Lecture Note Sketches

Spectral Methods for Partial Differential Equations

Hermann Riecke

Engineering Sciences and Applied Mathematics

h-riecke@northwestern.edu

June 3, 2009

Contents

Index

References

[1] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang. *Spectral methods in fluid dynamics*. Springer, 1987.

[2] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang. *Spectral methods: fundamentals in single domains*. Springer, 2006.

[3] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang. *Spectral methods: evolution to complex geometries and applications to fluid dynamics*. Springer, 2007.

[4] M. Charalambides and F. Waleffe. Gegenbauer tau methods with and without spurious eigenvalues. *SIAM J. Num. Anal.*, 47(1):48–68, 2008.

[5] D. Gottlieb and S. A. Orszag. *Numerical Analysis of Spectral Methods: Theory and Applications*. 1977.

[6] R. A. Horn and C. R. Johnson. *Topics in matrix analysis*. Cambridge University Press, 1994.

[7] A.-K. Kassam and L. N. Trefethen. Fourth-order time-stepping for stiff pdes. *SIAM J. Sci. Comput.*, 26:1214, 2005.

[8] B. J. Matkowsky and D. O. Olagunju. Propagation of a pulsating flame front in a gaseous combustible mixture. *SIAM J. Appl. Math.*, 39(2):290–300, 1980.

[9] A. Palacios, G. H. Gunaratne, M. Gorman, and K. A. Robbins. Cellular pattern formation in circular domains. *Chaos*, 7(3):463–475, September 1997.

1 Motivation and Introduction

Central step when solving partial differential equations: approximate derivatives in space and time. Focus here on spatial derivatives.

Finite difference approximation of (spatial) derivatives:

- Accuracy depends on order of approximation ⇒ number of grid points involved in the computation (width of 'stencil')
- For higher accuracy use higher-order approximation
  ⇒ use more points to calculate derivatives
- function is approximated *locally* by polynomials of increasing order

To get maximal order use **all** points in system
⇒ approximate function *globally* by polynomials

More generally:

- approximate function by *suitable global* functions $f_k(x)$

  $$u(x) \;=\; \sum_{k=1}^{\infty} u_k f_k(x)$$

  $f_k(x)$ need not be polynomials
- calculate derivative of $f_k(x)$ analytically: exact
  ⇒ error completely in expansion

Notes:

- For smooth functions the order of the approximation of the derivative is **higher than any power**.
- high derivatives not problematic

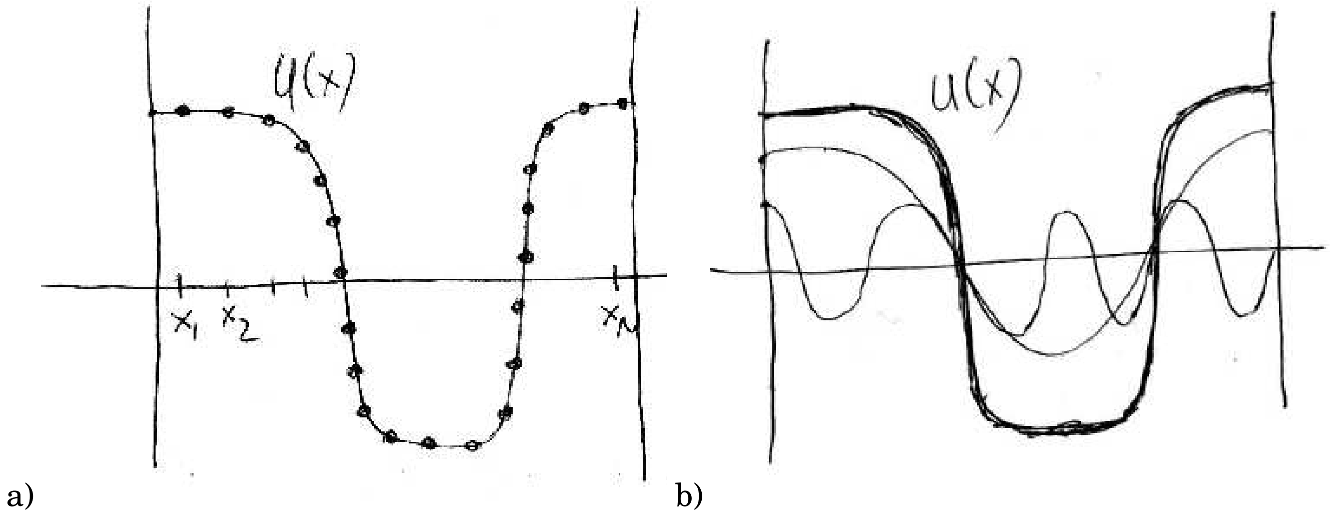


Figure 1: a) finite differences: local approximation $u = u_1, u_2, ...u_N$. Unknowns: values at grid points. b) spectral: global approximation . Unknowns: Fourier amplitudes

Note: in pseudo-spectral methods again values at grid points used although expanded in a set of global functions

Thus:

- **Study approximation of functions by sets of other functions**
- Impact of spectral approach on treatment of temporal evolution

We will use Fourier modes and Chebyshev polynomials.

Recommended books (for reference)

- *Spectral Methods* by C. Canuto, M.Y. Hussaini, A. Quarteroni, and T.A. Zang, Springer. They have written three books. [1, 2, 3]. The two new ones are not expensive.
- *Spectral Methods in MATLAB* by L.N. Trefethen, SIAM, ISBN 0898714656. Not expensive.
- *Chebyshev and Fourier Spectral Methods* by J.P. Boyd, Dover (2001). Available online at `http://www-personal.umich.edu/~jpboyd/BOOK_Spectral2000.html` and is also not expensive to buy.

1.1 Review of Linear Algebra

Motivation: Functions can be considered as vectors

$\Longrightarrow$ consider approximation of vectors by other vectors

Definition: V is a real (complex) vector space if for all $\mathbf{u}, \mathbf{v} \in V$ and all $\alpha, \beta \in R(C)$

$$\alpha\mathbf{u} + \beta\mathbf{v} \in V$$

Examples:

a) $R^3 = \{(x, y, z)|x, y, z \in R\}$ is a real vector space

b) C^n is a complex vector space

c) all continuous functions form a vector space:
$\alpha f(x) + \beta g(x)$ is a continuous function if $f(x)$ and $g(x)$ are

d) The space $V = \{f(x)|\text{continuous}, 0 \leq x \leq L, f(0) = a, f(L) = b\}$ is only a vector space for $a = 0 = b$. Why?

Definition: For a vector space V $< \cdot, \cdot >$: $V \times V \to C$ is called a scalar product or inner product iff

$$\begin{aligned} < u, v > &= < v, u >^* \\ < \alpha u + \beta v, w > &= \alpha^* < u, w > + \beta^* < v, w >, \qquad \alpha, \beta \in C \\ < u, u > &\geq 0 \\ < u, u > = 0 &\Leftrightarrow u = 0. \end{aligned}$$

Notes:

- $< u, v >$ is often written as $\mathbf{u}^+ \cdot \mathbf{v}$ with $\mathbf{u}^+$ denoting the transpose and complex conjugate of $\mathbf{u}$.
- $\mathbf{v}$ is a column vector, $\mathbf{u}^+$ is a row vector

Examples:

a) in R^3: $< u, v > = \sum_{i=1}^{3} u_i v_i$ is a scalar product

b) in $L_2 \equiv \{f(x)| \int_{-\infty}^{\infty} |f(x)|^2 \, dx < \infty\}$

$$< u, v > \quad = \quad \int_{-\infty}^{\infty} u^*(x) v(x) \, dx$$

is a scalar product.

Notes:

- $u(x)$ can be considered the "$x - th$ component" of the abstract vector $\mathbf{u}$.
- $< u, u > \equiv ||u||$ defines a norm.
- scalar product satisfies Cauchy-Schwartz inequality

  $$| < u, v > | \leq ||u|| \, ||v||$$

  (since the cosine of the angle between the vectors is smaller than 1)

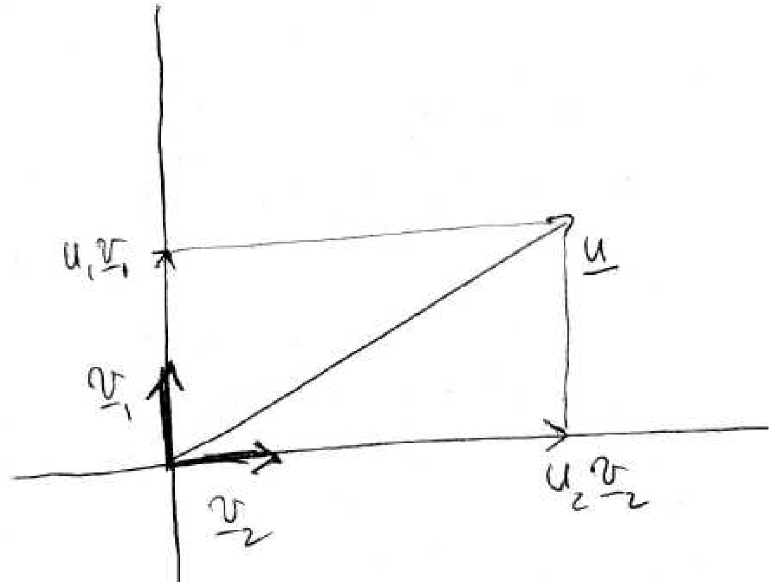
Definition: The set $\{\mathbf{v}_1, ..., \mathbf{v}_N\}$ is called an orthonormal complete set (or basis) of V if any vector $\mathbf{u} \in V$ can be written as

$$\begin{aligned} \mathbf{u} &= \sum_{k=1}^{N} u_k \mathbf{v}_k, \\ \text{with} \qquad \mathbf{v}_i^+ \cdot \mathbf{v}_j &\equiv < \mathbf{v}_i, \mathbf{v}_j >= \delta_{ij}. \end{aligned}$$

Calculate the coefficients u_i:

$$< \mathbf{v}_j, \mathbf{u} >= \sum_k u_k < \mathbf{v}_j, \mathbf{v}_k >= \sum_k u_k \delta_{kj} = u_j$$

Example: projections in R^2



$u_1 \mathbf{v}_1 =< \mathbf{v}_1, \mathbf{u} > \mathbf{v}_1$ is the *projection* of $\mathbf{u}$ onto $\mathbf{v}_1$.

Projections take one vector and transform it into another vector:

Definition: $L : V \to V$ is called a linear transformation iff

$$L(\alpha \mathbf{v} + \beta \mathbf{w}) = \alpha L \mathbf{v} + \beta L \mathbf{w}$$

Definition: A linear transformation $P : V \to V$ is called a projection iff

$$P^2 = P$$

Examples:

1. $P_v = N^{-1} \mathbf{v}\, \mathbf{v}^+$ with $N = \mathbf{v}^+ \cdot \mathbf{v}$ is a projection onto $\mathbf{v}$:

$$\begin{aligned} P_v \mathbf{u} &= \mathbf{v} \frac{\mathbf{v}^+ \cdot \mathbf{u}}{\mathbf{v}^+ \cdot \mathbf{v}} \\ P_v^2 \mathbf{u} = \mathbf{v} \frac{\mathbf{v}^+}{\mathbf{v}^+ \cdot \mathbf{v}} \cdot \left(\mathbf{v} \frac{\mathbf{v}^+ \cdot \mathbf{u}}{\mathbf{v}^+ \cdot \mathbf{v}} \right) &= \mathbf{v} \frac{\mathbf{v}^+ \cdot \mathbf{u}}{\mathbf{v}^+ \cdot \mathbf{v}} = P_v \mathbf{u} \end{aligned}$$

Notes:

- v can be thought of as a column vector and $\mathbf{v}^+$ a row vector
  $\Rightarrow \mathbf{v}^+ \cdot \mathbf{v}$ is a scalar while $\mathbf{v}\,\mathbf{v}^+$ is a projection operator
- $\mathbf{v}^+ \cdot \mathbf{u}/\mathbf{v}^+ \cdot \mathbf{v}$ is the length of the projection of u onto v

2. Let $\{\mathbf{v}_i, i = 1..N\}$ be a complete orthonormal set

$$\mathbf{u} = \sum_{k=1}^{N} (\mathbf{v}_k^+ \cdot \mathbf{u})\mathbf{v}_k = (\sum_{k=1}^{N} \mathbf{v}_k \mathbf{v}_k^+) \cdot \mathbf{u}$$

thus we have

$$\sum_{k=1}^{N} \mathbf{v}_k \mathbf{v}_k^+ = I$$

i.e. the sum over all projections onto a complete set yields the identity transformation: **completeness** of the set v

3. A linear transformation L can be represented by a matrix:

$$(L\mathbf{u})_i = \mathbf{v}_i^+ L \sum_{j=1}^{N} u_j \mathbf{v}_j = \sum_j \mathbf{v}_i^+ L \mathbf{v}_j \, u_j = \sum_j L_{ij} u_j$$

with

$$L_{ij} = \mathbf{v}_i^+ L \mathbf{v}_j$$

The identity transformation is given by the matrix

$$I_{ij} = \mathbf{v}_i^+ (\sum_k \mathbf{v}_k \mathbf{v}_k^+) \mathbf{v}_j = \sum_k \delta_{ik}\delta_{kj} = \delta_{ij}$$

can write this also as

$$I_{ij} = \sum_k \underbrace{\mathbf{v}_i^+ \mathbf{v}_k}_{i^{th}-\text{component of } \mathbf{v}_k} \cdot \underbrace{\left(\mathbf{v}_j^+ \mathbf{v}_k\right)^+}_{\text{cc of } j^{th}-\text{component of } \mathbf{v}_k} \tag{1}$$

Note: The matrix elements L_{ij} depend on the choice of the basis

Getting back to functions: Vector spaces formed by functions often cannot be spanned by a finite number of vectors, i.e. **no finite** set $\{\mathbf{v}_1, ..., \mathbf{v}_N\}$ suffices ⇒ need to consider **sequences** and **series** of vectors. We will not dwell on this sophistication.

2 Approximation of Functions by Fourier Series

Periodic boundary conditions are well suited to study phenomena that are not dominated by boundaries. For periodic functions it is natural to attempt approximations by Fourier series.

Consider the set of functions $\{\phi_k(x) = e^{ikx} | k \in N\}$. It forms a complete orthogonal set of $L_2[0, 2\pi]$.

1. Orthogonal

$$\phi_k^+ \cdot \phi_l \equiv < \phi_k, \phi_l > = \int_0^{2\pi} (e^{ikx})^* e^{ilx}\, dx = 2\pi \delta_{lk}$$

   as before e^{ikx} is the x^{th}-component of ϕ_k

2. Complete:

for any $u(x) \in L_2[0, 2\pi]$ there exist $\{u_k | k \in N\}$

$$\lim_{N\to\infty} ||u(x) - \sum_{k=-N}^{N} u_k \phi_k(x)||^2 = 0$$

i.e.

$$\lim_{N\to\infty} \int_0^{2\pi} |u(x) - \sum_{k=-N}^{N} u_k e^{ikx}|^2\, dx = 0$$

with the Fourier components given by

$$u_k = \frac{1}{2\pi} < \phi_k^+, u > = \frac{1}{2\pi} \int_0^{2\pi} e^{-ikx} u(x)\, dx$$

Note:

- Completeness $\sum_{k=1}^{N} \mathbf{v}_k \mathbf{v}_k^+ = I$ (cf (1)) implies

$$\lim_{N\to\infty} \sum_{|k|=0}^{\infty} \phi_k(x)\phi_k^+(x') = \lim_{N\to\infty} \sum_{|k|=0}^{N} e^{ik(x-x')} = 2\pi \sum_{l=-\infty}^{\infty} \delta(x - x' + 2\pi l). \qquad (2)$$

Definition: The spectral projection $P_N u(x)$ of $u(x)$ is defined as

$$P_N u(x) = \sum_{|k|=0}^{N} u_k \phi_k(x).$$

Thus,

$$\lim_{N\to\infty} ||u(x) - P_N u(x)||^2 = 0.$$

Notes:

- P_N is a projection, i.e. $P_N^2 = P_N$ (see homework)
- P_N projects $u(x)$ onto the subspace of the lowest $2N + 1$ Fourier modes

- $||P_N u(x)||^2 = 2\pi \sum_{|k|=0}^{N} |u_k|^2$:

$$\begin{aligned} ||P_N u(x)||^2 &= < P_N u, P_N u > \\ &= < \sum_{|k|=0}^{N} u_k \phi_k(x), \sum_{|l|=0}^{N} u_l \phi_l(x) > \\ &= \sum_{kl} u_k^* u_l < \phi_k(x), \phi_l(x) > \\ &= \sum_{kl} u_k^* u_l \, 2\pi \, \delta_{kl} \\ &= 2\pi \sum_{|k|=0}^{N} |u_k|^2. \end{aligned}$$

- Parseval identity extends this to the limit $N \rightarrow \infty$:

- $$||u||^2 = \lim_{N\rightarrow\infty} ||P_N u||^2 = \lim_{N\rightarrow\infty} 2\pi \sum_{|k|=0}^{\infty} |u_k|^2$$

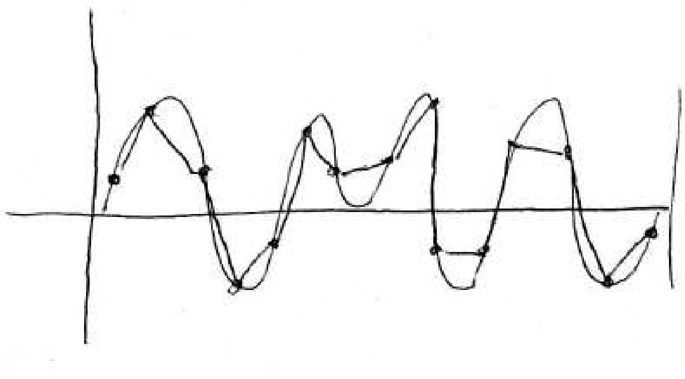
  i.e. the L_2-norm of a vector is given by the sum of the squares of its components for any orthonormal complete set. Thus, as more components are included the retained "energy" approaches the full energy.

  Proof: we have
  $$\lim_{N\rightarrow\infty} ||u(x) - P_N u(x)||^2 = 0$$
  and want to conclude $||u(x)||^2 = \lim_{N\rightarrow\infty} ||P_N u(x)||^2$.
  Consider
  $$\begin{aligned} (||u|| - ||v||)^2 &= ||u||^2 + ||v||^2 - 2||u|| \, ||v|| \\ &\leq ||u||^2 + ||v|^2 - 2| < u, v > | \end{aligned}$$
  using Schwartz inequality $| < u, v > | \leq ||u|| \, ||v||$ (projection is smaller than the whole vector).
  Now use $2| < u, v > | \geq 2Re(< u, v >) =< u, v > + < v, u >$ (note $< u, v >$ is in general complex).
  Then
  $$||u||^2 - ||v||^2 \leq ||u||^2 + ||v|^2 - < u, v > - < v, u > = < u - v, u - v > = ||u - v||^2.$$
  Get Parseval identity with $v = P_N u$.

2.1 Convergence of Spectral Projection

Convergence of Fourier series depends strongly on the function to be approximated

The highest wavenumber needed to approximate a function well surely depends on the number of "wiggles" of that function.



Definition: The total variation $\mathcal{V}(u)$ of a function $u(x)$ on $[0, 2\pi]$ is defined as

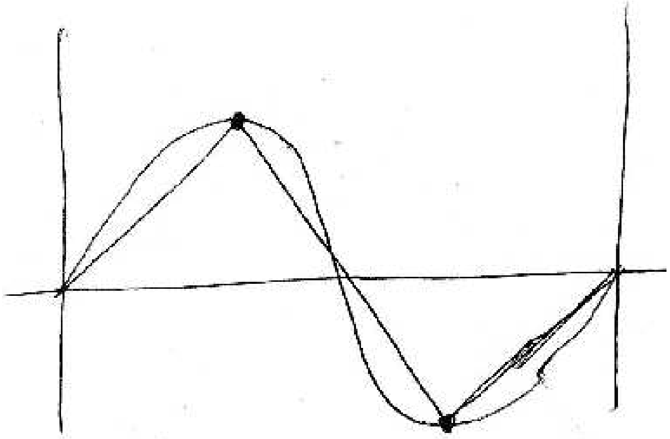
$$\mathcal{V}(u) = \sup_n \sup_{0=x_0<x_1<\ldots<x_n=2\pi} \sum_{i=1}^{n} |u(x_i) - u(x_{i-1})|$$

Notes:

- the supremum is defined as the lowest upper bound
- for supremum need only consider x_i at extrema

Examples:

1. $u(x) = \sin x$ on $[0, 2\pi]$ has $\mathcal{V}(u) = 4$



2. variation of $u(x) = \sin \frac{1}{x}$ is unbounded on $(0, 2\pi]$.

Results: One has for the spectral projection:

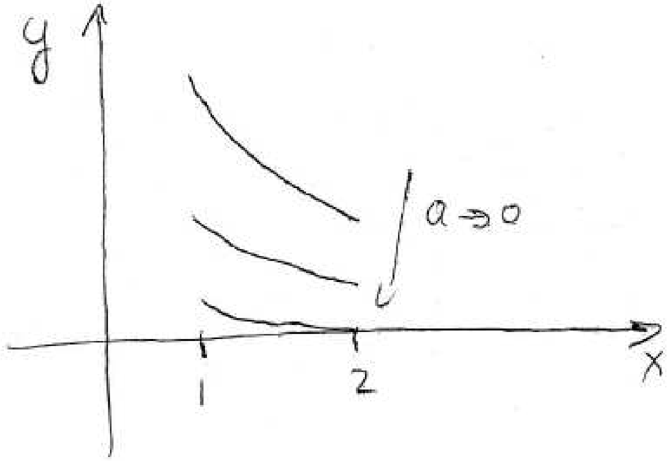
1. $u(x)$ continuous, periodic and of bounded variation
   $\Rightarrow P_N u$ converges *uniformly* and *pointwise* to u:

$$\lim_{N\to\infty} \max_{x\in[0,2\pi]} \left| u(x) - \sum_{|k|=0}^{N} e^{ikx} u_k \right| = 0$$

Notes:

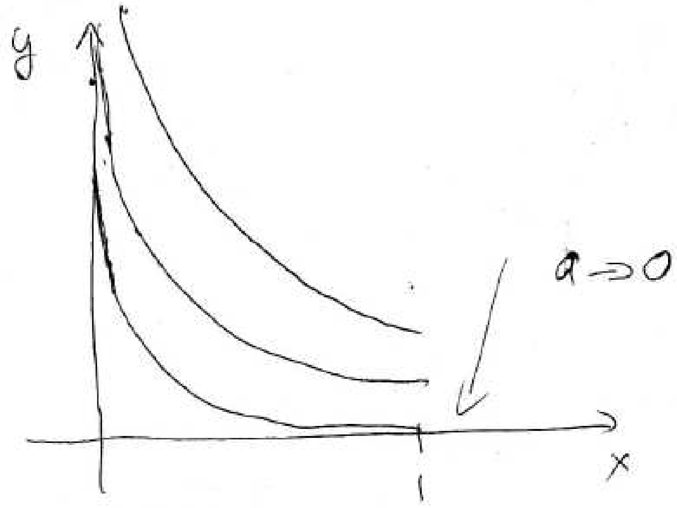
- example for uniform and non-uniform convergence:
  consider $u(x) = \frac{a}{x}$
  - on $[1,2]$ $\quad \lim_{a\to 0} u(x) = 0$ converges uniformly

$$\max_{x\in[1,2]} \left| \frac{a}{x} \right| = a \to 0$$



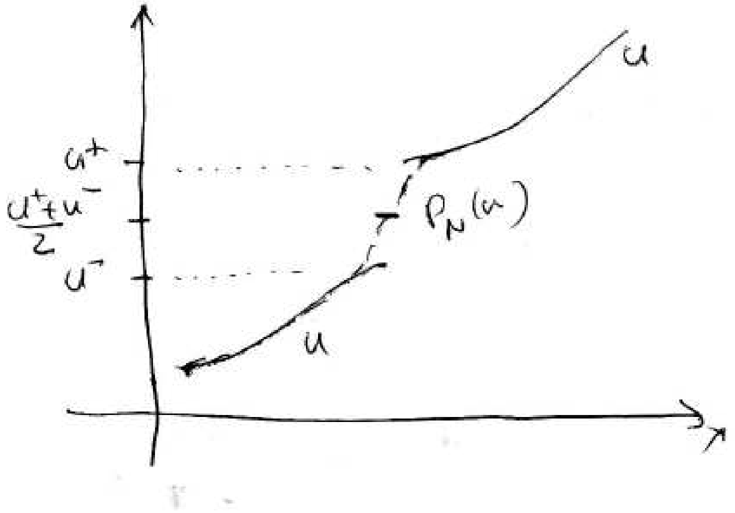
  - on $(0,1)$ $\quad \lim_{a\to 0} u(x) = 0$ converges but **not** uniformly

$$\max_{x\in(0,1)} \left| \frac{a}{x} \right| = \text{does not exist} \qquad \sup_{x\in(0,1)} \left| \frac{a}{x} \right| = \infty$$



Thus:
uniform convergence of Fourier approximation ⇒ there is an upper bound for error along the *whole* function (upper bound on *global* error).

2. $u(x)$ of bounded variation
$\Rightarrow P_N u$ converges pointwise to $\frac{1}{2}(u^+(x) + u^-(x))$ for any $x \in [0, 2\pi]$ where at discontinuities $u^\pm(x) = u(x \pm \epsilon)$
Note: even if $u(x)$ is discontinuous $P_N u(x)$ is always continuous for finite N



(a)

Figure 2: The spectral approximation is continuous even if the function to be approximated is discontinuous.

3. For $u(x) \in L_2$ the projection $P_N u$ converges in the mean,

$$\lim_{N\to\infty} \int_{-\infty}^{\infty} |u(x) - \sum_k \phi_k u_k|^2 \, dx = 0$$

but possibly $u(x_0) \neq P_N u(x_0)$ at isolated values of x_0, i.e. pointwise convergence except for possibly a "set of measure 0" (consisting of discontinuities and square-integrable singularities)

4. $u(x)$ continuous and periodic: $P_N u$ need not necessarily converge for all $x \in [0, 2\pi]$
Note: What could go 'wrong'? Are there functions that are periodic and continuous but have unbounded variation?
consider $u(x) = x \sin \frac{1}{x}$ on $[-\frac{1}{\pi}, \frac{1}{\pi}]$ (note $\sin \frac{1}{x}$ is not defined at $x = 0$)
$u(x)$ is continuous: $\lim_{x\to 0} x \sin \frac{1}{x} = 0$
$u(x)$ is periodic on $[-\frac{1}{\pi}, \frac{1}{\pi}]$
$u(x)$ not differentiable at $x = 0$: $u'(x) = \sin \frac{1}{x} - \frac{1}{x} \cos \frac{1}{x}$

Decay Rate of Coefficients:

The error $||u - P_N u|| = \sum_{|k|>N} |u_k|^2$ is determined by u_k for $|k| > N$ (cf. Parseval identity). Question: how fast does the error decrease as N is increased?

$\Rightarrow$ consider u_k for $k \to \infty$

$$\begin{aligned}
2\pi\, u_k &= <\phi_k, u> = \int_0^{2\pi} e^{-ikx} u(x)\, dx \\
&= \frac{i}{k} e^{-ikx} u(x)|_0^{2\pi} - \frac{i}{k}\int_0^{2\pi} e^{-ikx}\frac{du}{dx}\, dx \\
&= \frac{i}{k}(u(2\pi^-) - u(0^+)) - \frac{i}{k} < \phi_k, \frac{du}{dx} > \\
&\dots \\
&= \frac{i}{k}(u(2\pi^-) - u(0^+)) + \dots + (-1)^{r-1}(\frac{i}{k})^r \left(\left.\frac{d^{r-1}u}{dx^{r-1}}\right|_{2\pi^-} - \left.\frac{d^{r-1}u}{dx^{r-1}}\right|_{0^+} \right) + (-1)^r(\frac{i}{k})^r < \phi_k, \frac{d^r u}{dx^r} > .
\end{aligned}$$

Use Cauchy-Schwarz $| < \phi_k, \frac{d^r u}{dx^r} > | \leq ||\phi_k||\, ||\frac{d^r u}{dx^r}||$ as long as $||\frac{d^r u}{dx^r}|| < \infty$ (using $||\phi_k|| = \sqrt{2\pi}$):

$$|u_k| \leq \left| \frac{1}{2\pi k}\left(u(2\pi^-) - u(0^+)\right)\right| + \dots + \frac{1}{2\pi}\left| (\frac{1}{k})^r \left(\left.\frac{d^{r-1}u}{dx^{r-1}}\right|_{2\pi^-} - \left.\frac{d^{r-1}u}{dx^{r-1}}\right|_{0^+} \right)\right| + \left| \frac{1}{\sqrt{2\pi}k^r} ||\frac{d^r u}{dx^r}|| \right| .$$

Thus:

- for non-periodic functions

  $$|u_k| = \mathcal{O}\left(\frac{1}{k}(u(2\pi^-) - u(0^+))\right)$$

- for $C^\infty-$functions whose derivatives are all periodic *iterate* integration by parts *indefinitely*:

  $$|u_k| \leq \frac{1}{\sqrt{2\pi}k^r}\, ||\frac{d^r u}{dx^r}|| \qquad \text{for any } r \in N.$$

  Decay in k faster than any power law. One can show that

  $$u_k \sim e^{-\alpha |k|}$$

  with 2α being given by the strip of analyticity of $u(x)$ when extended to the complex plane (cf. Boyd, theorem 5, p.45).

  Example:

  With $z \equiv x + iy$ consider $u(z) = \tanh(\xi \sin\, z)$ along the imaginary axis:

  $$\tanh(\xi \sin iy) = \frac{\sinh(\xi i \sinh y)}{\cosh(\xi i \sinh y)} = \frac{i\, \sin(\xi \sinh y)}{\cos(\xi \sinh y)}$$

  has a first singularity at $y^\pm$ with $\xi \sinh y^\pm = \pm\frac{1}{2}\pi$. Strip of analyticity has width $2\alpha = y^+ - y^- \sim \frac{1}{\xi}$. The steeper $u(x)$ at $x = 0$ the narrower the strip of analyticity and the slower the decay of the Fourier coefficients.

- Cauchy-Schwarz estimate too soft: iteration possible as long as

$$\left| < \phi_k, \frac{d^r u}{dx^r} > \right| \quad < \quad \infty$$

(i.e. $\frac{d^r u}{dx^r} \in L_1$, see e.g. Benedetto: *Real Analysis*):
Thus

$$\left. \begin{array}{c} \frac{d^l u}{dx^l} \text{ periodic for } 0 \leq l \leq r-2 \\ \\ \frac{d^r u}{dx^r} \in L_1 \end{array} \right\} \Rightarrow u_k = \mathcal{O}\left(\frac{1}{k^r}\right)$$

Note:

- only $\frac{d^{r-2}u}{dx^{r-2}}$ has to be periodic because boundary contribution of $\frac{d^{r-1}u}{dx^{r-1}}$ is of the same order as that of the integral over $\frac{d^r u}{dx^r}$

Examples:

1. $u(x) = (x-\pi)^2$ is C^∞ in $(0, 2\pi)$, but derivative is not periodic:

   $$u_k = \frac{1}{2\pi} \int_0^{2\pi} e^{-ikx} (x-\pi)^2 \, dx = \frac{2}{k^2}$$

   origin for only quadratic decay are the boundary terms:

   $$u_k = -\frac{i}{2\pi k} \int_0^{2\pi} e^{-ikx} \frac{du}{dx} \, dx = \frac{1}{2\pi} \frac{1}{k^2} (u'(2\pi^-) - u'(0^+)) + \frac{1}{2\pi} \frac{1}{k^2} \int_0^{2\pi} e^{-ikx} u''(x) dx = \frac{2}{k^2}$$

   since $u'(2\pi^-) = 2\pi = -u'(0^+)$ and $\int_0^{2\pi} e^{-ikx} u''(x) dx = 0$.

2. $u(x) = x^2 - \theta(x-\pi)\left((x-2\pi)^2 - x^2\right)$ should be similar:
   periodic, but discontinuity of derivative
   1^{st} derivative has jump, 2^{nd} derivative has a $\delta-$function, 3^{rd} derivative involves the derivative of the δ-function: $\langle \phi_k, \delta'(x) \rangle = \mathcal{O}(k)$.

Estimate Convergence Rate of Spectral Approximation

Consider approximation for $u(x)$

$$E_N^2 \equiv ||u - P_N u||^2 = \sum_{|k|>N} |u_k|^2 = \sum_{|k|>N} |u_k|^2 \frac{|k|^{2r}}{|k|^{2r}} < \frac{1}{N^{2r}} \sum_{|k|>N} |u_k|^2 \, |k|^{2r}$$

If $\frac{d^r u}{dx^r}$ exists and is square-integrable then the sum converges and is bounded by the norm $||\frac{d^r u}{dx^r}||^2$:

$$\sum_{|k|>N} |u_k|^2 \, |k|^{2r} < \sum_{|k|=0}^{\infty} |k|^{2r} \, |u_k|^2 = ||\frac{d^r u}{dx^r}||^2$$

Thus:

$$||u - P_N u||^2 \leq \frac{1}{N^{2r}} ||\frac{d^r u}{dx^r}||^2.$$

For $u(x) \in C^\infty$ with all derivatives periodic the inequality holds for any r

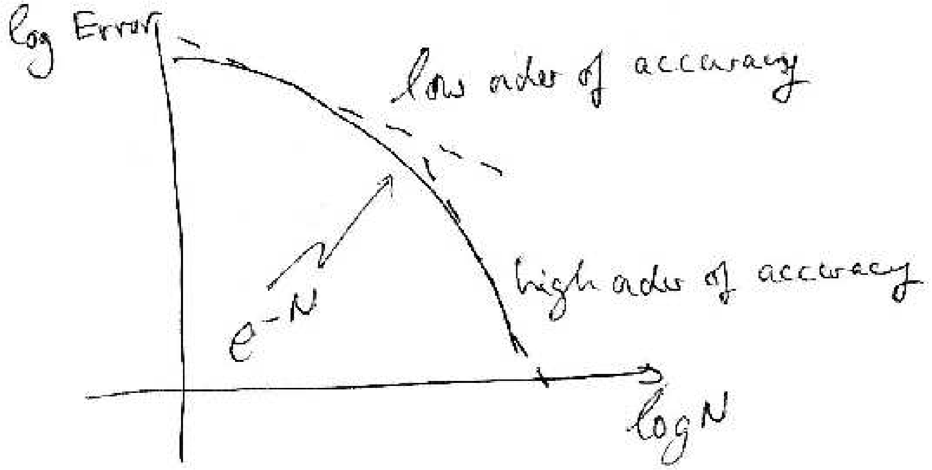
$$||u - P_N u||^2 \leq \inf_r \frac{1}{N^{2r}} ||\frac{d^r u}{dx^r}||^2 \tag{3}$$

Notes:

- The order of convergence depends on the *smoothness* of the function (highest square-integrable derivative)
- For $u(x) \in C^\infty$: $u_k \sim e^{-\alpha|k|}$
  $\Rightarrow$ one gets convergence faster than any power: **spectral or infinite-order accuracy**:

  $$||u - P_N u||^2 = \sum_{|k|>N} |u_k|^2 \sim 2e^{-\alpha(N+1)} \sum_{k=0}^{\infty} \left(e^{-\alpha}\right)^k = 2e^{-\alpha(N+1)} \frac{1}{1-e^{-\alpha}} = \left(\frac{2e^{-\alpha}}{1-e^{-\alpha}}\right) e^{-\alpha N}$$

  with 2α being the width of the strip of analyticity of $u(x)$ when $u(x)$ is continued analytically into the complex plane (cf. Trefethen Theorem 1c, p.30, Boyd theorem 5, p.45)



Spectral Approximation:

- convergence becomes faster with increasing N
- high-order convergence only for sufficiently large N

Finite-Difference Approximation:

- order of convergence fixed

- *Effective exponent* of convergence depends on N:
  Note: in general

  $$||\frac{d^r u}{dx^r}||^2 \rightarrow \infty \text{ faster than exponentially for } r \rightarrow \infty$$

  - Example

    $$||\frac{d^r e^{iqx}}{dx^r}|| = q^r ||e^{iqx}||$$

    Thus, for simple complex exponential $||\frac{d^r}{dx^r} e^{iqx}||$ grows exponentially in r.
  - For functions that are *not* given by a *finite* number of Fourier modes the norm has to grow with r faster than exponentially:
    show by contradiction

    $$\text{If} \qquad ||\frac{d^r u}{dx^r}||^2 \propto \eta^{2r} \qquad \text{then} \qquad E_N \propto \left(\frac{\eta}{N}\right)^{2r}$$

    Can then pick a *fixed* $N > \eta$ to get

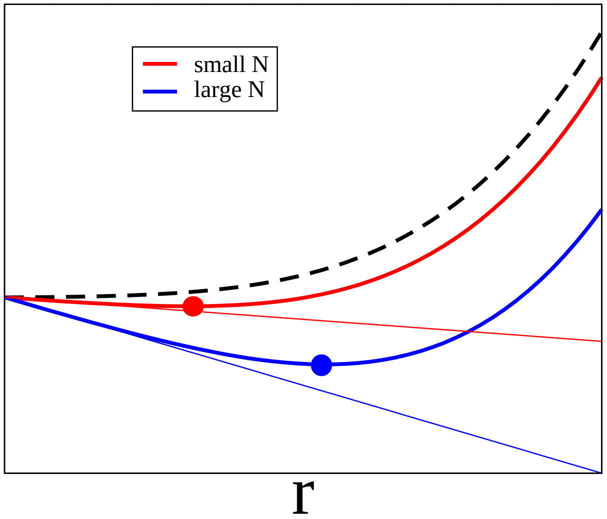
    $$\inf_r \; E_N = 0$$

    $\Rightarrow$ approximation is exact for *finite* N in contradiction to assumption..

Now consider

$$\ln E_N \leq \ln \left(\inf_r \frac{1}{N^{2r}} ||\frac{d^r u}{dx^r}||^2 \right) = \inf_r \left(\ln ||\frac{d^r u}{dx^r}||^2 - 2r \ln N \right)$$

$||\frac{d^r u}{dx^r}||^2$ grows faster than exponential $\Rightarrow \ln ||\frac{d^r u}{dx^r}||^2$ grows faster than linearly for large r



$\Rightarrow$ can pick N sufficiently large that for small r denominator N^r grows faster in r
$\Rightarrow$ error estimate decreases with r
for larger r the exponential N^r does not grow fast enough
$\Rightarrow$ error estimate grows with r
value of r at the minimum gives *effective exponent* for decrease in error in this regime of N.
With increasing N the minimum in the error estimate (solid circle in the figure) is shifted to larger r
$\Rightarrow$ effective order of accuracy increases with N :

Note:
Spectral approximation guaranteed to be superior to finite difference methods only in highly accurate regime

Approximation of Derivatives

Given $u(x)=\sum u_k e^{ikx}$ the derivatives are given by

$$\frac{d^n u}{dx^n}=\sum_{|k|=0}^{\infty}(ik)^n u_k e^{ikx}$$

if the series for the derivative converges (again, convergence in the mean)

Note:

- not all square-integrable functions have square-integrable derivatives

$$\frac{d\theta}{dx}=\delta(x)$$

- if series for $u(x)$ converges *uniformly* then its 1^{st} derivative still converges (possibly not uniformly)
- convergence for $\frac{d^q u}{dx^q}$ is a power of N^q slower than that for u since one can take only q fewer derivatives of it than of u,

$$\frac{d^q u}{dx^q}=\sum_k (ik)^q u_k\, e^{ikx}$$

  coefficients $(ik)^q u_k$ decay more slowly than u_k itself.
  the estimate (3) gets weakened by

$$||\frac{d^q u}{dx^q}-P_N\frac{d^q u}{dx^q}||^2\leq \inf_r \frac{1}{N^{2r-2q}}||\frac{d^r u}{dx^r}||^2 \qquad \text{for } r>q$$

- Periodic boundary conditions: non-periodic derivative $\frac{d^r u}{dx^r}$ equivalent to discontinuous $\frac{d^r u}{dx^r}$, i.e. $\frac{d^{r+1} u}{dx^{r+1}}$ not square-integrable

2.2 The Gibbs Phenomenon

Consider convergence in more detail for $u(x)$ piecewise continuous

$$P_N u(x)=\sum_{|k|=0}^{N} u_k e^{ikx}=\frac{1}{2\pi}\int_0^{2\pi}\sum_{|k|=0}^{N} e^{-ikx'+ikx}u(x')\,dx'$$

P_N can be written more compact using

$$D_N(s)\equiv\sum_{|k|=0}^{N} e^{iks}=\frac{\sin(N+\frac{1}{2})s}{\sin(\frac{1}{2}s)}.$$

This identity can be shown by multiplying by the denominator:

$$\left(e^{i\frac{1}{2}s}-e^{-i\frac{1}{2}s}\right)\left[e^{-iNs}+e^{-i(N-1)s}+...+e^{iNs}\right]=e^{i(N+\frac{1}{2})s}-e^{-i(N+\frac{1}{2})s}$$

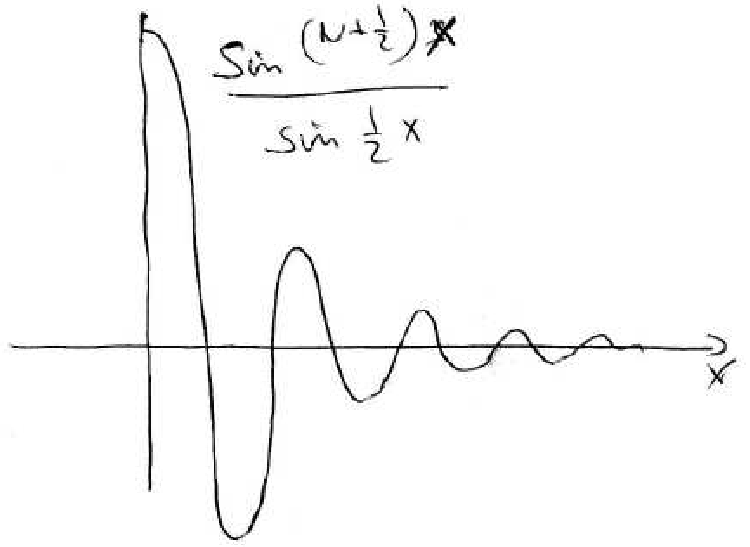
Insert

$$\begin{aligned} P_Nu(x) &= \frac{1}{2\pi}\int_0^{2\pi}\frac{\sin\left[(N+\frac{1}{2})(x-x')\right]}{\sin\left[\frac{1}{2}(x-x')\right]}u(x')\,dx' \\ &\underbrace{=}_{\text{use } t=x-x'} \frac{1}{2\pi}\int_{x-2\pi}^{x}\frac{\sin(N+\frac{1}{2})t}{\sin\frac{1}{2}t}u(x-t)\,dt \end{aligned}$$

Use the completeness of the Fourier modes

$$\lim_{N\to\infty}D_N(s)=\sum_{|k|=0}^{\infty}e^{iks}=2\pi\sum_{l=-\infty}^{\infty}\delta(s+2\pi l)$$

⇒ for large N the sum $D_N(s)$ is negligible except near $s=2\pi l$, $l=0,\pm1,\pm2,...$.



Assume $u(x)$ is discontinuous at x_0

$$u(x_0^-)=u^- \quad u(x_0^+)=u^+$$

Consider in particular points close to the discontinuity

$$x=x_0+\frac{\Delta x}{N+\frac{1}{2}}, \qquad \left|\frac{\Delta x}{N+\frac{1}{2}}\right|\ll 1,$$

and use that $D_N(t)$ decays rapidly away from $t=0$

$$P_Nu(x_0+\frac{\Delta x}{N+\frac{1}{2}})\approx\frac{1}{2\pi}\int_{-\epsilon}^{\epsilon}\frac{\sin(N+\frac{1}{2})t}{\sin\frac{1}{2}t}u(x_0+\frac{\Delta x}{N+\frac{1}{2}}-t)dt$$

Approximate $u(x)$ in the integrand by u^+ and u^-, respectively,

$$P_N u(x_0 + \frac{\Delta x}{N+\frac{1}{2}}) \approx \frac{1}{2\pi}\, u^+ \int_{-\epsilon}^{\frac{\Delta x}{N+\frac{1}{2}}} \frac{\sin(N+\frac{1}{2})t}{\frac{1}{2}t} dt + \frac{1}{2\pi}\, u^- \int_{\frac{\Delta x}{N+\frac{1}{2}}}^{\epsilon} \frac{\sin(N+\frac{1}{2})t}{\frac{1}{2}t} dt$$

Now write $s = (N+\frac{1}{2})t$ and consider $N \to \infty$ for fixed ϵ

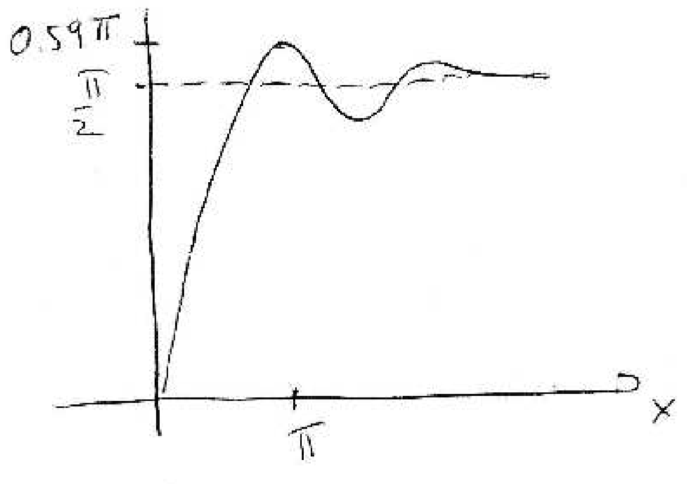
$$\begin{aligned} \int_{-(N+\frac{1}{2})\epsilon}^{\Delta x} \frac{\sin s}{s}\, ds &\to \int_{-\infty}^{\Delta x} \frac{\sin s}{s}\, ds \\ &= \int_{-\infty}^{0} \frac{\sin s}{s}\, ds + \int_{0}^{\Delta x} \frac{\sin s}{s}\, ds \\ &= \frac{\pi}{2} + Si(\Delta x) \end{aligned}$$

with $Si(x)$ the sine integral and $\lim_{x\to\infty} Si(x) = \pi/2$.
Similarly:

$$\begin{aligned} \int_{\Delta x}^{\epsilon(N+\frac{1}{2})} \frac{\sin s}{s}\, ds &\to \int_{\Delta x}^{\infty} \frac{\sin s}{s}\, ds \\ &= \frac{1}{2}\int_{-\infty}^{\infty} \frac{\sin s}{s}\, ds + \int_{\Delta x}^{0} \frac{\sin s}{s}\, ds \\ &= \frac{\pi}{2} - Si(\Delta x) \end{aligned}$$

Thus

$$P_N u(x_0 + \frac{\Delta x}{N+\frac{1}{2}}) \approx \frac{1}{2}(u^+ + u^-) + \frac{1}{\pi} Si(\Delta x)(u^+ - u^-)$$



Note:

- Maximal overshoot is 9% of the jump (independent of N)

$$P_N u(x_0 + \frac{\pi}{N+\frac{1}{2}}) - u^+ = (u^+ - u^-)\left(\frac{1}{\pi}Si(\pi) - \frac{1}{2}\right) = (u^+ - u^-)\, 0.09$$

- Location of overshoot at $x_0 + \frac{\pi}{N+\frac{1}{2}}$ converges to jump position x_0. Everywhere else series converges pointwise to $u(x)$
- the maximal error does not decrease: convergence is not *uniform* in x; but convergence in the L_2-norm, since area between $P_N u$ and u goes to 0.
- Smooth oscillation can indicate severe problem: *unresolved* discontinuity.
  To capture true discontinuity finite differences may be better.
- Smooth step (e.g. $\tanh x/\xi$):
  as long as step is not resolved expect behavior like for discontinuous function
  slow convergence and Gibbs overshoot (⇒HW), only when enough modes are retained to resolve the step the exponential convergence will set in.

2.3 Discrete Fourier Transformation

We had continuous Fourier transformation

$$u(x) = \sum_{|k|=0}^{\infty} e^{ikx} u_k$$

with

$$u_k = \frac{1}{2\pi}\int_0^{2\pi} e^{-ikx} u(x) dx$$

Consider evolution equation

$$\frac{\partial u}{\partial t} = F(u, \frac{\partial u}{\partial x})$$

Our goal was to do the time-integration completely in Fourier space since our variables are the Fouriermodes ⇒ need Fourier components F_k

Consider linear PDE:

- $F(u, \frac{\partial}{\partial x}) = \partial_x^2 u$

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

Insert Fourier expansion and project onto $\phi_k = e^{ikx}$

$$\frac{du_k}{dt} = -k^2 u_k$$

Consider nonlinear PDEs:

- Polynomial: $F(u) = u^3$

$$\begin{aligned} F_k &= \tfrac{1}{2\pi} \quad \int u(x)^3 e^{-ikx}\, dx = \frac{1}{2\pi}\int dx\, e^{-ikx} \sum_{k_1} e^{ik_1x} u_{k_1} \sum_{k_2} e^{ik_2x} u_{k_2} \sum_{k_3} e^{ik_3x} u_{k_3} \\ &= \quad \sum_{k_1}\sum_{k_2} u_{k_1} u_{k_2} u_{k-k_1-k_2} \end{aligned}$$

  convolution requires N^2 multiplication of three numbers, compared to a *single* such multiplication
  for $r^{th}-$order polynomial need N^{r-1}operations: **slow!**

- General nonlinearities, e.g.
  coupled pendula

$$F(u) = \sin(u) = 1 - \frac{1}{3!}u^3 + \frac{1}{5!}u^5 + ...$$

  Arrhenius law in chemical reactions

$$F(u) = e^u = \sum_{l=0}^{\infty} \frac{1}{l!}\, u^l$$

  arbitrarily high powers of u, cannot use convolution

Evaluate nonlinearities in real space:

need to transform *efficiently* between real space and Fourier space

Discrete Fourier transformation:

Question: will we loose spectral accuracy with only $2N$ grid points in integral?

trapezoidal rule $\frac{1}{2}11111..11\frac{1}{2}$ with $2N$ collocation points

$$x_j = \frac{2\pi}{2N} j, \qquad \Delta x = \frac{2\pi}{2N}, \qquad x_{2N} = x_0$$

$$\begin{aligned} \tilde{u}_k &= \frac{1}{c_k}\frac{1}{2N}\left(\frac{1}{2}e^{-ikx_0}u(x_0) + \sum_{j=1}^{2N-1} e^{-ikx_j}u(x_j) + \frac{1}{2}e^{-ikx_{2N}}u(x_{2N})\right) \\ &\underbrace{=}_{\text{for periodic } u(x)} \frac{1}{c_k}\frac{1}{2N}\sum_{j=0}^{2N-1} e^{-ikx_j}u(x_j) \end{aligned}$$

High wavenumbers:

Calculate high wavenumber components

$$\begin{aligned}
\tilde{u}_{N+m} &= \frac{1}{2N}\sum_{j=0}^{2N-1} \underbrace{e^{-iN\frac{2\pi}{2N}j}}_{e^{-i\pi j}}\, e^{-imx_j}\, u(x_j) \\
&= \frac{1}{2N}\sum_{j=0}^{2N-1} e^{+i\pi j}\, e^{-imx_j}\, u(x_j) \\
&= \tilde{u}_{-N+m}
\end{aligned}$$

- thus: $\tilde{u}_N = \tilde{u}_{-N}$
- there are only $2N$ independent amplitudes
  $\Rightarrow$ limited range of relevant wave numbers: $-N \leq k \leq N$

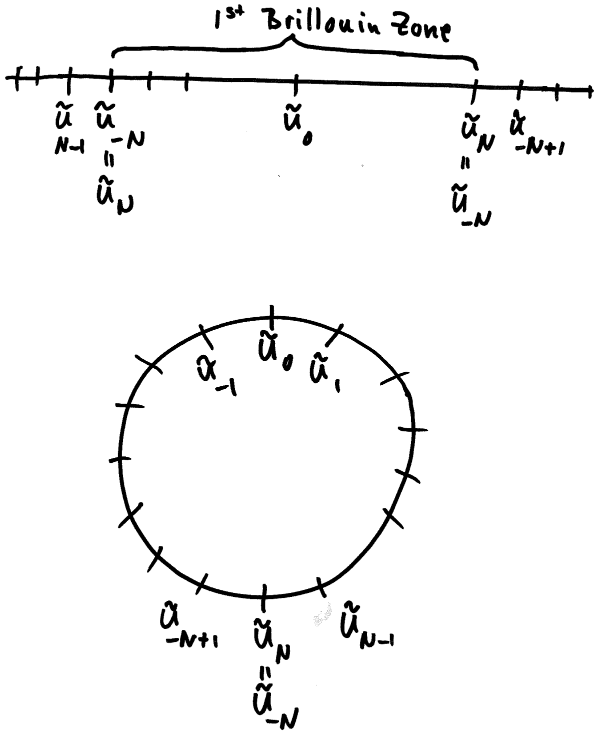


Figure 3: For a discrete spatial grid the Fourier space is periodic.
a) 1^{st} Brillouin zone, b) periodic representation of Fourier space.

- Fourier space is periodic ⇔ spatial grid is discrete rather than continuous
  This is the converse of the Fourier spectrum becoming discrete when the real space is made periodic (rather than infinite)
- Two possible treatments:
  1. restrict $-N \leq k \leq N-1$ (somewhat asymmetric)
     in Matlab: $(\tilde{u}_0, \tilde{u}_1, ...\tilde{u}_N, \tilde{u}_{-N+1}, \tilde{u}_{-N+2}, ..., \tilde{u}_{-1})$

2. in these notes we set

$$\tilde{u}_N = \tilde{u}_{-N} = \frac{1}{2}\frac{1}{2N}\sum_{j=0}^{2N-1} e^{iNx_j}u(x_j)$$

i.e.

$$c_N = c_{-N} = 2 \qquad \text{and} \qquad c_j = 1 \qquad \text{for} \qquad j \neq \pm N$$

Inverse Transformation

$$I_N(u(x_j)) = \sum_{k=-N}^{N} \tilde{u}_k e^{ikx_j}$$

Orthogonality:

$$< \phi_k, \phi_l >_N = \frac{1}{2N}\sum_{j=0}^{2N-1} e^{i(l-k)\frac{2\pi}{2N}j} = \sum_{l-k=-\infty}^{\infty} \delta_{l-k,2Nm} \tag{4}$$

Notation:
$< .,. >_N$ denotes the scalar product of functions defined only at N discrete points x_j

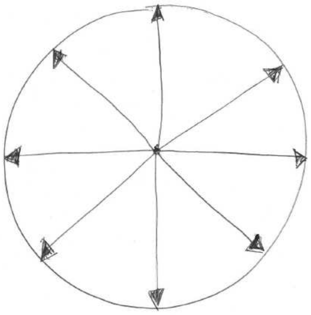


Figure 4: Cancellation of the Fourier modes in the sum. Here $N = 4$ and $l - k = 1$

Note:

- $< \phi_k, \phi_l >_N \neq 0$ if $k - l$ is *any* multiple of $2N$ and not only for $k = l$ (cf. completeness relation (2))
  high wavenumbers are not necessarily perpendicular to low wavenumbers

Interpolation property

Consider $I_N(u)$ on the grid

$$I_N(u(x_l)) \;\; = \;\; \sum_{k=-N}^{N} \tilde{u}_k e^{ikx_l}$$

$$= \sum_{k=-N}^{N} \frac{1}{2N}\frac{1}{c_k}\sum_{j=0}^{2N-1} e^{-ikx_j}u(x_j)e^{ikx_l} \qquad \text{interchange sums to get } \delta\text{-function}$$

$$= \frac{1}{2N}\sum_{j=0}^{2N-1} u(x_j) \sum_{r\equiv k+N=0}^{2N} e^{i(r-N)\frac{2\pi}{2N}(l-j)}\frac{1}{c_{r-N}}$$

in the r-sum: for $r = 2N$ we have $e^{i\pi(l-j)}\frac{1}{2}$ and for $r = 0$ we have $e^{-i\pi(l-j)}\frac{1}{2}$

⇒ using (4) the sum adds up to $2N\delta_{lj}e^{-i\pi(l-j)}$ (note that $|l-j| < 2N$)

Thus

$$I_N(u(x_l)) = \frac{1}{2N}\sum_{j=0}^{2N-1} u(x_j)\, 2N\delta_{jl} = u(x_l).$$

Notes:

- On the grid x_j the function $u(x)$ is represented *exactly* by $I_N(u(x))$; **no information lost on the grid**
- $I_N(u(x))$ is often called *Fourier interpolant*.

2.3.1 Aliasing

For the discrete Fourier transform the function is defined only on the grid:
what happens to the high wavenumbers that cannot be represented on that grid?

Consider $u(x) = e^{i(r+2N)x}$ with $0 < |r| < N$.

Continuous Fourier transform: $P_N u = 0$ since the wavenumber is higher than N.

Discrete Fourier transform:

$$u(x_j) = e^{i(2N+r)\frac{2\pi}{2N}j} = e^{ir\frac{2\pi}{2N}j} = e^{irx_j}$$

On the grid $u(x)$ looks like e^{irx}:

$$I_N(u(x_j)) \;\; = \;\; e^{irx_j} \neq 0$$

$u(x)$ is folded back into the 1^{st} Brillouin zone.

Notes:

- highest wavenumber that is resolvable on the grid: $|k| = N$

$$e^{\pm iN\frac{2\pi}{2N}j} = (-1)^j$$

- in CFT unresolved modes are set to 0
- in DFT unresolved modes modify the resolved modes: **Aliasing**

Relation between CFT (u_k) and DFT ($\tilde{u}_k$) coefficients:

$$\begin{aligned}
\tilde{u}_k &= \frac{1}{2N}\frac{1}{c_k}\sum_{j=0}^{2N-1} e^{-ikx_j}u(x_j) \\
&= \frac{1}{2N}\frac{1}{c_k}\sum_{l=-\infty}^{\infty}\sum_{j=0}^{2N-1} e^{i(l-k)\frac{2\pi}{2N}j}u_l \\
&= \frac{1}{c_k}\sum_{l=-\infty}^{\infty}\sum_{m=-\infty}^{\infty} \delta_{l-k,2Nm}u_l \\
\tilde{u}_k &= \frac{1}{c_k}u_k + \frac{1}{c_k}\sum_{|m|=1}^{\infty} u_{k+2Nm}
\end{aligned}$$

The sum contains the aliasing terms from higher harmonics that are not represented on the grid.

High wavenumbers look like low wavenumbers and contribute to low-k amplitudes

Error $||u - I_N u||^2$**:**

$$\begin{aligned}
I_N u &= \sum_{k=-N}^{N} \tilde{u}_k e^{ikx} = \sum_{k=-N}^{N}\left\{\frac{1}{c_k}u_k + \frac{1}{c_k}\sum_{|m|=1}^{\infty} u_{k+2Nm}\right\}e^{ikx} \\
&= P_N u + R_N u
\end{aligned}$$

$$||u-I_N u||^2 = ||\underbrace{u-P_N u}_{\text{all modes have } |k|>N} - \underbrace{R_N u}_{\text{all modes have } |k|\leq N}||^2 \underbrace{=}_{\text{orthogonality}} ||u-P_N u||^2 + ||R_N u||^2$$

Interpolation error is *larger* than projection error.

Decay of coefficients:

if CFT coefficients decay exponentially, $u_k \sim e^{-\alpha|k|}$, so will the DFT coefficients:

$$\tilde{u}_k \sim \frac{1}{c_k}e^{-\alpha|k|} + \frac{1}{c_k}\sum_{|m|=1}^{\infty} e^{-\alpha|k+2Nm|} \underbrace{\sim}_{\text{geometric series}} \sim \frac{1}{c_k}e^{-\alpha|k|} + \frac{1}{c_k}\frac{2e^{-2\alpha N}}{1-e^{-2\alpha N}} \qquad \text{for} \qquad k \ll N$$

Thus:

The asymptotic convergence properties of the DFT are essentially the same as those of the CFT ⇒ homework assignment

2.3.2 Differentiation

Main reason for spectral approach: derivatives

For CFT one has: projection and differentiation *commute* :

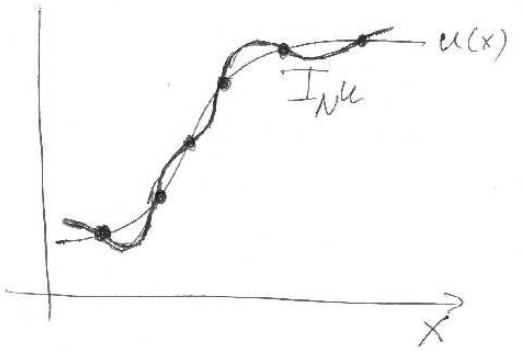
$$\begin{aligned}
\frac{d}{dx}(P_N u) &= \sum_{k=-N}^{N} iku_k e^{ikx} \\
P_N(\frac{du}{dx}) &= \sum_{k=-N}^{N} (\frac{du}{dx})_k e^{ikx} \\
&= \sum_{k=-N}^{N} \frac{1}{2\pi} \int e^{-ikx'} \frac{du}{dx'} dx' \, e^{ikx} \qquad \text{using i.b.p. :} \\
&= \sum_{k=-N}^{N} \frac{1}{2\pi} ik \int e^{-ikx'} u(x') dx' \, e^{ikx} \\
&= \frac{d}{dx}(P_N u)
\end{aligned}$$

For DFT interpolation and differentiation *do not commute:*

$$\frac{d}{dx}(I_N u) \neq I_N(\frac{du}{dx}).$$

i.e. $\frac{d}{dx}(I_N u)$ does not give the exact values of $\frac{du}{dx}$ on the grid points.

$I_N u$ does not agree with u between grid points $\Rightarrow$ its derivative does not agree with the derivative of u on the grid points, but $I_N(\frac{du}{dx})$ does interpolate $\frac{du}{dx}$.



Asymptotically, the errors of $I_n(\frac{du}{dx})$ and of $\frac{d}{dx}I_N(u)$ are of the same order.

Implementation of Discrete Fourier Transformation

Steps for calculating derivatives at a given point:

i) Transform method

1. calculate $\tilde{u}_k$ from values at collocation points x_j:

$$\tilde{u}_k = \frac{1}{2N}\frac{1}{c_k}\sum_{j=0}^{2N-1} e^{-ikx_j}u(x_j)$$

2. for $r^{th}-$derivative

$$\frac{d^r u}{dx^r} \Rightarrow (ik)^r \tilde{u}_k$$

3. back-transformation at collocation points

$$\frac{d^r}{dx^r}I_N(u(x_j)) = \sum_{k=-N}^{N}(ik)^r\tilde{u}_k e^{ikx_j}$$

Notes:

- seems to require $\mathcal{O}(N^2)$ operations
  compared to $\mathcal{O}(N)$ operations for finite differences
- for $N = 2^l 3^m 5^n ...$ DFT can be done in $\mathcal{O}(N\ln N)$ operations using **fast Fourier transform**[1]
- for u real: $\tilde{u}_k = \tilde{u}^*_{-k} \Rightarrow$ need to calculate only half the $\tilde{u}_k$:
  special FFT that stores the real data in a complex array of half size
  N independent variables: $\tilde{u}_0$ and $\tilde{u}_N$ real, $\tilde{u}_1$,...,$\tilde{u}_{N-1}$ complex

ii) Matrix multiplication method

$\frac{d^r}{dx^r}I_N(u)$ is linear in $u(x_j) \Rightarrow$ can write it as matrix multiplication

$$\begin{aligned}\frac{d^r}{dx^r}I_N(u(x_j)) &= \sum_{k=-N}^{N}(ik)^r\tilde{u}_k e^{ikx_j} \qquad \text{interchange sums}\\ &= \sum_{l=0}^{2N-1}\left(\sum_{k=-N}^{N}(ik)^r\frac{1}{2N}\frac{1}{c_k}e^{ik(x_j-x_l)}\right)u(x_l)\end{aligned}$$

write in terms of vectors and matrix

$$\begin{pmatrix} u(x_0) \\ ... \\ u(x_{2N-1})\end{pmatrix} = \mathbf{u} \qquad \frac{d^r}{dx^r}I_N(\mathbf{u}) = \begin{pmatrix} ... \\ u^{(r)}(x_j) \\ \\ ...\end{pmatrix}$$

Then first derivative

$$\mathbf{u}^{(1)} = \mathbf{D}\mathbf{u}$$

[1] In matlab functions FFT and IFFT.

with

$$D_{jl} = \frac{1}{2N} \sum_{k=-N}^{N} ik \frac{1}{c_k} e^{ik\frac{2\pi}{2N}(j-l)} = \begin{cases} \frac{1}{2}(-1)^{j+l} \cot(\frac{j-l}{2N}\pi) & \text{for} \quad j \neq l \\ 0 & \text{for} \quad j = l \end{cases}$$

Higher derivatives

$$\mathbf{u}^{(r)} = \mathbf{D}^r \mathbf{u}$$

Notes:

- **D** is $2N \times 2N$ matrix ($j, l = 0, ..., 2N-1$)
- **D** is anti-symmetric: $D_{lj} = -D_{jl}$
- matrix multiplication is expensive: N^2 operations
  but multiplication can be *vectorized*, i.e. different steps of multiplication/addition are done simultaneously for different numbers in the matrix

Eigenvalues of Pseudo-Spectral Derivative:

Fourier modes with $|k| \leq N-1$ are represented exactly

$$\mathbf{D}e^{ikx} = ik\, e^{ikx} \qquad \text{for} \qquad |k| \leq N-1$$

$\Rightarrow$ plane waves e^{ikx} must be eigenvectors with eigenvalues

$$\lambda_k = ik = 0, \pm 1i, \pm 2i, ..., \pm(N-1)i$$

D has $2N$ eigenvalues: one missing

$tr\mathbf{D} = 0 \Rightarrow \sum_k \lambda_k = 0 \Rightarrow$ last eigenvalue $\lambda_N = 0$

can see that also via: $e^{iN\frac{2\pi}{2N}j} = (-1)^j = e^{-iN\frac{2\pi}{2N}j} \Rightarrow$ eigenvalue must be independent of the sign of $N \Rightarrow \lambda_N = 0$

Interpretation: consider PDE

$$\frac{\partial u}{\partial t} = \frac{\partial u}{\partial x} \qquad \text{with} \qquad u = e^{i\omega t + ikx}$$

Frequency ω numerically determined by **Du**: $\omega = \lambda_k$

For $|k| \leq N-1$ the solution is a traveling wave with direction of propagation given by sign of k.

For $k = \pm N$ one has $u(x_j) = (-1)^j$: does not define a direction of propagation $\Rightarrow \omega \equiv \lambda_k = 0$.

Note:

One gets a vanishing eigenvalue also using the transform method:

$$(-1)^j = \tilde{u}_N e^{iN\frac{2\pi}{2N}j} + \tilde{u}_{-N} e^{-iN\frac{2\pi}{2N}j} \qquad \text{with } \tilde{u}_N = \tilde{u}_{-N}$$

thus

$$\frac{d}{dx} P_N \left((-1)^j\right) = iN\tilde{u}_N e^{iNx_j} + (-iN)\tilde{u}_{-N} e^{-iNx_j} = 0.$$

3 Fourier Methods for PDE: Continuous Time

Consider PDE

$$\frac{\partial u}{\partial t} = S(u) \equiv F(u, \frac{\partial u}{\partial x}, \frac{\partial^2 u}{\partial x^2}, ...)$$

The operator $S(u)$ can be nonlinear

Two methods

1. Pseudo-spectral:
$$u \Rightarrow I_N u$$
Spatial derivatives in Fourier space
Nonlinearities in real space
temporal evolution performed in real space or in Fourier space:
i.e. unknowns to be updated are the $u(x_j)$ in real space or the $\tilde{u}_k$ in Fourier space

2. Galerkin method
$$u \Rightarrow P_N u$$
completely in Fourier space: spatial derivatives, nonlinearities and temporal updating are all done in Fourier space

3.1 Pseudo-spectral Method

Method involves the steps

1. introduce collocation points x_j and $u(x_j)$

2. transfrom numerical solution $u(x_j) \Rightarrow \tilde{u}_k$ to Fourier space

3. evaluate derivatives using $\tilde{u}_k$

4. transform back into real space and evaluate nonlinearities

5. evolve in time either in real space or in Fourier space
$$\frac{d}{dt} I_N(u) = S(I_N(u))$$

Note:

$I_N(u)$ is not the spectral interpolant of the exact solution u since solving PDE induces errors:

1. taking the spectral interpolant of the exact solution $\hat{u}$ yields

$$I_N\left(\frac{d}{dt}\hat{u}\right) = I_N\left(S(\hat{u})\right).$$

Using

$$\frac{d}{dt}I_N(u) = I_N\left(\frac{d}{dt}u\right)$$

the pseudospectral solution satisfies

$$I_N\left(\frac{d}{dt}u\right) = S(I_N(u)) \neq I_N\left(S(u)\right)$$

since spatial derivative does not commute with I_N

2. time-stepping introduces errors beyond the spectral approximation.

Examples:

1. Wave equation

$$\partial_t u = \partial_x u$$

a) Using FFT

$$\partial_t u(x_j) = \partial_x I_N(u(x_j)) = \sum_{k=-N}^{N} ik\tilde{u}_k e^{ikx_j}$$

Note: $\tilde{u}_k$ and the sum over k (=back-transformation) are evaluated via two FFTs.
b) Using multiplication with spectral differentiation matrix D,

$$\partial_t u(x_j) = \sum_l D_{jl}u(x_l)$$

2. Variable coefficients

$$\partial_t u = c(x)\partial_x u$$

a)

$$\partial_t u(x_j) = c(x_j)\,\partial_x I_N(u(x_j))$$

multiply by wave speed in real space
b)

$$\partial_t u(x_j) = c(x_j)\sum_m D_{jm}u(x_m).$$

3. Reaction-diffusion equation

$$\partial_t u = \partial_x^2 u + f(u)$$

a) using FFT

$$\partial_t u(x_j) = \partial_x^2 I_N(u(x_j)) + f(u(x_j)) = -\sum_{k=-N}^{N} k^2\tilde{u}_k e^{ikx_j} + f(u(x_j))$$

b) matrix multiplication

$$\partial_t u(x_j) \quad = \quad \sum_m D^{(2)}_{jm}u(x_m) + f(u(x_j)) \qquad \text{with} \qquad D^{(2)}_{jm} = \sum_l D_{jl}D_{lm}.$$

4. Burgers equation

$$\begin{aligned} \partial_t u &= u\partial_x u \\ &= \frac{1}{2}\partial_x(u^2) \qquad \text{in conservation form} \end{aligned}$$

consider both types of nonlinearities[2] $\alpha u\partial_x u + \beta\partial_x(u^2)$
a)

$$\alpha u(x_j)\partial_x I_N(u(x_j)) = \alpha u(x_j)\sum_{k=-N}^{N} ik\,\tilde{u}_k e^{ikx_j}$$

$$\begin{aligned} \beta\,\partial_x I_N(u^2(x_j)) &= \beta \sum_{k=-N}^{N} ik\,\tilde{w}_k e^{ikx_j} \\ \tilde{w}_k &= \frac{1}{2N}\sum_{j=0}^{2N-1} e^{-ikx_j}\,u^2(x_j) \end{aligned}$$

b)

$$\partial_t u(x_j) = \alpha\,u(x)\,\mathbf{D}u + \beta\mathbf{D}\begin{pmatrix} u(x_0)^2 \\ \dots \\ u(x_{2N-1})^2 \end{pmatrix}$$

Notes:

- spectral methods will lead to Gibbs oscillations near the shock
- pseudo-spectral methods: on the grid the oscillations may not be visible; may need to plot function between grid points as well, but derivatives show oscillations
- all sums over Fourier modes k or grid points j should be done via FFT.

3.2 Galerkin Method

Equation solved completely in Fourier space

1. plug

$$u(x) = \sum_{k=-N}^{N} u_k e^{ikx}$$

into $\partial_t u = S(u)$

2. project equation onto first $2N$ Fourier modes ($-N \le l \le N$)

$$\partial_t u_l \equiv \frac{1}{2\pi}\int_0^{2\pi} e^{-ilx}\partial_t u(x)\,dx = \frac{1}{2\pi}\int_0^{2\pi} e^{-ilx}\,S(u(x))\,dx$$

[2] **Note:** For smooth functions the two formulations are equivalent.Burgers equation develops *shocks* at which the solution becomes discontinuous: formulations not equivalent, need to satisfy entropy condition, which corresponds to adding a viscous term $\nu\partial_x^2 u$ and letting $\nu \to 0$.

More generally, retaining N modes from a complete set of functions $\{\phi_k(x)\}$

$$\begin{aligned} u(x) &= \sum_{k=1}^{N} u_k \phi_k(x) \\ < \phi_l, \partial_t u > &= < \phi_l, S(u) > \qquad \text{for} \qquad 1 \leq l \leq N \\ < \phi_l, \partial_t u - S(u) > &= 0 \end{aligned}$$

Residual (=error) $\partial_t u - S(u)$ has to be orthogonal to all basis functions that were kept:

$$P_N \left(\partial_t P_N u - S(P_N u) \right) = 0$$

optimal choice within the space of N modes that is used in the expansion

Note: for Galerkin the integrals are calculated exactly either analytically or numerically with *sufficient resolution* (number of grid points →∞)

Examples:

1. Variable-coefficient wave equation

$$\partial_t u = c(x)\, \partial_x u$$

$$\begin{aligned} \partial_t u_m &= \int_0^{2\pi} e^{-imx} c(x) \sum_{k=-N}^{N} ik\, u_k e^{ikx} dx \\ &= \sum_{k=-N}^{N} C_{mk}\, iku_k \\ C_{mk} &= \int_0^{2\pi} e^{i(k-m)x} c(x) dx \end{aligned}$$

Note: although equation is linear, there are $\mathcal{O}(N^2)$ operations through variable coefficient (C_{mk} is in general not diagonal).

2. Burgers equation

$$\partial_t u = \alpha u \partial_x u + \beta \partial_x (u^2)$$

$$\begin{aligned} \alpha u \partial_x u &= \alpha \sum_{k=-N}^{N} \sum_{l=-N}^{N} u_k\, il u_l\, e^{i(k+l)x} \\ \beta \partial_x u^2 &= \beta \sum_{k=-N}^{N} \sum_{l=-N}^{N} i(k+l)\, u_k u_l\, e^{i(k+l)x} \end{aligned}$$

project onto e^{-imx} ⇒integral gives $\delta_{k+l,m}$ and $\sum_l$ yields $l \Rightarrow m-k$

$$\partial_t u_m = \sum_{k=-N}^{N} i(\alpha(m-k) + \beta m) u_k u_{m-k} \tag{5}$$

Note: again $\mathcal{O}(N^2)$ operations in each time step.

Comparison:

- Nonlinear problems:
  Galerkin: effort increases with degree of nonlinearity because of convolution
  pseudo-spectral: effort mostly in transformation to and from Fourier space: FFT essential
- Variable coefficients:
  Galerkin requires matrix multiplication, pseudospectral only scalar multiplication
- error larger in pseudo-spectral, but same scaling of error with N
- Unresolved modes:
  Pseudo-spectral has aliasing errors: unresolved modes spill into equations for resolved modes
  Nonlinearities generate high-wavenumber modes: their aliasing can be removed by taking more grid points ($\frac{3}{2}-$rule) or by phase shifts
- Grid effects:
  pseudo-spectral method breaks the translation symmetry, can lead to pinning of fronts
  Galerkin method does not break translation symmetry.
- Newton method for unstable fixed points or implicit time stepping:
  quite clear for Galerkin code: (5) is simply a set of coupled ODEs, not so obvious to implement for pseudo-spectral code, since back- and forth-transformations are needed.

4 Temporal Discretization

Consider

$$\partial_t u = S(u)$$

Two possible goals:

1. interested in steady state: transient towards steady state not relevant
   only spatial resolution relevant
2. initial-value problem: interested in complete evolution
   temporal error has to be kept as small as spatial error

If transient evolution is relevant then spectral accuracy in space best exploited if high temporal accuracy is obtained as well: seek high-order temporal schemes

4.1 Review of Stability

Consider ODE

$$\partial_t u = \lambda u \tag{6}$$

Definitions:

1. A scheme is *stable* if there are constants C, α, T, and δ such

   $$||u(t)|| \leq Ce^{\alpha t}||u(0)||$$

   for all $0 \leq t \leq T$, $0 < \Delta t < \delta$. The constants C and α have to be independent of Δt.

2. A scheme is *absolutely stable* if

   $$||u(t)|| < \infty \qquad \text{for all } t.$$

   Note:

   - The concept of absolute stability is only useful for differential equations for which the exact solution is bounded for all times.
   - absolute stability closely related to Neumann stability

3. The *region A of absolute stability* is given by the region A the complex plane defined by

   $$A = \{\lambda\Delta t \in C \,|\, ||u(t)|| \text{ bounded for all } t\}$$

Notes:

- for $\lambda \in R$ the ODE (6) corresponds to a parabolic equation like $\partial_t u = \partial_x^2 u$ in Fourier space
- for $\lambda \in iR$ the ODE (6) corresponds to a hyperbolic equation like $\partial_t u = \partial_x u$ in Fourier space

For a PDE one can think in terms of a system of ODEs coupled through differentiation matrices,

$$\partial_t \mathbf{u} = L\mathbf{u}$$

e.g. for $\partial_t u = \partial_x u$ one has $L = D$.

Assume L can be diagonalized

$$SLS^{-1} = \Lambda \qquad \text{with } \Lambda \text{ diagonal}$$

Then

$$\partial_t S\mathbf{u} = \Lambda S\mathbf{u}$$

Thus:
Stability requires that all eigenvalues λ of L are in the region of absolute stability of the scheme.

Note:

- highest Fourier eigenvalues
  - for simple wave equation: $\lambda_{max} = \pm i\,(N-1)$
  - for diffusion equation: $\lambda_{max} = -N^2$

Side Remark: Stability condition after diagonalization in terms of $S\mathbf{u}$,

$$||S\mathbf{u}(t)|| < Ce^{\alpha t}||S\mathbf{u}(0)||$$

We need

$$||\mathbf{u}(t)|| < \tilde{C}e^{\alpha t}||\mathbf{u}(0)||$$

If S is unitary, i.e. if $S^{-1} = S^{+}$we have

$$||S\mathbf{u}|| = ||u||$$

For Fourier modes spectral differentiation matrix is *normal*

$$D^{+}D = DD^{+}$$

$\Rightarrow$ D can be diagonalized by unitary matrix

(Not the case for Chebyshev basis functions used later)

Thus: for Fourier method it is sufficient to consider scalar equation (6).

4.2 Adams-Bashforth Methods

Based on rewriting in terms of integral equation

$$u^{n+1} = u^n + \int_{t_n}^{t_{n+1}} F(t', u(t'))dt'$$

Explicit method: approximate $F(u)$ by polynomial that interpolates $F(u)$ over last l time steps[3] and extrapolate to the interval $[t_n, t_{n+1}]$.

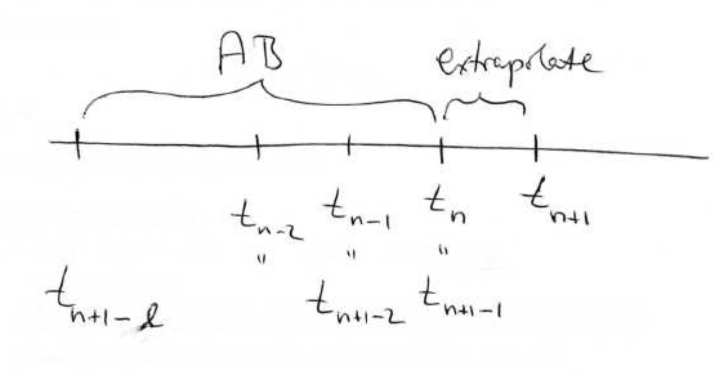


Figure 5: Adams-Bashforth methods interpolate $F(u)$ over the interval $[t_{n-l}, t_n]$ and then extrapolate to the interval $[t_n, t_{n+1}]$.

[3] Figure has wrong label for first grid point.

Consider

$$\partial_t u = F(u)$$

$$\begin{aligned}
\textbf{AB1:} \qquad u^{n+1} &= u^n + \Delta t F(u^n) \\
\textbf{AB2:} \qquad u^{n+1} &= u^n + \Delta t \left(\frac{3}{2} F(u^n) - \frac{1}{2} F(u^{n-1}) \right)
\end{aligned}$$

Note:

- AB1 identical to forward Euler

Stability:

Consider $F(u) = \lambda u$ with $\lambda \in C$

AB1:

$$z = 1 + \Delta t \lambda$$

$$|z|^2 = (1 + \lambda_r \Delta t)^2 + \lambda_i^2 \Delta t^2$$

Stability limit given by $|z|^2 = 1$:

$$\textbf{AB1=FE:} \qquad (1 + \lambda_r \Delta t)^2 + \lambda_i^2 \Delta t^2 = 1$$

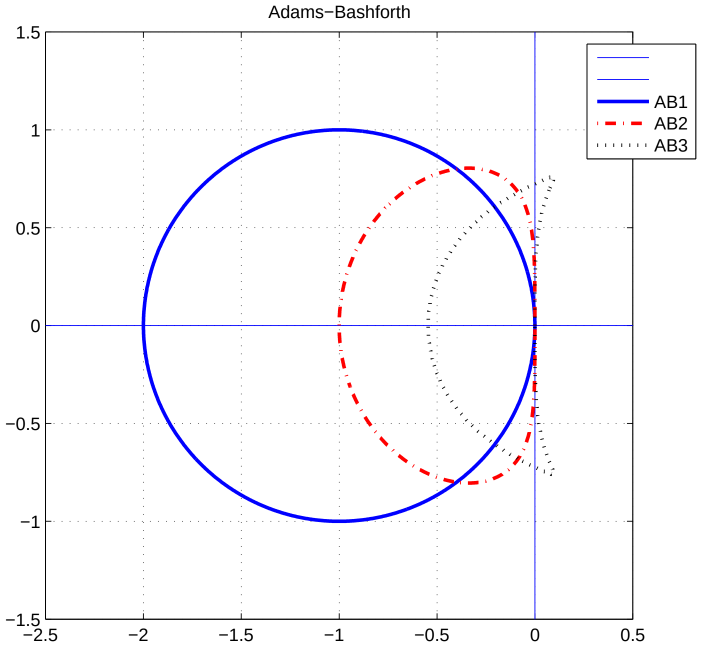
To plot stability limit parametrize $z = e^{i\theta}$ and plot $\lambda\,\Delta t \equiv (\lambda_r(\theta) + i\lambda_i(\theta))\Delta t$

AB1:

$$\lambda\,\Delta t = z - 1$$

AB2:

$$\lambda\,\Delta t \quad = \quad \frac{z-1}{\frac{3}{2} - \frac{1}{2z}}$$



Notes:

- AB1=FE and AB2 are not absolutely stable for purely dispersive equations $\lambda_r = 0$
- AB3 and AB4 are absolutely stable even for dispersive equations $\lambda_r = 0$
- AB1 and AB2: the stability limit is tangential to $\lambda_r = 0$: for $\lambda_r = 0$ exponential growth rate goes to 0 for $\Delta t \to 0$ at fixed number of modes (i.e. fixed λ). For fixed t_{max} we can choose Δt small enough to limit the growth of solution.

$$\begin{array}{lrcl}
\text{AB1:} & \text{for } \lambda_r & = & 0 \qquad |z|^2 = 1 + \lambda_i^2 \Delta t^2 \\
& |z|^{\frac{t_{max}}{\Delta t}} & = & (1 + \lambda_i^2 \Delta t^2)^{\frac{1}{2}\frac{t_{max}}{\Delta t}} \leq e^{\frac{1}{2}\lambda_i^2 \Delta t^2 \frac{t_{max}}{\Delta t}} \qquad \text{need} \qquad \Delta t \ll \mathcal{O}(\lambda_i^{-2})
\end{array}$$

for simple wave equation one has then

$$\Delta t \ll \mathcal{O}(N^{-2})$$

i.e. AB1 is stable for 'diffusive scaling'

$$\begin{array}{lrcl}
\text{AB2:} & \text{for } \lambda_r & = & 0 \qquad z = 1 + i\lambda_i \Delta t - \frac{1}{2}\lambda_i^2 \Delta t^2 + \frac{1}{4}\lambda_i^3 \Delta t^3 - \frac{1}{8}\lambda_i^4 \Delta t^4 \\
& |z|^2 & = & 1 + \frac{1}{2}\lambda_i^4 \Delta t^4 \\
& |z|^{\frac{t_{max}}{\Delta t}} & \leq & e^{\frac{1}{4}\lambda_i^4 \Delta t^4 \frac{t_{max}}{\Delta t}} \qquad \text{need} \qquad \Delta t \ll \mathcal{O}(\lambda_i^{-\frac{4}{3}}) = \mathcal{O}(N^{-\frac{4}{3}})
\end{array}$$

For simple wave equation one gets

$$\Delta t \ll \mathcal{O}(N^{-\frac{4}{3}})$$

which is less stringent than AB1=FE.
The growth may be less of a problem for spectral methods since one would like to balance the temporal error with the spatial error

$$\Delta t^p \sim e^{-\alpha N}$$

one may have to choose therefore quite small Δt just to achieve the desired accuracy, independent of the stability condition.
But: growth rate is largest for largest wavenumbers k: high Fourier modes tend to 'creep in'.

- Diffusion equation: FE stability limit for $\lambda_i = 0$ and $\lambda_r = -k^2 < 0$:

$$\Delta t < \frac{2}{|\lambda_r|} = \frac{2}{k_{max}^2} = \frac{2}{N^2}$$

for central difference scheme

$$\Delta t < \frac{1}{2}\Delta x^2 = \frac{1}{2}\left(\frac{2\pi}{2N}\right)^2 \approx \frac{5}{N^2}$$

The scaling of stability limit is the same, but finite-difference scheme has slightly larger prefactor, i.e. it has a slightly larger stability range. But it needs smaller Δx to achieve the same spatial accuracy.

Comment on Implementation

Consider

$$\partial_t u = \partial_x^2 u + f(u)$$

Forward Euler

$$u^{n+1} = u^n + \Delta t\, \partial_x^2 u^n + \Delta t\, f(u^n)$$

Want to evaluate derivative in Fourier space ⇒ FFT

1. If we do the temporal update in Fourier space

   $$\tilde{u}_k^{n+1} = \tilde{u}_k^n + \Delta t(-k^2)\tilde{u}_k^n + \Delta t\, \mathcal{F}_k(f(u^n))$$

   where $\mathcal{F}_k(f(u^n))$ is the k^{th}-mode of the Fourier transform of $f(u^n)$
   After updating $\tilde{u}_k^{n+1}$ transform back to $u^{n+1}(x_j)$ and calculate $f(u_j^{n+1})$ for next Euler step.

2. If we do the temporal update in real space
   First transform back into real space and do time the step there

   $$u_j^{n+1} = u_j^n + \Delta t \partial_x^2 I_N(u) + \Delta t\, f(u_j)$$

   Note: the choice between these two types of updates is quite common, not only in forward Euler.

4.3 Adams-Moulton-Methods

seek highly stable schemes: implicit scheme

→ in the polynomial interpolation of $F(u)$ for the integral in

$$u^{n+1} = u^n + \int_{t_n}^{t_{n+1}} F(t', u(t'))dt' \tag{7}$$

include t_{n+1}. This makes the scheme implicit.

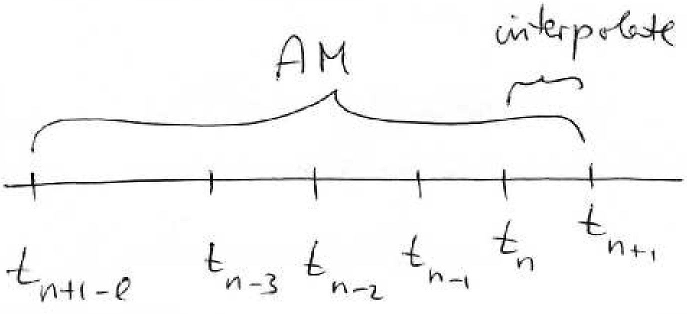
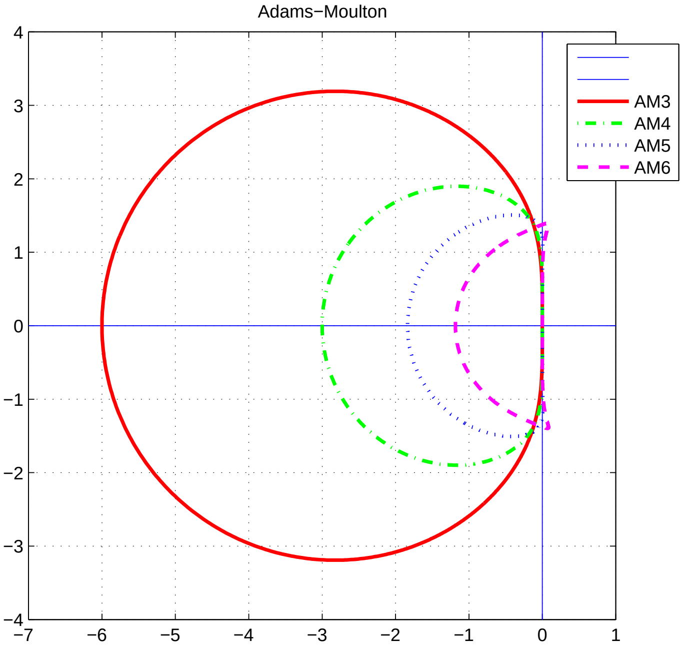


Figure 6: Adams-Moulton methods interpolate $F(u)$ over the interval $[t_{n+1-l}, t_{n+1}]$, which includes the new time step.

$$\begin{aligned} \text{Backwards Euler :} \qquad u^{n+1} &= u^n + \Delta t F(u^{n+1}) \\ \text{Crank-Nicholson :} \qquad u^{n+1} &= u^n + \frac{1}{2}\Delta t \left(F(u^{n+1}) + F(u^n)\right) \\ 3^{rd} \text{ order Adams-Moulton:} \qquad u^{n+1} &= u^n + \frac{1}{12}\Delta t \left(5F(u^{n+1}) + 8F(u^n) - F(u^{n-1})\right) \end{aligned}$$



Note:

- Region of stability *shrinks* with increasing order
- Only backward Euler and Crank-Nicholson are *unconditionally stable*
- AM3 and higher have *finite stability limit:* we do not get a high-order unconditionally stable schem with AM.

Implementation of Crank-Nicholson

Consider the wave equation

$$\partial_t u = \partial_x u$$

$$\left(1 - \frac{1}{2}\Delta t\,\partial_x\right) u^{n+1} = \left(1 + \frac{1}{2}\Delta t\,\partial_x\right) u^n$$

With matrix multiply method

$$\sum_l \left(1 - \frac{1}{2}\Delta t D_{jl}\right) u^{n+1}(x_l) = \sum_l \left(1 + \frac{1}{2}\Delta t D_{jl}\right) u^n(x_l)$$

would have to *invert full matrix*: slow

With FFT or for Galerkin insert $u(x)=\sum_k e^{ikx}\tilde{u}_k$ and project equation onto ϕ_k: $\int_0^{2\pi} dx\, e^{-ikx}$...

$$\begin{aligned}\left(1-\frac{1}{2}\Delta t\, ik\right)\tilde{u}_k^{n+1} &= \left(1+\frac{1}{2}\Delta t\, ik\right)\tilde{u}_k^{n}\\ \tilde{u}_k^{n+1} &= \frac{1+\frac{1}{2}\Delta t\, ik}{1-\frac{1}{2}\Delta t\, ik}\,\tilde{u}_k^{n}\end{aligned}$$

Note:

- Since derivative operator is diagonal in Fourier space, inversion of operator on l.h.s. is simple:
  time-stepping in *Fourier space* yields **explicit code although implicit scheme**.
  This is *not possible for finite differences*.
- With variable wave speed one would have

  $$\left(1-\frac{1}{2}\Delta t\, c(x)\,\partial_x\right)u^{n+1}=\left(1+\frac{1}{2}\Delta t\, c(x)\,\partial_x\right)u^{n}$$

  ⇒FFT does not lead to diagonal form: wavenumbers of $u(x)$ and of $c(x)$ couple
  ⇒projection leads to convolution of $c(x)$ and $\partial_x u^{n+1}$: expensive
- The scheme does not get more involved in higher dimensions
  e.g. for diffusion equation in two dimensions

  $$\partial_t u=\nabla^2 u$$

  one gets

  $$\tilde{u}_{kl}^{n+1}=\frac{1-\Delta t\,(k^2+l^2)}{1+\Delta t(k^2+l^2)}\,\tilde{u}_{kl}^{n}$$

  That is to be compared with the case of finite differences where implicit schemes in higher dimensions become much slower since the band width of the matrix becomes large ($\mathcal{O}(N)$ in two dimensions, worse yet in higher dimensions).

Note:

- make scheme explicit by combining Adams-Moulton with Adams-Bashforth to predictor-corrector
  replace the unknown u^{n+1} in the integrand of (7) of the AM-scheme by an estimate based on AB, which can be lower order than the AM-scheme:

  $$\left.\begin{array}{ll}\text{AB: predictor} & \mathcal{O}(\Delta t^{n-1})\\ \text{AM: corrector} & \mathcal{O}(\Delta t^{n})\end{array}\right\}\Rightarrow\mathcal{O}(\Delta t^{n})$$

  each time step requires *two* evaluations of r.h.s ⇒not worth if expensive
  Advantage: scheme has same accuracy as AB of $\mathcal{O}(\Delta t^{n})$ but greater range of stability with same storage requirements

4.4 Semi-Implicit Schemes

Often time step is limited by instabilities due to linear derivative terms but not due to nonlinear terms:

Treat

- linear derivative terms implicitly
- nonlinear terms explicitly

Note: implicit treatment of nonlinear terms would require matrix inversion at each time step

Example: Crank-Nicholson-Adams-Bashforth (CNAB)

Consider

$$\partial_t u = \partial_x^2 u + f(u)$$

$$\frac{u^{n+1} - u^n}{\Delta t} = \frac{1}{2}\partial_x^2 u^{n+1} + \frac{1}{2}\partial_x^2 u^n + \frac{3}{2} f(u^{n+1}) - \frac{1}{2} f(u^n) + \mathcal{O}(\Delta t^3)$$

$$\left(1 - \frac{1}{2}\Delta t D^2\right) u^{n+1} = \left(1 + \frac{1}{2}\Delta t D^2\right) u^n + \Delta t \left(\frac{3}{2} f(u^{n+1}) - \frac{1}{2} f(u^n)\right)$$

3 Steps:

- FFT $\mathcal{F}$ of r.h.s.
- divide by $(1 + \frac{1}{2}\Delta t k^2)$
- do inverse FFT of r.h.s. $\Rightarrow u_j^{n+1}$

$$u_j^{n+1} = \mathcal{F}^{-1}\left(\frac{1}{1 + \frac{1}{2}\Delta t k^2}\left\{\left(1 - \frac{1}{2}\Delta t\, k^2\right)\mathcal{F}(u_i^n) + \Delta t \mathcal{F}\left(\frac{3}{2} f(u_i^n) - \frac{1}{2} f(u_i^{n-1})\right)\right\}\right)$$

or written as

$$\tilde{u}_k^{n+1} = \frac{1}{1 + \frac{1}{2}\Delta t k^2}\left\{\left(1 - \frac{1}{2}\Delta t\, k^2\right)\mathcal{F}(u_i^n) + \Delta t\left(\frac{3}{2} f_k(u_i^n) - \frac{1}{2} f_k(u_i^{n-1})\right)\right\}$$

4.5 Runge-Kutta Methods

Runge-Kutta methods can be considered as approximations for the integral equation

$$u^{n+1} = u^n + \int_{t_n}^{t_{n+1}} F(t', u(t'))dt'$$

with approximation of F based purely on times $t' \in [t_n, t_{n+1}]$.

Runge-Kutta 2:

trapezoidal rule for integral

$$\int_{t_n}^{t_{n+1}} F(t', u(t'))dt' = \frac{1}{2}\Delta t \left(F(t_n, u^n) + F(t_{n+1}, u^{n+1})\right) + \mathcal{O}(\Delta t^3)$$

approximate u^{n+1} with forward Euler (its error contributes to the error in the overall scheme at $\mathcal{O}(\Delta t^3)$.

Improved Euler method (Heun's method)

$$\begin{aligned} k_1 &= F(t_n, u^n) \\ k_2 &= F(t_n + \Delta t, u^n + \Delta t\, k_1) \\ u^{n+1} &= u^n + \frac{1}{2}\Delta t\,(k_1 + k_2) + \mathcal{O}(\Delta t^3) \end{aligned}$$

Other version : mid-point rule ⇒ modified Euler:

$$u^{n+1} = u^n + \Delta t F\left(t + \frac{1}{2}\Delta t, u^n + \frac{1}{2}\Delta t F(t_n, u^n)\right)$$

Note:

- Runge-Kutta methods of a given order are not unique (usually free parameters)

General Runge-Kutta scheme:

$$\begin{aligned} u^{n+1} &= u^n + \Delta t \sum_{l=0}^{s} \gamma_l F_l \\ F_0 &= F(t_n, u^n) \\ F_l &= F(t_n + \alpha_l \Delta t, u^n + \Delta t \sum_{m=0}^{l} \beta_{lm} F_m) \qquad 1 \leq l \leq s \end{aligned}$$

Notes:

- Scheme has $s+1$ *stages*
- $F(u)$ is evaluated at intermediate times $t_n + \alpha_l \Delta t$ and at suitably chosen intermediate values of the function u.
- For $\beta_{ll} \neq 0$ scheme is implicit
- Coefficients α_l, β_{lm}, γ_l determined by requiring highest order of accuracy:
in general this does not determine the coefficients uniquely

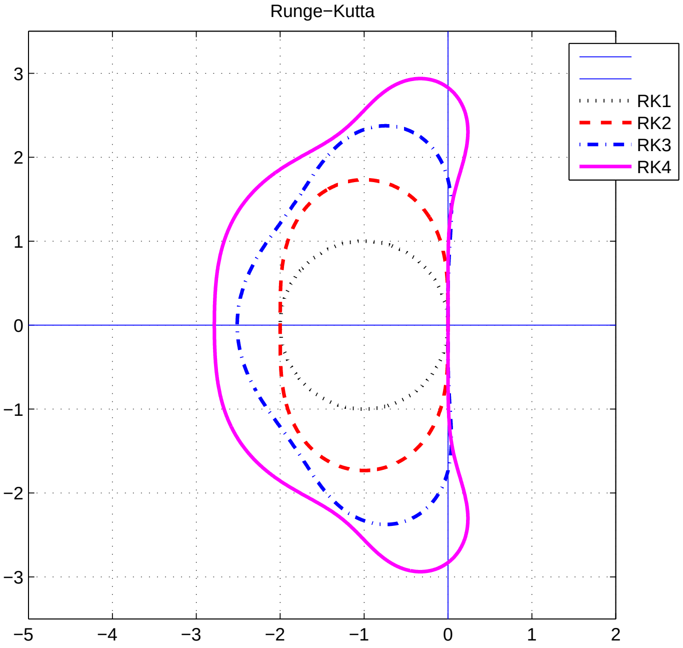
Runge-Kutta 4

corresponds to Simpson's rule ($\frac{1}{6}(1\ 4\ 1)$)

$$\begin{aligned}
k_1 &= F(t_n, u^n) \\
k_2 &= F(t_n + \frac{1}{2}\Delta t, u^n + \frac{1}{2}\Delta t\ k_1) \\
k_3 &= F(t_n + \frac{1}{2}\Delta t, u^n + \frac{1}{2}\Delta t\ k_2) \\
k_4 &= F(t_n + \Delta t, u^n + \Delta t\ k_3) \\
u^{n+1} &= u^n + \frac{1}{6}\Delta t\,(k_1 + 2k_2 + 2k_3 + k_4) + \mathcal{O}(\Delta t)^5
\end{aligned}$$

Note:

- to push the error to $\mathcal{O}(\Delta t^5)$ the middle term in Simpson's rule has to be split up into two different terms.



Notes:

- stability regions expand with increasing order
- RK4 covers parts of imaginary and of real axis: suited for *parabolic* and *hyperbolic* problems

4.6 Operator Splitting

For linear wave equation or diffusion equation we have exact solution in Fourier space,

$$\partial_t u = \partial_x^2 u \qquad \Rightarrow \qquad \tilde{u}_k^n = \tilde{u}_k(0)\, e^{-k^2 t_n}$$

Can we make use of that for more general problems?
For finite differences we discussed

$$\partial_t u = (L_1 + L_2) u$$

solution approximated as

$$\begin{aligned} u^{n+1} &= e^{(L_1+L_2)\Delta t} u^n \\ &= e^{L_1 \Delta t} e^{L_2 \Delta t} u^n + \mathcal{O}(\Delta t^2) \end{aligned}$$

this corresponds to

$$\partial_t u = L_2 u \qquad \text{and then} \qquad \partial_t u = L_1 u$$

alternating integration of each equation for a *full* time step Δt
Apply to reaction-diffusion equation

$$\begin{aligned} \partial_t u &= \partial_x^2 u + f(u) \\ & \quad L_1 u \sim \partial_x^2 u \qquad L_2 u \sim f(u) \end{aligned}$$

Treat $L_2 u$ in real space, e.g. forward Euler

$$u^*(x_j) = u^n(x_j) + \Delta t\, f(u^n(x_j))$$

Treat $L_1 u$ in Fourier space

$$\tilde{u}_k^{n+1} = e^{-k^2 \Delta t} \tilde{u}_k^* \qquad \textbf{exact!!}$$

Written together**:**

$$\tilde{u}_k^{n+1} = e^{-k^2 \Delta t} \left(u_k^n + \Delta t\, f_k(u_l^n) \right)$$

Notes:

- could use any other suitable time-stepping scheme for nonlinear term: higher-order would be better
- **But:** operator splitting error arises.
  Could improve
  $$e^{(L_1+L_2)\Delta t} u^n = e^{\frac{1}{2} L_1 \Delta t} e^{L_2 \Delta t} e^{\frac{1}{2} L_1 \Delta t} u^n + \mathcal{O}(\Delta t^3)$$
  If intermediate values need not be available the $\frac{1}{2}\Delta t-$steps can be combined:
  $$\begin{aligned} u^{n+2} &= e^{\frac{1}{2} L_1 \Delta t} e^{L_2 \Delta t} e^{\frac{1}{2} L_1 \Delta t} e^{\frac{1}{2} L_1 \Delta t} e^{L_2 \Delta t} e^{\frac{1}{2} L_1 \Delta t} u^n + \mathcal{O}(\Delta t^3) = \\ &= e^{\frac{1}{2} L_1 \Delta t} e^{L_2 \Delta t} e^{L_1 \Delta t} e^{L_2 \Delta t} e^{\frac{1}{2} L_1 \Delta t} u^n + \mathcal{O}(\Delta t^3) \end{aligned}$$
  approximate $e^{L_2 \Delta t}$ by second-order scheme (rather than forward Euler) to get over-all error of $\mathcal{O}(\Delta t^3)$.
- time-stepping is done in real space *and* in Fourier space
- to get higher order one would have to push the operator splitting error to higher order.

4.7 Exponential Time Differencing and Integrating Factor Scheme

Can we avoid the operator-splitting error altogether?

Consider again reaction-diffusion equation

$$\partial_t u = \partial_x^2 u + f(u)$$

without reaction the equation can be integrated exactly in Fourier space

$$u_k^{n+1} = e^{-k^2\Delta t} u_k^n$$

Go to Fourier space ('Galerkin style')

$$\partial_t u_k = -k^2 u_k + f_k(u) \tag{8}$$

Here $f_k(u)$ is $k-$component of Fourier transform of nonlinear term $f(u)$

To assess a good approach to solve (8) it is good to consider simpler problem yet:

$$\partial_t u = \lambda u + F(t) \tag{9}$$

where u is the Fourier mode in question and F plays the role of the coupling to the other Fourier modes.

We are in particular interested in efficient ways to deal with the fast modes with large, positive λ because they set the stability limit:

1. If the overall solution evolves on the fast time scale set by λ, accuracy requires a time step with $|\lambda \Delta t| \ll 1$ and an explicit scheme should be adequate.
2. If the overall solution evolves on a slower time scale $\tau \gg 1/|\lambda|$, which is set by Fourier modes with smaller wavenumber (i.e. $F(t)$evolves slowly in time) then one would like to take time steps with $|\lambda|\Delta t = \mathcal{O}(1)$ or even larger without sacrificing accuracy, i.e. one would like to be limited only by the condition $\Delta t \ll \tau$.
   In particular, for $F = const.$ one would like to obtain the exact solution $u^{\infty}_{exact} = -F/\lambda$ with large time steps.

Use integrating factor to rewrite (9) as

$$\partial_t \left(u e^{-\lambda t} \right) = e^{-\lambda t} F(t)$$

which is equivalent to

$$u^{n+1} = e^{\lambda \Delta t} u^n + e^{\lambda \Delta t} \int_0^{\Delta t} e^{-\lambda t'} F(t+t') dt'.$$

Need to approximate integral. To leading order it is tempting to write

$$u^{n+1} = e^{\lambda \Delta t} u^n + e^{\lambda \Delta t} \Delta t \, F(t).$$

This yields the forward Euler implementation of the *integrating-factor scheme*.

For $F = const.$ this yields the fixed point

$$u^{\infty}_{IF} \left(1 - e^{\lambda \Delta t} \right) = \Delta t \, e^{\lambda \Delta t} \, F.$$

But:

- for $-\lambda\Delta t \gg 1$ one has $u^{\infty}_{IF} \rightarrow 0$ independent of F and definitely not $u^{\infty}_{IF} \rightarrow u^{\infty}_{exact} \equiv -F/\lambda$. To get a good approximation of the correct fixed point u^{∞}_{exact} one therefore still needs $|\lambda|\Delta t \ll 1$!

Note:

- even for simple forward Euler fixed point ($u^{n+1} = u^n$) would be obtained exactly for large Δt (disregarding stability)

$$u^{n+1} = u^n + \Delta t \ (\lambda u^n + F)$$

Problem: Even if F evolves slowly, for large λ the integrand still evolves quickly over the integration interval: to assume the integrand is constant is a *poor approximation.*

Instead: assume only F is evolving slowly and integrate the exponential explicitly

$$u^{n+1} = e^{\lambda\Delta t}u^n + e^{\lambda\Delta t}F(t_n)\,\frac{1}{\lambda}\left(1 - e^{-\lambda\Delta t}\right)$$

This yields the forward Euler implementation of the *exponential time differencing scheme,*

$$u^{n+1} = e^{\lambda\Delta t}u^n + \Delta t\, F(t_n)\left(\frac{e^{\lambda\Delta t} - 1}{\lambda\Delta t}\right)$$

Notes:

- now, for $F = const$ and $-\lambda\Delta t \rightarrow \infty$ one gets the exact solution $u^{\infty}_{ETD} \rightarrow -F/\lambda$.
- for $|\lambda|\Delta t \ll 1$ one gets back the usual forward Euler scheme $(e^{\lambda\Delta t} - 1)/\lambda\Delta \rightarrow 1$.

For the nonlinear diffusion equation one gets for ETDFE

$$u^{n+1}_k = e^{-k^2\Delta t}u^n_k + \Delta t\, F_k(u_l(t))\left(\frac{1 - e^{-k^2\Delta t}}{k^2\Delta t}\right)$$

where in general $F_k(u_l(t))$ depends on all Fourier modes u_k.

For higher-order accuracy in time use better approximations for the integral (see Cox & Matthews, J. Comp. Physics 176 (2002) 430, and Kassam & Trefethen, SIAM J. Sci. Comput. 26 (2005) 1214, for a detailed discussion of various schemes and quantitative comparisons for ODEs and PDEs. The latter paper includes two matlab programs for Fourier and Chebyshev spectral implementations).

The 4^{th}-order Runge-Kutta version reads (using $c \equiv \lambda\Delta t$)

$$\begin{aligned}
u_{1k} &= u^n_k\, E_1 + \Delta t\, F_k(\mathbf{u}^n, t_n)\, E_2 \\
u_{2k} &= u^n_k\, E_1 + \Delta t\, F_k(\mathbf{u}_1, t_n + \frac{1}{2}\Delta t)\, E_2 \\
u_{3k} &= u_{1k}\, E_1 + \Delta t\, \left(2F_k\left(\mathbf{u}_2, t_n + \frac{1}{2}\Delta t\right) - F_k\left(\mathbf{u}^n, t_n\right)\right)\, E_2 \\
u^{n+1}_k &= u^n_k E^2_1 + \Delta t \cdot G \\
G &= F_k\left(\mathbf{u}^n, t_n\right)\, E_3 + 2\left(F_k\left(\mathbf{u}_1, t_n + \frac{1}{2}\Delta t\right) + F_k\left(\mathbf{u}_2, t_n + \frac{1}{2}\Delta t\right)\right)\, E_4 + \qquad (10) \\
&\quad + F_k\left(\mathbf{u}_3, t_n + \Delta t\right)\, E_5
\end{aligned}$$

with

$$\begin{aligned}
E_1(c) &= e^{c/2} \qquad E_2(c) = \frac{e^{c/2}-1}{c} \\
E_3(c) &= \frac{-4-c+e^c\left(4-3c+c^2\right)}{c^3} \\
E_4(c) &= \frac{2+c+e^c\left(-2+c\right)}{c^3} \\
E_5(c) &= \frac{-4-3c-c^2+e^c\left(4-c\right)}{c^3}
\end{aligned}$$

For $|c| < 0.2$ the factors $E_{3,4,5}(c)$ can become quite inaccurate due to cancellations:

$$E_5(c) = \frac{1}{c^3}\left(-4-3c-c^2+\left(1+c+\frac{1}{2}c^2+\frac{1}{6}c^3+\ldots\right)(4-c)\right) = \frac{1}{6}+\mathcal{O}(c)$$

For small values of c it is therefore better to replace $E_{3,4,5}$ by their Taylor expansions

$$\begin{aligned}
E_2(c) &= \frac{1}{2}+\frac{1}{8}c+\frac{1}{48}c^2+\frac{1}{384}c^3+\frac{1}{3840}c^4+\frac{1}{46080}c^5+\frac{1}{645120}c^6+\frac{1}{10321920}c^7 \\
E_3(c) &= \frac{1}{6}+\frac{1}{6}c+\frac{3}{40}c^2+\frac{1}{45}c^3+\frac{5}{1008}c^4+\frac{1}{1120}c^5+\frac{7}{51840}c^6+\frac{1}{56700}c^7 \\
E_4(c) &= \frac{1}{6}+\frac{1}{12}c+\frac{1}{40}c^2+\frac{1}{180}c^3+\frac{1}{1008}c^4+\frac{1}{6720}c^5+\frac{1}{51840}c^6+\frac{1}{453600}c^7 \\
E_5(c) &= \frac{1}{6}+0\,c-\frac{1}{120}c^2-\frac{1}{360}c^3-\frac{1}{1680}c^4-\frac{1}{10080}c^5-\frac{1}{72576}c^6-\frac{1}{604800}c^7
\end{aligned}$$

Alternatively, one can evaluate the coefficients via complex integration using the Cauchy integral formula [7]

$$f(z) = \frac{1}{2\pi i}\oint_{\mathcal{C}} \frac{f(t)}{t-z}\,dt \tag{11}$$

if $f(z)$ is analytic inside $\mathcal{C}$ which encloses z. Since the singularities of $E_i(c)$ at $c = 0$ are removable and since $\mathcal{C}$ can be chosen to remain a finite distance away from $c = 0$ the Cauchy integral formula (11) can be used to evaluate $E_i(c)$ even in the vicinity of $c = 0$.

Note:

- diffusion and any other linear terms retained in the eigenvalue λ of the linear operator are treated exactly
- no instability arises from the linear terms for *any* Δt : unconditionally stable
- to evaluate $F_k(\mathbf{u}_1, t_n + \frac{1}{2}\Delta t)$:

$$u_{1k} \quad \overset{\text{inverse FFT}}{\overbrace{\rightarrow}} \quad u_1(x_j) \quad \overset{\text{insert into } F}{\overbrace{\rightarrow}} \quad F(\mathbf{u}_1, t_n+\frac{1}{2}\Delta t) \overset{\text{FFT}}{\overbrace{\rightarrow}} F_k(\mathbf{u}_1, t_n+\frac{1}{2}\Delta t)$$

- if the PDE involves multiple components (e.g. u and v in a two-component reaction-diffusion system) at each stage of the RK4-scheme one needs to determine the analogous quantities u_{ik} and v_{ik} with $i = 1, 2, 3$ in parallel, i.e. one needs to determine both u_{1k} and v_{1k} before one can proceed to u_{2k} and v_{2k} etc.

- large wave numbers are strongly damped, as they should be (this is also true for operator splitting)
  compare with Crank-Nicholson (in CNAB, say)
  $$u_k^{n+1} = \frac{1-\frac{1}{2}\Delta t\, k^2}{1+\frac{1}{2}\Delta t\, k^2}\, u_k^n$$
  for large $k\Delta t$
  $$u_k^{n+1} = -(1 - \frac{4}{\Delta t\, k^2} + ...)u_k^n$$
  which exhibits *oscillatory* behavior and *slow* decay.
  Note that backward Euler also damps high-wavenumber oscillations, but it is only first order
  $$u_k^{n+1} = \frac{1}{1+\Delta t k^2}\, u_k^n \rightarrow \frac{1}{\Delta t k^2}\, u_k^n \quad \text{for} \quad |k| \rightarrow \infty.$$

Note:

- some comments on the 4th-order integrating factor scheme are in Appendix B.

4.8 Filtering

In some problems it is not (yet) possible to resolve all scales

- shock formation (cf. Burgers equation last quarter)
- fluid flow at high Reynolds numbers (turbulence): energy is pumped in at low wavenumbers (e.g. by motion of the large-scale walls), but only very high wavenumbers experience significant damping, since for low viscosity high shear is needed to have significant damping.

In these cases aliasing and Gibbs oscillations can lead to problems.

Aliasing and Nonlinearities

Nonlinearities generate high wavenumbers

$$u(x)^2 = \sum_{l=-N}^{N} \sum_{k=-N}^{N} u_l u_k e^{i(k+l)x}$$

p-th order polynomial generates wavenumbers up to $\pm pN$. On the grid of $2N$ points not all wavenumbers can be represented ⇒ Fourier interpolant $I_N(u(x))$ keeps only $\pm N$: higher wavenumber *aliased* into that range.

Example:

on grid $x_j = \frac{2\pi}{2N} j$ with only 2 grid points per wavelength $\frac{2\pi}{q}$ with $q = N$

$$u(x_j) = \cos qx_j = \cos N \frac{2\pi}{2N} j = \cos(\pi j) = (-1)^j$$

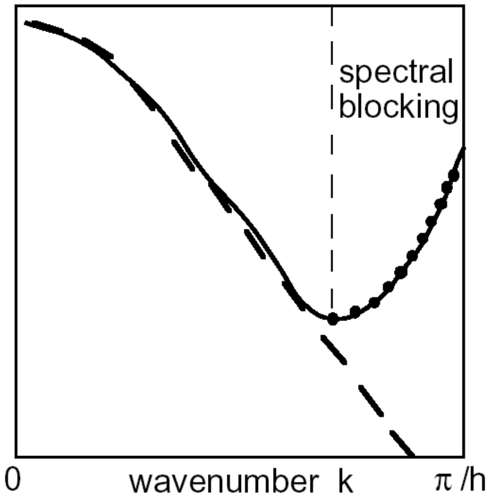
$$u(x_j)^2 = \cos^2 qx_j = (+1)^j = 1 \qquad \cos^2 qx_j \text{ is aliased to a constant on that grid}$$

Note: in a linear equation no aliasing arises during the simulation since no high wavenumbers are generated (aliasing only initially when initial condition is reduced to the discrete spatial grid)

Aliasing can lead to *spectral blocking*:

If dissipation occurs essentially only at the very high *unresolved* wavenumbers:

- dissipation is missing
- aliased high wavenumbers feed energy into the lower, weakly damped wavenumbers
- energy piles up most noticeably at the high-end of the resolved spectrum ($|k| = N$) because there the correct energy is smallest (*relative* error largest)
- pile up can lead to instability



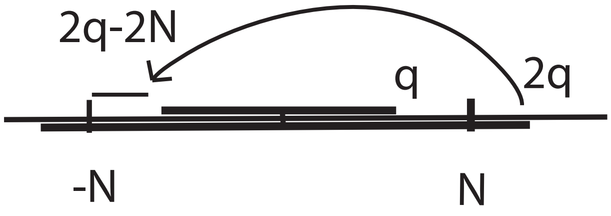
(from J.P. Boyd *Chebyshev and Fourier Spectral Methods, p. 2107)*

If resolution cannot be increased to the extent that high wavenumbers are resolved, improvement can be obtained by **filtering out** those wavenumbers that would be aliased into the lower spectrum.

Quadratic nonlinearities lead to doubling of wavenumbers:

The interval $[-q_{max}, q_{max}]$ is mapped into $[-2q_{max}, 2q_{max}]$

$$[-q_{max}, q_{max}] \rightarrow [-2q_{max}, 2q_{max}]$$



Require that the mapped wavenumber interval does not alias into the original wavenumber interval

$$2q_{max} - 2N \leq -q_{max}$$

i.e. require

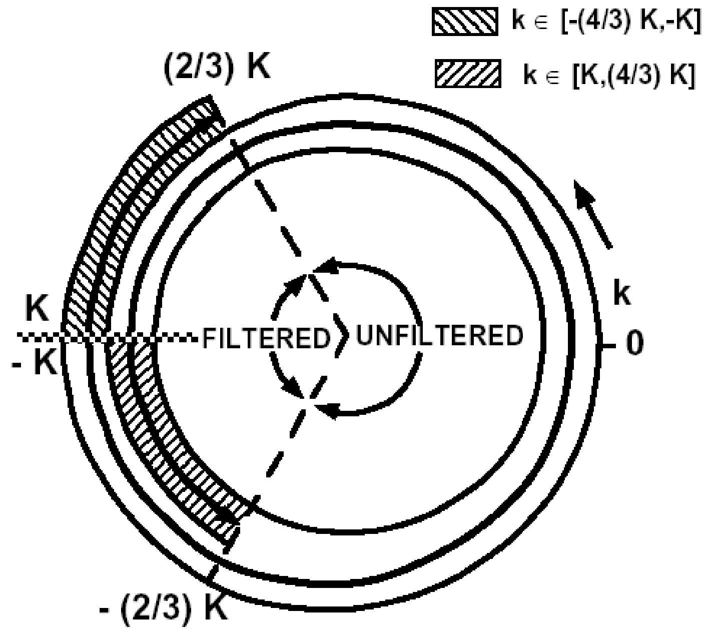
$$q_{max} \leq \frac{2}{3}N$$

More generally: for p^{th}-order nonlinearity choose

$$q_{max} = \frac{p+1}{2}N$$

Algorithm:

1. FFT: $u_i \rightarrow \tilde{u}_k$
2. take derivatives
3. filter out high wavenumbers: $\tilde{u}_k = 0$ for $|k| > \frac{p+1}{2}N$
4. inverse FFT: $\tilde{u}_k \rightarrow u_i$; this function does not contain any 'dangerous' high wavenumbers any more
5. evaluate nonlinearities $u_i \rightarrow u_i^p$
6. back to 1.



(from J.P. Boyd *Chebyshev and Fourier Spectral Methods, p. 212)*

Orszag's 2/3-rule:

For *quadratic* nonlinearity set the highest $N/3$ Fourier-modes to 0 in each time step ***just before*** the back-transformation to the spatial grid:

- evaluating the *quadratic* nonlinearity (which is done in real space):
  - the 'good' wavenumbers $[0, \frac{2}{3}N]$ contained in $u(x)$ generate the wavenumbers $[0, \frac{4}{3}N]$ of which the interval $[N, \frac{4}{3}N]$ will be aliased into $[-N, -\frac{2}{3}N]$ and therefore will contaminate the highest $N/3$ modes (analogously for $[0, -\frac{2}{3}N]$).
  - the 'bad', highest $N/3$ modes $[\frac{2}{3}N, N]$ generate wavenumbers $[\frac{4}{3}N, 2N]$ which are aliased into $[-\frac{2}{3}N, 0]$ and would contaminate the 'good' wavenumbers.
- setting the highest $N/3$ modes to 0 avoids contamination of good wavenumbers; no need to worry about contaminating the high wavenumbers that later are set to 0 anyway.

Alternative view:

For a *quadratic* nonlinearity, to represent the wavenumbers $[-N, N]$ without aliasing need $\frac{3}{2} \cdot 2N$ grid points:
want $3N$ grid points for integrals $\Rightarrow$ before transforming the Fourier modes $[-N, N]$ back to real space need to pad them with zeroes to the range $[-\frac{3}{2}N, \frac{3}{2}N]$.
Thus: To avoid aliasing for quadratic nonlinearity need 3 grid points per wavelength

$$\cos qx_j = \cos(N \frac{2\pi}{3N} j) = \cos(2\pi \frac{j}{3})$$

Notes:

- for higher nonlinearities larger portions of the spectrum have to be set to 0.
- instead of step-function filter can use smooth filter, e.g.

$$F(k) = \begin{cases} 1 & |k| \leq k_0 \, (= \frac{2}{3}N) \\ e^{-(|k|^n - |k_0|^n)} & |k| > k_0 \end{cases} \tag{12}$$

  with $n = 2, 4$.
- $\frac{2}{3}$–rule (and the smooth version) makes the pseudo-spectral method more similar to the projection of the Galerkin approach
- does not remedy the missing damping of high wavenumbers, but reduces the (incorrect) energy pumped into the weakly damped wave numbers.

Gibbs Oscillations

Oscillations due to insufficient resolution can contaminate solution even away from the sharp step/discontinuity: can be improved by *smoothing*

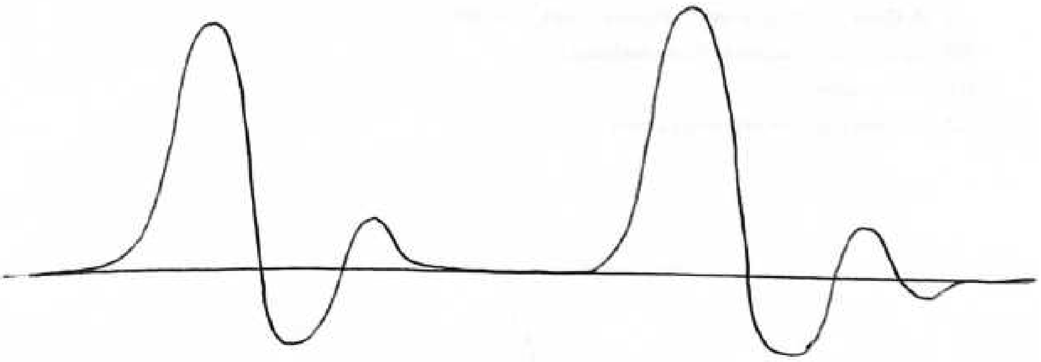
Approximate derivatives, since they are more sensitive to oscillations (function itself does not show any oscillations on the grid)

$$\partial_x u \Rightarrow \sum_{k=-N}^{N} ik\,\tilde{u}_k e^{ikx} \qquad \text{filter to} \qquad \sum_{k=-N}^{N} ik\,F(k)\,\tilde{u}_k e^{ikx}$$

with $F(k)$ as in (12).

Note:

- result is different than simply reducing number of modes since the number of grid points for the transformation is still high
- filter could also smooth away *relevant* oscillations ⇒ loose important features of solution
  e.g. interaction of localized wave pulses: oscillatory tails of the pulses determine the interaction between the pulses, smoothing would kill interaction



Notes:

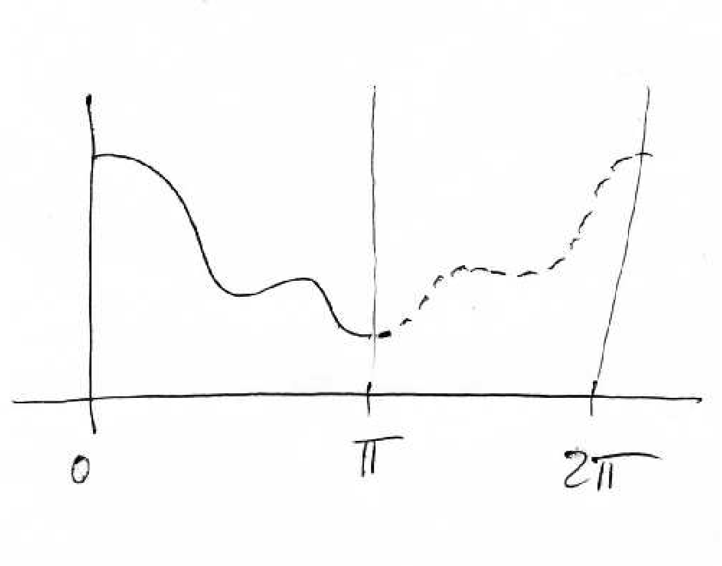
- It is always better to resolve the solution
- Filtering and smoothing make no distinction between numerical artifacts and physical features
- Shocks would better be treated with adaptive grid

5 Chebyshev Polynomials

Goal: approximate functions that are *not* periodic

5.1 Cosine Series and Chebyshev Expansion

Consider $h(\theta)$ on $0 \leq \theta \leq \pi$



extend to $[0, 2\pi]$ to generate periodic function by reflection about $\theta = \pi$

$$g(\theta) = \begin{cases} h(\theta) & 0 \leq \theta \leq \pi \\ h(2\pi - \theta) & \pi \leq \theta \leq 2\pi \end{cases}$$

Then

$$g(\theta) = \sum_{k=-\infty}^{\infty} \bar{g}_k e^{ik\theta} = \sum_{k=-\infty}^{\infty} \bar{g}_k (\cos k\theta + i \sin k\theta)$$

Reflection symmetry: $\sin\theta$ drops out

$$g(\theta) = \sum_{k=-\infty}^{\infty} \bar{g}_k \cos k\theta = \sum_{k=0}^{\infty} g_k \cos k\theta$$

with

$$g_k = \bar{g}_k \qquad \text{for} \qquad k = 0 \qquad g_k = 2\bar{g}_k \qquad \text{for} \qquad k > 0$$

$$\bar{g}_k = \frac{1}{2\pi} \int_0^{2\pi} e^{-ik\theta} g(\theta) d\theta = \frac{1}{\pi} \int_0^{\pi} \cos k\theta g(\theta) d\theta \qquad \text{reflection symmetry}$$

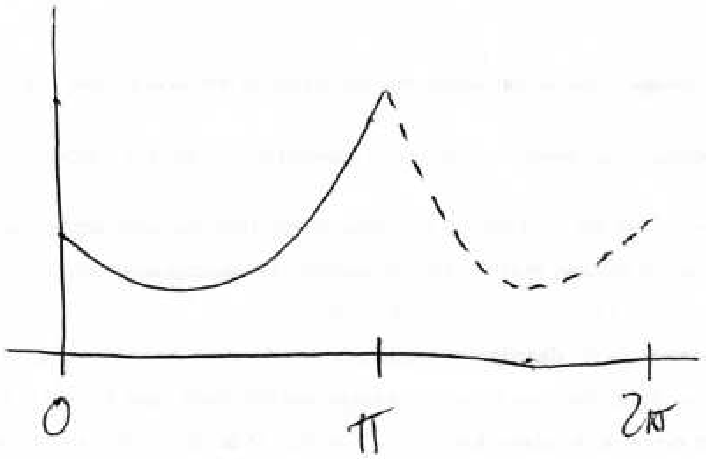
Write as

$$g_k = \frac{1}{\pi} \frac{2}{c_k} \int_0^{\pi} \cos k\theta \, g(\theta) d\theta \qquad \text{with} \qquad c_k = \begin{cases} 2 & \text{for } k = 0 \\ 1 & \text{for } k > 0 \end{cases}$$

This is the *cosine transform*.

Notes:

- Convergence of the cosine series depends on the odd derivatives at $\theta = 0$ and $\theta = \pi$



- If $\frac{dg}{d\theta} \neq 0$ at $\theta = 0$ or $\theta = \pi$ then $g_k = \mathcal{O}(k^{-2})$ even if function is perfectly smooth in $(0, \pi)$:

$$\begin{aligned} g_k &= \frac{2}{\pi c_k} \int_0^\pi \cos k\theta \, g(\theta) d\theta \qquad \text{i.b.p} \\ &= \frac{2}{\pi c_k} \frac{1}{k} \sin k\theta \, g(\theta) \Big|_0^\pi - \frac{2}{\pi c_k} \frac{1}{k} \int_0^\pi \sin k\theta \frac{d}{d\theta} g(\theta) d\theta \qquad \text{i.b.p} \\ &= \frac{2}{\pi c_k} \frac{1}{k^2} \cos k\theta \frac{d}{d\theta} g(\theta) \Big|_0^\pi - \frac{2}{\pi c_k} \frac{1}{k^2} \int_0^\pi \cos k\theta \frac{d^2}{d\theta^2} g(\theta) d\theta \end{aligned}$$

  boundary terms vanish for all k only if

$$g'(0) = 0 = g'(\pi)$$

  Since $\cos k\pi = (-1)^k$ non-zero slopes at the endpoints cannot cancel for all k.

- in general, only odd derivatives of $g(\theta)$ contribute to boundary terms:

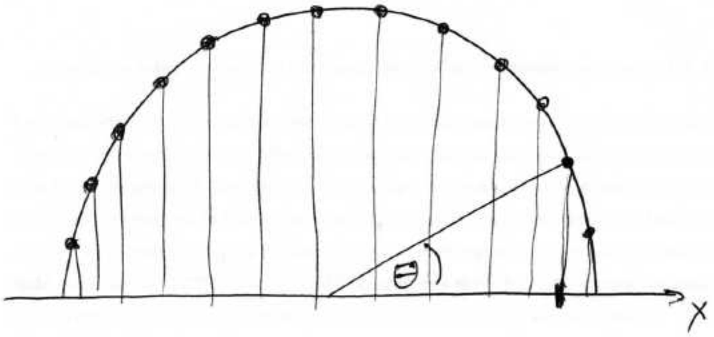
$$\frac{1}{k^{l+1}} \cos k\theta \, \frac{d^l}{d\theta^l} g(\theta) \Big|_0^\pi \qquad \text{for } l \text{ odd}$$

Thus:

- for general boundary conditions Fourier (=cosine) series converges badly: Gibbs phenomenon

5.2 Chebyshev Expansion

To get the derivative of the function *effectively* to vanish at the boundaries stretch the coordinates at the boundaries infinitely strongly. This can be achieved by parametrizing x using the angle θ on a circle:



Consider $f(x)$ on $-1 \leq x \leq 1$

Transform to $0 \leq \theta \leq \pi$ using $x = \cos\theta$, $g(\theta) = f(\cos(\theta))$

Function is now parametrized by θ instead of x

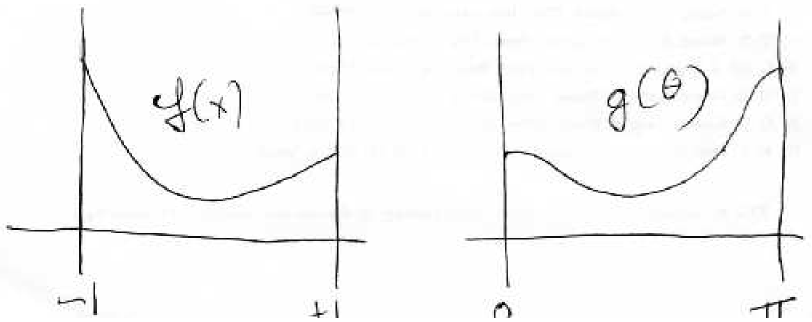
Consider Fourier series for $g(\theta)$

$$g'(\theta) = -f'(\cos\theta)\sin\theta \qquad \Rightarrow \frac{dg}{d\theta} = 0 \qquad \text{at}\, \theta = 0, \pi$$

Generally: *all* odd derivatives of $g(\theta)$ vanish at $\theta = 0$ and $\theta = \pi$.

Proof: $\cos\theta$ is even about $\theta = 0$ and about $\theta = \pi \Rightarrow f(\cos\theta)$ is also even about those points $\Rightarrow$ all odd derivatives vanish at $\theta = 0, \pi$.

Thus: the convergence of the approximation to $g(\theta)$ by a cosine-series does not depend on the boundary conditions on $f(x)$



$$f(x) \quad = \quad g(\theta) = \sum_{k=0}^{\infty} g_k \cos k\theta \qquad \text{extension of } g \text{ to } 2\pi \text{ is even}$$

$$= \sum_{k=0}^{\infty} g_k \, \cos(k \, \arccos x)$$

Introduce Chebyshev polynomials

$$\begin{aligned} T_k(x) &= \cos(k \, \arccos x) = \cos k\theta \\ f(x) &= \sum_{k=0}^{\infty} f_k \, T_k(x) \end{aligned}$$

Properties of Chebyshev Polynomials

- $T_k(x)$ is a $k^{th}-$order polynomial
  show recursively:

  $$\begin{aligned} T_0(x) &= 1 \qquad T_1(x) = x \\ T_{n+1}(x) &= \cos\left((n+1)\arccos x\right) = \cos\left((n+1)\theta\right) \end{aligned}$$

  Trig identities:

  $$\begin{aligned} \cos\left((n+1)\theta\right) &= \cos n\theta \cos\theta - \sin n\theta \sin\theta \\ \cos\left((n-1)\theta\right) &= \cos n\theta \cos\theta + \sin n\theta \sin\theta \end{aligned}$$

  cancel $\sin n\theta \sin\theta$ by adding and use $\cos(\theta) = T_1(x) = x$,

  $$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$$

  Note: recursion relation useful for computation of $T_n(x)$

- $T_n(x)$ even for n even, odd otherwise
- $T_n(x) = \sum_j a_j x^j \qquad \Rightarrow \qquad a_j$ have alternating signs
- the expansion coefficients are given by

  $$f_k = g_k = \frac{1}{\pi}\frac{2}{c_k}\int_0^{\pi} g(\theta)\cos k\theta \, d\theta$$

  rewrite in terms of x:

  $$\begin{aligned} \theta &= \arccos x \qquad d\theta = \frac{1}{\sqrt{1-x^2}}\,dx \\ f_k &= \frac{2}{\pi c_k}\int_{-1}^{1} f(x)T_k(x)\frac{1}{\sqrt{1-x^2}}dx \\ & c_k = \begin{cases} 2 & k=0 \\ 1 & k>0 \end{cases} \end{aligned}$$

- The convergence of $f(x)$ in terms of $T_k(x)$ is the same as that of $g(\theta)$ in terms of the cosine-series. In particular, boundary values are irrelevant (replace x by $\cos\theta$ in $f(x)$)

- The Chebyshev polynomials are orthogonal in the weighted scalar product

$$< T_k, T_l > \equiv \int_{-1}^{1} T_k(x) T_l(x) \frac{1}{\sqrt{1-x^2}} dx = c_k \frac{\pi}{2} \delta_{kl}$$

- The weight $\sqrt{1-x^2}^{-1}$ is singular but

$$\int_{-1}^{1} \frac{1}{\sqrt{1-x^2}} dx$$

  is finite.

- Derivatives of $T_k(x)$:

$$\frac{d}{dx} \text{ is not diagonal for basis of } T_k(x)$$

$$\frac{d}{dx} T_k(x) \neq \lambda T_k(x)$$

  in particular: the order of the polynomial changes upon differentiation. Considering $\frac{d}{d\theta}\cos(n \pm 1)\theta$ one gets

$$\begin{aligned} \frac{d}{dx} T_{k\pm 1} &= \frac{d}{d\theta} \cos(k \pm 1)\theta \, \frac{d\theta}{dx} \\ &= -(k \pm 1) \frac{1}{\frac{dx}{d\theta}} \left(\sin k\theta \cos\theta \pm \cos k\theta \sin\theta \right) \\ \frac{1}{k+1} \frac{d}{dx} T_{k+1}(x) - \frac{1}{k-1} \frac{d}{dx} T_{k-1}(x) &= \frac{1}{\sin\theta} \left(\sin k\theta \cos\theta + \cos k\theta \sin\theta - \sin k\theta \cos\theta + \cos k\theta \sin\theta \right) \end{aligned}$$

  thus

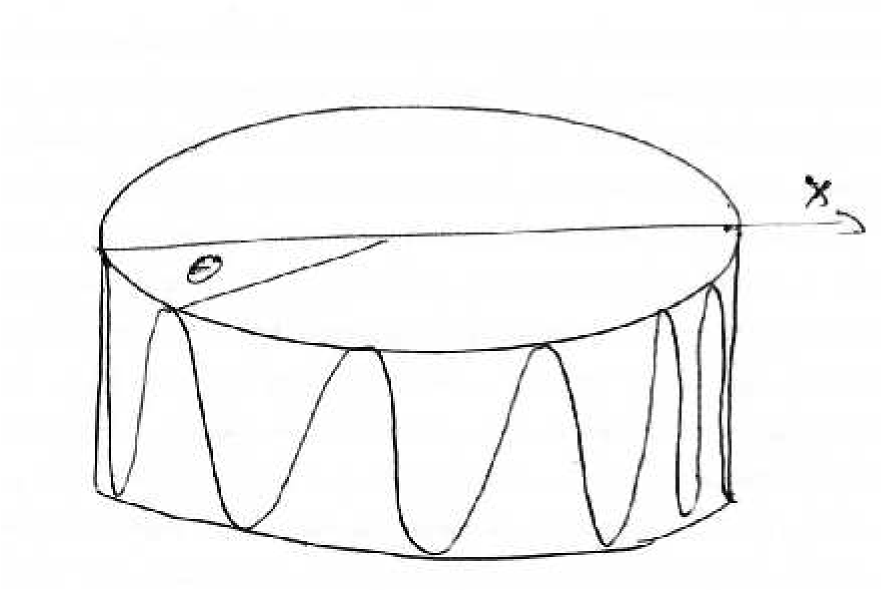
$$2T_k(x) = \frac{1}{k+1} \frac{d}{dx} T_{k+1}(x) - \frac{1}{k-1} \frac{d}{dx} T_{k-1}(x)$$

  Thus: differentiation more difficult than for Fourier modes.

- Zeroes of $T_k(x)$

$$\begin{aligned} T_k(x) &= \cos(k \arccos x) = \cos k\theta \\ \Rightarrow & \quad T_k(x) \text{ has } k \text{ zeroes in } [-1, 1] \\ & \quad k\theta_l = (2l - 1)\frac{\pi}{2} \qquad l = 1, ..., k \\ & \quad x_l = \cos \frac{2l-1}{2k} \pi \end{aligned}$$

  The zeroes *cluster* near the boundaries.



- Extrema of $T_k(x)$ (Chebyshev points)

$$\begin{aligned} k\theta_l &= l\pi \qquad x_l = \cos\frac{l}{k}\pi \qquad l = 1, ..., k \\ T_k(x_l) &= (-1)^l \end{aligned}$$

  Extrema are also clustered at boundary
  Chebyshev polynomial look like a cosine-wave wrapped around a cylinder and viewed from the side

- Transformation to $\theta = \arccos x$ places more points close to boundary: small neighborhood dx is blown up in $d\theta$

$$x = \cos\theta \qquad d\theta = -\frac{1}{\sin\theta}dx$$

$$\Rightarrow \qquad d\theta \to \infty \text{ for } \theta \to 0, \pi \frac{df}{d\theta} \to 0$$

  all derivatives vanish at boundary: no Gibbs phenomenon for non-periodic functions

- understanding of properties of functions often aided by knowing what eigenvalue problem they solve: what is the eigenvalue problem that has the $T_k(x)$ as solutions?

$$T_k(x) = \cos k\theta \qquad \frac{d^2}{d\theta^2}\cos k\theta = -k^2 \cos k\theta$$

  rewrite in terms of $x = \cos\theta$

$$\frac{d}{d\theta} = -\sin\theta \frac{d}{dx} = -\sqrt{1-x^2}\frac{d}{dx}$$

  thus $T_k(x)$ satisfies the Sturm-Liouville problem

$$\sqrt{1-x^2}\frac{d}{dx}\left(\sqrt{1-x^2}\frac{d}{dx}T_k(x)\right) + k^2 T_k(x) = 0$$

with boundary conditions: $T_k(x)$ bounded at $x = \pm 1$
Note: Sturm-Liouville problem is singular: coefficient of highest derivative vanishes at boundary ⇒ no boundary *values* specified but only boundedness
The singularity is the origin of hte good boundary resolution (no Gibbs). Fourier series is solution of regular Sturm-Liouville problem

6 Chebyshev Approximation

Approximate $f(x)$ on $a \le x \le b$ using Chebyshev polynomials

Again depending on the evaluation of the integrals

- Galerkin expansion
- Pseudospectral expansion

6.1 Galerkin Approximation

$$P_N u(x) = \sum_{k=0}^{N} u_k T_k(x)$$

with

$$u_k = \frac{2}{\pi}\frac{1}{c_k}\int_{-1}^{+1}\frac{1}{\sqrt{1-x^2}}u(x)T_k(x)dx$$

Note:

- need to transform first from interval $a \le t \le b$ to $-1 \le x \le +1$ using
$$x = \frac{2t-(a+b)}{b-a}$$

Transformation to $\theta = \arccos x$ showed

$$u_k = \mathcal{O}(k^{-r}) \qquad \text{if} \qquad u \in C^{r-1} \qquad (\partial_x^r u \in L_1)$$

i.e. if r^{th} derivative is still integrable (may be a $\delta-$function)

Show this directly in x:

$$\frac{\pi c_k}{2}u_k = \int \frac{1}{\sqrt{1-x^2}}u(x)T_k(x)dx$$

using $k^2 T_k(x) = -\sqrt{1-x^2}\frac{d}{dx}(\sqrt{1-x^2}\frac{d}{dx}T_k)$

$$\begin{aligned}
\frac{\pi c_k}{2}u_k &= -\frac{1}{k^2}\int \frac{1}{\sqrt{1-x^2}}u(x)\sqrt{1-x^2}\frac{d}{dx}\left(\sqrt{1-x^2}\,\frac{d}{dx}T_k(x)\right)dx = \\
&= -\frac{1}{k^2}u(x)\sqrt{1-x^2}\frac{d}{dx}T_k\bigg|_{-1}^{+1} + \frac{1}{k^2}\int_{-1}^{+1}\frac{du}{dx}\sqrt{1-x^2}\,\frac{d}{dx}T_k(x)dx = \qquad \text{since } u(x) \text{ bounded} \\
&= \frac{1}{k^2}\left\{\frac{du}{dx}\sqrt{1-x^2}T_k(x)\bigg|_{-1}^{+1} - \int_{-1}^{+1}\frac{d}{dx}\left(\frac{du}{dx}\sqrt{1-x^2}\right)T_k(x)dx\right\}
\end{aligned}$$

Note:

even without the 2^{nd} integration by parts it seems that $u_k = \mathcal{O}(k^{-2})$
$\Rightarrow$ it seems that even for $\frac{d^2u}{dx^2} \notin L_1$ one gets $u_k = \mathcal{O}(k^{-2})$

But:

$$\frac{d}{dx}T_k(x) = \frac{d}{dx}\cos(k \arccos x) = \mathcal{O}(k)$$

$\Rightarrow$ for $\frac{du}{dx} \in L_1$ and $\frac{d^2u}{dx^2} \notin L_1$:

$$u_k = \mathcal{O}(\frac{1}{k^2}\frac{d}{dx}T_k(x)) = \mathcal{O}(\frac{1}{k})$$

Again, convergence of Chebyshev approximation can be shown to be

$$||P_N u(x) - u(x)|| \leq \frac{C}{N^q}||u||_q$$

with $||u||$ being the usual L_2-norm (with weight $\sqrt{1-x^2}^{-1}$ and $||u||_q$ being the q^{th} Sobolev norm

$$||u||_q^2 = ||u||^2 + ||\frac{du}{dx}||^2 + ... + ||\frac{d^qu}{du^q}||^2$$

For derivatives one gets

$$||\frac{d^ru}{dx^r} - \frac{d^r}{dx^r}P_N u|| \sim ||u - P_N u||_r \leq \frac{C}{N^{\frac{1}{2}+q-2r}}||u||_q$$

Note:

- for each derivative the convergence decreases by **two** powers of N; in Fourier expansion each derivative lowered the convergence only by a single power in N.
- for $C^\infty-$functions one still has *spectral accuracy*, i.e. exponential convergence
- the estimate for the r^{th} derivative is not precisely for the derivative but for the $r-$Sobolev norm (cf. [1] for details)
- rule of thumb: for each wavelength of a periodic function one needs at least 3 Chebyshev polynomials to get reasonable approximation.

6.2 Pseudo-Spectral Approximation

For Galerkin approximation the projection integral

$$u_k = \frac{2}{\pi c_k}\int_0^\pi u(\cos\theta)\cos k\theta d\theta$$

has to be calculated exactly (e.g. analytically)

For pseudospectral approximation calculate integral based on a finite number of collocation points.

Strategy: find most accurate integration formula for the functions in question

Here: $u(\cos\theta)$ is even in $\theta \Rightarrow u(\cos\theta)\cos k\theta$ has expansion in $\cos n\theta$

$\Rightarrow$ need to consider only $\cos n\theta$ when discussing integration method

Analytically we have

$$\int_0^\pi \cos n\theta \, d\theta = \pi\,\delta_{n0}$$

Similar to Fourier case: use trapezoidal rule

$$\int_0^\pi g(\theta)\,d\theta \Rightarrow \sum_{j=0}^{N} g(\frac{\pi j}{N})\frac{\pi}{N\hat{c}_j} \qquad \text{with} \qquad \hat{c}_j = \begin{cases} 2 & j=0,N \\ 1 & \text{otherwise} \end{cases}$$

Show: Trapezoidal rule is exact for $\cos l\theta$, $l = 0, ..., 2N-1$

1. $l = 0$

$$\int d\theta \Rightarrow \sum_{j=0}^{N} g(\frac{\pi j}{N})\frac{\pi}{N\hat{c}_j} = \frac{\pi}{2N} + (N-1)\frac{\pi}{N} + \frac{\pi}{2N}$$

2. l even

$$\cos l\theta_j = \frac{1}{2}(e^{il\theta_j} + e^{-il\theta_j}) \qquad \text{with } \theta_j = \frac{\pi}{N}j$$

$$\Rightarrow \qquad \sum_{j=0}^{N}\frac{1}{\hat{c}_j}e^{il\frac{\pi}{N}j} \underbrace{=}_{e^{il\pi}=e^0 \text{ for } l \text{ even}} \sum_{j=1}^{N}\left(e^{il\frac{\pi}{N}}\right)^j$$

$$= e^{il\frac{\pi}{N}}\,\frac{1-e^{il\pi}}{1-e^{il\frac{\pi}{N}}} = 0 \qquad \text{using} \qquad \sum_{j=1}^{N} q^j = q\,\frac{1-q^N}{1-q}$$

Note: for $l = 2N$ the denominator vanishes:

$$\cos 2N\frac{\pi}{N}j \;=\; 1 \Rightarrow \sum \neq 0 \qquad \text{trapezoidal rule not correct}$$

3. l odd:
$\cos l\theta$ odd about $\theta = \frac{\pi}{2}$

$$\cos l\theta_j = -\cos l\theta_{N-j}$$

$$\begin{aligned} \cos l\theta_j &= \cos l\frac{\pi}{N}j \\ \cos l\theta_{N-j} &= \cos\left(l\frac{\pi}{N}N - l\frac{\pi}{N}j\right) = -\cos\left(-l\frac{\pi}{N}j\right) \end{aligned}$$

$$\Rightarrow \sum_{j=0}^{N}\cos l\theta_j = 0$$

Transform in $x-$coordinates

$$\int_{-1}^{1} \frac{p(x)}{\sqrt{1-x^2}} dx = \int_0^{\pi} p(\cos\theta) d\theta = \sum_{j=0}^{N} p(\cos\frac{\pi}{N}j)\, \frac{\pi}{N\hat{c}_j}$$

Note:

This can also be viewed as a Gauss-Lobatto integration

$$\int_{-1}^{1} p(x)w(x)dx = \sum_{j=0}^{N} p(x_j)w_j$$

with points $x_j = \cos\frac{\pi}{N}j$ and weights $w_j = \frac{\pi}{N\hat{c}_j}$

Gauss-Lobatto integration is exact for polynomials up to degree $2N-1$:

- degree $2N-1$ polynomials have $2N$ coefficients
- $2N$ parameters to choose:
  w_j for $j = 0, ..., N$ and x_j for $j = 1, ..., N-1$ since $x_0 = -1$ and $x_N = +1$

The x_j are roots of a certain polynomial $q(x) = p_{N+1}(x) + ap_N(x) + bp_{N-1}(x)$ with a and b chosen such that $q(\pm 1) = 0$

Note: for the scalar product one needs the integral to be exact up to order $2N$ since each factor can be a N^{th}-order polynomial $\Rightarrow$ see (13) below

Summarizing:

pseudo-spectral coefficients given by

$$\tilde{u}_k = \frac{2}{Nc_k} \sum_{j=0}^{N} u(x_j)\, T_k(x_j) \frac{1}{\hat{c}_j}$$

with

$$\hat{c}_i = \begin{cases} 2 & i = 0, N \\ \\ 1 & 1 \leq i \leq N-1 \end{cases}$$

again highest mode resolvable on the grid given by

$$T_N(x_j) = \cos\left(N \arccos\left(\cos\frac{\pi}{N}j\right)\right) = \cos\pi j = (-1)^j$$

Remember origin of c_k

$$c_N = 2 \qquad \text{as in Fourier expansion in } \theta$$

$$c_0 = 2 \qquad \text{since only for } k \neq 0 \text{ two exponentials } e^{\pm ikx} \text{ contribute to } \cos kx$$

Note:

- need not distinguish between c_k and $\hat{c_j}$: from now on $\hat{c}_j = c_j$

Notes:

- transformation can be written as matrix multiplication

$$\tilde{u}_k = \sum_{j=0}^{N} C_{kj} u(x_j)$$

  with

$$\begin{aligned} C_{kj} &= \frac{2}{N c_k c_j} T_k(x_j) = \frac{2}{N c_k c_j} \cos\left(k \arccos(\cos \frac{\pi}{N} j)\right) \\ &= \frac{2}{N c_k c_j} \cos(\frac{k j \pi}{N}) \end{aligned}$$

- the inverse transformation is

$$u(x_j) = \sum_{k=0}^{N} T_k(x_j)\, \tilde{u}_k = \sum_{k=0}^{N} \left(C^{-1}\right)_{jk} \tilde{u}_k$$

  with

$$\left(C^{-1}\right)_{jk} = T_k(x_j) = \cos \frac{\pi j k}{N}$$

- transformation seemingly $\mathcal{O}(N^2)$: but there are again fast transforms (see later).
- discrete orthogonality

$$\sum_{j=0}^{N} T_l(x_j) T_k(x_j) \frac{1}{c_j} = \frac{N}{2} c_l \delta_{lk}$$

  since for $l + k \leq 2N - 1$ the integration is exact

$$\sum_{j=0}^{N} T_l(x_j) T_k(x_j) w_j = \int T_l(x) T_k(x) \frac{1}{\sqrt{1-x^2}} dx = c_k \frac{\pi}{2} \delta_{lk} \qquad \text{note: } w_j = \frac{\pi}{c_j N}$$

  for $l + k = 2N$: since $l, k \leq N$ one has $l = N = k$: $T_N(x_j) = (-1)^j$

$$\Rightarrow \qquad \sum_{j=0}^{N} T_N(x_j) T_N(x_j) \frac{1}{c_j} = N \tag{13}$$

  although T_N^2 is not a constant (only on the grid).

The pseudospectral approximant interpolates the function on the grid

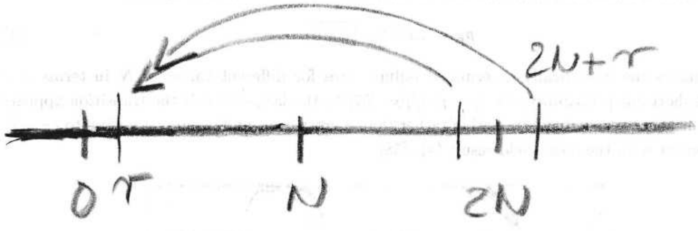
$$I_N u(x_l) = \sum_{k=0}^{N} \tilde{u}_k T_k(x_l) = \sum_{k=0}^{N} \sum_{j=0}^{N} \frac{2}{N c_k c_j} u(x_j) T_k(x_j) T_k(x_l)$$

use $T_k(x_j) = \cos k \arccos x_j = \cos k\frac{\pi j}{N} = T_j(x_k)$ and orthogonallity

$$\Rightarrow I_N u(x_l) = \sum_{j=0}^{N} \frac{2}{Nc_j} u(x_j) \sum_{k=0}^{N} \frac{1}{c_k} T_j(x_k) T_l(x_k) = \sum_{j=0}^{N} u(x_j) \frac{c_l}{c_j} \delta_{jl} = u(x_l)$$

Aliasing:

As with Fourier modes the pseudosprectral approximation has aliasing errors:



In Fourier we have aliasing from $2N + r$ to r and from $-2N + r$ to r . The mode $-2N + r$ is also contained in the Chebyshev mode $\cos(2N - r)\theta$. Therefore $2N - r$ also aliases into r.

Consider $T_{2mN \pm r}(x)$ on grid $x_j = \cos\frac{\pi j}{N}$

$$\begin{aligned} T_{2mN\pm r}(x_j) &= \cos\left((2mN \pm r)\arccos(\cos\frac{\pi j}{N})\right) = \cos\left((2mN \pm r)\frac{\pi j}{N}\right) = \\ &= \cos 2m\frac{N\pi j}{N}\cos r\frac{\pi j}{N} \mp \underbrace{\sin 2m\frac{N\pi j}{N}}_{0}\sin r\frac{\pi j}{N} = \cos r\frac{\pi j}{N} \end{aligned}$$

Thus: $T_{\pm r+2mN}$ is aliased to $T_r(x)$ on the grid.

Coefficients of T_k are determined by all contributions that look like T_k on the grid

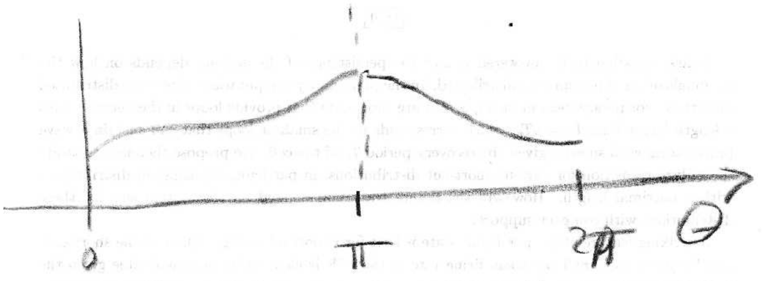
$$\tilde{u}_k = u_k + \sum_{m=1}^{\infty} u_{2mN\pm k}$$

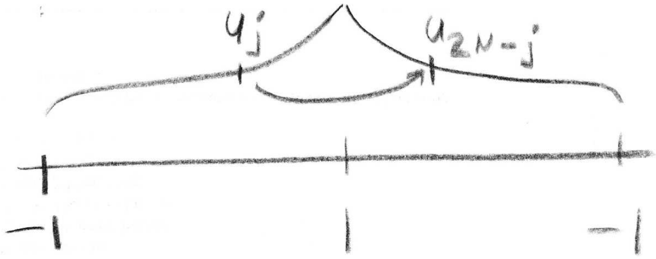
6.2.1 Implementation of Fast Transform

The $\tilde{u}_k$ can be obtained using FFT for $u(x)$ real

Extend $u(\cos\theta)$ from $[0, \pi]$ to $[0, 2\pi]$ in '$\theta-$space':

extended $f(\cos\theta)$ is periodic in $\theta \Rightarrow$ FFT





extension

$$u_j = \begin{cases} u(x_j) & 0 \le j \le N \\ u(x_{2N-j}) & N+1 \le j \le 2N \end{cases}$$

Note:

- in Matlab the extension can be done easily using the command FLIPDIM

Coefficients are given by

$$\tilde{u}_k = \frac{2}{Nc_k}\sum_{j=0}^{N} u(x_j)T_k(x_j)\frac{1}{c_j} = \frac{2}{Nc_k}\sum_{j=0}^{N} u(x_j)\cos(k\frac{\pi j}{N})\frac{1}{c_j} \tag{14}$$

Rewrite the sum in terms of the extension (using that $\cos$ and u are even with respect to $\theta = 0, \pi$

$$\sum_{j=1}^{N-1} u_j \cos\frac{\pi jk}{N} \underbrace{=}_{j=2N-r \quad r=2N-j} \sum_{r=N+1}^{2N-1} \underbrace{u_{2N-r}}_{u_r} \cos\left(\frac{\pi k}{N}(2N-r)\right) = \sum_{r=N+1}^{2N-1} u_r \cos\frac{\pi kr}{N}$$

thus considering factor $1/c_j$ in (14)

$$\begin{aligned} \tilde{u}_k &= \frac{2}{Nc_k}\frac{1}{2}\left\{u_0\cos 0 + u_N\cos\pi k + 2\sum_{j=1}^{N-1} u_j\cos\frac{\pi jk}{N}\right\} = \\ &= \frac{2}{Nc_k}\frac{1}{2}\left\{u_0\cos 0 + u_N\cos\pi k + \sum_{j=1}^{N-1} u_j\cos\frac{\pi jk}{N} + \sum_{j=N+1}^{2N-1} u_j\cos\frac{\pi jk}{N}\right\} \\ &= \frac{1}{Nc_k}\sum_{j=0}^{2N-1} u_j\cos\frac{\pi jk}{N} = \frac{1}{Nc_k}\underbrace{Re\left\{\sum_{j=0}^{2N-1} u_j e^{i\frac{j\pi k}{N}}\right\}}_{FFT} \end{aligned}$$

Notes:

- here the ordering of grid points is $x = \cos\theta$
  therefore $u_0 = u(+1)$ and $u_N = u(-1)$

Reorder:

$$z_j = \cos\theta_{N-j} \qquad \text{then} \qquad z_0 = -1 \quad z_N = +1$$

$$\begin{aligned} T_k(z_j) &= \cos\left(k \arccos\cos\theta_{N-j}\right) = \cos\left(k(N-j)\frac{\pi}{N}\right) \\ &= \cos k\pi \, \cos\frac{kj\pi}{N} + \sin k\pi \, \sin\frac{kj\pi}{N} = (-1)^k \cos\frac{kj\pi}{N} \end{aligned}$$

Thus:

$$T_k(z_j) = (-1)^k T_k(x_j)$$

expressing the fact that reflecting about the y-axis ($x \rightarrow -x$) amounts to switching sign of the *odd* Chebyshev polynomials but leaving the even T_k unchanged.

Relation to FFT is changed

$$\begin{aligned} \tilde{u}_k &= \frac{2}{Nc_k}\sum_{j=0}^{N} u(x_j)T_k(x_j)\frac{1}{c_j} \underbrace{=}_{\textbf{relabeling}} \frac{2}{Nc_k}\sum_{j=0}^{N} u(z_j)T_k(z_j)\frac{1}{c_j} \\ &= (-1)^k\frac{2}{Nc_k}\sum_{j=0}^{N} u(z_j)\cos\frac{kj\pi}{N}\,\frac{1}{c_j} = (-1)^k\frac{1}{Nc_k}\underbrace{Re\left\{\sum_{j=0}^{2N-1}\hat{u}_j\, e^{i\frac{j\pi k}{k}}\right\}}_{\textbf{FFT}} \end{aligned}$$

where

$$\hat{u}_0 = u(-1) \qquad \hat{u}_N = u(+1) \qquad \hat{u}_{2N} = u(-1)$$

$\Rightarrow$ with natural ordering FFT yields $(-1)^k\tilde{u}_k$.

6.3 Derivatives

Goal: approximate derivative of $u(x)$ by derivative of interpolant $I_N u(x)$

Need $\frac{d}{dx}T_k(x)$ in terms of $T_k(x)$. We had

Recursion Relation

$$\begin{aligned} \frac{d}{dx}T_{m+1}(x) &= (m+1)\left\{2T_m(x) + \frac{1}{m-1}\frac{d}{dx}T_{m-1}(x)\right\} \qquad m \geq 2 \\ \text{with} \qquad & \frac{d}{dx}T_0(x) = 0 \qquad \frac{d}{dx}T_1(x) = T_0 \end{aligned}$$

Note:

- $\frac{d}{dx}T_{m-1}$ contains even lower T_l etc.: $\frac{d}{dx}T_m$ contains contributions from many T_k

First Derivative

Expand the derivative of the interpolant in $T_k(x)$

$$\frac{d}{dx}\left(I_N u(x)\right) = \sum_{k=0}^{N} \tilde{u}_k \frac{d}{dx} T_k(x) = \sum_{k=0}^{N} b_k T_k(x)$$

To determine b_l project derivative onto $T_l(x)$

$$\sum_{k=0}^{N} \tilde{u}_k \int_{-1}^{+1} T_l(x) \frac{d}{dx} T_k(x) \frac{1}{\sqrt{1-x^2}} dx = \sum_{k=0}^{N} b_k \underbrace{\int_{-1}^{1} T_l(x) T_k(x) \frac{1}{\sqrt{1-x^2}} dx}_{\delta_{lk}\frac{\pi}{2}c_k} = \frac{\pi}{2} c_l\, b_l$$

Note:

- here $c_0 = 2$ and $c_N = 1$ since full projection, integrand evaluated not only at discrete grid points (we get an analytic result for the b_k)

Use

$$\int_{-1}^{1} T_l(x) \frac{d}{dx}\left(T_k(x)\right) \frac{1}{\sqrt{1-x^2}} dx = \begin{cases} \quad 0 & l \geq k \\ 0 \quad k > l & k + l \text{ even} \\ k\pi \quad k > l & k + l \text{ odd} \end{cases}$$

Proof:

1. $l \geq k$
   degree of $\frac{d}{dx}T_k$is $k-1 \Rightarrow$ can be expressed by sum of T_j with $j < l$; scalar product vanishes since $T_k \perp T_j$ for $j \neq k$
2. $k + l$ even $\Rightarrow$ l and k both even or both odd $\Rightarrow$ $T_l \frac{d}{dx} T_k$ odd $\Rightarrow$ integral vanishes
3. $k + l$ odd, $k > l$: prove by induction
   write $k = l + 2r - 1$, $r = 1, 2, 3, ...$
   (a) $r = 1$, $k = l + 1$
   first $l \neq 0$

$$< T_l \frac{d}{dx} T_{l+1} > \underbrace{=}_{\text{recursion for } \frac{d}{dx}T_{l+1}} (l+1) \left\{ 2 < T_l T_l > + \frac{1}{l-1} \underbrace{< T_l \frac{d}{dx} T_{l-1} >}_{=0 \text{ since } l-1<l} \right\} = 2(l+1)\frac{\pi}{2}$$

now $l = 0$

$$< T_0 \frac{d}{dx} T_1 > = < T_0 T_0 > = \pi$$

(b) induction step: assume

$$< T_l \frac{d}{dx} T_{l+2r-1} >= \underbrace{(l+2r-1)}_{k} \pi, \qquad r \geq 1$$

$$\begin{aligned}
\langle T_l \frac{d}{dx} T_{l+2(r+1)-1} \rangle &= \langle T_l (l+2r+1) \left(2T_{l+2r} + \frac{1}{l+2r-1} \frac{d}{dx} T_{l+2r-1} \right) \rangle \\
&= \frac{l+2r+1}{l+2r-1} \langle T_l \frac{d}{dx} T_{l+2r-1} \rangle = \frac{l+2r+1}{l+2r-1} (l+2r-1)\pi = (l+2r+1)\pi \\
&= (l+2(r+1)-1)\pi
\end{aligned}$$

Thus:

$$b_l = \frac{2}{c_l} \sum_{k=l+1\,;\,k+l \text{ odd}}^{N} k \tilde{u}_k$$

Notes:

- calculation of single coefficient b_l is $\mathcal{O}(N)$ operations instead of $\mathcal{O}(1)$ for Fourier $\Rightarrow$ calculation of complete derivative seems to require $\mathcal{O}(N^2)$ operation
- b_l depends only on $\tilde{u}_k$ with $k > l$: only polynomials with higher degree contribute to a given power of x of the derivative

Determine b_l recursively:

$$\frac{c_l}{2} b_l = (l+1)\tilde{u}_{l+1} + \sum_{k=l+3\,;\,k+l \text{ odd}}^{N} k \tilde{u}_k = (l+1)\tilde{u}_{l+1} + \frac{c_{l+2}}{2} b_{l+2}$$

Thus

$$\begin{aligned}
b_N &= 0 \\
b_{N-1} &= 2N\tilde{u}_N \\
c_l b_l &= 2(l+1)\tilde{u}_{l+1} + b_{l+2} \qquad 0 \leq l \leq N-2
\end{aligned}$$

Note:

- here $c_N = 1$ since full integral $\Rightarrow$ no factor c_{l+2} for $l \leq N-2$
- recursion relation requires only $\mathcal{O}(N)$ operations for all N coefficients
- recursion relation *cannot* be parallelized or vectorized:
  evaluation of b_l requires knowledge of b_k with $k > l$:
  - cannot evaluate all coefficients b_l simultaneously on parallel computers
  - cannot start evaluating product involving b_l without finishing first calculation for b_k with $k > l$

Higher Derivatives

calculate higher derivatives recursively

$$\frac{d^n}{dx^n}u(x) = \frac{d}{dx}\left(\frac{d^{n-1}}{dx^{n-1}}u(x)\right)$$

i.e. given

$$\frac{d^{n-1}}{dx^{n-1}}I_N(u(x)) = \sum_{k=0}^{N} b_k^{(n-1)}T_k(x)$$

one gets

$$\frac{d^n}{dx^n}I_N(u(x)) = \sum_{k=0}^{N} b_k^{(n-1)}\frac{d}{dx}T_k(x) = \sum_{k=0}^{N} b_k^{(n)}T_k(x)$$

with

$$\begin{aligned} b_N^{(n)} &= 0 \\ b_{N-1}^{(n)} &= 2Nb_N^{(n-1)} \\ c_l b_l^{(n)} &= 2(l+1)b_{l+1}^{(n-1)} + b_{l+2}^{(n)} \end{aligned}$$

Note:

- to get n^{th} derivative effectively have to calculate all derivatives up to n

6.3.1 Implementation of Pseudospectral Algorithm for Derivatives

Combine the steps: given $u(x)$ at the collocation points x_j calculate $\partial_x^n u$ at x_j

I. Transform Method

1. Transform to Chebyshev amplitudes

$$\tilde{u}_k = \frac{2}{Nc_k}\sum_{j=0}^{N} u(x_j)\cos\frac{jk\pi}{N}\,\frac{1}{c_j}$$

2. Calculate derivatives recursively

$$\begin{aligned} b_N^{(n)} &= 0 \\ b_{N-1}^{(n)} &= 2Nb_N^{(n-1)} \\ c_l b_l^{(n)} &= 2(l+1)b_{l+1}^{(n-1)} + b_{l+2}^{(n)} \end{aligned}$$

3. Transform back to real space at x_j

$$\partial_x^n I_N(u(x_j)) = \sum_{k=0}^{N} b_k^{(n)}\cos\frac{jk\pi}{N}$$

Note:

- steps 1. and 3. can be performed using FFT

FFT for back transformation

forward transformation was

$$\tilde{u}_k = \frac{2}{Nc_k}\sum_{j=0}^{N} u(x_j)\cos\frac{jk\pi}{N}\frac{1}{c_j} = \frac{1}{Nc_k}Re\left\{\sum_{j=0}^{2N-1} u_j e^{i\frac{\pi jk}{N}}\right\} \tag{15}$$

the last sum can be done as forward FFT

For first derivative at x_j we need

$$\sum_{k=0}^{N} b_k \cos\frac{jk\pi}{N}$$

1. extend b_j
$$b_r = b_{2N-r} \qquad \text{for} \qquad r = N+1, ..., 2N-1$$
2. need factors c_j (cf. (15)): redefine b_j
$$\hat{b}_0 = 2b_0 \qquad \hat{b}_N = 2b_N \qquad \hat{b}_j = b_j \qquad \text{for } j \neq 0, N$$
$$\sum_{k=0}^{N} b_k \cos\frac{jk\pi}{N} = \sum_{k=0}^{N} \hat{b}_k \cos\frac{jk\pi}{N}\frac{1}{c_k} = \frac{1}{2}\underbrace{Re\left\{\sum_{k=0}^{2N-1} \hat{b}_k e^{i\frac{jk\pi}{N}}\right\}}_{\textbf{FFT}}$$

Last sum can again be done as *forward* FFT.

Notes:

- backward transformation uses the same FFT as the forward transformation. more precisely, because only real part is taken the sign of i does not matter
- again for natural ordering want derivative at $z_j = \cos\frac{\pi}{N}(N-j)$: need
$$\hat{b}_k \cos\left(\frac{k\pi}{N}(N-j)\right) = (-1)^k \hat{b}_k \cos\frac{kj\pi}{N}$$
$\Rightarrow$ replace
$$\hat{b}_k \to (-1)^k \hat{b}_k$$

II. Matrix Multiply Approach

As in Fourier case derivative is linear in $u(x_j) \Rightarrow$ can be written as matrix multiplication

$$\partial_x I_N(u(x_j)) = \sum_{k=0}^{N} D_{jk} u(x_k)$$

D_{jk} gives contribution of $u(x_k)$ to derivative at x_j

The polynomial $I_N(u(x))$ interpolates u on the grid x_j. Since the order of I_N is equal to the number of grid points, this polynomial is unique. Therefore start by seeking the polynomial that interpolates $u(x_j)$ and then take its derivative.

Construct interpolating polynomial from polynomials $g_k(x)$ satisfying

$$g_k(x_j) \;=\; \delta_{jk}$$

$$u(x_j) = \sum_{k=0}^{N} g_k(x_j) u(x_k)$$

$$\partial_x u(x)|_{x_j} = \sum_{k=0}^{N} \partial_x g_k(x)\bigg|_{x_j} u(x_k) \equiv \sum_{k=1}^{N} D_{jk} u(x_k)$$

Construct the polynomial noting that Chebyshev polynomial $T_N(x)$ has extrema at all x_j for $1 \le j \le N-1$

$$\frac{d}{dx} T_N(x_j) = 0 \qquad \text{for } j = 1, ...N-1$$

Note: $\frac{d}{dx}T_N$ has $N-1$ zeroes since it has order $N-1$

$$g_k(x) = \underbrace{\frac{(-1)^{k+1}}{N^2 c_k}}_{\text{normalization}} \; \overbrace{(1-x^2)}^{\text{vanishes at } x_{0,N}} \; \overbrace{\frac{d}{dx} T_N(x)}^{\text{vanishes at } x_j} \; \underbrace{\frac{1}{x - x_k}}_{\text{cancels } (x-x_k) \text{ in numerator}}$$

Notes:

- $\sum u(x_k) g_k(x)$ interpolates u on the grid
- $g_k(x)$ is indeed a polynomial since denominator is cancelled by $\frac{d}{dx}T_N$, which vanishes at the x_k
- $g_k(x)$ is a Lagrange polynomial

$$L_k^{(N)}(x) = \prod_{k \neq m = 0}^{N} \frac{x - x_m}{x_k - x_m} = \frac{x - x_0}{x_k - x_0} \cdots \frac{x - x_{k-1}}{x_k - x_{k-1}} \frac{x - x_{k+1}}{x_k - x_{k+1}} \cdots \frac{x - x_{N-1}}{x_k - x_{N-1}} \frac{x - x_N}{x_k - x_N} \qquad 0 \le k \le N$$

Take derivative of $g_k(x)$

$$\frac{d}{dx} I_N u(x_j) = \sum_{k=0}^{N} u(x_k) g_k'(x_j) = \sum_{k=0}^{N} D_{jk} u(x_k)$$

For natural ordering $z = \cos\theta_{N-j} = \cos\frac{N-j}{N}\pi$, i.e. $z_0 = -1$ and $z_N = 1$, one gets

$$\begin{aligned} D_{jk} &= \frac{c_j}{c_k}(-1)^{j+k}\frac{1}{x_j - x_k} \qquad \text{for } j \neq k \\ D_{jj} &= -\frac{x_j}{2\,(1-x_j^2)} \qquad \text{for } j \neq 0, N \\ D_{00} &= -\frac{2N^2+1}{6} \qquad D_{NN} = +\frac{2N^2+1}{6} \end{aligned} \tag{16}$$

Notes:

- differentiation matrix is *not* skew-symmetric

$$D_{jk} \neq D_{kj} \qquad \text{since } D_{jj} \neq 0 \text{ and } \frac{c_j}{c_k}$$

- $||D|| = \mathcal{O}(N^2)$ because of clustering of points at the boundary clear for D_{00} and D_{NN}.
  smallest grid distance is $\mathcal{O}(N^{-2})$, e.g., for $|j - N| \ll N$

$$1 - z_j = 1 - \cos(\theta_{N-j}) = 1 - (1 - \frac{(N-j)^2}{N^2}\pi^2 + \ldots) = \mathcal{O}(N^{-2})$$

  ⇒ stability condition will involve N^{-2} instead of N^{-1}
  ⇒ more restrictive than Fourier modes

- higher derivatives obtained via $\mathbf{D}^n$

Note:

- it turns out that the numerical accuracy of the matrix-multiply approach using D as formulated in (16) is quite prone to numerical round-off errors. D has to satisfy

$$\sum_{j=0}^{N} D_{ij} = 0 \qquad \forall j$$

  reflecting that the derivative of a constant vanishes.
  A better implementation is

$$\begin{aligned} D_{jk} &= \frac{c_j}{c_k}(-1)^{j+k}\frac{1}{x_j - x_k} \qquad \text{for } j \neq k \\ D_{jj} &= -\sum_{j\neq k=0}^{N} D_{jk} \end{aligned} \tag{17}$$

$$\tag{18}$$

7 Initial-Boundary-Value Problems: Pseudo-spectral Method

We introduced Chebyshev polynomials to deal with general boundary conditions. Implement them now

7.1 Brief Review of Boundary-Value Problems

Depending on character of equation we need to pose/may pose different number of boundary conditions at different locations.

7.1.1 Hyperbolic Problems

characterized by *traveling* waves: boundary conditions depend on characteristics:

Boundary condition to be posed on incoming characteristic variable but not on outgoing characteristic variable. Solution blows up if boundary condition is posed on wrong variable.

1. Scalar wave equation

$$\partial_t u = \partial_x u \qquad u(x,0) = u_0(x) \qquad -1 \leq x \leq +1$$

   wave travels to the left

$$u(x,t) = u(x+vt)$$

   distinguish boundaries;

   (a) $x = -1$: outflow boundary $\Rightarrow$ u is outgoing variable
   requires and allows *no* boundary condition

   (b) $x = +1$: inflow boundary $\Rightarrow$ u is incoming variable
   needs and allows single boundary condition

2. System of wave equations

$$\partial_t \mathbf{u} = \mathbf{A}\partial_x \mathbf{u}$$

   diagonalize A to determine characteristic variables
   Example:

$$\begin{aligned} \partial_t u &= \partial_x v \\ \partial_t v &= \partial_x u \end{aligned}$$

   Taking sum and difference

$$U_l = u + v \qquad U_r = u - v$$

$$\partial_t U_{l,r} = \pm \partial_x U_{l,r}$$

   (a) $x = -1$: only U_r is incoming, only U_r accepts boundary condition

   (b) $x = +1$: only U_l is incoming, only U_l accepts boundary condition

   Physical boundary conditions often not in terms of characteristic variables
   Example:

$$u = u^{\pm} \qquad \text{at } x = \pm 1 \qquad v \text{ unspecified}$$

   at $x = -1$:

$$\begin{aligned} U_r(-1) &= u^- - v(-1) = u^- - \frac{1}{2}\left(U_l(-1) - U_r(-1)\right) \\ U_r(-1) &= 2u^- - U_l(-1) \end{aligned}$$

7.1.2 Parabolic Equations

No characteristics, boundary conditions at each boundary

Example:

$$\partial_t u = \nabla \cdot \mathbf{j} = \nabla \cdot \nabla u = \Delta u$$

Typical boundary conditions:

1. Dirichlet
$$u = 0$$
2. Neumann (no flux boundary condition)
$$\partial_x u = 0$$
3. Robin boundary conditions
$$\alpha u + \beta \partial_x u = g(t)$$

7.2 Pseudospectral Implementation

Pseudospectral: we have grid points ⇒ boundary values available
discuss using matrix-multiply approach

Explore: simple wave equation

$$\partial_t u = \partial_x u \qquad u(x=1,t) = g(t)$$

discretize

$$\partial_t u_i = \sum_{j=0}^{N} D_{ij} u_j \qquad \text{with } u_j = u(x_j)$$

Notes:

- spatial derivative calculated using all points
⇒ derivatives available at boundaries without introducing the virtual points that appeared when using finite differences
$$\partial_x u_0 = \frac{1}{2\Delta x}(u_1 - u_{-1})$$
- boundary condition *seems* not necessary: it looks as if u_N could be updated without making use of $g(t)$.
But: PDE would be *ill-posed* without boundary conditions
⇒ scheme should blow up! (see later)

Correct implementation

$$\begin{aligned} \partial_t u_i &= \sum_{j=0}^{N} D_{ij} u_j \qquad i = 0, ..., N-1 \\ u_N &= g(t) \end{aligned}$$

Note:

- although u_N is not updated using the PDE, it can still be used to calculate the derivative at the other points.

Express scheme in terms of *unknown variables* only: $u_0, u_1, ...u_{N-1}$
Define reduced $n \times n-$differentiation matrix

$$D_{ij}^{(N)} = D_{ij} \qquad i, j = 0, ..., N-1$$

i.e. N^{th} row and column of D_{ij} are omitted.

$$\begin{aligned} \partial_t u_i &= \sum_{j=0}^{N-1} D_{ij}^{(N)} u_j + D_{iN} u_N \qquad i = 0, ..., N-1 \\ u_N &= g(t) \end{aligned}$$

Notes:

- boundary conditions modify differentiation matrix
- in general equation becomes inhomogeneous

7.3 Spectra of Modified Differentiation Matrices

With $\mathbf{u} = (u_0, ..., u_{N-1})$ PDE becomes inhomogeneous system of ODEs

$$\partial_t \mathbf{u} = \mathbf{D}^{(N)} \mathbf{u} + \mathbf{d} \qquad \text{with } d_i = D_{iN} g(t)$$

For simplicity assume vanishing boundary values: $\mathbf{d} = 0$
Stability properties determined by eigenvalues λ_j of modified differentiation matrix $\mathbf{D}^{(N)}$

$$\partial_t \mathbf{u}_j = \lambda_j \mathbf{u}_j$$

Reminder:

- region of absolute stability of scheme for eigenvalue λ_j

$$\{\lambda_j \, \Delta t \in \mathbb{C} | \mathbf{u}_j \text{ bounded for all } t\}$$

- scheme is asymptotically stable if it is absolutely stable for all eigenvalues of $\mathbf{D}^{(N)}$

7.3.1 Wave Equation: First Derivative

What are the properties of $\mathbf{D}^{(N)}$?

Review of Fourier Case

- eigenvalues of $\mathbf{D}_F$ are ik, $|k| = 0, 1, ...N-1$. All eigenvalues are purely imaginary and the eigenvalue 0 is double.
- $\mathbf{D}_F$ is normal $\Rightarrow$ can be diagonalized by unitary matrix $\mathbf{U}$

$$\mathbf{U}^{-1}\mathbf{D}\mathbf{U} = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \cdots & \\ & & & \lambda_N \end{pmatrix} \equiv \mathcal{D}$$

  with $||\mathcal{D}|| = ||\mathbf{D}||$ and $||\mathbf{U}^{-1}\mathbf{u}|| = ||\mathbf{u}||$
  $\Rightarrow$ $||\mathbf{u}||$ is bounded by the same constant as $||\mathbf{U}^{-1}\mathbf{u}||$, independent of N
  $\Rightarrow$ sufficient to look at scalar equation.

Properties of $\mathbf{D}^{(N)}$ for Chebyshev

- eigenvalues of $\mathbf{D}^{(N)}$ are not known analytically
- eigenvalues of $\mathbf{D}^{(N)}$ have *negative real part*

$$\begin{aligned} \partial_t \mathbf{u} &= \mathbf{D}^{(N)}\mathbf{u} \qquad \text{well-posed} \\ \partial_t \mathbf{u} &= -\mathbf{D}^{(N)}\mathbf{u} \qquad \text{ill-posed} \end{aligned}$$

  in ill-posed case boundary condition should be at $x = -1$ but it is posed at $x = +1$
  Example: $N = 1$

$$\mathbf{D}^{(N)} = D_{00} = -\frac{2+1}{6} = -\frac{1}{2}$$

$$\partial_t u_0 = -\frac{1}{2}u_0 \qquad \text{bounded; boundary condition on } u_1$$

  For boundary condition at $x = -1$ introduce $\mathbf{D}^{(0)}$

$$D^{(0)}_{ij} = D_{ij} \qquad i, j = 1, ..., N$$

  Thus for

$$\partial_t u = -\partial_x u$$

$$\partial_t u_i = -\sum_{j=1}^{N} D^{(0)}_{ij} u_j + D_{i0} g(t) \qquad \text{for } i = 1, ..., N$$

  Eigenvalues of $\mathbf{D}^{(0)}$ have *positive* real part
  Example: $N = 1$

$$\mathbf{D}^{(0)} = D_{NN} = +\frac{1}{2}$$

  Note:

  - in Fourier real part vanishes: ⇒ no blow-up
    periodic boundary conditions are well-posed for both directions of propagation

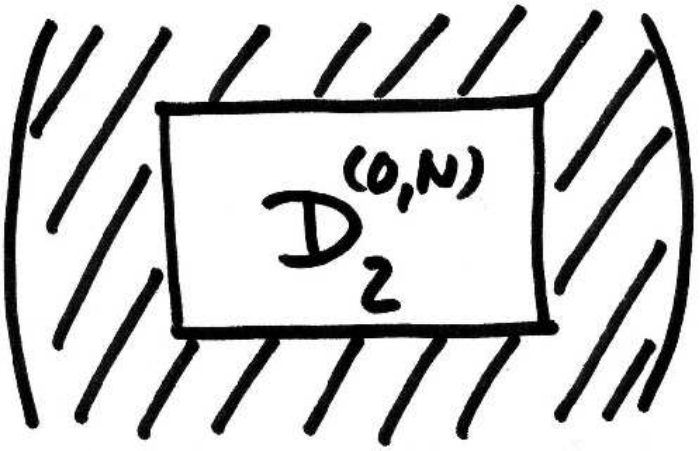
- $\mathbf{D}^{(N)}$ is not normal ($\mathbf{D}^{+}\mathbf{D} \neq \mathbf{D}\mathbf{D}^{+}$) ⇒ similarity transformation S to diagonal form not unitary
$$||\mathbf{u}|| \neq ||\mathbf{Su}||$$
For any *fixed* N $||\mathbf{u}||$ is bounded if $||\mathbf{Su}||$ is bounded
But constant relating $||\mathbf{u}||$ and $||\mathbf{Su}||$ could diverge for $N \to \infty$
⇒ stability is not guaranteed for $N \to \infty$ if scalar equation is stable.

- eigenvalues of $\mathbf{D}^{(N)}$ and $\mathbf{D}^{(0)}$ are $\mathcal{O}(N^2)$
⇒ stability limits for wave equation will involve
$$\Delta t \leq \mathcal{O}(N^{-2})$$
larger eigenvalues reflect the close grid spacing near the boundary, $\Delta x = \mathcal{O}(N^{-2})$

7.3.2 Diffusion Equation: Second Derivative

Consider
$$\partial_t u = \partial_x^2 u \qquad \alpha_{0,N}\, u + \beta_{0,N}\, \partial_x u = \gamma_{0,N} \qquad \text{at } x = \pm 1$$

a) Fixed Boundary Values $\alpha = 1$**,** $\beta = 0$
unknowns
$$u_1, u_2, ..., u_{N-1}$$
known
$$u_0 = \gamma_0 \qquad u_N = \gamma_N$$
Reduced $(n-1) \times (n-1)$ differentiation matrix for second derivative
$$D_{2,ij}^{(0,N)} = (D^2)_{ij} \qquad i, j = 1, ..., N-1$$



then
$$\partial_t u_i = \sum_{j=1}^{N-1} D_{2,ij}^{(0,N)}\, u_j + (D^2)_{i0}\gamma_0 + (D^2)_{iN}\gamma_N \qquad \text{for } i = 1, ..., N-1$$

Note:

- again the 2^{nd} derivative is calculated by using *all* values of u, including the fixed prescribed boundary values
- for transformation to $\tilde{u}_k$ via FFT use all grid points
information for $\partial_x^2 u$ is, however, discarded at the boundaries

Eigenvalues
exact eigenvalues of $\partial_x^2 u$ with $u(\pm 1) = 0$:

- $\sin qx$ is eigenfunction of ∂_x^2 for $q = \frac{\pi}{L} n = \frac{\pi}{2} n$. $\Rightarrow$ eigenvalues $\lambda_n = -\frac{\pi^2}{4} n^2$
- all functions that vanish at $x = \pm 1$ can be expanded in terms of $\sin qx$ with $q = \frac{\pi}{L} n = \frac{\pi}{2} n$
$\Rightarrow$ $\sin qx$ form a complete set $\Rightarrow$ no other eigenfunctions

eigenvalues of $\mathbf{D}_2^{(0,N)}$:

- all eigenvalues are real and negative
- eigenvalues are $\mathcal{O}(N^4)$ reflecting the small grid spacing near the boundaries.

b) Fixed Flux: $\alpha = 0$, $\beta = 1$
Need other modification of $\mathbf{D}^2$:

- u_0 and u_N now unknown $\Rightarrow$ $(n+1) \times (n+1)$ matrix
- $\partial_x u_0$ and $\partial_x u_N$ are known
$\Rightarrow$ $\partial_x u_i$ is calculated with $\mathbf{D}$ only for $i = 1, ..., N-1$

$$\hat{D}_{ij}^{(0,N)} = \begin{cases} D_{ij} & 1 \leq i \leq N-1 \\ 0 \quad i = 0 & \textbf{or} \qquad i = N \end{cases}$$

$$\partial_x u_i = \sum_{j=0}^{N} \hat{D}_{ij}^{(0,N)} u_j + \delta_{i,0}\gamma_0 + \delta_{i,N}\gamma_N \qquad i = 0, ..., N$$

- 2^{nd} derivative

$$\partial_x^2 u_i = \sum_{j=0}^{N} D_{ij} \partial_x u_j = \sum_{j,k=0}^{N} D_{ij} \hat{D}_{jk}^{(0,N)} u_k + D_{ij}\delta_{j,0}\gamma_0 + D_{ij}\delta_{j,N}\gamma_N$$

- Diffusion equation

$$\underbrace{\partial_t u_i = \sum_{j,k=0}^{N} D_{ij} \hat{D}_{jk}^{(0,N)} u_k}_{\text{apply e.g. Crank-Nicholson}} \quad + \quad \underbrace{D_{i0}\gamma_0 + D_{iN}\gamma_N}_{\text{inhomogeneous terms}}$$

$$\frac{1}{\Delta t}\left(\mathbf{u}^{n+1} - \mathbf{u}^n\right) = \theta \mathbf{D}\hat{\mathbf{D}}^{(0,N)}\mathbf{u}^{n+1} + (1-\theta)\mathbf{D}\hat{\mathbf{D}}^{(0,N)}\mathbf{u}^n + D_{i0}\gamma_0 + D_{iN}\gamma_N$$

Note:

- derivative at boundary is calculated also with spectral accuracy; in finite difference schemes they are one-sided: reduced accuracy
- Crank-Nicholson for fixed boundary values similar.

7.4 Discussion of Time-Stepping Methods for Chebyshev

Based on analysis of

$$\frac{du}{dt} = \lambda u$$

which scheme has range of Δt in which it is absolutely for given $\lambda \in \mathbb{C}$

Main aspect: not only $D_2^{(0,N)}$ but also $D^{(N)}$ has negative real part

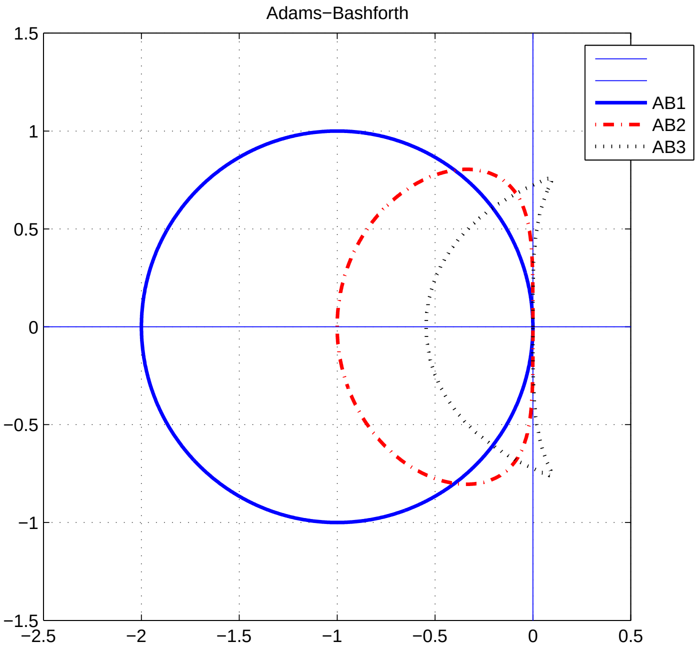
7.4.1 Adams-Bashforth

AB1= forward Euler
AB2

$$u^{n+1} = u^n + \Delta t \left(\frac{3}{2} f^n - \frac{1}{2} f^{n-1} \right)$$

AB3

$$u^{n+1} = u^n + \Delta t \left(\frac{23}{12} f^n - \frac{16}{12} f^{n-1} + \frac{5}{12} f^{n-2} \right)$$



Since the eigenvalues of the odd Chebyshev derivatives have non-zero (negative) real part all three schemes have stable regions not only for diffusion but also for wave equation.

Stability limits:

wave equation

$$\Delta t_{max} = \mathcal{O}(\frac{1}{N^2})$$

diffusion equation

$$\Delta t_{max} = \mathcal{O}(\frac{1}{N^4})$$

strong motivation for implicit scheme

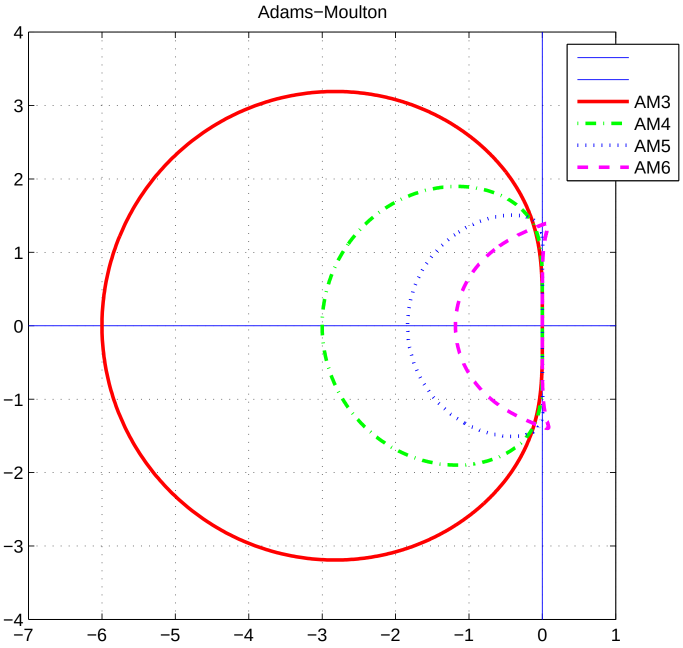
7.4.2 Adams-Moulton

AM1=backward Euler

AM2=Crank-Nicholson

AM3

$$u^{n+1} = u^n + \Delta t \left(\frac{5}{12} f^{n+1} + \frac{8}{12} f^n - \frac{1}{12} f^{n-1} \right)$$



backward Euler and Crank-Nicholson remain unconditionally stable for both equations

AM3: now stable for small Δt ; but still implicit scheme

Notes:

- Crank-Nicholson damps large wavenumbers only weakly, 2^{nd} order in time
- backward Euler damps large wavenumbers strongly: very robust, but only 1^{st}order in time
- if high wavenumbers arise from non-smooth initial conditions: take a few backward Euler steps

7.4.3 Backward-Difference Schemes

this class of schemes is obtained by obtaining interpolant for $u(t)$ and taking its derivative as the left-hand-side of differential equation

$$p_m(t) = \sum_{k=0}^{m-1} u(t_{n+1-k}) L_k^{(m)}(t)$$

with Lagrange polynomials

$$L_k^{(m)}(t) = \prod_{k \neq l=0}^{m-1} \frac{t - t_{n+1-l}}{t_{n+1-k} - t_{n+1-l}}$$

to get derivative

$$\left.\frac{du}{dt}\right|_{t_{n+1}} = \left.\frac{d}{dt} p_m(t)\right|_{t_{n+1}}$$

1. $m = 2$

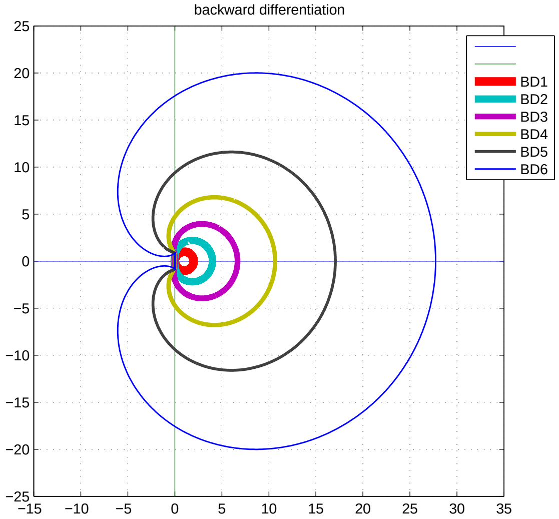
$$p_2(t) = \frac{u_{n+1} - u_n}{t_{n+1} - t_n}(t - t_n) + u_n$$

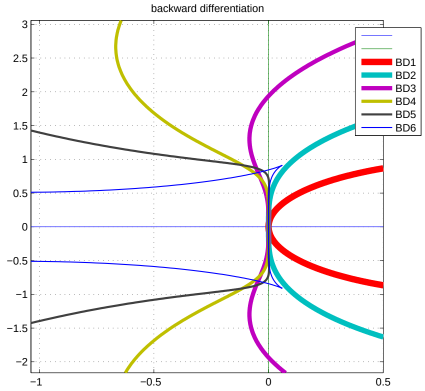
$$\left.\frac{d}{dt} p_2(t)\right|_{t_{n+1}} = \frac{u_{n+1} - u_n}{t_{n+1} - t_n} = f(u^{n+1})$$

   thus: BD1=backward Euler

2. $m = 3$ yields BD2

$$\frac{3}{2} u^{n+1} - 2u^n + \frac{1}{2} u^{n-1} = \Delta t\, f^{n+1}$$





Neumann Analysis for BD2:

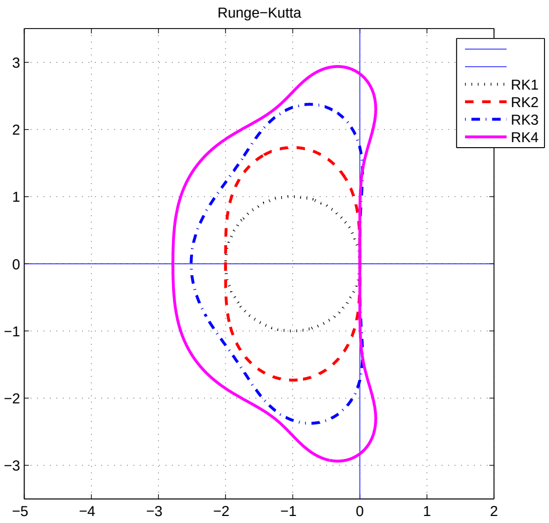
$$\frac{3}{2} z - 2 + \frac{1}{2z} - \Delta t \lambda z = 0$$

$$z_{1,2} = \frac{2 \pm \sqrt{1+2\Delta t\lambda}}{3-2\Delta t\lambda} \rightarrow \mp \frac{1}{\sqrt{2\Delta t|\lambda|}} \rightarrow 0 \qquad \text{for } \Delta t|\lambda| \rightarrow \infty$$

Note:

- BD1 and BD2 are unconditionally stable. BD3 and higher are not unconditionally stable
- BD2 damps high wavenumbers strongly (although not as strongly as BE) and is 2^{nd} order in time
  compared to Crank-Nicholson it needs more storage since it uses u^{n-1}

7.4.4 Runge-Kutta



For Chebyshev also RK2 stable for wave equation - was not the case for Fourier

7.4.5 Semi-Implicit Schemes

Consider diffusion equation with nonlinearity

$$\partial_t u = \underbrace{\partial_x^2 u}_{CN} + \underbrace{f(u)}_{AB2} \qquad u(x=0)=\gamma_0 \qquad u(x=L)=\gamma_N$$

$$u^{n+1} = u^n + \Delta t\left(\theta\partial_x^2 u^{n+1} + (1-\theta)\partial_x^2 u^n\right) + \Delta t\left(\frac{3}{2}f(u^n) - \frac{1}{2}f(u^{n-1})\right)$$

Calculate derivatives with differentiation matrix
⇒ boundary conditions enter

$$\partial_x^2 u_i = \sum_j D_{2,ij}^{(0,N)} u_j + D_{i0}^2\gamma_0 + D_{iN}^2\gamma_N \qquad i = 1, ..., N-1$$

remember

$$D^{(0,N)}_{2,ij} = (D^2)_{ij} \qquad i,j = 1,...,N-1$$

insert in scheme

$$\begin{aligned}\sum_j (\delta_{ij} - \Delta t\theta D^{(0,N)}_{2,ij})u^{n+1}_j &= \sum_j (\delta_{ij} + \Delta t(1-\theta)D^{(0,N)}_{2,ij})u^n_j + \Delta t\left(D^2_{i0}\gamma_0 + D^2_{iN}\gamma_N\right) + \\ &\quad \Delta t\left(\frac{3}{2}f_i(u^n) - \frac{1}{2}f_i(u^{n-1})\right) \qquad i = 1,...,N-1\end{aligned}$$

Notes:

- Need to invert $\delta_{ij} - \Delta t\theta D^{(0,N)}_{2,ij}$: constant matrix ⇒ only one matrix inversion
- in the algorithm $D^{(0,N)}_{2,ij}$ is effectively a $N-2 \times N-2$ matrix; but it is not the same matrix as D^2 for $N-2$ nodes!
- if boundary condition depends on time
  either CN
  $$\Delta t\left(\theta D^2_{i0}\gamma_0(t_{n+1}) + (1-\theta)D^2_{i0}\gamma_0(t_{n+1})\right)$$
  or AB2
  $$\Delta t\left(\frac{3}{2}D^2_{i0}\gamma_0(t_n) - \frac{1}{2}D^2_{i0}\gamma_0(t_{n-1})\right)$$

[4]

7.4.6 Exponential Time-Differencing

Consider again

$$\partial_t u = \partial^2_x u + f(u) \qquad 0 < x < L \qquad \text{b.c. at } x = 0, L$$

Using the Chebyshev differentiation matrix $\mathbf{D}^2$ this can be integrated formally

$$\mathbf{u}^{n+1} = e^{\mathbf{D}^2\Delta t}\mathbf{u}^n + e^{\mathbf{D}^2\Delta t}\int_0^{\Delta t} e^{-\mathbf{D}^2 t'}\mathbf{f}(t+t')dt'$$

where f denotes the vector $(f_1, \ldots, f_N)$.

For a EDTFE we approximate this as

$$\mathbf{u}^{n+1} = e^{\Delta t\mathbf{D}^2}\mathbf{u}^n + \Delta t\,\mathbf{E}_0(\Delta t\mathbf{D}^2)\,\mathbf{f}(t_n) \tag{19}$$

[4] Include CNAB for Chebyshev with FFT:

$$(I + \Delta tD)^{-1} = (FF^{-1} + \Delta tF^{-1}\tilde{D}F)^{-1} = (F^{-1}(I + \Delta t\tilde{D})F)^{-1} = F^{-1}(I + \Delta t\tilde{D})^{-1}F$$

with

$$\mathbf{E}_0(\Delta t\mathbf{D}^2) = (\Delta t\mathbf{D})^{-2}\left(\mathbf{I} - e^{-\mathbf{D}^2\Delta t}\right)$$

As in the Fourier case the evaluation of $\mathbf{E}_0$ suffers from round-off through cancellations. Even worse cancellations for $\mathbf{E}_i$ in EDTRK4 (cf. (10)). Using Taylor's formula is not straightforward. Use Cauchy integral formula for matrices **A** [6, 7].

Consider

$$\Phi(\mathbf{A}) = \frac{1}{2\pi i}\oint_{\mathcal{C}} f(t)\,(t\mathbf{I}-\mathbf{A})^{-1}\,dt$$

Assume A can be diagonalized

$$\mathbf{A} = \mathbf{S}\mathbf{\Lambda}\mathbf{S}^{-1}$$

with $\mathbf{\Lambda} = diag(\lambda_1, \lambda_2, \ldots, \lambda_n)$

$$\begin{aligned}(t\mathbf{I}-\mathbf{A})^{-1} &= \left(t\mathbf{S}\mathbf{I}\mathbf{S}^{-1} - \mathbf{S}\mathbf{\Lambda}\mathbf{S}^{-1}\right)^{-1} = \left\{\mathbf{S}\,(t\mathbf{I}-\mathbf{\Lambda})\,\mathbf{S}^{-1}\right\}^{-1}\\ &= \mathbf{S}\,(t\mathbf{I}-\mathbf{\Lambda})^{-1}\,\mathbf{S}^{-1} = \mathbf{S}\,diag\left(\frac{1}{t-\lambda_1},\ldots,\frac{1}{t-\lambda_n}\right)\mathbf{S}^{-1}\end{aligned}$$

Since S does not depend on $t \in \mathbb{C}$

$$\begin{aligned}\Phi(\mathbf{A}) &= \mathbf{S}\,\frac{1}{2\pi i}\oint_{\mathcal{C}} f(t) diag\left(\frac{1}{t-\lambda_1},\ldots,\frac{1}{t-\lambda_n}\right)dt\,\mathbf{S}^{-1}\\ &= \mathbf{S}\,diag\left(\frac{1}{2\pi i}\oint_{\mathcal{C}} f(t)\frac{1}{t-\lambda_1}dt,\ldots,\frac{1}{2\pi i}\oint_{\mathcal{C}} f(t)\frac{1}{t-\lambda_n}dt\right)\mathbf{S}^{-1}\end{aligned}$$

If $\mathcal{C}$ encloses λ_i

$$\frac{1}{2\pi i}\oint_{\mathcal{C}} f(t)\frac{1}{t-\lambda_i}dt = f(\lambda_i)$$

If $\mathcal{C}$ encloses **all** eigenvalues of A one gets

$$\Phi(\mathbf{A}) = \mathbf{S}\;diag\left(f(\lambda_1),\ldots,f(\lambda_n)\right)\mathbf{S}^{-1} = f(\mathbf{A})$$

and

$$f(\mathbf{A}) = \frac{1}{2\pi i}\oint_{\mathcal{C}} f(t)\,(t\mathbf{I}-\mathbf{A})^{-1}\,dt \qquad (20)$$

Notes:

- sample code for the Allen-Cahn equation, $f(u) = u - u^3$, is in the appendix of [7]
- the contour integral can be evaluated using the trapezoidal rule
- simplest contour is a circle with radius R centered at $t = 0$
- eigenvalues of $\mathbf{D}^2$ grow like $N^4 \Rightarrow R$ has to be chosen large enough
  - e^t grows and oscillates rapidly for ranges of large complex t (cf. (10))

  - more integration points are needed for the integral
  - possibly other contour shapes preferrable (e.g. elliptic close to real axis or parabolic)

Boundary conditions:

1. Fixed boundary values:
$$u_0 = \gamma_0 \qquad u_N = \gamma_N$$
using the modified differentiation matrix $\mathbf{D}_2^{(0,N)}$ we have $N-1$ unknowns $u_1, \ldots, u_{N-1}$,
$$\partial_t u_i = \sum_{j=1}^{N-1} D_{2,ij}^{(0,N)}\, u_j + (D^2)_{i0}\gamma_0 + (D^2)_{iN}\gamma_N + f_i(\mathbf{u}) \qquad \text{for } i = 1, ..., N-1$$
Two possibilities:
   (a) Shift solution to make boundary conditions homogeneous
   $$\mathbf{u} = \mathbf{U} + \mathbf{u}_b$$
   with
   $$u_b(x) = \gamma_0 + (\gamma_N - \gamma_0)\,\frac{x}{L}$$
   $\mathbf{U}$ satisfies now Dirichlet boundary conditions and can be determined using (19) or its RK4 version with $\mathbf{D}^2$ replaced by $\mathbf{D}_2^{(0,N)}$
   (b) Can include the inhomogenous terms $(D^2)_{i0}\gamma_0 + (D^2)_{iN}\gamma_N$ in $\mathbf{f}$.

2. Fixed flux boundary conditions
$$\partial_x u = \gamma_{0,N} \qquad \text{at} \quad x = 0, L$$
$$\partial_t u_i = \sum_{j,k=0}^{N} D_{ij}\hat{D}_{jk}^{(0,N)} u_k + D_{i0}\gamma_0 + D_{iN}\gamma_N + f_i(\mathbf{u}) \qquad \text{for } i = 0, \ldots, N$$
For $\gamma_0 \neq \gamma_N$ the transformation to Neumann condition would induce an additional term since $\partial_x^2 u_b \neq 0$.
Probably it is preferrabe to include the inhomogeneous terms in $\mathbf{f}$.

8 Initial-Boundary-Value Problems: Galerkin Method

Galerkin method:
unknowns are the expansion coefficients, no spatial grid is introduced

Implementation of boundary conditions is different for Galerkin and for pseudospectral:

- pseudospectral: we have grid points ⇒ boundary values available
- Galerkin: no grid points, equations obtained by projections
⇒ modify expansion functions or projection

8.1 Review Fourier Case

$$\partial_t u = Su \qquad 0 \leq x \leq 2\pi \qquad \text{periodic b.c.}$$

Expand u

$$P_N(u) = \sum_{k=-N}^{N} u_k(t)e^{ikx}$$

replace u by projection $P_N(u)$ in PDE

$$\partial_t P_N(u) - S\, P_N(u) = 0$$

the expansion coefficients are determined by the condition that equation be satisfied in subspace spanned by the e^{ikx}, $-N \leq k \leq N$, i.e. error orthogonal to that subspace
Project onto e^{ilx}, $-N \leq l \leq N$

$$\langle e^{ilx}, \partial_t P_N(u) - S\, P_N(U)\rangle = 0$$

Orthogonality of e^{ilx}-modes

$$\partial_t u_l - \int_0^{2\pi} e^{-ilx} S\, P_N(u) = 0$$

e.g. for $S = \partial_x$

$$\begin{aligned} \partial_t u_l - \int e^{-ilx} \sum_k (ik) u_k e^{ikx} &= 0 \\ \partial_t u_l - il u_l &= 0 \end{aligned}$$

Notes:

- no aliasing error since transforms are calculated exactly
- nonlinear terms and space-dependent terms require convolution: slow
- no grid: preserves translation symmetry
- boundary conditions:
  each Fourier mode satisfies the boundary conditions *individually*

8.2 Chebyshev Galerkin

Consider

$$\partial_t u = \partial_x u \qquad -1 \leq x \leq +1, \qquad u(x=+1,t) = g(t)$$

Expand

$$P_N(u) = \sum_{k=0}^{N} u_k(t) T_k(x)$$

project back onto $T_l(x)$

$$\langle T_l, \partial_t P_N(u) - \partial_x P_N(u)\rangle = 0$$

$$\partial_t u_l(t) = \sum_{k=0}^{N} \langle T_l(x), \partial_x T_k(x)\rangle \, u_k(t)$$

with

$$\langle u_1(x), u_2(x)\rangle = \int_{-1}^{+1} u_1(x) u_2(x) \frac{1}{\sqrt{1-x^2}}\, dx$$

Where are the boundary conditions?

Note:

- the $T_k(x)$ *do not* satisfy the boundary conditions individually

8.2.1 Modification of Set of Basis Functions

Construct new complete set of functions, each of which satisfies the boundary conditions.

Example: Dirichlet condition $g(t) = 0$

Since

$$T_k(x = +1) = 1$$

introduce

$$\hat{T}_k(x) = T_k(x) - T_0(x), \qquad k \geq 1$$

each $\hat{T}_k$ satisfies boundary condition.

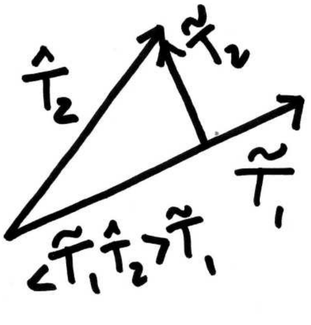
Note:

- modified functions may not be orthogonal any more

$$\langle \hat{T}_l(x), \hat{T}_k(x)\rangle = \underbrace{\langle T_k, T_l\rangle}_{\propto \delta_{kl}} - \underbrace{\langle T_k T_0\rangle}_{=0} - \underbrace{\langle T_0 T_l\rangle}_{=0} + \underbrace{\langle T_0 T_0\rangle}_{=\pi}$$

- could orthogonalize the set with Gram-Schmidt procedure

$$\begin{aligned} \tilde{T}_1 &= \hat{T}_1 \\ \tilde{T}_2 &= \hat{T}_2 - \langle \tilde{T}_1 \hat{T}_2\rangle \, \tilde{T}_1 \\ \tilde{T}_3 &= \hat{T}_3 - \langle \tilde{T}_1 \hat{T}_3\rangle \tilde{T}_1 - \langle \tilde{T}_2 \hat{T}_3\rangle \tilde{T}_2 \\ &\ldots \end{aligned}$$



- procedure is not very flexible, expansion functions have to be changed whenever boundary conditions are changed.

8.2.2 Chebyshev Tau-Method

To be satisfied

$$\begin{aligned} \partial_t u &= \partial_x u \\ u(+1,t) &= g(t) \end{aligned}$$

i.e. boundary condition represents one more condition on the expansion coefficients
$\Rightarrow$ introduce 1 extra unknown

Expand in $N+2$ modes

$$P_{N+1}(u) = \sum_{k=0}^{N} u_k T_k(x) + u_{N+1} T_{N+1}(x)$$

Project PDE onto T_0,...$T_N \Rightarrow N+1$ equations

$$\langle T_l, \partial_t P_{N+1}(u) - \partial_x P_{N+1}(u) \rangle = 0 \qquad 0 \leq l \leq N$$

satisfy boundary condition

$$\sum_{k=0}^{N} u_k T_k(x=+1) + u_{N+1} T_{N+1}(x=+1) = g(t)$$

Use orthogonality

$$c_l \partial_t u_l = \sum_{k=0}^{N+1} u_k \langle T_l, \partial_x T_k \rangle$$

and $T_k(x=1) = 1$

$$\sum_{k=0}^{N+1} u_k = g(t)$$

Thus: $N+1$ equations for $N+1$ unknowns. Should work.

Note:

- For p boundary conditions expand in $N+1+p$ modes and project PDE onto first $N+1$ modes and use remaining p modes to satisfy boundary conditions.

Spurious Instabilities

$\tau-$method can lead to spurious instabilities and eigenvalues.

Example: incompressible Stokes equation in two dimensions

$$\partial_t \mathbf{v} = -\frac{1}{\rho} \nabla p + \nu \Delta \mathbf{v} \qquad \nabla \cdot \mathbf{v} = 0$$

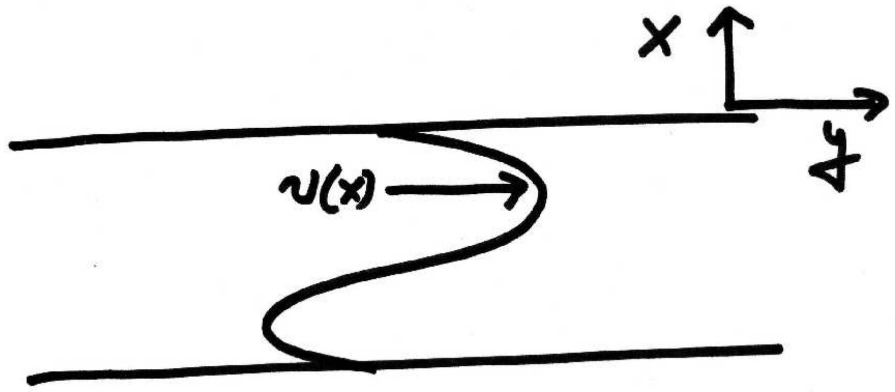
Introduce streamfunction ψ and vorticity ζ

$$\begin{aligned} \mathbf{v} &= (-\partial_y \psi, \partial_x \psi) = -\nabla \times (\psi \hat{k}) \\ \zeta &= (\nabla \times \mathbf{v})_z = \nabla^2 \psi \end{aligned}$$

eliminate pressure from Stokes by taking curl

$$\begin{aligned} \partial_t \zeta &= \nu \Delta \zeta \\ \zeta &= \nabla^2 \psi \end{aligned}$$

Consider parallel channel flow with v depending only on the transverse coordinate x: $\mathbf{v} = \mathbf{v}(x)$



$$\begin{aligned} \partial_t \zeta &= \nu \partial_x^2 \zeta & (21) \\ \zeta &= \partial_x^2 \psi & (22) \end{aligned}$$

Boundary conditions at $x = 0, L$

$$\begin{aligned} v_x &= 0 \qquad \Rightarrow \partial_y \psi = 0 \\ v_y &= 0 \qquad \Rightarrow \partial_x \psi = 0 \end{aligned}$$

Boundary condition $\partial_y \psi$ implies ψ is constant along the wall. If there is not net flux through the channel then ψ has to be equal on both sides of the channel

$$\psi = 0 \qquad x = 0, L$$

Can combine both equations (21,22) into single equation for ψ

$$\partial_t \partial_x^2 \psi = \nu \partial_x^4 \psi$$

with 4 boundary conditions

$$\psi = 0 \qquad \partial_x \psi = 0 \qquad \text{at } x = 0, L$$

Ansatz

$$\psi = e^{\sigma t}\Psi(x)$$
$$\sigma\partial_x^2\Psi = \nu\partial_x^4\Psi$$

Expand

$$\Psi(x) = \sum_{k=0}^{N}\Psi_k T_k(x) \qquad \partial_x^2\Psi = \sum_{k=0}^{N} b_k^{(2)} T_k(x) \qquad \partial_x^4\Psi = \sum_{k=0}^{N} b_k^{(4)} T_k(x)$$

Results for eigenvalues

N	σ_1	σ_2
10	-9.86966	$4,272$
15	-9.86960	$29,439$
20	-9.86960	$111,226$

Notes:

- spurious positive eigenvalues
$$\sigma_{max} = \mathcal{O}(N^4)$$
- scheme is *unconditionally unstable*, useless for time integration
o.k. to determine eigenvalues as long as spurious eigenvalues are recognized

Rephrase problem [5, 4]

expand

$$\begin{aligned}\psi &= e^{\sigma t}\sum_k \psi_k T_k(x)\\ \zeta &= e^{\sigma t}\sum_k \zeta_k T_k(x)\end{aligned}$$

in PDE

$$\begin{aligned}\sigma\zeta_k &= \nu\zeta_k^{(2)}\\ \zeta_k &= \psi_k^{(2)}\end{aligned}$$

where $\zeta_k^{(2)}$ and $\psi_k^{(2)}$ are coefficients of expansion of $2^{nd}-$derivative

Previously all boundary conditions were imposed on first equation

Physically:
impose no slip condition $v_y = 0$ on Stokes equation

$$\begin{aligned}\sigma\zeta_k \quad &= \quad \nu\zeta_k^{(2)} \qquad 0 \le k \le N-2\\ \partial_x\psi \quad (x=\pm 1) &= \quad 0 \qquad N-1 \le k \le N\end{aligned}$$

impose incompressibility on vorticity equation

$$\begin{aligned}\zeta_k &= \psi_k^{(2)} \qquad 0 \le k \le N-2\\ \psi(x=\pm 1) &= 0 \qquad N-1 \le k \le N\end{aligned}$$

This scheme is stable.

9 Iterative Methods for Implicit Schemes

Consider as simple example nonlinear diffusion equation

$$\partial_t u = \partial_x^2 u + f(u)$$

with Crank-Nicholson for stability or for Newton

$$\frac{u^{n+1} - u^n}{\Delta t} = \theta \partial_x^2 u^{n+1} + (1-\theta)\partial_x^2 u + \theta f(u^{n+1}) + (1-\theta) f(u^n)$$

linearize $f(u^{n+1})$ (for reduced Newton, i.e. only single Newton step)

$$f(u^{n+1}) = f(u^n + u^{n+1} - u^n) = f(u^n) + (u^{n+1} - u^n) f'(u^n) + ...$$

and discretize derivatives (Chebyshev or Fourier or finite differences)

$$\partial_x^2 u \Rightarrow \mathbf{D}_2 u$$

then

$$\left((\frac{1}{\Delta t} - \theta f'(u^n))\mathbf{I} - \theta \mathbf{D}_2 \right) \mathbf{u}^{n+1} = \left((\frac{1}{\Delta t} - \theta f'(u^n))\mathbf{I} + (1-\theta)\mathbf{D}_2 \right) \mathbf{u}^n + f(u^n)$$

Notes:

- in linear case matrix on l.h.s. is constant ⇒ only single matrix inversion
- in general:
  - matrix inversion in each time step
  - for full Newton matrix changes after each iteration
- finite differences: in one dimension only tri-diagonal matrix
- pseudospectral: matrix is *full*, inversion requires $\mathcal{O}(N^3)$ operations
- implicit treatment of nonlinearity is in particular important when nonlinearity contains spatial derivatives, otherwise in many cases sufficient to treat nonlinear term explicitly (e.g. CNAB)

9.1 Simple Iteration

Goal: replace 'solving a matrix equation' by 'multiplying by matrix', which is faster

Consider matrix equation

$$\mathbf{A}\mathbf{x} = \mathbf{b}$$

Seek iterative solution scheme

$$\mathbf{x}_{n+1} = \mathbf{x}_n + \mathbf{g}(\mathbf{x}_n)$$

need to chose g(x) to get convergence to solution

$$\mathbf{x}_{n+1} = \mathbf{x}_n \qquad \Leftrightarrow \qquad \mathbf{A}\mathbf{x}_n = \mathbf{b}$$

simplest attempt

$$\mathbf{g}(\mathbf{x}) = \mathbf{b} - \mathbf{A}\mathbf{x}$$

$$\mathbf{x}_{n+1} = (\mathbf{I} - \mathbf{A})\mathbf{x}_n + \mathbf{b} \equiv \mathbf{G}\mathbf{x} + \mathbf{b}$$

check whether solution is a *stable* fixed point: consider evolution of error

$$\delta_{\mathbf{n}} = \mathbf{x}_n - \mathbf{x}_e$$

$$\begin{aligned}\delta_{n+1} &= \mathbf{x}_{n+1} - \mathbf{x}_e = (\mathbf{I} - \mathbf{A})\mathbf{x}_n + \underbrace{\mathbf{b}}_{\mathbf{A}\mathbf{x}_e} - \mathbf{x}_e \\ &= (\mathbf{I} - \mathbf{A})(\mathbf{x}_n - \mathbf{x}_e) = (\mathbf{I} - \mathbf{A})\delta_n\end{aligned}$$

thus

$$\delta_{n+1} = \mathbf{G}\delta_n$$

Estimate convergence

$$||\delta_{n+1}|| \leq ||\mathbf{G}||\,||\delta_n||$$

and

$$||\delta_n|| \leq ||\mathbf{G}||^n\,||\delta_0||$$

convergence in the vicinity of the solution guaranteed for

$$||\mathbf{G}|| \leq \alpha < 1$$

If δ_n is eigenvector of $\mathbf{G}$

$$\delta_{n+1} = \mathbf{G}\delta_n = \lambda_i \delta_n$$

$\Rightarrow$ need $\lambda_i \leq \alpha < 1$ for all eigenvalues λ_i

Define *spectral radius* of $\mathbf{G}$

$$\rho(\mathbf{G}) = \max_i |\lambda_i|$$

then we have

$$\text{iteration converges iff} \qquad \rho(\mathbf{G}) \leq \alpha < 1$$

Define convergence rate $\mathcal{R}$ as inverse of number of iterations to decrease δ by factor e

$$\rho(\mathbf{G})^{\frac{1}{\mathcal{R}}} = \frac{1}{e}$$

$$\mathcal{R} = -\ln \rho(\mathbf{G}) > 0$$

Note:

- for special initial conditions that lie in a direction that contracts faster one could have faster convergence. The rate $\mathcal{R}$ is guaranteed.

- for poor initial guess: possibly no convergence at all.

For Crank-Nicholson (in the linear case)

$$\mathbf{A} = \frac{1}{\Delta t}\mathbf{I} - \theta \mathbf{D}_2$$

thus

$$\mathbf{G} = \mathbf{I} - \mathbf{A} = (1 - \frac{1}{\Delta t})\mathbf{I} + \theta \mathbf{D}_2$$

Eigenvalues of G:

$$\begin{aligned} \rho(\mathbf{G}) &= \mathcal{O}(N^2) \qquad \text{Fourier} \\ \rho(\mathbf{G}) &= \mathcal{O}(N^4) \qquad \text{Chebyshev} \end{aligned}$$

$\rho(\mathbf{G}) \gg 1$ *no convergence*.

9.2 Richardson Iteration

Choose g(x) more carefully

$$\mathbf{g}(\mathbf{x}) = \omega\,(\mathbf{b} - \mathbf{A}\mathbf{x})$$

Iteration

$$\mathbf{x}_{n+1} = \mathbf{x}_n + \omega\,(\mathbf{b} - \mathbf{A}\mathbf{x}_n) = \mathbf{G}\mathbf{x}_n + \omega \mathbf{b}$$

with iteration matrix

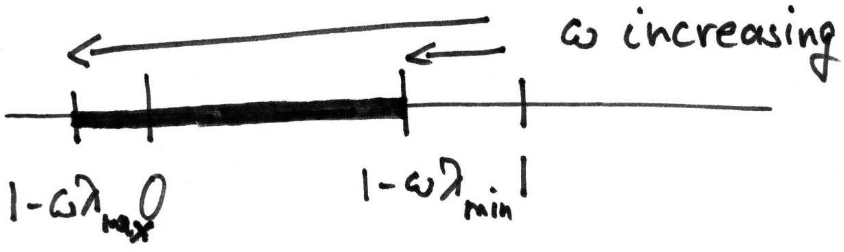
$$\mathbf{G} = \mathbf{I} - \omega \mathbf{A}$$

Choose free parameter ω such that that $\rho(\mathbf{G})$ is minimal, i.e.

$$\max_i |1 - \omega \lambda_i| \text{ minimal}$$

$\mathbf{A} = \frac{1}{\Delta t}\mathbf{I} - \theta \mathbf{D}_2$ has only positive eigenvalues

$$\mathcal{O}(1) = \lambda_{min} \leq \lambda \leq \lambda_{max} = \mathcal{O}(N^{2,4})$$



optimal choice

$$\begin{aligned} 1-\omega\lambda_{max} &= -(1-\omega\lambda_{min}) \\ \omega_{opt} &= \frac{2}{\lambda_{min}+\lambda_{max}} \end{aligned}$$

optimal spectral radius

$$\rho(\mathbf{G})_{min} = \max_i |1-\omega\lambda_i| = 1-\omega_{opt}\lambda_{min} = \frac{\lambda_{max}-\lambda_{min}}{\lambda_{max}+\lambda_{min}}$$

Spectral condition number

$$\kappa = \frac{\lambda_{max}}{\lambda_{min}}$$

$$\rho(\mathbf{G})_{min} = \frac{\kappa-1}{\kappa+1} < 1$$

Notes:

- Richardson iteration can be made to converge by suitable choice of ω independent of spectral radius of original matrix
- Fourier and Chebyshev have large κ

  $$\kappa = \mathcal{O}(N^{2,4}) \quad \Rightarrow \quad \rho \text{ very close to } 1$$

- in Crank-Nicholson

  $$A_{ij} = \left[\frac{1}{\Delta t} - \theta f'(\mathbf{u}^n)\right]\delta_{ij} - \theta D_{2,ij}$$

  the D_2-part corresponds to calculating the second derivative ⇒ can be done using FFT rather than matrix multiplication.

9.3 Preconditioning

Range of eigenvalues of G very large ⇒ *slow* convergence

Further improvement of g(x)

$$\mathbf{x}_{n+1} = \mathbf{x}_n + \omega \underbrace{\mathbf{M}^{-1}}_{\text{preconditioner}} (\mathbf{b} - \mathbf{A}\mathbf{x}_n)$$

Iteration matrix

$$\mathbf{G} = \mathbf{I} - \omega\mathbf{M}^{-1}\mathbf{A}$$

Goal: minimize range of eigenvalues of G

Note:

- optimal would be $\mathbf{M}^{-1} = \mathbf{A}^{-1}$ then $\mathbf{G} = 0$ ⇒ instant convergence
  that is the original problem
- find M that is easy to invert and is close to A, i.e. has similar spectrum
  ⇒ use M from finite difference approximation

9.3.1 Periodic Boundary Conditions: Fourier

For simplicity discuss using simpler problem

$$\partial_t u = \partial_x^2 u \qquad \text{with periodic b.c.}$$

backward Euler:

- spectral ⇒ $\mathbf{A}$, use Fourier because of boundary conditions
- finite differences ⇒ M

Finite differences

$$\frac{1}{\Delta t}(u_j^{n+1} - u_j^n) = \frac{1}{\Delta x^2}\left(u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}\right)$$

written as

$$\mathbf{M}\mathbf{u}^{n+1} = \mathbf{u}^n$$

with

$$\mathbf{M} = \begin{pmatrix} \frac{1}{\Delta t} + \frac{2}{\Delta x^2} & -\frac{1}{\Delta x^2} & 0 & -\frac{1}{\Delta x^2} \\ -\frac{1}{\Delta x^2} & \frac{1}{\Delta t} + \frac{2}{\Delta x^2} & \frac{1}{\Delta x^2} & 0 \\ 0 & \ldots & \ldots & \ldots \\ -\frac{1}{\Delta x^2} & 0 & -\frac{1}{\Delta x^2} & \frac{1}{\Delta t} + \frac{2}{\Delta x^2} \end{pmatrix}$$

Spectral

$$\mathbf{A} = \frac{1}{\Delta t}\mathbf{I} - \mathbf{D}_2$$

Eigenvalues of $\mathrm{M}^{-1}\mathbf{A}$:

M and A have same eigenvectors e^{ilx}
⇒ eigenvalues satisfy

$$\lambda_{\mathbf{M}^{-1}\mathbf{A}} = \frac{\lambda_{\mathbf{A}}}{\lambda_{\mathbf{M}}}$$

eigenvalues of M:

$$M_{ij}e^{ilx_j} = \left(\frac{1}{\Delta t} - \frac{e^{il\Delta x} - 2 + e^{-il\Delta x}}{\Delta x^2}\right) e^{ilx}$$

$$\lambda_{\mathbf{M}} = \frac{1}{\Delta t} + \frac{2}{\Delta x^2}\left(1 - \cos l\Delta x\right)$$

eigenvalues of A

$$\lambda_{\mathbf{A}} = \frac{1}{\Delta t} + l^2$$

⇒

$$\begin{aligned} \lambda_{\mathbf{M}^{-1}\mathbf{A}} &= \frac{\frac{1}{\Delta t} + l^2}{\frac{1}{\Delta t} + \frac{2}{\Delta x^2}\left(1 - \cos l\Delta x\right)} = \\ &= \frac{\frac{\Delta x^2}{\Delta t} + \Delta x^2 l^2}{\frac{\Delta x^2}{\Delta t} + 2(1 - \cos l\Delta x)} \end{aligned}$$

range of eigenvalues

$$\begin{array}{lll} l & \rightarrow & 0 \qquad \lambda_{\mathbf{M}^{-1}\mathbf{A}} \rightarrow 1 \qquad \text{when } \dfrac{\Delta x^2}{\Delta t} \text{ dominates} \\ l & \rightarrow & \dfrac{N}{2} \qquad \Delta x^2 l^2 \rightarrow \left(\dfrac{2\pi}{N}\dfrac{N}{2}\right)^2 = \pi^2 \qquad 1 - \cos l\Delta x \rightarrow 2 \qquad \lambda_{\mathbf{M}^{-1}\mathbf{A}} \rightarrow \dfrac{\pi^2}{4} \end{array}$$

Thus:

- ratio of eigenvalues is $\mathcal{O}(1) \Rightarrow$ fast convergence of iteration.

In practice

$$\mathbf{x}_{n+1} = \mathbf{x}_n + \omega \mathbf{M}^{-1}(\mathbf{b} - \mathbf{A}\mathbf{x}_n)$$

is solved as

$$\mathbf{M}\left(\mathbf{x}_{n+1} - \mathbf{x}_n\right) = \omega\left(\mathbf{b} - \mathbf{A}\mathbf{x}_n\right)$$

Notes:

- for Fourier case (periodic boundary conditions) M is almost tri-diagonal , equation can be solved fast
- for Chebyshev case: also tri-diagonal, but grid points are not equidistant, need finite difference approximation on the same grid

$$\begin{array}{rcl} \partial_x^2 u & = & \dfrac{2}{\Delta x_j(\Delta x_j + \Delta x_{j-1})} u_{j+1} - \dfrac{2}{\Delta x_j \Delta x_{j-1}} u_j + \dfrac{2}{\Delta x_{j-1}(\Delta x_j + \Delta x_{j-1})} u_{j-1} \\ \text{with} & & \Delta x_j = x_{j+1} - x_j \end{array} \qquad \textbf{(23)}$$

  again eigenvalues of $\mathbf{M}^{-1}\mathbf{A}$ can be shown to be $\mathcal{O}(1)$

- for $\kappa \approx 3$ one has $\rho = \frac{\kappa-1}{\kappa+1} \approx \frac{1}{2} \Rightarrow \delta_n = \delta_1 2^{-n}$
  thus

$$\frac{\delta_n}{\delta_1} \approx 10^{-4} \text{ for } n \approx 12$$

  $\Rightarrow$ implicit method with computational effort not much more than explicit

- the matrix multiplication should be done with *fast transform,* e.g. for Fourier

$$\mathbf{A}\mathbf{x}_n = \left(\frac{1}{\Delta t}\mathbf{I} - \mathbf{D}_2\right)\mathbf{x}_n = \frac{1}{\Delta t}\mathbf{x}_n - \mathcal{F}^{-1}\left(-k^2\mathcal{F}(\mathbf{x}_n)\right)$$

9.3.2 Non-Periodic Boundary Conditions: Chebyshev

Need to consider modified matrices, e.g. $D_2^{(0,N)}$, and also in finite differences

1. fixed values $u_{0,N} = \gamma_{0,N}$
   $\Rightarrow$ only $N-1$ unknowns
   Chebyshev: use $D_2^{(0,N)}$

   $$\sum_j \left[\frac{\delta_{ij}}{\Delta t} - \alpha D_{2,ij}^{(0,N)} \right] u_j^{n+1} = r.h.s. + D_{i0}^2 \gamma_0 + D_{iN}^2 \gamma_N$$

   finite differences[5]

   $$\begin{pmatrix} \frac{1}{\Delta t} - \frac{2\alpha}{\Delta x^2} & \frac{\alpha}{\Delta x^2} & 0 & 0 \\ \frac{\alpha}{\Delta x^2} & \frac{1}{\Delta t} - \frac{2\alpha}{\Delta x^2} & \frac{\alpha}{\Delta x^2} & 0 \\ 0 & & ... & \\ 0 & 0 & \frac{\alpha}{\Delta x^2} & \frac{1}{\Delta t} - \frac{2\alpha}{\Delta x^2} \end{pmatrix} = \begin{pmatrix} \\ r.h.s. \end{pmatrix} + \begin{pmatrix} \frac{-1}{\Delta x^2}\gamma_0 \\ 0 \\ ... \\ \frac{-1}{\Delta x^2}\gamma_N \end{pmatrix}$$

2. fixed flux $\partial_x u_{0,N} = \gamma_{0,N}$
   Chebyshev:

   $$\partial_x u_i = \sum_j \hat{D}_{ij}^{(0,N)} u_j + \delta_{i0}\gamma_0 + \delta_{iN}\gamma_N$$

   with

   $$\hat{D}^{(0,N)} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ & & D & \\ & & & \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

   then

   $$\partial_x^2 u_i = \underbrace{\sum_{jk} D_{ij} \hat{D}_{jk}^{(0,N)} u_k}_{\Rightarrow \text{ l.h.s.}} + \underbrace{D_{i0}\gamma_0 + D_{iN}\gamma_N}_{\text{known} \Rightarrow \text{ r.h.s.}}$$

   finite differences:
   introduce virtual points: u_{-1} and u_{N+1}

   $$\partial_x u_0 = \frac{u_1 - u_{-1}}{2\Delta x} = \gamma_0 \qquad \Rightarrow \qquad u_{-1} = u_1 - 2\Delta x \gamma_0$$

   $\Rightarrow$ equation for u_0 is modified

   $$\begin{aligned} \partial_x^2 u_0 &= \frac{u_1 - 2u_0 + u_{-1}}{\Delta x^2} = \frac{u_1 - 2u_0 + (u_1 - 2\Delta x \gamma_0)}{\Delta x^2} \\ &= \underbrace{\frac{-2}{\Delta x^2} u_0 + \frac{2}{\Delta x^2} u_1}_{\text{l.h.s.}} - \underbrace{\frac{2}{\Delta x}\gamma_0}_{\text{r.h.s.}} \end{aligned}$$

[5] The matrix is actually not correct. One has to take into account the non-equidistant grid (cf. (23)).

M is tridiagonal

$$\mathbf{M} = \begin{pmatrix} \frac{1}{\Delta t} - \frac{2}{\Delta x^2} & \frac{\mathbf{2}}{\Delta x^2} & 0 & 0 \\ \frac{1}{\Delta x^2} & \frac{1}{\Delta t} - \frac{2}{\Delta x^2} & \frac{1}{\Delta x^2} & 0 \\ 0 & & \cdots & \\ 0 & & & \end{pmatrix}$$

Notes:

- this leads apparently to eigenvalues $\lambda_{\mathbf{M}^{-1}\mathbf{A}}$in the range $\mathcal{O}(1)$ to $\mathcal{O}(\frac{1}{N}) \Rightarrow \kappa$ becomes large with N, convergence not good.
- apparently better to use $\hat{D}_{ij}^{(0,N)}$ only to calculate derivative for the boundary points and to calculate $\partial_x^2 u$ using D^2 for interior points (see Streett (1983) as referenced in [2] in Sec. 5.2)

Back to reaction-diffusion equation

$$\partial_t u = \partial_x^2 u + f(u)$$

Newton for Crank-Nicholson yields

$$\underbrace{\left[\frac{1}{\Delta t}\mathbf{I} - \alpha \mathbf{D}_2 - \alpha \mathbf{I} \frac{df(u^n)}{du} \right]}_{\mathbf{A}} \mathbf{u}^{n+1} = \text{r.h.s.}$$

Note:

- $\mathbf{A}$ depends on $\mathbf{u}^n \Rightarrow$ eigenvalues depend on $\mathbf{u}^n$ and therefore also on time
  $\Rightarrow$ eigenvalues are in general *not known*
  $\Rightarrow$ choice of ω is not straightforward: *trial and error* 'technique'

9.3.3 First Derivative

Consider simpler problem

$$\frac{du}{dx} = f(x) \qquad \text{with periodic b.c.}$$

i.e.

$$\sum_j D_{ij} u_j = f_i$$

Try usual central differences for finite-difference preconditioning of Fourier differentiation matrix

$$\frac{u_{j+1} - u_{j-1}}{2\Delta x} \quad \Longrightarrow \quad \lambda_{\mathbf{M}} = \frac{2i \sin l\Delta x}{2\Delta x}$$

then

$$\lambda_{\mathbf{M}^{-1}\mathbf{A}} = \frac{il\Delta x}{i \sin l\Delta x} \quad \text{with} \quad -\pi \leq l\Delta x \leq +\pi$$

since $\sin \pi = 0$ one has

- $\lambda_{\mathbf{M}^{-1}\mathbf{A}}$ unbounded $\Rightarrow$ κ unbounded
- no convergence

Possibilities:

1. Could omit higher modes (Orszag)

$$\tilde{u}_k^{(c)} = \begin{cases} \tilde{u}_k & |k| \leq \frac{2N}{3} \\ 0 & \frac{2N}{3} < |k| \leq N \end{cases}$$

   and calculate derivative with $\tilde{u}^{(c)}$

$$\frac{du_j}{dx} = \sum_{k=-N}^{N} ik\tilde{u}_k^{(c)}$$

   Now $l\Delta x \leq \frac{2}{3}\pi$ and range of $\lambda_{\mathbf{M}^{-1}\mathbf{A}}$ is $1 \leq \lambda_{\mathbf{M}^{-1}\mathbf{A}} \leq \frac{2\pi}{3}\sin\frac{2\pi}{3} \approx 2.4$.
   Omitting these modes would be consistent with anti-aliasing for a quadratic nonlinearity.

2. Want $\sin\frac{1}{2}l\Delta x$ instead of $\sin\Delta x$
   Use *staggered* grid: evaluate derivatives and differential equation at $x_{j+1/2}$ but based on the values at the grid points x_j
   Finite differences

$$\left.\frac{du}{dx}\right|_{x_{j+\frac{1}{2}}} = \frac{u_{j+1}-u_j}{\Delta x} = e^{ilx_{j+\frac{1}{2}}}\frac{e^{\frac{1}{2}il\Delta x} - e^{-\frac{1}{2}il\Delta x}}{\Delta x} \qquad \Rightarrow \lambda_{\mathbf{M}} = \frac{2i\,\sin\frac{1}{2}l\Delta x}{\Delta x}$$

   Spectral

$$\left.\frac{du}{dx}\right|_{x_{j+\frac{1}{2}}} = \sum_{l=-N}^{N} il\tilde{u}_k e^{il(x_j+\frac{1}{2}\frac{\pi}{N})} \quad \Rightarrow \quad \lambda_{\mathbf{A}} = il$$

   thus

$$\lambda_{\mathbf{M}^{-1}\mathbf{A}} = \frac{\frac{1}{2}l\Delta x}{\sin\frac{1}{2}l\Delta x} \qquad 1 \leq \lambda_{\mathbf{M}^{-1}\mathbf{A}} \leq \frac{\pi}{2}$$

For wave equation one would get similar problem with central-difference preconditioning

$$\lambda_{\mathbf{M}^{-1}\mathbf{A}} = \frac{\frac{\Delta x}{\Delta t} + il\Delta x}{\frac{\Delta x}{\Delta t} + i\sin l\Delta x} \quad \text{with} \quad -\pi \leq l\Delta x \leq +\pi$$

In implicit scheme Δt may be much larger than Δx:
again $\lambda_{\mathbf{M}^{-1}\mathbf{A}}$ has very large range $\Rightarrow$ poor convergence
Use same method.

Note:

- one-sided difference would not have this problem either:

$$\frac{u_{j+1}-u_j}{\Delta x} \Rightarrow \lambda_M = \frac{e^{il\Delta x}}{\Delta x}$$

$$\lambda_{\mathbf{M}^{-1}\mathbf{A}} = \frac{il\Delta x}{e^{il\Delta x}}$$

10 Spectral Methods and Sturm-Liouville Problems

Spectral methods:

- expansion in complete set of functions
- which functions to choose?

To get complete set consider eigenfunctions of a Sturm-Liouville problem

$$\frac{d}{dx}\left(p(x)\frac{d}{dx}\phi\right) - q(x)\phi + \lambda \qquad \underbrace{w(x)}_{\text{weight function}} \qquad \phi = 0 \qquad -1 \leq x \leq 1$$

with

$$p(x) > 0 \quad \text{in } -1 < x < 1 \qquad w(x), q(x) \geq 0$$

- regular:
$$p(-1) \neq 0 \neq p(+1)$$
- singular:
$$p(-1) = 0 \quad \text{and/or} \quad p(+1) = 0$$

Boundary conditions are homogeneous:

- regular
$$\alpha_\pm \phi(\pm 1) + \beta_\pm \frac{d\phi(\pm 1)}{dx} = 0 \tag{24}$$
- singular
$$p(x)\frac{d\phi}{dx} \to 0 \text{ for } x \to \pm 1 \tag{25}$$
ϕ cannot become too singular near the boundary

Sturm-Liouville problems have non-zero solutions only for certain values of λ: eigenvalues λ_n

Define scalar product:

$$\langle u, v\rangle_w = \int_{-1}^{+1} w(x)u^*(x)v(x)dx$$

eigenfunctions ϕ_k form an *orthonormal complete* set

$$\langle \phi_k, \phi_l\rangle = \delta_{lk}$$

Examples:

1. $p(x) = 1 = w(x)$ and $q(x) = 0$

$$\frac{d^2}{dx^2}\phi + \lambda\phi = 0 \qquad \text{Fourier, regular Sturm-Liouville problem}$$

2. $p(x) = \sqrt{1-x^2}$, $q(x) = 0$, $w(x) = \frac{1}{\sqrt{1-x^2}}$

$$\frac{d}{dx}\left(\sqrt{1-x^2}\frac{d}{dx}\phi\right) + \lambda\frac{1}{\sqrt{1-x^2}}\phi = 0 \qquad \text{Chebyshev, singular}$$

Expand solutions

$$u(x) = \sum_{k=0}^{\infty} u_k\phi_k(x)$$

with

$$u_k = \int w(x)\phi^*(x)u(x)\,dx \qquad \text{projection}$$

Consider convergence of expansion in L_2-norm

$$||u(x) - \sum_{k}^{N} u_k\phi_k(x)|| \to 0 \quad \text{for } N \to \infty$$

Note:

- pointwise convergence only for *almost all* x

Truncation error

$$||\sum_{k=N+1}^{\infty} u_k\phi_k(x)||$$

depends on decay of u_k with k

Want spectral accuracy

$$u_k \leq \mathcal{O}\left(\frac{1}{k^r}\right) \quad \text{for all } r$$

Under what condition is spectral accuracy obtained?

Consider

$$u_k = \int w(x)\phi^*(x)\,u(x)\,dx$$

Previously (Fourier and Chebyshev) did integration by parts.
Use Sturm-Liouville problem

$$w(x)\phi_k^*(x) = \frac{1}{\lambda_k}\left[q\phi_k^* - \frac{d}{dx}\left(p\frac{d\phi^*}{dx}\right)\right]$$

$$\begin{aligned}
u_k &= \frac{1}{\lambda_k}\int u\left\{q\phi_k^* - \frac{d}{dx}\left(p\frac{d}{dx}\phi_k^*\right)\right\}dx = \\
&= \frac{1}{\lambda_k}\int uq\phi_k^* dx + \frac{1}{\lambda_k}\left\{-up\frac{d}{dx}\phi_k^*\Big|_{\pm1} + \int \frac{du}{dx}\,p\,\frac{d\phi^*}{dx}\,dx\right\} = \\
&= \frac{1}{\lambda_k}\int uq\phi_k^* dx + \frac{1}{\lambda_k}\left\{-up\frac{d}{dx}\phi_k^*\Big|_{\pm1} + \frac{du}{dx}p\phi_k^*\Big|_{\pm1} - \int \frac{d}{dx}\left(\frac{du}{dx}p\right)\phi_k^* dx\right\}
\end{aligned}$$

Boundary terms vanish if

$$p\left\{u\frac{d\phi^*}{dx} - \frac{du}{dx}\phi_k^*\right\}\Big|_{\pm1} = 0$$

- regular case

$$\frac{d}{dx}\phi_k^*(\pm1) = -\frac{\alpha_\pm}{\beta_\pm}\phi_k^*(\pm1)$$

$$p\left\{-u\frac{\alpha_\pm}{\beta_\pm}\phi_k^* - \frac{du}{dx}\phi_k^*\right\}\Big|_{\pm1} = 0$$

  thus: u has to satisfy the same strict boundary conditions as ϕ_k

- singular case

$$p\frac{d}{dx}\phi_k \to 0 \quad \text{at boundary}$$

  ⇒ require

$$\phi_k p\frac{du}{dx} \to 0 \quad \text{at boundary}$$

  need only same weak condition on u as on ϕ

$$p\frac{du}{dx} \to 0 \quad \text{at boundary}$$

For large k (cf. Fourier case $\lambda_k = -k^2$ and $d\phi_k/dx = ik\phi_k$)

$$\lambda_k = \mathcal{O}(k^2) \qquad \frac{d\phi_k}{dx} = \mathcal{O}(k)$$

⇒ if boundary conditions are not met one gets

$$u_k = \mathcal{O}(\frac{1}{k})$$

For spectral accuracy *necessary* but *not sufficient*:
u satisfies same boundary conditions as ϕ

To consider higher orders use $L\phi_k = \lambda_k w\phi_k$ to rewrite compact (cf. [2]):

$$u_k = \langle\phi_k, u\rangle_w = \frac{1}{\lambda_k}\langle\frac{1}{w}L\phi_k, u\rangle_w$$

if ϕ and u satisfy the same boundary conditions, then they are in the same function spaces and $\frac{1}{w}L$ is self-adjoint (in explicit calculation above, the w cancel and one can perform the usual integration by parts)

$$u_k = \frac{1}{\lambda_k}\langle\phi_k, \frac{1}{w}Lu\rangle_w = \frac{1}{\lambda_k^2}\langle\frac{1}{w}L\phi_k, \frac{1}{w}Lu\rangle_w = \frac{1}{\lambda_k^2}\langle\phi_k, \frac{1}{w}L\frac{1}{w}Lu\rangle_w$$

The last step can be done if $\frac{1}{w}Lu$ satisfies the same boundary conditions as ϕ.
Introducing

$$u_{(m)} = \frac{1}{w}Lu_{(m-1)}$$

can write

$$u_k = \frac{1}{\lambda_k^r}\langle\phi_k, u_{(r)}\rangle = \mathcal{O}\left(\frac{1}{\lambda_k^r}\right)$$

if

- the $u_{(m)}$ satisfy same boundary conditions as ϕ for all $0 \leq m \leq r-1$
- $u_{(r)}$ is integrable

Conclusion:

- regular Sturm-Liouville problem: since $\left(\frac{1}{w}L\right)^r u$ has to satisfy the boundary conditions these boundary conditions (24) are a very restrictive condition.
  Fourier case is a regular Sturm-Liouville problem: for spectral accuracy we needed that *all* derivatives satisfy periodic boundary conditions.
- singular Sturm-Liouville problem: singular boundary conditions (25) only impose a condition on regularity, do not prescribe any boundary values themselves

Simple example:

$$\partial_t u = \partial_x^2 u + f(x,t) \qquad u(0) = 0 = u(\pi)$$

Could use sine-series

$$u = \sum_k a_k e^{\sigma t} \sin kx$$

since they satisfy related eigenvalue problem

$$\lambda\phi = \partial_x^2\phi \qquad \phi = 0 \quad \text{at } x = 0, \pi$$

But: this is a regular Sturm-Liouville problem with $L = \partial_x^2$ and $w = 1$

Spectral convergence only if

$$u_{(r)}(0) = 0 = u_{(r)}(\pi) \qquad \text{for all } r \tag{26}$$

i.e. if *all* even derivatives have to vanish at the boundary

Most functions that satisfy the original boundary conditions $u(0) = 0 = u(\pi)$ *do not* satisfy the additional conditions (26)

e.g. stationary solution for $f(x,t) = c$

$$u = \frac{1}{2}cx^2 - \frac{1}{2}c\pi x$$

of course $\partial_x^2 u(x = \pm 1) = c \neq 0$.

In fact, expanding in a sine-series one gets

$$a_k = \frac{1}{\pi k^3}\left((-1)^k - 1\right)$$

Thus:

- Expansions in *natural* eigenfunctions of a problem are only good if they satisfy a *singular* Sturm-Liouville problem.
- If they do not satisfy a singular Sturm-Liouville problem one most likely will not get spectral convergence even if the functions look very natural for the problem

11 Spectral Methods for Incompressible Fluid Dynamics

Navier-Stokes equations for fluids arise in a wide range of application.
In many situations the fluid velocities are much smaller than the speed of sound. Density variations can then often be assumed to propagate infinitely fast: the fluid can be assumed to be incompressible,

$$\begin{aligned} \partial_t \mathbf{u} + \mathbf{u} \cdot \nabla \mathbf{u} &= -\nabla p + \mathbf{f} + \nu \Delta \mathbf{u} \\ \nabla \cdot \mathbf{u} &= 0 \end{aligned}$$

Boundary conditions (no-slip condition and wall impermeable)

$$\mathbf{u} = 0 \qquad \text{on boundary}$$

External forces (or imposed pressure gradients) are included in f.

The effectively infinite wave speed leads to numerical challenges.

Mathematically:

- pressure appears in momentum equation, but does not have an evolution equation of its own and has no boundary condition at the walls
- divergence-free condition is an algebraic condition on the velocity, poses constraint on the momentum equation
- could write momentum equation in terms of the vorticity $\omega = \nabla \times \mathbf{u}$
  this would get rid of the pressure, but there is no convincing boundary condition for vorticity
- divergence-free
  - $\Rightarrow$ can introduce streamfunction
    boundary conditions can be tricky (can lead to spurious, destabilizing eigenvalues, cf. Sec.8.2.2)

For concreteness consider flows with boundaries only in one direction, e.g. flow between two plates:

- 1 or 2 directions (x and z) can be approximated by periodic boundary conditions
- no-slip boundary conditions in one direction (y)

There are a number of different approaches that have been taken, we discuss only a few selected ones. Most are formulated in terms of the *primitive variables* $(\mathbf{u}, p)$.

- coupled method: solve momentum equation and incompressibility simultaneously
- Galerkin method with divergence-free basis functions
- operator-splitting methods

Central aspects [3]:

- effectively infinite sound speed requires an *implicit* treatment of the pressure
- viscosity term has highest derivative: often also treated implicitly

The discussion here is following [3].

11.1 Coupled Method

treat u and p simultaneously in coupled equations, usually use *semi-implicit* method

For a first-order method one would get (with an imposed pressure gradient $p_x\mathbf{e}_x$ to drive the flow)

$$\begin{aligned} \frac{1}{\Delta t}\mathbf{u}^{n+1} - \nu\Delta\mathbf{u}^{n+1} + \nabla p^{n+1} &= \mathbf{f}^{n+1} + p_x\mathbf{e}_x + \frac{1}{\Delta t}\mathbf{u}^n - \mathbf{u}^n\cdot\nabla\mathbf{u}^n \qquad (27)\\ \nabla\cdot\mathbf{u}^{n+1} &= 0\\ \mathbf{u}^{n+1} &= 0 \qquad \text{on boundary} \end{aligned}$$

Derivatives implemented via ik in the x-direction and via Chebyshev differentiation matrix in y-direction:

$$\mathbf{u} = \sum \mathbf{U}_k(y,t)e^{ikx} = \sum \tilde{\mathbf{U}}_{km}(t)T_m(y)e^{ikx} \qquad p = \sum P_k(y,t)e^{ikx} = \sum \tilde{P}_{km}(t)T_m(y)e^{ikx} \quad (28)$$

With (28) and $\mathbf{U} = (U,V)$ the Navier-Stokes equation (27) becomes

$$\begin{aligned} \frac{1}{\Delta t}\mathbf{U}_k^{n+1} + \nu k^2\mathbf{U}_k^{n+1} - \nu\partial_y^2\mathbf{U}_k^{n+1} + ikP_k^{n+1}\mathbf{e}_x + \partial_y P_k^{n+1}\mathbf{e}_y &= \mathbf{r}_k \qquad (29)\\ ikU_k^{n+1} + \partial_y V_k^{n+1} &= 0 \end{aligned}$$

with

$$\mathbf{r}_k = \frac{1}{\Delta t}\mathbf{U}_k^n - (\mathbf{u}^n\cdot\nabla\mathbf{u}^n)_k - (p_x\mathbf{e}_\mathbf{x})_k \qquad (30)$$

and boundary condition

$$\mathbf{U}_k^{n+1}(y=\pm1) = 0$$

System can be solved

- directly with iterative method (precondition for the y-derivative)
- using the *influence matrix* method (Kleiser-Schumann)

Discuss here the influence matrix method.

For U one gets from (29)

$$-\nu U'' + \lambda U + ikP = r_x \qquad (31)$$

with

$$U(y=\pm1) = 0$$

and

$$\lambda = \frac{1}{\Delta t} + \nu k^2.$$

For V one gets

$$-\nu V'' + \lambda V + P' = r_y \qquad (32)$$

with boundary condition

$$V(y=\pm1) = 0. \qquad (33)$$

Once P is known U and V can be determined from (31, 32).

To get an equation for the pressure eliminate $\mathbf{U}^{n+1}$ from (27) by taking its divergence and using incompressibility (drop subscript k and superscript $n = 1$)

$$P'' - k^2 P = \nabla \cdot \mathbf{r} \tag{34}$$

We do not have a boundary condition for the pressure. Instead, using $\nabla \cdot \mathbf{U} = 0$ and $\partial_x U(y = \pm 1) = 0$ one gets the boundary condition on V'

$$V'(y = \pm 1) = 0 \tag{35}$$

Thus, the P-equation is coupled to the V-equation through this additional boundary conditions. Need to compute P and V simultaneously using (34,35,32,33)

$$\mathcal{L}\begin{pmatrix} P \\ V \end{pmatrix} = \mathbf{b} \qquad V(y = \pm 1) = 0 = V'(y = \pm 1) \tag{36}$$

with

$$\mathcal{L} = \begin{pmatrix} \partial_y^2 - k^2 & 0 \\ -\partial_y & \nu\partial_y^2 - \lambda \end{pmatrix} \qquad \mathbf{b} = \begin{pmatrix} \nabla \cdot \mathbf{r} \\ -r_y \end{pmatrix}$$

Slightly strange boundary conditions:

- 2^{nd}-order ODE for P but no boundary condition for P
- 2^{nd}-order ODE for V but 4 boundary conditions for V

Consider auxiliary problem, assuming there is a boundary condition for P,

$$\mathcal{L}\begin{pmatrix} P \\ V \end{pmatrix} = \mathbf{b} \qquad P(y = \pm 1) = P_\pm \qquad V(y = \pm 1) = 0 \tag{37}$$

(36) can be solved by solving 3 versions of (37):

$$\mathcal{L}\begin{pmatrix} P_p \\ V_p \end{pmatrix} = \mathbf{b} \qquad P_p(y = \pm 1) = 0 \qquad V_p(y = \pm 1) = 0 \tag{38}$$

$$\mathcal{L}\begin{pmatrix} P_+ \\ V_+ \end{pmatrix} = 0 \qquad P_+(y = +1) = 1 \qquad P_+(y = -1) = 0 \qquad V_+(y = \pm 1) = 0 \tag{39}$$

$$\mathcal{L}\begin{pmatrix} P_- \\ V_- \end{pmatrix} = 0 \qquad P_-(y = +1) = 0 \qquad P_-(y = -1) = 1 \qquad V_-(y = \pm 1) = 0 \tag{40}$$

Expand the solution to (36) as

$$\begin{pmatrix} P \\ V \end{pmatrix} = \begin{pmatrix} P_p \\ V_p \end{pmatrix} + \delta_+ \begin{pmatrix} P_+ \\ V_+ \end{pmatrix} + \delta_- \begin{pmatrix} P_- \\ V_- \end{pmatrix} \tag{41}$$

and impose the boundary condition of (36)

$$\underbrace{\begin{pmatrix} V_+'(+1) & V_-'(+1) \\ V_+'(-1) & V_-'(-1) \end{pmatrix}}_{M} \begin{pmatrix} \delta_+ \\ \delta_- \end{pmatrix} = -\begin{pmatrix} V_p'(+1) \\ V_p'(-1) \end{pmatrix}$$

Since $\mathcal{L}$ does not depend on the flow $(\mathbf{U}, P)$ the solutions to (39) and to (40) do not depend on the flow:

- $\begin{pmatrix} P_+ \\ V_+ \end{pmatrix}$ and $\begin{pmatrix} P_- \\ V_- \end{pmatrix}$ need to be calculated only once at the beginning of the code
- the influence matrix M can also be calculated initially

Procedure:

1. Compute $\begin{pmatrix} P_p \\ V_p \end{pmatrix}$ which depends on the flow via the inhomogeneity b
2. Compute $\delta_\pm$, which provides the correct boundary conditions $P_\pm$ for (37)
$$P_\pm = \delta_\pm$$
3. With $\delta_\pm$ the solution to (36) is given by (41) (no need to solve (37) explicitly).

Notes:

- in the spectral approach the differential equations in y will be solved using Chebyshev polynomials
- discussion above was done for continuous differentiation operators, not for discrete differentiation (pseudo-spectral collocation points) ⇒ the solution to the equations obtained from taking the divergence of the NS-equation (i.e. (34,32,35)) does not guarantee a divergence-free solution. Error is estimated to be (with N_y grid points in y-direction)
$$\mathcal{O}\left(\frac{N_y}{\nu\Delta t}\tilde{U}_{kN_y}, \frac{N_y}{\nu\Delta t}\tilde{U}_{kN_y-1}\right)$$
- Correction (τ-correction step) improves also stability limit
- With and without τ-correction code achieves spectral accuracy in space.

11.2 Operator-Splitting Methods

A common way to split the Navier-Stokes equations is into a *velocity step*

$$\frac{1}{\Delta t}\left(\mathbf{u}^{n+1/2} - \mathbf{u}^n\right) - \nu\Delta\mathbf{u}^{n+1/2} = -\mathbf{u}^n \cdot \nabla\mathbf{u} - p_x\mathbf{e}_x \tag{42}$$

with a boundary condition

$$\mathbf{u}^{n+1/2}(y = \pm 1) = \mathbf{g}^{n+1/2}$$

with $\mathrm{g}^{n+1/2}$ to be discussed later. The intermediate velocity field $\mathbf{u}^{n+1/2}$ is not divergence-free. This is achieved with the *pressure step*

$$\begin{aligned} \frac{1}{\Delta t}\left(\mathbf{u}^{n+1} - \mathbf{u}^{n+1/2}\right) + \nabla p^{n+1} &= 0 \\ \nabla \cdot \mathbf{u}^{n+1} &= 0 \end{aligned} \tag{43}$$

with boundary condition (again $\mathbf{u} = (u, v)$)

$$v^{n+1}(y = \pm 1) = 0$$

Note

- counting boundary conditions:
  - after Fourier transformation momentum equation is an algebraic equation in the x-component and a first-order ODE for p in the y-component
  - the incompressibility condition is a first-order ODE
  - for two first-order ODEs expect only two boundary conditions, v at both sides. Not possiblte to impose also boundary conditions on u
- in this formulation time-stepping is only first-order (Euler)
- $\mathbf{u}^{n+1}$ is divergence-free but does not satisfy the no-slip condition exactly:

$$u_{slip} \equiv u(y = \pm 1) \neq 0$$

  - for $\mathbf{g}^{n+1/2} = 0$ one has $u_{slip} = \mathcal{O}(\Delta t)$
  - modified boundary conditions can improve accuracy

$$g_x^{n+1/2} = \Delta t \partial_x p^n \qquad g_y^{n+1/2} = 0 \qquad \Rightarrow \qquad u_{slip} = \mathcal{O}(\Delta t^2)$$

    higher-order conditions are possible

For expansion in Chebyshev modes relevant:

- pressure enters equation only via its gradient
- $T_N'(x_j) = 0$ at all $x_j = \cos\frac{\pi j}{N}$
  $\Rightarrow$ pressure mode p_N does not affect flow field and results in spurious mode

To avoid spurious pressure mode use only $N-1$ Chebyshev modes

$$p(x,y,t) = \sum_k \sum_{m=0}^{N-1} \tilde{P}_{km}(t) T_m(y)\, e^{ikx}$$

and solve the pressure step using the *staggered grid points* as collocation points

$$y_{j+1/2} = \cos\frac{\pi(j+\frac{1}{2})}{N} \qquad j = 0 \ldots N-1$$

The velocity field is expanded as usually

$$\mathbf{u}(x,y,t) = \sum_k \sum_{m=0}^{N} \tilde{\mathbf{U}}_{km}(t) T_m(y)\, e^{ikx}$$

and for the velocity step the usual collocation points are used

$$y_j = \cos\frac{\pi j}{N} \qquad j = 0 \ldots N$$

Notes:

- The pressure mode $\tilde{P}_{00}$ also does not affect the flow. However, a spatially homogeneous pressures is also physically irrelevant ⇒ indeterminacy of $\tilde{P}_{00}$ does not pose a problem.
- Since two different grids are used one needs to interpolate $(\mathbf{u}, p)$ from one grid to the other by evaluating the T_m at the respective grid points. This introduces additional steps in the algorithm (some slowing down).

Velocity Step

drop again subscript k

$$\begin{aligned} \nu\partial_y^2\mathbf{U}^{n+1/2} - \lambda\mathbf{U}^{n+1/2} &= -\mathbf{r} \\ \mathbf{U}^{n+1/2}(y=\pm1) &= \mathbf{g}^{n+1/2}(y=\pm1) \end{aligned}$$

with $\lambda = \frac{1}{\Delta t} + \nu k^2$ and **r** as in (30)

Determine **U** using Chebyshev τ-method using the usual (Gauss-Lobatto) collocation points y_j.

Pressure Step

For transformation between the grids write

$$\hat{U} = (U(y_0), U(y_1), \dots, U(y_N))^t \qquad \hat{V} = (V(y_0), V(y_1), \dots, V(y_N))^t$$

and

$$\hat{P} = \left(P(y_{1/2}), P(y_{3/2}), \dots, P(y_{N-1/2})\right)^t$$

Need to compute the Chebyshev coefficients $\tilde{\mathbf{U}}$ and $\tilde{P}$ for **U** and P based on the values at the respective grid points

$$\tilde{U} = \mathbf{C}_0\hat{U} \qquad \tilde{V} = \mathbf{C}_0\hat{V}$$

and

$$\tilde{P} = \mathbf{C}_+\hat{P}$$

where $\mathbf{C}_0$ and $\mathbf{C}_+$ are the appropriate matrices

Velocity divergence needed on staggered grid points

$$\nabla\cdot\mathbf{u} \rightarrow D\hat{\mathbf{U}} \equiv \left(\mathbf{C}_+^{-1}\mathbf{C}_0\right)\left[ik\hat{U} + \mathbf{C}_0^{-1}D\mathbf{C}_0\hat{V}\right]$$

where D computes the derivative from the Chebyshev coefficients

Pressure gradient needed on regular grid points in momentum equation

$$\nabla p \rightarrow G\hat{P} \equiv \left(\mathbf{C}_0^{-1}\mathbf{C}_+\right)\left(ik\hat{P}, \mathbf{C}_+^{-1}D\mathbf{C}_+\hat{P}\right)$$

Pressure step (43) becomes

$$\begin{aligned} \hat{\mathbf{U}}^{n+1} &= \hat{\mathbf{U}}^{n+1/2} - \Delta t\, G\hat{P} \qquad \text{at interior points } y_j, \quad j = 1\dots N-1 && (44) \\ D\hat{\mathbf{U}}^{n+1} &= 0 \qquad \text{at } y_{j+1/2}, j = 0\dots N-1 && (45) \end{aligned}$$

with

$$\hat{U}_x^{n+1} = \hat{U}_x^{n+1/2} - \Delta t \left(G\hat{P} \right)_x \qquad \text{at } y = \pm 1 \tag{46}$$

$$\hat{U}_y^{n+1} = 0 \qquad \text{at } y = \pm 1 \tag{47}$$

Rewrite these equations to obtain an equation for the pressure. To make use of the divergence condition (45) combine (44) with (47)

$$\hat{\mathbf{U}}^{n+1} = Z \left(\hat{\mathbf{U}}^{n} - \Delta t \, G\hat{P} \right) \qquad \text{at } y_j, \quad j = 0 \ldots N$$

where the matrix Z sets the boundary values of y-componet to 0

Then one can use the divergence condition (45) to eliminate $\hat{\mathbf{U}}^{n+1}$ and obtains an equation for the pressure

$$D\ Z\ G\ \hat{P} = \frac{1}{\Delta t} D\ Z\ \hat{\mathbf{U}}^{n+1/2}$$

Once the pressure is known $\hat{\mathbf{U}}^{n+1}$ can be determined directly from (44-47).

Note:

- for more details on operator-splitting and other schemes for incompressible Navier-Stokes see [2, 3]

A Insertion: Testing of Codes

A few suggestions for how to test codes and identify bugs:

- test each term individually if possible
  - set all but one coefficient in the equation to 0:
    does the code behave *qualitatively* as expected from the equation?
  - compare *quantitatively* with simple *analytical* solutions (possibly with some coefficients set to 0)
- code 'blows up':
  - is it a 'true blow-up': exact solution should not blow up
  - is the blow-up reasonable for this type of scheme for this problem? Stability? Does decreasing dt increase/decrease the growth?
  - is the blow-up a coding error?
- track variables:
  use only few modes so you can *print out*/*plot* what is going on in each time step
- if the code seems not to do what it should it often is a good idea to vary the parameters and see whether the behavior of the code changes as expected (e.g. if a parameter was omitted in an expression the results may not change at all even though the parameters are changed); the response of the code to parameter changes may give an idea for where the error lies.

B Details on Integrating Factor Scheme IFRK4

Some more details for the integrating-factor scheme (keeping in mind that it is usually not as good as the exponential time differencing scheme):

Rewrite (8) with integrating factor e^{k^2t}

$$\partial_t(e^{k^2t}u_k) = k^2e^{k^2t}u_k + e^{k^2t}\partial_t u_k = e^{k^2t}f_k(u) \tag{48}$$

Introduce auxiliary variable $v_k(t) = e^{k^2t}u_k(t)$

$$\partial_t v_k = e^{k^2t}f_k(e^{-l^2t}v_l) \tag{49}$$

Note:

- for nonlinear f the Fourier coefficient f_k depends on all Fourier modes of v

It is natural to consider now suitable time-integration methods to solve equation (49)

Example: Forward Euler

$$\begin{aligned} v_k^{n+1} &= v_k^n + \Delta t\, e^{k^2 t} f_k(e^{-k^2 t} v_k^n) \\ e^{k^2(t+\Delta t)} u_k^{n+1} &= e^{k^2 t} u_k^n + \Delta t\, e^{k^2 t} f_k(u_k^n) \\ u_k^{n+1} &= e^{-k^2 \Delta t}\left(u_k^n + \Delta t\, f_k(u_k^n)\right) \end{aligned}$$

Note:

- with forward Euler integrating factor generates same scheme as the operator-splitting scheme above
- diffusion and other linear terms are treated exactly
- no instability arises from linear term for *any* Δt
- large wave numbers are strongly damped, as they should be (this is also true for operator splitting)
  compare with Crank-Nicholson (in CNAB, say)
  $$u_k^{n+1} = \frac{1 - \frac{1}{2}\Delta t k^2}{1 + \frac{1}{2}\Delta t k^2} u_k^n$$
  for large $k\Delta t$
  $$u_k^{n+1} = -(1 - \frac{4}{\Delta t k^2} + ...) u_k^n$$
  oscillatory behavior and *slow* decay.
- FFT is done on nonlinear term rather than the linear derivative term (cf. operator splitting)
- **But:** fixed points in u depend on the time step Δt and are not computed correctly for large Δt, whereas without the integrating factor the fixed points of the numerical scheme agree exactly with those of the differential equation.

Notes:

- It turns out that the prefactor of the error term is relatively large in particular compared to the exponential time differencing scheme (cf. Boyd, *Chebyshev and Fourier Spectral Methods*[6])

Details for Runge-Kutta:

In Fourier space

$$\partial_t u_k = -k^2 u_k + f_k(u)$$

[6] See also Cox and Matthews, J. Comp. Phys. 176 (2002) 430, who give a detailed comparison and a further advanced method *exponential time differencing*.

For $v_k = e^{k^2 t} u_k$ then

$$\partial_t v_k = e^{k^2 t} f_k(v_l e^{-l^2 t}) = F_k(t, v_l)$$

Note: $F_k(t, v_l)$ depends explicitly on time even if $f(u)$ does not!

Then

$$\begin{aligned}
k_{1k} &= \Delta t F_k(t_n, v_l^n) = \\
&= \Delta t\, e^{k^2 t_n} f_k(v_l^n e^{-l^2 t_n}) = \Delta t e^{k^2 t_n} f_k(u_l^n) \\
k_{2k} &= \Delta t F_k(t_n + \frac{1}{2}\Delta t, v_l^n + \frac{1}{2} k_{1l}) = \\
&= \Delta t\, e^{k^2 (t_n + \Delta t/2)} f_k((v_l^n + \frac{1}{2} k_{1l})\, e^{-l^2 (t_n + \Delta t/2)}) \\
&= \Delta t\, e^{k^2 (t_n + \Delta t/2)} f_k(v_l^n e^{-l^2 t_n} e^{-l^2 \Delta t/2} + \frac{1}{2} k_{1l} e^{-l^2 (t_n + \Delta t/2)}) \\
&= \Delta t\, e^{k^2 (t_n + \Delta t/2)} f_k(u_l^n e^{-l^2 \Delta t/2} + \frac{1}{2} k_{1l} e^{-l^2 (t_n + \Delta t/2)})
\end{aligned}$$

Growing exponentials become very large for large k. Introduce

$$\begin{aligned}
\bar{k}_{1k} &= k_{1k} e^{-k^2 t_n} \\
\bar{k}_{2k} &= k_{2k} e^{-k^2 (t_n + \Delta t/2)} \\
\bar{k}_{3k} &= k_{3k} e^{-k^2 (t_n + \Delta t/2)} \\
\bar{k}_{4k} &= k_{4k} e^{-k^2 (t_n + \Delta t)}
\end{aligned}$$

Then

$$\begin{aligned}
\bar{k}_{1k} &= \Delta t\, f_k(u_l^n) \\
\bar{k}_{2k} &= \Delta t\, f_k(u_l^n e^{-l^2 \Delta t/2} + \frac{1}{2} \bar{k}_{1l} e^{-l^2 \Delta t/2}) \\
&= \Delta t\, f_k \left((u_l^n + \frac{1}{2} \bar{k}_{1l}) e^{-l^2 \Delta t/2} \right) \\
\bar{k}_{3k} &= \Delta t\, f_k \left(u_l^n e^{-l^2 \Delta t/2} + \frac{1}{2} \bar{k}_{2l} \right) \\
\bar{k}_{4k} &= \Delta t\, f_k \left(u_l^n e^{-l^2 \Delta t} + \bar{k}_{3l} e^{-l^2 \Delta t/2} \right)
\end{aligned}$$

$$v_k^{n+1} = v_k^n + \frac{1}{6} \left(k_{1k} + 2 k_{2k} + 2 k_{3k} + k_{4k} \right)$$

$$u_k^{n+1} e^{k^2 (t_n + \Delta t)} = u_k^n e^{k^2 t_n} + \frac{1}{6} e^{k^2 t_n} \left(\bar{k}_{1k} + 2 \bar{k}_{2k} e^{k^2 \Delta t/2} + 2 \bar{k}_{3k} e^{k^2 \Delta t/2} + \bar{k}_{4k} e^{k^2 \Delta t} \right)$$

Thus

$$u_k^{n+1} = u_k^n e^{-k^2 \Delta t} + \frac{1}{6} \left(\bar{k}_{1k} e^{-k^2 \Delta t} + 2 \bar{k}_{2k} e^{-k^2 \Delta t/2} + 2 \bar{k}_{3k} e^{-k^2 \Delta t/2} + \bar{k}_{4k} \right)$$

Note

- In each of the four stages go to real space to evaluate nonlinearity and then transfrom back to Fourier space to get its Fourier components in order to evaluate $\bar{k}_{ik}, i = 1..4$.

C Chebyshev Example: Directional Sensing in Chemotaxis

Levine, Kessler, and Rappel have introduced a model to explain the ability of amoebae (e.g. Dictyostelium discoideum) to sense chemical gradients very sensitively despite the small size of the amoeba (see PNAS 103 (2006) 9761).

The model consists of an activator A, which is generated in response to the external chemical that is to be sensed. The activator is bound to the cell membrane and constitutes the output of the sensing activity (and triggers chemotactic motion), and a diffusing inhibitor B. The inhibitor can attach itself to the membrane (its concentration is denoted B_m) where it can inactivate A.

The model is given by

$$\frac{\partial B}{\partial t} = D\nabla^2 B \qquad \text{inside the cell } -1 < x < +1$$

with boundary ocndition

$$D\frac{\partial B}{\partial n} = k_a S - k_b B.$$

Here $\partial/\partial n$ is the outward normal derivative. In a one-dimension system its sign is opposite on the two sides of the system, $\partial/\partial n = -\partial/\partial x$ at $x = -1$ whereas $\partial/\partial n = +\partial/\partial x$ at $x = +1$. The reactions of the membrane bound species are given by

$$\begin{aligned}\frac{dA}{dt} &= k_a S - k_{-a}A - k_i A B_m \\ \frac{dB_m}{dt} &= k_b B - k_{-b}B_m - k_i A B_m\end{aligned}$$

To implement the boundary conditions with Chebyshev polynomials (using the matrix multiplication approach):

$$\begin{aligned}\frac{\partial B_i}{\partial x} &= \sum_{j=0}^{N} D_{ij}B_j \qquad \text{for} \quad i = 1, \ldots, N-1 \\ \frac{\partial B_0}{\partial x} &= -\frac{1}{D}\left(k_a S_0 - k_b B_0\right) \\ \frac{\partial B_N}{\partial x} &= \frac{1}{D}\left(k_a S_N - k_b B_N\right)\end{aligned}$$

The second derivative is then given by

$$D\frac{\partial^2 B_i}{\partial x^2} = D\sum_{j=1}^{N-1}\sum_{k=0}^{N} D_{ij}D_{jk}B_k - D_{i0}\left(k_a S_0 - k_b B_0\right) + D_{iN}\left(k_a S_N - k_b B_N\right)$$

which can be written as

$$D\frac{\partial^2 B_i}{\partial x^2} = \sum_{k=0}^{N} \tilde{D}_{ik}B_k + k_a\left(-D_{i0}S_0 + D_{iN}S_N\right)$$

with

$$\tilde{D}_{ik} = D \sum_{j=1}^{N-1} D_{ij} D_{jk} - b \begin{pmatrix} -D_{i0} & 0 & 0 & D_{iN} \\ -D_{i0} & \ldots & \ldots & D_{iN} \\ -D_{i0} & \ldots & \ldots & D_{iN} \\ -D_{i0} & 0 & 0 & D_{iN} \end{pmatrix}$$

The equations on the membrane are nonlinear. The implementation of Crank-Nicholson is then done most easily not completely implicitly, i.e. no full Newton iteration sequence is performed to solve the nonlinear equations. Instead only a single iteration is performed (semi-implicit) This is equivalent to expanding the terms at the new time around those at the old time. Specifically

$$\begin{aligned} \alpha A^{n+1} B^{n+1} + (1-\alpha) A^n B^n &= \alpha \left((A^n + \Delta A)(B^n + \Delta B) \right) + (1-\alpha) A^n B^n = \\ &= \alpha \left(A^n B^n + A^n \Delta B + B^n \Delta A + \mathcal{O}(\Delta A \Delta B) \right) + (1-\alpha) A^n B^n = \\ &= \alpha \left(A^{n+1} B^n + A^n B^{n+1} \right) + (1-2\alpha) A^n B^n + \mathcal{O}(\Delta A \Delta B). \end{aligned}$$

Ignoring the term $\mathcal{O}(\Delta A \Delta B)$ is often good enough.

D Background for Homework: Transitions in Reaction-Diffusion Systems

Many systems undergo transitions from steady state to oscillatory ones or from spatially homogeneous ones to states with spatial structure (periodic or more complex)

Examples:

- buckling of a bar or plate upon uniform compression (Euler instability)
- convection of a fluid heated from below: thermal instability through bouyancy or temperature-dependence of surface tension
- fluid between two rotating concentric cylinders: centrifual instability
- solid films adsorbed on substrates with different crystaline structure (cf. Golovin's recent coloquium)
- surface waves on a vertically vibrated liquid
- various chemical reactions: Belousov-Zhabotinsky
  - oscillations:
    in the 1950s Belousov could not get his observations published because the journal reviewers thought such temporal structures were not 'allowed' by the second law of thermodynamics
  - spatial structure:
    Turing suggested (1952) that different diffusion rates of competing chemicals could lead to spatial structures that could underly the formation of spatial structures in biology (segmentation of yellow-jackets, patterning of animal coats...)

Common to these systems is that the temporal or spatial structures arise through instabilities of a simpler (e.g. homogeneous) state. Mathematically, these instabilities are bifurcations at which new solutions come into existence.

General analytical approach:

1. find simpler *basic* state
2. identify instabilities of basic state
3. derive simplified equations that describe the structured state in the *weakly nonlinear* regime
   leads to equations for the amplitude of the unstable modes characterizing the structure: Ginzburg-Landau equations

In homework consider simple model in one spatial dimension for chemical reaction involving two species

$$\begin{aligned} \partial_t u &= D_1 \partial_x^2 u + f(u,v) \\ \partial_t v &= D_2 \partial_x^2 u + g(u,v) \end{aligned}$$

'Brusselator' (introduced by Glansdorff and Prigogine, 1971, from Brussels) does not model any specific reaction, it is just s very simple rich model

$$\begin{aligned} f(u,v) &= A - (B+1)\,u + u^2 v \\ g(u,v) &= Bu - u^2 v \end{aligned}$$

with A and B external control parameters. Keep in the following A fixed and vary B.

For all parameter values there is a simple homogeneous steady state

$$u = A \qquad v = \frac{B}{A}$$

This state may not be *stable* for all values of B: study stability by considering small perturbations

$$\begin{aligned} u &= A + U \\ v &= \frac{B}{A} + V \end{aligned}$$

Inserting in original equation

$$u^2 v = AB + 2BU + A^2 V + U^2 \frac{B}{A} + 2AUV + U^2 V$$

$$\begin{aligned} \partial_t U &= D_1 \partial_x^2 U + (B-1)U + A^2 V + F(U,V) \\ \partial_t V &= D_2 \partial_x^2 V - BU - A^2 V - F(U,V) \end{aligned}$$

with

$$F(U,V) = \frac{B}{A} U^2 + 2AUV + U^2 V$$

Linear stability: omit $F(U,V)$, which is negligible for infinitesimal U and V

$$\begin{pmatrix} \partial_t U \\ \partial_t V \end{pmatrix} = \begin{pmatrix} D_1\partial_x^2 U \\ D_2\partial_x^2 V \end{pmatrix} + \underbrace{\begin{pmatrix} B-1 & A^2 \\ -B & -A^2 \end{pmatrix}}_{\mathbf{M}_0} \begin{pmatrix} U \\ V \end{pmatrix}$$

Exponential ansatz

$$\begin{pmatrix} U \\ V \end{pmatrix} = e^{\sigma t} e^{iqx} \mathcal{A} \begin{pmatrix} U_0 \\ V_0 \end{pmatrix} \tag{50}$$

$$\mathbf{M}(\sigma,q) \begin{pmatrix} U_0 \\ V_0 \end{pmatrix} \equiv \begin{pmatrix} -\sigma - D_1q^2 + B - 1 & A^2 \\ -B & -\sigma - D_2q^2 - A^2 \end{pmatrix} \begin{pmatrix} U_0 \\ V_0 \end{pmatrix} = 0$$

has only a solution if

$$\det \mathbf{M}(\sigma,q) = 0$$

$$\sigma^2 + \sigma \underbrace{\left((D_1+D_2)q^2 + A^2 - B + 1\right)}_{\alpha(q)} + \underbrace{A^2(B-1) + q^2\left(A^2D_1 + (1-B)D_2\right) + D_1D_2q^4}_{\beta(q)} = 0$$

This gives a relation

$$\sigma = \sigma(q)$$

Instability occurs if

$$\Re(\sigma) \equiv \sigma_r > 0 \qquad \text{for some } q$$

In this model two possibilities for onset of instability

- $\sigma = i\omega$ with $q = 0$: oscillatory instability leading to ***Hopf*** bifurcation
  expect oscillations to arise with frequency ω
  occurs for $\alpha(q=0) = 0$

  $$B_c^{(H)} = 1 + A^2 \qquad \omega_c = \sigma_i$$

- $\sigma = 0$ with $q \neq 0$: instability sets in first at a specific $q = q_c$ (critical wavenumber)
  expect spatial structure to arise with wavenumber q_c
  occurs for $\beta(q_c) = 0$

  $$B_c^{(T)} = \left(1 + A\sqrt{\frac{D_1}{D_2}}\right)^2 \qquad q_c^2 = \frac{A}{\sqrt{D_1D_2}}$$

  here used $\sigma(q_c, B_c^{(T)}) = 0$ as well as $\left.\frac{d\sigma}{dq}\right|_{q_{c,B_c^{(T)}}} = 0$ to get the value where the *first* mode becomes unstable.

For small amplitude $\mathcal{A}$ one can do a weakly nonlinear analysis, expanding the equations in $\mathcal{A}$ and $B - B_c^{(H,T)}$ to obtain a Ginzburg-Landau equation for the complex amplitude $\mathcal{A}$,

$$\partial_T \mathcal{A} = \delta\, \partial_X^2 \mathcal{A} + \mu\mathcal{A} - \gamma|\mathcal{A}|^2\mathcal{A}$$

For Hopf bifurcation δ, μ, and γ are complex, for Turing bifurcation they are real.

In the original exponential ansatz (50) amplitude $\mathcal{A}$ is constant. It turns out one can allow $\mathcal{A}$ allow to vary slowly in space and time. The Ginzburg-Landau equation has simple spatially/temporally periodic solutions

$$\mathcal{A} = \mathcal{A}_0 e^{i\omega t} e^{iqx}$$

with

$$\mathcal{A}_0^2 = \frac{\mu_r - \delta_r q^2}{\gamma_r} \qquad \omega = \mu_i - \delta_i q^2 - \gamma_i |\mathcal{A}|^2$$

This leads to solutions for U and V of the form

$$\left(\begin{array}{c} U \\ V \end{array} \right) = e^{i(\omega_c+\omega)t} e^{i(q_c+q)x} \mathcal{A}_0 \left(\begin{array}{c} U_0 \\ V_0 \end{array} \right) + h.o.t.$$

In the homework the system has non-trivial boundaries: affects the onset of the instabilities. In this case one gets interesting behavior already for values of B that are slightly below B_c. Instabilities can arise at boundaries, which then can interact with the instabilities in the interior of the system.

E Background for Homework: Pulsating Combustion Fronts

Consider a one-dimensional combustible fluid in which the reactants are well mixed (premixed) and in which the concentration of a rate-limiting reactant is given by Y. The temperature of the fluid is given by T. A simple reaction with Arrhenius kinetics is then described by

$$\begin{aligned} \partial_{\hat{t}} T &= \kappa \partial_{\hat{x}}^2 t + q\, Y\, k(T) \\ \partial_{\hat{t}} Y &= D \partial_{\hat{x}}^2 Y - Y\, k(T) \end{aligned}$$

with the reaction term

$$k(T) = k_0 e^{-\frac{E}{k_B T}}$$

with E the activation energy and k_B the Boltzmann constant.

Boundary conditions

$$T(0,t) = T_l \qquad T(L,t) = T_r$$

$$Y(0,t) = Y_l \qquad Y(L,t) = Y_r$$

and initial conditions

$$T(x,0) = T_0 \qquad Y(x,0) = Y_0$$

Make dimensionless

$$C = \frac{Y}{Y_0}$$

and

$$\theta = \frac{T - T_{ad}}{T_{ad} - T_0} \qquad T_{ad} = T_0 + qY_0$$

i.e.

$$T = T_a + qY_0\theta$$

Insert into Arrhenius law

$$\begin{aligned} e^{-E/k_BT} &= e^{-E/k_BT_a}e^{E/k_B(1/T_a-1/T)} \\ &= k(T_a)exp\left(\frac{E}{k_B}\frac{1}{T_aT}(T-T_a)\right) \\ &= k(T_a)exp\left(\frac{E}{k_BT_a}\frac{qY_0\theta}{T_a+qY_0\theta}\right) \\ &= k(T_a)exp\left(\frac{Z\theta}{1+\delta\theta}\right) \end{aligned}$$

with the Zeldovich number Z given by

$$Z = \frac{E}{k_BT_a}\frac{qY_0}{T_a} \qquad \text{and} \qquad \delta = \frac{qY_0}{T_a}$$

This results in the final equations

$$\partial_t\theta = \partial_x^2\theta + Ce^{\frac{Z\theta}{1+\delta\theta}} \tag{51}$$

$$\partial_t C = \frac{1}{Le}\partial_x^2 C - Ce^{\frac{Z\theta}{1+\delta\theta}} \tag{52}$$

with Lewis number given by

$$Le = \frac{\kappa}{D}$$

Initial conditions

$$C = 1 \qquad \theta = -1$$

and boundary conditions

$$\theta = \theta_{l,r} \qquad C = C_{l,r}$$

For very large activation energy (Z large) the reaction front can be replaced by an internal layer and one can treat the outer solution analytically. A linear stability analysis shows that for $Le > 1$ and the Zeldovich number above a certain value of $Z_c(Le)$ the steadily propagating front becomes unstable to oscillations and a transition to pulsating fronts occur [8]. In two-dimensional versions of (51,52) instabilities to cellular flames arise for $Le < 1$ (cf. Fig. 7).

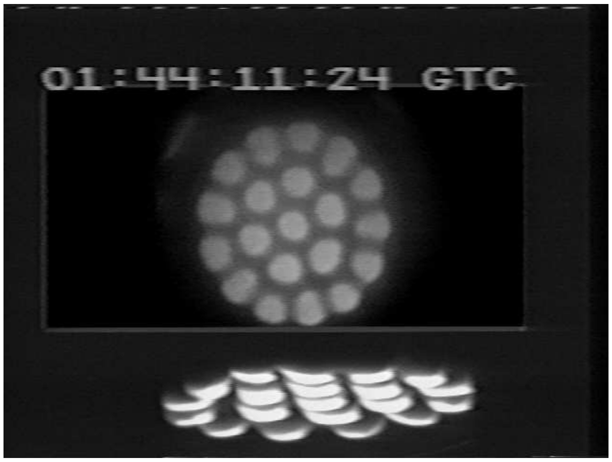


Figure 7: Cellular flame on a porous plug burner (from http://vip.cs.utsa.edu/flames/overview.html see also [9]).