Part III: Numerical Solution of DEs - Revision

Lectures by Arieh Iserles, notes by James Moore

1 ODEs: General principles

1.1 Problem formulation

Problem: Solve $\mathbf{y}' = \mathbf{f}(\mathbf{y})$, $\mathbf{y}(\mathbf{0}) = \mathbf{y}_0$, where $\mathbf{f}$ is *analytic* (i.e. around any point $\mathbf{x}$, there is an open neighbourhood on which $\mathbf{f}$ has a convergent Taylor series). Since $\mathbf{f}$ is analytic and $\mathbf{y}' = \mathbf{f}$, $\mathbf{y}$ is analytic.

WLOG we can take the system to be *autonomous*. If $\mathbf{y}' = \mathbf{f}(t, \mathbf{y})$, then defining $\mathbf{w} = (t, \mathbf{y})$, we have $\mathbf{w}' = (1, \mathbf{f}(\mathbf{w}))^T$.

1.2 Numerical solutions and order

Notation: Let $h > 0$ be a time step. Denote a numerical approximation to $\mathbf{y}(nh)$ by $\mathbf{y}_n$ (this notation suggests our numerical scheme gives $\mathbf{y}(\mathbf{0})$ exactly). Sometimes we write $t_n = nh$; then $\mathbf{y}(t_n) \approx \mathbf{y}_n$.

Definition: A numerical method is of *order* p if for all n, $\mathbf{y}_{n+1} - \tilde{\mathbf{y}}(t_{n+1}) = O(h^{p+1})$, where $\tilde{\mathbf{y}}$ is the exact solution of $\mathbf{y}' = \mathbf{f}$, $\mathbf{y}(nh) = \mathbf{y}_n$.

Slogan: The *order* of a method measures the *local error* committed.

Theorem: An order p method commits a global error of order $O(h^p)$; that is, $\mathbf{y}_{n+1} - \mathbf{y}(t_{n+1}) = O(h^p)$, where $\mathbf{y}$ solves $\mathbf{y}' = \mathbf{f}$, $\mathbf{y}(\mathbf{0}) = \mathbf{y}_0$ as usual.

Proof: Add up all local errors: $n \cdot O(h^{p+1}) = (t_n/h) \cdot O(h^{p+1}) = O(h^p)$, since $t_n = nh$ is fixed.

1.3 Higher derivatives and Taylor methods

Notation: Often we need to use $\mathbf{y}''$, $\mathbf{y}'''$, etc. These can be obtained from the differential equation $\mathbf{y}' = \mathbf{f}$ in terms of $\mathbf{f}$ by repeated application of the chain rule.

We write $\mathbf{f}_k(\mathbf{y}) = \mathbf{y}^{(k)}$. Then:

$$\mathbf{f}_0(\mathbf{y}) = \mathbf{y}, \quad \mathbf{f}_1(\mathbf{y}) = \mathbf{f}, \quad \mathbf{f}_2(\mathbf{y}) = \frac{\partial \mathbf{f}}{\partial \mathbf{y}}\mathbf{f}, \quad \text{etc.}$$

The natural way to approximate $\mathbf{y}((n+1)h)$ is to use its *Taylor series*:

$$\mathbf{y}((n+1)h) = \sum_{k=0}^{\infty} \frac{1}{k!} h^k \mathbf{f}_k(\mathbf{y}(nh)).$$

This gives the natural numerical method (by truncating and replacing with the numerical approximation scheme):

Method: The *Taylor method* is

$$\mathbf{y}_{n+1} = \sum_{k=0}^{p} \frac{1}{k!} h^k \mathbf{f}_k(\mathbf{y}_n).$$

$p = 1$ is the *forward Euler method*: $\mathbf{y}_{n+1} = \mathbf{y}_n + h\mathbf{f}(\mathbf{y}_n)$.

Theorem: The *Taylor method* is of order p.

Proof: Since $\tilde{\mathbf{y}}(nh) = \mathbf{y}_n$, we have:

$$\mathbf{y}_{n+1} - \tilde{\mathbf{y}}(t_{n+1}) = \sum_{k=0}^{p} \frac{h^k}{k!} \mathbf{f}_k(\mathbf{y}_n) - \sum_{k=0}^{\infty} \frac{h^k}{k!} \mathbf{f}_k(\mathbf{y}_n) = O(h^{p+1}). \quad \square$$

1.4 Operatorial interpretations

Definition: The *differential operator* is D, defined by $D\mathbf{g}(t) = \mathbf{g}'(t)$. The *shift operator* (by h) is defined by $E\mathbf{g}(t) = \mathbf{g}(t+h)$.

Numerical methods approximate the shift operator E and its powers.

Theorem: $E = e^{hD}$.

Proof: For analytic $\mathbf{g}$, we have:

$$E\mathbf{g}(t) = \mathbf{g}(t+h) = \sum_{k=0}^{\infty} \frac{h^k}{k!} D^k \mathbf{g}(t) = e^{hD}\mathbf{g}(t),$$

by Taylor's theorem. $\square$

Thus numerical methods must approximate e^{hD}. Letting $R(z) = e^z + O(z^{p+1})$, we have:

Method: The 'R' *method* is $\mathbf{y}_{n+1} = R(hD)\mathbf{y}_n$.

Theorem: The 'R' method is of order p.

Proof: We have:

$$\mathbf{y}_{n+1} - \tilde{\mathbf{y}}(t_{n+1}) = R(hD)\mathbf{y}_n - e^{hD}\mathbf{y}_n = O(h^{p+1}),$$

since $R(z) = e^z + O(z^{p+1})$. □

Taylor methods are a special case when $R(z)$ is a polynomial. It's also possible to get other methods by choosing $R(z)$ not a polynomial (e.g. *rational methods* - see below).

1.5 A-stability

Definition: The *linear stability domain* $\mathcal{D}$ of a method is the set of all $z = h\lambda \in \mathbb{C}$ for which $\lim\limits_{n\to\infty} y_n = 0$, where y_n is the numerical solution of the one-dimensional equation $y' = \lambda y$, $y(0) = 1$, with step-size h.

Definition: A method is *A-stable* if $\mathbb{C}^- \subseteq \mathcal{D}$.

Why this definition? The equation $y' = \lambda y$ has a solution decaying to zero iff $\lambda \in \mathbb{C}^-$, which occurs iff $h\lambda \in \mathbb{C}^-$ for $h > 0$.

Theorem: The Taylor methods are not A-stable.

Proof: For the equation $y' = \lambda y$, we have $y'' = \lambda y' = \lambda^2 y$, $y''' = \lambda y'' = \lambda^3 y$, etc. Hence $f_k(y) = \lambda^k y$. Thus Taylor's method is:

$$y_{n+1} = y_n \sum_{k=0}^{p} \frac{(h\lambda)^k}{k!}.$$

So we have domain of stability:

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

For z very large, real and negative, the condition clearly fails. So not A-stable. □

Example: This example show why A-stability is important. Suppose we are solving the system:

$$\mathbf{y}' = \begin{pmatrix} -1 & 1 \\ 0 & -100 \end{pmatrix} \mathbf{y}, \qquad \mathbf{y}(0) = \mathbf{y}_0.$$

One solution component decays as e^{-t} (gently) and the other decays as e^{-100t} (very quickly). For the numerical solution to decay in the e^{-100t} component, we need $-100h$ ($\lambda = -100$) to lie in the method's linear stability domain. If the method is not A-stable, we require h to be very, very small (order $O(1/100)$) to get a chance of decaying.

Definition: An equation for which non-A-stable methods require depressed step length to converge to zero is called a *stiff equation*.

1.6 Rational methods

We now consider choosing $R(z)$ in the 'R' method as a rational function.

Theorem: When

$$R(z) = \sum_{k=0}^{M} p_k z^k \bigg/ \sum_{k=0}^{N} q_k z^k,$$

the 'R' method reduces to the *rational method*:

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

Proof: The method is:

$$\mathbf{y}_{n+1} = \left(\sum_{k=0}^{M} p_k h^k D^k \right) \left(\sum_{k=0}^{N} q_k h^k D^k \right)^{-1} \mathbf{y}_n,$$

and hence the result follows. □

We want $R(z) = e^z + O(z^{p+1})$. When R is rational, it is called a *Padé approximation* to e^z.

Definition: Given a function f, analytic at the origin, the $[M/N]$ *Padé approximation* $R_{M/N}(z)$ to the function f is the quotient of an Mth degree polynomial over an Nth degree polynomial:

$$R_{M/N}(z) = \frac{P_{M/N}(z)}{Q_{M/N}(z)}$$

such that $R_{M/N}(z) = f(z) + O(z^{M+N+1})$.

Theorem: The Padé approximations to a function f exist are unique.

Proof: Existence not proved in this course. For uniqueness, suppose $R_{M/N}(z) = P_{M/N}(z)/Q_{M/N}(z) = f(z) + O(z^{M+N+1})$, $\tilde{R}_{M/N}(z) = \tilde{P}_{M/N}(z)/\tilde{Q}_{M/N}(z) = f(z) + O(z^{M+N+1})$. Subtracting:

$$\frac{P_{M/N}(z)}{Q_{M/N}(z)} - \frac{\tilde{P}_{M/N}(z)}{\tilde{Q}_{M/N}(z)} = O(z^{M+N+1}).$$

Cross-multiplying we have:

$$P_{M/N}(z)\tilde{Q}_{M/N}(z) - \tilde{P}_{M/N}(z)Q_{M/N}(z) = O(z^{M+3N+1}).$$

But the degree of the polynomial on the left is $M+N$; hence it must be the zero polynomial. Result follows. □

In the special case of e^z, it's possible to prove existence of the Padé approximations (via messy induction). The form is $R_{M/N}(z) = P_{M/N}(z)/Q_{M/N}(z)$, where

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

Definition: The $[M/N]$ *rational method* is the rational method for which $R(z) = R_{M/N}(z)$, the $[M/N]$ Padé approximation to e^z.

When we use the Padé approximation for our rational method, we trivially get:

Theorem: The order of the $[M/N]$ rational method is $M+N$.

Examples of rational methods: The most common rational methods are:

- *Backward Euler* ($[0/1]$): $\mathbf{y}_{n+1} = \mathbf{y}_n + h\mathbf{f}(\mathbf{y}_{n+1})$;
- *Trapezoidal rule* ($[1/1]$):
$$\mathbf{y}_{n+1} = \mathbf{y}_n + \frac{1}{2}h\left(\mathbf{f}(\mathbf{y}_n) + \mathbf{f}(\mathbf{y}_{n+1})\right).$$

Theorem: The rational method defined by the function $R(z)$ is A-stable iff (a) all the poles of R reside in $\mathbb{C}^+ = \{z \in \mathbb{C} : \mathrm{Re}(z) > 0\}$ and (b) $|R(iy)| \le 1$ for all $y \in \mathbb{R}$.

Proof: When solving the equation $y' = \lambda y$, we have $f_k(y) = \lambda^k y$, as usual. Thus the rational method is $y_{n+1} = R(h\lambda)y_n$. Hence the method is A-stable iff $|R(h\lambda)| < 1$ for all $\lambda \in \mathbb{C}^-$.

Suppose this holds. Then no poles allowed in $\mathbb{C}^-$, with none on the imaginary axis (otherwise we can tend to imaginary axis from LHP and get arbitrarily large modulus). So (i) holds. (ii) holds by continuity of R, since again we can tend to the imaginary axis from the LHP.

Conversely, suppose we have (i) and (ii). Then R is analytic in the LHP, and on its boundary obeys $|R(z)| \le 1$. By the *maximum modulus principle* then, $|R(z)| < 1$ throughout the LHP, and we're done. □

Theorem (Wanner, Hairer and Nørsett): The $[M/N]$ Padé rational method is A-stable iff $M \le N \le M+2$.

Proof: Not in course. One inequality in one direction is easy to see, though, since $|R(z)| \sim |z|^{M-N}$ for $|z|$ large, so necessary that $M \le N$ for A-stability. □

Example: Consider the rational method:

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

ORDER: We know that the Padé approximation provides the optimal order by definition, so the order of this method can be no more than the $[2/2]$ Padé method, i.e. $2+2=4$. We find that the order is $p=2$ if $a - 2b \ne 1/3$, $p = 3$ if $a - 2b = 1/3$ and $a \ne 1/2$, and $p=4$ if $a = 1/2$, $b = 1/12$.

STABILITY: We use the maximum modulus principle technique. Check iy condition first; the condition $|R(iy)| \le 1$ is equivalent to $|R(iy)|^2 \le 1$, which in our case is equivalent to

$$\left(1 - \left(b - a + \frac{1}{2}\right)y^2\right)^2 + (1-a)^2y^2 \le (1 - by^2)^2 + a^2y^2.$$

Solving, we find the condition: $\left(a - \frac{1}{2}\right)\left(2b - a - \frac{1}{2}\right) \le 0$.

The poles of $R(z)$ are at: $z = \dfrac{a \pm \sqrt{a^2 - 4b}}{2b}$. We need these to be in $\mathbb{C}^+$. Considering the various cases, we find the necessary and sufficient condition is $a > 0$, $b > 0$.

---

2 ODEs: Multi-step methods

2.1 Method and order

Method: A *multi-step* method is a method of the form:

$$\sum_{l=0}^{m} \rho_l \mathbf{y}_{n+l} = h \sum_{l=0}^{m} \sigma_l \mathbf{f}(\mathbf{y}_{n+l}), \qquad \rho_m = 1.$$

Notation: We define the standard polynomials:

$$\rho(w) = \sum_{l=0}^{m} \rho_l w^l, \qquad \sigma(w) = \sum_{l=0}^{m} \sigma_l w^l.$$

Theorem: A multi-step method is of order p if $\rho(w) = \sigma(w)\log(w) + O(|1-w|^{p+1})$.

Proof (intuition - actually wrong): In terms of operators, the multi-step method can be characterised as $\rho(E)\mathbf{y}_n = h\sigma(E)\mathbf{f}(\mathbf{y}_n)$ $(*)$. On the *exact* solution $\tilde{\mathbf{y}}$ of $\mathbf{y}' = \mathbf{f}$, $\mathbf{y}(nh) = \mathbf{y}_n$, we have:

$$(\rho(E) - \sigma(E)\log(E))\,\tilde{\mathbf{y}}(nh) = O(h^{p+1}),$$

since $e^{hD} = E$ on the exact solution. We've also acquired an error from $\rho(w) - \log(w)\sigma(w) = O(|w-1|^{p+1})$; since $E = 1 + O(h)$, this is the correct error.

Subtract the numerical scheme $(*)$. Then:

$$(\rho(E) - \sigma(E)\log(E))\left(\tilde{\mathbf{y}}(nh) - \mathbf{y}_n\right) = O(h^{p+1}).$$

So by the inverse function theorem (take $E = I + O(\Delta x)$ and expand on LHS), the method is of order p. □

In practice, it is much easier to substitute $w = e^{\theta}$ in this condition, and examine the behaviour as $\theta \to 0$.

2.2 The Dahlquist equivalence theorem

Definition: Let $\{\mathbf{y}_{n,h}\}_{n=0}^{T/h}$ denote a numerical solution of an equation on an interval $[0, T]$ with step size h. We say the numerical method is *convergent* if for all T, and for all $t \in [0, T]$, for any sequence $n_k \in \mathbb{N}$ obeying $n_k \cdot h \to t$ as $k \to \infty$, we have

$$\mathbf{y}_{n_k} \to \mathbf{y}(t) \qquad \text{as } h \to 0.$$

Definition: The function ρ *obeys the roots condition* if (i) all of its zeroes are in $|w| \leq 1$, (ii) the zeroes on $|w| = 1$ are simple.

Theorem (Dahlquist Equivalence): A multi-step method is convergent if and only if (i) its order is $p \geq 1$; (ii) it satisfies the root condition.

Proof: We prove that convergence implies the root condition only.

Suppose the method is convergent. Since the method is convergent for all choices of equation and all initial values, WLOG choose the scalar equation $y' \equiv 0$ and $y(0) = 1$. Then the multi-step method reduces to:

$$\sum_{l=0}^{m} \rho_l y_{n+l} = 0.$$

This is a linear difference equation. Let its characteristic equation have roots ω_j with multiplicities μ_j respectively. Then the solution is:

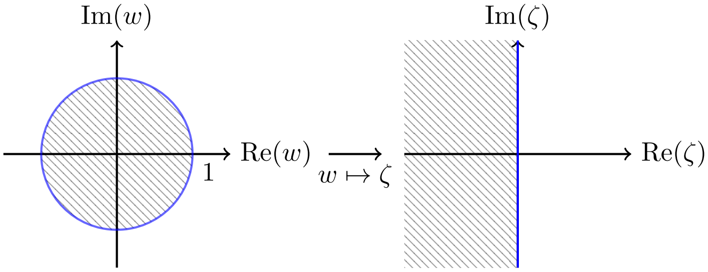
$$y_n = \sum_j \sum_{i=0}^{\mu_j - 1} n^i \omega_j^n \xi_{i,j},$$

for some $\xi_{i,j}$ independent of n. If $|\omega_i| > 1$ for some i, then $y_n \to \infty$ as $n \to \infty$. If $|\omega_i| = 1$ and $\mu_i \geq 2$ for some i, then y_n has polynomial growth. Either way, does not tend to true solution $y(t) \equiv 1$. So root condition necessary. □

2.3 Order and Dahlquist's first barrier

Idea: Dahlquist's first barrier tells us the maximum order of a multi-step method using polynomials of degree m.

The proof converts the root condition disk $|w| \leq 1$ to the LHP via the map $w = (\zeta + 1)/(\zeta - 1)$:



Theorem: Convergence of a multi-step method implies the order must obey $p \leq 2\lfloor (m+2)/2 \rfloor$.

Proof: Since p be the order of our method, so that

$$\rho(w) - \log(w)\sigma(w) = c(w-1)^{p+1} + O(|w-1|^{p+2})$$

for $c \neq 0$. Motivated by the map $w = (\zeta+1)/(\zeta-1)$, taking the unit disk to the LHP, we define the functions:

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

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

We note the following few facts about $R(\zeta)$:

- Note that $r_m = 2^{-m}\rho(1)$. Convergence implies $p \geq 1$, so letting $w \to 1$ in $\rho(w) - \log(w)\sigma(w) = O(|w-1|^{p+1})$ gives $\rho(1) = 0$. Thus $r_m = 0$ and $\deg(R(\zeta)) \leq m - 1$.
- Note that $r_{m-1} = 2^{-m}(2\rho'(1) - m\rho(1)) = 2^{1-m}\rho'(1)$. Convergence implies root condition, so 1 must be a simple zero of ρ; hence $\rho'(1) \neq 0$, and thus $r_{m-1} \neq 0$, i.e. $\deg(R(\zeta)) = m - 1$.
- Since the interior of the disk is mapped to $\mathbb{C}^-$, and the boundary is mapped to $i\mathbb{R}$, we know (from the root condition) that all the zeroes of R are in $\mathbb{C}^-$ or on $i\mathbb{R}$, and the zeroes on $i\mathbb{R}$ must be simple.

  Let ξ_1, ... , ξ_M, $\xi_{M+1} \pm i\nu_{M+1}$, ..., $\xi_N \pm i\nu_N$ be the zeroes of $R(\zeta)$. By the above, $\xi_j \leq 0$. Then

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

  Multiply everything out; since $-\xi_j \geq 0$ and $\nu_j^2 \geq 0$, all (non-zero) coefficients will have same sign as r_{m-1}.

We're now ready to prove the main result. Rewrite the order condition in terms of R and S, by substituting $w = (\zeta+1)/(\zeta-1)$. We then consider the limit $\zeta \to \infty$ (which is equivalent to $w \to 1$):

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

Define $G(\zeta) = (\log((\zeta+1)/(\zeta-1)))^{-1}$. Note that as $|\zeta| \to \infty$, we have

$$G(\zeta) = \left(\log\left(1 + \frac{2}{\zeta-1}\right)\right)^{-1} \sim \left(\log\left(1+\frac{2}{\zeta}\right)\right)^{-1} \sim \frac{\zeta}{2}.$$

Hence:

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

We note we may also write $G(\zeta)$ as a Laurent series:

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

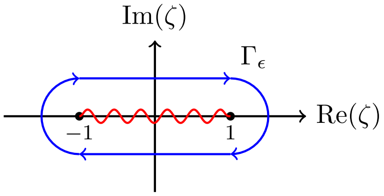
Note there are no even terms since $G(\zeta) = -G(\zeta)$. By Cauchy's integral formula, the coefficients are given by:

$$g_{2l+1} = \frac{1}{2\pi i}\oint_{\Gamma_\epsilon} v^{2l} G(v)\, dv$$

$$= \frac{1}{2\pi i}\int_{-1}^{1} v^{2l}\left\{\frac{1}{\log\left(\frac{1+v}{1-v}\right) + i\pi} - \frac{1}{\log\left(\frac{1+v}{1-v}\right) - i\pi}\right\} dv$$

$$= -\int_{-1}^{1} \frac{v^{2l}\, dv}{\log\left(\frac{1+v}{1-v}\right)^2 + \pi^2} < 0,$$

where Γ_ϵ was the clockwise contour encircling the branch cut from -1 to 1:



Go back to $R(\zeta)G(\zeta) - S(\zeta)$. Since R and S are polynomials, and S has degree m, R degree $m-1$, we can write:

$$R(\zeta)G(\zeta) - S(\zeta) = \sum_{l=-m}^{\infty} e_l \zeta^{-l}.$$

First non-zero e_l tells us order, as per order condition above. All non-zero r_l have same sign, and all g_{2l+1} are negative, hence for $m = 2s$ even (i.e. go up to r_{2s-1}):

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

and for $m = 2s+1$ odd (i.e. go up to r_{2s}):

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

The barrier simplifies (and decreases) if we are using an *explicit* method, i.e. $\sigma_m = 0$.

Theorem: The first Dahlquist barrier for an explicit method is $p \leq m$.

Proof: Not in course. □

2.4 Adams methods

Question: Can we attain the barriers? *Answer:* Yes.

However, when m is even, $m+2$ (for implicit methods) is attainable, but it turns out that all zeroes of ρ live on $|w| = 1$, which is unhealthy. It's better to choose $p = m+1$, for m even or odd, regardless.

We can achieve $p = m+1$ with the *implicit Adams methods:*

Method: The *implicit Adams methods*, called the *Adams-Moulton methods* choose $\rho(w) = w^{m-1}(w-1)$, and then choose $\sigma(w)$ such that the highest possible order implicit method is generated.

If we wish to work wholly with explicit methods, we can attain the barrier $p = m$ with *explicity Adams methods*:

Method: The *explicit Adams methods*, called the *Adams-Bashforth methods*, choose $\rho(w) = w^{m-1}(w-1)$, and then choose $\sigma(w)$ such that the highest possible order explicit method is generated.

Theorem: Implicit Adams methods are of order $m+1$, and explicit Adams methods are of order m. (So implicit Adams methods attain the Dahlquist barrier for m odd, and explicit Adams methods always attain the Dahlquist barrier.)

Proof: Not in course. □

Example: The first two Adams-Moulton methods have $\sigma(w) = \frac{1}{2}(w+1)$, and then $\sigma(w) = \frac{1}{12}(5w^2 + 8w - 1)$. The first two Adams-Bashforth methods have $\sigma(w) = 1$, and then $\sigma(w) = \frac{1}{2}(3w-1)$.

2.5 A-stability and the second barrier

Theorem: Consider the multi-step method defined by the polynomials $\rho(w)$, $\sigma(w)$. Let $T(z,w) = \rho(w) - z\sigma(w)$. The method is A-stable iff for all $\lambda \in \mathbb{C}^-$, the zeroes of $T(\lambda, w) = 0$ are in $|w| < 1$.

Proof: When applied to $y' = \lambda y$, $y(0) = 1$, the multi-step method becomes:

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

This is a difference equation with characteristic polynomial $T(\lambda, \cdot)$. Let its zeroes be $\omega_i(\lambda)$ of multiplicities $\mu_i(\lambda)$. Then

$$y_n = \sum_j \sum_{i=0}^{\mu_j(\lambda)-1} n^i \omega_j(\lambda)^n \xi_{i,j},$$

where $\xi_{i,j}$ are constants independent of n. We see the stability domain is the set of λ for which all the zeroes of $T(\lambda, w) = 0$ are in $|w| < 1$. The result follows. □

Theorem (Dahlquist's second barrier): A-stability of a multi-step method implies $p \leq 2$. Moreover, the second-order method with the least error constant is the trapezoidal rule.

Proof: Not in course. □

2.6 Example application

Consider the multi-step method:

$$\mathbf{y}_{n+3} - (1+2\alpha)\mathbf{y}_{n+2} + (1+2\alpha)\mathbf{y}_{n+1} - \mathbf{y}_n = \frac{1}{6}h\left((5+\alpha)\mathbf{f}(\mathbf{y}_{n+3}) - (4+8\alpha)\mathbf{f}(\mathbf{y}_{n+2}) + (11-5\alpha)\mathbf{f}(\mathbf{y}_{n+1})\right).$$

Convergence: For convergence, we know from Dahlquist equivalence that we need to check (i) the order and (ii) the root condition.

ORDER: In the polynomials $\rho(w)$ and $\sigma(w)$, it's best to substitute $w = e^\theta$ and consider $\theta \to 0$. We have:

$$\rho(e^\theta) - \theta\sigma(e^\theta) = e^{3\theta} - (1+2\alpha)e^{2\theta} + (1+2\alpha)e^{\theta} - 1$$

$$-\frac{1}{6}\theta\left((5+\alpha)e^{3\theta} - (4+8\alpha)e^{2\theta} + (11-5\alpha)e^{\theta}\right).$$

Simplifying, we find:

$$\rho(e^\theta) - \theta\sigma(e^\theta) = -\frac{1}{12}(5+\alpha)\theta^4 - \frac{1}{360}(2260 + 461\alpha)\theta^5 + O(\theta^6).$$

Hence the order is $p = 3$ if $\alpha \neq -5$, and $p = 4$ if $\alpha = -5$. In particular, $p \geq 1$.

ROOT CONDITION: The relevant polynomial is

$$\rho(w) = w^3 - (1+2\alpha)w^2 + (1+2\alpha)w - 1.$$

Since order is $p \geq 1$, we know that 1 is a root. To find the other roots, we must solve:

$$w^2 - 2\alpha w + 1 = 0.$$

This has solutions $w = \alpha \pm \sqrt{\alpha^2 - 1}$, and we need $|w| \leq 1$, $w \neq 1$. We identify two cases: $|\alpha| > 1$ and $|\alpha| < 1$ (if $\alpha = \pm 1$ we get double roots). Analysing the conditions separately, it's clear that $|\alpha| < 1$ is the necessary and sufficient condition.

Thus the method is convergent if and only if $|\alpha| < 1$.

A-STABILITY: Note that the method is of order at least 3 for all values of α. So Dahlquist's second barrier implies this method is not A-stable.

2.7 Multi-step multiderivative methods

Multi-step methods can be extended to include higher derivatives, like rational and Taylor methods. Their A-stability can be constrained by:

Theorem (Wanner-Hairer-Nørsett): Consider a multi-step N-derivative method. Then A-stability implies that $p \leq 2N$. The $2N$-order A-stable method with the least error constant is the 1-step $[N/N]$ Padé method.

Proof: Not in course. □

Note this Theorem reduces to Dahlquist's second barrier immediately by setting $N = 1$, and recalling that the trapezoidal rule is the $[1/1]$ Padé method.

2.8 A-stability of 2-step methods

2-step methods have a simple stability analysis because we need to analyse a quadratic $T(\lambda, w) = a(\lambda)w^2 + b(\lambda)w + c(\lambda)$. The analysis uses the *Cohn-Schur criterion*:

Theorem (Cohn-Schur): 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) $(|a|^2 - |c|^2)^2 \geq |a\bar{b} - b\bar{c}|^2$; (c) if (b) is obeyed as an equality then $|b| < 2|a|$.

Proof: Not required. □

Example: Consider quadratic methods with order $p \geq 2$ such that $h\mathbf{f}(\mathbf{y}_{n+2})$ has coefficient $3/4$. It's possible to show these are parametrised as:

$$\rho(w) = w^2 - (1+a)w + a, \quad \sigma(w) = \frac{3}{4}w^2 - \frac{1}{2}aw + \left(\frac{1}{4} - \frac{1}{2}a\right).$$

Convergence requires $-1 \leq a < 1$ by the root condition. Using the Cohn-Schur criterion, it's also possible to show that the method is A-stable for all values of a in this range.

2.9 Other concepts of stability

Definition: A method is *A(α)-stable* if its linear stability domain $\mathcal{D}$ contains a wedge of angle 2α in $\mathbb{C}^-$ (so A-stability is equivalent to A($90°$)-stability).

This is sufficient for most practical applications.

2.10 Backwards differentiation formulae

Motivation: For $|\lambda| \gg 1$, we have $T(\lambda, w) \approx -\lambda\sigma(w)$. This implies that the 'best' choice for stability chooses σ with all its zeroes at the origin, i.e. $\sigma(w) = \sigma_m w^m$.

Method: An order m *backwards differentiation formula* is a method with $\sigma(w) = \sigma_m w^m$, and order m.

Examples: The first three BDFs are:

- $m = 1$: $\rho(w) = w - 1$, $\sigma(w) = w$;
- $m = 2$: $\rho(w) = w^2 - \frac{4}{3}w + \frac{1}{3}$, $\sigma(w) = \frac{2}{3}w^2$;
- $m = 3$: $\rho(w) = w^3 - \frac{18}{11}w^2 + \frac{9}{11}w - \frac{2}{11}$, $\sigma(w) = \frac{6}{11}w^3$.

The $m = 1, 2$ methods are A-stable, and $m = 3$ is A($86°$)-stable (it can't be A-stable by the second barrier, but it's pretty close!).

Since BDFs are very stable, they are standard in the solution of *stiff equations*.

Theorem: BDFs are convergent iff $m \leq 6$.

Proof: Not in course. □

2.11 $R^{[1]}$ and $R^{[2]}$ methods

Convergence is a minimum requirement; any decent method will also be $R^{[1]}$ and $R^{[2]}$.

Definition: A method is $R^{[1]}$ if the existence and boundedness of the numerical limit $\hat{\mathbf{y}} = \lim_{n\to\infty} \mathbf{y}_n$ (for any $h > 0$) implies $\hat{\mathbf{y}}$ is a fixed point of the exact ODE: $\mathbf{f}(\hat{\mathbf{y}}) = \mathbf{0}$.

Idea: $R^{[1]}$ methods get fixed points of an ODE right.

Definition: A method is $R^{[2]}$ if, for all equations $\mathbf{y}' = \mathbf{f}(\mathbf{y})$, there exists no solution sequence (with any constant step-size $h > 0$) such that both $\hat{\mathbf{y}}_o = \lim_{n\to\infty} \mathbf{y}_{2n+1}$ and $\hat{\mathbf{y}}_e = \lim_{n\to\infty} \mathbf{y}_{2n}$ exist, are bounded, and $\hat{\mathbf{y}}_o \neq \hat{\mathbf{y}}_e$ (such solution sequence is necessarily false!).

Idea: $R^{[2]}$ methods can't oscillate forever.

Theorem: (i) Convergent multi-step methods are $R^{[1]}$; (ii) a convergent multi-step method defined by coprime polynomials ρ, σ, is $R^{[2]}$ if $\sigma(-1) = 0$.

Proof: (a) Suppose that the limit of $\mathbf{y}_n$ as $n \to \infty$ exists and is bounded; denote it by $\hat{\mathbf{y}}$. Taking the limit as $n \to \infty$ in the multi-step method, we have:

$$\left(\sum_{l=0}^{m} \rho_l\right)\hat{\mathbf{y}} = h\left(\sum_{l=0}^{m} \sigma_l\right)\mathbf{f}(\hat{\mathbf{y}}). \qquad (*)$$

Since the method is assumed to be convergent, we have $\rho(1) = 0$. Hence the LHS of $(*)$ is zero. We're left with $0 = h\sigma(1)\mathbf{f}(\hat{\mathbf{y}})$. It remains to show that $\sigma(1) \neq 0$.

Take the derivative of the condition $\rho(w) - \log(w)\sigma(w) = O(|w-1|^{p+1})$. This gives us:

$$\rho'(w) - \frac{\sigma(w)}{w} - \log(w)\sigma'(w) = O(|w-1|^p).$$

Let $w \to 1$, then $\rho'(1) - \sigma(1) = 0$. If $0 = \sigma(1)$, we'd need $\rho'(1) = 0$. But then 1 would be a double root of ρ, violating the root condition.

(b) Show the converse. Suppose $\mathbf{y}_{2n+1} \to \hat{\mathbf{y}}_o$ and $\mathbf{y}_{2n} \to \hat{\mathbf{y}}_e$ with $\hat{\mathbf{y}}_o \neq \hat{\mathbf{y}}_e$.

Considering odd and even terms in the method, and taking the limit we have:

$$\rho_0\hat{\mathbf{y}}_e + \rho_1\hat{\mathbf{y}}_o + ... + \rho_m\hat{\mathbf{y}}_{e/o} = h\left(\sigma_0\mathbf{f}(\hat{\mathbf{y}}_e) + ... + \sigma_m\hat{\mathbf{y}}_{e/o}\right)$$
$$\rho_0\hat{\mathbf{y}}_o + \rho_1\hat{\mathbf{y}}_e + ... + \rho_m\hat{\mathbf{y}}_{o/e} = h\left(\sigma_0\mathbf{f}(\hat{\mathbf{y}}_o) + ... + \sigma_m\hat{\mathbf{y}}_{o/e}\right)$$

Subtracting we have:

$$\rho(-1)(\hat{\mathbf{y}}_e - \hat{\mathbf{y}}_o) = h\sigma(-1)\left(\mathbf{f}(\hat{\mathbf{y}}_e) - \mathbf{f}(\hat{\mathbf{y}}_o)\right).$$

Suppose $\sigma(-1) = 0$. Then since $\hat{\mathbf{y}}_e \neq \hat{\mathbf{y}}_o$, we have $\rho(-1) = 0$. Contradiction, since σ and ρ are relatively prime. So $\sigma(-1) \neq 0$. □

2.12 Switching methods

To end this section, let's consider the benefit of switching up methods.

Example: Consider $\mathbf{y}' = \mathbf{f}(t, \mathbf{y})$ solved by

$$\mathbf{y}_{2n+1} = \mathbf{y}_{2n} + h\mathbf{f}(t_{2n}, \mathbf{y}_{2n}),$$
$$\mathbf{y}_{2n+2} = \mathbf{y}_{2n+1} + h\mathbf{f}(t_{2n+2}, \mathbf{y}_{2n+2}),$$

i.e. swapping backward and forward Euler. This is an *order* 2 method at the points t_{2n} even though the individual methods are only order 1. It's easy to see this, since on elimination of $\mathbf{y}_{2n+1}$, we have:

$$\mathbf{y}_{2n+2} = \mathbf{y}_{2n} + h\left(\mathbf{f}(t_{2n}, \mathbf{y}_{2n}) + \mathbf{f}(t_{2n+2}, \mathbf{y}_{2n+2})\right).$$

This is the trapezoidal rule, so is of order 2, and is also A-stable.

---

3 Implementation of ODE methods

3.1 Solving non-linear algebraic equations

Implicit methods require solution of algebraic equations at each step. We need to consider:

Notation: The *setup costs* for an iterative method are denoted $\mathbf{C}_S$. The *iteration costs* are denoted $\mathbf{C}_I$.

There are two possible ways of deciding when to stop iterations:

Definition: If we continue to iterate a method until the error is beneath a specified tolerance, we say we are *iterating to convergence*. This is written PC^∞. If we execute a small number m of fixed iterations, and abandon the process unless the error is below tolerance, we are using the PC^m method.

Clearly if $\mathbf{C}_S \gg \mathbf{C}_I$, we should use PC^∞; otherwise use PC^m.

One way of solving algebraic equations such as $\mathbf{y} - \beta h\mathbf{f}(\mathbf{y}) = \mathbf{v}$, where $\mathbf{v}$ is known, is by *direct iteration*:

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

This is a special case of searching for a fixed point $\mathbf{x}^{[j+1]} = \mathbf{g}(\mathbf{x}^{[j]})$ for some $\mathbf{g}$. We are guaranteed to find a solution by:

Banach's Contraction Mapping Theorem:
Suppose $||\mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{y})|| \leq L||\mathbf{x} - \mathbf{y}||$ for some $0 < L < 1$ for all $\mathbf{x}$, $\mathbf{y}$ in a ball of radius $r > 0$ around the first iterate: $||\mathbf{x} - \mathbf{x}^{[0]}|| \leq r$, $||\mathbf{y} - \mathbf{x}^{[0]}|| \leq r$. Then if $||\mathbf{g}(\mathbf{x}^{[0]}) - \mathbf{x}^{[0]}|| \leq (1-L)r$, we have:

(i) $||\mathbf{x}^{[j]} - \mathbf{x}^{[0]}|| \leq r$;

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

(iii) $\mathbf{x}^*$ is the unique fixed pt. of $\mathbf{g}$ in $S_r = \{||\mathbf{x} - \mathbf{x}^{[0]}|| \leq r\}$.

Proof: We first prove $||\mathbf{x}^{[j+1]} - \mathbf{x}^{[j]}|| \leq L^j(1-L)r$ by induction. Clearly true for $j = 0$, and then $||\mathbf{x}^{[j+1]} - \mathbf{x}^{[j]}|| =$

$$||\mathbf{g}(\mathbf{x}^{[j]}) - \mathbf{g}(\mathbf{x}^{[j-1]})|| \leq L||\mathbf{x}^{[j]} - \mathbf{x}^{[j-1]}|| \leq L^j(1-L)r,$$

so done. Therefore: $||\mathbf{x}^{[j]} - \mathbf{x}^{[0]}|| =$

$$\left|\left|\sum_{i=0}^{j-1}(\mathbf{x}^{[i+1]} - \mathbf{x}^{[i]})\right|\right| \leq \sum_{i=0}^{j-1} L^i(1-L)r = (1-L^j)r \leq r,$$

so all iterates lie in S_r, proving (i). Note that $\mathbf{x}^{[j]}$ is Cauchy, since for all k, we have

$$||\mathbf{x}^{[k+j]} - \mathbf{x}^{[j]}|| = \left|\left|\sum_{i=0}^{k-1}(\mathbf{x}^{[j+i+1]} - \mathbf{x}^{[j+i]})\right|\right| \leq L^j r \to 0,$$

as $j \to \infty$. Hence it converges to some limit $\mathbf{x}^*$, which clearly must be a fixed point of $\mathbf{g}$, proving (ii).

Finally, suppose $\mathbf{x}^o$ is a fixed point of $\mathbf{g}$ in S_r. Then

$$||\mathbf{x}^* - \mathbf{x}^o|| = ||\mathbf{g}(\mathbf{x}^*) - \mathbf{g}(\mathbf{x}^o)|| \leq L||\mathbf{x}^* - \mathbf{x}^o||.$$

Since $0 < L < 1$, we need $\mathbf{x}^* = \mathbf{x}^o$, and (iii) follows. □

For implicit methods, $\mathbf{g}(\mathbf{x}) = \mathbf{v} + \beta h\mathbf{f}(\mathbf{x})$, hence

$$\mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{y}) = \beta h\mathbf{f}(\mathbf{x}) - \beta h\mathbf{f}(\mathbf{y}) \approx \beta h(\mathbf{x} - \mathbf{y})\frac{\partial \mathbf{f}}{\partial \mathbf{y}}.$$

Hence $L \approx h|\beta|||\partial\mathbf{f}/\partial\mathbf{y}||$. So for stiff equations, getting $L < 1$ may require very small step-size h. In that case, we may instead use *Newton-Raphson*.

Suppose we wish to solve $\mathbf{x} = \mathbf{g}(\mathbf{x})$, and we have a decent guess $\tilde{\mathbf{x}}$. Then

$$\mathbf{x} = \mathbf{g}(\tilde{\mathbf{x}}) + \frac{\partial \mathbf{g}(\tilde{\mathbf{x}})}{\partial \mathbf{x}}(\mathbf{x} - \tilde{\mathbf{x}}) + \cdots$$
$$\Rightarrow \left(I - \frac{\partial \mathbf{g}(\tilde{\mathbf{x}})}{\partial \mathbf{x}}\right)\mathbf{x} = \mathbf{g}(\tilde{\mathbf{x}}) - \frac{\partial \mathbf{g}(\tilde{\mathbf{x}})}{\partial \mathbf{x}}\tilde{\mathbf{x}}$$

On rearrangement, we obtain:

Newton-Raphson: The *Newton-Raphson iteration* is defined by

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

In practice, this is expensive, since the inverse matrix needs to be re-evaluated at every iteration, and we need to solve a new linear system at every step. These problems are cured by:

Modified Newton-Raphson: The *MNR iteration* is defined by

$$\mathbf{x}^{[j+1]} = \mathbf{x}^{[j]} - \left(I - \frac{\partial \mathbf{g}(\mathbf{x}^o)}{\partial \mathbf{x}}\right)^{-1} (\mathbf{x}^{[j]} - \mathbf{g}(\mathbf{x}^{[j]})),$$

where $\mathbf{x}^o$ is fixed, say $\mathbf{x}^o = \mathbf{x}^{[0]}$. We now have

$$L = \left(I - \beta h \frac{\partial \mathbf{f}(\mathbf{x}^o)}{\partial \mathbf{x}}\right)^{-1},$$

which is small for all h, so this is a good approach for stiff equations.

We use direct iteration for non-stiff equations, and MNR for stiff equations. What about PC^∞ versus PC^m?

$\mathbf{C}_S$ are small for direct iteration, but $\mathbf{C}_I$ is large. The opposite is true for MNR. Hence non-stiff equations should use PC^∞, and stiff equations should use PC^m.

3.2 Methods of error control

Error control allows us to control local error and choose step length so that the error is within a specified tolerance.

The Milne device: Let c_P, and c_C be the error constants of a *predictor* (explicit) and *corrector* (implicit) pair of methods of equal order p. Then:

$$\begin{aligned}\mathbf{y}_{n+1}^{(P)} &= \mathbf{y}(t_{n+1}) + c_P h^{p+1} \mathbf{y}^{(p+1)}(t_n) + ...,\\ \mathbf{y}_{n+1}^{(C)} &= \mathbf{y}(t_{n+1}) + c_C h^{p+1} \mathbf{y}^{(p+1)}(t_n) + ...,\end{aligned}$$

and so subtracting:

$$\mathbf{y}_{n+1}^{(P)} - \mathbf{y}_{n+1}^{(C)} \approx (c_P - c_C) h^{p+1} \mathbf{y}^{(p+1)}(t_n).$$

Eliminating $\mathbf{y}_{n+1}^{(P)}$ from the first and third equations, we find:

$$||\mathbf{y}_{n+1}^{(C)} - \mathbf{y}(t_{n+1})|| \approx \left|\frac{c_C}{c_P - c_C}\right| ||\mathbf{y}_{n+1}^{(P)} - \mathbf{y}_{n+1}^{(C)}||.$$

Hence it is possible to estimate the error in the corrector method.

Deferred correction: By example. Consider the trapezoidal rule:

$$\mathbf{y}_{n+1} = \mathbf{y}_n + \frac{1}{2} h \left(\mathbf{f}(\mathbf{y}_n) + \mathbf{f}(\mathbf{y}_{n+1})\right).$$

It's trivial to show the error is $-\frac{1}{12} h^3 \mathbf{y}'''(t_n) + O(h^4)$. Define:

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

Then $\mathbf{s}(\mathbf{y}_{n-1}, \mathbf{y}_n, \mathbf{y}_{n+1}) = -\frac{1}{12} h^3 \mathbf{y}'''(t_n) + O(h^4)$; this is the actual error, up to $O(h^4)$! So we retain a $\mathbf{y}_{n-1}$, and wait for the next step $\mathbf{y}_{n+1}$ (*defer*) to estimate the error in $\mathbf{y}_n$.

The Zadunaisky device: Given a p-order numerical solution $\mathbf{y}_j$, choose a vector of polynomials $\mathbf{q}$ s.t. $\deg(\mathbf{q}) = p$ and interpolates $\mathbf{y}$ at the last $p+1$ grid points.

Define the defect $\mathbf{d}(t) = \mathbf{q}'(t) - \mathbf{f}(\mathbf{q}(t))$. Since $\mathbf{q}(t) = \mathbf{y}(t) + O(h^{p+1})$ (order of the method, and $\mathbf{q}$ agrees with $\mathbf{y}$ at last $p+1$ points), and $\mathbf{y}' = \mathbf{f}(\mathbf{y})$, we see that $\mathbf{d}(t) = O(h^p)$.

Consider the *auxiliary system* $\mathbf{z}' = \mathbf{f}(\mathbf{z}) + \mathbf{d}(t)$. Since $\mathbf{d}(t) = O(h^p)$, this system is within $O(h^{p+1})$ of the original ODE, and $\mathbf{q}(t)$ solves it exactly. Solve this system numerically, and use $\mathbf{z}_{n+1} - \mathbf{q}(t_{n+1})$ to estimate $\mathbf{y}_{n+1} - \mathbf{y}(t_{n+1})$, the error.

3.3 Gear's automatic integration

Suppose we have a family of m-step methods for $m = 1, 2, ..., m^*$, each of order $p_m = m + K$ (e.g. $K = 1$ for Adams-Moulton, $K = 0$, $m* = 6$ for BDF), and each with error constant c_m.

Gear's method:

1. Start iteration with $m = 1$.
2. At the nth step, working with the m-step method, evaluate the error estimates:

   $$\begin{gathered} E_j \approx c_j h^{j+K+1} \mathbf{y}^{(j+K+1)}(t_n),\\ j \in I_m := \{m-1, m, m+1\} \cap \{1, 2, ..., m^*\}.\end{gathered}$$

   This could be achieved by a Zadunaisky-style method, by interpolating numerical points with an polynomial, and using the $(j+K+1)$th derivative of the polynomial to estimate E_j.

   Note we do this for the main m-step method, and its neighbours, $m \pm 1$.
3. Of the $j \in I_m$, determine which j^* is such that E_{j^*} is below tolerance, but h (the step-size) is largest.
4. Change to that method and step-size and calculate the next value. Iterate.

Note that we must retain enough past values for error-control and step-size management.

4 ODEs: Runge-Kutta methods

4.1 Motivation and definition

We can convert $\mathbf{y}'(t) = \mathbf{f}(t, \mathbf{y}(t))$ into an integral equation, and approximate the integral using a *quadrature formula*:

$$\begin{aligned}\mathbf{y}(t_0 + h) &= \mathbf{y}_0 + \int_0^h \mathbf{f}(t_0 + \tau, \mathbf{y}(t_0 + \tau))d\tau \\ &\approx \mathbf{y}_0 + h\sum_{l=1}^{s} b_l \mathbf{f}(t_0 + c_l h, \mathbf{y}(t_0 + c_l h)).\end{aligned}$$

When we turn this into a numerical scheme, we obtain:

Method: An s-stage *Runge-Kutta method* is a method of the form:

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

where the c_l obey the condition (which we will see is required for the method to be of order at least 1):

$$c_l = \sum_{j=1}^{s} a_{l,j}.$$

The parameters can written as two vectors and a matrix: A, $\mathbf{b}$ and $\mathbf{c}$, where $A_{ij} = a_{ij}$. The c_l condition can then be written as $\mathbf{c} = A\mathbf{1}$, where $\mathbf{1}$ is the vector of 1's.

Notation: Runge-Kutta methods are specified by their *Butcher tableau*:

$$\begin{array}{c|c} \mathbf{c} & A \\ \hline & \mathbf{b}^T \end{array}$$

Definition: A Runge-Kutta method is

- *explicit* (ERK) if A is strictly lower triangular;
- *diagonally-implicit* (DIRK) if A is lower triangular;
- *singly-diagonally-implicit* (SDIRK) if A is lower triangular and all diagonal elements $a_{l,l}$ are equal (and non-zero);
- *implicit* (IRK) otherwise.

The benefit of SDIRK over DIRK is we save setup costs in modified Newton Raphson.

4.2 Naïve order analysis

Theorem: For Runge-Kutta methods of order less than or equal to 5, there is no loss of generality in working with scalar, autonomous equations when finding the order.

Proof: Not in course. □

Example: Consider the general 3-stage ERK, with Butcher tableau:

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ c_2 & a_{21} & 0 & 0 \\ c_3 & a_{31} & a_{32} & 0 \\ \hline & b_1 & b_2 & b_3 \end{array}$$

WLOG, we can work with scalar, autonomous ODEs by the above. Let all quantities be evaluated at (t_n, y_n). Then:

$$\begin{aligned}k_1 = f,\ \ k_2 = f(y + ha_{21}, k_1) &= f + c_2 h f_y f + \frac{1}{2}h^2 c_2^2 f_{yy} f^2 + \cdots, \\ k_3 &= f(y + h(a_{31}k_1 + a_{32}k_2)) \\ &= f + hc_3 f_y f + h^2(c_2 a_{32} f_y^2 f + \frac{1}{2}c_3^2 f_{yy} f^2) + \cdots\end{aligned}$$

and so

$$\begin{aligned}y_{n+1} &= y + h(b_1 + b_2 + b_3)f + h^2(b_2c_2 + b_3c_3)ff_y \\ &+h^3\left(\frac{b_2c_2^2 + b_3c_3^2}{2}f_{yy}f^2 + b_3c_2a_{32}f_y^2 f\right) + O(h^4).\end{aligned}$$

Compare with the expansion of $y(t_n + h)$, using $f' = f_y f$, $f'' = f_y y f^2 + f_y^2 f$, etc, to get conditions:

$$\begin{aligned}b_1 + b_2 + b_3 = 1, &\qquad b_2c_2 + b_3c_3 = \frac{1}{2}, \\ b_2c_2^2 + b_3c_3^2 = \frac{1}{3}, &\quad b_3c_2a_{32} = \frac{1}{6}.\end{aligned}$$

Example: Consider the general 2-stage IRK, with Butcher tableau:

$$\begin{array}{c|cc} c_1 & a_{11} & a_{12} \\ c_2 & a_{21} & a_{22} \\ \hline & b_1 & b_2 \end{array}$$

It's possible to show that for this to be third order, we need $b_1 + b_2 = 1$, $b_1c_1 + b_2c_2 = \frac{1}{2}$, $\mathbf{b}^T A\mathbf{c} = \frac{1}{6}$ and $b_1c_1^2 + b_2c_2^2 = \frac{1}{3}$.

4.3 Error control of RK methods

RK methods have a specific error control device called *embedding*. We put the RK method into a larger one, with higher order:

$$\tilde{A} = \begin{pmatrix} A & \mathbf{0} \\ \mathbf{a}^T & \tilde{a} \end{pmatrix}, \qquad \tilde{\mathbf{c}} = \begin{pmatrix} \mathbf{c} \\ \tilde{c} \end{pmatrix}$$

The first stages are identical to the smaller method, and the additional stage can be used to control the error.

4.4 A review of interpolation and quadrature

Order analysis can be made easier by converting a Runge-Kutta method to a *collocation method*. These are tied to *numerical quadrature* and *polynomial interpolation*, so we will review these subjects first.

Polynomial interpolation

Definition: Let $c_1, ..., c_s$ be a set of interpolation points. Define $\omega(t) = \prod_{l=1}^{s}(t-c_l)$ and $\omega_k(t) = \omega(t)/(t-c_k)$. The *Lagrange cardinal polynomials* for this set of points are then:

$$L_k(t) = \frac{\omega_k(t)}{\omega_k(c_k)}, \quad k = 1, ..., s.$$

The L_k are each polynomials of degree $s-1$ such that $L_k(c_k) = 1$ and $L_j(c_k) = 0$ for $j \neq k$.

A degree $s-1$ polynomial interpolation to the function $f(t)$ using the points c_l is then clearly:

$$f(t) \approx \sum_{k=1}^{s} f(c_k)L_k(t).$$

Numerical quadrature

Definition: A *quadrature formula* with weight $w \geq 0$ on an interval (a, b) is an approximation of the form:

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

The quadrature formula is of *order* p if it is correct on all polynomials of degree $p-1$ (i.e. for all functions $g \in \mathbb{P}_{p-1}$).

It is possible to show that the zeroes of orthogonal polynomials optimise quadrature:

Theorem: Let $c_1, ..., c_s$ be the zeroes of the sth degree orthogonal polynomial p_s on interval (a, b) with weight $w \geq 0$, i.e.

$$\int_a^b q(\tau)p_s(\tau)w(\tau)d\tau = 0, \text{ for all } q \in \mathbb{P}_{s-1}.$$

Then the quadrature formula with these values of c_l, and values of b_l generated from the linear system:

$$\sum_{l=1}^{s} b_l c_l^j = \int_a^b \tau^j w(\tau)d\tau, \qquad (*)$$

for $j = 0, 1, ..., s-1$: (i) is exactly of order $2s$, and (ii) any other quadrature formula is of order $\leq 2s-1$.

Proof: (i) Let $v \in \mathbb{P}_{2s-1}$. Then $v = p_s q + \tilde{v}$ for some $q, \tilde{v} \in \mathbb{P}_{s-1}$ by polynomial division. The integral then beomces:

$$\int_a^b (p_s(\tau)q(\tau) + \tilde{v}(\tau))w(\tau)\, d\tau = \int_a^b \tilde{v}(\tau)w(\tau)\, d\tau.$$

Since c_l are zeroes of p_s, our quadrature formula gives:

$$\sum_{l=0}^{s} b_l \left(p_s(c_l)q(c_l) + \tilde{v}(c_l)\right) = \sum_{l=0}^{s} b_l \tilde{v}(c_l).$$

According to $(*)$ the b_l are such that the formula is exact for polynomials in $\mathbb{P}_{s-1}$. So the answers must agree.

To prove the formula is *exactly* of order $2s$, and no more, trial $g = p_s^2 \in \mathbb{P}_{2s}$. Then our integral is:

$$\int_a^b (p_s(\tau))^2 w(\tau)\, d\tau > 0,$$

but our formula is

$$\sum_{l=1}^{s} b_l (p_s(c_l))^2 = 0,$$

(ii) Assume we have a quadrature formula of order $2s$. Then the formula is exact on $v = p_s q \in \mathbb{P}_{2s-1}$, with $q = L_m$, for $m \in \{1, ..., s\}$. Then the quadrature formula gives $b_m p_s(c_m) = 0$; this must be zero for agreement with the integral of v.

Note $b_m \neq 0$, else the quadrature formula omitting the point c_m gives a formula of order $2s$ by assumption, which is greater than is possible (we'd get an overdetermined b_l system). So c_m must be roots of p_s for order $2s$. □

From the proof above, we see that:

Theorem: Quadrature is of exactly order $r+s$, for $r \in \{0, 1, ..., s\}$ iff $b_1, ..., b_s$ are chosen as in the above Theorem, i.e.

$$\sum_{l=1}^{s} b_l c_l^j = \int_a^b \tau^j w(\tau)d\tau, \quad j = 0, 1, ..., s-1, \qquad (*)$$

but instead the c_l are such that

$$\int_a^b \tau^j \omega(\tau)w(\tau)\, d\tau = 0, \quad j = 0, ..., r-1,$$

$$\int_a^b \tau^r \omega(\tau)w(\tau)d\tau \neq 0,$$

where $\omega(t) = \prod_{k=1}^{s}(t-c_k)$. (I.e. *some* of the c_k are roots of an orthogonal polynomial, but the rest are random.)

In particular, the highest order quadrature formula on $(0,1)$, for $w \equiv 1$, occurs when c_l are roots of the shifted Legendre polynomial P_s (shifted from $[-1,1]$ to $[0,1]$).

4.5 Collocation methods and order

Method: The *collocation method* is defined as follows. Let $c_1, ..., c_s \in [0,1]$ be distinct interpolation points. Let $\mathbf{y}_n$ be the numerical solution at step n. Using polynomial interpolation, find a vector of s-degree polynomials $\mathbf{u}$ obeying:

$$\mathbf{u}(t_n) = \mathbf{y}_n, \qquad \mathbf{u}'(t_n + c_l h) = \mathbf{f}(t_n + c_l h, \mathbf{u}(t_n + c_l h)).$$

This is possible since s-degree polynomials have $s+1$ free parameters, and there are $s+1$ conditions here. Then let:

$$\mathbf{y}_{n+1} = \mathbf{u}(t_n + h).$$

The collocation method can be re-formulated as a Runge-Kutta method via:

Theorem: The collocation method is identical to the s-stage Runge-Kutta method:

$$a_{k,l} = \int_0^{c_k} L_l(\tau) d\tau, \qquad b_l = \int_0^1 L_l(\tau) d\tau,$$

where the $\mathbf{c}$ vector in the Runge-Kutta method is just the vector of collocation points.

Proof: The polynomial $\mathbf{u}'$ coincides with its $(s-1)$st degree Lagrange interpolation polynomial. So we must have:

$$\mathbf{u}'(t) = \sum_{j=1}^{s} L_j\left(\frac{t-t_n}{h}\right)\mathbf{u}'(t_n + c_j h)$$
$$= \sum_{j=1}^{s} \frac{\omega((t-t_n)/h)}{\omega_j(c_j)}\mathbf{u}'(t_n + c_j h)$$

(note we have shifted the points at which we interpolate, but retained the L_j corresponding to c_j; this amounts to a shift in the argument of L_j). By the definition of $\mathbf{u}$, we then have:

$$\mathbf{u}'(t) = \sum_{j=1}^{s} \frac{\omega((t-t_n)/h)}{\omega_j(c_j)}\mathbf{f}\left(t_n + c_j h, \mathbf{u}(t_n + c_j h)\right).$$

Integration then yields (after a change of variables):

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

Define $\mathbf{k}_j = \mathbf{f}(t_n + c_j h, \mathbf{u}(t_n + c_j h))$. The formula for $\mathbf{u}(t)$ then gives:

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

as per the definition of $a_{l,j}$ in the Theorem, and hence

$$\mathbf{k}_j = \mathbf{f}\left(t_n + c_j h, \mathbf{y}_n + h\sum_{j=1}^{s} a_{l,j}\mathbf{k}_j\right).$$

This is most of Runge-Kutta; finally need

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

Collocation methods have an easy order analysis, so it's easier to find the order of some Runge-Kutta methods by reframing them as collocation methods. The collocation methods' order comes straight from numerical quadrature, in two parts:

(1) Lemma (Alekseev-Gröbner): Let $\mathbf{u}$ be a smooth function such that $\mathbf{u}(t_0) = \mathbf{y}(t_0)$, where $\mathbf{y}$ solves $\mathbf{y}' = \mathbf{f}(t, \mathbf{y})$. Then

$$\mathbf{u}(t) - \mathbf{y}(t) = \int_{t_0}^{t} \Phi\left(t, \tau, \mathbf{u}(\tau)\right)\left(\mathbf{f}(\tau, \mathbf{u}(\tau)) - \mathbf{u}'(\tau)\right) d\tau,$$

where Φ is a matrix obeying the ODE $\frac{d\Phi}{dt} = \frac{\partial \mathbf{f}}{\partial \mathbf{y}}\Phi(\tau) = I$.

Proof: Not in course. □

(2) Theorem: The order of the collocation method with collocation points $c_1, ..., c_s$ is identical to the order of the quadrature formula on $(0,1)$ with weight 1 at the interpolation points $c_1, ..., c_s$.

Proof: We have:

$$\mathbf{y}_{n+1} - \tilde{\mathbf{y}}(t_{n+1}) = \mathbf{u}(t_{n+1}) - \tilde{\mathbf{y}}(t_{n+1})$$
$$= \int_{t_n}^{t_{n+1}} \Phi(t_{n+1}, \tau, \mathbf{u}(\tau))\left(\mathbf{f}(\tau, \mathbf{u}(\tau)) - \mathbf{u}'(\tau)\right) d\tau,$$

by the Alekseev-Gröbner Lemma. Use the quadrature formula with points $t_n + c_1 h$, $t_n + c_2 h$, ... , $t_n + c_s h$, and weight $w \equiv 1$. Then $\mathbf{y}_{n+1} - \tilde{\mathbf{y}}(t_{n+1})$

$$= \sum_{l=1}^{s} b_l \Phi(t_{n+1}, t_n + c_l h, \mathbf{u}(t_n + c_l h))\left(\mathbf{d}(t_n + c_l h)\right) + \text{error},$$

where $\mathbf{d}(t) = \mathbf{f}(t, \mathbf{u}(t)) - \mathbf{u}'(t)$. But by definition of the function $\mathbf{u}$, we have $\mathbf{d}(t_n + c_l h) = 0$.

Thus local error of collocation method is the same as the local error of the quadrature on $[t_n, t_{n+1}]$ with points $t_n + c_l h$ and weight $w \equiv 1$.

We now need to relate this to orders. We are attempting to approximate:

$$\int_{t_n}^{t_{n+1}} g(\tau)\ d\tau = \int_0^h g(\tau)\ d\tau,$$

where we've just shifted limits and redefined g. Expand g in a Taylor series about 0:

$$\int_0^h \left(g(0) + \tau g'(0) + ... + \frac{\tau^{p-1}}{(p-1)!} g^{(p)}(0) \right)\ d\tau + O(h^{p+1}).$$

Assuming the quadrature formula is of order p, it is exact on polynomials of degree $p-1$, so we see the error committed is $O(h^{p+1})$. Thus the quadrature formula's order is the same as the numerical method's order.

We've worked with quadrature on $[t_n, t_{n+1}]$ and with points $t_n + c_l h$. Translating and rescaling, we get the result on $[0,1]$ with points c_l as required. □

Corollary: The highest order s-stage Runge-Kutta method coming from collocation takes collocation points at shifted Legendre points. This is called the *Gauss-Legendre Runge-Kutta method*, and is of order $2s$.

Example: Consider the Runge-Kutta method:

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

This has $c_1 = 0$, $c_2 = \frac{1}{2}$, $c_3 = 1$, so the suggested collocation polynomial $\omega(\tau) = \tau(\tau - \frac{1}{2})(\tau - 1)$. We check this collocation gives rise to RK form:

$$b_1 = 2\int_0^1 \left(\tau - \frac{1}{2}\right)(\tau - 1)\ d\tau = 1/6,$$

and similarly for b_1, b_2, a_{11}, a_{21}, a_{22}, a_{23}, $a_{3l} = b_l$. Using the test with the integral

$$\int_0^1 \tau^j \omega(\tau)\ d\tau,$$

we see that the method is of order 4 (since the integral is zero for $j = 1$ and non-zero for $j = 1$).

4.6 A-stability of Runge-Kutta methods

Theorem: Define:

$$R(\lambda) = \mathbf{b}^T (I - \lambda A)^{-1}(I - \lambda(A - I))\mathbf{1}.$$

An RK method is A-stable iff $|R(\lambda)| < 1$ for all $\lambda \in \mathbb{C}^-$.

Proof: Apply the RK method to $y' = \lambda y$, $y(0) = 1$ and $h = 1$. We see that the vector $\mathbf{k} = (k_1, k_2, ..., k_n)$ obeys:

$$\mathbf{k} = \lambda(\mathbf{1} + A\mathbf{k}),$$

and so $\mathbf{k} = \lambda(I - \lambda A)^{-1}\mathbf{1}$. It follows that $y_{n+1} = y_n + \mathbf{b}^T\mathbf{k} = R(\lambda)y_n$ (using sum of b_l is 1 for order greater than 1) and then condition follows. □

In particular, we note that $R(z)$ is a rational function in $\mathbb{P}_{s/s}$, since we can write the inverse of a matrix as:

$$(I - \lambda A)^{-1} = \frac{\operatorname{adj}(I - \lambda A)}{\det(I - \lambda A)}.$$

This immediately gives:

Theorem: The Gauss-Legendre RK is A-stable.

Proof: $R \in \mathbb{P}_{s/s}$ and approximates e^z to order $2s$ for Gauss-Legendre. Hence it must be the (unique) $[s/s]$ Padé method, which is A-stable by the Wanner-Hairer-Nørsett Theorem. □

Example 1: For an ERK, we have:

$$\det(I - \lambda A) = \det \begin{pmatrix} 1 & \cdots & 0 \\ * & \ddots & 0 \\ * & \cdots & 1 \end{pmatrix} = 1.$$

and hence $R(\lambda) \in \mathbb{P}_s$ (just a polynomial). So it can at most approximate e^z to order s: $p \leq s$.

Example 2: Consider a s-stage RK method coming from collocation polynomial:

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

where $\tilde{P}_s$ is the sth Legendre polynomial shifted to $[0,1]$. Clearly this method is at least of order $s + (s-1) = 2s - 1$.

We choose α, β such that A is invertible and $\mathbf{b}^T A^{-1}\mathbf{1} = 1$. From here, we can deduce A-stability. Write $R(\lambda)$ as:

$$R(\lambda) = 1 + \lambda\mathbf{b}^T(I - \lambda A)^{-1}\mathbf{1} = 1 + \mathbf{b}^T\left(\frac{1}{\lambda} - A\right)^{-1}\mathbf{1}.$$

Let $\lambda \to \infty$. Then $R(\infty) = 0$ by our choice of α, β. So $R(\lambda)$ must be of the form $\mathbb{P}_{s-1}/\mathbb{P}_s$. So we know from the order, it is the Padé approximation, and hence we can deduce A-stability (by Wanner-Hairer-Nørsett).

4.7 Non-linear stability of RK methods

A-stability is a property of *linear* equations (we apply the numerical method to $y' = \lambda y$) and hence is rather restrictive. Sometimes, we want to say something about non-linear stability.

One instance when we can say something about stability is when the solution of an equation is *dissipative*:

Definition: The solution of $\mathbf{y}' = \mathbf{f}(\mathbf{y})$ is called *dissipative* if $||\mathbf{u}(t) - \mathbf{v}(t)||$ is monotonically non-increasing for any two solutions $\mathbf{u}(t)$ and $\mathbf{v}(t)$, $t \geq 0$.

In practice, it is more convenient to work with a condition leading to dissipative behaviour:

Theorem: Suppose the function $\mathbf{f}$ obeys $\langle \mathbf{u} - \mathbf{v}, \mathbf{f}(\mathbf{u}) - \mathbf{f}(\mathbf{v}) \rangle \leq 0$. Then the solution of $\mathbf{y}' = \mathbf{f}(\mathbf{y})$ is dissipative.

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

$$\frac{1}{2}\phi'(t) = \langle \mathbf{u} - \mathbf{v}, \mathbf{u}' - \mathbf{v}' \rangle = \langle \mathbf{u} - \mathbf{v}, \mathbf{f}(\mathbf{u}) - \mathbf{f}(\mathbf{v}) \rangle \leq 0. \quad \square$$

Definition: A method is *algebraically stable* if it preserves *monotonicity* of a non-linear equation; that is, if the equation $\mathbf{y}' = \mathbf{f}(\mathbf{y})$ obeys

$$\langle \mathbf{f}(\mathbf{y}) - \mathbf{f}(\mathbf{z}), \mathbf{y} - \mathbf{z} \rangle < 0,$$

for any two solutions $\mathbf{y}$, $\mathbf{z}$, then the corresponding numerical solutions obey $||\mathbf{y}_{n+1} - \mathbf{z}_{n+1}|| \leq ||\mathbf{y}_n - \mathbf{z}_n||$.

Butcher's Theorem: An RK method is algebraically stable if and only if $\mathbf{b} \geq \mathbf{0}$ and M is positive semi-definite, where M is the $s \times s$ matrix with entries

$$M_{ij} = b_i a_{ij} + b_j a_{ji} - b_i b_j.$$

Proof: Let $\mathbf{k}_1, ..., \mathbf{k}_s$ be the stages for a numerical solution $\mathbf{u}_n$ and $\mathbf{l}_1, ..., \mathbf{l}_s$ for a numerical solution $\mathbf{v}_n$. Then:

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

Thus for algebraic stability we require

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

For $j = 1, ..., s$, set $\mathbf{d}_j = \mathbf{k}_j - \mathbf{l}_j$,

$$\mathbf{p}_j = \mathbf{u}_n + h\sum_{i=1}^{s} a_{ji}\mathbf{k}_i, \qquad \mathbf{q}_j = \mathbf{v}_n + h\sum_{i=1}^{s} a_{ji}\mathbf{l}_i.$$

Then $\mathbf{k}_j = \mathbf{f}(\mathbf{p}_j)$ and $\mathbf{l}_j = \mathbf{f}(\mathbf{q}_j)$. We now use this notation to bound the first term in $(*)$. We have that: $\left\langle \mathbf{u}_n - \mathbf{v}_n, \sum_j b_j \mathbf{d}_j \right\rangle =$

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

by dissipative behaviour of $\mathbf{f}$. Thus we can bound $(*)$ by:

$$\sum_{i,j} \mathbf{d}_j^T \left(b_i b_j - b_j a_{ji} - b_i a_{ji} \right) \mathbf{d}_i = -\sum_{i,j} \mathbf{d}_i^T M_{ij} \mathbf{d}_j.$$

Let D be the matrix with columns $\mathbf{d}_1$, $\mathbf{d}_2$, ... , $\mathbf{d}_s$, and let $\boldsymbol{\delta}_1^T$, ... , $\boldsymbol{\delta}_d^T$ be the rows of D. Then $\sum_{i,j} \mathbf{d}_i^T M_{ij} \mathbf{d}_j =$

$$\sum_{i,j,k} D_{ik} M_{ij} D_{jk} = \sum_k \sum_{i,j} D_{ik} M_{ij} D_{jk} = \sum_k \boldsymbol{\delta}_k^T M \boldsymbol{\delta}_k \geq 0. \quad \square$$

4.8 $R^{[1]}$ and $R^{[2]}$ properties of RK methods

As we discussed earlier, it is important for a method to be $R^{[1]}$ and $R^{[2]}$. Here are further examples of analysis in the case of RK methods.

Example 1: The RK method

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

is *not* $R^{[1]}$. It is sufficient to demonstrate this for one equation, say $y' = \kappa y(1 - y)$. Applying the method, we find that the numerical method gives us additional fixed points, namely $y = 2/\kappa h$ and $y = (2 + h\kappa)/h\kappa$.

Example 2: The *implicit midpoint rule*

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

is actually a Runge-Kutta method, as can be seen by setting $\mathbf{k} = \mathbf{f}\left(t_n + \frac{1}{2}h, \frac{1}{2}(\mathbf{y}_n + \mathbf{y}_{n+1}\right)$. This reduces the method to the RK method:

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

Since the first shifted Legendre polynomial on $[0, 1]$ is $P_1(x) = 2x - 1$, this is a Gauss-Legendre RK, so is A-stable. It's straightforward to see the method is also $R^{[1]}$ and $R^{[2]}$.

5 Synthesis of FDMs

5.1 Finite difference calculus

Definition: Given a sequence y_n, we define the operators:

$$\begin{aligned} Ey_n &= y_{n+1} && \textit{- shift operator;}\\ \Delta_+ y_n &= y_{n+1} - y_n && \textit{- forward difference operator;}\\ \Delta_- y_n &= y_n - y_{n-1} && \textit{- backward difference operator;}\\ \Delta_0 y_n &= y_{n+\frac{1}{2}} - y_{n-\frac{1}{2}} && \textit{- central difference operator;}\\ \mu_0 y_n &= \frac{1}{2}\left(y_{n+\frac{1}{2}} + y_{n-\frac{1}{2}}\right) && \textit{- averaging operator.} \end{aligned}$$

Note that Δ_0 and μ_0 are not defined on a grid, but even combinations of the two are, e.g. Δ_0^2 and $\Delta_0\mu_0$. Given that $y_n = y(nh)$ for y analytic in $\mathbb{R}$, we define the *differential operator* by $Dy_n = y'(nh)$.

Theorem: All of the above operators commute.

Proof: All operators can be expressed in terms of one another, e.g.

$$E = e^{hD}, \qquad \mu_0 = \frac{1}{2}(E^{1/2} - E^{-1/2}),$$
$$E = 2\mu_0^2 - I + 2\mu_0\sqrt{\mu_0^2 - I},$$

etc, so must trivially commute. □

5.2 Approximation of D^s

When solving PDEs, we want to approximate D^s for some power s.

Theorem: We can approximate D^s as:

(i) In terms of forward differences,

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

(ii) In terms of backward differences,

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

(iii) In terms of central differences,

$$D = \frac{4}{h}\sum_{j=0}^{\infty}\frac{(-1)^j}{2j+1}\binom{2j}{j}\left(\frac{1}{4}\Delta_0\right)^{2j+1},$$

then raising both sides to the sth power.

Proof: (i) and (ii) following immediately by using $D = \log(E)/h = \log(I + \Delta_+)/h = \log(I - \Delta_-)/h$ and expanding the logarithm.

(iii) is more difficult. We notice that $\Delta_0^2 y_n = y_{n+1} - 2y_n + y_{n-1}$, and so:

$$D = \frac{2}{h}\log\left(\frac{1}{2}\Delta_0 + \sqrt{I + \frac{1}{4}\Delta_0^2}\right).$$

Let $g(z) = \log(z + \sqrt{1+z^2})$ so that $D = \frac{2}{h}g(\frac{1}{2}\Delta_0)$. Note:

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

Integrate and use the boundary condition $g(0) = 0$ to find the result. □

Using these series expansions, we can make approximations. The order of the approximation can be determined using the fact that $\Delta_+ = E - I = O(h)$, $\Delta_- = I - E^{-1} = O(h)$ and $\Delta_0^2 = e^{hD} - 2I + e^{-hD} = O(h^2) \Rightarrow \Delta_0 = O(h)$.

For example, the truncations in (i) and (ii) both commit an error of order $O(h^3)$.

(iii) can be used immediately to find D^s for s even, since we'll just get even powers of Δ_0 in the answer. E.g. a simple truncation for D^2 is:

$$D^2 y_n \approx \frac{1}{h^2}\left(\Delta_0^2 - \frac{1}{12}\Delta_0^4\right)y_n.$$

This commits an error of order $O(h^4)$.

For odd s, we multiply our approximation to D^s by the identity, written as:

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

Example: An order $O(h^4)$ approximation to D obtained from central differences is:

$$Dy_n \approx \frac{1}{h}\left(\frac{1}{12}y_{n-2} - \frac{2}{3}y_{n-1} + \frac{2}{3}y_{n+1} - \frac{1}{12}y_{n+2}\right).$$

Note that central differences typically provide much better order because they are more symmetric about the point we are trying to approximate.

This does NOT mean they are useful in all cases though. Sometimes the PDE has an inherent asymmetry, e.g. the advection equation $u_t = u_x$, $u(x,0) = \phi(x)$ has solution $u(x,t) = \phi(x+t)$, i.e. the solution propagates from left to right. So it's much better to use *slanted* rather than *central* differences in this case.

5.3 Computational stencil calculus

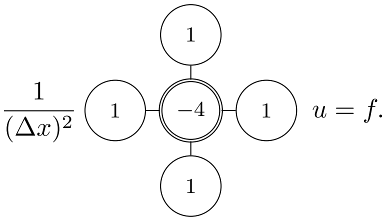
Consider solving the *Poisson equation* $\nabla^2 u = f$. We let $u_{k,l} \approx u(k\Delta x, l\Delta x)$, i.e. we work on a square grid of spacing Δx. We wish to approximate $\nabla^2 = D_x^2 + D_y^2$. The first term in the central difference approximation to D^2 is just $\frac{1}{(\Delta x)^2}\Delta_0^2$, so let's just use central differences:

Definition: The *five-point formula* is the numerical method:

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

This is cumbersome to write, so we introduce:

Notation: We represent numerical finite difference operators using the self-explanatory *computational stencil* notation. The five-point formula in this notation is:



The double circle denotes the central point of the approximation.

Stencils can be added and multiplied, corresponding to addition and composition of the underlying finite difference operators. The rules are as follows:

Stencil addition: Overlay the stencils at the centre, and add things in corresponding circles.

Stencil multiplication: Put the centre of the second stencil on each point of the first. Perform multiplication at each point. If we get contributions from two iterations of this procedure, sum the contributions.

5.4 Semi & full-discretisation

Definition: *Full-discretisation* (FD) discretises both time and space in unison, whereas *semi-discretisation* discretises only space, leaving an ODE system.

Example: Consider $u_t = u_{xx}$. Discretising the spatial derivative with a central difference gives the SD method:

$$u_m' = \frac{1}{(\Delta x)^2}(u_{m-1} - 2u_m + u_{m+1}).$$

We can then discretise the time derivative to get an FD method. Using a forward difference gives *Euler's method*, and using the trapezoidal rule to solve the ODE system above gives the *Crank-Nicolson method*.

6 Order analysis of FDMs

In previous section, concentrated on building FDMs. Assuming we have one, how do we find its order?

6.1 Order analysis of boundary value probs

Theorem: The numerical method for the Poisson equation given by:

$$(\Delta x)^{-2}\sum_{(i,j)\in I} a_{i,j}u_{k+i,l+j} = f_{k,l},$$

where I is the set of stencil points, is of order p iff

$$\sum_{(i,j)\in I} a_{i,j}x^i y^j - \log^2(x) - \log^2(y) = O((\Delta x)^{p+3}),$$

where $x, y = 1 + O(\Delta x)$.

Proof: Define the finite difference operator:

$$\mathcal{L}_{\Delta x} = L_{\Delta x}(E_x, E_y), \qquad L_{\Delta x}(x,y) = \sum_{(i,j)\in I} a_{i,j}x^i y^j.$$

so that the method is $(\Delta x)^{-2}\mathcal{L}_{\Delta x}u_{k,l} = f_k, l$.

The exact solution obeys $\nabla^2 u = (D_x^2 + D_y^2)u = f$, which is equivalent to $(\Delta x)^{-2}L(E_x,E_y)u = f$, where

$$L(x,y) = \log^2(x) + \log^2(y).$$

The condition in the Theorem then says:

$$L_{\Delta x}(E_x,E_y) - L(E_x,E_y) = O((\Delta x)^{p+3}).$$

Let $\tilde{u}_{k,l} = u(k\Delta x, l\Delta x)$ (i.e. sample the exact solution at grid points). Then inserting this into the numerical method we find:

$$\begin{aligned}(\Delta x)^{-2}\mathcal{L}_{\Delta x}\tilde{u}_{k,l} - f_{k,l} &= \underbrace{\nabla^2\tilde{u}_{k,l} - f_{k,l}}_{0,\text{ since exact}} + O((\Delta x)^{p+1}) \\ &= O((\Delta x)^{p+1}).\end{aligned}$$

Subtract the numerical solution $(\Delta x)^{-2}\mathcal{L}_{\Delta x}u_{k,l} - f_{kl} = 0$:

$$(\Delta x)^{-2}\mathcal{L}_{\Delta x}(\tilde{u} - u) = O((\Delta x)^{p+1}).$$

In simple geometries, with nice boundary conditions, we can invert the operator on the LHS. It is of order:

$$\begin{aligned}(\Delta x)^{-2}\mathcal{L}_{\Delta x} &= (\Delta x)^{-2}L_{\Delta x}(E_x,E_y) \\ &= (\Delta x)^{-2}L(E_x,E_y) + O((\Delta x)^{p+1}) \\ &= 2(\Delta x)^{-2}\left(\log^2(1+O(\Delta x))\right) + O((\Delta x)^{p+1}) \\ &= O(1).\end{aligned}$$

where we've used $E_x, E_y = 1 + O(\Delta x)$, so when we invert we still get:

$$\tilde{u} - u = O((\Delta x)^{p+1}). \qquad \square$$

Example 1: Recall the 5-point formula:

$$\frac{1}{(\Delta x)^2}\begin{array}{ccc} & 1 & \\ 1 & -4 & 1 \\ & 1 & \end{array}\; u = f.$$

In practice, it's convenient to write $x = e^{i\theta}$ and $y = e^{i\psi}$, so that $\theta, \psi = O(\Delta x)$ in the expansion. Then we need to consider:

$$-4 + 2\cos(\theta) + 2\cos(\psi) + \theta^2 + \psi^2 = O(\theta^4, \psi^4).$$

Note we've used the useful trick of turning everything into trig functions. Hence the method is of order 1.

Example 2: Consider the 9-point formula:

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

We need to consider:

$$-\frac{10}{3}+\frac{4}{3}\cos(\theta)+\frac{4}{3}\cos(\psi)+\frac{1}{3}\cos(\theta+\psi)+\frac{1}{3}\cos(\theta-\psi)+\theta^2+\psi^2.$$

A short calculation shows this is $O(\theta^4, \psi^4)$, and again it's of order 1 (so the error is of order $O((\Delta x)^2)$). Indeed the error constant is given by:

$$L_{\Delta x}(e^{i\theta}, e^{i\psi}) - L(e^{i\theta}, e^{i\psi}) = \frac{1}{12}(\theta^2+\psi^2)^2 + O((\Delta x)^6),$$

which shows that we're actually solving

$$\left(1 + \frac{1}{12}(\Delta x)^2\nabla^2\right)\nabla^2 u = f,$$

up to error of order $O((\Delta x)^4)$. In particular, when $f \equiv 0$, we can 'invert' the first operator, and instead be solving $\nabla^2 u = 0$, to order $O((\Delta x)^4)$!

So the nine-point formula has order 1 when solving Poisson's equation, but 3 when solving Laplace's equation!

6.2 The Mehrstellenverfahren

This device extends the benefit of the 9-point formula (and other numerical methods) to Poisson's equation.

Method (Mehrstellenverfahren): Let $\mathcal{M}_{\Delta x}$ be a finite difference approximation to the operator

$$I + \frac{1}{12}(\Delta x)^2\nabla^2,$$

with error $O((\Delta x)^4)$. That is, suppose that

$$\mathcal{M}_{\Delta x} - I - \frac{1}{12}(\Delta x)^2\nabla^2 = O((\Delta x)^4).$$

Then if $\mathcal{L}_{\Delta x}$ is the nine-point formula stencil, we solve numerically:

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

That is, we are solving the equation

$$\left(I + \frac{1}{12}(\Delta x)^2\nabla^2\right)\nabla^2 u = \left(I + \frac{1}{12}(\Delta x)^2\nabla^2\right)f,$$

up to error of order $O((\Delta x)^2)$. Now when we invert the operator, we're left with Poisson's equation!

To construct $\mathcal{M}_{\Delta x}$, we need to use the synthesis methods from above. We need only approximate ∇^2 to order $O(1)$, since the $O((\Delta x)^2)$ terms will just fall out. We have:

$$I + \frac{1}{12}(\Delta x)^2\nabla^2 \approx$$
$$I + \frac{1}{12}(\Delta x)^2\left(\frac{1}{(\Delta x)^2}(\Delta_{0,x}^2 + \Delta_{0,y}^2)\right) + O((\Delta x)^4).$$

Applying this to some $f_{k,l}$ (or noticing this is just 1 added to $1/12$ of the five-point formula stencil), we see that we should use the stencil:

$$\mathcal{M}_{\Delta x} = \begin{array}{ccc} & \frac{1}{12} & \\ \frac{1}{12} & \frac{2}{3} & \frac{1}{12} \\ & \frac{1}{12} & \end{array}\; .$$

The results extend to d-dimensions, in which we have:

Theorem: Let $L_{\Delta x}(E_{x_1}, ..., E_{x_d})$ be the finite difference operator approximating ∇^2 within error $O((\Delta x)^2)$:

$$L_{\Delta x}(\mathbf{x}) = -\frac{2}{3}(2d+1)+\frac{2}{3}\sum_{k=1}^{d}\left(x_k+\frac{1}{x_k}\right)+\frac{2}{3}\cdot\frac{1}{2^d}\prod_{k=1}^{d}\left(x_k+\frac{1}{x_k}\right).$$

If the Mehrstellenverfahren

$$M_{\Delta x}(\mathbf{x}) = 1 - \frac{1}{6}d + \frac{1}{12}\sum_{k=1}^{d}\left(x_k + \frac{1}{x_k}\right),$$

is used, then the numerical scheme $L_{\Delta x}u_{\mathbf{k}} = (\Delta x)^2 M_{\Delta x}f_{\mathbf{k}}$ approximates the solution of $\nabla^2 u = f$ to $O((\Delta x)^4)$.

Proof: Let $L(E_{x_1}, ..., E_{x_d}) = (\Delta x)^2\nabla^2$ in d-dimensions. Proof just consists of showing $L_{\Delta x}(e^{i\theta_1}, ..., e^{i\theta_d}) = L + \frac{1}{12}L^2 + O((\Delta x)^6)$ and $M_{\Delta x}(e^{i\theta_1}, ..., e^{i\theta_d}) = 1 + \frac{1}{12}L + O((\Delta x)^4)$. □

6.3 Order analysis of equations of evolution

Definition: A *PDE of evolution* is an equation of the form

$$u_t = \frac{\partial^L u}{\partial x^L}.$$

Numerically, we solve with a time step Δt and a spatial step Δx. Define the *Courant number* as $\mu = \Delta t/(\Delta x)^L$.

A general semi-discretisation of a PDE of evolution is:

$$u_m' - \frac{1}{(\Delta x)^L}\sum_{k=-r}^{s} \alpha_k u_{m+k} = 0.$$

We obtain this by approximating $\partial^L/\partial x^L$. We can calculate its order using:

Definition: The *symbol* of the method is the Laurent polynomial:

$$h(z) = \sum_{k=-r}^{s} \alpha_k z^k.$$

Theorem: The method is of order p iff

$$h(z) - \log^L(z) = O(|z-1|^{p+1}),$$

and the Courant number is constant.

Proof: Let $\tilde{u}$ be the exact solution, i.e. $D_t\tilde{u} = D_x^L\tilde{u}$. Apply the numerical method to $\tilde{u}$:

$$\begin{aligned}\tilde{u}_m' - \frac{1}{(\Delta x)^L}\sum_{k=-r}^{s}\alpha_k\tilde{u}_{m+k} &= \left(D_t - \frac{1}{(\Delta x)^L}\sum_{k=-r}^{s}\alpha_k E_x^k\right)\tilde{u}_m \\ &= \frac{1}{(\Delta x)^L}\left(\log^L(E_x) - \sum_{k=-r}^{s}\alpha_k E_x^k\right)\tilde{u}_m \\ &= \frac{1}{(\Delta x)^L}O(|E_x-1|^{p+1})\tilde{u}_m,\end{aligned}$$

by condition in Theorem. Now $E_x = 1 + O(\Delta x)$, so we have that

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

Subtract the numerical method from this to get:

$$e_m' - \frac{1}{(\Delta x)^L}\sum_{k=-r}^{s}\alpha_k e_{m+k} = O((\Delta x)^{p+1-L}),$$

where $e_m = \tilde{u}_m - u_m$.

Now when we solve these equations numerically, we'll find $\partial/\partial t = O(1/\Delta t)$. So multiplying up and inverting, we see $e_m = O((\Delta x)^{p+1})$, and thus the method is of order p iff $\mu = \Delta t/(\Delta x)^L$ is constant. □

The method applies equally well to general full discretisations:

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

where $\gamma_k = \gamma_k(\mu)$, $\delta_k = \delta_k(\mu)$ and $\sum_{k=-r}^{s}\gamma_k(0) \neq 0$.

Definition: The *symbol* of a full-discretisation is:

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

Theorem: The full-discretisation is of order p iff

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

Proof: Proceeds just as in semi-discretisation case. □

Example 1: Consider the advection equation $u_t = u_x$ ($L = 1$) solved by

$$u_m' = \frac{1}{2\Delta x}(u_{m+1} - u_{m-1}).$$

The symbol is $h(z) = \frac{1}{2}(z - 1/z)$. Let $z = e^{i\theta}$ and consider $\theta \to 0$, as usual. Then

$$h(e^{i\theta}) - i\theta = i\sin(\theta) - i\theta = O(\theta^3).$$

So method is of order 2.

Example 2: Consider the full-discretisation of the advection equation:

$$u_m^{n+1} = u_m^n + \frac{\mu}{2}\left(u_{m+1}^n - u_{m-1}^n\right).$$

The symbol is

$$H(z;\mu) = 1 + \frac{\mu}{2}\left(z - \frac{1}{z}\right).$$

Set $z = i\theta$. Then:

$$H(e^{i\theta};\mu) = 1 + i\mu\sin(\theta) - e^{\mu i\theta} = O(\theta^2).$$

So method is of order 1.

Why did the order reduce? We used *central difference* in the semi-discretisation, which is of order 2 in ODE theory. But then we used forward Euler to approximate the time derivative, which is of order 1 in ODE theory. We need the Courant number to be constant: $\mu = \Delta t/\Delta x$, so for advection we need to use methods of the *same* order.

This is not true for the diffusion equation, $L = 2$, where $\mu = \Delta t/(\Delta x)^2$. Here, we can afford to use a smaller spatial order.

7 Convergence of FDMs

7.1 Well-posedness

Consider the PDE of evolution $u_t = \mathcal{L}u$, where $\mathcal{L}$ is linear and has zero BCs. The solution is

$$u = \mathcal{E}(t)u_0,$$

where $\mathcal{E}$ is the *evolution operator* and $u_0 = u(x,0)$ is the initial condition.

Definition: The equation is *well-posed* if for all times $T > 0$, there exists a constant C_T such that $||\mathcal{E}(t)||_{\text{op}} \leq C_T$ for all $t \in [0,T]$.

Example 1: The advection equation $u_x = u_t$ with initial data specified on the real line, $u(x,0) = u_0(x)$, is well-posed, since the solution is $u(x,t) = u(x+t,0) = u_0(x+t)$, and hence

$$||u(\cdot,t)||_2^2 = \int_{-\infty}^{\infty} |u_0(x+t)|^2 \, dx = \int_{-\infty}^{\infty} |u_0(x)|^2 = ||u_0||_2^2. \;(*)$$

It follows $||\mathcal{E}||_{\text{op}} = 1$. (Note if we were on an interval $[0,1]$, we wouldn't be able to use $(*)$. Instead, we get $||u(\cdot,t)|| \leq ||u_0||$, i.e. the equation is dissipative!)

Example 2: The diffusion equation $u_t = u_{xx}$ with zero BCs on $[-\pi,\pi]$ is well-posed, since for initial conditions $u(x,0) = \sum_{m=-\infty}^{\infty} \alpha_m e^{imx}$, we have by separation of variables:

$$u(x,t) = \sum_{m=-\infty}^{\infty} \alpha_m e^{imx-m^2t}.$$

Therefore,

$$||u(x,t)|| = \sqrt{\sum_{m=-\infty}^{\infty} |\alpha_m|^2 e^{-2m^2t}} \leq ||u(x,0)||,$$

and so $||\mathcal{E}||_{\text{op}} \leq 1$.

Note that if $t \mapsto -t$, we get the *reversed diffusion equation* $u_t = -u_{xx}$. This is not well-posed, since given any $[0,T]$ and C_T, we can always find an initial condition $\alpha_m = 0$ for all $m \neq \tilde{m}$, so that

$$u(x,t) = e^{i\tilde{m}+\tilde{m}^2t^2}.$$

Choosing $\tilde{m}$ large enough, we can make $||u||$ arbitrarily large.

7.2 Generality

We've consistently worked with zero BCs and homogeneous equations. What if we relaxed these conditions?

INHOMOGENEITY: Suppose $u_t = \mathcal{L}u + f$ with zero BCs, and assume it is well-posed with $||f|| \leq c$. We can obtain the solution to this problem from the homogeneous problem $v_t = \mathcal{L}v$ using *Duhamel's formula:*

Theorem (Duhamel): The solution to $u_t = \mathcal{L}u + f$ is given by:

$$u(x,t) = \mathcal{E}(t)u_0 + \int_0^t \mathcal{E}(t-\tau)f(\cdot,\tau)d\tau,$$

where $v(t) = \mathcal{E}(t)v(0)$ solves $v_t = \mathcal{L}v$.

NON-ZERO BCS: We can always just *subtract* BCs. Let w be an arbitrary function satisfying the BCs. Define $v = u - w$; this has zero BCs. Then

$$v_t = u_t - w_t = \mathcal{L}u + f - w_t = \mathcal{L}v + (\mathcal{L}w - w_t + f).$$

This is an inhomogeneous problem with zero BCs, so applying Duhamel, we reduce to the zero BCs, homogeneous case.

So we can always assume homogeneous and zero BCs WLOG.

7.3 Convergence, stability and Lax-equiv.

In general, a full-discretised system for a PDE of evolution can be written as:

$$\mathbf{u}_{\Delta x}^{n+1} = \mathcal{A}_{\Delta x}\mathbf{u}_{\Delta x}^n + \tilde{\mathbf{f}}_{\Delta x}^n.$$

Here, $\mathbf{u}_{\Delta x}^{n+1}$ is a vector containing our sample points, and the information about the method is contained in the matrix $\mathcal{A}_{\Delta x}$.

Important: Both the vector length and matrix dimensions *depend on the spatial step size*, Δx.

Definition: An FD numerical method for a PDE of evolution is called *convergent* if for every compact interval in time $[0,T]$, and spatial region Ω (i.e. we are working on $\Omega \times [0,T]$) we have that for all $\mathbf{x} \in \Omega$, $t \in [0,T]$, any sequence $(\mathbf{k}_i, l_i) \in \mathbb{N}^d \times \mathbb{N}$ such that

$$\mathbf{k}_i\Delta x \to \mathbf{x}, \qquad l_i\Delta t \to t,$$

with constant Courant number $\mu = \Delta t/(\Delta x)^m$, we have $u_{\mathbf{k}_i}^{l_i} \to u(\mathbf{x},t)$ as $i \to \infty$, uniformly on $\Omega \times [0,T]$.

Stability requires us to consider the norm of $\mathbf{u}_{\Delta x}^{n+1}$. We need to account for the fact its length depends on Δx. This is achieved by defining:

$$||\mathbf{u}_{\Delta x}^n||_{\Delta x} = \sqrt{\Delta x \sum_m |u_{\Delta x,m}^n|^2}.$$

This works because in the limit as $\Delta x \to 0$, the right hand side tends to an integral, i.e. the L^2 norm, as it should.

This allows us to make the definition of stability:

Definition: We say that an FD numerical method for a PDE of evolution is *(Lax) stable* if for all $[0,T]$, and any $n \in \mathbb{N}$ such that $n\Delta t \in [0,T]$, we have that $||\mathcal{A}_{\Delta x}^n||_{\Delta x}$ is uniformly bounded when $\Delta x \to 0$ and μ is kept constant.

Here, the notation $\mathcal{A}_{\Delta x}^n$ mean $\mathcal{A}_{\Delta x}$ raised to the nth power.

This is equivalent to $||\mathbf{u}^n||_{\Delta x}$ being bounded under the progression to the same limit. Equivalently, for all T, and any $n \in \mathbb{N}$ such that $n\Delta t \in [0,T]$, there exists $C_T > 0$ such that $||\mathbf{u}^n||_{\Delta x} \le C_T||\mathbf{u}^0||_{\Delta x}$ as $\Delta x \to 0$ with μ fixed.

Slogan: Stability means the numerical method is *uniformly well-posed* as $\Delta x \to 0$.

Exactly the same definitions apply for the semi-discretised scheme

$$\mathbf{u}'_{\Delta x} = \mathcal{P}_{\Delta x}\mathbf{u}_{\Delta x} + \tilde{\mathbf{f}}_{\Delta x}(t).$$

This has formal solution:

$$\mathbf{u}_{\Delta x}(t) = e^{t\mathcal{P}_{\Delta x}}\mathbf{u}_{\Delta x}(0) + \int_0^t e^{(t-\tau)\mathcal{P}_{\Delta x}}\tilde{\mathbf{f}}_{\Delta x}(\tau)d\tau,$$

using Duhamel's principle. So $e^{t\mathcal{P}_{\Delta x}}$ plays the role of $\mathcal{A}_{\Delta x}$ here. Thus we have:

Definition: Convergence means the solution of the ODE system tens to the solution of the PDE system when $\Delta x \to 0$, uniformly in Δx and $t \in [0,T]$.

Definition: Stability means that $||\exp(t\mathcal{P}_{\Delta x})||$ is uniformly bounded for all $t \ge 0$, $\Delta x \to 0$.

For both fully-discretised and semi-discretised numerical methods, we have:

Theorem (Lax equivalence): For linear well-posed PDEs of evolution, convergence is equivalence to stability and order ≥ 1.

Proof: Not in course. □

8 Stability analysis of FDMs

To use Lax equivalence, need to know about *order* (which we've already studied) and *stability*. We need to know how to deal with stability more easily.

8.1 Spectral properties of normal matrices

Definition: A matrix A is *normal* if $AA^\dagger = A^\dagger A$.

Theorem: A matrix is normal iff it has a complete set of orthonormal (in L^2) eigenvectors.

Proof: Elementary fact from undergraduate. □

Hence, $A = Q^\dagger DQ$ for Q unitary and D diagonal.

Definition: The *spectral radius* $\rho(A)$ of a matrix A is the maximum modulus of the eigenvalues of A: $\rho(A) = \max_i |\lambda_i|$.

Theorem: For a normal matrix, we have $||A||_{\text{op}} = \rho(A)$.

Proof: Let $\mathbf{v}_i$ be an evector of A with eigenvalue λ_i Note that

$$||A||_{\text{op}} = \max_{\mathbf{v}\in V}\frac{||A\mathbf{v}||}{||\mathbf{v}||} \ge \frac{||A\mathbf{v}_i||}{||\mathbf{v}_i||} = |\lambda_i|.$$

So $||A||_{\text{op}}$ is greater than the modulus of all the evalues. Hence:

$$\begin{aligned}\rho(A) \le ||A||_{\text{op}} &= ||Q^\dagger DQ||_{\text{op}} \le ||Q^\dagger||_{\text{op}}||D||_{\text{op}}||Q||_{\text{op}}\\ &= ||D||_{\text{op}} = \rho(A),\end{aligned}$$

since for unitary Q, $||Q\mathbf{v}||_2 = ||\mathbf{v}||_2$, and thus $||Q||_{\text{op}} = 1$. □

Theorem: For general A, we have $||A||_{\text{op}} = \sqrt{\rho(A^\dagger A)}$.

Proof: We have $||A||_{\text{op}}^2 =$

$$\max_{||\mathbf{v}||_2=1}(A\mathbf{v},A\mathbf{v}) = \max_{||\mathbf{v}||_2=1}(A^\dagger A\mathbf{v},\mathbf{v}) \le \max_{||\mathbf{v}||=1}\sqrt{||A^\dagger A\mathbf{v}||_2}\sqrt{||\mathbf{v}||_2},$$

where in the last step, we used the Cauchy-Schwarz inequality. The result follows immediately. □

8.2 Eigenvalue analysis of FD schemes

Theorem: Suppose $\mathcal{A}_{\Delta x}$ is normal for all Δx, and there exists $\alpha \ge 0$ such that $\rho(\mathcal{A}_{\Delta x}) \le e^{\alpha\Delta t}$. Then the FD method with matrix $\mathcal{A}_{\Delta x}$ is stable.

Proof: Diagonalise $\mathcal{A}_{\Delta x} = Q_{\Delta x}^{-1}D_{\Delta x}Q_{\Delta x}$, with $Q_{\Delta x}$ unitary, and $D_{\Delta x}$ diagonal. Then

$$||\mathcal{A}_{\Delta x}^n|| = ||Q_{\Delta x}^{-1}D_{\Delta x}^nQ_{\Delta x}|| \le ||Q_{\Delta x}|| \cdot ||Q_{\Delta x}^{-1}|| \cdot ||D_{\Delta x}||^n.$$

Now $||D_{\Delta x}|| = \rho(D_{\Delta x}) =$ max evalue of $D_{\Delta x}$, since $D_{\Delta x}$ is diagonal, so normal. But the max evalue of $D_{\Delta x}$ is the max evalue of $\mathcal{A}_{\Delta x}$, so $||D_{\Delta x}|| = ||\mathcal{A}_{\Delta x}|| = \rho(\mathcal{A}_{\Delta x})$.

But $||\mathcal{A}_{\Delta x}|| = \rho(\mathcal{A}_{\Delta x}) \leq e^{\alpha\Delta t}$ is given. So

$$||\mathcal{A}^n_{\Delta x}|| \leq ||Q_{\Delta x}|| \cdot ||Q^{-1}_{\Delta x}|| \cdot e^{n\alpha\Delta t} \leq e^{\alpha T},$$

since $n\Delta t \leq T$ and $||Q_{\Delta x}|| = 1$ since $Q_{\Delta x}$ is normal. □

The reverse of the above Theorem is also true. If the evalues are not uniformly bounded, there exists an eigenvector $\mathbf{v}_{\Delta x}$ for which $||\mathcal{A}^n_{\Delta x}\mathbf{v}_{\Delta x}||$ is not uniformly bounded, so $||\mathcal{A}^n_{\Delta x}||$ is not uniformly bounded. So:

Slogan: If $\mathcal{A}_{\Delta x}$ is normal, stability is equivalent to the evalues of $\mathcal{A}_{\Delta x}$ being uniformly bounded.

8.3 Examples of FD eigenvalue analysis

Very often, $\mathcal{A}_{\Delta x}$ takes a special form which makes the application of the above Theorem easy.

Definition: A matrix is called *Toeplitz* if it is constant along the diagonals.

Definition: A matrix is called *TST* if it is Toeplitz, symmetric and tridiagonal. It then has the form $a_{k,k} = \alpha$, $a_{k,k\pm 1} = \beta$, with all other entries zero.

Theorem: Let A be a $d \times d$ TST matrix with entries $a_{k,k} = \alpha$, $a_{k,k\pm1} = \beta$. Then the eigenvalues of A are

$$\lambda_k = \alpha + 2\beta\cos\left(\frac{k\pi}{2d+2}\right),$$

with corresponding evectors $v_{k,l} = \sin(\pi kl/(2d+2))$.

Proof: Just substitute answer in to check. □

Theorem: All TST matrices commute.

Proof: The evectors are independent of α, β in the above Theorem, hence all TST matrices have a joint eigenbasis, and thus must commute. □

Example 1: Consider Euler's method for the diffusion equation on $0 \leq x \leq 1$:

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

Here,

$$\mathcal{A}_{\Delta x} = \begin{pmatrix} 1-2\mu & \mu & \cdots & 0 \\ \mu & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mu \\ 0 & \cdots & \mu & 1-2\mu \end{pmatrix},$$

so $\mathcal{A}_{\Delta x}$ is TST with evalues:

$$\lambda_k = 1 - 4\mu\sin^2(k\pi/(2d+2)).$$

Therefore $\rho(\mathcal{A}_{\Delta x}) = |1-4\mu| \leq e^{\alpha\Delta t}$ is required for some $\alpha > 0$, and as $\Delta x \to 0$. But constant μ requires $\Delta t \to 0$ as $\Delta x \to 0$, so $|1-4\mu| \leq 1$ is an equivalent condition.

Thus method is stable iff $\mu \leq \frac{1}{2}$.

Example 2: Consider Crank-Nicolson: $u^{n+1}_m =$

$$u^n_m + \frac{1}{2}\mu\left(u^n_{m-1} - 2u^n_m + u^n_{m+1} + u^{n+1}_{m-1} - 2u^{n+1}_m + u^{n+1}_{m+1}\right).$$

Now $\mathcal{A}_{\Delta x} = B^{-1}A$, where

$$B = \begin{pmatrix} 1+\mu & -\frac{1}{2}\mu & \cdots & 0 \\ -\frac{1}{2}\mu & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & -\frac{1}{2}\mu \\ 0 & \cdots & -\frac{1}{2}\mu & 1+\mu \end{pmatrix},$$

$$A = \begin{pmatrix} 1-\mu & \frac{1}{2}\mu & \cdots & 0 \\ \frac{1}{2}\mu & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \frac{1}{2}\mu \\ 0 & \cdots & \frac{1}{2}\mu & 1+\mu \end{pmatrix}.$$

So both A and B are TST matrices. B^{-1} has the same evectors as B, so is also TST. $B^{-1}A$ also has the same evectors, so must be TST (and hence normal), with evalues:

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

and so $\mu > 0$ is the condition for stability.

Eigenvalues analysis is **wrong** if $\mathcal{A}_{\Delta x}$ is not normal. We'll consider an example, and use:

The Gerschgorin Theorem: Let A be a $d \times d$ matrix with entries a_{ij}. The *Gerschgorin disks* are defined by:

$$\mathbb{S}_k = \left\{ z \in \mathbb{C} : |z - a_{k,k}| \leq \sum_{\substack{j=1 \\ j\neq k}}^{d} |a_{k,j}| \right\}.$$

Let $\sigma(A)$ be the spectrum of A. Then:

(i) $\sigma(A) \subset \bigcup_{k=1}^{d} \mathbb{S}_k$;

(ii) When r disks form a connected region, there are r eigenvalues in that region.

Example: Consider the advection equation on $[0,1]$ solved with Euler's method

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

Here, $\mathcal{A}_{\Delta x}$ is bidiagonal, so the eigenvalues are all $1-\mu$. Thus $\rho(\mathcal{A}_{\Delta x}) = |1-\mu|$. But since $\mathcal{A}_{\Delta x}$ is *not* normal, the condition $|1-\mu| \le 1 \Rightarrow \mu \le 2$ is only *necessary* for stability.

Indeed, consider a general $d \times d$ bidiagonal matrix:

$$A_d = \begin{pmatrix} a & b & \cdots & 0 \\ 0 & a & \ddots & \vdots \\ \vdots & \ddots & \ddots & b \\ 0 & \cdots & 0 & a \end{pmatrix}$$

This gives:

$$A_d^T A_d = \begin{pmatrix} a^2 & ab & \cdots & 0 \\ ab & a^2+b^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & ab \\ 0 & \cdots & ab & a^2+b^2 \end{pmatrix}$$

Recall $||A_d||^2 = \rho(A_d^T A_d)$. Now use the Gerschgorin Theorem. The intersections of the Gerschgorin disks of $A_d^T A_d$ with the real line are $[a^2-ab, a^2+ab]$ and $[(|a|-|b|)^2, (|a|+|b|)^2]$. Hence $||A_d||^2 = \rho(A_d^T A_d) \le (|a|+|b|)^2$ by (i).

Bounding in the other direction, note that if $\mathbf{v}_d$ is chosen to be $v_{d,k} = (\operatorname{sgn}(a/b))^{k-1}$, we have

$$A_d \mathbf{v}_d = (|a|+|b|)^2 \mathbf{v}_d - \begin{pmatrix} |ab|+b^2 \\ 0 \\ \vdots \\ 0 \\ |ab| \end{pmatrix}.$$

Taking the norm of both sides, we see

$$||A_d \mathbf{v}_d|| \ge \left| (|a|+|b|)^2 ||\mathbf{v}_d|| - ||(|ab|+b^2, 0, \cdots, 0, |ab|)^T|| \right|.$$

Divide by $||\mathbf{v}_d|| = \sqrt{d}$ and consider the limit as $d \to \infty$. We then see that $||A_d \mathbf{v}_d|| / ||\mathbf{v}_d|| \ge |a|+|b|$ in the limit. So $||A_d||_{\text{op}} \ge |a|+|b|$.

Thus $||A_d|| = |a|+|b|$ by the two bounds. In our advection equation case, $||\mathcal{A}_{\Delta x}|| = |1-\mu|+\mu$, and hence the method is stable iff $0 < \mu \le 1$. So eigenvalue analysis got it wrong!

8.4 Eigenvalue analysis of SD methods

Eigenvalue analysis extends simply to semi-discretised methods.

Theorem: Let $\mathcal{P}_{\Delta x}$ be the matrix of a semi-discrete scheme, and suppose it is normal. If there exists $\beta \in \mathbb{R}$ such that $\lambda \le \beta$ for all $\lambda \in \sigma(\frac{1}{2}(\mathcal{P}_{\Delta x} + \mathcal{P}_{\Delta x}^*))$ as $\Delta x \to 0$, then the SD method is stable.

Proof: For SD schemes, $e^{t\mathcal{P}_{\Delta x}}$ plays the same role as $\mathcal{A}_{\Delta x}$. Write $\mathcal{P}_{\Delta x} = Q_{\Delta x}^\dagger D_{\Delta x} Q_{\Delta x}$ for $D_{\Delta x}$ diagonal, $Q_{\Delta x}$ unitary, so that:

$$e^{t\mathcal{P}_{\Delta x}} = Q_{\Delta x}^\dagger e^{tD_{\Delta x}} Q_{\Delta x}.$$

Take the operator norm of both sides to get $||e^{t\mathcal{P}_{\Delta x}}|| = ||e^{tD_{\Delta x}}||$. Since $|e^{x+iy}| = |e^x|$, sufficient to bound real part of eigenvalues, i.e. bound $\lambda \in \sigma\left((\frac{1}{2}\left((\mathcal{P}_{\Delta x} + \mathcal{P}_{\Delta x}^*)\right)\right)$. □

8.5 Fourier analysis

Definition: The *Fourier transform* of $\{v_m\}_{m\in\mathbb{Z}}$ is

$$\hat{v}(\theta) = \sum_{m=-\infty}^{\infty} v_m e^{-im\theta}.$$

Theorem: The Fourier transform is an $\ell_2 \to L_2$ *isometry*, characterised by *Parseval's identity*:

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

Proof: Elementary fact from undergraduate. □

Using Fourier theory, we can analyse schemes for linear PDEs with constant coefficients and *Cauchy initial data*, i.e. initial value given on all of $\mathbb{R}$, with no BCs:

Theorem: An FD scheme is stable iff the symbol H of the method obeys $|H(e^{i\theta};\mu)| \le 1$ for all θ.

Proof: The general FD method is:

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

Multiply by $e^{-im\theta}$ and sum from $m = -\infty$ to ∞. Break up $e^{-im\theta} = e^{ik\theta} e^{-i(m+k)\theta}$, and then shift indices in the m sum on the LHS, to end up with:

$$\left(\sum_{k=-r}^{s} \gamma_k e^{ik\theta}\right) \hat{u}^{n+1} = \left(\sum_{k=-r}^{s} \delta_k e^{ik\theta}\right) \hat{u}^n.$$

So $\hat{u}^{n+1}(\theta) = H(e^{i\theta};\mu)\hat{u}^n(\theta) = H^{n+1}(e^{i\theta};\mu)\hat{u}^0(\theta)$. If $|H| \leq 1$, we have $||\hat{u}^n||^2 =$

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} |\hat{u}^n(\theta)|^2 \, d\theta = \frac{1}{2\pi}\int_{-\pi}^{\pi} |H(e^{i\theta};\mu)|^{2n}|u^0(\theta)|^2 \, d\theta \leq ||\hat{u}^0||^2.$$

Using Parseval's identity, we have in real space $||\mathbf{u}^{n+1}|| \leq ||\mathbf{u}^0||$. This is equivalent to stability.

Conversely, suppose that $|H| > 1$ for some value. Choose α, β such that $|H(e^{i\theta};\mu)| > 1+\epsilon$ for all $\theta \in (\alpha,\beta)$, and consider the function with Fourier transform

$$\hat{u}(\theta) = \begin{cases} 1 \text{ for } \theta \in (\alpha,\beta) \\ 0 \text{ otherwise.} \end{cases}$$

We see that

$$||\hat{u}^n||^2 = \frac{1}{2\pi}\int_{\alpha}^{\beta} |H(e^{i\theta};\mu)|^{2n}|\hat{u}^0(\theta)|^2 \, d\theta \geq ||\hat{u}^0||^2(1+\epsilon)^{2n}.$$

Let $n \to \infty$, then $||\hat{u}^n||^2 \to \infty$. By Parseval's identity, $||\mathbf{u}^n||$ is also unbounded. □

Theorem: An SD scheme is stable iff the symbol h of the method obeys $\mathrm{Re}(h(e^{i\theta})) \leq 0$ for all θ.

Proof: Similar story. □

Example 1: Consider the diffusion equation $u_t = u_{xx}$ approximated by the *Crandall method*:

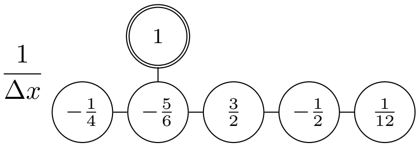
$$\left(\frac{1}{12} - \frac{1}{2}\mu\right)u_{m-1}^{n+1} + \left(\frac{5}{6} + \mu\right)u_m^{n+1} + \left(\frac{1}{12} - \frac{1}{2}\mu\right)u_{m+1}^{n+1}$$
$$= \left(\frac{1}{12} + \frac{1}{2}\mu\right)u_{m-1}^{n} + \left(\frac{5}{6} - \mu\right)u_m^{n} + \left(\frac{1}{12} + \frac{1}{2}\mu\right)u_{m+1}^{n}.$$

Fourier analysis says we should consider the symbol

$$H(e^{i\theta},\mu) = \frac{(1+6\mu)e^{-i\theta} + (10-12\mu) + (1+6\mu)e^{i\theta}}{(1-6\mu)e^{-i\theta} + (10+12\mu) + (1-6\mu)e^{i\theta}}$$
$$= \frac{(6-2\sin^2(\theta/2)) - 12\mu\sin^2(\theta/2)}{(6-2\sin^2(\theta/2)) + 12\mu\sin^2(\theta/2)}.$$

This is clearly ≤ 1, so stable.

Example 2: Consider the SD method



The symbol of the method is

$$h(e^{i\theta}) = -\frac{1}{4}e^{-i\theta} - \frac{5}{6} + \frac{3}{2}e^{i\theta} - \frac{1}{2}e^{2i\theta} + \frac{1}{12}e^{3i\theta}.$$

Taking the real part, we have

$$\mathrm{Re}(h(e^{i\theta})) = -\frac{5}{6} + \frac{5}{4}\cos(\theta) - \frac{1}{2}\cos(2\theta) + \frac{1}{12}\cos(3\theta)$$
$$\leq -\frac{5}{6} + \frac{5}{4} - \frac{1}{2} + \frac{1}{12} = 0,$$

using the bound $|\cos(\theta)| \leq 1$. Hence the method is stable.

8.6 Equivalence of analyses

The eigenvalue and Fourier approaches are equivalent, as can be seen by considering *Toeplitz operators*.

Definition: A bi-infinite Toeplitz matrix, i.e. a matrix with components $(T_{kl})_{k,l\in\mathbb{Z}}$ obeying $T_{kl} = t_{k-l}$ (constant along diagonals) for some sequence $(t_n)_{n\in\mathbb{Z}}$, is called a *Toeplitz operator*.

The *symbol* of a Toeplitz operator is the Laurent series:

$$t_T(z) = \sum_{k=-\infty}^{\infty} t_k z^k.$$

The set of all Toeplitz operators is denoted $\mathcal{T}$.

Theorem: If $T, S \in \mathcal{T}$, the following hold:

(i) $aT \in \mathcal{T}$ for $a \in \mathbb{R}$, and $t_{aT} = at_T$;

(ii) $T + S \in \mathcal{T}$, and $t_{T+S} = t_T + t_S$;

(iii) $TS \in \mathcal{T}$, and $t_{TS} = t_T t_S$;

(iv) If T is invertible, $T^{-1} \in \mathcal{T}$ and $t_{T^{-1}} = 1/t_T$.

Proof: Not in course. □

Toeplitz operators are important to use as they are the limits of Toeplitz matrices as we enter infinite dimensions. We know we are interested in the spectra of Toeplitz matrices, so we are also interested in the spectra of Toeplitz operators.

We first need to decide what we mean by an eigenvalue of a Toeplitz operator. However:

Theorem: For a finite matrix A, $A - \lambda I$ not invertible is equivalent to $A\mathbf{v} = \lambda\mathbf{v}$ for some non-zero $\mathbf{v}$ (i.e. these are two equivalent definitions of an eigenvalue). For an infinite matrix A, the implication is only one way: $A\mathbf{v} = \lambda\mathbf{v}$ for non-zero $\mathbf{v}$ implies $A - \lambda I$ not invertible, but *the reverse is not true*.

Proof: Not in course. □

Definition: The *spectrum* $\sigma(A)$ of an infinite matrix A is the set of λ for which $A - \lambda I$ is not invertible. The *point spectrum* $\sigma_p(A)$ is the set of λ for which there exists non-zero $\mathbf{v}$ with $A\mathbf{v} = \lambda\mathbf{v}$. By the above Theorem, $\sigma(A) \subseteq \sigma_p(A)$.

Definition: The *norm* of an infinite matrix is $||A|| = \max\limits_{z\in\sigma(A)} |z|$.

Theorem: For a Toeplitz operator T, the spectrum is given by:

$$\sigma(T) = \{t_T(e^{i\theta}) : -\pi \leq \theta \leq \pi\},$$

where t_T is the symbol of T.

Proof: Not in course. □

So for an FD method $\mathcal{B}\mathbf{u}^{n+1} = \mathcal{C}\mathbf{u}^n$ using an infinite number of points, with $\mathbb{B}$ and $\mathbb{C}$ Toeplitz operators, we have $\mathbf{u}^{n+1} = \mathcal{B}^{-1}\mathcal{C}\mathbf{u}^n$, and so the spectrum of the matrix is

$$\begin{aligned}\sigma(\mathcal{B}^{-1}\mathcal{C}) &= \{t_{\mathcal{B}^{-1}\mathcal{C}}(e^{i\theta}) : -\pi \leq \theta \leq \pi\} \\ &= \left\{\frac{t_{\mathcal{C}}(e^{i\theta})}{t_{\mathcal{B}}(e^{i\theta})} : -\pi \leq \theta \leq \pi\right\},\end{aligned}$$

using the properties of Toeplitz operators. Notice $t_{\mathcal{C}}/t_{\mathcal{B}} = H$, the symbol of the FD method! So eigenvalue analysis is the same as Fourier analysis for infinite matrices.

Why is normalcy a condition in finite analysis then? It is due to the Theorem:

Theorem: Let T_n be the nth principal minor of the Toeplitz operator T. If T_n is normal for each n, then $\sigma(T_n) \to \sigma(T)$ as $n \to \infty$.

Proof: Not in course. □

This is the bridge between eigenvalue and Fourier analysis. For non-normal operators, $\sigma(T_n) \not\to \sigma(T)$ in general, e.g.:

Example: Consider the advection equation scheme $u_m^{n+1} = (1-\mu)u_m^n + \mu u_{m+1}^n$. From Fourier analysis, the spectrum of the bi-infinite Toeplitz matrix is $\sigma(T) = \{1-\mu+\mu e^{i\theta} : -\pi \leq \theta \leq \pi\}$. From evalue analysis, we've seen that any principal $n \times n$ minor is

$$T_n = \begin{pmatrix} 1-\mu & \mu & \cdots & 0 \\ 0 & 1-\mu & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mu \\ 0 & \cdots & 0 & 1-\mu \end{pmatrix}$$

But $\sigma(T_n) = \{1-\mu\}$ **does not** tend to $\sigma(T)$ as $n \to \infty$.

8.7 Fourier analysis of multi-step methods

Fourier immediately extends to multi-step methods:

Example: Consider the *leapfrog method*:

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

Taking the Fourier transform, we get

$$\begin{aligned}\hat{u}^{n+1} &= \hat{u}^{n-1} + \mu(e^{i\theta} - e^{-i\theta})\hat{u}^n \quad \Rightarrow \\ \hat{u}^{n+1} &- 2i\mu\sin(\theta)\hat{u}^n - \hat{u}^{n-1} = 0.\end{aligned}$$

This has characteristic equation $z^2 - 2i\mu\sin(\theta)z - 1 = 0$. It's possible to use Cohn-Schur to bound the roots now, but in this case we can do it directly:

$$z = i\mu\sin(\theta) \pm \sqrt{1-\mu^2\sin^2(\theta)},$$

so stable for $-1 \leq \mu \leq 1$ ($|z|^2 \equiv 1$), unstable for $\mu > 1$, since $\theta = \pi/2$ gives $z = i\mu \pm i\sqrt{\mu^2-1} \Rightarrow |z| > 1$.

Note we don't exclude negative μ. Why?

For *systems* of PDEs, we get e.g. $\mathbf{u}_t = A\mathbf{u}_x$, A a matrix. Solving with, say, leapfrog, we get

$$\mathbf{u}_m^{n+1} = \mathbf{u}_m^{n-1} + \mu A(\mathbf{u}_{m+1}^n - \mathbf{u}_{m-1}^n).$$

If $A = VDV^{-1}$ for $D = \mathrm{diag}\{d_1, ..., d_M\}$ diagonal, then the equations decouple to scalar equations in some basis:

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

We end up with an 'effective' Courant number μd_l, which could be negative if $d_l < 0$.

9 Influence of boundary conditions

9.1 The Strang condition

Theorem: Stability in the presence of zero BCs is equivalent to:

(i) Fourier stability, i.e. $|H| \leq 1$ or $\mathrm{Re}(h) \leq 1$.

(ii) The *Strang condition*: $\sum\limits_{k=-r}^{s} \gamma_k z^k$ has r zeroes within $|z| < 1$ and s zeroes outside this region, and no zeroes are on $|z| = 1$. (Note that we allow SD methods with a LHS generalised to include a sum.)

Proof: Not in course. □

This is a surprising condition, since r and s are *not uniquely defined*. We can always multiply the symbol's numerator and denominator by powers of z to change r and s. However, $r+s$ *is* defined uniquely which allows this Theorem to work.

9.2 Trefethen's theory

Suppose we are solving $u_t = u_x$ by a *conservative scheme*, i.e. $|H(e^{i\theta};\mu)| \equiv 1$. We consider a wavelike solution to the numerical scheme:

$$u_m^n = e^{i(\xi m\Delta x + \omega(\xi) n\Delta t)}.$$

Definition: ξ is the *wavenumber*, $\omega(\xi)$ is the *phase velocity* and $c(\xi) = \omega'(\xi)$ is the *group velocity*.

Both are defined only when $|\xi| \leq \pi/\Delta x$, since our discrete sampling can't tell larger wavenumbers apart.

Group velocity $c(\xi)$ has an interpretation in physical systems: it is the rate at which energy associated with wavenumber ξ propagates in the system.

Example: Recall that $u_t = u_x$ has solution $u(x,t) = u_0(x+t)$, i.e. the solution propagates from right to left at constant speed 1 (and hence so must energy). Assume we solve $u_t = u_x$ with Crank-Nicolson. We find:

$$c(\xi) = \frac{\cos(\xi\Delta x)}{1 + \frac{1}{4}\mu^2\sin^2(\xi\Delta x)}.$$

For small ξ, everything is flowing in the right direction. But $c(\xi)$ changes sign in $|\xi| \leq \pi/\Delta x$ - so some wavenumbers are transported in the *wrong direction* by the numerical flow!

The idea of *Trefethen's theory* is to use a *boundary scheme* as well as an *internal scheme* to solve the equation. Everything is fine as long as the boundary scheme *and* internal scheme do not send a wave rightward at the same time; then, we have an unbounded increase in energy in the internal system.

That is, we require that there is no solution which has non-negative group velocity at the boundary and internally.

Example: If we use the boundary scheme $u_0^{n+1} = u_2^{n-1} + (\mu - 1)(u_2^n - u_0^n)$ with Crank-Nicolson, we can find the solution $u_m^n = (-1)^m = e^{i\pi m}$. This has group velocity 0, hence unstable for all values.

Alternatively, the boundary scheme $u_0^{n+1} = u_1^n$ has wavelike solution $e^{i\omega(n+1)\Delta t} = e^{i(\xi\Delta x + \omega n\Delta t)} \Rightarrow \omega = \xi/\mu \Rightarrow c = 1/\mu$, at the boundary. But $1/\mu > 1$ for $\mu < 1$, whereas $c(\xi) \in [-1,1]$ in above. So no possibility of rightward propagating wave, hence stable when $0 < \mu < 1$.

9.3 Periodic boundary conditions

Definition: A PDE in u has periodic BCs on $[0,1]$ if all derivatives of u must match at 0 and 1, i.e. $u(0,t) = u(1,t)$, $u_x(0,t) = u_x(1,t)$, etc.

Periodic boundary conditions product *circulant matrices* when we use numerical methods. These are Toeplitz matrices of the form:

$$\begin{pmatrix} f_0 & f_1 & \cdots & f_{M-1} \\ f_{M-1} & f_0 & \cdots & f_{M-2} \\ \vdots & \ddots & \ddots & \vdots \\ f_1 & \cdots & f_{M-1} & f_0 \end{pmatrix}$$

Theorem: The evalues and evectors of a circulant are

$$\lambda_l = \sum_{k=0}^{M-1} f_k e^{2\pi i k l/M}, \quad \{v_{l,k} = e^{2\pi i k l/M}\},$$

respectively.

Proof: Just substitute in and check. □

Since the evectors are orthonormal, circulants are *normal*, so normal eigenvalue analysis applies.

Example: Consider $u_t = u_x$ solved by the SD method:

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

with initial conditions $u(x,0) = \psi(x)$, $0 \leq x \leq 1$ and BCs $u(0,t) = u(1,t)$, $t \geq 0$. The method can be written as $\mathbf{u}' = A\mathbf{u}$, where

$$A = \begin{pmatrix} -\frac{3}{2} & 2 & -\frac{1}{2} & 0 & \cdots & 0 & 0 \\ 0 & -\frac{3}{2} & 2 & -\frac{1}{2} & \cdots & 0 & 0 \\ 0 & 0 & -\frac{3}{2} & 2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -\frac{1}{2} & 0 & 0 & 0 & \cdots & -\frac{3}{2} & 2 \\ 2 & -\frac{1}{2} & 0 & 0 & \cdots & 0 & -\frac{3}{2} \end{pmatrix}.$$

This is a circulant and hence the evalues are

$$\lambda_k = -\frac{3}{2} + 2e^{2\pi i k/M} - \frac{1}{2}e^{4\pi i k/M},$$

from which we see $\mathrm{Re}(\lambda_k) \leq 0$ immediately (using $|\cos(\theta)| \leq 1$). Hence stable.

9.4 The energy method

If all else fails, we can use the *energy method*. The principle is to make a direct estimate of $||\mathbf{u}||_{\Delta x}$.

Consider $u_t = a(x)u_x$ as an example with zero BCs, solved by

$$u_m' = \frac{a_m}{2\Delta x}(u_{m+1} - u_{m-1}).$$

Suppose $a(x)$ is Lipschitz, i.e. there exists α such that $|a(x) - a(y)| \le \alpha|x - y|$ for all x, y. We have, as usual,

$$||\mathbf{u}||_{\Delta x} = \left((\Delta x) \sum_{m=1}^{M-1} u_m^2 \right)^{1/2}.$$

Note $u_0 = u_M = 0$ since we have zero BCs. Then:

$$\frac{d}{dt}||\mathbf{u}||_{\Delta x}^2 = 2(\Delta x) \sum_{m=1}^{M-1} u_m u_m' = \sum_{m=1}^{M-1} a_m u_m (u_{m+1} - u_{m-1})$$

$$= \sum_{m=1}^{M-1} (a_m - a_{m+1}) u_m u_{m+1} \quad \text{(shift } m \mapsto m+1 \text{ in 2nd term)}$$

$$\le \alpha(\Delta x) \sum_{m=1}^{M-1} |u_m u_{m+1}| \quad \text{(Lipschitz property)}$$

$$\le \alpha ||\mathbf{u}||_{\Delta x}^2 \quad \text{(Cauchy-Schwarz inequality).}$$

Hence $||\mathbf{u}(t)||_{\Delta x}^2 \le e^{\alpha t}||\mathbf{u}(0)||_{\Delta x}^2$, and so we have uniform boundedness, and hence stability.

9.5 Some examples

Example 1: Consider the equation

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

for $x \in [0, 1]$ and zero BCs, where $a(x) \ge 0$.

A scheme we might use to approximate the equation is of the form $\mathbf{u}' = \mathcal{DAD}\mathbf{u}$, where the matrices are $N \times N$, $\mathcal{D}$ is skew-symmetric and $\mathcal{A}$ is diagonal, given by $\mathcal{A}_{m,m} = a(m/(N+1))$.

STABILITY: To prove stability, note

$$(\mathcal{DAD})^\dagger = (\mathcal{DAD})^T = \mathcal{D}^T \mathcal{A}^T \mathcal{D}^T = (-\mathcal{D})\mathcal{A}(-\mathcal{D}) = \mathcal{DAD}.$$

Therefore, $\mathcal{DAD}$ is Hermitian, and is thus a normal matrix, so just need to check evalues are non-positive.

Let $\mathbf{x}$ be any vector. Then:

$$\mathbf{x}^T \mathcal{DAD}\mathbf{x} = -(\mathcal{D}\mathbf{x})^T \mathcal{A}(\mathcal{D}\mathbf{x}) = -\mathbf{y}^T \mathcal{A}\mathbf{y} \ge 0,$$

where we've set $\mathbf{y} = \mathcal{D}\mathbf{x}$, and the inequality follows since $\mathcal{A}$ is positive definite. So the eigenvalues of $\mathcal{DAD}$ are all non-positive.

An example of such a method would be to take $\mathcal{D}$ to be central differences approximating $\partial/\partial x$. We then have:

$$\mathcal{D} = \frac{1}{2\Delta x} \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ -1 & 0 & 1 & \cdots & 0 \\ 0 & -1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}.$$

Since central differences commit a second order error for ODEs, we get a second order method overall.

NON-ZERO BCS: Recall non-zero BCs are the same as inclusion of an inhomogeneous term f, by Duhamel's principle. The solution of the equation is then:

$$u(x,t) = e^{t\mathcal{L}}u(x,0) + \int_0^t e^{(t-\tau)\mathcal{L}} f(x,\tau)\, d\tau,$$

where $\mathcal{L} = \partial_x(a(x)\partial_x \cdot)$. Computing the integral using trapezoidal rule for quadrature suggests the numerical scheme:

$$\mathbf{u}^{n+1} = e^{(\Delta t)\mathcal{DAD}}\mathbf{u}^n + \frac{\Delta t}{2}\left(e^{(\Delta t)\mathcal{DAD}}\mathbf{f}^n + \mathbf{f}^{n+1} \right).$$

Use of the trapezoidal rule suggests this commits an error $O((\Delta t)^2)$.

Example 2: Consider the parabolic equation

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

on $[0, 1]$ with zero BCs and an initial condition at $t = 0$.

The exact solution of this equation can be obtained by separation of variables; we find:

$$u(x,t) = \sum_{n=1}^{\infty} C_n e^{(\kappa - n^2\pi^2)t} \sin(n\pi x).$$

where the C_n depend on the initial condition. We see that $u \to 0$ as $t \to \infty$ for every initial condition iff $\kappa < \pi^2$.

Solving numerically, we might use the SD scheme

$$u_m' = \frac{1}{(\Delta x)^2}(u_{m-1} - 2u_m + u_{m+1}) + \kappa u_m, \quad m = 1, ..., M.$$

In terms of a matrix, this method can be written as $\mathbf{u}' =$

$$\left(\frac{1}{(\Delta x)^2} \begin{pmatrix} -2 & 1 & 0 & \cdots & 0 \\ 1 & -2 & 1 & \cdots & 0 \\ 0 & 1 & -2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & -2 \end{pmatrix} + \kappa \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \right) \mathbf{u}$$

$$= \begin{pmatrix} \kappa - \frac{2}{(\Delta x)^2} & \frac{1}{(\Delta x^2)} & 0 & \cdots & 0 \\ \frac{1}{(\Delta x)^2} & \kappa - \frac{2}{(\Delta x)^2} & \frac{1}{(\Delta x)^2} & \cdots & 0 \\ 0 & \frac{1}{(\Delta x)^2} & \kappa - \frac{2}{(\Delta x)^2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \kappa - \frac{2}{(\Delta x)^2} \end{pmatrix} \mathbf{u}.$$

This is TST, so we know its eigenvalues are

$$\kappa - \frac{2}{(\Delta x)^2} + \frac{2}{(\Delta x)^2} \cos\left(\frac{\pi k}{M+1}\right), \qquad k = 1, ..., M.$$

To reproduce the behaviour that $u \to 0$ as $t \to \infty$, we want all evalues to be negative. So we need

$$\kappa < \frac{4}{(\Delta x)^2} \sin^2\left(\frac{\pi k}{2(M+1)}\right), \text{ for } k = 1, 2, ..., M$$
$$\Rightarrow \quad \kappa < \frac{4}{(\Delta x)^2} \sin^2\left(\frac{\pi}{2(M+1)}\right).$$

Writing $M + 1 = 1/\Delta x$, we find that:

$$\kappa < \frac{4}{(\Delta x)^2} \sin^2\left(\frac{\pi \Delta x}{2}\right) = \pi^2 - \frac{\pi^2 (\Delta x)^2}{6} + \cdots .$$

So we need κ to be *just less* than π^2 in the numerical scheme.

10 Analysis of non-linear PDEs

10.1 Typical example

We focus on the *hyperbolic conservation law*:

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

This has three important features:

1. Characteristics: The solution is *constant along characteristics*. That is, if $u(x_0, 0) = u_0$, say, then along the ling $x = x_0 + f'(u_0)t$ the solution retains its value, i.e. $u(x_0 + f'(u_0)t, t) = u_0$.

2. Shocks: Since the slopes of characteristics can vary, they may clash; this is called a *shock*. There is clearly a discontinuity in the solution at shocks. The flow is completely into the shock and no information leaves it.

   It is possible to show that if $\Gamma(t)$ is a parametric representation of a shock, then the *Rankine-Hugeniot equation* is obeyed:

   $$\frac{d\Gamma(t)}{dt} = \frac{[f(u)]}{[u]},$$

   where $[w]$ denotes the jump across the shock, i.e.

   $$[u] = \lim_{\text{RH}}(u(t)) - \lim_{\text{LH}}(u(t)).$$

3. Rarefactions: The characteristics could depart too quickly away from one another, leaving a void called a *rarefaction fan*. There are multiple ways of patching the solution here (the solution is not unique), but only one has an important physical significance:

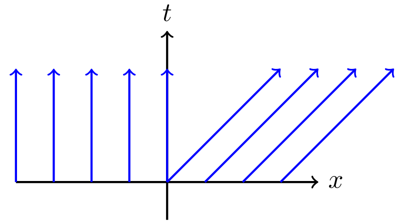
   Definition: The *entropy condition* is

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

Theorem: Provided the entropy condition is obeyed, and RH is used to resolve shocks, the solution of (†) exists, is unique and is bounded by the initial data; that is, there exists c such that

$$||u|| \le c||u(x, 0)||.$$

Example: Burgers' equation takes $f(u) = \frac{1}{2}u^2$, i.e. $u_t + uu_x = 0$. Consider initial data $u(x, 0) = 1$ for $x \ge 0$ and $u(x, 0) = 0$ for $x < 0$. We see the characteristics leave a rarefaction fan:



One way of filling the void is to use

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

The entropy condition, since $F(u) = \frac{1}{3}u^3$, and so the entropy condition is

$$uu_t + u^2 u_x \le 0.$$

It's trivial to check that our choice of u satisfies this, and hence this is the unique solution obeying the entropy condition inside the rarefaction fan.

10.2 Godunov's method & improvements

Method: *Godunov's method* approximates the initial condition $u_0(x)$ by a step function. We replace the initial condition by $u(x,0) = \tilde{u}_0(x)$, where $\tilde{u}_0(x) = u_0((m+\frac{1}{2})\Delta x)$ for $m\Delta x \leq x < (m+1)\Delta x$ (i.e. replace function by its midpoint value on each interval).

We now have a *Riemann problem*, which we can solve explicitly. If $u(x,0) = a$ for all $a \in [x_0, x_1)$, then the characteristic flow gives $u(x,t) = \tilde{u}_0(x - f'(a)t) = a$ for $x_0 \leq x - f'(a)t < x_1$.

We continue to advance t until more than two characteristics clash (we can resolve a single shock using the RH equation), and also so as not to open up rarefaction fans too much. We then just resample (provided $\Delta t < \Delta x \max |f'(\tilde{u}_0)|$, we have enough data), and iterate. It's possible to show this gives a first order method.

Method: *Van Leer's method* improves this by approximating by a piecewise linear function instead of a step function. This gives a second order method.

Method: *Glimm's method* improves this by instead of sampling $\tilde{u}_0$ at midpoints, choosing randomly on each interval. This has limited practical use, but allows us to prove the earlier Theorem about the uniqueness and existence of the solution of (†).

10.3 The Enquist-Osher method

Assume f is strictly convex, and we are solving the Cauchy problem. Since f is strictly convex, there exists a unique minimum $\overline{u} \in \mathbb{R}$ such that $f'(\overline{u}) = 0$. This $\overline{u}$ is called the *sonic point* or *stagnation point* of $f(u)$.

Definition: The *Enquist-Osher switches* are

$$f_-(y) = f(\min\{y, \overline{u}\}), \quad f_+(y) = f(\max\{y, \overline{u}\}).$$

If $y < \overline{u}$, $f_-(y) = f(y)$ and $f_+(y) = f(\overline{u}) =$ constant, and if $y > \overline{u}$, $f_+(y) = f(y)$ and $f_-(y) = f(\overline{u}) =$ constant. So one switch is always f, and one is always constant.

Because of discontinuity of shocks, we want a method that does not take points from both sides of the shock. This can be achieved by using the switches:

Method: The *Engquist-Osher method* is defined by:

$$u_m' = -\frac{1}{\Delta x}(\Delta_+ f_-(u_m) + \Delta_- f_+(u_m)).$$

If $u_{m-1}, u_m, u_{m+1} > \overline{u}$, then $\Delta_+ f_- = 0$, so we just get backward difference. If $u_{m-1}, u_m, u_{m+1} < \overline{u}$, we just get forward difference. So the Engquist-Osher method indeed allows for discontinuity at shocks.

Theorem: The Engquist-Osher method is stable.

Proof: Since the equation is non-linear, we basically have to use the energy method. We have:

$$\frac{1}{2}\frac{d}{dt}||\mathbf{u}||^2 = \underbrace{-\sum_{m=-\infty}^{\infty} u_m \Delta_+ f_-(u_m)}_{=:B_1} - \underbrace{\sum_{m=-\infty}^{\infty} u_m \Delta_- f_+(u_m)}_{=:B_2}.$$

We show $B_1 < 0$, $B_2 < 0$ separately.

We'll use the fact

$$\sum_{m=-\infty}^{\infty} \int_{u_m}^{u_{m+1}} y f_-'(y)\, dy = 0.$$

To prove this, integrate by parts, and then recognise we have a telescoping sum:

$$\sum_{m=-\infty}^{\infty} (u_{m+1} f_-(u_{m+1}) - u_m f_-(u_m)) - \int_{u_m}^{u_{m+1}} f_-(y)\, dy.$$

We get zero since $u_m \to 0$ as $m \to \pm\infty$ (in particular, the integral in the sum vanishes.

Now go back to B_1:

$$\begin{aligned} B_1 &= -\sum_{m=-\infty}^{\infty} u_m (f_-(u_{m+1}) - f_-(u_m)) \\ &= -\sum_{m=-\infty}^{\infty} u_m \int_{u_m}^{u_{m+1}} f_-'(y)\, dy. \end{aligned}$$

Add on zero using our special integral

$$B_1 = \sum_{m=-\infty}^{\infty} \int_{u_m}^{u_{m+1}} (y - u_m) f_-'(y)\, dy.$$

We claim that each term in the sum is non-positive. There are two cases:

(i) $u_{m+1} \geq u_m$. Then $y - u_m \geq 0$, and $f_-'(y) \leq 0$, since $f_-(y)$ is always decreasing. So negative.

(ii) $u_{m+1} \leq u_m$. Then $y - u_m \leq 0$, and $f_-'(y) \leq 0$. So negative again (get extra minus from limits flipping).

Similarly $B_2 < 0$. □