Notes from Numerical Solutions to Differential Equations

Cambridge Part III Mathematical Tripos 2012-2013

Lecturer: Arieh Iserles

Vivak Patel

April 28, 2013

Contents

1 Single-Step Methods

1.1 Introduction

1. Notation

   (a) Equivalent notations for a differential equation:

   $$\text{Differential Eq.: } y' = f(t,y) \qquad \text{Initial Condition: } y(0) = y_0 \in \mathbb{R}^d$$
   $$\text{Differential Eq.: } y' = f(y) \qquad \text{Initial Condition: } y(0) = y_0 \in \mathbb{R}^{d+1}$$

   (b) Higher Order Derivatives:

   $$f_0(y) = y$$
   $$f_1(y) = f(y) = \frac{\partial y}{\partial t}$$
   $$f_2(y) = f'(y)f(y)$$

   (c) Approximate Solutions: let $h = \Delta t > 0$ be a step size. y_n is the approximate solution to a method at time $t = nh$. (i.e. $y_0 = y(0), y_1 \approx y(h), \ldots, y_n \approx y(nh)$)

2. Conditions on f

   (a) Lipschitz Condition on f guarantees a unique solution to an ODE. That is, f must satisfy for all $x, y \in \mathbb{R}^d$:

   $$\|f(t,x) - f(t,y)\| \leq \lambda \|x - y\|$$

   (b) Analytic Condition on f is much stronger and requires f to have smooth derivatives of all orders.

3. Definitions of Order, Convergence and A-stability

   Definition 1.1. *Given the above notation:*

   (a) ***Order****: Let $y_{n+1} = \mathcal{Y}_n(h, f, y_0, y_1, \ldots, y_n)$ be a method. $\mathcal{Y}_n$ is of order p is $\forall n \in \{0, 1, \ldots\}$ and for $\tilde{y}' = f(t, \tilde{y})$ with initial condition $y(\tilde{nh}) = y_n$:*

   $$\tilde{y}(nh + h) - y_{n+1} = \mathcal{O}\left(h^{p+1}\right)$$

   (i.e. Order deals with the local errors between each step.)

   (b) ***Convergence****: Suppose we approximate an ODE on $[t_0, t_0 + t^*]$ by splitting the intervals into as many equidistant intervals of size h as possible. A method $y_{n+1} = \mathcal{Y}_n(h, f, y_0, \ldots, y_n)$ is convergent if:*

   $$\lim_{h \to 0} \max_{n \in \{0, \ldots, \lfloor \frac{t^*}{h} \rfloor\}} \|y_n - y(nh)\| = 0$$

   (c) ***Linear Stability Domain****: Apply a method $\mathcal{Y}_n$ to $y' = \lambda y$ with condition $y(0) = 1$. The linear stability domain, D, of a method $\mathcal{Y}_n$ is*

   $$D = \left\{\lambda \in \mathbb{C} | \lim_{n \to 0} y_n = 0\right\}$$

   (d) ***A-stability****: A method is A-stable is $\mathbb{C}^- \subseteq D$*

1.2 Taylor Methods

1. Euler's Method (Linear Approximation)

   (a) Method

      i. $y(t_0 + h) \approx y_1 = y_0 + hf(t_0, y_0)$

      ii. General Step: $y(t_0 + nh) \approx y_n = y_{n-1} + hf(t_0 + (n-1)h, y_{n-1})$

   (b) Convergence

   Proposition 1.1. *Euler's method converges.*

   Proof. Given that f is analytic or at least Lipschitz:

      i. We expand y_n by its definition and $y(t_n)$ by Taylor's Theorem, where $t_n = nh + t_0$.

$$\begin{aligned} y_n &= y_{n-1} + hf(t_{n-1}, y_{n-1}) \\ y(t_n) &= y(t_n) + hf(t_{n-1}, y(t_{n-1})) + \mathcal{O}\left(h^2\right) \end{aligned}$$

      ii. Letting $e_n = y_n - y(t_n)$ and using Lipschitz with some $\lambda > 0$, we have that:

$$\begin{aligned} \|y_n - y(t_n)\| &= \|e_n\| \\ &\leq \|e_{n-1}\| + h \|f(t_{n-1}, y(t_{n-1} + e_{n-1})) - f(t_{n-1}, y(t_{n-1}))\| + ch^2 \\ &\leq \|e_{n-1}\| + h\lambda \|e_{n-1}\| + ch^2 \\ &\leq (1 + h\lambda) \|e_{n-1}\| + ch^2 \end{aligned}$$

      iii. We can therefore upper bound $\|e_n\|$ recursively, so as $h \to 0$:

$$\begin{aligned} \|e_n\| &\leq (1 + h\lambda) \|e_{n-1}\| + ch^2 \\ &\leq (1 + h\lambda)^2 \|e_{n-2}\| + (1 + h\lambda)ch^2 + ch^2 \\ &\leq \left[(1 + h\lambda)^{n-1} + (1 + h\lambda)^{n-2} + \cdots + 1\right] ch^2 \\ &= ch^2 \left[\frac{(1 + h\lambda)^n - 1}{\lambda h}\right] \\ &\leq \frac{ch}{\lambda} \left[(1 + h\lambda)^n - 1\right] \\ &\to 0 \end{aligned}$$

   □

2. Generalised Taylor Methods

   (a) Taylor Method

      i. Note that by Taylor's Theorem: $y(t_{n+1}) = \sum_{k=0}^{\infty} \frac{f_k(t_n, y(t_n))}{k!} h^k$

      ii. The generalised Taylor method truncates the Taylor Series and approximates $y(t_n) \approx y_n$, so that:

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

(b) Operators

i. Differential Operator: $Dg(t) = g'(t)$

ii. Shift operator: $Eg(t) = g(t+h)$

iii. If g is analytic, by Taylor's theorem:

$$\begin{aligned} Eg(t) &= \sum_{k=0}^{\infty} \frac{h^k}{k!} g^{(k)}(t) \\ &= \sum_{k=0}^{\infty} \frac{h^k}{k!} D^k g(t) \\ &= \exp(hD) g(t) \end{aligned}$$

(c) Generalised Taylor Methods

i. Method: Let $R(z) = \sum_{k=0}^{\infty} r_k z^k$ be an operator such that $R(z) - \exp(z) = \mathcal{O}\left(z^{p+1}\right)$ .. Then the Generalised Taylor Method is:

$$y_{n+1} = R(hD) f(t_n, y_n)$$

ii. Order of Generalised Taylor Methods

Proposition 1.2. *The generalised Taylor method has order p.*

Proof. Let $R(z)$ be defined as above as an operator with order p.

A. First we expand y_{n+1} and $\tilde{y}(t_{n+1})$ using Taylor's theorem for the latter.

$$\begin{aligned} y_{n+1} = R(hD) f(t_n, y_n) &= (\exp hD + \mathcal{O}\left((hD)^{p+1}\right)) f(t_n, y_n) \\ &= \sum_{k=0}^{\infty} \frac{f_k(t_n, y_n)}{k!} h^k + \mathcal{O}\left(h^{p+1}\right) \\ \tilde{y}(t_{n+1}) &= \sum_{k=0}^{\infty} \frac{f_k(t_n, \tilde{y}(t_n))}{k!} h^k \\ &= \sum_{k=0}^{\infty} \frac{f_k(t_n, y_n)}{k!} h^k \end{aligned}$$

B. The last line followed by the definition of $\tilde{y}$ and its initial condition. Taking the difference in some norm, we will have:

$$\|y_{n+1} - \tilde{y}(t_{n+1})\| = \mathcal{O}\left(h^{p+1}\right)$$

□

iii. Linear Stability domain of the Taylor Method

Proposition 1.3. *The linear stability domain of the Taylor Method is:*

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

Proof. Starting with the test function $y' = \lambda y$, we see that $f_k(y) = \lambda^k y$. So:

$$y_n = \left[\frac{(h\lambda)^k}{k!}\right]^n$$

In order for $\lim_{n\to\infty} y_n = 0$, we need: $\left[\frac{(h\lambda)^k}{k!}\right] < 1$. Therefore, we have the above stability domain for which $z = h\lambda$. □

iv. Since D is bounded, $\mathbb{C}^- \not\subseteq D$. So the Taylor method is not A-stable.

1.3 Rational Methods

1. Trapezoidal Method:

   (a) Method: $y_{n+1} = y_n + \frac{1}{2}h\left[f(y_n) + f(y_{n+1})\right]$

   (b) Convergence

   Proposition 1.4. *The trapezoidal method is convergent.*

   Proof. Again we will use Lipschitz continuity on functions f and y for a λ_1 and λ_2 respectively.

   i. We first expand each term as always:

$$y_{n+1} = y_n + \frac{1}{2}h\left[f(y_n) + f(y_{n+1})\right]$$
$$y(t_{n+1}) = y(t_n) + hf(t_n, y(t_n)) + \sum_{k=2}^{\infty} \frac{f_k(t_n, y(t_n)}{k!}h^k$$

   ii. Defining the error e_n as we did before, and taking the difference of the expansions, (and changing notation because we are lazy), we have:

$$\begin{aligned} e_{n+1} =& e_n + \frac{1}{2}h\left[f(y_n) - f(y(t_n))\right] + \frac{1}{2}h\left[f(y_{n+1}) - f(y(t_n))\right] \\ & - \sum_{k=2}^{\infty} \frac{f_k(t_n, y(t_n)}{k!}h^k \\ =& e_n + \frac{1}{2}h\left[f(y_n) - f(y(t_n))\right] + \frac{1}{2}h\left[f(y_{n+1}) - f(y(t_{n+1}))\right] \\ & + \frac{1}{2}h\left[f(y(t_{n+1})) - f(y(t_n))\right] - \sum_{k=2}^{\infty} \frac{f_k(t_n, y(t_n)}{k!}h^k \end{aligned}$$

   iii. Taking the modulus, using triangle inequality and Lipschitz, then for some λ:

$$\begin{aligned} \|e_{n+1}\| &\leq \|e_n\| + \frac{1}{2}\lambda_1\|e_n\| + \frac{1}{2}h\lambda_1\|e_{n+1}\| + \frac{1}{2}h^2\lambda_1\lambda_2 + \mathcal{O}\left(h^2\right) \\ &\leq \|e_n\|\left[\frac{1 + \frac{1}{2}h\lambda}{1 - \frac{1}{2}h\lambda}\right] + h^2\left[\frac{\frac{1}{2}\lambda^2 + c}{1 - \frac{1}{2}h\lambda}\right] \end{aligned}$$

iv. Computing $\|e_1\|, \|e_2\|, \ldots$ and extrapolating a general pattern, we have that:

$$\|e_n\| \leq h\left[\frac{1}{2}\lambda + \frac{c}{\lambda}\right]\left[\left(\frac{1+\frac{1}{2}h\lambda}{1-\frac{1}{2}h\lambda}\right)^n - 1\right]$$

v. As $h \to 0$, we see that the method is convergent.

□

2. General Rational Methods

(a) Method

i. Consider the rational operator $R(z) = \frac{\sum_{k=0}^{M} p_k z^k}{\sum_{k=0}^{N} q_k z^k}$

ii. The corresponding method is

$$\sum_{k=0}^{N} q_k h^k f_k(y_{n+1}) = \sum_{k=0}^{M} p_k h^k f_k(y_n)$$

(b) Stability

i. Linear Stability

Proposition 1.5. *The linear stability domain is*

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

Proof. Using the test function $y' = \lambda y$, we note that $f_k = \lambda^k y$. We can compute

$$y_n = (R(h\lambda))^n$$

Therefore, as $n \to \infty$, $y_n \to 0$ if $|R(\lambda h)| < 1$ □

ii. A-stability (proven later)

Lemma 1.1. *the rational method, $R(z)$ is A-stable if and only if all poles of $R(z)$ reside in $\mathbb{C}^+$, and $|R(iy)| \leq 1$ for all $y \in \mathbb{R}$*

3. Maximal Order Rational Methods: Pade Approximations

(a) Method: The $[M/N]$ Pade Approximation is $R_{[M/N]} = \frac{P_{[M/N]}}{Q_{[M/N]}}$

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

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

(b) Order

Lemma 1.2. $R_{[M/N]}(z) = \exp(z) + \mathcal{O}\left(z^{M+N+1}\right)$ *and no $[M/N]$ function can do better.*

Corollary 1.1. *The $[M/N]$ Pade approximation is of order $M+N$*

(c) Stability

Theorem 1.1. *(Wanner, Hairer, Norsett) The $[M/N]$ Pade method is A-stable if and only if $M \leq N \leq M+2$*

2 Multi-Step Methods

2.1 Adam's Method

1. General Idea: given the past values of a numerical solution, we can use them to approximate $f(t, y)$ as an integrable function $p(t)$, and hence approximate

$$y(t_{n+s}) \approx y_{n+s} = y_{n+s-1} + \int_{t_{n+s-1}}^{t_{n+s}} p(t)dt$$

2. Determining $p(t)$

   (a) We assume that y_{n+m} where $m = 0, 1, \ldots, s-1$ are available and $y_{n+m} - y(t_{n+m}) = \mathcal{O}\left(h^{s+1}\right)$

   (b) We use interpolation theory to show that given the Lagrange interpolation polynomials $p_m(t)$:

$$p(t) = \sum_{m=0}^{s-1} p_m(t) f(t_{n+m}, y_{n+m})$$

   (c) From our assumptions and approximation theory, we have that for $t \in [t_{n+s-1}, t_{n+s}]$:

$$p(t) - f(t, y) = \mathcal{O}\left(h^s\right)$$

   (d) Therefore:

$$y_{n+s} = y_{n+s-1} + h \sum_{m=0}^{s-1} b_m f(t_{n+m}, y_{n+m})$$

   where

$$b_m = \frac{1}{h} \int_{t_{n+s-1}}^{t_{n+s}} p_m(t)dt = \frac{1}{h} \int_0^h p_m(t_{n+s-1} + \tau)d\tau$$

   (e) The order is $p = s$.

2.2 Order and Convergence of Multi-step Methods

1. General Multi-step Methods

   (a) Method: For $n = 0, 1, \ldots$:

$$\sum_{m=0}^{s} a_m y_{n+m} = h \sum_{m=0}^{s} b_m f(t_{n+m}, y_{n+m})$$

      i. a_m, b_m for $m = 0, 1, \ldots, s$ are given constants independent of h, n and the differential equation

      ii. $a_s = 1$ is typical to normalise the method

   (b) If $b_s = 0$ the method is explicit, else it is implicit

(c) a_m and b_m are selected using a variety of criterion, so we must rely on order and convergence to determine the value of a particular method

(d) Method in terms of Operators

i. Recall: $\log(E) = hD$. Let $\rho(\omega) = \sum_{m=0}^{s} a_m \omega^m$ and $\sigma(\omega) = \sum_{m=0}^{s} b_m \omega^m$

ii. Therefore, we have that:

$$\rho(E)y(t_n) \approx \sum_{m=0}^{s} a_m y_{n+m}$$

$$\log(E)\sigma(E)y(t_n) \approx h \sum_{m=0}^{s} b_m f(t_{n+m}, y_{n+m})$$

2. Order

(a) Redefinition of Order

Definition 2.1. *A multi-step method is of order p if and only if for all sufficiently smooth functions y (one must exist and none can do better):*

$$\psi(t, y) = \sum_{m=0}^{s} a_m y(t + mh) - h \sum_{m=0}^{s} b_m y'(t + mh) = \mathcal{O}\left(h^{p+1}\right)$$

(b) In terms of operators, the multi-step method is of order p if and only if:

$$[\rho(E) - \log(E)\sigma(E)]\, y(t_n) = \mathcal{O}\left((h^{p+1})\right)$$

(c) Necessity and Sufficiency for a Multi-step Method to have order p

Proposition 2.1. *A multi-step method is of order p:*

i. If and only if for $c \neq 0$ as $\omega \to 1$:

$$\rho(\omega) - \log(\omega)\sigma(\omega) = c(\omega - 1)^{p+1} + \mathcal{O}\left(|\omega - 1|^{p+1}\right) + \mathcal{O}\left(|\omega - 1|^{p+2}\right)$$

ii. then it is necessary and sufficient that:

$$\sum_{m=0}^{s} a_m = 0$$

$$\sum_{m=0}^{s} m^k a_m = k \sum_{m=0}^{s} m^{k-1} b_m \text{ for } k = 1, \ldots, p$$

$$\sum_{m=0}^{s} m^{p+1} a_m \neq p + 1 \sum_{m=0}^{s} m^p b_m$$

Proof. The strategy of this proof is to prove necessity and sufficiency in both directions of part (b) yielding part (a)

i. Suppose the method is of order p. Then we expand in Taylor series and use the definition of order to get (b):

$$\begin{aligned}\psi(t,y) &= \sum_{m=0}^{s} a_m \sum_{k=0}^{\infty} \frac{f_k(t,y)}{k!}(mh)^k - h\sum_{m=0}^{s} b_m \sum_{k=1}^{\infty} \frac{f_k(t,y)}{k!}(mh)^{k-1} \\ &= \left(\sum_{m=0}^{s} a_m\right) y(t) + \sum_{m=0}^{s}\left(\sum_{k=1}^{\infty} \frac{f_k h^k}{k!}\left(a_m m^k - k b_m m^{k-1}\right)\right) \\ &= \left(\sum_{m=0}^{s} a_m\right) y(t) + \sum_{k=1}^{\infty} \frac{f_k h^k}{k!} \sum_{m=0}^{s}\left(a_m m^k - k b_m m^{k-1}\right)\end{aligned}$$

For this to have order p, only the terms with $k \geq p+1$ can be non-zero, the rest must be 0 yielding result (b).

ii. Suppose that $\rho(\omega) - \log(\omega)\sigma(\omega) = c(\omega-1)^{p+1} + \mathcal{O}\left(|\omega-1|^{p+2}\right)$ The trick is to make the substitution $\omega = \exp(z)$ and note that as $\omega \to 1$, $\exp(z) \to 1+z$ and $z \to 0$. Therefore, the right hand side of the supposition will become $cz^{p+1} + \mathcal{O}\left(z^{p+2}\right)$. Playing now with the left hand side:

$$\begin{aligned}\rho(e^z) - z\sigma(e^z) &= \sum_{m=0}^{s}\left(a_m \sum_{k=0}^{\infty} \frac{(mz)^k}{k!} - b_m \sum_{k=0}^{\infty} \frac{m^k z^{k+1}}{k!}\right) \\ &= \sum_{m=0}^{s} a_m + \sum_{k=1}^{\infty} \frac{z^k}{k!}\left(\sum_{m=0}^{s} a_m m^k - b_m k m^{k-1}\right)\end{aligned}$$

Setting this equal to $cz^{p+1} + \mathcal{O}\left(z^{p+2}\right)$, we get the conditions of part (b).

□

3. Convergence and Order

Definition 2.2. *A polynomial obeys the* ***Root Condition*** *if all of its zeros are in the unit disc and all zeros on the unit circle are of multiplicity* 1.

Theorem 2.1. *Dahlquist Equivalence Theorem: Suppose the error in* $y_1, \ldots, y_{s-1}$ *tends to* 0 *as* $h \to 0$. *The multi-step method is convergent if and only if it has order* $p \geq 1$ *and the polynomial* ρ *obeys the root condition.*

Theorem 2.2. *Dahlquist First Barrier: The maximal order of an s-step method is at most* $2\lfloor \frac{s+2}{2} \rfloor$ *for implicit schemes, and s for explicit schemes.*

4. Attaining higher-order, convergent multi-step methods

(a) We first select $\rho(\omega)$ to coincide with the Dahlquist Equivalence Theorem and so that it has order p:

i. For convergence, $\rho(\omega)$ must obey the root condition

ii. For order p, $\rho(1) = \sum_{m=0}^{s} a_m = 0$

(b) Using the fact that a method has order p if and only if as $\omega \to 1$:

$$\rho(\omega) - \log(\omega)\sigma(\omega) - \mathcal{O}\left(|\omega - 1|^{p+1}\right)$$

we can solve for σ by expanding the $\log(\omega)$ term and letting $\omega \to 1$:

$$\sigma(\omega) = \frac{\rho(\omega)}{\log(\omega)} + \mathcal{O}\left(|\omega - 1|^{p}\right)$$

(c) Expand the fraction, (note substituting $\xi + 1 = \omega$ helps) and select $\sigma(\omega)$ to be the polynomial that matches the expansion up to order p.

(d) Select $p = s + 1$ for implicit methods and $p = s$ for explicit methods.

5. Backwards Differentiation Formula

(a) Method: An s-step BDF has $\sigma(\omega) = \beta\omega^s$ and it has order s

(b) Computing β and ρ:

Lemma 2.1. *Suppose we have an s-step BDF:*

$$\beta = \left(\sum_{m=1}^{s} \frac{1}{m}\right)^{-1}$$

$$\rho(\omega) = \sum_{m=1}^{s} \frac{1}{m}\omega^{s-m}(\omega - 1)^m$$

(c) Order and Convergence:

Proposition 2.2. *For BDF methods: $\rho(\omega)$ obeys the root condition (and is hence convergent) if and only if $1 \leq s \leq 6$*

2.3 Implementation of Multi-step Methods

3 Runge-Kutta Methods

3.1 Gaussian Quadrature

1. Quadrature:

$$\int_a^b f(\tau)\omega(\tau)d\tau \approx_{j=1}^{v} b_j f(c_j)$$

(a) $\omega(\tau) > 0$ on (a, b), and

i. $\int_a^b \omega(\tau) < \infty$

ii. For $j = 1, 2, \ldots$, we have $\left|\int_a^b \tau^j \omega(\tau)d\tau\right| < \infty$

iii. ω is called the weight function

(b) $b_1, \ldots, b_v$ and $c_1, \ldots, c_v$ are independent of f but depend on a, b and ω

i. b_j are quadrature weights

ii. c_j are quadrature nodes

(c) If f is continuously differentiable p times, then

$$\left| \int_a^b f(\tau)\omega(\tau)d\tau - \sum_{j=1}^{v} b_j f(c_j) \right| < c \sup_{t\in[a,b]} \left| f^{(p)}(t) \right|$$

2. Order of a Quadrature

Lemma 3.1. *Given a distinct set of nodes $c_1, \ldots, c_v$, it is possible to find unique $b_1, \ldots, b_v$ such that the quadrature is of order $p \geq v$*

3. Orthogonal Polynomials

(a) Weight functions define inner products:

$$\langle f, g \rangle = \int_a^b f(\tau)g(\tau)\omega(\tau)d\tau$$

(b) Definition of Orthogonal Polynomials

Definition 3.1. *$p_m \in \mathbb{P}_m \setminus \{0\}$ is an m^{th} orthogonal polynomial if $\langle p_m, \hat{p} \rangle = 0$ for all $\hat{p} \in \mathbb{P}_{m-1}$*

(c) Zeros of m^{th} orthogonal polynomial

Lemma 3.2. *All m zeros of p_m reside in (a, b) and they are simple.*

(d) Vandermonde System

Definition 3.2. *The v^{th} Vandermonde system has $b_1, \ldots, b_v$ unknowns. Given $c_1, \ldots, c_j$, the system is formed by using the basis for $\mathbb{P}_{v-1}$:*

$$\sum_{j=1}^{v} b_j c_j^m = \int_a^b \tau^m d\tau \qquad m = 0, \ldots, v-1$$

(e) Maximal Order Quadrature: Gaussian Quadrature

Theorem 3.1. *Let $c_1, \ldots, c_v$ be the zeros of the v^{th} monic (first coefficient is 0) orthogonal polynomial, p_v, and $b_1, \ldots, b_v$ be the solutions to the Vandermonde System. This is called Gaussian Quadrature. Moreover:*

i. The quadrature method used for $\hat{p} \in \mathbb{P}_{2v-1}$ has order $2v$

ii. No quadrature can exceed this order

3.2 Explicit Runge-Kutta Schemes

1. ERK Method

(a) Our goal is to approximate the transformed integral using quadrature:

$$\begin{aligned} y(t_{n+1}) &= y(t_n) + \int_{t_n}^{t_{n+1}} f(\tau, y(\tau))d\tau \\ &= y(t_n) + h\int_0^1 f(t_n + h\tau, y(t_n + h\tau))d\tau \end{aligned}$$

(b) The method is almost:

$$y_{n+1} = y_n + h\sum_{j=1}^{v} b_j f(t_n + c_j h, y(t_n + c_j h)) \qquad n = 0, 1, \ldots$$

(c) However, we do not know $y(t_n + c_j h)$ for $j = 1, \ldots, v$ so we approximate them explicitly for some choice of $a_{j,i}$ and require $c_1 = 0$:

$$\begin{aligned}
\xi_1 &= y_n \\
\xi_2 &= y_n + h a_{2,1} f(t_n, \xi_1) \\
\xi_3 &= y_n + h a_{3,2} f(t_n + c_2 h, \xi_2) + h a_{3,1} f(t_n, \xi_1) \\
&\vdots \\
\xi_j &= y_n + h\sum_{i=1}^{j-1} a_{j,i} f(t_n + c_i h, \xi_i)
\end{aligned}$$

(d) Therefore, the v-stage method is:

$$y_{n+1} = y_n + h\sum_{j=1}^{v} b_j f(t_n + c_j h, \xi_j)$$

2. Important relationship between $a_{j,i}$ and c_j: Note that we are constrained by the fact that if we want to recover the solution to the scalar autonomous equation $y = f(y) = y'$:

$$\begin{aligned}
y(t_n + c_j h) &\approx y_n + f(t_n) c_j h \approx y_n + y_n c_j h \\
f(\xi_1) &= f(y_n) = y_n \\
f(\xi_2) &= f(y_n + h a_{2,1} f(y_n)) = y_n + h a_{2,1} y_n \\
f(\xi_j) &= f(y_n + h\sum_{i=1}^{j-1} a_{j,i} f(\xi_i)) = y_n + y_n h \left(\sum_{i=1}^{j-1} a_{j,i}\right)
\end{aligned}$$

Therefore:

$$\sum_{i=1}^{j-1} a_{j,i} = c_j$$

3. Selecting the RK Matrix (i.e. $a_{j,i}$)

Example 3.1. *Suppose we have a scalar, autonomous ODE $y = f(y) = y'$ which we want to solve using an ERK with 3-stages of order 3. Let $f = f(y_n)$, and $f_y = f_y(y_n)$, etc.. We compute $k_j = f(\xi_j)$ instead of just ξ_j, since they are equal. By Taylor' theorem:*

$$\begin{aligned}
k_1 &= f(\xi_1) = f(y_n) = f \\
k_2 &= f(y_n + h a_{2,1} f) = f + f_y(y_n) h c_2 f + \frac{1}{2} h^2 c_2^2 f_{yy} f^2 + \mathcal{O}\left(h^3\right) \\
k_3 &= f + h c_3 f_y f + h^2 (c_2 a_{3,2} f_y^2 f + \frac{1}{2} c_3^2 f_y y f^2) + \mathcal{O}\left((h^3)\right)
\end{aligned}$$

The method is then:

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

Comparing this to the expansion about $y(t_n + h)$ (using chain rule $\partial_t f = f_y f$):

$$y(t_n + h) \approx y_n + hf + \frac{1}{2}h^2 f_y f + h^3 \left(\frac{1}{6} f_{yy} f^2 + \frac{1}{6} f_y^2 f \right) + \mathcal{O}\left(h^4\right)$$

We have the following system:

$$\begin{aligned} b_1 + b_2 + b_3 &= 1 & b_2 c_2 + b_3 c_3 &= \frac{1}{2} \\ b_2 c_2^2 + b_3 c_3^2 &= \frac{1}{3} & b_3 c_2 a_{3,2} &= \frac{1}{6} \end{aligned}$$

This system does not have a unique solution.

(a) This example illustrates that the RK matrix can be selected using (tedious) Taylor expansion, but other ways such as graph theory (Elementary Differentials)

(b) The matrix for the order and stages is not unique

3.3 Implicit Runge-Kutta Schemes

1. Method:

   (a) For $j = 1, \ldots, v$, the ξ_j are implicitly defined:

   $$\xi_j = y_n + h \sum_{i=1}^{v} a_{j,i} f(t_n + c_i h, \xi_i)$$

   (b) The method is then:

   $$y_{n+1} = y_n + h \sum_{j=1}^{v} b_j f(t_n + c_j h, \xi_j)$$

2. Order

   (a) We can evaluate the order for an IRK with $p \leq 3$ using the one-dimensional autonomous equation $y' = f(y)$

   (b) We can determine the exact order using $y' = y$ (scalar)

3.4 Collocation and IRK Methods

1. Collocation Methods

   (a) Let $c_1, \ldots, c_v$ be v distinct collocation parameters (usually in $[0, 1]$)

(b) We want a v^{th} degree polynomial u with vector coefficients such that:

i. $u(t_n) = y_n$

ii. $u'(t_n + c_j h) = f(t_n + c_j h, u(t_n + c_j h))$ for $j = 1, \ldots, v$

(c) We approximate $y(t_{n+1}) \approx y_{n+1} = u(t_{n+1})$

2. Collocation is a case of IRK

Lemma 3.3. *Set* $q(t) = \prod_{j=1}^{v}(t - c_j)$ *and* $q_l(t) = \frac{q(t)}{t-c_l}$ *for* $l = 1, \ldots, v$ *and let:*

(a) $a_{j,i} = \int_0^{c_j} \frac{q_i(\tau)}{q_i(c_j)} d\tau$

(b) $b_j = \int_0^1 \frac{q_j(\tau)}{q_j(c_j)} d\tau$

The collocation is the same as the IRK methods (c, A, b^T).

3. Defects and Convergence

(a) Given an ODE we can find a smooth, differentiable solution (candidate) v through approximation and pertubation

(b) We define the defect to be $d(t) = v'(t) - f(t, v(t))$ and we expect small values in $\|d\|$ to indicate small error

(c) Alekseev-Grobner Theorem

Theorem 3.2. *Let* v *be a smooth differentiable function that obeys* $v(t_0) = y_0$. *Then for* $t \geq t_0$*:*

$$v(t) - y(t) = \int_{t_0}^{t} \Phi(t - \tau, v(t - \tau)) d\tau$$

where $\Phi(t, v)$ *is the matrix of partial derivatives of the solution* $\omega' = f(t, \omega)$, $\omega = v$

4. Order and Gauss-Legrende

(a) Order of a collocation method

Theorem 3.3. *Suppose that for* $q(t) = \prod_{j=1}^{v}(t - c_j)$, $\int_0^1 q(t) t^j dt = 0$ *for* $j = 0, \ldots, m - 1 \leq v - 1$. *Then the collocation method corresponding to this choice of* $c_1, \ldots, c_v$ *is of order* $v + m$

(b) Maximal Order: Gauss-Legrende Collocation/IRK method

Corollary 3.1. *Let* $c_1, \ldots, c_v$ *be the zeros of polynomials* $\tilde{p}_v \in \mathbb{P}_v$ *orthogonal with respect to* $w(t) = 1$ *and* $0 \leq t \leq 1$. *The underlying* v*-stage IRK/collocation method (Gauss-Legrende) is of order* $2v$.

4 Stiff Equations

4.1 Linear Stability and A-stability

Stiff equations will typically result in instability unless step-size is very small. So stable methods are ideal because they can easily handle this requirement.

1. Linear Stability Domain

   (a) Definition

   Definition 4.1. *The LSD is*

   $$D = \{z \in \mathbb{C} : h\lambda = z, y' = \lambda y, \lim_{n\to\infty} y_n = 0\}$$

   (b) The LSD is the set of all $h\lambda$ that recover the correct asymptotic behaviour of $y' = \lambda y$, $y(0) = 1$ when the ODE is stable (i.e. $Re(\lambda) < 0$).

2. A-stability

   (a) Definition

   Definition 4.2. *A method is A-stable if* $\mathbb{C}^- \subseteq D$

   (b) If a method is A-stable, h can be selected for accuracy/error control without regard to asymptotic behaviour

4.2 A-stability of RK and Rational Methods

1. Notation

   (a) We apply the RK method to $y' = \lambda y$ with $y(0) = 1$, and find:

   $$\xi_j = y_n + h\sum_{i=1}^{v} a_{j,i} f(\xi_i) = y_n + h\sum_{i=1}^{v} a_{j,i}\lambda\xi_i$$

   or in vector notation:

   $$\xi = \mathbf{1}y_n + h\lambda A\xi$$

   yielding the method:

   $$y_{n+1} = y_n + h\lambda b^T\xi$$

   (b) If $I - \lambda hA$ is nonsingular then:

   $$y_{n+1} = \left[1 + h\lambda b^T (I - h\lambda A)^{-1}\mathbf{1}\right] y_n$$

2. Linear Stability Domain

   (a) Note that the RK methods behave as a rational function in terms of stability, and so the following results apply to rational and RK methods

   (b) Equivalence of rational methods and RK methods:

   Lemma 4.1. *Every RK method has a corresponding rational function* $f \in \mathbb{P}_{v/v}$ *(i.e. the numerator and denominator have degree* v*) such that* $y_n = [r(h\lambda)]^n$*. Moreover, if it is an ERK method,* $r \in \mathbb{P}_v$

   (c) Stability of ERK methods (by Maximum Modulus Principle)

   Lemma 4.2. *ERK methods are not A-stable*

(d) Criteria for A-stability or Rational/IRK method

Lemma 4.3. *Let $r(z) \in \mathbb{P}_{v/v} \setminus \{constants\}$: $|r(z)| < 1,\ \forall z \in \mathbb{C}^-$ if and only if all poles of r have real part positive and $|r(it)| \leq 1$ for all $t \in \mathbb{R}$*

3. Order and Optimality for RK and Rational Methods

(a) $r(z)$ operator and $\exp(z)$ operator

Lemma 4.4. *Suppose $y_n = [r(\lambda h)]^n$ for some method for the linear equation $y' = \lambda y$ with $y(0) = 1$. Suppose the method is of order p. Then $r(z) = \exp(z) + \mathcal{O}\left(z^{p+1}\right)$.*

(b) Pade Approximations

i. Pade approximations are rational functions that match e^z to the highest order

ii. Existence, Uniqueness, and form of Pade Approximation

Theorem 4.1. *Given any $\alpha, \beta > 0$, $\exists! r_{\alpha/\beta} \in \mathbb{P}_{\alpha/\beta}$ such that $r_{\alpha/\beta}$ is of order $\alpha + \beta$ (and no function can do better), and it can be written as:*

$$r_{\alpha/\beta} = \frac{p_{\alpha/\beta}}{q_{\alpha/\beta}}$$

where

$$p_{\alpha/\beta}(z) = \sum_{k=0}^{\alpha} \binom{\alpha}{k} \frac{(\alpha+\beta-k)!}{(\alpha+\beta)!} z^k$$
$$q_{\alpha/\beta}(z) = p_{\beta/\alpha}(-z)$$

Wanner-Hairer-Norsett Theorem on A-stability of Pade Approximations

Theorem 4.2. *Pade approximations correspond to A-stable methods if and only if $\alpha \leq \beta \leq \alpha + 2$*

Corollary 4.1. *The Gauss-Legrende IRK methods are A-stable for all $v \geq 1$.*

4.3 A-stability of Multi-step Methods

1. Difference Equations: suppose we have $\sum_{m=0}^{s} g_m x_{m+n} = 0$. Suppose $w_1, \ldots, w_q$ are the solutions with multiplicity $k_1, \ldots, k_q$. Then the solution for $n = 0, 1, \ldots$ is:

$$x_n = \sum_{i=1}^{q} \left(\sum_{j=0}^{k_i - 1} c_{i,j} n^j \right) w_i^n$$

Derivation. We have that $g_s x_{s+n} + \cdots g_0 x_n = 0$.

(a) We guess the solution $x_n = w^n c$, so we have:

$$(g_s w^s + \cdots + g_0) w^n = 0$$

(b) From our supposition, $w_1, \ldots, w_q$ solve the polynomial equation with their respective multiplicities. As we do with repeated eigenvalues for differential equations: $n^j w_i^n$ for $j = 1, \ldots, k_i - 1$ are solutions along with w_i.

(c) So the general solution is the sum over all of these with $c_{i,j}$ as arbitrary constants.

□

2. A-Stability of Multi-step Methods

Lemma 4.5. *The multi-step method is A-stable if and only if $|w_i(z)| < 1$ for $i = 1, \ldots, q(z)$ for all $z \in \mathbb{C}^{-1}$, where $w_i(z)$ are the zeros of difference equations corresponding to the multi-step methods.*

Proof. First we derive the difference equation and its solution before proving the statement. We consider $y' = \lambda y$:

(a) The difference equation:

$$\begin{aligned} \sum_{m=0}^{s} a_m y_{m+n} &= h \sum_{m=0}^{s} b_m f(y_{m+n}) \\ &= h\lambda \sum_{m=0} b_m y_{m+n} \\ 0 = \sum_{m=0}^{s} (a_m - (h\lambda) b_m) y_{m+n} &= \sum_{m=0}^{s} (a_m - z b_m) y_{n+m} \end{aligned}$$

(b) Given that the zeros occur at $w_i(z)$ with multiplicity $k_i(z)$ for $i = 1, \ldots, q(z)$, the general solutions i:

$$y_n = \sum_{i=1}^{q(z)} \left(\sum_{j=1}^{k_i(z)-1} c_{i,j} n^j \right) w_i(z)^n$$

(c) For this solution to converge as $n \to \infty$ to 0, for all i, $|w_i(z)| < 1$.

□

3. Equivalent, but easier to verify formulation for A-stability

Lemma 4.6. *The multi-step method is A-stable if and only if $b_s > 0$ and $|w_j(it)| \leq 1$ for $j = 1, \ldots, q(z)$ and $t \in \mathbb{R}$*

4. Confirming the second part requires the Cohn-Schur Criterion:

Lemma 4.7. *Zeros of a complex-coefficient quadratic polynomial $\alpha\omega^2 + \beta\omega + \gamma$ are in the closed complex disc if and only if:*

$$\begin{aligned} &|\alpha| \geq |\gamma| \\ &||\alpha|^2 - |\gamma|^2| \geq |\alpha\bar{\beta} - \beta\bar{\gamma}| \\ &\alpha = \gamma \neq 0 \implies |\beta| \leq 2|\alpha| \end{aligned}$$

5. Implications

   (a) The two-step BDF is A-stable.

   (b) Dahlquist Second Barrier for Multi-step Methods

   Theorem 4.3. *The highest order of an A-stable multi-step methods is* 2.

6. $A(\alpha)$-stability

   (a) Let $v_\alpha = \{\rho e^{i\theta} : \rho > 0, |\theta + \pi| < \alpha \in (0, \pi)\} \subset \mathbb{C}^-$

   (b) A method is $A(\alpha)$-stable if for some specified α measured in degrees, $v_\alpha \subset D$

   (c) All s-step BDF methods for $s \leq 6$ are $A(\alpha)$-stable

5 Strategies for Error Control

5.1 Multi-step Methods

1. Milne Device

   (a) Consider the error in the multi-step method, which we get from the Order theorem:

$$y(t_{n+s}) - y_{n+s} = c_P h^{p+1} y^{(p+1)}(t_{n+s}) + \mathcal{O}\left(h^{p+2}\right)$$

   (b) We do not know $y^{(p+1)}$ so instead we use another method with comparable order to approximate this value:

$$y(t_{n+s}) - x_{n+s} = c_E h^{p+1} y^{(p+1)}(t_{n+s}) + \mathcal{O}\left(h^{p+2}\right)$$

   (c) Combining the two terms we can compute the approximate global error:

$$y(t_{n+s}) - y_{n+s} \approx \frac{c_P}{c_P - c_E}\left(x_{n+s} - y_{n+s}\right)$$

   (d) For every computational step we check to see if:

$$\kappa = \left|\frac{c_P}{c_P - c_E}\right| \|x_{n+s} - y_{n+s}\| < h\delta$$

   Where δ is some specified tolerance, multiplied by h to improve global error control

      i. If $\kappa > h\delta$ we halve h and try again
      ii. If $\kappa < \frac{1}{10} h\delta$ (for example), we can double h

2. Zandunaisky Device

   (a) Given a p order solution sequence $(y_j)_{j=0}^n$, we choose a polynomial $\mathbf{q}$ of degree p that interpolate y over the previous $p+1$ grid points.

   (b) Let $d(t) = q'(t) - f(q(t))$. let the auxiliary system be $z' = f(z) = d(t)$ with the past values $(y_j)_{j=0}^n$.

(c) Then:

i. Since $q(t) = y(t) + \mathcal{O}\left(h^{p+1}\right)$, substituting in for $y' - f(y) = 0$:

$$q'(t) - \mathcal{O}\left(h^p\right) - f(q(t)) = 0 \implies d(t) = \mathcal{O}\left(h^p\right)$$

ii. We can use the auxiliary equation to compute z_{n+1} and subtract $q(t_{n+})$ to estimate $y_{n+1} - y(t_{n+1})$

3. Gear's Automatic Integration

(a) Suppose we have a family of m step methods for $m = 1, \ldots, M$ each of order $p_m = m + K$ with error constant c_m

(b) Process

i. Commence iteration with $m = 1$

ii. At the n^{th} step, working working with the m-step method, evaluate the error estimates:

$$E_j \approx c_j h^{j+K+1} y^{(j+K+1)}(t_n)$$

$$j \in I_m = \{m-1, m, m+1\} \cap \{1, \ldots, M\}$$

iii. Check that $\|y_{n+1} - y(t_{n+1})\|$ is below tolerance using error E_m

iv. Using remaining E_j find the method in I_m that will produce a result within tolerance but has the largest step size

v. Change to the new method and step size, use interpolation to re-grid computed values

(c) Disadvantages

i. Must retain many past values

ii. We cannot increase step size too soon after a a previous increase

5.2 Embedded Runge-Kutta

1. Naive Approach

(a) use two RK methods (one of order p and another of order $> p$) to estimate the error:

$$y_{n+1} = \tilde{y}(t_{n+1}) + lh^{p+1} + \mathcal{O}\left(h^{p+1}\right)$$
$$x_{n+1} = \tilde{y}(t_{n+1}) + \mathcal{O}\left(h^{p+2}\right)$$

(b) Then $lh^{p+1} \approx y_{n+1} - x_{n+1}$ and set $\kappa = \|y_{n+1} - x_{n+1}\|$

(c) This requires at lest twice as many calculations!

2. Embedded RK: We embed the p-order v stage method into a $p < \tilde{p}$ order $v < \tilde{v}$-step system, thus decreasing the cost of computation. So if c and A are the old RK matrix and nodes:

$$\tilde{A} = \begin{pmatrix} A & 0 \\ \tilde{a}' & \tilde{a} \end{pmatrix}$$
$$\tilde{\mathbf{c}} = \begin{pmatrix} c \\ \tilde{c} \end{pmatrix}$$

6 Finite Difference Schemes

6.1 Finite Differences and Operators

1. Motivation: replace derivatives with linear combinations of discrete functions

2. Operators (living in $\{z_k\}_{k=-\infty}^{\infty}$

   (a) Basic Operators: mapping from $z_k \to z_k$ except for the last three:

      i. Shift operator: $(\mathcal{E}z)_k = z_{k+1}$
      ii. Forward Difference operator: $(\Delta_+ z)_k = z_{k+1} - z_k$
      iii. Backward Difference operator: $\Delta_- z)_k = z_k - z_{k-1}$
      iv. Central Difference operator: $(\Delta_0 z)_k = z_{k+1/2} - z_{k-1/2}$
      v. Average operator: $(\mathcal{Y}_0 z)_k = \frac{1}{2}\left(z_{k+1/2} + z_{k-1/2}\right)$
      vi. Differential operator: $Dz_k = Dz(hk) = z'(hk)$

   (b) Properties

      i. Linearity: Let $\tau \in \{\mathcal{E}, \Delta_+, \Delta_-, \Delta_0, \mathcal{Y}_0, D\}$. Let $w, z \in \mathbb{R}^{\mathbb{Z}}$, and $a, b \in \mathbb{R}$. Then:

$$\tau(aw + bz) = a\tau(w) + b\tau(z)$$

      ii. Functions of operators: Let $g(x) = \sum_{j=0}^{\infty} a_j x^j$ be an arbitrary analytic function. Noting that $\mathcal{E} - I, \mathcal{Y}_0 - I, \Delta_0, \Delta_-, \Delta_+, hD$ all tend to 0 as $h \to 0^+$, we can expand g about $\mathcal{E} - I, \mathcal{Y}_0 - I, \Delta_0, \Delta_-, \Delta_+, hD$
      iii. Commutativity: all operators can be expressed in terms of $\mathcal{E}$ so they commute

$$\begin{aligned}
\Delta_+ &= \mathcal{E} - I \\
\Delta_- &= I - \mathcal{E}^- \\
\Delta_0 &= \mathcal{E}^{1/2} - \mathcal{E}^{-1/2} \\
\mathcal{Y}_0 &= \frac{1}{2}\left(\mathcal{E}^{1/2} + \mathcal{E}^{-1/2}\right) \\
D &= \frac{1}{h}\log\mathcal{E}
\end{aligned}$$

3. Forward and Backward Representations of D

   (a) Our primary goal is to represent D in different ways, and expand using the taylor series to get approximations of D to certain orders:

(b) First Derivative replacement by Δ_+ or Δ_-

$$\begin{aligned} D =& \frac{1}{h}\log(I+\Delta_+) \\ =& \frac{1}{h}\left[\Delta_+ - \frac{1}{2}\Delta_+^2 + \frac{1}{3}\Delta_+^3\right] + \mathcal{O}\left(h^3\right) \\ D =& \frac{-1}{h}\log(I-\Delta_-) \\ =& \frac{1}{h}\left[\Delta_- + \frac{1}{2}\Delta_-^2 + \frac{1}{3}\Delta_-^3\right] + \mathcal{O}\left(h^3\right) \end{aligned}$$

(c) Higher Order derivative replacement by Δ_+ or Δ_-: one strategy is to use multiplication to compute $s = 1, 2, 3$ and then find a pattern. The following are derived in this fashion:

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

4. Central Difference Formula (Representation of D)

(a) Note the following properties of $\mathcal{E}$ and $\mathcal{Y}_0$:

$$\begin{aligned} 0 =& \mathcal{E} - \Delta_0\mathcal{E}^{1/2} - I \\ \implies \sqrt{\mathcal{E}} =& \frac{\Delta_0 \pm \sqrt{\Delta_0^2 + 4I}}{2} \\ \implies D =& \frac{2}{h}\log\left[\frac{1}{2}\Delta_0 + \sqrt{I + \frac{1}{4}\Delta_0^2}\right] \end{aligned}$$

Secondly,

$$\begin{aligned} 4\mathcal{Y}_0^2 =& \mathcal{E} + 2I + \mathcal{E}^- \\ \Delta_0^2 =& \mathcal{E} - 2I + \mathcal{E}^- \\ \implies \mathcal{Y}_0 =& \left[I + \frac{1}{4}\Delta_0^2\right]^{1/2} \\ \implies I =& \mathcal{Y}_0\left[I + \frac{1}{4}\Delta_0^2\right]^{-1/2} \end{aligned}$$

Finally,

$$\Delta_0\mathcal{Y}_0 = \frac{1}{2}\left[\mathcal{E} - \mathcal{E}^-\right]$$

(b) We consider the Taylor expansion/General Binomial expansion of D for Central differences. Let $g(\xi) = \log(\xi + \sqrt{1+\xi^2})$.

$$\begin{aligned}
g'(\xi) =& \frac{1}{\sqrt{1+\xi^2}} = \sum_{j=0}^{\infty} (-1)^j \binom{2j}{j} \left(\frac{1}{2}\xi\right)^{2j} \\
g(\xi) - g(0) =& \sum_{j=0}^{\infty} (-1)^j \binom{2j}{j} \left(\frac{1}{2}\right)^{2j} \int_0^{\xi} \tau^{2j} d\tau \\
g(\xi) - 0 =& 2\sum_{j=0}^{\infty} \frac{(-1)^j}{2j+1} \binom{2j}{j} \left[\frac{1}{2}\xi\right]^{2j+1} \\
g(\frac{1}{2}\Delta_0) =& 2\sum_{j=0}^{\infty} \frac{(-1)^j}{2j+1} \binom{2j}{j} \left[\frac{1}{4}\Delta_0\right]^{2j+1}
\end{aligned}$$

(c) We therefore have for the first and even-ordered higher derivatives:

$$\begin{aligned}
D &= \frac{4}{h} \sum_{j=0}^{\infty} \frac{(-1)^j}{2j+1} \binom{2j}{j} \left[\frac{1}{4}\Delta_0\right]^{2j+1} \\
D^s &= \frac{1}{h^s} \left(\Delta_0^s - \frac{1}{24} s \Delta_0^{s+2} + \frac{1}{5760} s(5s+22) \Delta_0^{s+4} + \cdots \right)
\end{aligned}$$

(d) For higher-ordered odd derivatives we make use of $\mathcal{Y}_0$:

$$\begin{aligned}
I =& \mathcal{Y}_0 \left(I + \frac{1}{4}\Delta_0^2 \right)^{-1/2} \\
=& \mathcal{Y}_0 g'(\frac{1}{4}\Delta_0) \\
=& \mathcal{Y}_0 \sum_{j=0}^{\infty} (-1)^j \binom{2j}{j} \left(\frac{1}{4}\Delta_0\right)^{2j}
\end{aligned}$$

So for Higher-ordered odd derivatives, we can use:

$$\begin{aligned}
D^s =& I \frac{1}{h^s} \left(\Delta_0^s - \frac{1}{24} s \Delta_0^{s+2} + \frac{1}{5760} s(5s+22) \Delta_0^{s+4} + \cdots \right) \\
=& \frac{1}{h^s} \mathcal{Y}_0 \Delta_0 \left(\Delta_0^{s-1} - \frac{1}{24}(s+3) \Delta_0^{s+1} + \cdots \right)
\end{aligned}$$

6.2 Five-point Formula for $\nabla^2 u = f$

1. Poisson Equation
   (a) PDE: $\nabla^2 u = f$ for $(x, y) \in \Omega$ where $\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$
   (b) Dirichlet BC: $u(x, y) = \phi(x, y)$ for $(x, y) \in \partial\Omega$
2. Grid
   (a) We create a grid of equal spacing (Δx) in the x and y direction in the closure of Ω.

(b) Set Notation

i. $\Omega_{\Delta x} = \{(x_0 + k\Delta x, y_0 + l\Delta x) \in cl(\Omega) : k, l \in \mathbb{Z}\}$ are the grid points, for some (x_0, y_0) in the grid

ii. $I_{\Delta x} = \{(k, l) : (x_0 + k\Delta x, y_0 + l\Delta x) \in cl(\Omega)\}$

iii. $I^o_{\Delta x} = \{(k.l) : (x_0 + k\Delta x, y_0 + l\Delta x) \in \Omega\}$

(c) Point Classification

i. Boundary Point: a grid point on $\partial\Omega$

ii. Internal Point: a grid point inside Ω which is not a boundary or near-boundary point

iii. Near Boundary Point: a grid point with at least one neighbour not in $cl(\Omega)$.

3. Discretised Approximation

(a) We can discretise ∇^2 using the second derivative form for central differences:

$$\begin{aligned}\nabla^2 =& D_x^2 + D_y^2 \\ =& \frac{1}{(\Delta x)^2}\left(\Delta_x^2 + \Delta_y^2\right) + \frac{1}{(\Delta x)^2}\mathcal{O}\left(\Delta_x^4 + \Delta_y^4\right) \\ =& \frac{1}{(\Delta x)^2}\left(\Delta_x^2 + \Delta_y^2\right) + \mathcal{O}\left((\Delta x)^2\right)\end{aligned}$$

(b) Explicitly applied to the Poisson equation we have for some point $u_{k,l} = u(x_0 + k\Delta x, y_0 + l\Delta x)$ and $f_{k,l} = f(x_0 + k\Delta x, y_0 + l\Delta x)$:

$$\begin{aligned}f_{k,l} =& \frac{1}{(\Delta x)^2}\left(\Delta_x^2 + \Delta_y^2\right) u_{k,l} \\ =& \frac{1}{(\Delta x)^2}\left(u_{k+1,l} - 2u_{k,l} + u_{k-1,l} + u_{k,l+1} - 2u_{k,l} + u_{k,l-1}\right) \\ =& \frac{1}{(\Delta x)^2}\left(u_{k+1,l} + u_{k-1,l} + u_{k,l+1} + u_{k,l-1} - 4u_{k,l}\right)\end{aligned}$$

6.3 Error Analysis

1. Suppose we have differential operator

$$\mathcal{L} = \sum_{i,j \in I'} a'_{i,j} D_x^i D_y^j$$

2. Nothing that $D = \frac{1}{\Delta x}\log E$, we let:

$$L(E_x, E_y) = \sum_{i,j \in I'} a'_{i,j} \left(\log E_x\right)^i \left(\log E_y\right)^j$$

and

$$L_\Delta(E_x, E_y) = \sum_{i,j \in I} a_{i,j} E_x^i E_y^j$$

3. Poisson Example with Five Point Formula

Example 6.1. *We have that* $\nabla^2 = D_x^2 + D_y^2$. *Then:*

$$L(E_x, E_y) = (\log E_x)^2 + (\log E_y)^2$$

And for the five point approximation, we have:

$$L_\Delta(E_x, E_y) = E_x + E_x^{-1} + E_y + E_y^{-1} - 4$$

So we have the following:

$$\nabla^2 = (\Delta x)^{-2} L(E_x, E_y) = (\Delta x)^{-2} L_\Delta(E_x, E_y) + \mathcal{O}\left((\Delta x)^2\right)$$

4. Supposing that $\mathcal{L} = (\Delta x)^{-s} L$ and computing p for $L - L_\Delta = \mathcal{O}\left((\Delta x)^p\right)$, the error, given nice boundary conditions is then:

$$(\Delta x)^{-s} (L - L_\Delta) = \mathcal{O}\left((\Delta x)^{p-s}\right)$$

Example 6.2. *For the five point formula, the error is* $(\Delta x)^2$ *since:*

$$L_\Delta - L = \mathcal{O}\left((\Delta x)^4\right)$$

Example 6.3. *For the nine-point formula, the error is* $(\Delta x)^2$ *for the Poisson Equation* $\nabla^2 u = f$ *and* $(\Delta x)^4$ *for the Laplace Equation* $\nabla^2 u = 0$, *since:*

$$L_\Delta = L + \frac{1}{12} L^2 + \mathcal{O}\left((\Delta x)^6\right)$$

So that:

$$L_\Delta - \frac{1}{12} L_\Delta^2 = L + \mathcal{O}\left((\Delta x)^6\right)$$

To get the Poisson equation to have the same error as Laplace, we consider solving $\nabla^2 u - f = 0$ *instead, giving the system corresponding to:*

$$\left(1 - \frac{(\Delta x)^2}{12} L_\Delta\right) \left(L_\Delta u_{k,l} - (\Delta x)^2 f_{k,l}\right) = 0$$

5. Methodology: We can determine the error by simply Taylor expanding $L_\Delta(e^{i\theta}, e^{i\psi})$ and replacing the resulting terms on the right with $L(e^{i\theta}, e^{i\psi})$

6.4 Equations of Evolution

6.5 Stability Analysis

6.6 Examples and Practical Considerations

7 Finite Element Method

7.1 Principles of FEM Methods and Two-point BVP

7.1.1 Principles

1. Approximate solution in a finite dimensional space: $\phi_0 + \overset{\circ}{\mathbb{H}}_m \subset \mathbb{H}$

2. Retain only essential boundary conditions
3. Choose and approximation such that the defect is orthogonal to the space $\overset{\circ}{\mathbb{H}}_m$, or, equivalently, so that the variational problem is minimised in $\overset{\circ}{\mathbb{H}}_m$
4. Integrate by parts to depress differentiability requirements of $\overset{\circ}{\mathbb{H}}_m$
5. Choose functions which span a finite dimensional space (i.e. $\overset{\circ}{\mathbb{H}}_m$) such that they have small, supports in the domain of interest

7.1.2 Two-point Boundary Value Problem

1. The Two-Point Boundary Value Problem. For $x \in [0, 1]$:

$$-\frac{d}{dx}\left[a(x)\frac{du}{dx}\right] + b(x)u = f$$

   (a) Functions a, b, f are known
   (b) $a(x) > 0$, $b(x) \geq 0$ on $(0, 1)$
   (c) $a'(x)$ exists on $(0, 1)$

2. Dirichlet Boundary Conditions: $u(0) = \alpha$ and $u(1) = \beta$

7.1.3 Illustration of Principles with Two-point BVP

1. Approximate the Solution in a finite-dimensional function space:
   (a) We want to approximate u by $u_m = \phi_0(x) + \sum_{l=1}^{m} \gamma_l \phi_l(x)$ for $x \in [0, 1]$
   (b) ϕ_0 is required to satisfy the boundary conditions, and ϕ_l must vanish at the boundaries
   (c) ϕ_l define a space $\overset{\circ}{\mathbb{H}}_m$ and are linearly independent
2. To retain essential boundary conditions: see the next section on Variational Problems
3. Orthogonality of Defect to $\overset{\circ}{\mathbb{H}}_m = \text{span}(\phi_l)$ to our choice of approximation
   (a) For some choice of $\gamma_1, \ldots, \gamma_m$ we consider the defect:

$$d_m(x) = -\frac{d}{dx}\left[a(x)\frac{du_m}{dx}\right] + b(x)u_m(x) - f(x)$$

   (b) We want the defect to be orthogonal to $\overset{\circ}{\mathbb{H}}_m$ (i.e. minimising the defect using our function space)

(c) This results in the Galerkin Equations (using the Euclidean inner-product): for $k = 1, \ldots, m$

$$\begin{aligned}
0 =& \langle d_m, \phi_k \rangle \\
=& \int_0^1 d_m \phi_k dx \\
=& \int_0^1 - \left[a(x) \left(\phi_0 + \sum_{l=1}^m \gamma_l \phi_l \right)' \right]' \phi_k + b(x) \left(\phi_0 + \sum_{l=1}^m \gamma_l \phi_l \right) \phi_k \\
& - f \phi_k \\
=& \int_0^1 \left\{ - \left[a(x) \phi_0' \right]' + b(x) - f \right\} \phi_k + \sum_{l=1}^m \gamma_l \left\{ - \left[a(x) \phi_l' \right]' + b(x) \phi_l \right\} \phi_k
\end{aligned}$$

Equivalently:

$$\sum_{l=1}^m \gamma_l \int_0^1 - \left[a \phi_l' \right]' \phi_k + b \phi_l \phi_k = \langle f, \phi_k \rangle - \left(\int_0^1 - \left[a \phi_0' \right]' \phi_k + b \phi_0 \phi_k \right)$$

4. We now use integration by parts to reduce the integrability requirements of ϕ_l and hence the space $\overset{\circ}{\mathbb{H}}_m$

   (a) Notice that by integration by parts we have and vanishing end points:

   $$- \int_0^1 [a \phi_l']' \phi_k = \int_0^1 a \phi_l' \phi_k'$$

   (b) Therefore we have for $k = 1, \ldots, m$:

   $$\sum_{l=1}^m a_{k,l} \gamma_l = \int_0^1 f(x) \phi_k(x) dx - a_{k,0}$$

   where

   $$a_{k,l} = \int_0^1 a \phi_l' \phi_k' + b \phi_l \phi_k$$

5. Choosing Function Space

   (a) Sparse Method: Select $\phi_1, \ldots, \phi_m$ to be the most dense in $\mathcal{H}$; thus, $\|u_m - u\|$ is as small as possible.

   (b) FE Method: Select $\phi_1, \ldots, \phi_m$ to reduce computational costs:

      i. Each function ϕ_k is supported on some small set $\mathbb{E}_k \subset (0, 1)$

      ii. And $\mathbb{E}_k \cap \mathbb{E}_l = \emptyset$ when $k \neq l$ for as many (k, l) as possible

   Example 7.1. *Chapeau Function*

      i. The Chapeau Function: $\psi(x) = (1 - |x|)_+$

      ii. We select $\psi_k(x) = \psi(\frac{x}{h} - k)$ *for* $k = 1, \ldots, m$ *for* $h = \frac{1}{1+m}$

iii. We have for the supports that:

$$\mathbb{E}_k \cap \mathbb{E}_l = \begin{cases} [(k-1)h, (k+1)h] & \text{if } k = l \\ [(k-1)h, kh] & \text{if } k = l+1 \\ [kh, (k+1)h] & \text{if } k = l-1 \\ \emptyset & \text{if } |k-l| \geq 2 \end{cases}$$

iv. Therefore, the system resulting from the Galerkin Equations

$$\sum_{l=1}^{m} a_{k,l}\gamma_l = \langle f, \phi_k \rangle - a_{k,0}$$

for unknown γ_l becomes tridiagonal since $a_{k,l} = 0$ if $|k-l| \geq 2$. This requires only $\mathcal{O}(m)$ integrals to be calculated, allowing the system to be solved much faster.

7.2 Variational Formulation

1. Definition of a Variational Problem

   Definition 7.1. *Let $\mathbb{H}$ be a function space. Given a functional $\mathcal{J} : \mathbb{H} \to \mathbb{R}$, finding $u \in \mathbb{H}$ such that $\mathcal{J}(u) = \inf_{v \in \mathbb{H}} \mathcal{J}(v)$ is a variational problem.*

2. Variational problems and Euler-Lagrange formulations can be translated between each other

   Example 7.2. *Consider the two-point boundary value problem: For a, b, f on $[0,1]$ with $a(x) > 0$ and $b(x) \geq 1$, $\mathbb{H}$ the space of functions with $u(0) = \alpha$ and $u(1) = \beta$, and $\int_0^1 v^2 d\tau < \infty$ and $\int_0^1 (v')^2 d\tau < \infty$, the equivalent variational problem minimises:*

$$\mathcal{J}(v) = \int_0^1 a[v']^2\tau + bv^2 - 2fvd\tau$$

   (a) First we must look at the function space $\mathbb{H}$:
      i. $\mathbb{H}$ is not a linear space
      ii. For $u \in \mathbb{H}$ let $\overset{\circ}{\mathbb{H}} = \{v - u : v \in \mathbb{H}\}$. The elements of $\overset{\circ}{\mathbb{H}}$ obey the zero-boundary conditions and is a vector space
      iii. We call $\mathbb{H} = u + \overset{\circ}{\mathbb{H}}$ and affine space

   (b) Then we consider Bilinear Forms:
      i. Definition of Bilinear Forms

      Definition 7.2. *Given $\mathcal{L}$, a linear differential operator, a bilinear form $\tilde{a}(\cdot,\cdot)$ is defined as $\tilde{a}(v,w) = \langle \mathcal{L}v, w \rangle$ for $v, w \in \mathbb{H}$.*

      ii. Definition of properties of Bilinear Forms

      Definition 7.3. *Let $\mathcal{L}$ be a linear differential operator with bilinear form $\tilde{a}$*

      A. $\mathcal{L}$ is self-adjoint if $\tilde{a}(v,w) = \tilde{a}(w,v)$ for all $v, w \in \overset{\circ}{\mathbb{H}}$

B. $\mathcal{L}$ *is elliptic if* $\tilde{a}(v,v) > 0$ *for all* $v \in \overset{\circ}{\mathbb{H}}$

C. $\mathcal{L}$ *is positive definite if it is self-adjoint and elliptic.*

(c) The following theorem lets us interconvert between the Euler-Lagrange and Variational Problem formulations of a differential equation:

Theorem 7.1. *If* $\mathcal{L}$ *is positive definite then:*

i. $\mathcal{J}(v) = \tilde{a}(v,v) - 2\langle f, v\rangle$ *is the variational form of the Euler-Lagrange equation* $\mathcal{L}u = f$

ii. The weak solution of the Euler-Lagrange Problem is the minimum of the variational problem

3. Essential and Natural Boundary Conditions

Example 7.3. *Consider the following initial value problem for* $x \in [0,1]$*:*

$$\begin{cases} -u'' = f(x) \\ u'(0) = \gamma_0 \\ u(1) = 0 \end{cases}$$

We will multiply through by a function v*, depress the differentiability requirements of* u*:*

$$\int -u''v = \int fv$$

$$-u'(1)v(1) + u'(0)v(0) + \int_0^1 u'v' = \int fv$$

We can set $u'(0) = \gamma_0$*. This is a natural boundary condition because it will still be in the variational problem. But we cannot do the same for* $-u'(1)v(1)$*. We have the essential condition* $v(1) = 0$ *mirroring the behaviour of* $u(1)$*.*

$$\gamma_0 v(0) + \int_0^1 u'v' = \int fv$$

$$\tilde{a}(u,v) = \int_0^1 u'v'$$

$$L(v) = \int_0^1 fvdx - \gamma_0 v(0)$$

According to the theorem we have the following functional:

$$\mathcal{J}(v) = \int_0^1 u'v'dx - 2\left(\int_0^1 fvdx - \gamma_0 v(0)\right)$$

Example 7.4. *Consider the following initial value problem for* $x \in [0,1]$*:*

$$\begin{cases} -(k(x)u')' = f(x) \\ k(0)u'(0) = (u(0) - \gamma_0)\alpha \\ u(1) = 0 \end{cases}$$

We repeat the same process:

$$-\int (ku')'v = \int fv$$
$$-k(1)u'(1)v(1) + k(0)u'(0)v(0) + \int ku'v' = \int fv$$

The first term requires that we impose the essential boundary condition $v(1) = 0$. *The second term is natural, so that the equation becomes:*

$$\int ku'v'dx + \alpha u(0)v(0) = \int fvdx + \alpha\gamma_0 v(0)$$
$$\tilde{a}(u, v) = L(v)$$

The variational problem is then:

$$\mathcal{J}(v) = \left(\int k(v')^2 dx + \alpha v(0)^2\right) - 2\left(\int fvdx + \alpha\gamma_0 v(0)\right)$$

Message: If we can incorporate the boundary condition into the variational problem, then it is natural. If we must impose the boundary condition to get a nice variational problem, it is an essential condition.

7.3 Ritz and Generalised Galerkin Method

1. Ritz Method: ($\mathcal{L}$ is positive definite)

   (a) Method: Let $\phi_0 \in \mathbb{H}$, and $\phi_1, \ldots, \phi_m \in \overset{\circ}{\mathbb{H}}$ be linearly independent. We seek $\gamma = [\ \gamma_1\ \cdots\ \gamma_m\ ]^T \in \mathbb{R}^m$ to minimise

   $$\mathcal{J}_m(\gamma) = \mathcal{J}\left(\phi_0 + \sum_{l=1}^{m} \gamma_l \phi_l\right)$$

   (b) Setting $\nabla J_m = 0$, we have the Galerkin Equations for which:

   $$a_{k,l} = \tilde{a}(\phi_k, \phi_l)$$

   Derivation.

   $$\begin{aligned}\mathcal{J}_m(\gamma) = \int a\left[\phi_0' + \sum_{l=1}^{m}\gamma_l\phi_l'\right]^2 + b\left[\phi_0 + \sum_{l-1}^{m}\gamma_l\phi_l\right]^2 \\ - 2f\left[\phi_0 + \sum_{l-1}^{m}\gamma_l\phi_l\right]\end{aligned}$$
   $$\begin{aligned}\frac{\partial J_m}{\partial \gamma_k} = \int 2a\left[\phi_0' + \sum_{l=1}^{m}\gamma_l\phi_l'\right]\phi_k' + 2b\left[\phi_0 + \sum_{l-1}^{m}\gamma_l\phi_l\right]\phi_k \\ - 2f\phi_k\end{aligned}$$

Setting the integral equal to 0 and checking the second derivative to ensure a minimum:

$$\sum_{m=1}^{l} \gamma_l \int a\phi_l'\phi_k' + b\phi_l\phi_k = \int f\phi_k - \int a\phi_0'\phi_k' + b\phi_0\phi_k$$

$$\sum_{m=1}^{1} \gamma_l \tilde{a}(\phi_l, \phi_k) = \langle f, \phi_k \rangle - \tilde{a}(\phi_0, \phi_k)$$

□

(c) note that since $\mathcal{L}$ is self-adjoint, $a_{k,l} = a_{l,k}$

2. The Generalised Galerking Method: (for linear and non-linear operators) Find γ that minimises for $k = 1, \ldots, m$:

$$\tilde{a}\left(\phi_0 + \sum_{l=1}^{m} \gamma_l \phi_l \ , \ \phi_k \right) - \langle f, v \rangle = 0$$

7.4 Error

1. From calculus of variation we can prove that the minimiser of the variational problem exists and is unique, so the solution to the Galerkin equations exists and is unique, equivalently.

2. Definitions

Definition 7.4. *Given a bilinear form $\tilde{a}$:*

(a) A special case of the Sobolev norm: $\|v\|_{\mathbb{H}} = \|v\|^2 + \sqrt{\tilde{a}(v,v)}$

(b) $\tilde{a}$ is bounded if $\exists \delta > 0$ such that $|\tilde{a}(v,w)| \leq \delta \|v\|_{\mathbb{H}} \|u\|_{\mathbb{H}}$

(c) $\tilde{a}$ is coercive if $\exists \kappa > 0$ such that $\tilde{a}(v,v) \geq \kappa \|v\|_{\mathbb{H}}^2$

3. Lax-Milgram Theorem and Cea lemma on Error of Galerkin solution

Theorem 7.2. *Let $\mathcal{L}$ be linear, and the corresponding bilinear form be bounded and coercive. Let $\mathbb{V} < \overset{\circ}{\mathbb{H}}$ be a closed subspace. Then, $\exists! \tilde{u} \in \phi_0 + \mathbb{V}$ such that:*

(a) $\tilde{a}(\tilde{u}, v) - \langle f, v \rangle = 0$ for $v \in \mathbb{V}$

(b) for the weak solution u of $\mathcal{L}u = f$

$$\|\tilde{u} - u\|_{\mathbb{H}} \leq \frac{\delta}{\kappa} \inf \{\|v - u\|_{\mathbb{H}} : v \in \phi_0 + \mathbb{V}\}$$

Remark 7.1. *The second result is important: we know that $u \in \phi_0 + \overset{\circ}{\mathbb{H}}$. Let $\overset{\circ}{\mathbb{H}} = \mathbb{V}$. then this theorem gives the distance between u and the space $\phi_0 + \overset{\circ}{\mathbb{H}}_m$. Since we don not know u, we can bound the error with some distance between an arbitrary $w \in \phi_0 + \overset{\circ}{\mathbb{H}}$ and the space $\phi_0 + \overset{\circ}{\mathbb{H}}_m$*

7.5 Convergence

1. By Cea's Lemma, convergence of FEm requires that for an infinite sequence of $\left(\overset{\circ}{\mathbb{H}}_{m_i}\right)_{i=1}^{\infty} = \overset{\circ}{\mathbb{H}}$, where dim $\left(\overset{\circ}{\mathbb{H}}_{m_i}\right)$, and (m_i) is monotonically increasing to infinity, then

$$\lim_{i\to\infty} \|u_{m_i} - u\|_{\mathbb{H}} = 0$$

where u_{m_i} is the Galerkin solution in the space $\phi_0 + \overset{\circ}{\mathbb{H}}_{m_i}$

2. It is sufficient of show that $\forall v \in \overset{\circ}{\mathbb{H}}$,

$$\lim_{i\to\infty} \inf_{w\in\overset{\circ}{\mathbb{H}}_{m_i}} \|v - w\|_{\mathbb{H}} = 0$$

7.6 Practical Considerations for Finite Elements

1. Recall: we want $\phi_1^{[i]}, \ldots, \phi_{m_i}^{[i]} \in \overset{\circ}{\mathbb{H}}_{m_i}$ to have small supports so that we have minimal intersections in Ω.

2. Accordingly, we partition Ω into elements such that:

$$cl(\Omega) = \bigcup_{\alpha=1}^{n_i} cl(\Omega_\alpha^{[i]})$$

where

   (a) $\Omega_\alpha^{[i]} \cap \Omega_\beta^{[i]} = \emptyset$ for $\alpha \neq \beta$

   (b) $\phi_j^{[i]}$ are only allowed to span a few elements

   (c) The support of $\phi_j^{[i]}$ must contain whole elements

3. Requirements for selecting $\overset{\circ}{\mathbb{H}}_{m_i}$

   (a) $\overset{\circ}{\mathbb{H}}_{m_i} \subset \overset{\circ}{\mathbb{H}}$ so that the basis of the approximating space satisfies $\mathcal{L}$

   (b) Each finite element must be spanned by enough functions to approximate arbitrary functions

   (c) As i increases, and partitions re refined,

$$\lim_{i\to\infty} \max_{\alpha=1,\ldots,n_i} diam\left(\Omega_\alpha^{[i]}\right) = 0$$

   (d) As $i \to \infty$, the geometry of $\Omega_\alpha^{[i]}$ remain simple

4. Properties of FE Functions

   (a) Definitions

   Definition 7.5. *Let* $\overset{\circ}{\mathbb{H}}_m$ *be spanned by* $\Phi_m = \{\phi_1, \ldots, \phi_m\}$.

i. Φ_m *is of smoothness* q *if each* phi_j *is smoothly differentiable* $q-1$ *times and the* q^{th} *derivative exists in the sense of distributions*

ii. Φ_m *has accuracy* p *if inside each element we can represent a* p *degree polynomial as a linear combination of elements of* Φ_m

(b) Triangular Finite Elements

Example 7.5. *Consider piece-wise linear functions with "interpolation" conditions at the corner of the triangles represented by* $\bullet$*. These are called Pyramid Functions:* $\phi(x,y) = \alpha + \beta x + \gamma y$*. They have accuracy* $p = 1$*. To obtain a smoothness of* $q = 1$*, we need continuity across the edges. Since we have three degrees of freedom in selecting* $\phi(x,y)$*, across each edge, we can achieve a smoothness of* $q = 1$*. We can further require at interpolation points represented by* $\odot$ *that the function and its* x *and* y *derivatives match up across two elements.*

(c) Quadrilateral Finite Elements

Example 7.6. *Typically, bipolynomial functions are applied:* $\phi(x,y) = \psi_1(x)\psi_2(y)$*. An example would be the Pagoda Function.*

Natural Boundary Conditions

In contrast to essential boundary conditions, which are built into the solution space, natural boundary conditions are built into the weak form.

Example 1. Consider the 1-D Poisson equation with a Neumann boundary condition on the left and a homogeneous Dirichlet (essential) boundary condition on the right,

$$\begin{aligned} -\frac{d^2u}{dx^2} &= f(x), \quad 0 < x < 1, \\ \frac{du}{dx}(0) &= \gamma_0, \\ u(1) &= 0. \end{aligned} \tag{1}$$

If we multiply the ODE by $v(x)$ and integrate by parts, we obtain

$$\int_0^1 f(x)v(x)\,dx = -\int_0^1 \frac{d^2u}{dx^2}v(x)\,dx = \int_0^1 \frac{du}{dx}\frac{dv}{dx}\,dx - \frac{du}{dx}(1)v(1) + \underbrace{\frac{du}{dx}(0)}_{\gamma_0}v(0).$$

Suppose we impose the same right (essential) boundary condition on v as the solution u satisfies, so $v(1) = 0$. Then we obtain

$$\int_0^1 f(x)v(x)\,dx = \int_0^1 \frac{du}{dx}\frac{dv}{dx}\,dx + \gamma_0 v(0). \tag{2}$$

If we define

$$\begin{aligned} \mathcal{V}_0 &= \{v \in H^1[0,1] \,:\, v(1) = 0\} \\ a(u,v) &= \int_0^1 \frac{du}{dx}\frac{dv}{dx}\,dx \\ L(v) &= \int_0^1 f(x)v(x)\,dx - \gamma_0 v(0), \end{aligned}$$

then we obtain from (2) the weak form of the above Poisson equation: Find $u \in \mathcal{V}_0$ such that

$$a(u,v) = L(v) \quad \forall v \in \mathcal{V}_0.$$

The weak solution is also the minimizer over $\mathcal{V}_0$ of the energy functional

$$J(v) = \frac{1}{2}a(v,v) - L(v).$$

Remark. As before, one can show that a is bilinear, symmetric, coersive, and bounded on $\mathcal{V}_0$. Hence, the abstract theory of boundary value problems can be applied to show the existence, uniqueness, and continuous dependence of a weak solution in the Hilbert space $\mathcal{V}_0 \subset H^1[0,1]$.

Remark. If we impose on the weak solution u some additional smoothness, then we can show that u solves the Poisson eqn (1). If $u \in C^2[0,1]$, we can apply integration by parts to (2) to obtain

$$\int_0^1 f(x)v(x)\,dx = -\int_0^1 \frac{d^2u}{dx^2}v(x)\,dx + \frac{du}{dx}(1)v(1) - \frac{du}{dx}(0)v(0) + \gamma_0 v(0) \tag{3}$$

for all $v \in \mathcal{V}_0$. But if (3) holds for all $v \in \mathcal{V}_0$, it must hold for all $v \in H_0^1[0,1]$, i.e., it holds if we impose the additional restriction $v(0) = 0$ on top of the essential boundary condition $v(1) = 0$. Then

$$\int_0^1 \left[-\frac{d^2u}{dx^2} - f(x)\right] v(x)\,dx = 0 \quad \forall v \in H_0^1[0,1],$$

which implies that

$$-\frac{d^2u}{dx^2} - f(x) = 0 \quad \forall x \in [0,1],$$

so the ODE in (1) holds. If we substitute this into (3) we obtain

$$\frac{du}{dx}(1)v(1) + \left(\gamma_0 - \frac{du}{dx}(0)\right) v(0) = 0 \quad \forall v \in \mathcal{V}_0.$$

The condition $v \in \mathcal{V}_0$ forces $v(1) = 0$, so that

$$\left(\gamma_0 - \frac{du}{dx}(0)\right) v(0) = 0 \quad \forall v \in \mathcal{V}_0.$$

But $v \in \mathcal{V}_0$ allows $v(0)$ to vary. By picking $v(0) = 1$, we enforce the left boundary condition,

$$\gamma_0 - \frac{du}{dx}(0) = 0.$$

Example 2. Consider the 1-D steady-state diffusion equation with "radiation" boundary condition on the left and a homogeneous Dirichlet boundary condition on the right,

$$\begin{aligned} -\frac{d}{dx}\left(\kappa(x)\frac{du}{dx}\right) &= f(x), \quad 0 < x < 1, \\ \kappa(0)\frac{du}{dx}(0) &= \alpha(u(0) - \gamma_0), \\ u(1) &= 0, \end{aligned} \tag{4}$$

where α is a positive parameter. In the context of heat transfer, the left boundary condition means that the heat flux across the boundary is proportional to the difference between the boundary temperature $u(0)$ and some ambient temperature γ_0. If we multiply the ODE by $v(x)$ and integrate by parts and apply the boundary conditions for u and assume $v(1) = 0$ (essential BC), we obtain

$$\int_0^1 f(x)v(x)\,dx = \int_0^1 \kappa(x)\frac{du}{dx}\frac{dv}{dx}\,dx - \kappa(1)\frac{du}{dx}(1)\underbrace{v(1)}_{0} + \underbrace{\kappa(0)\frac{du}{dx}(0)}_{\alpha(u(0)-\gamma_0)} v(0)$$

and hence

$$\int_0^1 \kappa(x)\frac{du}{dx}\frac{dv}{dx} + \alpha u(0)v(0) = \int_0^1 f(x)v(x)\,dx + \alpha\gamma_0\, v(0).$$

The left hand side defines the bilinear form $a(u,v)$ and the right hand side gives the linear functional $L(v)$. The weak solution lies in the function space $\mathcal{V}_0 = \{v \in H^1[0,1] \,:\, v(1) = 0\}$. Clearly a is symmetric and it is easy to show that a is bounded with respect to the Sobolev H^1 norm. Coersivity can be established as long as the flux parameter α is nonnegative. The corresponding energy functional is

$$J(v) = \frac{1}{2}a(v,v) - L(v) = \int_0^1 \left[\frac{1}{2}\left(\frac{dv}{dx}\right)^2 - f(x)\right] v(x)\,dx + \frac{\alpha}{2}v(0)^2 - \alpha\gamma_0\, v(0).$$

Example 3. Natural Boundary Conditions in Higher Dimensions. Consider the steady-state diffusion equation with mixed boundary conditions

$$\begin{aligned} -\text{div}\,(\kappa(\mathbf{x})\nabla u) &= f(\mathbf{x}), && \mathbf{x} \in \Omega \\ u(\mathbf{x}) &= 0, && \mathbf{x} \in \Gamma_1 \quad \text{(essential BC)} \\ \kappa(\mathbf{x})\nabla u \cdot \mathbf{n}(\mathbf{x}) &= \gamma(\mathbf{x}), && \mathbf{x} \in \Gamma_0 \quad \text{(natural BC)} \end{aligned} \tag{5}$$

where $\mathbf{n}$ denotes the outward unit normal to the boundary and $\partial\Omega = \Gamma_1 + \Gamma_0$. By this we mean that Γ_1, Γ_0 form a *partition* of $\partial\Omega$, so that $\partial\Omega = \Gamma_1 \cup \Gamma_0$ and $\Gamma_1 \cap \Gamma_0 = \emptyset$. If we multiply the PDE by $v(\mathbf{x})$, apply Green's identity, apply the boundary conditions for u, and assume that v satisfies the (essential) homogeneous Dirichlet boundary conditions on Γ_1, we obtain

$$\begin{aligned} \int_\Omega \kappa(\mathbf{x})\nabla u \cdot \nabla v\, d\mathbf{x} &= \int_\Omega f(\mathbf{x})v(\mathbf{x})\,d\mathbf{x} + \int_{\partial\Omega} \kappa(\mathbf{x})\nabla u \cdot \mathbf{n}(\mathbf{x})\, v(\mathbf{x})\, dS \\ &= \int_\Omega f(\mathbf{x})v(\mathbf{x})\,d\mathbf{x} + \int_{\Gamma_1} \kappa(\mathbf{x})\nabla u \cdot \mathbf{n}(\mathbf{x})\, \underbrace{v(\mathbf{x})}_{0}\, dS + \int_{\Gamma_0} \gamma(\mathbf{x})v(\mathbf{x})\, dS. \end{aligned} \tag{6}$$

This gives us the weak form: Find $u \in \mathcal{V}_0$ such that

$$a(u,v) = L(v) \quad \forall v \in \mathcal{V}_0$$

where

$$\begin{aligned} \mathcal{V}_0 &= \{v \in H^1(\Omega) \,:\, v(\mathbf{x}) = 0\ \forall \mathbf{x} \in \Gamma_1\}. \\ a(u,v) &= \int_\Omega \kappa(\mathbf{x})\nabla u \cdot \nabla v\, d\mathbf{x} \\ L(v) &= \int_\Omega f(\mathbf{x})v(\mathbf{x})\,d\mathbf{x} + \int_{\Gamma_0} \gamma(\mathbf{x})v(\mathbf{x})\, dS \end{aligned}$$

As in the 1-D case, one can show that a is bilinear, symmetric, bounded, and coercive, so a unique weak solution u exists and minimizes $J(v) = a(v,v)/2 - L(v)$ over $v \in \mathcal{V}_0$. If $u \in C^2(\Omega)$, then one can show that the weak solution satisfies the PDE and the boundary conditions.

Exercises

1. Derive the weak form for the ODE BVP

$$\begin{aligned} -\frac{d}{dx}\left(\kappa(x)\frac{du}{dx}\right) &= f(x), \quad 0 < x < 1, \\ u(0) &= g_0, \\ \frac{du}{dx}(1) &= \gamma_1, \end{aligned}$$

   i.e., give the appropriate space $\mathcal{V}_0$ in which the weak solution lies, give the bilinear functional $a(u, v)$, and give the bounded linear functional $L(v)$ associated with this BVP. Note that you will need to apply a weak version of the superposition principle to handle the left boundary condition.

2. Implement the finite element method for the BVP in problem 1 with continuous piecewise linear "hat" basis functions to obtain an approximate solution with $\gamma_0 = 2$, $g_1 = 2e^1$, $\kappa = 1 + x^2$, and $f(x) = -2(1 + x)^2\, e^x$. The true solution is $u(x) = 2e^x$. Note that the left boundary condition is Dirichlet, but not homogeneous. Hand in the usual material, i.e., a description of your implementation, code listings, plots, and a summary of convergence results as you vary the grid spacing h.