Jacobs University Bremen **Spring Semester 2022**

Dr. D. Sinden

CA-MATH-804: Numerical Analysis[1]

Summary from 4 January 2024

Contents

[1]Note that the proofs for theorems marked with an * where presented in class.

1 Principles of Numerical Mathematics

Find x such that $F(x, d) = 0$ for a set of data, d and F, a functional relationship between x and d.

1.1 Well Posed Problems

Definition 1.1 (Well-Posed Problems). A problem is said to be **well-posed** if

- a solution exists,
- the solution is unique,
- the solution's behaviour changes continuously with the initial conditions.

A problem which does not have these properties is said to be **ill-posed**.

Definition 1.2 (Relative and Absolute Condition Numbers). The **relative condition number** of a problem is given by:

$$K(d) = \sup_{\delta d \in \mathcal{D}} \frac{\|\delta x\| \,/\, \|x\|}{\|\delta d\| \,/\, \|d\|}. \tag{1}$$

The **absolute condition number** is

$$K_{\text{abs}}(d) = \sup_{\delta d \in \mathcal{D}} \frac{\|\delta x\|}{\|\delta d\|}. \tag{2}$$

Consider a well-posed problem, then construct a sequence of approximate solutions via a sequence of approximate solutions and data, i.e. $F_n(x_n, d_n) = 0$

Definition 1.3 (Consistency). If the d is admissible for F_n, a numerical method $F_n(x_n, d_n) = 0$ is **consistent** if

$$\lim_{n \to \infty} F_n(x, d) \to F(x, d). \tag{3}$$

The method is strongly consistent if $F_n(x, d) = 0$ for all $n \geq 0$.

Given an approximate solution, x_n and solution x, the absolute and relative error are given by

$$E\left(x_n\right) = \left|x - x_n\right| \quad \text{and} \quad E_{\text{rel}}\left(x_n\right) = \frac{\left|x - x_n\right|}{|x|} \quad \text{if} \quad x \neq 0. \tag{4}$$

Definition 1.4 (Stability). **Stability** means that for any fixed n there exists a unique solution x_n for the data d_n and that the solution depends continuously on the data:

$$\forall\, \eta > 0 \quad \exists\, K = K\left(\eta, d_n\right) \quad \text{such that} \quad \|d_n\| < \nu \Rightarrow \|x_n\| < K\,\|d_n\|\,. \tag{5}$$

Definition 1.5 (Relative and Absolute Asymptotic Condition Numbers). If the sets of functions for $F_n(x_n, d_n) = 0$ and $F(x, d) = 0$ coincide, that is

$$K_n\left(d_n\right) = \sup_{\delta d_n \in \mathcal{D}_n} \frac{\|\delta x_n\| \,/\, \|x_n\|}{\|\delta d_n\| \,/\, \|d_n\|} \tag{6}$$

and

$$K_{n,\text{abs}}\left(d_n\right) = \sup_{\delta d_n \in \mathcal{D}_n} \frac{\|\delta x_n\|}{\|\delta d_n\|} \tag{7}$$

then the **relative asymptotic condition number** is

$$K^{\text{num}}(d) = \lim_{k\to\infty} \sup_{n\leq k} K_n\left(d_n\right). \tag{8}$$

The **absolute asymptotic condition number** is

$$K^{\text{num}}_{\text{abs}}(d) = \lim_{k\to\infty} \sup_{n\leq k} K_{n,\text{abs}}\left(d_n\right). \tag{9}$$

Definition 1.6 (Convergence). A method is **convergent** if and only if:

$$\forall \varepsilon > 0, \quad \exists\, n \quad \text{such that} \quad \|x(d) - x_n\left(d + \delta d_n\right)\| \leq \varepsilon. \tag{10}$$

Theorem 1 (Lax-Ritchmyer). A numerical algorithm converges if and only if it is consistent and stable.

Definition 1.7 (Inner Product). An **inner product** (sometimes called a scalar product) is a function $(\cdot,\cdot) : V \times V \to F$ which takes two members of a vector space V and maps them to a field, F (that is either the real or complex numbers) and has the following properties:

1. Symmetry: $(x, y) = (y, x)$, indeed, conjugate symmetry $(x, y) = \overline{(y, x)}$ (also called Hermitian).
2. Non-negativity: $(x, x) > 0$ for every $x \in \mathbb{R}^n$ and $(x, x) > 0$ if and only if $x = 0$, the zero vector.
3. Linearity: $(ax + by, z) = a\,(x, z) + b\,(y, z)$.

An inner product leads to notions of distance and angle.

Definition 1.8 (Orthogonality)**.** Two vectors are said to be **orthogonal** if $(x, y) = 0$.

Definition 1.9 (Norms and Semi-Norms)**.** An operator $\|\cdot\| : V \to \mathbb{R}$ is called a **norm** if

1. Non-negativity:
   - (i) $\|x\| \geq 0$ for every $x \in \mathbb{R}^n$
   - (ii) $\|x\| = 0$ if and only if $x = 0$, the zero vector.
2. Linearity: $\|\alpha x\| = |\alpha| \, \|x\|$.
3. Triangle Inequality: $\|x + y\| \leq \|x\| + \|y\|$.

An operator $|\cdot|_V : V \to \mathbb{R}$ which is linear, satisfies the triangle inequality but only satisfies the first condition of non-negativity is called a **semi-norm**.

Inner products can induce norms, that is $\|x\| = \sqrt{(x, x)}$. The inner product satisfies the Cauchy–Schwarz inequality

$$|(x, y)| \leq \|x\| \, \|y\| \,. \tag{11}$$

Let $p \geq 1$ be a real number. The ***p*-norm** (also called ℓ_p-norm) of vector $\boldsymbol{x} = (x_1, \ldots, x_n)$ is given by

$$\|\boldsymbol{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} . \tag{12}$$

2 Matrix Analysis

Matrix norms can be produced from the vector norms:

$$\|A\|_{p,q} = \sup_{\boldsymbol{x} \neq \mathbf{0}} \frac{\|A\boldsymbol{x}\|_p}{\|\boldsymbol{x}\|_q}. \tag{13}$$

and

$$\|A\|_p = \sup_{\boldsymbol{x} \neq \mathbf{0}} \frac{\|A\boldsymbol{x}\|_p}{\|\boldsymbol{x}\|_p}. \tag{14}$$

This is called an **induced matrix norm**. Note that any induced norm of the identity matrix is 1.

Without loss of generality, now consider the case when $\|x\| = 1$. There are three main types of p-norm:

$$\|A\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^{m} |a_{ij}|, \tag{15}$$

which is simply the maximum absolute column sum of the matrix. The **infinity norm** is given by

$$\|A\|_\infty = \max_{1 \leq i \leq m} \sum_{j=1}^{n} |a_{ij}| \tag{16}$$

which is simply the maximum absolute row sum of the matrix. In the special case of $p = 2$ the induced matrix norm is called the **spectral norm**.

The spectral norm of a matrix A is the largest singular value of A (i.e., the square root of the largest eigenvalue of the matrix $A^H A$, where A^H denotes the conjugate transpose of A

$$\|A\|_2 = \sqrt{\sigma_{\max}\left(A^H A\right)} \tag{17}$$

where $\sigma_{\max}(A)$ represents the largest singular value of the matrix A. Also,

$$\|A^* A\|_2 = \|A A^*\|_2 = \|A\|_2^2. \tag{18}$$

Related to the spectral norm is the **Frobenius norm** given by

$$\|A\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2}. \tag{19}$$

it can also be expressed as

$$= \sqrt{\operatorname{trace}\left(A^H A\right)} \tag{20}$$

where the trace is the sum of the diagonal elements of a matrix, a_{ii}, and

$$= \sqrt{\sum_{i=1}^{\min(n,m)} \sigma_i\left(A\right)}. \tag{21}$$

Theorem 2*. Let $A \in \mathbb{R}^{n\times n}$, then

1. $\lim_{k\to\infty} A^k = 0 \Leftrightarrow \rho(A) < 1$. Where $\rho(A)$ is the largest absolute value of the eigenvalues of A. This is called the **spectral radius**

2. The geometric series, $\sum_{k=0}^{\infty} A^k$ is convergent if and only if $\rho(A) < 1$. Then in this case, the sum is given by

$$\sum_{k=0}^{\infty} A^k = (I - A)^{-1}. \tag{22}$$

3. Thus, if $\rho(A) < 1$, the matrix $I - A$ is invertible and

$$\frac{1}{1 + \|A\|} \leq \left\|(I - A)^{-1}\right\| \leq \frac{1}{1 - \|A\|} \tag{23}$$

where $\|\cdot\|$ is an induced matrix norm such that $\|A\| < 1$.

Theorem 3*. Let $A \in \mathbb{R}^{n\times n}$ be non-singular and let $\delta A \in \mathbb{R}^{n\times n}$ be such that $\left\|A^{-1}\right\| \left\|\delta A\right\| < 1$. Furthermore, if $x \in \mathbb{R}^n$ is a solution to $Ax = b$, where $b \in \mathbb{R}^n$ and $b \neq 0$ and δx is such that

$$(A + \delta A)(x + \delta x) = b + \delta b \tag{24}$$

for a $\delta b \in \mathbb{R}^n$, then

$$(A + \delta A)(x + \delta x) \leq \frac{K(A)}{1 - K(A)\left\|\delta A\right\|_2 / \left\|A\right\|_2} \left(\frac{\|\delta b\|_2}{\|b\|_2} + \frac{\|\delta A\|_2}{\|A\|_2}\right). \tag{25}$$

Theorem 4*. Let $A \in \mathbb{R}^{n\times n}$ be non-singular and if $x \in \mathbb{R}^n$ is a solution to $Ax = b$, where $b \in \mathbb{R}^n$ and $b \neq 0$ and δx is such that

$$A(x + \delta x) = b + \delta b \tag{26}$$

then

$$\frac{1}{K(A)} \frac{\|\delta b\|}{\|b\|} \leq \frac{\|\delta x\|}{\|x\|} \leq K(A) \frac{\|\delta b\|}{\|b\|}. \tag{27}$$

Theorem 5. For $A \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$, assume $\|\delta A\| \leq \gamma \|A\|$ and $\|\delta b\| \leq \gamma \|b\|$ for some $\gamma \in \mathbb{R}^+$. Then, if $\gamma K(A) < 1$, then the following holds

$$\frac{\|x + \delta x\|}{\|x\|} \leq \frac{1 + \gamma K(A)}{1 - \gamma K(A)} \tag{28}$$

and

$$\frac{\|\delta x\|}{\|x\|} \le \frac{2\gamma K(A)}{1-\gamma K(A)}. \tag{29}$$

Theorem 6. For A, $C \in \mathbb{R}^{n\times n}$, let $R = AC - I$. If $\|R\|_2 < 1$ and

$$\left\|A^{-1}\right\| \le \frac{\|C\|}{1-\|R\|} \tag{30}$$

and

$$\frac{\|R\|}{\|A\|} \le \left\|C - A^{-1}\right\| \le \frac{\|C\|\,\|R\|}{1-\|R\|}. \tag{31}$$

In the framework of backwards a priori analysis we can interpret C as being the inverse of $A+\delta A$ (for a suitable unknown δA). We are thus assuming that $C(A+\delta A) = I$. This yields

$$\delta A = C^{-1} - A = -(AC - I)C^{-1} = -RC^{-1} \tag{32}$$

and, as a consequence, if $\|R\| < 1$ it turns out that

$$\begin{aligned} \left\|\delta A\right\| &\le \|R\| \left\|C^{-1}\right\| \\ &\le \frac{\|R\|\,\|A\|}{1-\|R\|}. \end{aligned} \tag{33}$$

3 Iterative Solutions for Matrix Inversion

Construct a scheme which solves the linear system $Ax = b$ by generating a sequence $\{x^{(n)}\}$ which approximates the solution, x, that is

$$\lim_{n\to\infty} x^{(n)} = x. \tag{34}$$

So that $x = A^{-1}b$. Split the matrix $A = P - N$ and solve

$$Px^{(n+1)} = Bx^{(n)} + f, \tag{35}$$

where P is called a **preconditioner** and $B = P^{-1}N$ is the **iteration matrix**.

An equivalent formulation is given by

$$x^{(k+1)} = x^{(k)} + P^{-1}r^{(k)} \tag{36}$$

where

$$r^{(k)} = b - Ax^{(k)} \tag{37}$$

is the **residual**.

Definition 3.1 (Consistency). An iterative method is said to be **consistent** if $x = Bx + f$, or equivalently,

$$f = (I - B)A^{-1}b. \tag{38}$$

Theorem 7. If an iterative scheme is consistent, then if and only if $\rho(B) < 1$ the method will converge for any initial guess $x^{(0)}$.

Definition 3.2 (Stationary Methods). The formulation can be written as

$$\begin{aligned} x^{(0)} &= F^{(0)}(A, b) \quad \text{and} \\ x^{(k+1)} &= F^{(k+1)}\left(x^{(k)}, x^{(k-1)}, \ldots, x^{(0)}, A, b\right). \end{aligned} \tag{39}$$

If the functions $F^{(k)}$ are independent of the number of iterations, then it is said to be **stationary**.

3.1 Jacobi Method

The Jacobi method decomposes the matrix A into diagonal, lower and upper triangular matrices $A = D + L + U$, and solves

$$Dx^{(n+1)} = -(L + U)x^{(n)} + b. \tag{40}$$

Element-wise this is

$$x_i^{(k+1)} = \frac{1}{a_{ii}}\left(b_i - \sum_{j=1, j\neq i}^{n} a_{ij}x_j^{(k)}\right). \tag{41}$$

Thus, the iterative scheme is

$$x^{(n+1)} = -D^{-1}(L+U)x^{(n)} + D^{-1}b. \tag{42}$$

As $L+U = A-D$, so the iteration matrix can be written as $B = I - D^{-1}A$.

3.2 Over-Relaxation of Jacobi Method

Also called the weighted Jacobi method. Introduce ω to solve

$$x_i^{(k+1)} = \frac{\omega}{a_{ii}}\left(b_i - \sum_{j=1, j\neq i}^{n} a_{ij}x_j^{(k)}\right) + (1-\omega)\,x^{(k)}. \tag{43}$$

3.3 Successive Over-Relaxation

Introduce ω to solve

$$(D+\omega L)\,x^{(n+1)} = -((\omega-1)D + \omega U)x^{(n)} + \omega b. \tag{44}$$

3.4 Gauss-Seidel

The Gauss-Seidel method decomposes the matrix A into diagonal, lower and upper triangular matrices $A = D + L + U$, and solves

$$(D+L)x^{(n+1)} = -Ux^{(n)} + b \tag{45}$$

Theorem 8. 1. If A is strictly diagonally dominant by rows, that is $|a_{ii}| > \sum_{j\neq i}|a_{ij}|$, the Jacobi and Gauss-Seidel methods are convergent.

2. If A and $2D - A$ are symmetric and positive definite, then the Jacobi method is convergent and the spectral radius of the iteration matrix B is equal to

$$\rho\,(B) = \|B\|_A = \|B\|_D \tag{46}$$

where $\|\cdot\|_A$ is the energy norm which is induced by the vector norm $\|x\|_A = \sqrt{x\cdot Ax}$

3. If and only if A is symmetric and positive definite, the Jacobi over-relaxation method is convergent if

$$0 < \omega < \frac{2}{\rho\,(D^{-1}A)}. \tag{47}$$

4. If and only if A is symmetric and positive definite, the Gauss-Seidel method is monotonically convergent with respect to the energy norm $\|\cdot\|_A$.

Theorem 9*. For any $\omega \in \mathbb{R}$ we have $\rho(B(\omega)) \geq |\omega - 1|$. Thus, SOR does not converge if either $\omega \leq 0$ or $\omega \geq 2$.

Theorem 10 (Ostrowski). If A is symmetric and positive definite, then the SOR method is convergent if and only if $0 < \omega < 2$. Furthermore, the convergence is monotonic with respect to the energy norm $\|\cdot\|_A$.

3.5 Gradient Descent

Consider the function $\Phi(y) : \mathbb{R}^n \mapsto \mathbb{R}$ which takes the form:

$$\Phi(y) = \frac{1}{2} y \cdot Ay - y \cdot b. \tag{48}$$

It can be shown that solving $Ax = b$ is equivalent to minimizing Φ.

If x is a solution to the linear system and minimizes $\Phi(x)$ then $\nabla\Phi(x) = 0$, so that $Ax - b = \nabla\Phi(x) = 0$.

Now express the function as

$$\begin{aligned} \Phi(y) &= \Phi(x + (y - x)) \\ &= \Phi(x) + \frac{1}{2} \|y - x\|_A^2 . \end{aligned} \tag{49}$$

Where $\|\cdot\|_A^2$ is the energy norm from the matrix A. Thus, from equation (49), it is possible to show that as the Hessian of the system, $\nabla^2\Phi = A$, is symmetric and positive-definite and x is a solution to the linear system and hence minimizes Φ, then if $\Phi(y) = 0$, so y is equal to x. That is the gradient descent provides a unique solution.

Gradient descent seeks to construct a scheme which updates the vector $x^{(k)}$ according to

$$x^{(k+1)} = x^{(k)} + \alpha^{(k)} d^{(k)} \tag{50}$$

where $d^{(k)}$ is the update direction and $\alpha^{(k)}$ is the step size at the k-th iterate.

Note that in contrast to the methods above, the gradient descent method is non-stationary as values d and α change at every iterate.

The idea is to let the search direction be the gradient of the function Φ

$$\begin{aligned} d^{(k)} &= -\nabla\Phi\left(x^{(k)}\right) \\ &= -\left(Ax^{(k)} - b\right) \\ &= b - Ax^{(k)} \\ &= r^{(k)}. \end{aligned} \tag{51}$$

The step size is found by differentiating Φ with respect to α and setting this to zero, so that

$$\alpha^{(k)} = \frac{r^{(k)} \cdot r^{(k)}}{r^{(k)} \cdot Ar^{(k)}}. \tag{52}$$

Theorem 11*. If A is symmetric and positive definite, then the gradient-descent method is convergent for any $x^{(0)}$ and

$$\left\| e^{(k+1)} \right\|_A \leq \frac{K(A)-1}{K(A)+1} \left\| e^{(k)} \right\|_A . \tag{53}$$

If we apply a preconditioner, i.e. multiplying both sides of the linear system from the left by P^{-1}, then the rescaled linear system is $\tilde{A}x = \tilde{b}$, where $\tilde{A} = P^{-1}A$ and $\tilde{b} = P^{-1}b$. Then the a good preconditioner will reduce the condition number of the new linear system.

3.6 Conjugate Gradient

Definition 3.3 (Conjugate Vectors)**.** If A is symmetric and positive definite, let the vectors u and v be ***A*-orthogonal** or **conjugate** if $u \cdot Av = 0$.

Lemma 3.4*. Choosing $p^{(k+1)}$ such that

$$p^{(k+1)} \cdot Ap^{(j)} = 0 \tag{54}$$

for $j = 0, \ldots, k$ leads to

$$p^{(j)} \cdot r^{(k+1)} = 0. \tag{55}$$

Lemma 3.5*. Setting

$$\beta^{(k)} = \frac{r^{(k+1)} \cdot Ap^{(k)}}{p^{(k)} \cdot Ap^{(k)}} \tag{56}$$

and

$$p^{(k+1)} = r^{(k+1)} - \beta^{(k)} p^{(k)} \tag{57}$$

then, for $j = 0, \ldots, k$, yields

$$p^{(k+1)} \cdot Ap^{(j)} = 0. \tag{58}$$

Theorem 12*. If $A \in \mathbb{R}^{n \times n}$ is a symmetric and positive definite matrix, and $b \in \mathbb{R}^n$, then the conjugate gradient method yields the exact solution of $Ax = b$ after n steps.

4 Interpolation

Numerical treatment of problems often involves the process of *discretization* - i.e. going from a continuous function to set of discrete points.

Interpolation provides a way of approximating continuous functions by discrete data.

Types of functions which can be used are:

- **Polynomial interpolation** : using a polynomial to approximate the data,
- **Trigonometric interpolation**: using polynomials of trigonometric functions,
- **Spline interpolation**: using a set of piecewise polynomials over subintervals of the data.

Theorem 13*. Given $n+1$ distinct points $x_0, x_1, \ldots, x_n$ and $n+1$ corresponding values $y_0, y_1, \ldots, y_n$ there exists a *unique* polynomial $\Pi_n \in \mathbb{P}_n$ such that for all $i = 0, \ldots, n$

$$\Pi_n\left(x_i\right) = y_i. \tag{59}$$

4.1 Lagrange Interpolation

Definition 4.1 (Lagrange Polynomials)**.** The **Lagrange form of an interpolating polynomial** is given by

$$\Pi_n\left(x\right) = \sum_{i=0}^{n} y_i l_i\left(x\right) \tag{60}$$

where $l_i \in \mathbb{P}_n$ such that $l_i\left(x_j\right) = \delta_{ij}$. The polynomials $l_i\left(x\right) \in \mathbb{P}_n$ for $i = 0, \ldots, n$, are called **characteristic polynomials** and are given by

$$l_i\left(x\right) = \prod_{j=0, j \neq i}^{n} \frac{x - x_j}{x_i - x_j}. \tag{61}$$

Theorem 14*. Let $x_0, x_1, \ldots x_n$ be $n+1$ distinct nodes and let x be a point belonging to the domain of a given function f. Let I_x be the smallest interval containing the nodes $x_0, x_1, \ldots x_n$ and x and assume that $f \in C^{n+1}\left(I_x\right)$. Then the interpolation error at the point x is defined and given by

$$\begin{aligned} E_n(x) &= f(x) - \Pi_n f(x) \\ &= \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega_{n+1}(x) \end{aligned} \tag{62}$$

where $f^{(n+1)}$ is the $(n+1)^{\text{th}}$ derivative of f, $\xi \in I_x$ and ω_{n+1} is the nodal polynomial of degree $n+1$, which is defined as

$$\omega_{n+1}(x) = \prod_{i=0}^{n} (x - x_i)\,. \tag{63}$$

4.2 Piecewise Lagrange Interpolation

Partition $\mathcal{T}_h$ of $[a,b]$ into K subintervals $I_j = [x_j, x_{j+1}]$ of length h_j such that $[a,b] = \bigcup_{j=0}^{K-1} I_j$. Let $h = \max_{0\leq j\leq K-1} h_j,$.

For $k \geq 1$, introduce on $\mathcal{T}_h$ the piecewise polynomial space

$$X_h^k = \left\{ v \in C^0(a,b) \,:\, v|_{I_j} \in \mathbb{P}_k\left(I_j\right) \quad \forall\, I_j \in \mathcal{T}_h \right\} \tag{64}$$

which is the space of the continuous functions over the interval $[a,b]$ whose restrictions on each I_j are polynomials of degree less than or equal to k.

Then, for any continuous function f in $[a,b]$, the piecewise interpolation polynomial $\Pi_h^k f$ coincides on each I_j with the interpolating polynomial of $f|_{I_j}$ at the $n+1$ nodes $\left\{ x_j^{(i)}, 0 \leq i \leq n \right\}$.

As a consequence, if $f \in C^{k+1}(a,b)$, then from (62) within each interval the following error estimate holds

$$\left\| f - \Pi_h^k f \right\|_\infty \leq C h^{k+1} . \left\| f^{(k+1)} \right\|_\infty . \tag{65}$$

Definition 4.2 (L^2 Space)**.** Define the $\mathbf{L^2}$ **function space** as the collection of all functions such that

$$\mathrm{L}^2(a,b) = \left\{ f : (a,b) \to \mathbb{R}, \int_a^b |f(x)|^2 \, \mathrm{d}x < +\infty \right\} \tag{66}$$

with the norm

$$\|f\|_{\mathrm{L}^2(a,b)} = \left(\int_a^b |f(x)|^2 \mathrm{d}x \right)^{1/2} . \tag{67}$$

This defines a norm for $\mathrm{L}^2(a,b)$. Note that integral of the function $|f|^2$ is in the Lebesgue sense - in particular, f needs not be continuous everywhere. Functions for which the integral is exists and is finite are called square integrable. Functions in L^2 are said to be square integrable.

Theorem 15*. Using Lagrange interpolation on each subinterval I_j using $n+1$ equally spaced nodes $\left\{ x_j^{(i)}, 0 \leq i \leq n \right\}$ with a small n. Then Π_n^k is the

piecewise interpolation polynomial.
Let $0 \leq m \leq k+1$, with $k \geq 1$ and assume that $f^{(m)} \in \mathrm{L}^2(a,b)$ for $0 \leq m \leq k+1$ then there exists a positive constant C, independent of h, such that

$$\left\| \left(f - \Pi_h^k f\right)^{(m)} \right\|_{\mathrm{L}^2(a,b)} \leq C h^{k+1-m} \left\| f^{(k+1)} \right\|_{\mathrm{L}^2(a,b)}. \tag{68}$$

In particular, for $k=1$ and $m=0$, or $m=1$

$$\left\| f - \Pi_h^1 f \right\|_{\mathrm{L}^2(a,b)} \leq C_1 h^2 \left\| f'' \right\|_{\mathrm{L}^2(a,b)} \tag{69a}$$

and

$$\left\| \left(f - \Pi_h^1 f\right)' \right\|_{\mathrm{L}^2(a,b)} \leq C_2 h \left\| f'' \right\|_{\mathrm{L}^2(a,b)} \tag{69b}$$

for two suitable positive constants C_1 and C_2.

5 Integration

If $f \in C^0(a,b)$, the quadrature error $E_n(f) = I(f) - I_n(f)$ satisfies

$$|E_n(f)| \leq \int_a^b |f(x) - f_n(x)| \, \mathrm{d}x \leq (b-a) \|f - f_n\|_\infty \tag{70}$$

Therefore, if for some n, $\|f - f_n\|_\infty < \varepsilon$, then $|E_n(f)| \leq \varepsilon(b-a)$.

The approximation of the function f_n must be easily integrable, which is the case if, for example, $f_n \in \mathbb{P}_n$. In this respect, a natural approach consists of using $f_n = \Pi_n f$, the interpolating Lagrange interpolatory polynomial of f over a set of $n+1$ distinct nodes $\{x_i\}$, with $i = 0, \ldots, n$. It follows that the approximation to the integral is

$$I_n(f) = \sum_{i=0}^{n} f(x_i) \int_a^b l_i(x) \mathrm{d}x \tag{71}$$

where l_i is the characteristic Lagrange interpolatory polynomial of degree n associated with node x_i. It is called the **Lagrange quadrature formula**, and is a special instance of the following, generalised, quadrature formula

$$I_n(f) = \sum_{i=0}^{n} \alpha_i f(x_i) \tag{72}$$

where the coefficients α_i of the linear combination are given by $\int_a^b l_i(x) \, \mathrm{d}x$. The above equation is a weighted sum of the values of f at the points x_i, for $i = 0, \ldots, n$. These points are said to be the nodes of the quadrature formula, while the $\alpha_i \in \mathbb{R}$ are its *coefficients* or *weights*. Both weights and nodes depend in general on n.

Another approximation of the function f leads to the **Hermite quadrature formula**

$$I_n(f) = \sum_{k=0}^{1} \sum_{i=0}^{n} \alpha_{ik} f^{(k)}(x_i) \tag{73}$$

where the weights are now denoted by α_{ik}. This depends on an evaluation of the function and its derivative.

Both the above are *interpolatory quadrature formula*, since the function f has been replaced by its interpolating polynomial (Lagrange and Hermite polynomials, respectively).

Define the **degree of exactness** of a quadrature formula as the maximum integer $r \geq 0$ for which

$$I_n(f) = I(f), \quad \forall f \in \mathbb{P}_r. \tag{74}$$

Any interpolatory quadrature formula that makes use of $n+1$ distinct nodes has degree of exactness equal to at least n. Indeed, if $f \in \mathbb{P}_n$, then $\Pi_n f = f$ and thus $I_n(\Pi_n f) = I(\Pi_n f)$.

The converse statement is also true, that is, a quadrature formula using $n+1$ distinct nodes and having degree of exactness equal at least to n is necessarily of interpolatory type.

5.1 Midpoint Rule

$$I_0 = (b-a) f\left(\frac{a+b}{2}\right). \tag{75}$$

5.2 Trapezoidal Rule

$$I_1 = \frac{b-a}{2}\left(f\left(a\right) + f\left(b\right)\right). \tag{76}$$

5.3 Simpson's Rule

$$I_2 = \frac{b-a}{6}\left(f\left(a\right) + 4f\left(\frac{a+b}{2}\right) + f\left(b\right)\right). \tag{77}$$

5.4 Gaussian Integration

Gaussian quadrature integrates a function by a suitable choice of both *nodes* and *weights*.

Theorem 16*. With the exact integral of f

$$I_g(f) = \int_{-1}^{1} f(x)g(x)\,\mathrm{d}x, \tag{78}$$

being $f \in C^0\left(-1,1\right)$, consider quadrature rules of the type

$$I_{n,g}(f) = \sum_{i=0}^{n} \alpha_i f(x_i) \tag{79}$$

where α_i are to be determined.
For a given $m > 0$, the quadrature $I_{n,g}$ has degree of exactness $d = n + m$ if and only if it is of interpolatory type and the nodal polynomial ω_{n+1} associated with the set of nodes $\{x_i\}$, is such that

$$\int_{-1}^{1} \omega_{n+1}(x)p(x)g(x)\,\mathrm{d}x = 0, \quad \forall\, p \in \mathbb{P}_{m-1}. \tag{80}$$

6 Finite Difference Methods

6.1 Green's functions

For a linear differential operator acting on u, that is $\mathcal{L}\left[u\left(x\right)\right]$, which has a differential equation of the form

$$\mathcal{L}\left[u\left(x\right)\right] = f\left(x\right), \tag{81}$$

then the **Green's function** for the operator $\mathcal{L}$, denoted by $G\left(x, s\right)$, can be used to solved the differential equation as

$$u(x) = \int^{x} G\left(x, s\right) f\left(s\right) \mathrm{d}s. \tag{82}$$

6.2 Finite Difference Methods

First discretize the domain and then approximate the governing equation to produce a linear system.

Definition 6.1 (Finite-Difference Quotients)**.** There are approximations to the first-order derivative at x_j

1. **Forward Difference Quotient:**

$$D_j^{+} u = \frac{u_{j+1} - u_j}{h} \tag{83}$$

2. **Backwards Difference Quotient:**

$$D_j^{-} u = \frac{u_j - u_{j-1}}{h} \tag{84}$$

3. **Central Difference Quotient:**

$$D_j^{0} u = \frac{u_{j+1} - u_{j-1}}{2h}. \tag{85}$$

With these, approximations to second-order derivatives can be constructed, for example:

$$\begin{aligned} D_j^{\pm} u &= \frac{D_j^{+} u - D_j^{-} u}{h} \\ &= \frac{\dfrac{u_{j+1} - u_j}{h} - \dfrac{u_j - u_{j-1}}{h}}{h} \\ &= \frac{u_{j+1} - 2u_j + u_{j-1}}{h^2}. \end{aligned} \tag{86}$$

Theorem 17 (Errors for Finite-Difference Quotients). The errors for the approximation of the derivatives are given by

1. $u(x_j) - D_j^+ u = -\frac{h}{2} u''(\xi)$ where $\xi \in (x_j, x_{j+1})$
2. $u(x_j) - D_j^+ u = \frac{h}{2} u''(\xi)$ where $\xi \in (x_{j-1}, x_j)$
3. $u(x_j) - D_j^+ u = -\frac{h^2}{6} u'''(\xi)$ where $\xi \in (x_{j-1}, x_{j+1})$
4. $u(x_j) - D_j^\pm u = -\frac{h^2}{24}\left(u^{(4)}(\xi_1) + u^{(4)}(\xi_2)\right)$ where $\xi_1 \in (x_{j-1}, x_j)$ and $\xi_2 \in (x_j, x_{j+1})$.

6.3 Stability Analysis

Let V_h be the set of discrete functions defined on the nodal points x_j and $V_h^0 \subset V_h$ contain the discrete functions $v_h \in V_h$ which vanish at x_0 and x_n, i.e. $v_0 = 0$ and $v_n = 0$.

Lemma 6.2 (*). Let $\mathcal{L}_h$ be the discretization of a linear differential operator which acts on $u_h \in V_h$, i.e. $\mathcal{L}_h[u_h]$. If the **discrete inner product** for both v_h and $w_h \in V_h$ is induced by the inner product, i.e. it is defined as

$$(v_h, w_h)_h = h \sum_{j=0}^{n} c_j v_j w_j \tag{87}$$

where $c_j = 1$ for $j = 1, \ldots n-1$ and $c_0 = c_n = \frac{1}{2}$ and a **norm** is defined as

$$\|v_h\|_h = \sqrt{(v_h, v_h)_h} \tag{88}$$

for a $v_h \in V_h$. Then the operator $\mathcal{L}_h$ is **symmetric**

$$(\mathcal{L}_h[v_h], w_h)_h = (v_h, \mathcal{L}_h[w_h])_h \quad \forall\, w_h,\, v_h \in V_h^0 \tag{89}$$

and **positive definite**, that is

$$(\mathcal{L}_h[v_h], v_h)_h \geq 0 \quad \forall\, v_h \in V_h^0 \tag{90}$$

and

$$(\mathcal{L}_h[v_h], v_h)_h = 0 \Longleftrightarrow v_h = 0. \tag{91}$$

Note that the that the discrete inner product is the Trapezium Rule, so

$$(w, v) = \int w(x) v(x)\, \mathrm{d}x \tag{92}$$

i.e. it approximates an integral.

Lemma 6.3 (*). For any $v_h \in V_h$

$$\|v_h\|_h \leq \frac{1}{\sqrt{2}} \left(h \sum_{j=0}^{n-1} \left(\frac{v_{j+1} - v_j}{h} \right)^2 \right)^{1/2}. \tag{93}$$

6.4 Convergence

The finite difference solution u_h can be characterised by a discrete Green's function. Define $G^k(x) \in V_h^0$ such that

$$\mathcal{L}_h \left[G^k(x) \right] = e^k(x) \tag{94}$$

where $e^k \in V_h^0$ satisfies $e^k(x_j) = \delta_{kj}$. Then

$$G^k(x_j) = hG(x_j, x_k). \tag{95}$$

Theorem 18*. Let

$$\|v_h\|_{h,\infty} = \max_{0 \leq j \leq n} |v_h(x_j)| \tag{96}$$

be the *discrete maximum norm.* Assume that $f \in C^2(0,1)$, then the nodal error, given by $e(x_j) = u(x_j) - u_h(x_j)$ satisfies:

$$\|u - u_h\|_{h,\infty} \leq \frac{h^2}{96} \|f''\|_\infty. \tag{97}$$

7 Distributions

Denote by $\mathrm{H}^s(a,b)$, for $s \geq 1$, the space of the functions $f \in C^{s-1}(a,b)$ such that $f^{(s-1)}$ is continuous and piecewise differentiable, so that $f^{(s)}$ exists unless for a finite number of points and belongs to $\mathrm{L}^2(a,b)$. The space $\mathrm{H}^s(a,b)$ is known as the Sobolev function space of order s and is endowed with the norm $\|\cdot\|_{H^s(a,b)}$ defined as

$$\|f\|_s = \left(\sum_{k=0}^{s} \left\| f^{(k)} \right\|^2_{\mathrm{L}^2(a,b)} \right)^{1/2} . \tag{98}$$

Let

$$C_0^\infty = \{\varphi \in C^\infty \,|\, \exists\, a, b \in (0,1) \quad \text{such that} \quad \varphi(x) = 0 \\ \text{for} \quad 0 \leq x < a \quad \text{or} \quad b < x \leq 1\} .$$

Then for a function $v \in \mathrm{L}^2(0,1)$ we say v' is the **weak derivative** (or **distributional derivative**) if

$$\int_0^1 v' \varphi \, \mathrm{d}x = - \int_0^1 v \varphi' \, \mathrm{d}x \quad \forall\, \varphi \in C_0^\infty\,(0,1) . \tag{99}$$

Of interest is

$$H^1(0,1) = \left\{ v \in \mathrm{L}^2(0,1) \, : \, v' \in \mathrm{L}^2(0,1) \right\} \tag{100}$$

where v' is the distributional derivative of v, and

$$H_0^1(0,1) = \left\{ v \in \mathrm{L}^2(0,1) \, : \, v' \in \mathrm{L}^2(0,1),\, v(0) = v(1) = 0 \right\} . \tag{101}$$

On H^1 there is the semi-norm:

$$|v|_{\mathrm{H}^1(0,1)} = \left(\int_0^1 \|v'(x)\|^2 \, \mathrm{d}x \right)^{1/2} = \|v'\|_{\mathrm{L}^2(0,1)} . \tag{102}$$

To see that it is a semi-norm and not a norm, consider v a constant, so $v' = 0$ thus $|v|_{\mathrm{H}^1(0,1)} = 0$ for $v \neq 0$ and thus by definition is a semi-norm, rather than a norm. Now consider the integral on functions in H_0^1, it is the case that if the integral is zero so the function is constant, but as it must be zero on the boundaries, so the function is zero and hence a norm.

8 Galerkin Method

Consider the elementary problem:

$$-\left(\alpha u'\right)' + \beta u' + \gamma u = f\left(x\right) \quad \text{on} \quad (0,1) \quad \text{with} \quad u(0) = u(1) = 0 \tag{103}$$

where α, β, $\gamma \in C^0\left(0,1\right)$ and $\alpha(x) \geq \alpha_0 > 0$ for all $x \in [0,1]$.

Next, on $\mathrm{L}^2(0,1)$, define the **scalar product**

$$(f,v) = \int_0^1 f v \,\mathrm{d}x \tag{104}$$

and a **bilinear form** $a : (\cdot,\cdot)$ which maps $H_0^1 \times H_0^1 \to \mathbb{R}$

$$a\left(u,v\right) = \int_0^1 \left(\alpha u'v' + \beta u'v + \gamma uv\right) \mathrm{d}x \tag{105}$$

and consider the **weak form** of the elementary problem:

$$\text{Find} \quad u \in H_0^1 \quad \text{such that} \quad a\left(u,v\right) = \left(f,v\right) \quad \forall\, v \in H_0^1\left(0,1\right). \tag{106}$$

Theorem 19. The following hold:

a) Let u be a C^2 be a solution of the elementary problem, then $u \in H_0^1$ also solves the weak form.

b) Let $u \in H_0^1$ be a solution of the weak problem. If and only if $u \in C^2\left(0,1\right)$ then u also solves the elementary problem.

Theorem 20 (Fundamental Theorem of the Calculus of Variations). Suppose that f is integrable on $(0,1)$ and

$$\int_0^1 \phi f \,\mathrm{d}x = 0 \quad \forall\, \phi \in C_0^\infty\left(0,1\right) \tag{107}$$

then $f = 0$.

Approximate H_0^1 by V_h. The **discrete weak problem** is then:

$$\text{Find a} \quad u_h \in V_h \quad \text{such that} \quad a\left(u_h, v_h\right) = \left(f, v_h\right) \quad \forall\, v_h \in V_h \tag{108}$$

Let $\{\varphi_1, \varphi_2, \ldots, \varphi_N\}$ be a basis of V_h, then, with $N = \mathrm{dim} V_h$, so that

$$u_h\left(x\right) = \sum_{j=1}^{N} u_j \varphi_j\left(x\right). \tag{109}$$

So the problem can be written as: Find $(u_1, \dots u_N) \in \mathbb{R}^N$ such that

$$\sum_{j=1}^{N} u_j a(\varphi_j, \varphi_i) = (f, \varphi_i) \quad i = 1, \dots, N. \tag{110}$$

Denote $a_{ij} = a(\varphi_j, \varphi_i)$ as the elements of the matrix A, let $u = (u_1, \dots, u_N)$ and $f = (f_1, \dots, f_N)$ be vectors where each entry is given by $f_i = f\varphi_i$, so that the problem is equivalent to solving the linear problem $Au = f$

Theorem 21 (Poincaré–Friedrich Inequality). Let $\Omega \subset \mathbb{R}^n$ be contained in n-dimensional cube of length s, then

$$\|v\|_{L^2(\Omega)} \leq s\,|v|_{H_0^1(\Omega)}\,. \tag{111}$$

For functions which are zero on the boundary a simplified form is

$$\int_a^b |v(x)|^2 \,\mathrm{d}x \leq C_p \int_a^b |v'(x)|^2 \,\mathrm{d}x \quad \forall\, v \in V_0 \tag{112}$$

Theorem 22*. Let

$$C = \frac{1}{\alpha_0}\left(\|\alpha\|_\infty + C_p^2 \|\gamma\|_\infty\right) \tag{113}$$

then

$$|u - u_h|_{H^1(0,1)} \leq C \min_{w_h \in V_h} |u - w_h|_{H^1(0,1)}\,. \tag{114}$$

Definition 8.1 (Coercivity and Continuity of Bilinear Forms)**.** A bilinear form $a(\cdot,\cdot)$ on V, with a norm $\|\cdot\|_V$, then a bilinear form is **coercive** if there exists an $\alpha_0 > 0$ such that

$$a(v, v) \geq \alpha_0 \|v\|_V^2 \quad \forall\, v \in V. \tag{115}$$

A bilinear form is said to be **continuous** if there exists an $M > 0$ such that

$$|a(u, v)| \leq M \|u\|_V \|v\|_V \quad \forall\, u, v \in V. \tag{116}$$

Theorem 23 (Lax–Milgram). If coercive and continuous, and the right hand side (f, v) satisfies the following inequality

$$|(f, v)| \leq K \|v\|_V \quad \forall\, v \in V. \tag{117}$$

Then the weak and discrete weak form problems admit unique solutions which satisfy

$$\|u\|_V \leq \frac{K}{\alpha_0} \quad \text{and} \quad \|u_h\|_V \leq \frac{K}{\alpha_0}. \tag{118}$$

Lemma 8.2 (Céa)**.** It is possible to show that

$$\|u - u_h\|_V \leq \frac{M}{\alpha_0} \min_{w_h \in V_h} \|u - w_h\|_V \,. \tag{119}$$

9 Finite Element Method

The finite element method (FEM) is a special technique for constructing a subspace V_h based on piecewise polynomial interpolation.

Introduce a partition $\mathcal{T}_h$ of $[0,1]$ into n subintervals $I_j = [x_j, x_{j+1}]$, $n \geq 2$, of width $h_j = x_{j+1} - x_j$, $j = 0, \ldots, n-1$, with $0 = x_0 < x_1 < \ldots < x_{n-1} < x_n = 1$ and let $h = \max_{\mathcal{T}_h} (h_j)$.

Since functions in $\mathrm{H}_0^1(0,1)$ are continuous it makes sense to consider for $k \geq 1$ the family of piecewise polynomials X_h^k introduced in (64) (where now $[a,b]$ must be replaced by $[0,1]$).

Any function $v_h \in X_h^k$ is a continuous piecewise polynomial over $[0,1]$ and its restriction over each interval $I_j \in \mathcal{T}_h$ is a polynomial of degree $\leq k$.

Considering the cases $k=1$ and $k=2$, set

$$V_h = X_h^{k,0} = \left\{ v_h \in X_h^k \,:\, v_h(0) = v_h(1) = 0 \right\}. \tag{120}$$

The dimension N of the finite element space V_h is equal to $nk-1$. In the following the two cases $k=1$ and $k=2$ will be examined.

To assess the accuracy of the Galerkin FEM, first notice that, due to Céa's lemma

$$\min_{w_h \in V_h} \|u - w_h\|_{\mathrm{H}_0^1(0,1)} \leq \left\|u - \Pi_h^k u\right\|_{\mathrm{H}_0^1(0,1)} \tag{121}$$

where $\Pi_h^k u$ is the interpolant of the exact solution $u \in V$ from the weak form of the governing equation. From inequality (121) estimating the Galerkin approximation error $\|u - u_h\|_{\mathrm{H}_0^1(0,1)}$ is then equivalent to estimating the interpolation error $\left\|u - \Pi_h^k u\right\|_{H_0^1(0,1)}$. When $k=1$, using (119) and the bounds on the interpolation errors (69)

$$\|u - u_h\|_{\mathrm{H}_0^1(0,1)} \leq \frac{M}{\alpha_0} C h \|u\|_{\mathrm{H}^2(0,1)} \tag{122}$$

provided that $u \in \mathrm{H}^2(0,1)$. This estimate can be extended to the case $k > 1$ as stated in the following convergence result.

Theorem 24. Let $u \in H_0^1(0,1)$ be the exact solution of

$$a(u,v) = f(v) \quad \forall\, v \in H_0^1(0) \tag{123}$$

and let $u_h \in V_h$ be it finite element approximation using a continuous piecewise polynomial of degree less than or equal to k, where $k \geq 1$. Furthermore, assume that $u \in \mathrm{H}^s(0,1)$ for some $s \geq 2$. Then the error is bounded as

$$\|u - u_h\|_{\mathrm{H}_0^1(0,1)} \leq \frac{M}{\alpha_0} C h^l \|u\|_{\mathrm{H}^{l+1}(0,1)} \tag{124}$$

where $l = \min(k, s-1)$. Additionally, under the same assumptions it is possible to show that

$$\|u - u_h\|_{\mathrm{L}^2(0,1)} \leq Ch^{l+1}\|u\|_{\mathrm{H}^{l+1}(0,1)}. \tag{125}$$

The error estimate shows that the Galerkin method is *convergent*, that is the approximation error tends to zero as $h \to 0$. The order of convergence is k.