f_{k,l} = \varphi_{k-l}$, $k, l \in \mathbb{Z}$, and we call the Laurent series $f(z) = \sum_{k=-\infty}^{\infty} \varphi_k z^k$ the *symbol* of F.

Theorem 24 *$\sigma(F) = \big\{f(\mathrm{e}^{\mathrm{i}\theta}) : |\theta| \le \pi\big\}$ and $\|F\|_2 = \max\{|f(\mathrm{e}^{\mathrm{i}\theta})| : |\theta| \le \pi\}$.*

[1] Have a good look at a functional analysis book if you wish to understand finer points of operatorial spectra. This, though, will not be required in the sequel…

Moreover, if A and B are bi-infinite Toeplitz then $\sigma(B^{-1}A)$ is the mapping of $|z| = 1$ under the *quotient* of their symbols. For the difference scheme (3.11) we obtain $\sigma(B^{-1}A) = \{H(\mathrm{e}^{\mathrm{i}\theta};\mu) \,:\, |\theta| \le \pi\}$, hence $|H| \le 1$ iff stability. Similar argument extends to SD schemes.

Suppose that F_d is a $d\times d$ Toeplitz matrix and $F = F_\infty$. In general it is not true that $\lim_{d\to\infty}\sigma(F_d) = \sigma(F)$ but (according to a theorem by Szegő) this is the case when the F_ds are all normal. This connects the present material to eigenvalue analysis.

Influence of boundary conditions. Presence of boundaries means that we need *boundary schemes*, some of which can be artificial (when the number of required values exceeds the number of boundary conditions). Although the Fourier condition is necessary, there are in general extra conditions to ensure sufficiency. Zero boundary conditions require the *Strang condition*, namely that, given $H = P/Q$ ($h = p/q$), the Laurent polynomial Q (or q) has precisely r zeros inside and s zeros outside the complex unit circle. This condition is a consequence of a *Wiener–Hopf factorization* of the underlying Toeplitz matrices.

Stability analysis in the presence of boundaries is due to Godunov & Riabienkǐ, Osher and in particular Gustaffson, Kreiss and Sundström (the *GKS theory*). It is far too complicated for elementary exposition. Fortunately, a more recent theory of Trefethen simplifies matters. Suppose that we are solving $u_t = u_x$ by the scheme (3.11) which is *conservative*: $|H(\mathrm{e}^{\mathrm{i}\theta};\mu)| \equiv 1$. We seek a solution of the form $u_m^n = \mathrm{e}^{\mathrm{i}(\xi m\Delta x + \omega(\xi) n\Delta t)}$. Here $\omega(\xi)$ is the *phase velocity* and $c(\xi) := \frac{\mathrm{d}}{\mathrm{d}\xi}\omega(\xi)$ is the *group velocity*. Both are defined for $|\xi| \le \frac{\pi}{\Delta x}$, wave numbers supported by the grid. Substituting the stipulated values of u_m^n into (3.11) yields

$$\mathrm{e}^{\mathrm{i}\omega(\xi)\mu\Delta x} = H(\mathrm{e}^{\mathrm{i}\xi\Delta x};\mu), \quad \text{hence} \quad c(\xi) = \frac{\mathrm{e}^{\mathrm{i}\xi\Delta x}}{\mu}\,\frac{\frac{\mathrm{d}}{\mathrm{d}z}H(\mathrm{e}^{\mathrm{i}\xi\Delta x};\mu)}{H(\mathrm{e}^{\mathrm{i}\xi\Delta x},\mu)}.$$

Example Crank–Nicolson: $\frac{1}{4}\mu u_{m-1}^{n+1} + u_m^{n+1} - \frac{1}{4}\mu u_{m+1}^{n+1} = -\frac{1}{4}\mu u_{m-1}^n + u_m^n + \frac{1}{4}\mu u_{m+1}^n$. Thus,

$$H(\mathrm{e}^{\mathrm{i}\theta};\mu) = \frac{1 + \mathrm{i}\frac{1}{2}\mu\sin\theta}{1 - \mathrm{i}\frac{1}{2}\mu\sin\theta} \quad \Rightarrow \quad c(\xi) = \frac{\cos(\xi\Delta x)}{1 + \mu^2\frac{1}{4}\sin^2(\xi\Delta x)}.$$

Note that $c(\xi)$ changes sign in $|\xi| \le \frac{\pi}{\Delta x}$. This means that some wave numbers are transported by the numerical 'flow' in *the wrong direction*! In fact, c being a derivative of a periodic function, it is easy to prove that, for every conservative (3.11), either $c(\xi)$ changes sign or it can't be uniformly bounded for all $\mu > 0$. Thus, some wave numbers are transported either in the wrong direction or with infinite speed or both!

Trefethen's theory. Suppose that, for some $|\xi| \le \frac{\pi}{\Delta x}$ we have $c(\xi) < 0$ and that this is also the case for the group velocity induced by the boundary scheme. Then the method is unstable. However, if the 'internal' scheme is stable *and* boundary schemes bar all ξ such that $c(\xi) < 0$ then (3.11) is stable.

Example (with CN): (i) $u_0^{n+1} = u_2^{n-1} + (\mu-1)(u_2^n - u_0^n)$. Both CN and the boundary scheme admit $u_m^n = (-1)^m = \mathrm{e}^{\mathrm{i}\pi m} \Rightarrow$ instability.
(ii) $u_0^{n+1} = u_1^n$. For ξ to be admitted by the boundary scheme we need $\mathrm{e}^{\mathrm{i}\omega\Delta t} = \mathrm{e}^{\mathrm{i}\xi\Delta x}$, hence $c \equiv \mu^{-1} > 1$ (where c is the 'boundary group velocity'), whereas the method's group velocity lives in $[-1,1]$. Hence stability.

Multistep methods. Fourier analysis can be easily extended to multistep schemes, e.g.
Leapfrog: $u_m^{n+1} = \mu(u_{m+1}^n - u_{m-1}^n) + u_m^{n-1}$,
Angled derivative: $u_m^{n+1} = -(1-2\mu)(u_{m+1}^n - u_m^n) + u_{m+1}^{n-1}$,
both for $u_t = u_x$. Thus, angled derivative gives

$$T(w,z;\mu) = w^2 + (1-2\mu)(z-1)w - z \qquad \Rightarrow \qquad T(z^\mu, z;\mu) = \mathcal{O}\big(|z-1|^3\big)\,,$$

hence the order is 2. Moreover, its Fourier transformation yields

$$\hat{u}^{n+1} + (1-2\mu)(\mathrm{e}^{\mathrm{i}\theta} - 1)\hat{u}^n - \mathrm{e}^{\mathrm{i}\theta}\hat{u}^{n-1} = 0.$$

This is a two-step recurrence relation and all its solutions are uniformly bounded iff the quadratic $T(\cdot, \mathrm{e}^{\mathrm{i}\theta}; \mu)$ obeys the root condition. The latter is true for all $\theta \in [-\pi, \pi]$ iff $0 \le \mu \le 1$.
Likewise, leapfrog is order-2 and stable for all $|\mu| \le 1$. The interest in negative values of μ is motivated by the extension of our analysis to the PDE system $\boldsymbol{u}_t = G\boldsymbol{u}_x$, where $\sigma(G) = \{\lambda_1, \ldots, \lambda_d\}$ is real (this ensures hyperbolicity and well-posedness). In this case stability requires that $\lambda_l \frac{\Delta t}{\Delta x}$ are in the 'stability set' for all $l = 1, \ldots, d$.

Periodic boundary conditions. These produce *circulants*, i.e. Toeplitz matrices of the form

$$F = \begin{bmatrix} f_0 & f_1 & f_2 & \cdots & f_{M-1} \\ f_{M-1} & f_0 & f_1 & \cdots & f_{M-2} \\ f_{M-2} & f_{M-1} & f_0 & \cdots & f_{M-3} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ f_1 & \cdots & f_{M-2} & f_{M-1} & f_0 \end{bmatrix}.$$

It is trivial to verify that $\lambda_l = \sum_{k=0}^{M-1} f_k \mathrm{e}^{\frac{2\pi i k l}{M}}$ and $\left\{ v_{l,k} = \mathrm{e}^{\frac{2\pi i k l}{M}} : k = 0, 1, \ldots, M-1 \right\}$ are an eigenvalue/eigenvector pair for $l = 0, 1, \ldots, M-1$. Hence normalcy, consequently stability requires that the symbol is bounded by 1 on roots of unity.

Group velocity and wave propagation. Figure 3.1 displays the evolution of a step function, as $u_t = u_x$ (with periodic boundary conditions) is discretized by two different methods with $\mu = \frac{3}{4}$. In the case of Crank–Nicolson we have already seen that $c(\xi)$ changes sign, hence some wave numbers (and all wave numbers are present in a discontinuity) are propagated backwards. It is easy to demonstrate, however, that the group velocity of the *box method*

$$(1+\mu)u_m^{n+1} + (1-\mu)u_{m+1}^{n+1} = (1-\mu)u_m^n + (1+\mu)u_{m+1}^n,$$

namely

$$c(\xi) = \frac{2}{(1+\mu^2) - (1-\mu^2)\cos\xi\Delta x},$$

always exceeds one for $\xi \neq 0$ and $\mu \le 1$. Hence wave numbers always propagate in the right direction but they do it too fast!

The energy method. (when all else fails…) Consists of direct estimates in the *energy* (a.k.a. *Euclidean*, a.k.a. *least squares*, a.k.a. ℓ_2) norm. For example, approximate $u_t = a(x)u_x$, with zero b.c., by the SD

$$u'_m = \frac{a_m}{2\Delta x}(u_{m+1} - u_{m-1}) \tag{3.12}$$

and let

$$\|\boldsymbol{u}\| = \left[(\Delta x) \sum_{m=1}^{M-1} u_m^2 \right]^{\frac{1}{2}}, \qquad \Delta x = \frac{1}{M}.$$

Hence, using the Cauchy–Schwarz inequality

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{u}\|^2 &= 2(\Delta x)\sum_{m=1}^{M-1} u_m u'_m = \sum_{m=1}^{M-1} a_m u_m (u_{m+1} - u_{m-1}) \\ &= \sum_{m=1}^{M-1} (a_m - a_{m+1}) u_m u_{m+1} \le \alpha(\Delta x)\sum_{m=1}^{M-1} |u_m u_{m+1}| \le \alpha\|\boldsymbol{u}\|^2, \end{aligned}$$

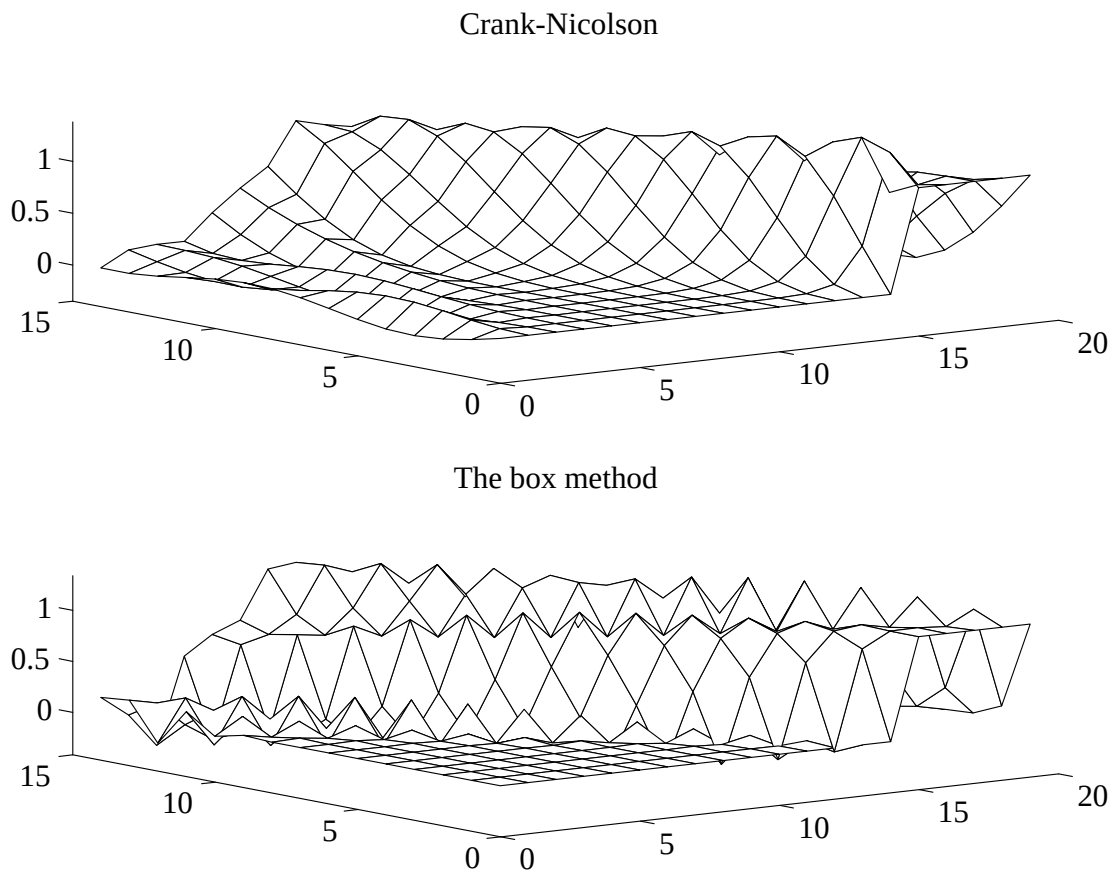


Figure 3.1 Propagation of a discontinuity by different numerical schemes.

provided that $|a(x) - a(y)| \leq \alpha|x - y|$ for some $\alpha > 0$. Consequently $\|\boldsymbol{u}(t)\| \leq \mathrm{e}^{\alpha t}\|\boldsymbol{u}(0)\|$ and we have uniform boundedness on compact intervals, hence stability of (3.12).

3.5 A nonlinear example: Hyperbolic conservation laws

We consider

$$\frac{\partial}{\partial t}u + \frac{\partial}{\partial x}f(u) = 0, \tag{3.13}$$

where f is a given function. It is accompanied by initial (and possibly boundary) conditions.

Applications. Equations (3.13) include as special case the Burgers' equation (i.e. $f(u) = \frac{1}{2}u^2$). Their multivariate generalizations include the equations of compressible invicid flow in Eulerian formulation and equations of gas dynamics.

Features of the solution. It is easy to see that the solution is *constant along characteristics* – given that $u(x_0, 0) = u_0$, say, the solution stays u_0 for all $t \geq 0$ at $x = x_0 + f'(u_0)t$. However, the slope

of each characteristic line is, in principle, different, and we can anticipate two problems:

1. **Shocks:** When two characteristics, 'carrying' distinct values, clash, a discontinuity occurs. It is a shock, i.e. the flow is always *into* the discontinuity and no information ever leaves it. Let $\Gamma(t)$ be a parametric representation of a shock. It is possible to show that it obeys the *Rankine–Hugeniot conditions*

$$\frac{\mathrm{d}\Gamma(t)}{\mathrm{d}t} = \frac{[f(u)]}{[u]} \qquad \text{where } [w] \text{ is the jump of } w \text{ across the shock.}$$

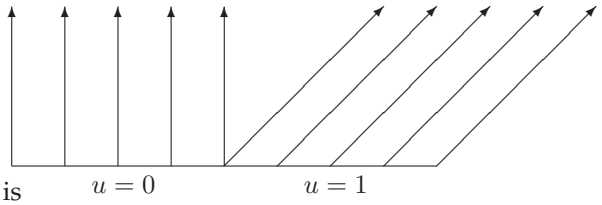
2. **Rarefactions:** When characteristics depart from each other, they create a 'void', a *rarefaction fan*. The solution can be patched across the void in a multitude of ways, but only one has physical significance. We envisage (3.13) as the solution of $u_t + (f(u))_x = \nu u_{xx}$, $\nu > 0$, as $\nu \to 0$ (νu_{xx} represents viscosity) and this implies the *entropy condition*

$$\tfrac{1}{2}\frac{\partial}{\partial t}u^2 + \frac{\partial}{\partial x}F(u) \le 0 \qquad \text{where} \qquad F(u) := \int_0^u y f'(y)\,\mathrm{d}y.$$

Provided that the RH and entropy conditions are obeyed, the solution of (3.13) exists, is unique and $\exists c > 0$ s.t. $\|u\| \le c\|\phi\|$, where $\phi(x) = u(x,0)$.
Note that **(a)** (3.13) is not a 'true' conservation law – the energy is lost at shocks; **(b)** Shocks are a feature of the equation and they are likely to occur even with smooth initial conditions. They have, incidentally, a physical interpretation.

Example: Burgers' equation $u_t + uu_x = 0$. It is easy to see that

$$u(x,0) = \begin{cases} 1 & : x \ge 0 \\ 0 & : x < 0 \end{cases} \quad \Rightarrow$$



The right way of filling the triangular void is

$$u(x,t) = \begin{cases} 1 & : 0 \le t \le x, \\ \frac{x}{t} & : 0 \le x \le t, \\ 0 & : x < 0,\ t \ge 0 \end{cases} .$$

Indeed, $f(u) = \frac{1}{2}u^2 \Rightarrow F(u) = \frac{1}{3}u^3 \Rightarrow$ the entropy condition is $uu_t + u^2u_x \le 0$. There is nothing to check when $0 \le t \le x$ or when $x < 0$, $t \ge 0$ and it is trivial to verify that $uu_t + u^2u_x = 0$ in the triangle $0 \le x \le t$.

Godunov's method. The main idea is to approximate the initial condition by a step function. Hence, we keep (3.13) intact, but replace $u(x,0) = \phi(x)$ with $u(x,0) = \tilde{\phi}(x)$, where $\tilde{\phi}(x) \equiv \phi\left(\left(m + \frac{1}{2}\right)\Delta x\right)$ for $m\Delta x \le x < (m+1)\Delta x$, $m \in \mathbb{Z}$. The latter is a *Riemann problem*, which can be solved explicitly. Specifically, let $u(x,0) = a$ for all $a \in [x_0, x_1)$. Then $u_t + f'(a)u_x = 0$ there, hence $u(x,t) = \tilde{\phi}(x - f'(a)t) = a$ for $x_0 \le x - f'(a)t < x_1$. We advance the solution in this fashion as long as no more than two characteristics clash (i.e., as long as we can resolve distinct shocks) – the length of a step depends on the steepness of the slopes – but also so as not to 'open up' rarefaction fans too much. Thus, having reached a new time level t_1, say, we have numerical values that *are not equidistributed*. We replace them by a step function (as long as $\Delta t < (\Delta x) \times \max|f'(\tilde{\phi})|$, i.e. the Courant number ≤ 1, we'll have enough data in each interval) and continue this procedure.

Improvement I – van Leer's method. Instead of a step function, it is possible to approximate ϕ by a piecewise linear function and derive the exact solution of (3.13) with this initial condition. This yields a second-order method (compared to Godunov's, which is first-order).

Improvement II – Glimm's method. Instead of letting $\tilde{\phi}$ be the value at the centre, we choose it at random (according to uniform distribution) in the interval. This is repeated in every step. Although this has only limited numerical merit, convergence of this procedure constitutes the proof (originally due to James Glimm) that (3.13), subject to RH and entropy conditions, possesses a (weak) solution. This is an example of an important use of 'pseudo-numerical' methods as a device to prove existence of solutions of nonlinear PDEs.

The Engquist–Osher method. Suppose that f is a strictly convex function. Thus, there exists a unique $\bar{u} \in \mathbb{R}$ s.t. $f'(\bar{u}) = 0$ (*sonic point*, *stagnation point*). We assume that the initial condition is in $L_2(\mathbb{R})$ and define the *EO switches*

$$f_-(y) := f(\min\{y, \bar{u}\}), \qquad f_+(y) := f(\max\{y, \bar{u}\}), \qquad y \in \mathbb{R}.$$

The EO method is the semidiscretization

$$u'_m = -\frac{1}{\Delta x}\left(\Delta_+ f_-(u_m) + \Delta_- f_+(u_m)\right), \qquad m \in \mathbb{Z}.$$

Stability. We use the energy method. Set

$$B_1 := -\sum_{m=-\infty}^{\infty} u_m \Delta_+ f_-(u_m), \qquad B_2 := -\sum_{m=-\infty}^{\infty} u_m \Delta_- f_+(u_m).$$

Thus,

$$\|\boldsymbol{u}\|^2 = (\Delta x)\sum_{m=-\infty}^{\infty} u_m^2 \qquad \Rightarrow \qquad \tfrac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{u}\|^2 = B_1 + B_2. \tag{3.14}$$

Integrating by parts,

$$\sum_{m=-\infty}^{\infty}\int_{u_m}^{u_{m+1}} y f'_-(y)\,\mathrm{d}y = \sum_{m=-\infty}^{\infty}\left\{u_{m+1}f_-(u_{m+1}) - u_m f_-(u_m) - \int_{u_m}^{u_{m+1}} f_-(y)\,\mathrm{d}y\right\}.$$

Telescoping series and exploiting $\lim_{m\to\infty} u_{\pm m} = 0$ (a consequence of the initial condition being L_2), we obtain

$$\sum_{m=-\infty}^{\infty}\int_{u_m}^{u_{m+1}} y f'_-(y)\,\mathrm{d}y = \lim_{m,k\to\infty}\left\{u_m f_-(u_m) - u_{-k}f_-(u_{-k}) - \int_{u_{-k}}^{u_m} f_-(y)\,\mathrm{d}y\right\} = 0.$$

Therefore,

$$B_1 = -\sum_{m=-\infty}^{\infty} u_m \int_{u_m}^{u_{m+1}} f'_-(y)\,\mathrm{d}y = \sum_{m=-\infty}^{\infty}\int_{u_m}^{u_{m+1}} (y-u_m) f'_-(y)\,\mathrm{d}y := \sum_{m=-\infty}^{\infty} I_m.$$

But

$$\left.\begin{array}{lllll} u_{m+1} \ge u_m & \Rightarrow & (y-u_m) \ge 0,\ f'_-(y) \le 0,\ u_m \le y \le u_{m+1} & \Rightarrow & I_m \le 0 \\ u_{m+1} \le u_m & \Rightarrow & (y-u_m) \le 0,\ f'_-(y) \le 0,\ u_{m+1} \le y \le u_m & \Rightarrow & I_m \le 0 \end{array}\right\} \Rightarrow\ B_1 \le 0.$$

Similarly, it follows that $B_2 \le 0$ and (3.14) implies that $\|\boldsymbol{u}\|$ is monotonically nonincreasing – hence stability.

3.6 Additional PDE problems and themes

Finite difference are the simplest means to solve PDEs but numerous other approaches are available. Thus, we can single out

- **Finite elements:** The theme of the next section: the equation is formulated either as a *variational problem* or in a *weak form.* In the first case we seek to minimize a functional $J(u)$, say, amongst all u in an appropriate function space $\mathcal{H}$, in the latter case we replace $\mathcal{L}(u) = f$ by the more general $\langle \mathcal{L}(u) - f, v \rangle = 0 \; \forall v \in \mathcal{H}$, where $\langle \cdot\,, \cdot \rangle$ is an inner product. The problem is solved by replacing $\mathcal{H}$ with an appropriate *finite-dimensional subspace,* which is spanned by functions with small local support ('finite elements').

  Finite elements lend themselves very well to complicated geometries and exotic boundary conditions.

- **Spectral methods:** Remember the method of *separation of variables* from basic PDE courses? A spectral method is, basically, an attempt to turn separation of variables into a numerical method. This leads to very effective methods for periodic boundary conditions in very simple geometries: typically we need to invert full $M \times M$ matrices, but the error decays at 'spectral' speed, as $\mathrm{e}^{-\alpha M}$ for some $\alpha > 0$. (For finite difference and finite elements it typically decays as $1/M^q$ for some $q > 0$.) The method is far less potent for exotic geometries or nonperiodic boundary conditions.

- **Boundary-integral methods:** Often, a PDE in a domain $\mathcal{D}$ can be replaced by an integral equation defined on $\partial\mathcal{D}$. Such an integral equation can be solved, e.g. by collocation methods. Although this is quite complicated and typically requires the inversion of linear systems with dense matrices, the redeeming feature of this approach is that the *dimension* of $\partial\mathcal{D}$ is one less than the dimension of $\mathcal{D}$.

- **Fast multipole methods:** The solution of a Poisson equation can be approximated by assuming that the potential is generated by a finite number of particles. Subsequently, far-field and near-field potentials are approximated differently, in a manner which lends itself to very fast nested computation.

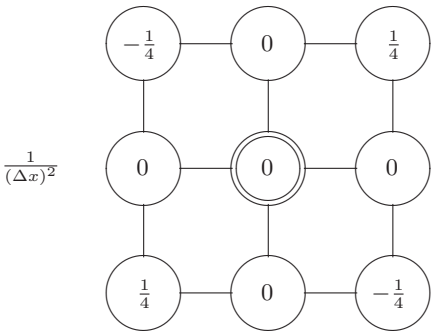
- **Multiresolution methods:** The solution of a PDE is approximated by functions (in particular, by wavelets) that automatically 'zoom' on parts of the solution where the action takes place and provide higher resolution there.

The list goes on and on: boundary-element methods, pseudospectral methods, finite-volume methods, particle methods, vorticity methods, gas-lattice methods, meshless methods. . . .

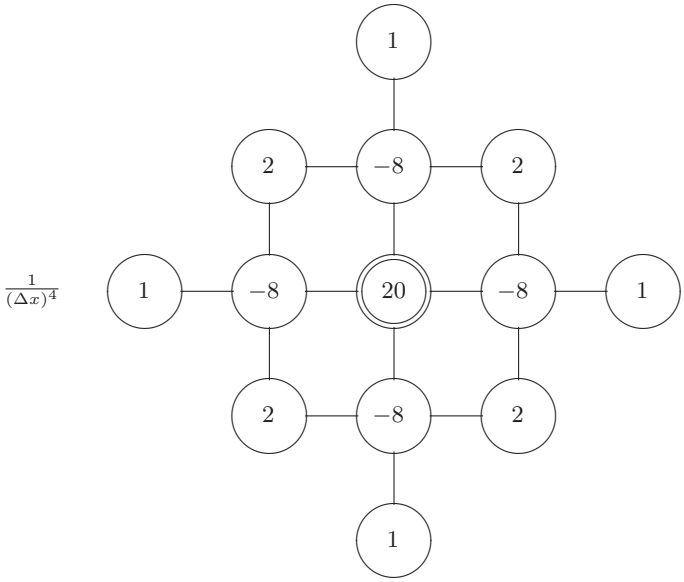
The choice of a method is just the beginning of a long procedure. Typically, it is a worthwhile idea to use *adaptive grids* or *domain decomposition.* The whole procedure also hinges upon our algorithm to solve the underlying algebraic system (cf. Section 5) and it typically takes place within parallel computer architecture.

Exercises

3.1 Determine the order of magnitude of the error of the finite-difference approximation to $\partial^2 u/\partial x \partial y$ which is given by the computational stencil

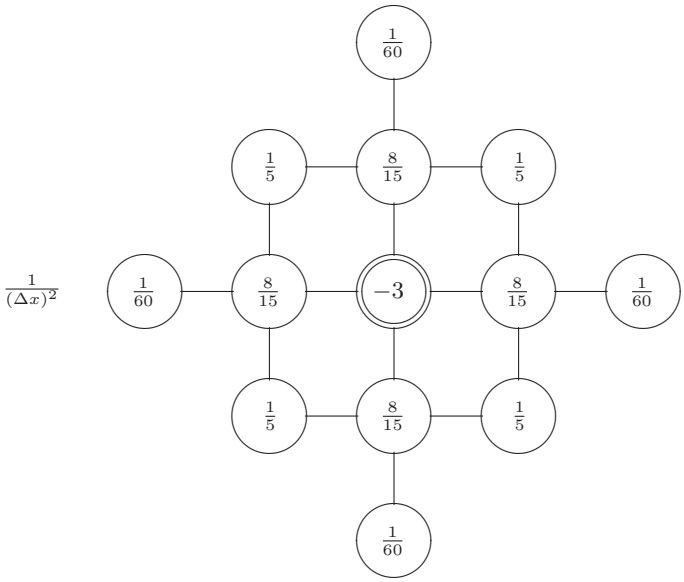


3.2 The biharmonic operator $\mathcal{L} = \nabla^4$ is approximated by the thirteen-point finite-difference stencil



1. Write down explicitly the underlying linear system for the values $u_{k,l}$.
2. What is the order of the approximation?
3. Discuss briefly the nature of boundary conditions that we can expect in the present case. Suggest amendments to the stencil at grid points near the boundary, to cater for these conditions.

3.3 A finite-difference operator $\mathcal{L}_\Delta$ is defined by the stencil

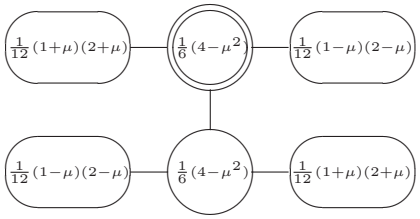


1. Prove that
$$\mathcal{L}_\Delta = \nabla^2 \left(I + \frac{1}{10}(\Delta x)^2\nabla^2 + \frac{1}{180}(\Delta x)^4\nabla^4 \right) + \mathcal{O}\big((\Delta x)^6\big)\,.$$
2. Find a finite-difference operator $\mathcal{M}_\Delta$ such that the *Mehrstellenverfahren*
$$\mathcal{L}_\Delta U_{k,l} + \mathcal{M}_\Delta F_{k,l} = 0$$
is an $\mathcal{O}\big((\Delta x)^6\big)$ approximation to $\nabla^2 u + f = 0$.

3.4 The diffusion equation $u_t = u_{xx}$ is being approximated by the FD method
$$\begin{aligned}&\left(\frac{1}{12} - \frac{1}{2}\mu\right) u_{m-1}^{n+1} + \left(\frac{5}{6} + \mu\right) u_m^{n+1} + \left(\frac{1}{12} - \frac{1}{2}\mu\right) u_{m+1}^{n+1} \\ &= \left(\frac{1}{12} + \frac{1}{2}\mu\right) u_{m-1}^{n} + \left(\frac{5}{6} - \mu\right) u_m^{n} + \left(\frac{1}{12} + \frac{1}{2}\mu\right) u_{m+1}^{n},\end{aligned}$$
where μ is the Courant number (the *Crandall method*, a.k.a. the *Morris & Gourlay method*).

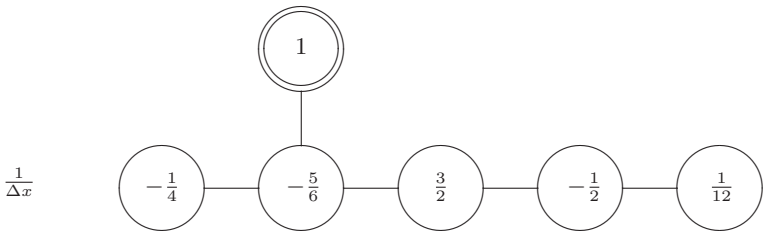
1. Prove that the method is of order 5 (in Δx).
2. Prove that the method is stable for the Cauchy problem by two alternative approaches: (i) Eigenvalue analysis; (ii) Fourier analysis.

3.5 Determine order and stability of the finite-difference implicit FD scheme



for the solution of the advection equation $u_t = u_x$.

3.6 Ditto for the SD scheme



3.7 The semi-discretization

$$\begin{aligned} u'_m &= \frac{1}{\Delta x}\left(-\frac{3}{2}u_m + 2u_{m+1} - \frac{1}{2}u_{m+2}\right), \quad m = 1, 2, \ldots, M-2,\ \Delta x = \frac{1}{M}, \\ u'_{M-1} &= \frac{1}{\Delta x}\left(-\frac{3}{2}u_{M-1} + 2u_M - \frac{1}{2}u_1\right), \\ u'_M &= \frac{1}{\Delta x}\left(-\frac{3}{2}u_M + 2u_1 - \frac{1}{2}u_2\right) \end{aligned}$$

approximates the solution of the advection equation $u_t = u_x$, given with initial conditions $u(x,0) = \psi(x)$, $0 \le x \le 1$, and *periodic* boundary conditions $u(0,t) = u(1,t)$, $t \ge 0$. Prove that the method is stable. *[Hint: You might formulate the problem in a matrix form, $\boldsymbol{u}' = \frac{1}{\Delta x}A\boldsymbol{u}$, say, and show that all eigenvectors of A are of the form $[1, \omega_l, \ldots, \omega_l^{M-1}]^\top$ for some complex numbers $\{\omega_l : l = 1, 2, \ldots, M\}$.]*

3.8 The parabolic equation

$$\frac{\partial}{\partial t}u = \frac{\partial^2}{\partial x^2}u + \kappa u$$

is given for $0 \le x \le 1$, $t \ge 0$, together with an initial condition at $t = 0$ and zero boundary conditions at $x = 0$ and $x = 1$.

1. Prove (by separation of variables or otherwise) that the exact solution of the equation tends to zero as $t \to \infty$ for every initial condition if and only if $\kappa < \pi^2$.
2. The equation is semidiscretized into

$$u'_m = \frac{1}{(\Delta x)^2}(u_{m-1} - 2u_m + u_{m+1}) + \kappa u_m, \qquad m = 1, 2, \ldots, M,$$

where $(M+1)\Delta x = 1$. Find the necessary and sufficient condition on κ that ensures that $\lim_{t\to\infty} u_m(t) = 0$, $m = 1, 2, \ldots, M$, for all possible initial conditions.

3.9 An FD approximation

$$\sum_{k=-r}^{s} \alpha_k(\mu) u_{m+k}^{n+1} = \sum_{k=-r}^{s} \beta_k(\mu) u_{m+k}^{n}$$

to $u_t = u_x$ is given. We assume the analyticity (as a function of μ) of all the coefficients as $\mu \to 0$. Set

$$H(z;\mu) := \frac{\sum_{k=-r}^{s} \beta_k(\mu) z^k}{\sum_{k=-r}^{s} \alpha_k(\mu) z^k}.$$

We further assume that $H(z,0) \equiv 1$ and define

$$h(z) := \frac{\partial}{\partial \mu} H(z;\mu)\bigg|_{\mu=0}.$$

1. Show that h is a rational function,

$$h(z) = \frac{\sum_{k=-r}^{s} b_k z^k}{\sum_{k=-r}^{s} a_k z^k},$$

say.

2. Given that the FD method is of order p, prove that the SD method

$$\sum_{k=-r}^{s} a_k u'_{m+k}(t) = \frac{1}{\Delta x} \sum_{k=-r}^{s} b_k u_{m+k}(t)$$

is of order $p^* \geq p$.

3. Prove that, subject to stability of the FD scheme as $\mu \to 0$, the SD method is stable.

3.10 The differential equation

$$\begin{aligned} &u_t = a(x)u_x, \qquad -\infty < x < \infty, \quad t > 0, \\ &u(x,0) = \phi(x), \qquad -\infty < x < \infty \end{aligned}$$

(*the Cauchy problem for the advection equation with a variable coefficient*) is solved by the SD method

$$u'_m = \frac{a_m}{2\Delta x}(u_{m+1} - u_{m-1}), \qquad m \in \mathbb{Z},$$

where $a_m = a(m\Delta x),\ m \in \mathbb{Z}$. We assume that $0 < C_1 < C_2$ exist such that

$$C_1 \leq a(x) \leq C_2, \qquad x \in \mathbb{R}.$$

1. Prove that the scheme for $u_t = a(x)u_x$ is a second-order method.
2. Set $v(x,t) := u(x,t)/\sqrt{a(x)}$ and $\boldsymbol{v} := A^{-\frac{1}{2}}\boldsymbol{u}$, where A is the bi-infinite *diagonal* matrix whose (l,l) component is a_l and $\boldsymbol{u}$ is the vector of the u_m's.
   Find a differential equation that is satisfied by v and show that the vector $\boldsymbol{v}$ is a second-order finite-difference SD approximation to the solution of that equation.
3. By considering the difference equation that is obeyed by $\tilde{V}$, prove that the method (for $\tilde{U}$) is stable. *Hint:*
   (a) Prove that the matrix B in the SD system $\boldsymbol{v}' = B\boldsymbol{v}$ is self-adjoint.
   (b) Apply the Lax equivalence theorem *twice*, once in each direction.

3.11 Find a stable second-order semidiscretization of the Cauchy problem for the diffusion equation with a variable coefficient,

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(a(x)\frac{\partial u}{\partial x}\right).$$

The coefficient a is assumed to be uniformly bounded and uniformly positive for all $x \in \mathbb{R}$.

3.12 The advection equation $u_t = u_x$ is semidiscretized by

$$u'_m = \tfrac{1}{\Delta x} \sum_{k=-r}^{s} \alpha_k u_{m+k},$$

where

$$\begin{aligned} \alpha_k &= \frac{(-1)^{k-1}}{k} \frac{r!s!}{(r+k)!(s-k)!}, \quad k = -r, \ldots, s,\ k \neq 0, \\ \alpha_0 &= \sum_{\substack{j=-r \\ j\neq 0}}^{s} \frac{(-1)^j}{j} \frac{r!s!}{(r+j)!(s-j)!} \end{aligned}$$

Prove that the method is of order $r+s$ and that any other method (with the same r and s) is of smaller order.

3.13 Let $r = s$ in the last question and evaluate the limit of α_k as $r \to \infty$. You may use (without proof) the *Stirling formula*

$$m! \approx \sqrt{2\pi m}\left(\frac{m}{\mathrm{e}}\right)^m, \qquad m \gg 1.$$

Consider the following situation: the equation $u_t = u_x$ is given in $[0,1]$ with *periodic* boundary conditions. Semidiscretize

$$u'_m = \tfrac{1}{\Delta x}\sum_{k\in\mathbb{Z}} \alpha_k u_{m+k},$$

where the α_ks are the aforementioned limits and we identify u_{m+k} with its 'wrap-around' grid point in $[0,1]$. Prove that this gives an *infinite-order* method. (This algorithm, due to Fornberg, is an example of a *pseudospectral* method.)

Bibliography

[1] A.R. Mitchell and D.F. Griffiths, *The Finite Difference Method in Partial Differential Equations*, Wiley, London, 1980.

[2] R.D. Richtmyer and K.W. Morton, *Difference Methods for Initial-Value Problems*, Interscience, New York, 1967.

[3] G.D. Smith, *Numerical Solution of Partial Differential Equations: Finite Difference Methods*, 3rd Edition **!!!**, Oxford Univ. Press, Oxford, 1985.

4 Finite elements

4.1 Guiding principles

It is convenient to single out five main ideas behid the finite element method (FEM):

1. Formulate a differential equation as a variational problem, e.g. by seeking a function that minimizes a nonlinear functional;
2. Reduce differentiability requirements in the above functional, using integration by parts;
3. Take appropriate care with boundary conditions while converting a differential to a variational problem and distinguish between *essential* and *natural* conditions;
4. Replace the underlying problem by an approximate one, restricted to a finite-dimensional space;
5. Choose a finite-dimensional space which is spanned by functions with a small support (*finite element functions*).

4.2 Variational formulation

Euler–Lagrange equations. Many differential equations start their life as variational problems: find a function u such that $I(u) = \min_{v\in\mathbb{H}} I(v)$. Here $\mathbb{H}$ is a function space, defined in terms of differentiability and boundary conditions, whereas I is a functional. Examples:

1. $I(u) = \int_a^b F(x,u,u_x)\,\mathrm{d}x \Rightarrow \frac{\partial F}{\partial u} - \frac{\mathrm{d}}{\mathrm{d}x}\frac{\partial F}{\partial u_x} = 0$;

2. $I(u,v) = \int_a^b F(x,u,v,u_x,v_x)\,\mathrm{d}x \Rightarrow \begin{cases} \frac{\partial F}{\partial u} - \frac{\mathrm{d}}{\mathrm{d}x}\frac{\partial F}{\partial u_x} = 0, \\ \\ \frac{\partial F}{\partial v} - \frac{\mathrm{d}}{\mathrm{d}x}\frac{\partial F}{\partial v_x} = 0. \end{cases}$;

3. $I(u) = \int_{\mathcal{D}} F(x,y,u,u_x,u_y)\,\mathrm{d}x\,\mathrm{d}y \Rightarrow \frac{\partial F}{\partial u} - \frac{\partial}{\partial x}\frac{\partial F}{\partial u_x} - \frac{\partial}{\partial y}\frac{\partial F}{\partial u_y} = 0$;

4. $I(u) = \int_a^b F(x,u,u_x,u_{xx})\,\mathrm{d}x \Rightarrow \frac{\partial F}{\partial u} - \frac{\mathrm{d}}{\mathrm{d}x}\frac{\partial F}{\partial u_x} + \frac{\mathrm{d}^2}{\mathrm{d}x^2}\frac{\partial F}{\partial u_{xx}} = 0.$

It is perhaps worthwhile to recall that variational problems are the origin to many – if not most – differential equations of physical significance.

The Ritz method. We *reverse* the reasoning behind the Euler–Lagrange equations, replacing a DE by a variational principle. E.g., $-(pu')' + qu = f$, where $p > 0$, $q \geq 0$ and initial conditions are given at $\{0,1\}$, is replaced by the problem of minimizing $I(u) = \int_0^1 (p(x)u'^2(x) + q(x)u^2(x) - 2f(x)u(x))\,\mathrm{d}x$. We seek the solution of the latter by letting $u(x) \approx \varphi_0(x) + \sum_{m=1}^M a_m\varphi_m(x)$, where $a_0\varphi_0$ caters for the boundary conditions, and minimizing I w.r.t. a_m, $m = 1,\ldots,M$. Let

$$\mathcal{I}(\boldsymbol{a}) = \int_0^1 \left\{ p(x)\left(\sum_{m=0}^M a_m\varphi_m'(x)\right)^2 + q(x)\left(\sum_{m=0}^M a_m\varphi_m(x)\right)^2 - 2f(x)\sum_{m=0}^M a_m\varphi_m(x) \right\} \mathrm{d}x.$$

At the minimum

$$\tfrac{1}{2}\frac{\partial \mathcal{I}(\boldsymbol{a})}{\partial a_k} = \sum_{m=0}^M a_m \int_0^1 \{p\varphi_m'\varphi_k' + q\varphi_m\varphi_k\}\,\mathrm{d}x - \int_0^1 f\varphi_k\,\mathrm{d}x = 0, \quad m = 1,\ldots,M. \tag{4.1}$$

This is a set of M linear equations in $a_1,\ldots,a_m$.

Weak solutions and function spaces. Let $\langle\,\cdot\,,\,\cdot\,\rangle$ be an inner product (e.g. the standard L_2 inner product $\langle f,g\rangle := \int_a^b f(x)g(x)\,\mathrm{d}x$). The *weak solution* of $\mathcal{L}u = f$ in the function space H is the function (to be precise, an equivalence class of functions) u s.t. $\langle \mathcal{L}u - f, v\rangle = 0$ for every $v \in H$. Weak and classical solution coincide when the latter exists, but weak solutions are more general, because, allowing integration by parts, they impose lesser differentiability requirements. The underlying space H is made out of the closure of all functions f and g for which $\langle \mathcal{L}f, g\rangle$ makes sense *after integration by parts*. This helps to highlight why weak solutions are more general than classical ones – they require less differentiability! For example, for $\mathcal{L} = \frac{\mathrm{d}^2}{\mathrm{d}x^2}$ we require the closure of *once-differentiable* functions. Spaces like this are called *Sobolev spaces* and they form an important special case of *Hilbert spaces*.

The Ritz method – again... $\mathcal{L}$ is said to be *self–adjoint* if $\langle \mathcal{L}u, v\rangle = \langle u, \mathcal{L}v\rangle\ \forall u,v,\in \mathbb{H}$, *elliptic* if $\langle \mathcal{L}v, v\rangle > 0\ \forall v \in \mathbb{H}\setminus\{0\}$ and *positive definite* if it is both self-adjoint and elliptic.

Theorem 25 *If $\mathcal{L}$ is positive definite then*

(a) *$\mathcal{L}u = f$ is the Euler–Lagrange equation of $I(v) = \langle \mathcal{L}v, v\rangle - 2\langle f, v\rangle$;*

(b) *The weak solution of $\mathcal{L}u = f$ is the unique minimum of I (and hence the weak solution is unique!).*

Proof Suppose that u is a (local) minimum of I and choose $v \in \mathbb{H}$ s.t. $u + \varepsilon v$ obeys boundary conditions. (Actually, we need only to stipulate that essential b.c.'s are obeyed – read on.) For

example, for Dirichlet b.c. we require that v obeys zero b.c. Note that every 'candidate' for solution can be expressed in this form. For small $|\varepsilon| > 0$

$$\begin{aligned} I(u) \le I(u+\varepsilon v) &= \langle \mathcal{L}(u+\varepsilon v), u+\varepsilon v\rangle - 2\langle f, u+\varepsilon v\rangle && \text{(optimality of } u\text{)} \\ &= \{\mathcal{L}u, u\rangle - 2\langle f, u\rangle\} + \varepsilon\{\langle \mathcal{L}v, u\rangle + \langle \mathcal{L}u, v\rangle - 2\langle f, v\rangle\} + \varepsilon^2\langle \mathcal{L}v, v\rangle && \text{(linearity)} \\ &= I(u) + 2\varepsilon\langle \mathcal{L}u - f, v\rangle + \varepsilon^2\langle \mathcal{L}v, v\rangle. && \text{(self-adjointness and linearity)} \end{aligned}$$

Suppose that $\exists v$ s.t. $\langle \mathcal{L}u - f, v\rangle \neq 0$. Then we can always choose $0 < |\varepsilon| \ll 1$ such that $\varepsilon\langle \mathcal{L}u - f, v\rangle + \varepsilon^2\langle \mathcal{L}v, v\rangle < 0$, hence contradiction. It follows that $\langle \mathcal{L}u - f, v\rangle = 0$ for all suitable v and u is a weak solution.

Suppose that another, distinct, weak solution, w, say, exists. Then, along the lines of aforementioned analysis and by ellipticity, letting $\varepsilon v = w - u$,

$$\begin{aligned} I(w) = I(u + (w-u)) &= I(u) + 2\langle \mathcal{L}u - f, w-u\rangle + \langle \mathcal{L}(w-u), w-u\rangle \\ &= I(u) + \langle \mathcal{L}(w-u), w-u\rangle > I(u) \end{aligned}$$

and, likewise, $I(u) < I(w)$, and this is a contradiction. □

Example Let $\mathcal{L} = -\frac{\mathrm{d}}{\mathrm{d}x}\left(p(x)\frac{\mathrm{d}}{\mathrm{d}x}\right) + q(x)$, where $p > 0$, $q \ge 0$, $x \in [0,1]$, and $p \in C^1$, with zero b.c. We employ the standard L_2 inner product. Since

$$\begin{aligned} \langle \mathcal{L}u, v\rangle &= \int_0^1 \{-(p(x)u'(x))' + q(x)u(x)\}v(x)\,\mathrm{d}x \\ &= \int_0^1 u(x)\{-(p(x)v'(x))' + q(x)v(x)\}\,\mathrm{d}x = \langle u, \mathcal{L}v\rangle, \end{aligned}$$

$\mathcal{L}$ is self-adjoint. It is elliptic, since

$$\langle \mathcal{L}v, v\rangle = \int_0^1 \{p(x)v'^2(x) + q(x)v^2(x)\}\,\mathrm{d}x > 0, \qquad v \neq 0.$$

Consequently, $I(v) = \int_0^1 (pv'^2 + qv^2 - 2fv)\,\mathrm{d}x$, as before.

Boundary conditions. There is a major discrepancy between DEs and variational problems – DEs require a full complement of boundary conditions, whereas variational problems can survive with less. Thus, in the last example, the DE b.c. $u(0) = \alpha$, $u(1) = \beta$ translate intact, whereas $u(0) = \alpha$, $u'(1) = 0$ translate into the variational b.c. $u(0) = \alpha$ – in general, *natural* b.c. are discarded and only *essential* b.c. survive. The story is more complicated for more general b.c.'s. Thus, for example, the two-point ODE $\frac{\partial F}{\partial u} - \frac{\mathrm{d}}{\mathrm{d}x}\frac{\partial F}{\partial u_x} = 0$ with the b.c. $\frac{\partial F}{\partial u_x} + \frac{\partial g_a}{\partial u} = 0$ (for $x = a$) and $\frac{\partial F}{\partial u_x} + \frac{\partial g_b}{\partial u} = 0$ (at $x = b$), where g_a and g_b are given functions of x and u, corresponds to the functional

$$I(u) = \int_a^b F(x, u, u_x)\,\mathrm{d}x + g_b(x,u)\Big|_{x=b} - g_a(x,u)\Big|_{x=a}.$$

The Galerkin method. It consists of seeking a weak solution in a finite-dimensional space $\mathbb{H}_M \subset H$, without an intermediate stage of Euler–Lagrange equations. If the space is spanned by the functions $\varphi_1, \ldots, \varphi_M$, this gives for $-(pu')' + qu = f$ exactly the same linear equations as (4.1).

The error in Galerkin's method. Suppose 0 b.c., $\mathcal{L}$ linear, and let $a(u,v) := \langle \mathcal{L}u, v\rangle$. Thus, a is a *bilinear form*. We say that a is *coercive* if $a(v,v) \ge \gamma\|v\|^2\ \forall v \in H$ and *bounded* if $|a(v,w)| \le \alpha\|v\|\cdot\|w\|$ $\forall v, w \in H$, where $\alpha, \gamma > 0$. Note that the Galerkin equations are $a(u,v) = \langle f, v\rangle$, $v \in \mathbb{H}_M$.

Theorem 26 (The Lax–Milgram theorem) *Suppose that $\mathcal{L}$ is linear, coercive and bounded. Then there exists a unique $u_M \in \mathbb{H}_M$ s.t. $a(u_M, v) = (f, v)\ \forall v \in \mathbb{H}_M$. In particular, letting $M \to \infty$, the DE itself possesses a unique weak solution in H. Furthermore, the error of the Galerkin method obeys*

$$\|u_M - u\| \le \frac{\alpha}{\gamma} \inf_{v \in \mathbb{H}_M} \|v - u\|.$$

Consequently, the error in the solution can be bounded by the error of *projecting* H functions into $\mathbb{H}_M$.
Note, incidentally, that (like Theorem 24), the last theorem is about analytic and numerical solutions alike!

4.3 Finite element functions

The main idea in the FE method is to choose $\mathbb{H}_M = \text{Span}\{\varphi_1, \ldots, \varphi_M\}$ as a linear combination of functions with *small local support*, since then the coupling between equations is weak, hence the linear (or, in general, nonlinear) algebraic systems that we need to solve are sparse. In other words, we partition the underlaying space into 'elements' $\mathcal{E}_1 \ldots, \mathcal{E}_N$ such that only few functions in $\{\varphi_1, \ldots, \varphi_M\}$ are nonzero in each $\mathcal{E}_l$, $l = 1, \ldots, N$.

With a great deal of handwaving... Suppose that the highest derivative present (in the variational functional $\mathcal{I}$ or the Galerkin functional a) is $s \ge 1$. Note that we use integration by parts to depress s as much as possible: $s = 1$ for Poisson, $s = 2$ for the biharmonic equation etc. We are using a basis of 'similar' functions $\Phi^{(M)} := \{\varphi_1^{(M)}, \ldots, \varphi_M^{(M)}\}$ (e.g., each such function is a linear translate of the same 'master function'). Denoting by h_M the *diameter* of the maximal element in the underlying partition, we assume that $\lim_{M\to\infty} h_M = 0$. We say that $\Phi^{(M)}$ is of *accuracy* p if, inside each element, we can represent an arbitrary pth degree polynomial as a linear combination of elements of $\Phi^{(M)}$. Moreover, we say that $\Phi^{(M)}$ is of *smoothness* q if each $\varphi_k^{(M)}$ is smoothly differentiable $(q-1)$ times and the qth derivative exists in the sense of distributions (i.e., almost everywhere). For example, translates of the *chapeau function* (a.k.a. the hat function)

$$\psi(x) = \begin{cases} 1 - |x| & : |x| \le 1, \\ 0 & : |x| \ge 1 \end{cases}$$

are of both accuracy and smoothness 1. It is possible to prove for elliptic problems that, for $p = q \ge s$,

$$\|u_M - u\| \le c h_M^{p+1-s} \|u^{(p+1)}\|.$$

Hence convergence.

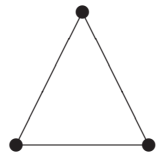
Larger $p = q$ can be obtained by using pth degree *B-splines*. Splines are C^{p-1} functions that reduce to a degree-p polynomial inside each element, and B-splines form a basis of splines with the least-possible support – each spline lives in just $p + 1$ elements.

$p = 1$ gives the chapeau functions, whereas $p = 2$ requires translates of

$$\psi(x) = \begin{cases} \frac{1}{6}x^2 & : 0 \le x \le 1, \\ -\frac{1}{6}(2x^2 - 6x + 3) & : 1 \le x \le 2, \\ \frac{1}{6}(x^2 - 6x + 9) & : 2 \le x \le 3, \\ 0 & : \text{otherwise.} \end{cases}$$

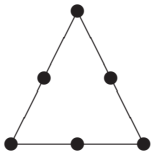
Finite elements in 2D. It is convenient to define 2D piecewise-polynomial elements in terms of interpolation conditions.

Triangular finite elements are convenient for complicated boundaries. Consider first piecewise-linears with interpolation conditions at •. These are the *pyramid* functions.

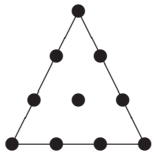


Inside each element the function is $\varphi(x,y) = \alpha + \beta x + \gamma y$ for some $\alpha, \beta, \gamma \in \mathbb{R}$. Thus, $p = 1$. To prove that also $q = 1$, we require continuity along the edges. This is trivial: along each edge, φ is a linear function in a single variable. hence, it is determined uniquely by the interpolation conditions along that edge, conditions that are shared by both elements that adjoin it!

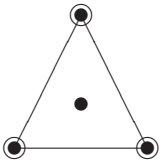
Higher values of p can be obtained with piecewise-quadratics and piecewise-cubics by specifying the interpolation conditions



and

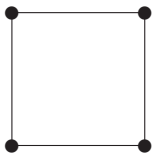


with $p = 2$ and $p = 3$ respectively. However, $q = 1$ for both. To increase q, we may use instead



where ◉ means that we are interpolating (with piecewise-cubics) both to a function value u and to u_x and u_y. The required smoothness follows by Hermite interpolation (to a function and its directional derivative, a linear combination of u_x and u_y).

Quadrilateral elements are best dealt with in the case of rectangles, employing *bipolynomial* functions $\varphi(x,y) = \varphi_1(x)\varphi_2(y)$. It is easy to verify now that their values along each edge are independent of their values elsewhere, hence automatically $q \geq 1$, provided that we have at least two interpolation conditions along each edge. A typical example is the *pagoda* function



4.4 Initial value problems

Example: $u_t = u_{xx}$, 0 b.c. Let $u_M(x,t) = \sum_{m=1}^{M} a_k(t)\varphi_k(x)$, where each φ_k is a translate of the chapeau function, supported on $\left(\frac{k-1}{M+1}, \frac{k+1}{M+1}\right)$. Requiring

$$\left\langle \frac{\partial u_M}{\partial t} - \frac{\partial^2 u_M}{\partial x^2}, \varphi_l \right\rangle = 0, \qquad l = 1, 2, \ldots, M,$$

gives the semidiscretized ODE system

$$\tfrac{1}{6}a'_{m-1} + \tfrac{2}{3}a'_m + \tfrac{1}{6}a'_{m+1} = \frac{1}{(\Delta x)^2}(a_{m-1} - 2a_m + a_{m+1}), \quad m = 1, \ldots, M, \quad \Delta x := \frac{1}{M+1}.$$

Of course, as soon as we are solving initial value problems, stability considerations apply to FEM just as they do to finite differences.

Exercises

4.1 Find a variational problem with the Euler–Lagrange equation

$$u_{xx} + u_{yy} = \mathrm{e}^u, \quad (x,y) \in \mathcal{D},$$

where $\mathcal{D}$ is a given, simply-connected, two-dimensional domain (the *radiation equation*).

4.2 Prove that the *biharmonic* operator $\mathcal{L} = \nabla^4$, acting in a simply-connected two-dimensional domain $\mathcal{D}$, with zero boundary conditions (on functions *and* derivatives) imposed along $\partial\mathcal{D}$, is positive-definite.

4.3 The solution of the equation $-u''(x) + f(x) = 0$, $x \in [0,1]$, with zero boundary conditions, is approximated by $u_n(x) = \sum_{k=1}^{n} a_k \phi_k(x)$, where the ϕ_k's are *hat functions*. Prove from basic assumptions that the error can be bounded by $|u(x) - u_n(x)| \leq \frac{C}{n+1}$, where C is a constant. Deduce convergence.

4.4 The two-point boundary-value problem

$$-\frac{\mathrm{d}}{\mathrm{d}x}\left(p(x)\frac{\mathrm{d}u(x)}{\mathrm{d}x}\right) + q(x)u(x) = 0, \quad x \in [0,1],$$

where $p(x) > 0$, $q(x) \geq 0$ and p is differentiable in $[0,1]$, is solved by the Ritz method with hat functions. Derive the linear algebraic system of equations (having discretised underlying integrals) for the following cases: (a) zero boundary conditions; (b) $u(0) = a$, $u(1) = b$; and (c) $u(0) = a$, $u'(1) = 0$.

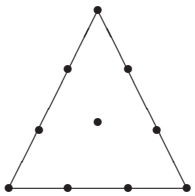
4.5 The Poisson equation in two variables with zero boundary conditions is solved in the unit square by the Ritz method. Given that pyramid functions form the finite-element basis, obtain the explicit form of the algebraic equations.

4.6 1. The diffusion equation $u_t = u_{xx}$ with Dirichlet initial conditions and zero boundary conditions in the strip $0 \leq x \leq 1$, $t \geq 0$, is solved in the following manner: we approximate $u(x,t)$ by $U(x,t) := \sum_{j=1}^{n} a_j(t)\phi_j(x)$, where each ϕ_j's is the hat function with support in $(\frac{j-1}{n}, \frac{j+1}{n})$, and use the Galerkin method to derive a system of ordinary differential equations for the functions $a_1, \ldots, a_n$. Prove that this system is of the form

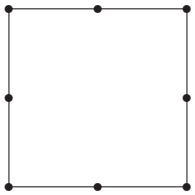
$$\tfrac{1}{6}a'_{j-1} + \tfrac{2}{3}a'_j + \tfrac{1}{6}a'_{j+1} = n^2\left(a_{j-1} - 2a_j + a_{j+1}\right), \quad j = 1, 2, \ldots, n.$$

2. The above equations are similar to semi-discretised finite differences and they admit similar concepts of stability. Thus, rewriting them as $A\boldsymbol{a}' = B\boldsymbol{a}$, prove that, for all initial conditions $\boldsymbol{a}(0)$, it is true that $\lim_{t\to\infty} \|\boldsymbol{a}(t)\| = 0$.

4.7 Show that the two-dimensional cubic polynomial has ten terms, to match the ten nodes in the triangle below. Prove that there is continuity across the edges and show the underlying orders of smoothness and accuracy.

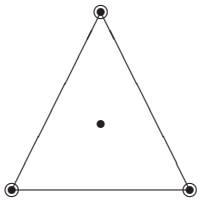


4.8 The square



has 8 interpolation points, rather than the usual 9 interpolation points for a bi-quadratic element. Therefore we remove the x^2y^2 term and keep $b_1 + b_2x + b_3y + b_4x^2 + b_5xy + b_6y^2 + b_7x^2y + b_8xy^2$. Find the function that equals 1 at $x = y = 0$ and vanishes at all other interpolation points.

4.9 We consider the interpolation pattern



where, as usual, • denotes interpolation to the *function value*, whereas ◉ stands for interpolation to *the function value and its first derivatives*. Show that a cubic has the right number of coefficients to fit this pattern and derive orders of smoothness and accuracy of the underlying finite element functions.

4.10 Form (by specifying interpolation points) a piecewise-linear basis of tetrahedral finite elements in $\mathbb{R}^3$. Derive its smoothness and accuracy orders.

Bibliography

[1] W. Hackbusch, *Elliptic Differential Equations. Theory and Numerical Treatment*, Springer-Verlag, Berlin, 1992.

[2] A.R. Mitchell and R. Wait, *The Finite Element Method in Partial Differential Equations*, Wiley, London, 1977.

[3] G. Strang and G. Fix, *An Analysis of the Finite Element Method*, Prentice–Hall, Englewood Cliffs, 1973.

5 Solution of sparse algebraic systems

5.1 Fast Poisson solvers

The Hockney method. We solve $\nabla^2 u = f$ with the 5-point formula in a rectangle. Hence $\mathcal{A}\boldsymbol{u} = \boldsymbol{b}$, where

$$\mathcal{A}=\begin{bmatrix} A & I & & 0 \\ I & A & I & \\ & \ddots & \ddots & \ddots \\ 0 & & I & A \end{bmatrix},\ \boldsymbol{u}=\begin{bmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \\ \vdots \\ \boldsymbol{u}_{N-1} \\ \boldsymbol{u}_N \end{bmatrix},\ \boldsymbol{b}=\begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \\ \vdots \\ \boldsymbol{b}_{N-1} \\ \boldsymbol{b}_N \end{bmatrix},\ A=\begin{bmatrix} -4 & 1 & & 0 \\ 1 & -4 & 1 & \\ & \ddots & \ddots & \ddots \\ 0 & & 1 & -4 \end{bmatrix},$$

the matrix A being $M \times M$. Let $A = QDQ$, where $Q_{k,l} = \sqrt{\frac{2}{M+1}} \sin \frac{kl\pi}{M+1}$, D is diagonal, $D_{k,k} = -4 + 2\cos\frac{2k\pi}{M+1}$. Set $\boldsymbol{v}_k := Q\boldsymbol{u}_k$, $\boldsymbol{c}_k = Q\boldsymbol{b}_k$, $k = 1, \ldots, N$. This yields

$$\begin{bmatrix} D & I & & 0 \\ I & D & I & \\ & \ddots & \ddots & \ddots \\ 0 & & I & D \end{bmatrix} \boldsymbol{v} = \boldsymbol{c}. \tag{5.1}$$

Recall that $\boldsymbol{u}$ and $\boldsymbol{b}$ (and hence $\boldsymbol{v}$ and $\boldsymbol{c}$) have been obtained by ordering the grid by *columns*. We now reorder it by *rows*, and this permutation yields $\tilde{\boldsymbol{v}}$ and $\tilde{\boldsymbol{c}}$ respectively. After the rearrangement

(5.1) becomes

$$\begin{bmatrix} \Gamma_1 & O & & \\ O & \Gamma_2 & O & \\ & \ddots & \ddots & \ddots \\ & & O & \Gamma_M \end{bmatrix} \tilde{\boldsymbol{v}} = \boldsymbol{c}, \text{ where } \Gamma_l = \begin{bmatrix} D_{l,l} & 1 & & 0 \\ 1 & D_{l,l} & 1 & \\ & \ddots & \ddots & \ddots \\ 0 & & 1 & D_{l,l} \end{bmatrix}_{N\times N}, k = 1, \dots, M.$$

Hence we decouple the system into M *tridiagonal* subsystems $\Gamma_l \tilde{\boldsymbol{v}}_l = \tilde{\boldsymbol{c}}_l$.
The computational cost of this algorithm:

- Forming the products $\boldsymbol{c}_k = Q\boldsymbol{b}_k$, $k = 1, \dots, N$ $\mathcal{O}\big(M^2N\big)$;
- Solving M tridiagonal $N \times N$ systems $\Gamma_l \tilde{\boldsymbol{v}}_l = \tilde{\boldsymbol{c}}_l$, $l = 1, \dots, M$ $\mathcal{O}(MN)$;
- Forming the products $\boldsymbol{u}_k = Q\boldsymbol{v}_k$, $k = 1, \dots, N$ $\mathcal{O}\big(M^2N\big)$.

Rearrangements are virtually cost-free, since in reality we store variables in a 2D array anyway. The 'bottleneck' are the $2N$ matrix/vector products.

Fast Fourier transform. The matrix/vector products are of the form

$$\sum_{l=1}^{N} x_l \sin \frac{kl\pi}{N+1} = \operatorname{Im} \sum_{l=0}^{N} x_l \exp \frac{\mathrm{i}kl\pi}{N+1}, \qquad k = 1, \dots, N,$$

and these can be performed efficiently by the FFT.
Let Π_n be the n-dimensional space of all bi-infinite complex sequences of period n, i.e. $\boldsymbol{x} = \{x_k\}_{k=-\infty}^{\infty}$, $x_{k+n} = x_k \ \forall k \in \mathbb{Z}$. Set $\omega_n := \exp \frac{2\pi\mathrm{i}}{n}$, a root of unity of degree n. The *discrete Fourier transform (DFT)* is a $\Pi_n \to \Pi_n$ bijection $\mathcal{F}_n$ defined by

$$\mathcal{F}_n \boldsymbol{x} = \boldsymbol{y}, \quad y_k = \frac{1}{n} \sum_{l=0}^{n-1} \omega_n^{-kl} x_l, \quad k \in \mathbb{Z}.$$

It is easy to verify that $\boldsymbol{y} \in \Pi_n$ and that the inverse mapping is

$$\mathcal{F}_n^{-1} \boldsymbol{y} = \boldsymbol{x}, \qquad x_l = \sum_{k=0}^{n-1} \omega_n^{kl} y_k, \quad l \in \mathbb{Z}.$$

On the face of it, DFT (or its inverse) require $\mathcal{O}\big(n^2\big)$ operations. However, by using the *fast Fourier transform (FFT)* it is possible to reduce this to $\mathcal{O}(n \log_2 n)$ in the special case when $n = 2^L$ for any $L \in \mathbb{Z}_+$.

Thus, let $n = 2^L$, $\tilde{\mathcal{F}}_n := n\mathcal{F}_n$ and

$$\boldsymbol{x}^{(E)} := \{x_{2j}\}_{j\in\mathbb{Z}}, \quad \boldsymbol{x}^{(O)} := \{x_{2j+1}\}_{j\in\mathbb{Z}}, \qquad \tilde{\boldsymbol{y}}^{(E)} := \tilde{\mathcal{F}}_{n/2} \boldsymbol{x}^{(E)}, \quad \tilde{\boldsymbol{y}}^{(O)} := \tilde{\mathcal{F}}_{n/2} \boldsymbol{x}^{(O)}.$$

Hence, letting $\tilde{\boldsymbol{y}} = \tilde{\mathcal{F}}_n \boldsymbol{x}$, $\omega^n = 1$ yields

$$\begin{aligned} \tilde{y}_k &= \sum_{j=0}^{2^L-1} \omega_{2^L}^{-kj} x_j = \sum_{j=0}^{2^{L-1}-1} \omega_{2^L}^{-2kj} x_{2j} + \sum_{j=0}^{2^{L-1}-1} \omega_{2^L}^{-(2j+1)k} x_{2j+1} \\ &= \sum_{j=0}^{2^{L-1}-1} \omega_{2^{L-1}}^{-kj} x_j^{(E)} + \omega_{2^L}^{-k} \sum_{j=0}^{2^{L-1}-1} \omega_{2^{L-1}}^{-kj} x_j^{(O)} = \tilde{y}_k^{(E)} + \omega_{2^L}^{-k} \tilde{y}_k^{(O)}, \quad k = 0, 1, \dots, n-1. \end{aligned}$$

Provided that $\tilde{\boldsymbol{y}}^{(E)}$ *and* $\tilde{\boldsymbol{y}}^{(O)}$ *are known*, it costs just n products to evaluate $\tilde{\boldsymbol{y}}$. This can be further reduced by a factor of two by noticing that, for $k \le n/2-1 = 2^{L-1}-1$,

$$\tilde{y}_{k+2^{L-1}} = \tilde{y}_k^{(E)} - \omega_{2^L}^{-k}\tilde{y}_k^{(O)}.$$

Hence the product $\omega_{2^L}^{-k}\tilde{y}_k^{(O)}$ need be performed only for $k \le n/2-1$.

The algorithm is now clear: starting from $\tilde{\mathcal{F}}_1$, we assemble vectors by synthesising even and odd parts and doubling the length. There are L such stages and in each we have $\frac{1}{2}n$ products, hence *in toto* $\frac{1}{2}nL = \frac{1}{2}n\log_2 n$ products.

Suppose that Hockney's method (with FFT) for 127×127 grid takes 1 sec. Then, comparing the number of operations, 'naive' (nonsparse) Gaussian elimination takes 30 days, 10 hours, 3 minutes and 10 sec.

Cyclic Odd-Even Reduction and Factorization (CORF). We wish to solve a block-TST system

$$T\boldsymbol{u}_{j-1} + S\boldsymbol{u}_j + T\boldsymbol{u}_{j+1} = \boldsymbol{b}_j, j = 1, 2, \ldots, N, \tag{5.2}$$

where $\boldsymbol{u}_0 = \boldsymbol{u}_{N+1} \equiv \boldsymbol{0}$. Multiply

$$\begin{aligned} T\boldsymbol{u}_{j-2} + S\boldsymbol{u}_{j-1} + T\boldsymbol{u}_j &= \boldsymbol{b}_{j-1} \qquad \text{by} \qquad T, \\ T\boldsymbol{u}_{j-1} + S\boldsymbol{u}_j + T\boldsymbol{u}_{j+1} &= \boldsymbol{b}_j \qquad \text{by} \qquad -S, \\ T\boldsymbol{u}_j + S\boldsymbol{u}_{j+1} + T\boldsymbol{u}_{j+2} &= \boldsymbol{b}_{j+1} \qquad \text{by} \qquad T \end{aligned}$$

and add the lot. This yields for $m = 1, 2, \ldots, \left[\frac{N}{2}\right]$

$$T^2\boldsymbol{u}_{2(m-1)} + (2T^2 - S^2)\boldsymbol{u}_{2m} + T^2\boldsymbol{u}_{2(m+1)} = T(\boldsymbol{b}_{2m-1} + \boldsymbol{b}_{2m+1}) - S\boldsymbol{b}_{2m}. \tag{5.3}$$

Note that (5.3), which is also a block-TST system, has half the equations of (5.2). Moreover, suppose that we know the solution of (5.3). Then we can fill-in the gaps in the solution of (5.2) by computing $S\boldsymbol{u}_{2m+1} = \boldsymbol{b}_{2m+1} - T(\boldsymbol{u}_{2m} + \boldsymbol{u}_{2m+2})$.
Provided that $N = 2^{L+1}$, $L \in \mathbb{Z}_+$, we can continue this procedure. Thus, letting $S^{(0)} := S$, $T^{(0)} := T$, $\boldsymbol{b}_j^{(0)} := \boldsymbol{b}_j$, $j = 1, \ldots, N$, we have for all $r = 0, 1, \ldots, L$,

$$S^{(r+1)} = 2\left(T^{(r)}\right)^2 - \left(S^{(r)}\right)^2, \qquad T^{(r+1)} = \left(T^{(r)}\right)^2, \tag{5.4}$$

and $\boldsymbol{b}_j^{(r+1)} = T^{(r)}\left(\boldsymbol{b}_{j-2^r}^{(r)} + \boldsymbol{b}_{j+2^r}^{(r)}\right) - S^{(r)}\boldsymbol{b}_j^{(r)}$. Half of equations are eliminated by each stage, and they can be recovered by solving

$$S^{(r-1)}\boldsymbol{u}_{j2^r-2^{r-1}} = \boldsymbol{b}_{j2^r-2^{r-1}}^{(r-1)} - T^{(r-1)}\left(\boldsymbol{u}_{j2^r} + \boldsymbol{u}_{(j-1)2^r}\right), \quad j = 1, \ldots, 2^{L-r}. \tag{5.5}$$

This is the method of *cyclic reduction*.

Suppose that S and T are themselves tridiagonal and that they commute. Then $S^{(1)}$ is quindiagonal, $S^{(2)}$ has 9 diagonals etc. In general, this causes not just fill-in but, far worse, ill-conditioning. It is easy to prove inductively that $S^{(r)}$ is a polynomial of degree 2^r in S and T,

$$S^{(r)} = \sum_{j=0}^{2^{r-1}} c_{2j} S^{2j} T^{2^r-2j} := P_r(S, T).$$

According to (5.4), $T^{(r)} = T^{2^r}$ and

$$P_{r+1}(s, t) = 2t^{2^{r+1}} - \left(P_r(s, t)\right)^2.$$

Thus, letting $\cos\theta := -\frac{1}{2}\,\frac{s}{t}$, we obtain by induction that $P_r(s,t) = -(2t)^{2^r}\cos 2^r\theta$ and we can factorize

$$P_r(s,t) = -\prod_{j=1}^{2^r}\left(s + 2t\cos\frac{(2j-1)\pi}{2^{r+1}}\right),$$

hence

$$S^{(r)} = \prod_{j=1}^{2^r}\left(S + 2T\cos\frac{(2j-1)\pi}{2^{r+1}}\right). \tag{5.6}$$

Using the factorization (5.6), we can solve (5.5) stably, as a sequence of tridiagonal systems.

Fast Poisson solver in a disc. Suppose that $\nabla^2 u = f$ is given in the unit disc $x^2 + y^2 < 1$, with Dirichlet b.c. conditions $u = g$ along the circle. We first translate from Cartesian to polar coordinates. Thus $v(r,\theta) = u(r\cos\theta, r\sin\theta)$, $\phi(r,\theta) = f(r\cos\theta, r\sin\theta)$ and

$$\frac{\partial^2}{\partial r^2}v + \frac{1}{r}\frac{\partial}{\partial r}v + \frac{1}{r^2}\frac{\partial^2}{\partial\theta^2}v = \phi, \qquad 0 < r < 1, \quad -\pi \le \theta \le \pi. \tag{5.7}$$

The boundary conditions are

1. $v(1,\theta) = g(\cos\theta, \sin\theta)$, $-\pi \le \theta \le \pi$, inherited from Cartesian coordinates;
2. $v(r,-\pi) = v(r,\pi)$, $0 < r < 1$;
3. $\frac{\partial}{\partial\theta}v(0,\theta) \equiv 0$, $-\pi \le \theta \le \pi$ (since the whole line $r = 0$ corresponds to a single value at the centre of the disc).

We Fourier transform (5.7): letting

$$\hat{v}_k(r) := \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{e}^{-ik\theta}v(r,\theta)\,\mathrm{d}\theta, \qquad k \in \mathbb{Z},$$

we obtain

$$\hat{v}_k'' + \frac{1}{r}\hat{v}_k' - \frac{k^2}{r^2}\hat{v}_k = \hat{\phi}_k, \qquad k \in \mathbb{Z}.$$

The b.c. are $\hat{v}_k(1) = \hat{g}_k$ (that's the obvious bit) and $\hat{v}_k(0) = 0$, $k \neq 0$, $\hat{v}_0'(0) = 0$. Each wavenumber yields an *uncoupled* ODE equation with 2-point b.c. We solve it by central differences (which yield tridiagonal algebraic systems). Provided that FFT is used to approximate the Fourier transforms (the numerical error of such procedure decays *exponentially* with n), the total cost is $\mathcal{O}\big(n^2\log_2 n\big)$, where n is the number of Fourier modes and of points in the discretization of the two-point boundary value problem.

Poisson's equation in general 2D regions. Embedding $\mathbb{R}^2$ in $\mathbb{C}$, we use conformal maps, exploiting the connection between solutions of the Laplace equation and analytic functions. In other words, if we can solve Poisson in the unit disc *and* know (or can find) the conformal map from the disc to our (simply connected) domain, we can synthetize the solution of $\nabla^2 u = f$ therein. Rich theory of numerical conformal maps exists, using mainly integral equations and… FFT. A recent alternative, relevant to other equations (e.g. the biharmonic equation) and to higher dimensions, is to use a *multipole method*, whereby the source term is approximated by a finite number of point sources, whose contribution (that is, 'individual' Green functions) is calculated by a clever mixture of near-field and far-field expansions.

5.2 Splitting of evolution operators

Although fast Poisson solvers can be made to cater for more general equations, and even applied to evolutionary equations in ≥ 2 space dimensions, in the latter case a better approach is to use *dimensional splitting*. Suppose, for simplicity, that $u_t = \mathcal{L}u$, with 0 b.c. Here $\mathcal{L}$ is linear and acts in (x, y). Semidiscretizing, we have $\boldsymbol{u}' = (A + B)\boldsymbol{u}$, where A and B contain the 'contribution' of differentiation in x and y. Note that *both A and B are banded (e.g. tridiagonal), up to known permutation*. Consequently, when approximating $\exp(\alpha A)\boldsymbol{v}$ and $\exp(\alpha B)\boldsymbol{v}$ by rational functions, we obtain banded algebraic systems.

The solution is $\boldsymbol{u}(t) = \exp(t(A + B))\boldsymbol{u}(0)$. Suppose that A and B commute. Then $\exp t(A + B) = \exp tA \times \exp tB$, hence $\boldsymbol{u}^{n+1}$ can be derived from $\boldsymbol{u}^n$ very fast. But this is typically false if A and B fail to commute. In that case we can approximate

1. $\mathrm{e}^{t(A+B)} \approx \mathrm{e}^{tA}\mathrm{e}^{tB}$ (Beam & Warming's splitting), first order.
2. $\mathrm{e}^{t(A+B)} \approx \mathrm{e}^{\frac{1}{2}tA}\mathrm{e}^{tB}\mathrm{e}^{\frac{1}{2}tA}$ (Strang's splitting), second order.
3. $\mathrm{e}^{t(A+B)} \approx \frac{1}{2}\mathrm{e}^{tA}\mathrm{e}^{tB} + \frac{1}{2}\mathrm{e}^{tB}\mathrm{e}^{tA}$ (parallel splitting), second order.

We say that a splitting is *stable* if all the coefficients (in front of the exponentials, A and B) are nonnegative. No stable splittings of order ≥ 3 are possible for linear equations (theorem by Sheng). An example of a third-order unstable splitting is

$$\mathrm{e}^{t(A+B)} \approx \mathrm{e}^{tA}\mathrm{e}^{tB} + \mathrm{e}^{\frac{1}{2}tA}\mathrm{e}^{-\frac{1}{2}tB}\mathrm{e}^{-\frac{1}{2}tA}\mathrm{e}^{\frac{1}{2}tB} - \mathrm{e}^{\frac{1}{2}tB}\mathrm{e}^{-\frac{1}{2}tA}\mathrm{e}^{-\frac{1}{2}tB}\mathrm{e}^{\frac{1}{2}tA}.$$

Inhomogeneous terms and nonzero boundary conditions can be accomodated by discretizing the integral in the variation of constants formula

$$\boldsymbol{u}^{n+1} = \mathrm{e}^{(\Delta t)(A+B)}\left\{\boldsymbol{u}^n + \int_0^{\Delta t} \mathrm{e}^{-\tau(A+B)} f(t_n + \tau)\,\mathrm{d}\tau\right\}$$

with the trapezoidal rule, say. Similar approach can be extended even to nonlinear equations.

An alternative to dimensional splitting is the more usual *operatorial splitting*, whereby we partition the multivariate evolution operator $\mathcal{L}$ into $\mathcal{L}_1 + \cdots + \mathcal{L}_s$, where each $\mathcal{L}_j$ depends on just one space variable. We then approximate $\mathcal{L}u = f$ by the s one-dimensional problems $\mathcal{L}u_j = f_j$. However, unlike dimensional splitting, boundary conditions may become a problem.

The Yošida device. Suppose that a numerical method is *time symmetric:* solving from t_0 to t_1, say, and then from t_1 back to t_0 produces *exactly* the initial value. (For example, the implicit midpoint method and, with greater generality, all Gauss–Legendre RK methods.) Such a method is *always* of an even order p, say (can you prove it?). Denote the *map* corresponding to the method by Ψ_h (where h is the step size), i.e. $\boldsymbol{y}_{n+1} = \Psi_h(\boldsymbol{y}_n)$. (Thus, time symmetry means that $\Psi_{-h} \circ \Psi_h = \mathrm{Id}$.) In that case, letting $\alpha = 1/(2-2^{1/(p+1)}) > 1$, it is true that $\Psi_{\alpha h} \circ \Psi_{(1-2\alpha)h} \circ \Psi_{\alpha h}$ is a map corresponding to a method of order $p+2$. Moreover, this new method is also time symmetric, hence the procedure can be repeated. In other words, we execute three steps to advance from t_n to $t_{n+1} = t_n + h$: a step forward to $t_{(n+\alpha)h} > t_{n+1}$, a step backward to $t_{(n+1-\alpha)h} < t_n$ and, finally, a step forward to t_{n+1}.

This procedure might destabilise a method when the underlying equation is itself unstable when integrated backwards (stiff systems, parabolic PDEs), but it is very useful for *conservative* equations.

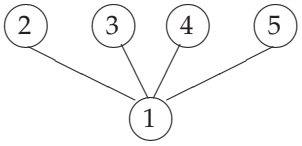
5.3 Sparse Gaussian elimination

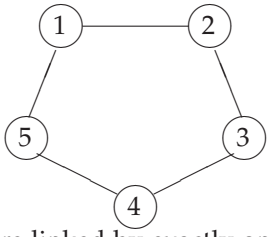
Cholesky factorization. A linear system $A\boldsymbol{x} = \boldsymbol{b}$, A symmetric and positive definite, is solved by *Cholesky factorization*, whereby we first find $A = LL^\top$, where L is lower triangular, and then solve $L\boldsymbol{y} = \boldsymbol{b}$ and $L^\top \boldsymbol{x} = \boldsymbol{y}$. In principle, this is equivalent to symmetric Gaussian elimination, except that we don't need to recompute L in the ubiquitous problem of solving the sequence $A\boldsymbol{x}^{[k+1]} = \boldsymbol{b}^{[k]}$, $k \in \mathbb{Z}_+$, with $\boldsymbol{x}^{[0]}$ given and $\boldsymbol{b}^{[k]}$ dependent on $\boldsymbol{x}^{[0]}, \ldots, \boldsymbol{x}^{[k]}$. The latter occurs, for example, in modified Newton–Raphson equations or when time-stepping an implicit discretization of a linear PDE of evolution with constant coefficients.

If A is banded then both the factorization and the 'backsolving' cost $\mathcal{O}(n)$ operations for an $n \times n$ matrix, whereas the count for general dense matrices is $\mathcal{O}(n^3)$ and $\mathcal{O}(n^2)$ respectively. Our problem: how to factorize a sparse matrix (not necessarily banded) whilst avoiding *fill-in*. We say that the factorization is *perfect* if it avoids fill-in altogether. Thus, factorization of tridiagonal systems is perfect.

Matrices and rooted trees. A *graph* is a collection $\mathcal{G} = \{V, E\}$ of *vertices* $V = \{1, 2, \ldots, n\}$ and *edges* $E \subseteq V^2$. We express a *sparsity pattern* of a symmetric $n \times n$ matrix A by defining a corresponding graph $\mathcal{G}$, letting $(i, j) \in E$ iff $a_{i,j} \neq 0$. Examples:

(i) $A = \begin{bmatrix} \times & \times & 0 & 0 & 0 \\ \times & \times & \times & 0 & 0 \\ 0 & \times & \times & \times & 0 \\ 0 & 0 & \times & \times & \times \\ 0 & 0 & 0 & \times & \times \end{bmatrix} \Rightarrow$ 

(ii) $A = \begin{bmatrix} \times & \times & \times & \times & \times \\ \times & \times & 0 & 0 & 0 \\ \times & 0 & \times & 0 & 0 \\ \times & 0 & 0 & \times & 0 \\ \times & 0 & 0 & 0 & \times \end{bmatrix} \Rightarrow$ 

(iii) $A = \begin{bmatrix} \times & \times & 0 & 0 & \times \\ \times & \times & \times & 0 & 0 \\ 0 & \times & \times & \times & 0 \\ 0 & 0 & \times & \times & \times \\ \times & 0 & 0 & \times & \times \end{bmatrix} \Rightarrow$ 

We say that $\mathcal{G}$ is a *tree* if each two distinct vertices are linked by exactly one path. Given $r \in V$, $(r, \mathcal{G})$ is a *rooted tree* and r is the *root*. Rooted trees admit partial ordering: Let $j \in V \setminus \{r\}$ and let $\{(r, i_1), (i_1, i_2), \ldots, (i_s, j)\}$ be the path joining r and j. We say that j is the *descendant* of $r, i_1, \ldots, i_s$ and that $r, i_1, \ldots, i_s$ are the *ancestors* of j. The rooted tree is *monotonically ordered* if each vertex is numbered before all its ancestors.

Theorem 27 (Parter) *Let $A = LL^\top$ be a positive definite $n \times n$ matrix with a graph that is a monotonically ordered rooted tree. Then $l_{k,j} = a_{k,j}/l_{j,j}$, $j + 1 \leq k \leq n$, $1 \leq j \leq n - 1$. Thus $a_{k,j} = 0 \Rightarrow l_{k,j} = 0$ and the matrix admits perfect factorization.*

Proof By induction. The statement is true for $j = 1$, since always $l_{k,1} = a_{k,1}/l_{1,1}$. Suppose that the theorem is true for $1 \leq j \leq q - 1$. Because of monotone ordering, for every $1 \leq i \leq n - 1$ there exists a unique k_i, $i + 1 \leq k_i \leq n$, s.t. $(i, k_i) \in E$. By the induction assumption, $l_{k,i}$ and $a_{k,i}$ share

the same sparsity structure $\forall 1 \le i \le q-1$, $i+1 \le k \le n$. But

$$l_{k,q} = \frac{1}{l_{q,q}} \left(a_{k,q} - \sum_{i=1}^{q-1} l_{k,i} l_{q,i} \right).$$

Since, for all $q+1 \le k \le n$,
$l_{k,i} \neq 0$ for some $1 \le i \le q-1 \quad \Rightarrow \quad (i,k) \in E \quad \Rightarrow \quad (i,q) \notin E \quad \Rightarrow \quad l_{q,i} = 0,$
$l_{q,i} \neq 0$ for some $1 \le i \le q-1 \quad \Rightarrow \quad (i,q) \in E \quad \Rightarrow \quad (i,k) \notin E \quad \Rightarrow \quad l_{k,i} = 0,$
it follows that $l_{k,q} = a_{k,q}/l_{q,q}$ and the theorem is true. □

Unless we can factorize perfectly, good startegies are (i) factorize with small fill-in; or (ii) block-factorize 'perfectly'.

Arranging a matrix im a narrow band. A typical approach of this kind is the *Cuthill–McKee algorithm*, which aims to order a sparse matrix so that nonzero entries are confined to a narrow band – needless to say, entries outside the band will not be filled in. Let $\beta_k(A) := k - \min\{l : a_{k,l} \neq 0\}$, $k = 1, \ldots, n$. The *bandwidth* of A is $\beta(A) := \max_{k=1,\ldots,n} \beta_k(A)$ and we aim to decrease it by relabelling elements of V.

Let us suppose that $\mathcal{G}$ is *connected*, i.e. that each two distinct vertices $k, l \in V$ are linked by a path, and define the *distance* $d_{k,l}$ as the length of the shortest path connecting k and l. The quantity $e_k := \max\{d_{k,l} : l \in V \setminus \{k\}\}$ is the distance from k to the 'remotest' vertex, whereas $\delta(\mathcal{G}) := \max_{k \in V} e_k$ is called the *eccentricity* of $\mathcal{G}$ – in essence, it is the length of the longest path(s) in the graph. Each $k \in V$ such that $e_k = \delta(\mathcal{G})$ is called a *peripheral vertex*.

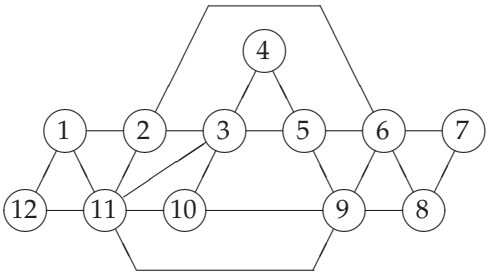
Let $\mathrm{Adj}\,(k) := \{l \in V \setminus \{k\}\,(k,l) \in E\}$ be the set of vertices that adjoin $k = 1, \ldots, n$. We let $\deg(k)$ be the number of elements in $\mathrm{Adj}\,(k)$. Note that the bandwidth can be expressed by using this terminology,

$$\beta(A) = \max\{0, \max\{k - l : l \in \mathrm{Adj}(k),\ k \in V\}\}.$$

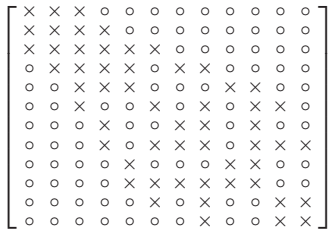
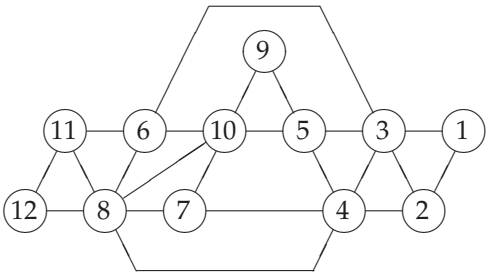
In the Cuthill–McKee algorithm we choose $k_1 \in V$ as a peripheral vertex, label it by "1" and proceed by induction. Having already labelled $k_1, \ldots, k_r$ as "1",…,"r" respectively, in such a manner that all the members of $\mathrm{Adj}\,(k_l)$ have been labelled for $l = 1, \ldots, q$, $q = 1, \ldots, r-1$, we find all the unlabelled elements of $\mathrm{Adj}\,(k_{q+1})$ and label them in an increasing order of degree.

Example: Our point of departure is a matrix with the graph

$$A = \begin{bmatrix}
\times & \times & \circ & \circ & \circ & \circ & \circ & \circ & \circ & \circ & \times & \times \\
\times & \times & \times & \circ & \circ & \times & \circ & \circ & \circ & \circ & \times & \circ \\
\circ & \times & \times & \times & \times & \circ & \circ & \circ & \circ & \times & \times & \circ \\
\circ & \circ & \times & \times & \times & \circ & \circ & \circ & \circ & \circ & \circ & \circ \\
\circ & \circ & \times & \times & \times & \times & \circ & \circ & \times & \circ & \circ & \circ \\
\circ & \times & \circ & \circ & \times & \times & \times & \times & \times & \circ & \circ & \circ \\
\circ & \circ & \circ & \circ & \circ & \times & \times & \times & \times & \circ & \circ & \circ \\
\circ & \circ & \circ & \circ & \times & \times & \times & \times & \times & \circ & \circ & \circ \\
\circ & \circ & \circ & \circ & \times & \times & \circ & \times & \times & \times & \times & \circ \\
\circ & \circ & \times & \circ & \circ & \circ & \circ & \times & \times & \times & \times & \circ \\
\times & \times & \times & \circ & \circ & \circ & \circ & \circ & \times & \times & \times & \times \\
\times & \circ & \circ & \circ & \circ & \circ & \circ & \circ & \circ & \circ & \times & \times
\end{bmatrix} \quad \Leftrightarrow$$



We commence with the peripherial node 7, thus 7∼"1". Since $\mathrm{Adj}\,(7) = \{6, 8\}$ and $\deg(8) = 3 < 5 = \deg(6)$, we label 8∼"2" and 6∼"3". The single unlabelled neighbour of 8 is labelled next, 9∼"4", and then we proceed to the unlabelled members of $\mathrm{Adj}\,(6)$. Since $\deg(2) = 4 = \deg(5)$, we have two options and we set, arbitrarily, 5∼"5" and 2∼"6". Next in line is 9, with two available neighbours – $\deg(10) = 3$, $\deg(11) = 6$, hence 10∼"7" and 11∼"8". Progressing in that manner, we finally obtain the Cuthill–McKee ordering,

$A =$  $\Leftrightarrow$ 

The bandwidth is now $\beta(A) = 5$, clearly superior to the original ordering. Superior but not optimal, however! Typically to combinatorial techniques, the best is the worse enemy of the good – finding the optimal ordering out of 12! combinations will take rather longer than solving the system with any ordering...

5.4 Iterative methods

Let $A\boldsymbol{x} = \boldsymbol{b}$. We consider iterative schemes $\boldsymbol{x}^{(k+1)} = H_k(\boldsymbol{x}^{(0)}, \ldots, \boldsymbol{x}^{(k)})$. There are three considerations: (a) does $\boldsymbol{x}^{(k)}$ converge to $\boldsymbol{x}$? (b) what is the speed of convergence? (c) what is the cost of each iteration?

Linear stationary one-step schemes. Here $H_k(\boldsymbol{x}^{(0)}, \ldots, \boldsymbol{x}^{(k)}) = H\boldsymbol{x}^{(k)} + \boldsymbol{v}$. It is known from introductory numerical analysis courses that convergence $\Leftrightarrow \rho(H) < 1$ and that true solution is obtained if $\boldsymbol{v} = (I - H)A^{-1}\boldsymbol{b}$.

Regular splittings. We split $A = P - N$, where P is a nonsingular matrix, and iterate

$$P\boldsymbol{x}^{(k+1)} = N\boldsymbol{x}^{(k)} + \boldsymbol{b}, \qquad k \in \mathbb{Z}_+. \tag{5.8}$$

In other words, $H = P^{-1}N$. (Of course, the underlying assumption is that a system of the form $P\boldsymbol{y} = \boldsymbol{c}$ can be solved easily.)

Theorem 28 *Suppose that both A and $P + P^\top - A$ are symmetric and positive definite. Then, with the above splitting, $\rho(H) < 1$.*

Proof Let λ and $\boldsymbol{v}$ be an eigenvalue and a corresponding eigenvector of $H = I - P^{-1}A$. Multiplying by P,

$$(I - P^{-1}A)\boldsymbol{v} = \lambda\boldsymbol{v} \qquad \Rightarrow \qquad (1 - \lambda)P\boldsymbol{v} = A\boldsymbol{v}.$$

In particular, $\det A \neq 0$ implies that $\lambda \neq 1$.
Since A is symmetric,

$$\mathbb{R} \ni \boldsymbol{v}^* A\boldsymbol{v} = (1 - \lambda)\boldsymbol{v}^* P\boldsymbol{v} = \overline{(1 - \lambda)\boldsymbol{v}^* P\boldsymbol{v}} = (1 - \bar{\lambda})\boldsymbol{v}^* P^\top \boldsymbol{v}.$$

Consequently (recall that $\lambda \neq 1$),

$$\left(\frac{1}{1 - \lambda} + \frac{1}{1 - \bar{\lambda}} - 1\right) \boldsymbol{v}^* A\boldsymbol{v} = \boldsymbol{v}^*(P + P^\top - A)\boldsymbol{v}. \tag{5.9}$$

Let $\boldsymbol{v} = \boldsymbol{v}_R + i\boldsymbol{v}_I$, $\boldsymbol{v}_R, \boldsymbol{v}_I \in \mathbb{R}^d$. Taking the real part in (5.9) yields

$$\frac{1 - |\lambda|^2}{|1 - \lambda|^2} \left\{\boldsymbol{v}_R^\top A\boldsymbol{v}_R + \boldsymbol{v}_I^\top A\boldsymbol{v}_I\right\} = \left\{\boldsymbol{v}_R^\top (P + P^\top - A)\boldsymbol{v}_R + \boldsymbol{v}_I^\top (P + P^\top - A)\boldsymbol{v}_I\right\}.$$

Since A and $P + P^\top - A$ are positive definite, necessarily $\boldsymbol{w}^\top A\boldsymbol{w}, \boldsymbol{w}^\top(P + P^\top - A)\boldsymbol{w} > 0$ for all $\boldsymbol{w} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$, and we deduce that $1 - |\lambda|^2 > 0$ for all $\lambda \in \sigma(I - P^{-1}A)$. Since this is true for all $\lambda \in \sigma(H)$, the theorem follows. □

Classical methods. Let $A = D - L_0 - U_0$, where $-L_0$, $-U_0$ and D are the lower $\triangle$, upper $\triangle$ and diagonal parts respectively. We assume that $D_{i,i} \neq 0\ \forall i$ and let $L := D^{-1}L_0$, $U := D^{-1}U_0$. The following regular splittings are important:

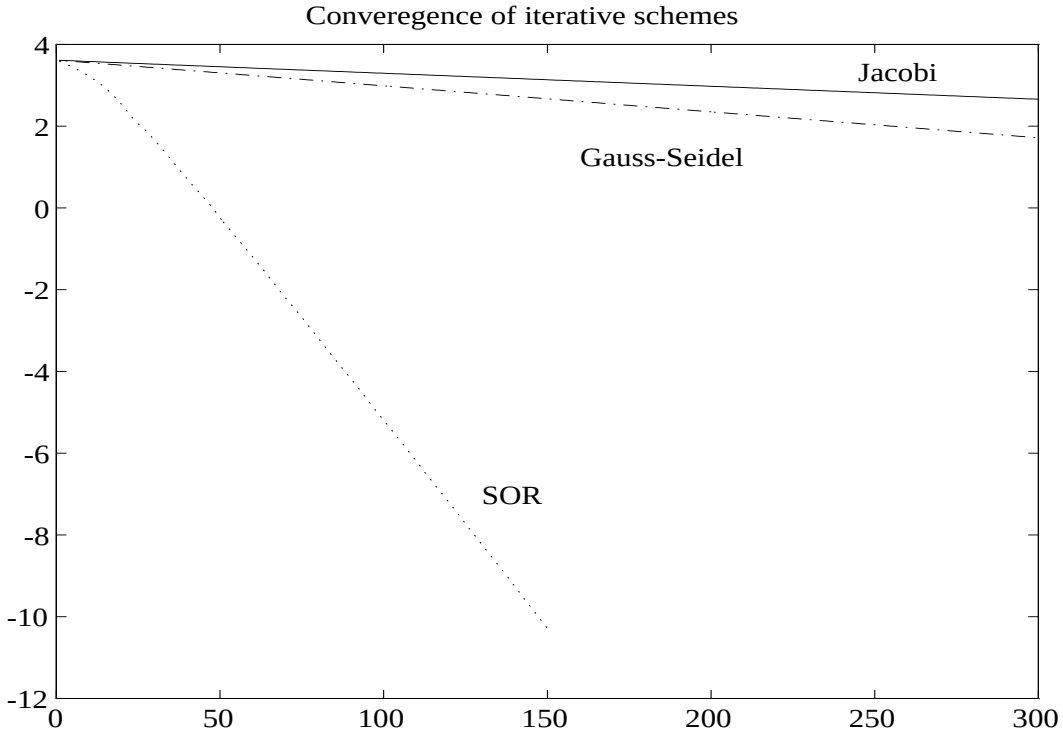


Figure 5.1 Convergence of iterative schemes. $\log_{10}$ of the error is displayed for the first 300 iterations.

Jacobi:
$P = D$ $\quad H = B := L + U$ $\quad \boldsymbol{v} = D^{-1}\boldsymbol{b}$,
Gauss–Seidel:
$P = D - L_0$ $\quad H = \mathcal{L} := (I - L)^{-1}U$ $\quad \boldsymbol{v} = (I - L)^{-1}D^{-1}\boldsymbol{b}$,
Successive over-relaxation (SOR):
$P = \omega^{-1}D - L_0$ $\quad H = \mathcal{L}_\omega := (I - \omega L)^{-1}((1 - \omega)I + \omega U)$ $\quad \boldsymbol{v} = \omega(I - \omega L)^{-1}D^{-1}\boldsymbol{b}$.

Figure 5.1 displays the growth in the number of significant digits in the first 300 iterations of the system

$$x_{k-1} - 2x_k + x_{k+1} = k, \qquad k = 1, 2, \ldots, n,$$

where $x_0 = x_{n+1} = 0$ and $n = 25$, by the three aforementioned methods, with the optimal choice of ω in SOR. Identical type of information features in Figure 5.2, except that now only SOR is used (with optimal choices of the parameter) and $n \in \{25, 50, 75, 100\}$. These figures illustrate behaviour that will be made more explicit in the text.

Example: The 5-point formula in square geometry with natural ordering:

$$\begin{aligned}
\text{J:}\quad & u_{m,n}^{(k+1)} = \tfrac{1}{4}\left(u_{m+1,n}^{(k)} + u_{m-1,n}^{(k)} + u_{m,n+1}^{(k)} + u_{m,n-1}^{(k)} - (\Delta x)^2 f_{m,n}\right),\\
\text{GS:}\quad & u_{m,n}^{(k+1)} = \tfrac{1}{4}\left(u_{m+1,n}^{(k)} + u_{m-1,n}^{(k+1)} + u_{m,n+1}^{(k)} + u_{m,n-1}^{(k+1)} - (\Delta x)^2 f_{m,n}\right),\\
\text{SOR:}\quad & u_{m,n}^{(k+1)} = u_{m,n}^{(k)} + \tfrac{\omega}{4}\left(-4u_{m,n}^{(k)} + u_{m+1,n}^{(k)} + u_{m-1,n}^{(k+1)} + u_{m,n+1}^{(k)} + u_{m,n-1}^{(k+1)} - (\Delta x)^2 f_{m,n}\right).
\end{aligned}$$

Given an $N \times N$ grid, the spectral radii are

$$\begin{aligned}
\text{J:}\quad & \rho(B) = \cos\frac{\pi}{N+1} \approx 1 - \tfrac{\pi^2}{2N^2},\\
\text{GS:}\quad & \rho(\mathcal{L}) = \cos^2\frac{\pi}{N+1} \approx 1 - \tfrac{\pi^2}{N^2},\\
\text{SOR:}\quad & \omega_{\text{opt}} = \frac{2}{1+\sin\frac{\pi}{N+1}}, \quad \rho(\mathcal{L}_{\omega_{\text{opt}}}) = \frac{1-\sin\frac{\pi}{N+1}}{1+\sin\frac{\pi}{N+1}} \approx 1 - \frac{2\pi}{N}.
\end{aligned}$$

Convergence. A matrix $A = (a_{k,l})_{k,l=1}^{d}$ is said to be *strictly diagonally dominant* provided that $|a_{k,k}| \geq \sum_{l\neq k} |a_{k,l}|$, $k = 1, \ldots, d$, and the inequality is sharp for at least one $k \in \{1, \ldots, d\}$.

Theorem 29 *If A is strictly diagonally dominant then the Jacobi method converges.*

Proof An easy consequence of the Gerschgorin disc theorem is that $\rho(B) < 1$, hence convergence. □

Similar theorem can be proved (with more effort) for Gauss–Seidel.

Theorem 30 (The Stein–Rosenberg theorem) *Suppose that $a_{k,k} \neq 0$, $k = 1, \ldots, n$, and that all the entries of B are nonnegative. Then*

either	$\rho(\mathcal{L}) = \rho(B) = 0$	*or*	$\rho(\mathcal{L}) < \rho(B) < 1$
or	$\rho(\mathcal{L}) = \rho(B) = 1$	*or*	$\rho(\mathcal{L}) > \rho(B) > 1.$

Hence, the Jacobi and Gauss–Seidel methods are either simultaneously convergent or simultaneously divergent.

In special cases, more powerful comparison is available:

Theorem 31 *Suppose that A is a tridiagonal matrix with a nonvanishing diagonal. Then $\lambda \in \sigma(B) \Rightarrow \lambda^2 \in \sigma(\mathcal{L})$, whereas $\mu \in \sigma(\mathcal{L})$, $\mu \neq 0 \Rightarrow \pm\sqrt{\mu} \in \sigma(B)$.*

Corollary 6 *For tridiagonal matrices Jacobi and Gauss–Seidel methods converge (or diverge) together and $\rho(\mathcal{L}) = [\rho(B)]^2$.*

Hence, if $\rho(B) < 1$ it is always preferable (in serial computer architecture) to use Gauss–Seidel within the above framework. Of course, the Gauss–Seidel iteration has the added advantage of having only half the computer memory requirements of Jacobi.

Convergence of SOR. Recall the association of sparse matrices with graphs in §5.3. Since we now wish to extend the framework to nonsymmetric matrices, we say that $(k, l) \in E$ if either $a_{k,l} \neq 0$

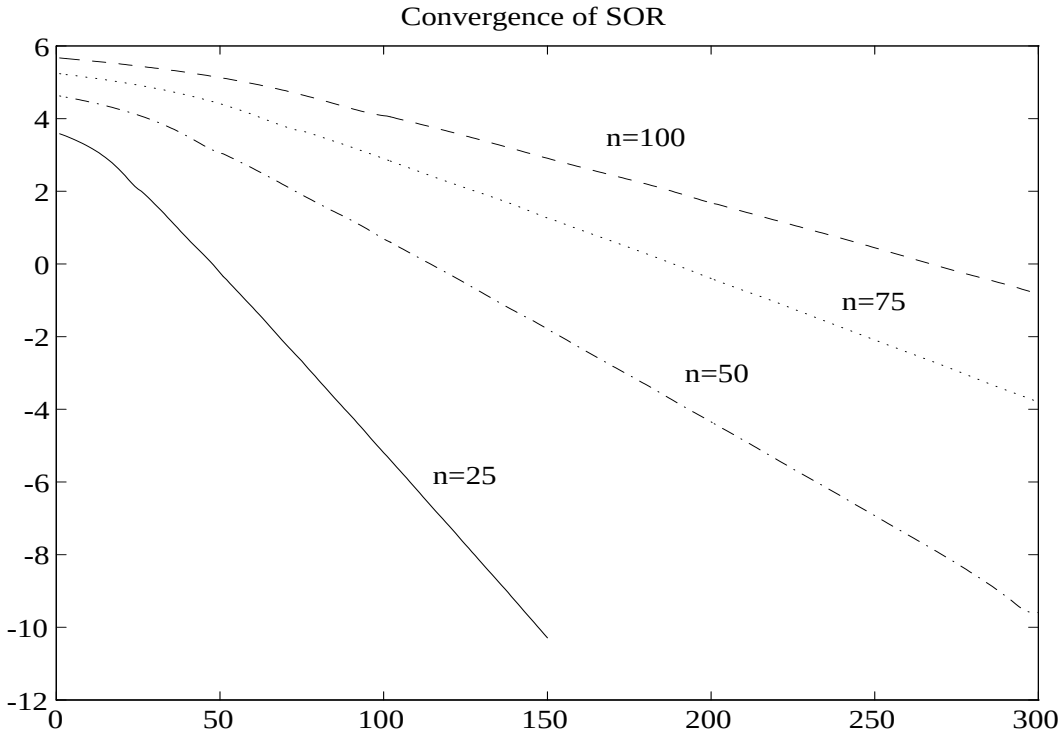


Figure 5.2 Convergence of SOR. $\log_{10}$ of the error is displayed for the first 300 iterations.

or $a_{l,k} \neq 0$.
$\boldsymbol{j} \in \mathbb{Z}^d$ is said to be an *ordering vector* if for every $(k,l) \in E$ it is true that $|j_k - j_l| = 1$. It is said to be a *compatible ordering vector* if, for all $(k,l) \in E$,

$$\begin{aligned} k \geq l+1 \qquad &\Rightarrow \qquad j_k - j_l = 1, \\ k \leq l-1 \qquad &\Rightarrow \qquad j_k - j_l = -1. \end{aligned}$$

Lemma 32 *If A has an ordering vector then there exists a permutation matrix P such that PAP^{-1} has a compatible ordering vector.*

Proof Since each similarity transformation by a permutation matrix is merely a relabelling of variables, the graph of $\tilde{A} := PAP^{-1}$ is simply $\tilde{\mathcal{G}} = \{\pi(V), \pi(E)\}$, where π is a permutation of $\{1,\ldots,d\}$. Thus, $\pi(V) = \{\pi(1),\ldots,\pi(d)\}$ etc.
Denote the ordering vector by J and let $i_k := j_{\pi^{-1}(k)}$, $k = 1,\ldots,d$. It is trivial that $\boldsymbol{i}$ is an ordering vector of $\tilde{A}$: if $(k,l) \in \pi(E)$ then $(\pi^{-1}(k), \pi^{-1}(l)) \in E$, hence $\left|j_{\pi^{-1}(k)} - j_{\pi^{-1}(l)}\right| = 1$. By definition of $\boldsymbol{i}$, this gives $|i_k - i_l| = 1$.
We choose a permutation π such that $i_1 \leq i_2 \leq \cdots \leq i_d$. In that event, given $(k,l) \in E$, $k \geq l+1$, we obtain $i_k - i_l = 1$. Similarly, $i_k - i_l = -1$ if $k \leq l-1$. □

Lemma 33 *Suppose that A has a compatible ordering vector. Then the function*

$$g(s,t) := \det\left(tL_0 + \frac{1}{t}U_0 - sD\right)$$

is independent of $t \in \mathbb{R} \setminus \{0\}$.

Proof It follows from the definition that the matrix $tL_0 + \frac{1}{t}U_0 - sD$ also possesses a compatible ordering vector.
According to the definition of the determinant,

$$g(s,t) = (-1)^d \sum_{\pi\in\Pi_d} (-1)^{|\pi|} \left(\prod_{k=1}^{d} a_{k,\pi(k)}\right) t^{d_L(\pi)-d_U(\pi)} s^{d-d_L(\pi)-d_U(\pi)}, \tag{5.10}$$

where Π_d is the set of all permutations of $\{1,2,\ldots,d\}$, $|\pi|$ is the sign of the permutation π, whereas $d_L(\pi)$ and $d_U(\pi)$ denote the number of values of elements l such that $l > \pi(l)$ and $l < \pi(l)$ respectively. Let $\boldsymbol{j}$ be a compatible ordering vector. Then, unless $a_{k,\pi(k)} = 0$ for some k,

$$d_L(\pi) = \sum_{\substack{l=1\\ \pi(l)<l}}^{d} \left(j_l - j_{\pi(l)}\right), \qquad d_U(\pi) = \sum_{\substack{l=1\\ \pi(l)>l}}^{d} \left(j_{\pi(l)} - j_l\right).$$

Hence, π being a permutation,

$$d_L(\pi) - d_U(\pi) = \sum_{\substack{l=1\\ l\neq\pi(l)}}^{d} (j_l - j_{\pi(l)}) = \sum_{l=1}^{d} (j_l - j_{\pi(l)}) = 0$$

and, by (5.10), g is independent of t. □

Theorem 34 *Provided that A has a compatible ordering vector, it is true that*
(a) *If $\mu \in \sigma(B)$ of multiplicity q then so is $-\mu$;*
(b) *Given $\mu \in \sigma(B)$ and $\omega \in (0,2)$, every λ that obeys the equation*

$$\lambda + \omega - 1 = \omega\mu\lambda^{\frac{1}{2}} \tag{5.11}$$

belongs to $\sigma(\mathcal{L}_\omega)$;
(c) *If $\omega \in (0,2)$ and $\lambda \in \sigma(\mathcal{L}_\omega)$ then there exists $\mu \in \sigma(B)$ so that (5.11) holds.*

Proof According to the last lemma,

$$\det(L_0 + U_0 + \mu D) = g(\mu, 1) = g(\mu, -1) = \det(-L_0 - U_0 + \mu D) = (-1)^d \det(L_0 + U_0 - \mu D).$$

Thus, **(a)** follows from

$$\det(B - \mu I) = \frac{\det(L_0 + U_0 - \mu D)}{\det D} = (-1)^d \det(B + \mu I).$$

Since $\det(I - \omega L) \equiv 1$, we have

$$\begin{aligned}\det(\mathcal{L}_\omega - \lambda I) &= \det\left((I - \omega L)^{-1}(\omega U + (1-\omega)I) - \lambda I\right)\\ &= \det(\omega U + \omega\lambda L - (\lambda + \omega - 1)I).\end{aligned} \tag{5.12}$$

If $\lambda = 0$ lies in $\sigma(\mathcal{L}_\omega)$ then $\det(\omega U - (\omega - 1)I) = 0$ means that $(\omega, \lambda) = (1, 0)$ obeys (5.11). Conversely, if $\lambda = 0$ satisfies (5.11) then $\omega = 1$ and $0 \in \sigma(\mathcal{L}_\omega)$. In the remaining case $\lambda \neq 0$, and then, by Lemma 32 and (5.12),

$$\frac{1}{\omega^d \lambda^{\frac{1}{2}d}} \det(\mathcal{L}_\omega - \lambda I) = \det\left(\lambda^{\frac{1}{2}} L + \lambda^{-\frac{1}{2}} U - \frac{\lambda + \mu - 1}{\omega \lambda^{\frac{1}{2}}} I\right) = \det\left(L + U - \frac{\lambda + \omega - 1}{\omega \lambda^{\frac{1}{2}}} I\right).$$

Hence $\lambda \in \sigma(\mathcal{L}_\omega)$ iff (5.11) holds for some $\mu \in \sigma(B)$. □

Corollary 7 *The existence of a compatible ordering vector implies $\rho(\mathcal{L}) = [\rho(B)]^2$. (Compare with Theorem 30.)*

Theorem 35 *$\rho(\mathcal{L}_\omega) < 1$ implies that $\omega \in (0, 2)$. Moreover, if A has a compatible ordering vector then SOR converges for all $\omega \in (0, 2)$ iff $\rho(B) < 1$ and $\rho(B) \subset \mathbb{R}$.*

Proof Letting $\sigma(\mathcal{L}_\omega) = \{\lambda_1, \ldots, \lambda_d\}$, we have

$$(-1)^d \prod_{k=1}^{d} \lambda_k = \det \mathcal{L}_\omega = \det(\omega U + (1 - \omega)I) + (1 - \omega)^d,$$

therefore

$$\rho(\mathcal{L}_\omega) \geq \left(|1 - \omega|^d\right)^{\frac{1}{d}} = |1 - \omega|.$$

Hence, $\rho(\mathcal{L}_\omega) < 1$ is possible only if $\omega \in (0, 2)$.

If A has a compatible ordering vector then, by Theorem 33, for all $\lambda \in \sigma(\mathcal{L}_\omega)$ there exists $\mu \in \sigma(B)$ such that $p(\sqrt{\lambda}) = 0$, where $p(z) = z^2 - \omega\mu z + \omega - 1$. Using the Cohn–Schur criterion, we can easily verify that both zeros of p lie in the open unit disc iff $\omega \in (0, 2)$ and $|\mu| < 1$. This completes the proof. □

Property A. It should be clear by now that the existence of a compatible ordering vector is important in convergence analysis of SOR. But how to check for this condition?

We say that the matrix A possesses *property A* if there exists a partition $V = S_1 \cup S_2$, $S_1 \cap S_2 = \emptyset$, such that for all $(k, l) \in E$ either $k \in S_1$, $l \in S_2$ or $l \in S_1$, $k \in S_2$.

Lemma 36 *A has property A iff it has an ordering vector.*

Proof If property A holds then we set $j_k = 1$ if $k \in S_1$ and $j_k = 2$ otherwise. Then, for every $(k, l) \in E$, it is true that $|j_k - j_l| = 1$, hence $\boldsymbol{j}$ is an ordering vector.

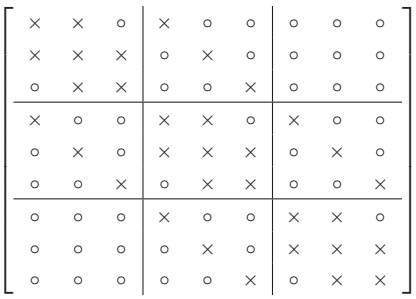
To prove in the other direction we let $\boldsymbol{j}$ be an ordering vector and set

$$S_1 := \{k \in V \,:\, j_k \text{ odd}\}, \qquad S_2 := \{k \in V \,:\, j_k \text{ even}\}.$$

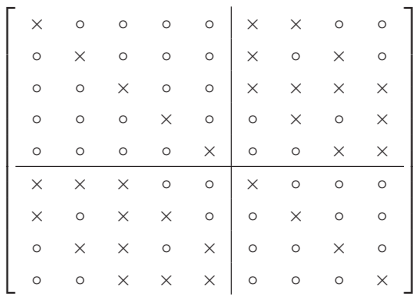
Clearly, $\{S_1, S_2\}$ is a partition of V. Moreover, provided that $(k, l) \in E$, it is easy to verify that k and l belong to distinct members of this partition. hence property A. □

It is, in general, easier to check for property A than for the existence of an ordering vector.

Example: The five-point formula on a rectangular grid.
It is usual to order the grid points columnwise – this is the *natural ordering*. An alternative is to let each (m, n) grid point into S_1 if $m + n$ is even and into S_2 otherwise. Subsequent columnwise ordering, first of S_1 and then of S_2, gives that *red-black ordering*. For example, a 3×3 grid yields



Natural ordering



Red-black ordering

It is easy to verify that the red-black ordering is consistent with property A. Hence, the five-point matrix in a rectangle (and, for that matter, any TST matrix) has property A and, according to Lemmas 31 and 35, can be permuted so that PAP^{-1} has a compatible ordering vector.

Optimal parameter ω. If A has a compatible ordering vector, $\sigma(B) \subset \mathbb{R}$ and $\bar{\mu} := \rho(B) < 1$, then it is possible to determine the value of ω_{opt} such that $\rho(\mathcal{L}_\omega) > \rho(\mathcal{L}_{\omega_{\text{opt}}})$ for all $\omega \neq \omega_{\text{opt}}$. Specifically,

$$\omega_{\text{opt}} = \frac{2}{1+\sqrt{1-\bar{\mu}^2}} = 1 + \left(\frac{\bar{\mu}}{1+\sqrt{1-\bar{\mu}^2}}\right)^2 \in (1,2).$$

To prove this, we exploit Theorem 33 to argue that, $\sigma(B)$ being real, each $\lambda \in \sigma(\mathcal{L}_\omega)$ is of the form

$$\lambda = \tfrac{1}{4}\left\{\omega|\mu| + \sqrt{(\omega\mu)^2 - 4(\omega-1)}\right\}^2.$$

Hence, as long as $2\left(1-\sqrt{1-\mu^2}\right)/\mu^2 \leq \omega < 2$, the argument of the square root is nonpositive, hence $|\lambda| = \omega - 1$. Thus, in this interval $\rho(\mathcal{L}_\omega)$ increases strictly monotonically in ω.
Otherwise, if $0 < \omega \leq 2\left(1-\sqrt{1-\mu^2}\right)/\mu^2$, then $|\lambda| = \frac{1}{4}[F(\omega,|\mu|)]^2$, where

$$F(\omega,t) := \omega t + \sqrt{(\omega t)^2 - 4(\omega-1)}.$$

It is trivial to ascertain that, for all $t \in [0,1)$ and ω in the aforementioned range, F increases strictly monotonically as a function of t. Thus, the spectral radius of $\mathcal{L}_\omega$ in $\omega \in (0, 2(1-\sqrt{1-\bar{\mu}^2})/\bar{\mu}^2] = (0, \omega_{\text{opt}}]$ is $\frac{1}{4}[F(\omega,\bar{\mu})]^2$.
Next we verify by elementary means that F decreases strictly monotonically as a function of ω. Thus, it attains its minimum at $\omega = \omega_{\text{opt}}$. We deduce that, for $\omega \in (0,2)$,

$$\rho(\mathcal{L}_\omega) = \begin{cases} \frac{1}{4}\left\{\omega\bar{\mu} + \sqrt{(\omega\bar{\mu})^2 - 4(\omega-1)}\right\}^2 & : 0 < \omega \leq \omega_{\text{opt}}, \\ \omega - 1 & : \omega_{\text{opt}} \leq \omega < 2 \end{cases}$$

and, for every $\omega \neq \omega_{\text{opt}}$,

$$\rho(\mathcal{L}_\omega) > \rho(\mathcal{L}_{\omega_{\text{opt}}}) = \left(\frac{\bar{\mu}}{1+\sqrt{1-\bar{\mu}^2}}\right)^2.$$

5.5 Multigrid

Consider the model problem of $\nabla^2 u = f$ in a square, solved with the 5-point formula and Gauss–Seidel. Denote by $u_{m,n}$ the exact solution of the linear system, let $\varepsilon^{(k)}_{m,n} = u^{(k)}_{m,n} - u_{m,n}$, and subtract

$$\begin{array}{rrcl} & -4u^{(k+1)}_{m,n} + u^{(k)}_{m+1,n} + u^{(k+1)}_{m-1,n} + u^{(k)}_{m,n+1} + u^{(k+1)}_{m,n-1} & = & (\Delta x)^2 f_{m,n}, \\ - & -4u_{m,n} + u_{m+1,n} + u_{m-1,n} + u_{m,n+1} + u_{m,n-1} & = & (\Delta x)^2 f_{m,n}, \\ \hline = & -4\varepsilon^{(k+1)}_{m,n} + \varepsilon^{(k)}_{m+1,n} + \varepsilon^{(k+1)}_{m-1,n} + \varepsilon^{(k)}_{m,n+1} + \varepsilon^{(k+1)}_{m,n-1} & = & 0. \end{array}$$

Note that $\varepsilon^{(k)}_{m,n}$ satisfies zero boundary conditions.

Hand-waiving, we assume that Fourier harmonics of the error are uncoupled (this would have been true had we had periodic boundary conditions!) and represent $\varepsilon^{(k)}_{m,n} = r^{(k)}_{\theta,\psi} \mathrm{e}^{\mathrm{i}(m\theta + n\psi)}$. Substituting in the above expression,

$$\left(4 - \mathrm{e}^{-i\theta} - \mathrm{e}^{-i\psi}\right) r^{(k+1)}_{\theta,\psi} = (\mathrm{e}^{\mathrm{i}\theta} + \mathrm{e}^{\mathrm{i}\psi}) r^{(k)}_{\theta,\psi}.$$

Thus, the *local attenuation* is

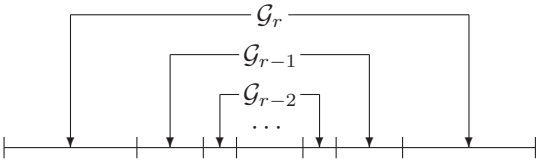
$$\rho^{(k+1)}_{\theta,\psi} := \left| \frac{r^{(k+1)}_{\theta,\psi}}{r^{(k)}_{\theta,\psi}} \right| = \left| \frac{\mathrm{e}^{\mathrm{i}\theta} + \mathrm{e}^{\mathrm{i}\psi}}{4 - \mathrm{e}^{\mathrm{i}\theta} - \mathrm{e}^{\mathrm{i}\psi}} \right|.$$

As $\min\{|\theta|, |\psi|\} = \mathcal{O}(\Delta x)$ it follows, unsurprisingly, that $\rho^{(k+1)}_{\theta,\psi} \approx 1 - c(\Delta x)^2$, the slow attenuation rate predicted by the GS theory. However,

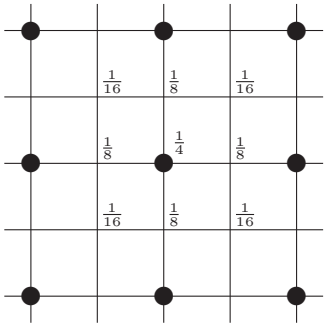
$$\max_{\frac{\pi}{2} \le \max\{|\theta|,|\psi|\} \le \pi} \rho^{(k+1)}_{\theta,\psi} = \rho_{\frac{\pi}{2}, \arctan\frac{3}{4}} = \tfrac{1}{2}.$$

In other words, *the upper half of Fourier frequencies attenuate very fast!*

The multigrid method is based on the realization that different grids cover the range of frequencies differently – thus, having nested grids $\mathcal{G}_r \subset \mathcal{G}_{r-1} \subset \cdots \subset \mathcal{G}_0$, the ranges of 'fast' frequencies are

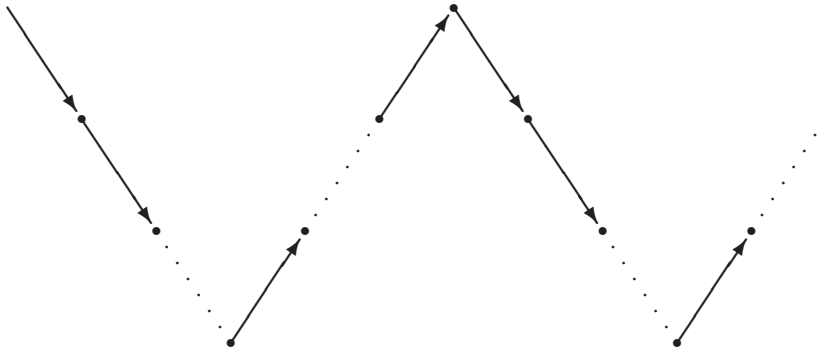


We cover the range of relevant Fourier numbers by this hierarchy of grids. A typical multigrid sweep starts at the *finest* grid, travels to coarsest (on which we solve directly) and back to finest. Each *coarsening* stage involves computing the residual vector $\boldsymbol{r}_h = \boldsymbol{b} - A_h \boldsymbol{x}_h$ and restricting it to the coarse grid $\boldsymbol{r}_{2h} = C\boldsymbol{r}_h$, where we are solving *the residual equation* $A_{2h}\boldsymbol{y}_{2h} = \boldsymbol{r}_{2h}$ etc. Likewise, *refinement* involves a prolongation $\boldsymbol{y}_h = P\boldsymbol{y}_{2h}$ and a correction $\boldsymbol{x}_h^{\text{new}} = \boldsymbol{x}_h + \boldsymbol{y}_h$. A good choice of a restriction matrix linearly combines 9 'fine' values according to the rule



and prolongation by piecewise-linear interpolation reverses this procedure.

The V-cycle Typical implementation of multigird is with the V-cycle



Each coarsening stage typically involves ≈ 5 iterations (i.e., not letting the slow asymptotic attenuation 'take over') and each refinement ≈ 3 iterations. We don't check for convergence, except on the finest grid. Typically, one or two sweeps of multigrid are sufficient for convergence and the cost is often *linear* in the number of variables.

5.6 Conjugate gradients

Let A be symmetric and positive definite. In order to solve $A\boldsymbol{x} = \boldsymbol{b}$ we choose $\nu \in \mathbb{Z}_+$ and let $f_\nu(\boldsymbol{u}) := \frac{1}{2}\boldsymbol{u}^\top A^\nu \boldsymbol{u} - \boldsymbol{b}^\top A^{\nu-1}\boldsymbol{u}$. Thus,

$$\nabla f_\nu(\boldsymbol{u}) = A^\nu \boldsymbol{u} - A^{\nu-1}\boldsymbol{b} = A^{\nu-1}\boldsymbol{r}(\boldsymbol{u}),$$

where $\boldsymbol{r}(\boldsymbol{u}) := A\boldsymbol{u} - \boldsymbol{b}$ is the *residual*. Since A is positive definite, so is $\nabla^2 f_\nu(\boldsymbol{u})$ and

$$\min f_\nu(\boldsymbol{u}) = f_\nu(\boldsymbol{x}) \quad \Leftrightarrow \quad \boldsymbol{r}(\boldsymbol{x}) = \boldsymbol{0} \quad \Leftrightarrow \quad A\boldsymbol{x} = \boldsymbol{b}.$$

$f_\nu(\boldsymbol{u})$ can be rewritten as $\frac{1}{2}(\boldsymbol{u}-\boldsymbol{x})^\top A^\nu(\boldsymbol{u}-\boldsymbol{x}) - \frac{1}{2}\boldsymbol{x}^\top A^\nu \boldsymbol{x}$, and its minimization is the same as minimizing

$$F_\nu(\boldsymbol{u}) := \tfrac{1}{2}(\boldsymbol{u}-\boldsymbol{x})^\top A^\nu(\boldsymbol{u}-\boldsymbol{x}) = \tfrac{1}{2}\boldsymbol{r}(\boldsymbol{u})^\top A^{\nu-2}\boldsymbol{r}(\boldsymbol{u}).$$

We iterate $\boldsymbol{u}^{(k+1)} = \boldsymbol{u}^{(k)} + \lambda_k \boldsymbol{d}^{(k)}$. Thus, the residual obeys $\boldsymbol{r}^{(k+1)} = \boldsymbol{r}^{(k)} + \lambda_k A\boldsymbol{d}^{(k)}$, $k \in \mathbb{Z}_+$. Substituting in F_ν yields the quadratic

$$F_\nu(\boldsymbol{u}^{(k+1)}) = F_\nu(\boldsymbol{u}^{(k)} + \lambda_k \boldsymbol{d}^{(k)}) = \tfrac{1}{2}\left\{ {\boldsymbol{r}^{(k)}}^\top A^{\nu-2}\boldsymbol{r}^{(k)} + 2\lambda_k {\boldsymbol{d}^{(k)}}^\top A^{\nu-1}\boldsymbol{r}^{(k)} + \lambda_k^2 {\boldsymbol{d}^{(k)}}^\top A^\nu \boldsymbol{d}^{(k)} \right\},$$

which, since A is pos. def., is minimized when

$$\lambda_k = -\frac{{\boldsymbol{d}^{(k)}}^\top A^{\nu-1}\boldsymbol{r}^{(k)}}{{\boldsymbol{d}^{(k)}}^\top A^\nu \boldsymbol{d}^{(k)}}.$$

The main idea of the conjugate gradient method lies in choosing the *search direction* $\boldsymbol{d}^{(k)} = -\boldsymbol{r}^{(k)} + \beta_k \boldsymbol{d}^{(k-1)}$, where

$$\beta_k = \frac{{\boldsymbol{r}^{(k)}}^\top A^\nu \boldsymbol{d}^{(k-1)}}{{\boldsymbol{d}^{(k-1)}}^\top A^\nu \boldsymbol{d}^{(k-1)}}.$$

It follows at once from the construction that $\boldsymbol{d}^{(k)\top} A^\nu \boldsymbol{d}^{(k-1)} = 0$ and it is possible to prove that, actually, $\boldsymbol{d}^{(k)\top} A^\nu \boldsymbol{d}^{(l)} = 0$ for all $l = 0, 1, \ldots, k-1$. In other words, the search directions are orthogonal w.r.t. the weighted inner product $(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^\top A^\nu \boldsymbol{v}$. In particular, it follows that (in exact arithmetic) the iteration must reach the exact solution in $\leq n$ iterations, where A is $n \times n$.

Preconditioned conjugate gradients. The effective number of iterations of the CG method depends on the number of *clusters of eigenvalues of* A – each iteration 'takes care' of a single cluster. An efficient method to cluster eigenvalues lies in *preconditioning* CG with Gaussian elimination: Suppose that we can solve easily by elimination a 'neighbour' system $\tilde{A}\tilde{\boldsymbol{x}} = \boldsymbol{b}$, where $\tilde{A} \approx A$. Solving $A\boldsymbol{x} = \boldsymbol{b}$ is identical to solving $\tilde{A}^{-1}A\boldsymbol{x} = \tilde{A}^{-1}\boldsymbol{b} = \tilde{\boldsymbol{x}}$ and the matrix $\tilde{A}^{-1}A$ is likely to have few clusters of eigenvalues (in the extreme – and nonsensical – case $\tilde{A} = A$ it has just one eigenvalue at 1). Practical computation requires Gaussian elimination of the 'easy' system in each CG step, to evaluate the residual: $\boldsymbol{r} := \tilde{A}^{-1}A\boldsymbol{x} - \tilde{\boldsymbol{x}} = \tilde{A}^{-1}(A\boldsymbol{x} - \boldsymbol{b})$.

Nonsymmetric matrices. A classical approach is to solve $A^\top A\boldsymbol{x} = \boldsymbol{b}$. Of course, we never actually form $A^\top A$ (which is both very expensive and destroys sparsity) – e.g., instead of $\boldsymbol{r}^{(k)\top} A^\top A \boldsymbol{d}^{(k-1)}$ (i.e. with $\nu = 1$) we evaluate the identical expression $(A\boldsymbol{r}^{(k)})^\top (A\boldsymbol{d}^{(k-1)})$.
A more modern approach is to use special nonsymmetric versions of CG (e.g. GMRES of Saad & Schultz). The latter can be represented in an equivalent form as follows. Consider the iterative procedure

$$\boldsymbol{x}^{(k+1)} = H\boldsymbol{x}^{(k)} + \boldsymbol{b}, \qquad k \in \mathbb{Z}_+. \tag{5.13}$$

Let $\boldsymbol{r}^{(k)} = (I - H)\boldsymbol{x}^{(k)} - \boldsymbol{b}$ be the *residual* in the kth iteration. Having performed m steps of (5.13), where $m \geq 2$ is a given integer, we seek $\alpha_0, \alpha_1, \ldots, \alpha_m \in \mathbb{R}$, $\sum_{k=0}^m \alpha_k = 1$, that minimize

$$\left\| \sum_{k=0}^m \alpha_k \boldsymbol{r}^{(k)} \right\|$$

($\|\cdot\|$ is the ℓ_2 norm) and restart (5.13) with the new initial value

$$\sum_{k=0}^m \alpha_k \boldsymbol{x}^{(k)}.$$

This procedure is repeated until – hopefully – convergence takes place. Note, however, that implementation of GMRES is considerably more complicated than naively solving (e.g. with Lagrange multipliers) the above minimization problem.[2]

Nonlinear equations. The CG method can be generalized to nonlinear systems of equations – in fact, it follows the same logic as other standard methods of unconstrained optimization (steepest descent, variable metric methods etc.): we first choose a direction for the next step and then select a good step-size. Widely used nonlinear CG algorithms are due to Fletcher & Reeves and to Polack & Ribiére.

Exercises

5.1 Show how to modify the Hockney method to evaluate numerically

1. A solution of the Poisson equation in a square with a *Mehlerstellenverfahren* scheme.
2. A solution of the Poisson equation with Dirichlet boundary conditions in a three-dimensional cube.

[2] Practical computation should avoid formation of inner products, which are prone to considerable roundoff errors.

5.2 Outline a fast solver for the Poisson equation with Dirichlet boundary conditions in cylindrical geometry.

5.3 Describe a fast solver for the Helmholtz equation $\nabla^2 u + \lambda u = 0,\ \lambda \geq 0$, with Dirichlet boundary conditions in the unit disc.

5.4 $F(t) := \mathrm{e}^{tA}\mathrm{e}^{tB}$ is the first-order *Beam–Warming splitting* of $\mathrm{e}^{t(A+B)}$.

1. Prove that
$$F(t) = \mathrm{e}^{t(A+B)} + \int_0^t \mathrm{e}^{(t-\tau)(A+B)} \left(\mathrm{e}^{\tau A}B - B\mathrm{e}^{\tau A}\right) \mathrm{e}^{\tau B}\,\mathrm{d}\tau.$$
[Hint: Find explicitly $G(t) := F'(t) - (A+B)F(t)$ and express the solution of the linear matrix ODE $F' = (A+B)F + G$, $F(0) = I$, using variation of constants.]

2. Suppose that a norm $\|\cdot\|$ is given and that there exist real constants μ_A, μ_B and μ_{A+B} such that
$$\|\mathrm{e}^{tA}\| \leq \mathrm{e}^{\mu_A t}, \qquad \|\mathrm{e}^{tB}\| \leq \mathrm{e}^{\mu_B t}, \qquad \|\mathrm{e}^{t(A+B)}\| \leq \mathrm{e}^{\mu_{A+B} t}.$$
Prove that
$$\|F(t) - \mathrm{e}^{t(A+B)}\| \leq 2\|B\| \frac{\mathrm{e}^{(\mu_A+\mu_B)t} - \mathrm{e}^{\mu_{A+B}t}}{\mu_A + \mu_B - \mu_{A+B}}.$$
Hence, if $\mu_A, \mu_B < 0$ then the splitting error remains relatively small even for large t.

5.5 Discuss the solution of PDEs of evolution (in several space variables) by the combination of semidiscretization, dimensional splitting and the ODE solver
$$\boldsymbol{y}_{n+1} = \mathrm{e}^{hA}\boldsymbol{y}_n + \frac{1}{2}h\left(\mathrm{e}^{hA} + I\right)(\boldsymbol{f}(\boldsymbol{y}_n) - A\boldsymbol{y}_n), \tag{5.14}$$
where A is the Jacobian matrix or an approxomation thereof. The method (5.14) can be justified as follows:
$$\boldsymbol{y}' = \boldsymbol{f}(\boldsymbol{y}) \approx A\boldsymbol{y} + (\boldsymbol{f}(\boldsymbol{y}_n) - A\boldsymbol{y}_n).$$
This, in tandem with the variation of constants formula and the trapezoidal rule for integration, motivate
$$\begin{aligned}\boldsymbol{y}(t_{n+1}) &\approx \mathrm{e}^{hA}\boldsymbol{y}_n + \int_0^h \mathrm{e}^{(h-\tau)A}\,\mathrm{d}\tau(\boldsymbol{f}(\boldsymbol{y}_n) - A\boldsymbol{y}_n)\\ &\approx \mathrm{e}^{hA}\boldsymbol{y}_n + \frac{1}{2}\left(\mathrm{e}^{hA} + I\right)(\boldsymbol{f}(\boldsymbol{y}_n) - A\boldsymbol{y}_n) := \boldsymbol{y}_{n+1}.\end{aligned}$$

5.6 Let $F(t) = \mathrm{e}^{tA/2}\mathrm{e}^{tB}\mathrm{e}^{tA/2}$ (the *Strang splitting*).

1. Prove that $F(t) = \mathrm{e}^{t(A+B)} + Ct^3 + \mathcal{O}\left(t^4\right)$ for some matrix $C = C(A,B)$.
2. Prove that there exists $\alpha \in \mathbb{R}$ such that $G(t) = \mathrm{e}^{t(A+B)} + \mathcal{O}\left(t^4\right)$, where
$$G(t) = F(\alpha t)F((1-2\alpha)t)F(\alpha t)$$
(the *Yošida device*).

5.7 We define the *bandwidth* of a symmetric n-by-n matrix A as
$$\beta(A) := \max_{i=1,\ldots,n} \beta_i(A),$$
where $\beta_i(A) := i - \min\{j\ :\ a_{i,j} \neq 0\},\ i = 1,\ldots,n$. Show that, subject to $\beta(A) << n$, if symmetric Cholesky factorization is applied to A with due regard to the banded structure then the operation count is of the order of magnitude of $(\beta(A))^2\, n$.

5.8 A two-dimensional Poisson equation is solved in a square by the nine-point formula. What is the graph of the underlying matrix?

5.9 A three-dimensional Poisson equation is solved in a cube by the seven-point formula (the 3D cousin of the five-point formula). Find the graph of the matrix.

5.10 Prove the convergence of Jacobi and Gauss–Seidel methods, as applied to the Poisson equation in a square, discretized with the five-point formula.

5.11 Given $A\boldsymbol{x} = \boldsymbol{b}$, we consider the following iterative scheme (*Richardson's method*): choosing arbitrary $\boldsymbol{x}^{(0)}$, let

$$\boldsymbol{x}^{(n+1)} = \boldsymbol{x}^{(n)} - \alpha_n(A\boldsymbol{x}^{(n)} - \boldsymbol{b}), \quad n = 0, 1, \ldots, \tag{5.15}$$

where $\alpha_0, \alpha_1, \ldots$ are real constants.

1. Let $\boldsymbol{r}_n := A\boldsymbol{x}^{(n)} - \boldsymbol{b}$ be the *remainder* in the nth iterant. Prove that
$$\boldsymbol{r}_n = \prod_{j=0}^{n-1}(I - \alpha_j A)\boldsymbol{r}_0.$$
2. Integer $m > 0$ is specified and we assume that $\alpha_{m+j} = \alpha_j$ for all $j = 0, 1, \ldots$. Set $p(z) := \prod_{j=0}^{m}(1 - \alpha_j z)$ and prove that (5.15) converges to the correct solution for every starting value if $|p(\lambda_k)| < 1$ for all $\lambda_k \in \sigma(A)$.
3. Let $\mathcal{D}$ be a complex, bounded, convex domain. We say that T_m is the *mth Chebyshev polynomial with respect to $\mathcal{D}$* if $T_m(\,\cdot\,, \mathcal{D})$ is mth degree, monic and
$$\max\{|T_m(z; \mathcal{D})| \,:\, z \in \mathcal{D}\} = \min_{\deg q = m,\ q \text{ monic}} \max\{|q(z)| \,:\, z \in \mathcal{D}\}.$$
Suppose that all that is known about the eigenvalues of A is that $\sigma(A) \subseteq \mathcal{D}$. Prove that the best choice of $\alpha_0, \ldots, \alpha_{m-1}$ is as the reciprocals of the zeros of $T_m(\,\cdot\,, \mathcal{D})$ and that the iteration converges if
$$\rho_m(\mathcal{D}) := \frac{\max\{|T_m(z; \mathcal{D})| \,:\, z \in \mathcal{D}\}}{|T_m(0, \mathcal{D})|} < 1.$$
4. Let T_m be the 'usual' Chebyshev polynomial (of the first kind), $T_m(\cos\theta) = \cos n\theta$. Then $2^{-m+1}T_m$ (the scaling renders the polynomial monic!) is precisely $T_m(\,\cdot\,, [-1, 1])$. Thereby find explicitly $T_m(\,\cdot\,, [a, b])$ and $\rho_m([a, b])$ for any real interval $[a, b]$.
5. Suppose that $\sigma(A) \in [a, b]$, where $0 < a < b$. Prove that the method (with the optimal choice of α_js) converges. Discuss briefly the speed of convergence.

Bibliography

[1] O. Axelsson, *Iterative Solution Methods*, Cambridge University Press, Cambridge, 1996.

[2] G.H. Golub & C.F. Van Loan, *Matrix Computations*, Johns Hopkins Press, Baltimore, 1989.

[3] D.M. Young, *Iterative Solution of Large Linear Systems*, Academic Press, New York, 1971.