Mat3110 UiO, Summary of course

H Hoel

Fall 2023

Overview

1. Iterative methods for solving $f(x) = 0$
2. Matrix factorizations
3. Norms on linear spaces
4. Numerical methods for eigenvalues
5. Polynomial interpolation
6. Approximation estimates
7. Numerical integration
8. Splines
9. Ordinary differential equations

Cou curriculum

- SM 1.1-1.4 iterative methods for scalar problems
- SM 2.1-2.7 and 2.9 solutions of linear systems of equations
- SM 3.2-3.3 efficient solution methods for matrices with structure
- SM 4.1-4.3 iterative methods for nonlinear systems of equations
- Lecture notes on numerical methods for eigenvalues and eigenvectors (excluding QR iteration)
- SM 6.2-6.5 Lagrange and Hermite interpolation
- SM 7.2-7.7 Newton-Cotes methods for numerical integration and extrapolation methods
- SM 8.2-8.5 Polynomial approximations in the infinity-norm
- SM 9.2-9.4 Polynomial approximations in the 2-norm
- SM 10.2 and 10.4-10.5 Gauss quadrature for numerical integration
- SM 11.2 and 11.4 Linear splines and natural cubic splines
- SM 12.1-12.3 and 12.5 and note on Runge–Kutta methods and A-Stability for initial value problems.
- The text The Monte Carlo method in a Nutshell by Fjordhold, Risebro and Hoel.

Fixed-point method

Solving $f(x) = 0$ for $f : \mathbb{R}^d \to \mathbb{R}^d$ can for some $g : \mathbb{R}^d \to \mathbb{R}^d$ be rephrased as fixed point problem

$$g(x) = x \quad \text{where} \quad g(\xi) = \xi \iff f(\xi) = 0.$$

Fixed-point method

$$x^{(k+1)} = g(x^{(k)}) \qquad k = 0, 1, \ldots$$

Order of convergence: Let $(x^{(k)}) \subset \mathbb{R}^d$ and suppose that

$$\|x^{(k)} - \xi\|_\infty > 0 \quad \forall k \qquad \text{and} \qquad \lim_{k\to\infty} \|x^{(k+1)} - \xi\|_\infty = 0.$$

Let $q \geq 1$ be the largest constant s.t.

$$\lim_{k\to\infty} \frac{\|x^{(k+1)} - \xi\|}{\|x^{(k)} - \xi\|^q} \leq C$$

for some $C > 0$, where we must have that $C \in (0, 1)$ if $q = 1$. Then the sequence is said to converge to ξ with order q.

Fixed point method II

Let $D \subset \mathbb{R}^d$ be closed and nonempty in slides that follow (could be $D = \mathbb{R}^d$) and fix norm $\|\cdot\|_\infty$.

Contraction mapping: $g : D \to \mathbb{R}^d$ is Lipschitz continuous if $\exists L > 0$ s.t.

$$\|g(x) - g(y)\|_\infty \leq L\|x - y\|_\infty \qquad \forall x, y \in D,$$

and if $L \in (0, 1)$, then g is called a **contraction**.

Theorem (Convergence)

Let mapping $g \in C(D, \mathbb{R}^d)$ satisfy $g(D) \subset D$ and be a contraction on D in $\infty-$norm. Then g has unique f.p. $\xi \in D$ and

$$x^{(k+1)} = g(x^{(k)}) \qquad k = 0, 1,$$

converges to ξ for any $x^{(0)} \in D$.

Proof ideas: Both exploint contraction of g:

Uniqueness: ξ, η f.p. $\implies \|\xi - \eta\|_\infty = \|g(\xi) - g(\eta)\|_\infty \leq \underbrace{L}_{<1} \|\xi - \eta\|_\infty \implies \xi = \eta.$

Existence of f.p.:

$$\begin{aligned}\|x^{(k+1)} - x^{(k)}\|_\infty &= \|g(x^{(k)} - g(x^{(k-1)})\|_\infty \\ &\leq L\|x^{(k)} - x^{(k-1)}\|_\infty \\ &\leq \ldots \leq L^k\|x^{(1)} - x^{(0)}\|_\infty\end{aligned}$$

Can use this to show that $(x^{(k)})$ is a Cauchy sequence in $(\mathbb{R}^d, \|\cdot\|_\infty)$, as for $m > n \geq 1$,

$$\begin{aligned}\|x^{(m)} - x^{(n)}\|_\infty &\leq \sum_{k=n}^{m-1} \|x^{(k+1)} - x^{(k)}\|_\infty \leq \sum_{k=n}^{m-1} L^k\|x^{(1)} - x^{(0)}\|_\infty \\ &\leq \frac{L^n}{1-L}\|x^{(1)} - x^{(0)}\|_\infty \to 0 \quad \text{as} \quad m, n \to \infty.\end{aligned} \tag{1}$$

Cauchy sequence has a limit $\xi := \lim_{k\to\infty} x_k \in D$ and limit is an f.p. as

$$\xi = \lim_{k\to\infty} x^{(k)} = \lim_{k\to\infty} x^{(k+1)} = \lim_{k\to\infty} g(x^{(k)}) \underbrace{=}_{g \text{ continuous}} g(\lim_{k\to\infty} x^{(k)}) = g(\xi).$$

How many iterations needed?

For $m = \infty$, inequality (1) tells us that

$$\|\xi - x^{(n)}\|_\infty \leq \frac{L^n}{1-L}\|x^{(1)} - x^{(0)}\|_\infty$$

Given $\epsilon > 0$, how large n is needed to ensure

$$\|\xi - x^{(n)}\|_\infty \leq \epsilon \quad ?$$

Above yields sufficient condition:

$$n = \left\lceil \frac{\ln(\|x^{(1)} - x^{(0)}\|_\infty) - \ln((1-L)\epsilon)}{\ln(1/L)} \right\rceil$$

where $\lceil x \rceil := \min\{z \in \mathbb{Z} \mid z \geq x\}$.

Jacobian of g: $J_g(x) \in \mathbb{R}^{d\times d}$ has entries defined by

$$J_g(x)_{ij} = \frac{\partial g_i}{\partial x_j}(x) \qquad 1 \le i, j \le d.$$

Theorem (Stable f.p.)

Let $g \in C(D, \mathbb{R}^d)$ with an f.p. $\xi \in D$. Assume $\exists N(\xi) \subset D$ s.t. $g \in C^1(N(\xi), \mathbb{R}^d)$ and $\|J_g(\xi)\|_\infty < 1$. Then ξ is a stable f.p. in the following sense:

$$\exists \epsilon > 0 \text{ and } \overline{B}_\epsilon \subset N(\xi) \text{ s.t. } g(\overline{B}_\epsilon(\xi)) \subset \overline{B}_\epsilon(\xi)$$

and fixed point sequence $x^{(k)} \to \xi$ as $k \to \infty$ for any $x^{(0)} \in \overline{B}_\epsilon(\xi)$.

Order of conv: Above result implies g is a local contraction mapping. When FP-method with g converges and f is a local/global contraction with Lipschitz const $L \in (0,1)$, then order of conv is at least $q = 1$, which can be deduced from

$$\frac{\|x^{(k+1)} - \xi\|_\infty}{\|x^{(k)} - \xi\|_\infty} = \frac{\|g(x^{(k)}) - g(\xi)\|_\infty}{\|x^{(k)} - \xi\|_\infty} \le L.$$

Things to know: compute iterations of fixed-point method on $\mathbb{R}^d$, how to use above theorems and how to compute how many iterations are sufficient to reach given accuracy constraint.

Newton's method for $f : \mathbb{R}^d \to \mathbb{R}^d$

$$d = 1\textbf{:} \quad x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})} \qquad k = 0, 1, \dots$$

$$d \geq 1\textbf{:} \quad x^{(k+1)} = x^{(k)} - (J_f(x^{(k)}))^{-1} f(x^{(k)}) \qquad k = 0, 1, \dots$$

This is a fixed point method with $g(x) = x - (J_f(x))^{-1} f(x)$.

Theorem

Let $\xi \in \mathbb{R}^d$ satisfy $f(\xi) = 0$, and suppose there exists an $N(\xi)$ s.t. $f \in C^2(N(\xi), \mathbb{R}^d)$ and that $J_f(\xi)$ is invertible. Then the Newton sequence converges to ξ if $x^{(0)}$ is sufficiently close to ξ, and order of convergence is at least $q = 2$.

Know to: Compute iterations with method in $\mathbb{R}^d$. Use theorem.

LU factorization

Is a factorization of $A \in \mathbb{R}^{n\times n}$ on the form

$$A = LU,$$

where $L, U \in \mathbb{R}^{n\times n}$ with L unit lower triangular (ult) and U upper triangular (ut).
If factorization exists, following must hold

$$a_{ij} = \sum_{k=1}^{\min(i,j)} \ell_{ik} u_{kj} \qquad 1 \le i, j \le n$$

Iterative formulas for rows of U and columns of L:
For m=1,. . . ,n: set $\ell_{mm} = 1$ and

$$u_{mj} = a_{mj} - \sum_{k=1}^{m-1} \ell_{mk} u_{kj} \qquad j = m, \dots, n$$

$$\ell_{im} = \frac{a_{im} - \sum_{k=1}^{m-1} \ell_{ik} u_{km}}{u_{mm}} \qquad i = m+1, \dots, n$$

LU-**factorization exists** whenever $u_{mm} \neq 0$ for all $m = 1, \dots, n-1$.

LU-factorization II

Sufficient condition: LU-factorization exists if $A^{(m)}$, the leading prinicipal submatrix of A of order m, is invertible for all $m = 1, \ldots, n-1$. (As this implies $u_{mm} \neq 0$ for all $m = 1, \ldots, n-1$. Why?)

Know how to: compute LU-factorization, know what it is used for, estimate compuational cost LU-factorization, know how to prove properties of upper and lower triangular matrices (if L and $\tilde{L}$ are ult of same size, then $L\tilde{L}$ is ult, L^{-1} is ult etc.)

PLU-factorization

Some square matrices are not LU factorizable, e.g.,

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \text{ why?}$$

But every square matrix is PLU-factorizable, where P is a permutation matrix. For example,

$$PA = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

is LU factorizable.

Know how to: How to use PLU-factorization to solve linear equations. Find P such that PA is LU-factorizable (that is, know how to PLU-factorize). Properties of permutation matrices.

p-norms on $\mathbb{R}^n$, and subordinate matrix norms

For $u \in \mathbb{R}^n$,

$$\|u\|_p := \begin{cases} \left(\sum_{i=1}^d |u_i|^p\right)^{1/p} & p \in [1, \infty) \\ \max_{i=1,\ldots,d} |u_i| & p = \infty \end{cases}$$

Know: Verify that these are norms, and that they are equivalent norms. Use Cauchy–Schwarz, Hölder's and Minkowski's inequalities. Prove Cauchy–Schwarz.

Subordinate matrix norms for $A \in \mathbb{R}^{n \times n}$:

$$\|A\|_p := \max_{v \in \mathbb{R}^n_*} \frac{\|Av\|_p}{\|v\|_p}$$

Norm is "easily" computable for some p-values:

$$\|A\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^n |a_{ij}|, \qquad \|A\|_2 = \max_{\lambda \in \sigma(A^T A)} \sqrt{\lambda}, \qquad \|A\|_\infty = \max_{i=1,\ldots,n} \sum_{j=1}^n |a_{ij}|.$$

Frequently used properties:

$$\|Av\|_p \le \|A\|_p \|v\|_p, \qquad \|AB\|_p \le \|A\|_p \|B\|_p$$

Know: Verify that these are norms. Use above properties.

Condition numbers applied to linear problems

How sensitive is solution x of $Ax = b$, for invertible A, to perturbations δb in b?
Fixing p-norm, for some $p \in [1, \infty)$, We estimate sensitivity in terms of relative condition error:

$$\sup_{\delta b \in \mathbb{R}^n_*} \frac{\|A^{-1}(b+\delta b) - A^{-1}b\|_p / \|A^{-1}b\|_p}{\|\delta b\|_p / \|b\|_p} \leq \|A^{-1}\|_p \|A\|_p$$

$\kappa_p(A) = \|A^{-1}\|_p \|A\|_p$ is called the (p-norm) condition number of matrix A.

Can show that for $A(x + \delta x) = b + \delta b$,

$$\underbrace{\frac{\|\delta x\|_p}{\|x\|_p}}_{\text{output rel. err.}} \leq \kappa_p(A) \underbrace{\frac{\|\delta b\|_p}{\|b\|_p}}_{\text{input rel. err.}} .$$

Know: How to show above inequalities. Be able to compute condition number and interpret condition number. Classify ill-conditioned problems, and use condition number to bound output error.

QR-factorization

If $A \in \mathbb{R}^{m \times n}$ with $m \geq n$, then $\exists Q \in \mathbb{R}^{m \times n}$ with $Q^T Q = I$ and an upper triangular $R \in \mathbb{R}^{n \times n}$ s.t.

$$A = QR$$

with R invertible when $\text{rank}(A) = n$.

How to obtain when $rank(A) = n$ (see notes for general):

1. Comlumn vector representation: $A = [a_1 \ a_2 \ldots a_n]$.
2. Gramm-Schmidt orthogonalization, For $k = 1, \ldots, n$:

$$c_k = a_k - \sum_{j=1}^{k-1} (a_k^T q_j) q_j, \qquad q_k = c_k / \|c_k\|_2$$

3. Set $Q = [q_1 \ q_2 \ldots q_n]$ and

$$R = Q^T A \qquad \text{(verify that it will be upper triangular and invertible)}$$

Know how to: Compute factorization, use for solving least squares problems $Ax = b$, and argue why QR-factorization is useful for least squares problems.

Positive definite matrices and Cholesky factorization

A matrix $A \in \mathbb{R}^{n \times n}_{sym}$ is called **positive definite** if

$$x^T A x > 0 \qquad \forall x \in \mathbb{R}^n_*.$$

- A is pos. def. iff all eigenvalues of A are strictly positive,
- and if A is pos. def. then
  - $\det(A) > 0$
  - $\det(A^{(m)}) > 0$ for all $m = 1, \ldots, n$
  - and one can find orthogonal eigenbasis for A ...

Know: verify properties of positive definite matrices.

Given $A \in \mathbb{R}^{n \times n}_{sym}$, a factorization of the form $A = LL^T$ where $L \in \mathbb{R}^{n \times n}$ is a lower triangular matrix is called a **Cholesky factorization** of A.

Sufficient condition: If A is positive definite, then there exists a Cholesky factorization for A.

Cholesky factorization

How to compute L in $A = LL^T$?: Similar constructive reasoning as for LU factorization.

Cholesky plays similar role as LU-factorization, to solve $Ax = b$ in this course, but it has more applications.

Know:

- Compute Cholesky factorization and how use it to solve $Ax = b$
- Estimate computational cost of both LU and Cholesky factorization for full and banded matrices (e.g. tridiagonal).

Gershgorin's theorem

Theorem (Gershgorin's circle theorem)

For $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ with $r_i = \sum_{j\neq i} |a_{ij}|$, it holds that any $\lambda \in \sigma(A)$ belongs to some Gershgorin disc, meaning $\lambda \in D_i := \{z \in \mathbb{C} \mid |z - a_{ii}| \leq r_i\}$ for at least one $i = 1, 2, \ldots, n$.

Extension of Gershgorin's thm: If the Gershgorin discs of a matrix $A \in \mathbb{R}^{n\times n}$ for some ordering satisfies that $B_1 = \cup_{i=1}^{k} D_i$ is disjoint from $B_2 = \cup_{i=k+1}^{n} D_i$ (meaning $B_1 \cap B_2 = \emptyset$), then k eigenvalues belong to B_1 and $n - k$ eigenvalues belong to B_2.

And if all discs are disjoint, then each disc contains one and only one eigenvalue.

Know: How to prove the above theorem, and how to use it and the extension to estimate spectrum of A, also in combination with similarity transformations $B = T^{-1}AT$.

Iteration methods (here presented without normalization of iter. vectors)

Power iteration (approx largest eigval):

$$x^{(k)} = Ax^{(k-1)}, \qquad \lambda^{(k)} = \frac{(x^{(k)})^T A x^{(k)}}{\|x^{(k)}\|_2^2}, \quad k = 1, 2, \ldots$$

Inverse iteration (approx smallest eigval):

$$x^{(k)} = A^{-1}x^{(k-1)}, \qquad \lambda^{(k)} = \frac{(x^{(k)})^T A x^{(k)}}{\|x^{(k)}\|_2^2}, \quad k = 1, 2, \ldots$$

Inverse iteration with shift

$$x^{(k)} = (A - \mu I)^{-1}x^{(k-1)}, \qquad \lambda^{(k)} = \frac{(x^{(k)})^T A x^{(k)}}{\|x^{(k)}\|_2^2}, \quad k = 1, 2, \ldots$$

Know: How to compute iterations in practice, what $\lambda^{(k)}$ converge towards in each case, what assumptions are sufficient to ensure convergence?

Theorem (Bauer–Fike)

For a diagonalizable matrix $A = T\Lambda T^{-1} \in \mathbb{R}^{n\times n}$ with $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$, and given a perturbation $\Delta A \in \mathbb{R}^{n\times n}$, then it holds for the any eigenvalue in the perturbed spectrum $\mu \in \sigma(A + \Delta A)$ that

$$\min_{\lambda\in\sigma(A)} |\mu - \lambda| \leq \underbrace{\|T\|_2 \|T^{-1}\|_2}_{=:\kappa_2(T)} \|\Delta A\|_2, \tag{2}$$

Know: How to use in practical computations.

Lagrange interpolation

Given interpolation points $\{(x_k, f(x_k))\}_{k=0}^n$, there exists a unique $p_n \in \mathcal{P}_n$ s.t.

$$p_n(x_k) = f(x_k) \qquad k = 0, \dots, n$$

and it is given by

$$p_n(x) = \sum_{k=0}^{n} L_k(x) f(x_k) \qquad \text{where} \quad L_K(x) := \begin{cases} \prod_{i=0, i\neq k}^{n} \frac{x - x_i}{x_k - x_i} & n \geq 1 \\ 1 & n = 1 \end{cases}$$

Approximation error: If $f \in C^{n+1}[a, b]$ and all $\{x_i\}_{i=0}^n \subset [a, b]$, then for all $x \in [a, b]$,

$$|f(x) - p_n(x)| \leq \frac{M_{n+1}}{(n+1)!} |\pi_{n+1}(x)|$$

where $M_{n+1} = \max_{y \in [a,b]} |f^{(n+1)}(y)|$ and $\pi_{n+1}(x) = \prod_{i=0}^n (x - x_i)$. And

$$|f'(x) - p_n'(x)| \leq \frac{M_{n+1}}{n!} \prod_{i=1}^{n} |x - \eta_i|$$

for some $\{\eta_i\}_{i=1}^n \subset (a, b)$ (that are independent of x).

Interpolation II

Know about Lagrange interp: Solve interpolation problems, prove uniqueness, bound approximation error of $p_n \approx f$ and $p_n' \approx f'$, sufficient conditions for uniform convergence $\|p_n - f\|_\infty$ when $n \to \infty$, and Runge's phenomenon.

Hermite interpolation: Given interpolation points $\{(x_k, f(x_k), f'(x_k))\}_{k=0}^n$, there exists a unique $p_{2n+1} \in \mathcal{P}_{2n+1}$ s.t.

$$p_{2n+1}(x_k) = f(x_k) \quad \text{and} \quad p_{2n+1}'(x_k) = f'(x_k) \qquad k = 0, \ldots, n$$

and it is given by

$$p_{2n+1}(x) = \sum_{k=0}^{n} H_k(x) f(x_k) + K_k(x) f'(x_k)$$

$$H_K(x) := (L_k(x))^2 (1 - 2L'(x_k)(x - x_k)), \qquad K_k(x) = (L_k(x))^2 (x - x_k)$$

Approx error: If $f \in C^{2n+2}[a,b]$ and all $\{x_i\}_{i=0}^n \subset [a,b]$, then for all $x \in [a,b]$,

$$|f(x) - p_{2n+1}(x)| \leq \frac{M_{2n+2}}{(2n+2)!} |\pi_{n+1}(x)|^2$$

Know: Compute Hermite interpolant, bound approx error.

Best approximation in ∞-norm

For $f \in C[a,b]$, we consider the ∞-norm

$$\|f\|_\infty = \max_{x \in [a,b]} |f(x)|,$$

and given $f \in C[a,b]$, we seek the minmax polynomial (best approximation in ∞-norm) of degree $\leq n$, meaning $p_n \in \mathcal{P}_n$ s.t.

$$\|f - p_n\|_\infty = \min_{q \in \mathcal{P}_n} \|f - q\|_\infty.$$

Result 1: For any n and $f \in C[a,b]$, there exists a unique minmax polynomial p_n.

Result 2: Weierstrass approx theorem implies that $\lim_{n\to\infty} \|p_n - f\| = 0$.

Question: How can one determine p_n in practice?
This is easy for $n = 0$, but not easy in general. We explore some features relating to minmax more generally.

Chebyshev polynomials

Oscillation thm If $f \in C[a,b]$ then $p_n \in \mathcal{P}_n$ minmax to f iff there exists $n+2$ critical points $x_0 < x_1 \ldots < x_n + 1$ in $[a,b]$ s.t.

$$|f(x_i) - p_n(x_i)| = \|f - p_n\|_\infty \qquad i = 0, 1, \ldots, n+1$$

and

$$f(x_i) - p_n(x_i) = -(f(x_{i+1}) - p_n(x_{i+1})) \qquad i = 0, \ldots, n$$

Chebyshev polynomials Are defined by $T_n(t) := \cos(n \cos^{-1}(t)) \in \mathcal{P}_n$ for $n = 0, 1, \ldots$ with exact degree of T_n equal to n.

Key property: $\|T_{n+1}\|_\infty = 1$ attained at points $y_k = \cos(k\pi/(n+1))$ $k = 0, \ldots, n+1$ with $T_{n+1}(y_k) = (-1)^k$.

Partial result minmax: For $[a,b] = [-1,1]$, $f(t) = t^{n+1}$ has minmax polynomial of degree $\leq n$ given by

$$p_n(t) = f(t) - 2^{-n} T_{n+1} \qquad \text{and} \quad \|p_n - f\|_\infty = 2^{-n}.$$

(as $f(t) - p_n(t) = 2^{-n} T_{n+1}(t)$ and RHS is a function satisfying oscillation thm conditions at points $\{y_k\}$).

Implication: For any $f \in \mathcal{P}_{n+1}$ on $[-1,1]$, we can find minmax of degree n.

Chebyshev interpolation points

T_{n+1} has zeros $t_i = \cos((i+1/2)\pi/(n+1))$ for $i = 0, \dots, n$. Can show that using $\{t_i\}_{i=0}^n \in [-1, 1]$ as interpolation points in Lagrange are ideal in the sense that they are the points minimizing magnitude of

$$\max_{t \in [-1,1]} \prod_{i=0}^{n} |t - t_i| = \max_{t \in [-1,1]} |\pi_{n+1}(t)| = 2^{-n}.$$

Moreover, if $f \in C^1[-1, 1]$, then Lagrange interpolation of f at Chebyshev interpolation points is very robust, satisfying that

$$\lim_{n \to \infty} \|p_n - f\|_\infty = 0.$$

Know: Oscillation theorem, define minmax polynomial of degree $\leq n$, compute minmax polynomial, and estimate error in special cases using Chebyshev polynomials. Describe Chebyshev interpolation points and benefits of using these points in Lagrange interpolation.

Best approximation in weighted 2-norm

Given a weight function $w \in C(a, b)$ that that is positive $w(x) > 0$ forall $x \in (a, b)$, and integrable $\int_a^b w(x)dx < \infty$, we introduced the space

$$L^2_w(a, b) := \{(\text{measurable})\ f : (a, b) \to \mathbb{R} \mid \int_a^b |f(x)|^2 w(x)dx < \infty\}.$$

Associated to this space we have the innner product

$$\langle f, g\rangle = \int_a^b f(x)g(x)w(x)dx \qquad \forall f, g \in L^2_w(a, b)$$

and weighted 2-norm $\|f\|_2 := \sqrt{\langle f, f\rangle}$.

Objective: Given $f \in L^2_w(a, b)$, find $p_n \mathcal{P}_n$ s.t.

$$\|f - p_n\|_2 = \inf_{q \in \mathcal{P}_n} \|f - q\|_2.$$

Such a p_n is called best approx to f in 2-norm of degree $\leq n$.

Approach:

1. Find polynomial orthonormal system $\{\phi_i\}_{i=0}^n$ for $\mathcal{P}_n$ with degree$(\phi_i) = i$ using Gram–Schmidt.
2. Compute best approximation

$$p_n = \sum_{i=0}^{n} \langle f, \phi_i \rangle \phi_i$$

Orthogonality result: For any $f \in L^2_w(a, b)$ and $n \geq 0$, the best approximation p_n is unique and $\langle f - p_n, q \rangle = 0$ for all $q \in \mathcal{P}_n$.

Error estimation:

$$\|f - p_n\|_2^2 = \|f\|_2^2 - \sum_{i=0}^{n} |\langle f, \phi_i \rangle|^2.$$

NB! Orthonormal system depends on interval (a, b) and $w(x)$.

Know: How to compute orthonormal system and best approx in 2-norm p_n of degree $\leq n$ given f, (a, b), and $w(x)$. Prove that p_n exists, is unique and above orthogonallity result.

Newton–Cotes rules

Is interpolation-based numerical integration:

$$I := \int_a^b f(x)dx \approx \int_a^b p_n(a)dx$$

where $p_n \in \mathcal{P}_n$ is polynomial satisfying

$$p_n(x_i) = f(x_i) \qquad i = 0, 1, \ldots, n$$

with $x_i = a + ih$ where $h = (b-a)/n$. By Lagrange interpolation

$$\int_a^b p_n(x)dx = \sum_{k=0}^{n} \underbrace{\int_a^b L_k(x)dx}_{=:w_k} f(x_k)$$

Hence $\int_a^b f(x)dx \approx \sum_{k=0}^n w_k f(x_k)$. Examples

$$n = 1: \frac{f(a) + f(b)}{2}(b-a), \qquad n = 2: \frac{f(a) + 4f((a+b)/2) + f(b)}{6}(b-a)$$

Approximation error:

$$|E_n(f)| = \left| I - \int_a^b p_n(x)dx \right| \leq \frac{M_{n+1}}{(n+1)!} \int_a^b |\pi_{n+1}(x)|dx$$

yields

$|E_1(f)| \leq \frac{M_2}{12}(b-a)^3,$ and for $n=2$ (improved to) $|E_2(f)| \leq \frac{M_4}{2880}(b-a)^5$

Composite Trapezoidal rule: For $m \geq 1$ let now $h = (b-a)/m$, $x_i = a + ih$ for $i = 0, 1 \ldots, m$ and set

$$T(m) = h\left(\frac{f(x_0) + f(x_m)}{2} + \sum_{i=1}^{m-1} f(x_i)\right)$$

Error: $|I - T(m)| \leq \frac{M_2(b-a)}{12}h^2$ ...and higher order under periodicity condition.

Composite Simpson rule

For $m \geq 1$ let now $h = (b-a)/2m$, $x_i = a + ih$ for $i = 0, 1 \ldots, 2m$ and set

$$S(m) = \frac{h}{3} \sum_{i=1}^{m} \Big(f(x_{2i-2}) + 4f(x_{2i-1}) + f(x_{2i}) \Big)$$

Error: $$|I - S(m)| \leq \frac{M_4 (b-a)^5}{2880 m^4} = \mathcal{O}(h^4).$$

Know: Construction of Newton–Cotes rules, error estimates and application of Trapezoidal and Simpson's rules. Same also for composite Trapezoidal and Simpson's rules.

Extrapolation methods for Newton–Cotes

Extrapolation methods: One can show that when f is sufficiently smooth,

$$I - T(m) = c_1 h^2 + c_2 h^4 + \mathcal{O}(h^6)$$

with $h = (b-a)/m$ and constants independent of $h > 0$.

Improve rate by **Richardson extrapolation:**

$$T_1(m) := \frac{4T(2m) - T(m)}{3}, \quad \text{yields} \quad I - T_1(m) = -\frac{c_2}{4} h^4 + \mathcal{O}(h^6)$$

Extends to Romberg integration: Set $T_0(m) := T(m)$ and

$$T_k(m) := \frac{4^k T_{k-1}(2m) - T_{k-1}(m)}{4^k - 1} \qquad k \geq 1 \quad \text{with} \quad |I - T_k(m)| = \mathcal{O}(h^{2k+2}).$$

Know: Construction and application of above extrapolation methods.

Gauss quadrature

Goal: Given weight function w and $f \in C[a,b]$, approximate

$$I := \int_a^b w(x)f(x)dx$$

Idea: Use Hermite interpolant $p_{2n+1} \in \mathcal{P}_{2n+1}$

$$p_{2n+1}(x) = \sum_{k=0}^{n} H_k(x)f(x_k) + K_k(x)f'(x_k) \approx f(x)$$

and choose interpolation points $\{x_i\}_{i=0}^n$ in smart way to obtain that

$$I \approx \int_a^b w(x)p_{2n+1}(x)dx = \sum_{k=0}^{n} \underbrace{\int_a^b w(x)(L_k(x))^2 dx}_{W_k} f(x_k)$$

Benefit: then only need to compute $n+1$ weights and function evaluations instead of expected $2n+2$.

Given $n \geq 0$:

1. Compute polynomial orthogonal basis $\phi_0, \ldots, \phi_{n+1}$ to $\mathcal{P}_{n+1}$ st $\deg(\phi_i) = i$. Let $\{x_k\}_{k=0}^{n}$ be zeros of ϕ_{n+1} (these are all distinct and in (a, b) by SM Thm 9.4).
2. Set, as before, $L_k = \prod_{i \neq k}(x - x_i)/(x_k - x_i)$ compute weights W_k and obtain Gauss rule using $n+1$ quad points by

$$G_n(a, b) := \sum_{k=0}^{n} W_k f(x_k)$$

Error: If $w \in C(a, b)$ is positive and integrable and $f \in C^{2n+2}[a, b]$ for some $n \geq 0$, then

$$|I - G_n(a, b)| \leq \int_a^b w(x)|f(x) - p_{2n+1}(x)|dx \leq \frac{M_{2n+2}}{(2n+2)!} \int_a^b w(x)(\pi_{n+1}(x))^2 dx$$

Composite Gauss rules for setting with $w \equiv 1$

1. Divide $[a, b]$ into m subintervals $[x_{i-1}, x_i]$ with $x_i = a + ih$, $i = 0, 1, \ldots, m-1$ and $h = (b-a)/m$.
2. Set

$$I = \sum_{i=1}^{m} \int_{x_{i-1}}^{x_i} f(x)dx \approx \sum_{i=1}^{m} G_n(x_{i-1}, x_i) =: G_{m,n}$$

$G_{m,n}$ uses m subintervals with $n+1$ quadrature points over each subinterval.

Example: Composite midpoint rule with $m \geq 1$,

$$G_{m,0} = \sum_{i=1}^{m} G_0(x_{i-1}, x_i) = h \sum_{i=1}^{n} f((x_{i-1} + x_i)/2).$$

Error estimate: $f \in C^{2n+2}[a, b] \implies |I - G_{m,n}| \leq \dfrac{M_{2n+2}(b-a)}{(2n+2)!2^{2n+2}} h^{2n+2} = \mathcal{O}(h^{2n+2})$

Comparison: At same computational budget, Newton–Cotes rule achieves $\mathcal{O}(h^{n+1})$ approx error.

Know: compute/construct $G_n(a, b)$ given w and (a, b) and how to estimate error $|I - G_n(a, b)|$. In setting $w \equiv 1$, extension to composite Gauss rule and computing $G_{m,n}$.

Monte Carlo integration

For square integrable $f : [0,1]^d \to \mathbb{R}^d$, we approximate

$$I(f) := \int_{[0,1]^d} f(x)dx$$

by Monte Carlo estimator

$$I_M(f) = \frac{1}{M}\sum_{m=1}^{M} f(X_m)$$

where $X_1, \ldots, X_M \sim U([0,1]^d)$ are mutually independent.

By the independence and identical distribution of X_i and the linearity of the expectation operator, we obtain the root-mean square error (RMSE)

$$\mathcal{E}_M := \sqrt{\mathbb{E}[(I_M(f) - \mathbb{E}[f(X)])^2]} = \frac{\sqrt{\mathrm{Var}[f(X)]}}{\sqrt{M}}$$

where $X \sim U([0,1]^d)$. Can further show that

$$\mathrm{Var}[f(X)] = \mathbb{E}[(f(X) - \mathbb{E}[f(X)])^2] \leq \frac{(\sup_{x\in[0,1]^d} f(x) - \inf_{x\in[0,1]^d} f(x))^2}{4}$$

This yields RMSE

$$\mathcal{E}_M = \frac{\sqrt{\mathrm{Var}[f(X)]}}{\sqrt{M}} \leq \frac{\sup_{x\in[0,1]^d} f(x) - \inf_{x\in[0,1]^d} f(x)}{2\sqrt{M}} = \mathcal{O}(M^{-1/2})$$

(last inequality useful when it's difficult to estimate $\mathrm{Var}[f(X)]$).

Alternative error bound: By Chebyshev inequalities we obtain for any $\epsilon > 0$ that

$$\mathbb{P}(|I_M(f) - I(f)| \geq \epsilon) \leq \frac{\mathrm{Var}[f(X)]}{\epsilon^2 M} \leq \frac{(\sup_{x\in[0,1]^d} f(x) - \inf_{x\in[0,1]^d} f(x))^2}{4\epsilon^2 M} = \mathcal{O}(M^{-1})$$

Convergence in probabilty: If $\mathrm{Var}[f(X)] < \infty$, then for any $\epsilon > 0$,

$$\lim_{M\to\infty} \mathbb{P}(|I_M(f) - I(f)| \geq \epsilon) = 0.$$

and also possible to show stronger result: $\mathbb{P}$-**almost sure** convergence

$$\mathbb{P}\Big(\lim_{M\to\infty} I_M(f) = I(f)\Big) = 0.$$

Error control through number of samples

Given $\epsilon > 0$, can ask how large M is needed to ensure that $\mathcal{E}_M \leq \epsilon$?

Answer: By previous slide, need M so large that

$$\frac{\mathrm{Var}[f(X)]}{M} \leq \epsilon^2 \implies M = \left\lceil \frac{\mathrm{Var}[f(X)]}{\epsilon^2} \right\rceil,$$

or alternatively (if $\mathrm{Var}[f(X)]$ is not computable),

$$\frac{(\sup_{x\in[0,1]^d} f(x) - \inf_{x\in[0,1]^d} f(x))^2}{4M} \leq \epsilon^2 \implies M = \left\lceil \frac{(\sup_{x\in[0,1]^d} f(x) - \inf_{x\in[0,1]^d} f(x))^2}{4\epsilon^2} \right\rceil$$

But can also ask, given $\epsilon > 0$ and $\delta \in (0,1)$, how large M is needed to ensure

$$\mathbb{P}(|I_M(f) - I(f)| \geq \epsilon) \leq \delta?$$

and, by previous slide, determine M by either

$$\frac{\mathrm{Var}[f(X)]}{\epsilon^2 M} \leq \delta, \quad \text{or} \quad \frac{(\sup_{x\in[0,1]^d} f(x) - \inf_{x\in[0,1]^d} f(x))^2}{4\epsilon^2 M} \leq \delta.$$

Monte Carlo integration

- Monte Carlo is said to overcome curse of dimensionality in the sense that its order of convergence for $I_M(f) \to I(f)$ does not depend on state-space dimension d and they do not depend on regularity of f as long as

$$\int_{[0,1]^d} |f(x)|^2 dx < \infty.$$

- This is different from classic quadrature methods, like Newton–Cotes or Gauss, as they depend both on d and the regularity of f.
- Monte Carlo is often more efficient and flexible than classic quadrature methods for numerical integration in high dimensions d.

Know: implement Monte Carlo integration for a given square integrable $f : [0,1]^d \to \mathbb{R}$, estimate number of samples needed to reach error bound, and know when method is useful.

Splines I

Piecewise polynomial approximation of $f : [a, b] \to \mathbb{R}$ over subintervals $[x_{i-1}, x_i]$ with the set of knots

$$a = x_0 < x_1 < \ldots < x_m = b$$

(Piecewise) linear spline interpolation: $s_L : [a, b] \to \mathbb{R}$ is piecewise linear function $s_L|_{[x_{i-1}, x_i]} \in \mathcal{P}_1$ over each interval, so two unkown coefficients per interval. Spline has $2m$ equal-to-f-at-knots constraints:

$$s_L(x_i-) = f(x_i) \quad \text{and} \quad s_L(x_i+) = f(x_i) \quad i = 1, \ldots, m-1 \qquad \text{and} \quad s_L(a) = f(a), \quad s$$

where $s_L(x-) := \lim_{\delta \downarrow 0} s_L(x + \delta)$ and $s_L(x+) := \lim_{\delta \downarrow 0} s_L(x + \delta)$.
Solution: For each interval and $x \in [x_{i-1}, x_i]$,

$$s_L(x) := \frac{x_i - x}{x_i - x_{i-1}} f(x_{i-1}) + \frac{x - x_{i-1}}{x_i - x_{i-1}} f(x_i)$$

Error bound: If $f \in C^2[a, b]$, then (by error estimates for Lagrange interpolation)

$$\max_{x \in [a,b]} |S_L(x) - f(x)| \leq \frac{\max_{x \in [a,b]} |f''(x)|}{8} h^2$$

where $h = \max_{i=1,\ldots,m} |x_i - x_{i-1}|$.

Natural cubic spline interpolation

Is function $s_2 : [a, b] \to \mathbb{R}$ that is piecewise cubic $s_2|_{[x_{i-1}, x_i]} \in \mathcal{P}_3$, so four unknown coefficients per interval.

Spline has $2m$ equal-to-f-at-knots constraints:

$$s_2(x_i-) = f(x_i), \quad s_2(x_i+) = f(x_i) \quad i = 1, \dots, m-1 \quad \text{and} \quad s_2(a) = f(a), \quad s_2(b) = f(b),$$

$2m - 2$ smoothig-conditions-at-knots constraints:

$$s_2'(x_i-) = s_2'(x_i+) \qquad s_2''(x_i-) = s_2''(x_i+) \qquad m = 1, \dots, m-1$$

and boundary constraints $s_2''(a) = 0$ and $s_2''(b) = 0$.

This yields $4m$ constraints for $4m$ unknowns and can be solved by writing $\sigma_i = s_2''(x_i)$ and integrating twice

$$s_2''(x) = \frac{x_i - x}{x_i - x_{i-1}} \sigma_{i-1} + \frac{x - x_{i-1}}{x_i - x_{i-1}} \sigma_i \qquad x \in [x_{i-1}, x_i].$$

Know: Given $f : [a, b] \to \mathbb{R}$ compute linear and obtain system of equations for determining $\sigma_0, \dots, \sigma_m$ for natural cubic splines.

Existence and uniqueness

Theorem (Existence and uniqueness)

Consider the IVP

$$y' = f(t, y) \qquad t \in [a, b], \quad y(a) = y_0 \in \mathbb{R}^d \tag{3}$$

with $f \in C([a, b] \times \mathbb{R}^d, \mathbb{R}^d)$ *Lipschitz in* y*. Then there exists a unique solution to* (3) *with* $y \in C^1([a, b], \mathbb{R}^d)$*.*

Theorem (Convergence of one-step method)

Consider the IVP (3) *with* f *Lipschitz in* y*. Let* $y_{n+1} = y_n + h\Phi(t_n, y_n; h)$ *with* $h = (b - a)/N$ *and* $t_n = a + nh$*, be an explicit one-step method with order of accuracy* $p \geq 1$ *(for particular IVP). Then it holds that*

$$\max_{n=0,1,\ldots,N} \|y_n - y(t_n)\| = \mathcal{O}(h^p).$$

Know: Application above theorems. Compute truncation error, consistency, global error, order of accuracy for given explicit or implicit Runge–Kutta one-step method applied to a given/particular IVP.

Runge–Kutta methods and A-Stability

- Know how to translate to translate Butcher tableau (b, c, A) into one-step method and oppositely, given one-step method (for up to $s = 2$ stages) into Butcher tableau.
- For explicit RK methods, know sufficient conditions on (b, c, A) to obtain consistency, and order of accuracy at least $p = 1$ and $p = 2$.
- For given RK method, be able to compute stability function $R(z)$, region of absolute stability and determine if method is A-stable or not.
- Be able to compute one or two solution iterations of RK-methods for higher-dimensional problems.
- Understand strengths and weaknesses of explicit and implicit RK methods (Key features: stiff problems, stability and computational cost of solution iterations.)