Numerik WS 2021/22

Ida Hönigmann

April 8, 2024

Contents

1 Grundbegriffe

1.1 Gegenstand der Numerischen Mathematik

Von der Realität bis zur Interpretation einer Simulation ist es ein langer Weg.

- **Mathematisches Modell** versucht mit Hilfe von mathematischen Formeln (idR. Differentialgl.) die Realität zu beschreiben.
- Die wenigsten Lösungen dieser mathematischen Modelle kann man exakt berechnen, d.h. man approximiert die exakte Lösung mittels **numerischer Simulation** am Rechner.
- Diese **numerische Lösung** wird dann interpretiert und man hofft, dass diese Interpretation die Realität beschreibt.

Jede numerische Simulation zerfällt in kleinere **numerische Probleme**, die geeignet zu lösen sind. Die elementarsten numerischen Probleme sind Gegenstand dieser Vorlesung.

Beispiel 1.
- *Wie approximiert man komplizierte Funktionen mittels einfacher Funktionen (z.B. stückweise Polynome)?*
- *Wie berechnet man Grenzwerte (z.B. Integral, Differential)?*
- *Wie löst man lineare / nichtlineare Gleichungen?*

Jede numerische Simulation ist fehlerbehaftet.

- **Modellfehler:** Das mathematische Modell vereinfacht die Realität.
- **Datenfehler:** Die Eingangsdaten einer Simulation stammen meistens aus physikalischen Messungen und haben daher eine gewisse Mess(un-)genauigkeit.
- **Rundungsfehler:** Auf Rechnern ersetzt die endliche Menge an Gleitkommazahlen das kontinuierliche $\mathbb{R}$, d.h. sowohl die Daten als auch die Rechnungen sind rundungsfehlerbehaftet.
- **Verfahrensfehler:** Viele Probleme werden mathematisch in unendlich dimensionalen Räumen oder mit Limiten formuliert. Beides steht im Rechner nicht zur Verfügung und muss diskretisiert werden.

In der Vorlesung liegt unser Hauptaugenmerk auf dem Verfahrensfehler und dem Aufwand zugehöriger Algorithmen.

1.2 Kondition und Stabilität

Betrachte ein abstraktes Problem. Werte $\Phi : X \to Y$ bei $x \in X$ aus, wobei X, Y geeignete normierte Räume sind. Die **Kondition eines Problems** besagt, wie stark Änderungen in x (z.B. Rundungsfehler) sich auf $\Phi(x)$ auswirken.

Definition 1. *Das Problem ist* ***schlecht konditioniert bzgl. absolutem Fehler****, wenn es eine kleine Störung $\tilde{x}$ von x gibt mit $||\Phi(x) - \Phi(\tilde{x})|| \gg ||x - \tilde{x}||$.*

Das Problem ist ***schlecht konditioniert bzgl. relativem Fehler****, falls $x \neq 0 \neq \Phi(x)$ und es ex. eine kleine Störung $\tilde{x}$ von x gibt mit $\frac{||\Phi(x)-\Phi(\tilde{x})||}{||\Phi(x)||} \gg \frac{||x-\tilde{x}||}{||x||}$.*

Andernfalls bezeichnet man das Problem als ***gut konditioniert (bzgl. abs./rel. Fehler)****.*

Bemerkung 1. *Ist Φ stetig differenzierbar, d.h. $\Phi(x) - \Phi(\tilde{x}) = D\Phi(x)(x - \tilde{x}) + o(||x - \tilde{x}||)$ für $\tilde{x} \to x$ so beschreibt die Ableitung $D\Phi(x) \in L(X, Y)$ wie stark sich Änderungen in x auf den Fehler auswirken.*

Deshalb bezeichnet man $\kappa_{abs}(x) = ||D\Phi(x)||$, $\kappa_{rel}(x) = \frac{||D\Phi(x)||\cdot||x||}{||\Phi(x)||}$ als ***Konditionszahlen (bzgl. abs./rel. Fehler)****, d.h. man ist gut konditioniert für $\kappa_{abs}, \kappa_{rel}$ vergleichsweise klein.*

Definition 2. *Es sei $\tilde{\Phi}$ eine algorithmische Umsetzung von Φ. Der Algorithmus $\tilde{\Phi}$ ist* ***instabil****, wenn es eine kleine Störung $\tilde{x}$ von x gibt, sodass* $\underbrace{||\Phi(x) - \tilde{\Phi}(\tilde{x})||}_{\textit{tatsätlicher Fehler im Rechner}} \gg \underbrace{||\Phi(x) - \Phi(\tilde{x})||}_{\textit{unvermeidlicher Fehler}}$.

Andernfalls ist der Algorithmus stabil.

Bemerkung 2. *Mir ist bewusst, dass die Symbolik $\gg$ ("wesentlich größer") ungenauer ist, als Sie es aus anderen Vorlesungen kennen. Aber "schlecht konditioniert" und "instabil" hängt halt an der Genauigkeit der Daten und den Erfordernissen des Nutzers!*

In der Vorlesung geht es primär um die Asymptotik, d.h. was könnte im Worst-Case passieren.

Erinnerung/Warnung: Die Arithmetik im Rechner erfüllt weder Assoziativität noch Distributivgesetz, d.h. die Reihenfolge (und Formulierung) der Rechenoperatoren spielt eine Rolle für Stabilität.

Beispiel 2 (schlechte Kondition bei Auslöschung)**.** *Als **Auslöschung** bezeichnet man das Phänomen, dass bei Subtraktion zweier annähernd gleicher Zahlen im Rechner die hinteren Ziffern (welche rundungsbehaftet sind) signifikant werden. Der relative Fehler kann sogar beliebig groß werden, d.h.* $\Phi : \mathbb{R}^2 \to \mathbb{R}, (x,y) \mapsto x - y$ *hat* $\kappa_{rel}(x,y) = \frac{\sqrt{2}||(x,y)||_2}{|x-y|} \gg 0$ *für* $x \approx y$*.*

Achtung: Oft ist ein Problem gut konditioniert, wird aber in Teilprobleme zerlegt (im Algorithmus) sodass der resultierende Algorithmus instabil wird.

Beispiel 3. *Werte* $\Phi(x) = \frac{1}{x+1} - \frac{1}{x}$ *für* $x \gg 0$ *aus.*

$$\begin{aligned} \Phi'(x) = -\frac{1}{(x+1)^2} + \frac{1}{x^2} &= \frac{(x+1)^2 - x^2}{x^2(x+1)^2} = \frac{2x+1}{x^2(x+1)^2} \\ \Phi(x) &= \frac{x-(x+1)}{x(x+1)} = \frac{-1}{x(x+1)} \\ \kappa_{rel}(x) = \frac{|\Phi'(x)| \cdot |x|}{|\Phi(x)|} &= \frac{(2x+1)}{x^2(x+1)^2} x^2(x+1) = 1 + \frac{x}{x+1} \leq 2 \end{aligned}$$

$\implies$ *gut konditioniert!*

Beispiel 4. *Es sei* $||.||$ *eine Norm auf* $\mathbb{K}^n$ *und wir verwenden dieselbe Notation für die induzierte Operatornorm* $||A|| := \sum_{x \in \mathbb{K}^n, x \neq 0} \frac{||Ax||}{||x||}$ *für* $A \in \mathbb{K}^{n \times n}$*.*

Ist A *invertierbar, so bezeichnet* $cond(A) := ||A|| \cdot ||A^{-1}||$ *die **Konditionszahl von** A **(bzgl.** $||.||$**)**. Betrachtet man das Lösungsproblem* $\Phi : \mathbb{K}^n \to \mathbb{K}^n, \Phi(b) = A^{-1}b$*, so gilt für die relative Konditionszahl (mit* $x = A^{-1}b$*)*

$$\kappa_{rel}(b) = \frac{||D\Phi(b)|| \cdot ||b||}{||\Phi(b)||} = \frac{||A^{-1}|| \cdot ||Ax||}{||x||} \leq cond(A),$$

*wobei die letzte Abschätzung **scharf ist**, d.h. es gilt Gleichheit für mindestens ein* b *(und ein* x*).*

1.3 Verfahrensfehler

Im Wesentlichen gibt es zwei Arten von Verfahrensfehlern

- **Abbruchfehler**, wenn ein konvergenter (aber unendlicher) Algorithmus nach endlich vielen Schritten abgebrochen wird.
- **Diskretisierungsfehler**, wenn eine kontinuierliche Größe durch eine diskrete vereinfacht wird, z.B. Differenzenquotienten statt Differenzialquotient.

Beispiel 5 (Abbruchfehler Heron-Verfahren)**.** *Für* $x > 0$ *def.* $y_1 := \frac{1}{2}(1+x), y_{n+1} := \frac{1}{2}(y_n + \frac{x}{y_n})$*.*

$$\begin{aligned} \implies y_{n+1}^2 - x = \frac{1}{4}(y_n^2 + 2x + \frac{x^2}{y_n}) - x = \frac{1}{4}(y_n - \frac{x}{y_n})^2 \geq 0 \\ \implies y_{n+1}^2 \geq x > 0 \text{ und } y_{n+1} - y_n = \frac{1}{2}(y_n + \frac{x}{y_n}) - y_n = \frac{x}{2y_n} - \frac{y_n}{2} = \frac{x - y_n^2}{2y_n} \leq 0 \\ \implies 0 < \sqrt{x} \leq y_{n+1} \leq y_n \implies y_n \to y \\ \implies y = \frac{1}{2}(y + \frac{x}{y}) \implies \frac{1}{2}y^2 = \frac{1}{2}y^2 + \frac{1}{2}x \implies y^2 = x \implies y = \sqrt{x} \end{aligned}$$

$\implies$ *(*y_n*) konvergiert monoton fallend gegen* $\sqrt{x}$*. Sobald* $y_n = \sqrt{x}$*, würde auch* $y_{n+1} = \sqrt{x}$ *gelten, d.h. endkonstante Folge.*

Später: Heron-Verfahren ist tatsächlich **quadratisch konvergent**, d.h. ex. $C > 0$ mit $|\sqrt{x} - y_{n+1}| \leq C|\sqrt{x} - y_n|^2$. $\implies$ schnelle konvergenz, weil sich korrekte Ziffern pro Schritt verdoppeln!

Beispiel 6 (Diskretesierungsfehler einseitiger Differenzenquotient)**.** *Sei $f : \mathbb{R} \to \mathbb{R}$ differenzierbar, $x \in \mathbb{R}$*

Ziel: Approximiere $\Phi := f'(x)$ durch den einseitigen Diffquot. $\Phi_h = \frac{f(x+h)-f(x)}{h}$

klar: $\Phi_h \to \Phi$ für $h \to 0$, aber die Konvergenz kann beliebig langsam sein. Man interessiert sich in der Numerik auch für Konvergenzraten bzgl. des Diskretisierungsparameters.

Für $f \in \mathcal{C}^2$ (lokal um x) gilt nach Mittelwertsatz

$$f'(x) - \frac{f(x+h)-f(x)}{h} = f'(x) - f'(\zeta) = f''(\xi)(x-\zeta)$$

mit Zwischenstellen $x \leq \xi \leq \zeta \leq x+h$

$$\implies |\Phi - \Phi_n| \leq ||f''||_{L^\infty(x,x+h)} h = \mathcal{O}(h)$$

d.h. hier Konvergenzrate 1 in h.

Definition 3. *Es sei Φ eine kontinuierliche Große mit Diskretisierung Φ_h für $h > 0$. Dann bezeichnet man eine Abschätzung der Form $|\Phi - \Phi_h| = \mathcal{O}(h^\alpha)$ als **a-priori Fehlerabschätzung** mit **Konvergenzrate** $\alpha > 0$ (auch **Konvergenzordung**).*

Natürlich interessiert sich die Numerik für Verfahren, bei denen $\alpha > 0$ möglich groß ist.

Beispiel 7 (zentraler Differenzenquotient)**.** *$f : \mathbb{R} \to \mathbb{R}$ diffbar, $x \in \mathbb{R}$, $\Phi := f'(x)$ und $\Phi_h := \frac{1}{2}\left(\frac{f(x+h)-f(x)}{h} + \frac{f(x)-f(x-h)}{h}\right)$*

klar: $\Phi_h \to \Phi$ für $h \to 0$, $|\Phi - \Phi_h| = \mathcal{O}(h)$ sofern $f \in \mathcal{C}^2$ (lokal um x).

Für $f \in \mathcal{C}^3$ (lokal um x) gilt mit Taylor $f(x+h) = f(x)) + hf'(x) + \frac{h^2}{2}f''(x) + \frac{h^3}{6}f'''(\zeta)$ für geeignete $x-h \leq \zeta_- \leq x \leq \zeta_+ \leq x+h$

$$\begin{aligned}
\implies \frac{f(x+h)-f(x)}{h} &= f'(x) + \frac{h}{2}f'(x) + \frac{h^2}{6}f'''(\zeta_-) \\
\frac{f(x)-f(x-h)}{h} &= f'(x) - \frac{h}{2}f'(x) + \frac{h^2}{6}f'''(\zeta_+) \\
\implies |\Phi - \Phi_h| &= \frac{h^2}{6}\frac{|f'''(\zeta_+) + f'''(\zeta_-)|}{2} = \mathcal{O}(h^2)
\end{aligned}$$

d.h. Konvergenzrate $\alpha = 2$.

$\implies$ höhere Genauigkeit für gleiches h bzw. gleiche Genauigkeit für größeres h.

Bemerkung 3. *Auslöschung tritt immer auf (insb. bei Diffquot.), aber sie wird abgemildert durch Verfahren höherer Ordnung. Eine andere Möglichkeit für Verfahren höherer Ordnung zur Approximation von $f'(x)$ ist die Verwendung von Polynomapproximation, d.h. $f \approx p$ Polynom und berechne $p'(x) \approx f'(x)$.*

2 Interpolation

Bei einem **Interpolationsproblem** sind im einfachsten Fall Paare (x_j, y_j) gegeben und eine "einfache" Funktion p mit $p(x_j) = y_j \forall j$ gesucht, z.B. Polynome, Splines (= stückweise Polynome), rationale Funktionen (= Quotienten von Polynomen). Verwandt, aber mathematisch schwieriger sind **Approximationsprobleme**. Dabei ist eine Funktion f und eine Norm $||.||$ gegeben, und es wird eine einfache Funktion p gesucht, die $||f-p||$ in dieser Klasse einfacher Fkt. minimiert. Oft ist dabei die Funktion f nur implizit gegeben, d.h. unbekannt.

2.1 Lagrange-Polynominterpolation

Problemstellung: Gegeben sind $n+1$ reelle **Stützstellen** $a \leq x_0 < ... < x_n \leq b$ und **Funktionswerte** $y_0, .., y_n \in \mathbb{K}$. Die **Lagrange-Interpolationsaufgabe** sucht ein Polynom $p \in \mathbb{P}_n = \{p(x) = \sum_{j=0}^n a_j x^j | a_0, ..., a_n \in \mathbb{K}\}$ vom Grad n mit $p(x_j) = y_j \forall j = 0, ..., n$

Lemma 1.

1. *$\mathbb{P}_n$ ist $\mathbb{K}$-Vektorraum mit $\dim \mathbb{P}_n = n+1$.*
2. *Die **Monome** $p_j(x) = x^j, j = 0, ..., n$ sind eine Basis von $\mathbb{P}_n$.*
3. *Die **Lagrange-Polynome** $L_j(x) = \prod_{k=0,k\neq j}^n \frac{x-x_k}{x_j-x_k} \in \mathbb{P}_n$ erfüllen $L_j(x_k) = \delta_{jk}$ für alle $j, k = 0, ..., n$ und bilden eine Basis von $\mathbb{P}_n$.*
4. *Die **Newton-Polynome** $q_j(x) = \prod_{k=0}^{j-1}(x-x_k) \in \mathbb{P}_j$ für $j = 0, ..., n$ bilden eine Basis von $\mathbb{P}_n$.*

Proof. klar: $\mathbb{P}_n$ ist $\mathbb{K}$-Vektorraum, $\dim \mathbb{P}_n \leq n+1$,
zz: $\{L_0, .., L_n\} \subseteq \mathbb{P}_n$ lin. unab.
Sei $\mu_0, ..., \mu_n \in \mathbb{K}$ mit $\sum_{j=0}^n \mu_j L_j(x) = 0 \forall x$
Für $x = x_k$ folgt

$$0 = \sum_{j=0}^n \mu_j \underbrace{L_j(x_k)}_{=\delta_{jk}} = \mu_k$$

$\implies$ lin. unab. laut Def. $\implies \dim \mathbb{P}_n \geq n+1 \implies$ Monome + Lagrange Pol. bilden Basis von $\mathbb{P}_n$.
zz. $\{q_0, ..., q_n\} \subseteq \mathbb{P}_n$ lin. unab.
Seien $\mu_0, ..., \mu_n \in \mathbb{K}$ mit $\sum_{j=0}^n \mu_j \underbrace{q_j(x)}_{=\prod_{k=0}^{j-1}(x-x_k)} = 0$.
Für $x = x_0$ folgt $\mu_0 q_0(x) = 0 \implies \mu_0 = 0$.
Für $x = x_1$ folgt $\mu_1 \underbrace{q_1(x)}_{\neq 0} = 0$, also $\mu_1 = 0$. Induktives Vorgehen zeigt $\mu_j = 0 \forall j$. □

Satz 1 (Eindeutigkeit + Existenz)**.** *Betrachte Lagrange-Interpolation zu Stützstellen $a \leq x_0 < ... < x_n \leq b$ und Funktionswerten $y_0, ..., y_n \in \mathbb{K}$. Dann existiert ein eindeutiges $p \in \mathbb{P}_n$ mit $p(x_j) = y_j \forall j$. Dieses wird gegeben durch $p = \sum_{j=0}^n y_j L_j$. Ist $\{q_0, ..., q_n\} \subseteq \mathbb{P}_n$ eine Basis von $\mathbb{P}_n$ und $p = \sum_{j=0}^n \lambda_j q_j$, so löst $\lambda = (\lambda_0, ..., \lambda_n)$ das lineare Gleichungssystem*

$$\underbrace{\begin{pmatrix} q_0(x_0) & ... & q_n(x_0) \\ \vdots & & \vdots \\ q_0(x_n) & ... & q_n(x_n) \end{pmatrix}}_{=:A} \lambda = \begin{pmatrix} y_0 \\ \vdots \\ y_n \end{pmatrix}$$

Die Matrix A ist regulär, d.h. λ ist die eindeutige Lösung.

Proof. Da $L_j(x_k) = \delta_{jk} \forall j, k$ ist offensichtlich, dass $p = \sum_{j=0}^n \mu_j L_j$ genau dann das Interpolationsproblem löst, wenn $\mu_j = y_j \forall j$. $\implies$ Eindeutigkeit + Existenz
Def. Lösungsoperator $\mathcal{P} : \mathbb{K}^{n-1} \to \mathbb{P}_n$ durch $(\mathcal{P}y)(x_j) = y_j \forall j = 0, ..., n \forall y \in \mathbb{K}^{n+1}$
$\implies$ wohldef, bijektiv
Def. Auswertungsoperator $\mathcal{A} : \mathbb{P}_n \to \mathbb{K}^{n+1}, p \mapsto (p(x_0), ..., p(x_n))$
$\implies$ wohldef, linear
$\mathcal{P} \circ \mathcal{A} =$ Identität, $\mathcal{A} \circ \mathcal{P} =$ Identität,
$\implies \mathcal{A} = \mathcal{P}^{-1}, \mathcal{P} = \mathcal{A}^{-1}$
$\implies A$ ist die darstellende Matrix $\mathcal{A}$. $\implies A$ ist regulär, da $\mathcal{A}$ bijektiv, linear. □

Bemerkung 4. *Die Konditionszahl cond(A) der sogenannten* ***Vandermonde-Matrix*** *A hängt stark von der Wahl der Basis ab. Für die Lagrange-Polynome wäre A die Identität. Für die Monome ist cond(A) in der Regel indiskutabel schlecht (hängt an der Wahl der x_j). Die Basiswahl beeinflusst auch die Besetzungsstruktur der Matrix.*

Beispiel 8. *Die Newton-Basis führt auf eine untere Dreiecksmatrix*

$$\begin{pmatrix} 1 & 0 & ... & 0 \\ 1 & q_1(x_1) & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & q_1(x_n) & ... & q_n(x_n) \end{pmatrix}$$

d.h. das lineare GLS kann in $\mathcal{O}(n^2)$ statt $\mathcal{O}(n^3)$ gelöst werden.

Lemma 2 (Horner-Schema)**.** *Sei $p(x) = \sum_{j=0}^n \lambda_j \underbrace{q_j(x)}_{\prod_{k=0}^{j-1}(x-x_k)}$*

Für einen Auswertungspunkt $x \in \mathbb{R}$ betrachte

- $y = \lambda_n$
- *for* $k = n-1 : -1_0$
- $y = (x - x_k)y + \lambda_k$

- *end*

$\implies$ *Der Algorithmus berechnet in $3n$ Operationen den Funktionswert $y = p(x)$.*

Satz 2 (Interpolationsfehlerdarstellung). *Sei $f \in \mathcal{C}^{n+1}[a,b], 0 \le m \le n, p \in \mathbb{P}_n$ mit $p(x_j) = f(x_j) \forall j = 0, ..., n$, wobei $a \le x_0 < ... < x_n \le b$*

$$\implies f^{(m)}(x) - p^{(m)}(x) = \frac{f^{(n+1)(\xi)}}{(n+1-m)!} \prod_{l=0}^{n-m} (x - \zeta_l),$$

wobei $\xi = \xi(m,x)$ und $\zeta_l = \zeta_l(m, x, x_0, .., x_n)$ in $[a,b]$
Für $m = 0$ gilt $\zeta_l = x_l \forall l$.

Proof. $e := f - p \in \mathcal{C}^{n+1}[a,b]$
$\implies$ e hat mindestens $n+1$ Nullstenen (bei x_l) $\implies$ e' hat mindestens n Nullstellen $\implies$ $e^{(m)}$ hat mindestens $n+1-m$ Nullstellen $a < \zeta_0 < ... < \zeta_{n-m} < b$
o.B.d.A. $x \notin \{\zeta_0, ..., \zeta_{n-m}\}$
Def. $F(y) := e^{(m)}(x)w(y) - e^{(m)}(y)w(x)$ mit $w(x) := \prod_{l=0}^{m-n}(y - \zeta_l)$
$\implies$ F hat $n+2-m$ Nullstellen $\implies$ $F^{(n+1-m)}$ hat mind. 1 Nullstelle ξ

$$0 = F^{(n+1-m)(\xi)} = \underbrace{e^{(m)}(x)}_{=f^{(m)}(x)-p^{(m)}(x)} \underbrace{w^{(n+1-m)}(\xi)}_{=(n+1-m)!} - \underbrace{e^{(n+1)}(\xi)}_{=f^{n+1}(\xi)} \underbrace{w(x)}_{=\prod_{l=0}^{n-m}(x-\zeta_l)}$$

□

Korollar 1 (Interpolationsfehler-Abschätzung). *Seien $f \in \mathcal{C}^{n+1}[a,b]$ reell- oder komplexwertig, $a \le x_0 < ... < x_n \le b, p \in \mathbb{P}_n$ mit $p(x_j) = f(x_j) \forall j = 0, ..., n, 0 \le m \le n$*

$$\implies ||f^{(m)} - p^{(m)}||_{L^\infty(a,b)} \le C_{\mathbb{K}} \frac{||f^{(n+1)}||_{L^\infty(a,b)}}{(n+1-m)!} (b-a)^{n+1-m}$$

mit $C_{\mathbb{K}} = 1$ für reelwertiges f, $C_{\mathbb{K}} = 2$ für komplexwertiges f.

Proof. klar für $\mathbb{K} = \mathbb{R}$.
Für $\mathbb{K} = \mathbb{C}$, betrachte $Re(f), Im(f) \in \mathcal{C}^{n+1}[a,b]$. □

Bemerkung 5. *Aus der Fehlerabschätzung und der Konvergenz der Exponentialreihe $exp(y) = \sum_{k=0}^{\infty} \frac{y^k}{k!}$ folgt, dass der Interpolationsfehler "schnell" konvergiert, sofern sich die Ableitungen $||f^{(k)}||_{L^\infty(a,b)}$ gut verhalten (z.B. $||f^{(k)}||_{L^\infty(a,b)} \le M < \infty$).*

Bemerkung 6. *Für $m = 1, a = x$ und $b = x + h$ folgt*

$$|f'(x) - p'(x)| \le ||f' - p'||_{L^\infty(x,x+h)} \le C_{\mathbb{K}} \frac{||f^{(n+1)}||_{L^\infty(x,x+h)}}{n!} h^n$$

d.h. besser als die Differenzquotienten aus Kapitel 1.

2.2 Cebysev-Knoten

Definition 4. *Sei $f : \mathbb{R} \to \mathbb{R}$ eine Funktion und $x \in \mathbb{R}$. Man nennt x eine* ***n-fache Nullstelle von*** *f, gdw. $f(x) = 0$ und f is lokal um x $(n-1)$-mal diffbar mit $f^{(k)}(x) = 0 \forall k = 1, ..., n-1$. Wir schreiben $n(f,x) \in \mathbb{N}_0$ für die Vielfachheit.*

Lemma 3. *Sei $p \in \mathbb{P}_n$ mit Nullstellen $x_1 < ... < x_k$ und $N := \sum_{j=1}^{n} n(f, x_j) > n$.*
$\implies$ *$p = 0$, d.h. ein nicht-triviales Polynom vom Grad n hat $\le n$ Nullstellen, wobei diese mit Vielfachheit gezählt werden.*

Proof. Induktion nach n.
Ind.anf.: $n = 0$, d.h. p ist konstant mit mind. einer Nullstelle $\implies p = 0$✓
Ind.hyp: Die Aussage gelte für alle Polynome $q \in \mathbb{P}_{n-1}$.
$p \in \mathbb{P}_n$ hat Nullstellen $x_1 < ... < x_k$ und $N = \sum_{j=1}^{k} n(p, x_j) > n$
$\implies$ $p \in \mathbb{P}_{n-1}$ hat Nullstellen $\zeta_1 < ... < \zeta_{k-1}$ mit $x_j < \zeta_j < x_{j+1}$ (nach MWS) und bei allen x_j mit $n(p, x_j) > 1$.

Für die Nullstellen von $p' \in \mathbb{P}_{n-1}$ gilt also

$$\sum_{j=1}^{k-1} \underbrace{n(p', \zeta_j)}_{\geq 1} + \sum_{j=1}^{k} \max\{n(p,x_j)-1,0\} \geq -1 + \underbrace{\sum_{j=1}^{k} \underbrace{(\max\{n(p,x_j)-1,0\}+1)}_{\geq n(p,x_j)}}_{=N>n} > n-1$$

$\implies p' = 0 \implies p$ konstant $\implies p = 0.$ $\square$

Bemerkung 7. *Aus der linearen Algebra wissen wir, dass sich jedes Polynom $p \in \mathbb{P}_n$ mit Nullstelle x_0 in der Form $p(x) = q(x)(x - x_0)$ schreiben lässt mit $q \in \mathbb{P}_{n-1}$, sog.* ***Polynomdivision****.*

Ferner gilt der **Fundamentalsatz der Algebra**: Für jedes $p \in \mathbb{P}_n$ existieren $x_1, ..., x_n \in \mathbb{C}$ und $\lambda \in \mathbb{C}$ mit $p(x) = \lambda \prod_{j=1}^{n}(x - x_j)$. Offensichtlich ist diese Aussage viel stärker als "mein Lemma".

Ziel: Für $m = 0$ gilt für alle $x \in [a, b]$

$$|f(x) - p(x)| \leq C_{\mathbb{K}} \frac{||f^{(n+1)}||_{L^\infty(a,b)}}{(n+1)!} \prod_{l=0}^{n} |x - x_l|,$$

wenn f glatt und $p \in \mathbb{P}_n$ Lagrange-Interpolationspolynom zu x_j.

Nun wollen wir die x_j so wählen, dass $\max_{x\in[a,b]} \prod_{l=0}^{n} |x - x_l|$ minimal wird.

Definition 5. *Für $n \in \mathbb{N}_0$ definiere die* ***Cebysev-Polynome (der ersten Art)*** *durch $T_n(t) := \cos(n \arccos t)$ auf $[-1, 1]$.*

Lemma 4. *1. $T_n(\cos(\Phi)) = \cos(n\Phi) \forall 0 \leq \Phi \leq \pi \forall n \in \mathbb{N}_0$*

2. Auf $[-1, 1]$ gilt $T_0(t) = 1$, $T_1(t) = t$, $T_{n+1}(t) = 2tT_n(t) - T_{n-1}(t) \forall n \in \mathbb{N}$

3. $T_n \in \mathbb{P}_n[-1, 1]$ mit Leitkoeffizient 2^{n-1} für $n \geq 1$

4. $||T_n||_{L^\infty(-1,1)} = 1$

5. T_n hat in $[-1, 1]$ genau $n + 1$ lokale Extrema $T_n(s_j^{(n)}) = (-1)^j$ mit $s_j^{(n)} = \cos\left(\frac{j\pi}{n}\right)$ für $j = 0, ..., n$

6. T_n hat in $[-1, 1]$ genau n einfache Nullstellen $T_n(t_j^{(n)}) = 0$, $t_j^{(n)} = \cos\left(\frac{(2j-1)}{n}\frac{\pi}{2}\right)$ für $j = 1, ..., n$

nur die sog. Drei-Term-Rekursion in (2). Whl: Additionstheorem des Cosinus: $\cos(x)+\cos(y) = 2\cos\left(\frac{x+y}{2}\right)\cos\left(\frac{x-y}{2}\right)$

$t = \cos(\Theta), x := (n + 1)\Theta, y := (n - 1)\Theta$

$$\implies \frac{x+y}{2} = n\Theta, \frac{x-y}{2} = \Theta$$

$$\implies T_{n+1}(t) + T_{n-1}(t) = \cos(\underbrace{(n+1)\Theta}_{x}) + \cos(\underbrace{(n-1)\Theta}_{y}) = \underbrace{2\cos\left(\underbrace{\frac{x+y}{2}}_{n\Theta}\right)}_{T_n(t)} \underbrace{\cos\left(\underbrace{\frac{x-y}{2}}_{\Theta}\right)}_{=t}$$

$\square$

Satz 3 (Optimalität der Cebysev-Knoten)**.** *Betrachte die affine Transformation $\Psi : [-1, 1] \to [a, b], \Psi(t) = \frac{1}{2}\{(a + b) + t(b - a)\}$.*

Seien $t_1^{(n+1)}, ..., t_{n+1}^{(n+1)}$ die Nullstellen von T_{n+1}.

$$\implies \min_{x_0,...,x_n\in[a,b]} \max_{x\in[a,b]} \prod_{j=0}^{n} |x - x_j| = \max_{x\in[a,b]} \prod_{j=0}^{n} |x - \Psi(t_{j+1}^{(n+1)})| = \left(\frac{b-a}{2}\right)^{n+1} \frac{1}{2^n}.$$

Die $\Psi(t_{j+1}^{(n+1)})$ für $j = 0, ..., n$ heißen ***Cebysev-Knoten in [a,b]****.*

Proof. 1. zz: $\max_{t\in[-1,1]} \prod_{j=0}^{n} |t - t_{j+1}^{(n+1)}| = \frac{1}{2^n}$

Lemma (iii) + (vi) $\implies T_{n+1}(t) = 2^n \prod_{j=0}^{n}(t - t_{j+1}^{(n+1)})$

Lemma (iv) $\implies 1 = ||T_{n+1}||_{L^\infty(-1,1)} = \max_{t\in[-1,1]} 2^n \prod_{j=0}^{n} |t - t_{j+1}^{(n+1)}|$

2. zz. $\frac{1}{2^n} \leq \inf_{t_0,...,t_n \in [-1,1]} \max_{t\in[-1,1]} \prod_{j=0}^n |t-t_j|$ (dann folgt die Behauptung für $[a,b]=[-1,1]$)
   Annahme: Ex. $t_0,...,t_n \in [-1,1]$ mit $w(t) := \prod_{j=0}^n (t-t_i)$ erfüllt

$$||w||_{L^\infty(-1,1)} = \max_{t\in[-1,1]} \prod_{j=0}^n |t-t_j| \lneq \frac{1}{2^n}.$$

   Definiere $p := \underbrace{\frac{1}{2^n} T_{n+1}}_{\in \mathbb{P}_{n+1}} - \underbrace{w}_{\in \mathbb{P}_{n+1}} \in \mathbb{P}_n$. Ferner $\frac{1}{2^n} T_{n+1}(s_j^{(n+1)}) = \frac{(-1)^j}{2^n}$ und $|w(s_j^{n+1})| < \frac{1}{2^n}$.

   $\implies p$ hat $n+1$ Vorzeichenwechsel $\implies n+1$ Nullstellen $\implies p=0 \implies w = \frac{1}{2^n} T_{n+1}$ ↯

3. klar: Ψ ist Bijektion von $[-1,1]$ auf $[a,b]$
   $\Psi(t) - \Psi(t_{j+1}^{(n+1)}) = \frac{1}{2}\{(t - t_{j+1}^{(n+1)})(b-a)\}$

$$\implies \max_{x\in[a,b]} \prod_{j=0}^n |x - \Psi(t_{j+1}^{(n+1)})| = \max_{t\in[-1,1]} \prod_{j=0}^n |\Psi(t) - \Psi(t_{j+1}^{(n+1)})| =$$

$$\left(\frac{b-a}{2}\right)^{n+1} \underbrace{\max_{t\in[-1,1]} \prod_{j=0}^n |t - t_{j+1}(n+1)|}_{=\frac{1}{2^n}}$$

□

2.3 Lebesgue-Konstante

Satz 4. *Seien $a \leq x_0 < ... < x_n \leq b$ Stützstellen mit zugehörigen Lagrange-Polynomen $L_0, ..., L_n \in \mathbb{P}_n$.*
Def. $I_n : \mathcal{C}[a,b] \to \mathbb{P}_n, I_n f := \sum_{j=0}^n f(x_j) L_j$.
$\implies$ I_n ist eine lineare Projektion auf $\mathbb{P}_n$ mit Operatornorm

$$||I_n|| := \sup_{f\in\mathcal{C}[a,b], f\neq 0} \frac{||I_n f||_{L^\infty(a,b)}}{||f||_{L^\infty(a,b)}} = \max_{x\in[a,b]} \sum_{j=0}^n |L_j(x)| =: \Lambda(x_0,...,x_n).$$

Die Zahl $\Lambda(x_0,...,x_n)$ heißt ***Lebesgue-Konstante****.*

Proof. I_n wohldef., linear ✓
Für $p \in \mathbb{P}_n$ gilt $I_n p = p$, da Polynominterpolation eine eindeutige Lsg. hat.
Für $f \in \mathcal{C}[a,b]$ mit $f \neq 0$ gilt

$$\frac{||I_n f||_{L^\infty(a,b)}}{||f||_{L^\infty(a,b)}} = \max_{x\in[a,b]} \left| \sum_{j=0}^n \frac{f(x_j)}{||f||_{L^\infty(a,b)}} L_j(x) \right| \leq \Lambda(x_0,...,x_n).$$

Um Gleichheit zu zeigen, wähle $x \in [a,b]$ mit $\sum_{j=0}^n |L_j(x)| = \Lambda(x_0,...,x_n)$. Wähle $f \in \mathcal{C}[a,b]$ als Polygonzug mit $||f||_{L^\infty(a,b)} \leq 1$ und $f(x_j) = sign L_j(x)$. $\implies$ Gleichheit bei obiger Abschätzung. □

Bemerkung 8. *Derselbe Beweis zeigt für $\tilde{I}_n : \mathbb{K}^{n+1} \to \mathbb{P}_k, \tilde{I}_n(y_0,...,y_n) := \sum_{j=0}^n y_j L_j$, dass $||\tilde{I}_n(y_0,...,y_n)||_{L^\infty(a,b)} \leq \Lambda \max_{j=0,...,n} |y_j|$ mit Gleichheit für spezielle $y_j = sign(L_j(x))$, wenn $x \in [a,b]$ mit $\sum_{j=0}^n |L_j(x)| = \max_{\tilde{x}\in[a,b]} \sum_{j=0}^n |L_j(x)|$.*
$\implies$ $\tilde{I}$ ist der (lineare) Lösungsoperator der Pol.int.

$$\implies ||\tilde{I}_n(y_0,...,y_n) - I_n(\tilde{y}_0,...,\tilde{y}_n)||_{L^\infty(a,b)} \leq \Lambda \max_{j=0,...,n} |y_j - \tilde{y}_i|$$

d.h. Λ ist die abstrakte Konditionszahl der Polynominterpolation.

Abschließend einige Bemerkungen zum Bestapproximationsproblem.

Lemma 5. *X normierter Raum, $Y \leq X$ endlich-dim. Teilraum, $x \in X$*
$\implies$ Ex. $y \in Y$ mit $||x-y||_X = \min_{\tilde{y}\in Y} ||x - \tilde{y}||_X$

Proof. Wähle Folge $(y_n)_{n\in\mathbb{N}} \subseteq Y$ mit

$$\lim_{n\to\infty} ||x-y_n||_X = \inf_{\tilde{y}\in Y} ||x-\tilde{y}||_X$$

$$\implies ||y_n||_X \le \underbrace{||x-y_n||_X}_{\text{glm. beschränkt wegen Konvergenz}} + ||x||_X \implies \sup_{n\in\mathbb{N}} ||y_n||_X \le M < \infty$$

Da Y endl.-dim., gilt der Satz von Bolzano-Weierstraß, d.h. $(y_n)_{n\in\mathbb{N}}$ hat eine konvergente Teilfolge. O.B.d.A. ex. $y \in Y$ mit $||y-y_n||_X \to 0$ für $n \to \infty$.

$||x-y||_X = \lim_{n\to\infty} ||x-y_n||_X = \inf_{\tilde{y}\in Y} ||x-\tilde{y}||_X$. □

Bemerkung 9. *Mit Satz über Lebesgue-Konstante und dem Lemma gilt für alle* $q \in \mathbb{P}_n$

$$\underbrace{||f-I_nf||_{L^\infty(a,b)}}_{\ge \min_{q\in\mathbb{P}_n} ||f-q||_{L^\infty(a,b)}} \le ||f-q||_{L^\infty(a,b)} + \underbrace{||I_n(f-q)||_{L^\infty(a,b)}}_{\le \Lambda ||f-q||_{L^\infty(a,b)}}$$

$$\implies \min_{q\in\mathbb{P}_n} ||f-q||_{L^\infty(a,b)} \le ||f-I_nf||_{L^\infty(a,b)} \le (1+\Lambda) \min_{q\in\mathbb{P}_n} ||f-q||_{L^\infty(a,b)}$$

Bemerkung 10. *Nach Satz von Weierstraß gilt* $\lim\limits_{n\to\infty} \min_{p\in\mathbb{P}_n} ||f-p||_{L^\infty(a,b)} = 0 \forall f \in \mathcal{C}[a,b]$.

Nach Satz von Faber gilt allerdings, dass es für jede Folge von Stützstellen $(x_0^{(n)}, ..., x_n^{(n)})_{n\in\mathbb{N}}$ *eine Funktion* $f \in \mathcal{C}[a,b]$ *mit der Eigenschaft, dass* $||f - I_n^{(n)} f||_{L^\infty(a,b)}$ *divergiert!*

Insbesondere muss also $\Lambda_n^{(n)} \to \infty$ *gelten!*

Bemerkung 11. *Für äquidistante Stützstellen divergiert* Λ_n *exponentiell schnell. Für Cebysev-Knoten gilt allerdings* $\Lambda_n = \mathcal{O}(\log n)$.

Bemerkung 12. *Der* ***Remez-Algorithmus*** *berechnet (in unendlich vielen Schritten) ein Polynom* $q \in \mathbb{P}_n$ *mit* $||f-q||_{L^\infty(a,b)} = \min_{p\in\mathbb{P}_n} ||f-p||_{L^\infty(a,b)} \forall f \in \mathcal{C}[a,b]$.

Startwert ist dafür der Cebysev-Interpoland.

Der ***Alternantensatz von Cebysev*** *zeigt, dass das Bestapprox.polynom* $q \in \mathbb{P}_n$ *bzgl.* $||.||_{L^\infty(a,b)}$ *in der Tat eindeutig ist.*

2.4 Auswertung von Interpolationspol.

Satz 5 (Neville-Verfahren)**.** *Seien* $a \le x_0 < ... < x_n \le b$ *Stützstellen mit Funktionswerten* $y_j \in \mathbb{K}$ *und* $p \in \mathbb{P}_n$ *mit* $p(x_j) = y_j \forall j = 0, ..., n, x \in [a,b]$ *Auswertungspunkt.*

Für $j, m \in \mathbb{N}_0$ *mit* $j + m \le n$, *definiere* $p_{j,m} \in \mathbb{P}_m$ *als eind. Int.polynom mit* $p(x_k) = y_k \forall k = j, ..., j+m$

$$p(x) = p_{0,n}(x)$$

$$p_{j,0}(x) = y_j$$

$$p_{j,m}(x) = \underbrace{\frac{(x-x_j)p_{j+1,m-1}(x) - (x-x_{j+m})p_{j,m-1}}{x_{j+m}-x_j}}_{=:q(x), q\in\mathbb{P}_m}$$

Proof. $q(x) = y_j, q(x_{j+m}) = y_{j+m}, q(x_k) = y_k, k = j+1, ..., n-m+1 \implies q = p_{j,m}$ □

Dieser Satz führt auf das induktive **Neville-Schema**

$$\begin{array}{ccccccc}
y_0 & = & p_{0,0}(x) & \to & p_{0,1}(x) & \to & \dots \quad p_{0,n}(x) = p(x) \\
 & & & \nearrow & & & \\
y_1 & = & p_{1,0}(x) & \to & p_{1,1}(x) & \nearrow & \\
y_2 & = & p_{2,0}(x) & \nearrow & & & \\
 & \vdots & & & & & \\
y_{n-1} & = & p_{n-1,0}(x) & \to & p_{n-1,1}(x) & & \\
y_n & = & p_{n,0}(x) & \nearrow & & &
\end{array}$$

Bemerkung 13.
- *Das Neville-Verfahren ist ein sog.* ***Einschritt-Verfahren****, d.h. eine "neue Spalte" nur mit Hilfe der vorausgegangenen Spalte berechnet.*
- *Wenn man "von oben nach unten rechnet", ist kein zusätzlicher Speicher nötig. In diesem Fall sollte man die "Diagonale" speichern.*

- *Man kann im Neville-Verfahren dann leicht einen neuen Punkt (x_{n+1}, y_{n+1}) hinzunehmen und erhält $p_{0,n+1}(x)$, indem man nur die neue Diagonale rechnet.*

Algorithmus 1 (Neville)**.** *Input: Stützstellen $a \leq x_0 < ... < x_n \leq b$, Funktionswerte $y_0, ..., y_n \in \mathbb{K}$, Auswertungspunkt $x \in \mathbb{R}$*

- *for $m = 1 : n$*
- *for $j = 0 : n - m$*
- $y_j = \frac{(x-x_j)y_{j+1}-(x-x_{j+m})y_j}{x_{j+m}-x_j}$
- *end*
- *end*

Output: $y_0 = p(x)$, wobei $p \in \mathbb{P}_n$ mit $p(x_j) = y_j \forall j$
klar: Speicherbedarf $n + 1$ (überschreiben von y-Vektor), Arithmetischer Aufwand $\frac{7}{2}n(n + 1)$.

Definition 6. *Sei $p = \sum_{j=0}^{n} \lambda_j x^j \in \mathbb{P}_n$. Dann bezeichnet man λ_n als **führenden Koeffizienten von** p **bzgl.** $\mathbb{P}_n$.*
*Falls $j = 0$ oder ($\lambda_j \neq 0$ und $\lambda_k = 0 \forall k > j$), so bezeichnet man λ_j als **Leitkoeffizient von** p.*

Satz 6 (Newtons Dividierte Differenzen)**.** *Seien $a \leq x_0 < ... < x_n \leq b$ Stützstellen, $y_j \in \mathbb{K}, p \in \mathbb{P}_n$ mit $p(x_j) = y_j \forall j = 0, ..., n$. Für $j, m \in \mathbb{N}_0$ mit $j + m \leq n$ definiere*

$$y_{j,0} := y_j y_{j,m} := \frac{y_{j+1,m-1} - y_{j,m-1}}{x_{j+m} - x_j}$$

$\Longrightarrow$

1. *$y_{j,m}$ ist der führende Koeff. von $p_{j,m} \in \mathbb{P}_m$ aus dem Neville-Verfahren.*
2. *Mit $\lambda_j := y_{0,j}$ gilt $p(x) = \sum_{j=0}^{n} \lambda_j \underbrace{\prod_{k=0}^{j-1}(x - x_k)}_{=q_j \in \mathbb{P}_j}$ d.h. die dividierten Differenzen geben die Koeffizienten des Int.pol. bzgl. Newton-Basis.*

Proof. $q_k := p_{0,k} - p_{0,k-1} \in \mathbb{P}_n$ mit führendem Koeff. $y_{0,k}$ und Nullstellen $x_0, ..., x_{n-1}$
$\Longrightarrow q_k = y_{0,k} \prod_{j=0}^{k-1}(x - x_j)$ nach Pol.div.

$$\Longrightarrow p = p_{0,k} = p_{0,0} + \sum_{k=1}^{n} \underbrace{(p_{j,k} - p_{0,k-1})}_{=q_k} = y_{0,0} + \sum_{k=1}^{n} y_{0,k} \prod_{j=0}^{k-1}(x - x_j) = \sum_{k=0}^{n} \underbrace{y_{0,k}}_{=\lambda_k} \prod_{j=0}^{k-1}(x - x_j)$$

□

Schema der dividierten Differenzen

$$\begin{array}{ccccccccc}
y_0 & = & y_{0,0} & \searrow & & & \\
y_1 & = & y_{1,0} & \rightarrow & y_{0,1} & & \\
 & \vdots & & \searrow & & & \\
 & \vdots & & \rightarrow & y_{1,1} & \rightarrow & y_{0,2} \\
 & \vdots & & & & & \ddots \\
y_{n-1} & = & y_{n-1,0} & \searrow & & & \\
y_n & = & y_{n,0} & \rightarrow & y_{n-1,1} & \rightarrow & \ldots & & y_{0,n}
\end{array}$$

$\Longrightarrow$ arithmetischer Aufwand $3\frac{n(n+1)}{2}$, um alle $y_{0,j}$ zu berechnen.

Bemerkung 14.

- *Die dividierten Differenzen sind ein Einschrittverfahren.*
- *Wenn man den y-Vektor überschreibt, braucht man keinen zusätzlichen Speicher.*
- *Das Verfahren löst das Vandermonde-System für die Newton-Basis, aber die Matrix aufzustellen.*

- *Die Auswertung von $p(x)$ erfolgt mit Horner-Schema und Aufwand $3n$ pro $x \in \mathbb{K}$.*

Algorithmus 2. *Input: Stützstellen $x_0 < ... < x_n$, Funktionswerte $y_0, ..., y_n \in \mathbb{K}$*

- *for $m = 1 : n$*
- *for $j = n - m : -1 : 0$*
- $y_{j,m} := \frac{y_{j+m} - y_{j,m-1}}{x_{j+m} - x_j}$
- *end*
- *end*

Output: Koeffizienten des Interpol.pl. $p \in \mathbb{P}_n$ $y_0, ..., y_n$ bzgl. Newton-Basis.

Bemerkung 15. *Will man das Interpolationspolynom $p(x)$ an N Stellen auswerten, so gilt für den Gesamtaufwand: Aufwand(Neville) $= \frac{7}{2}Nn(n+1)$, Aufwand(Div. Diff. + Horner) $= \underbrace{\frac{3}{2}n(n+1)}_{\text{div. Diff.}} + \underbrace{3Nn}_{\text{Horner}}$.*

Es gilt immer: Aufwand(Div. Diff. + Horner) $\leq$ Aufwand(Neville). Wenn man sich den Fortpflanzungsfehler anschaut dann sieht man aber, dass Neville weniger anfällig ist für Auslöschung.

In der Praxis verwendet man deshalb Neville für kleine N und Div. Diff. + Horner für große N.

2.5 Hermite-Polynominterpolation

Satz 7 (Wohlgestelltheit)**.** *Gegeben seien Stützstellen $a \leq x_0 < ... < x_n \leq b$, Funktionswerte $y_j^(k) \in \mathbb{K}$ für $j = 0, ..., n$ und $k = 0, ..., n_j \in \mathbb{N}_0$ (Lagrange $n_j = 0 \forall j$), Def $N := \left(\sum_{j=0}^{n}(n_j + 1)\right) - 1$*

$\implies$ *Ex. eind. $p \in \mathbb{P}_N$ mit $p^{(k)}(x_j) = y_j^{(k)} \forall j = 0, ..., n, \forall k = 0, ..., n_j$, wobei $p^{(0)} = p$.*

Proof. Betrachte den Auswertungsoperator $\mathcal{A} : \mathbb{P}_N \to \mathbb{K}^{N+1}, \mathcal{A}p := (p(x_0), ..., p^{(n_0)}(x_0), p(x_1), ..., p^{(n_1)}(x_1), ..., p^{(n_n)}(x_n))$

klar: $\mathcal{A}$ ist linear und $\dim \mathbb{P}_N = N + 1$

$\implies \mathcal{A}$ ist $\underbrace{\text{bijektiv}}_{=\text{Behauptung}}$, gdw. $\mathcal{A}$ $\underbrace{\text{injektiv}}_{\text{zu zeigen!}}$ $\underbrace{(\text{oder } \mathcal{A} \text{ ist surj})}_{\text{"schwierig"}}$.

Sei $p \in \mathbb{P}_N$ mit $\mathcal{A}p = 0$, d.h. x_j eine (n_j+1)-fache Nullstelle von $p \forall j$. $\implies p \in \mathbb{P}_N$ hat $\sum_{j=0}^{n}(n_j+1) = N+1$ viele Nst. (bzgl. Vielfachheit) $\implies p = 0$. □

Bemerkung 16.
- *Der vorausgegangene Beweis ist das "normale Beweisprinzip" für lineare Interpolationsaufgaben. Klar: Man kann die Interpolationsaufgabe insb. lineares Gleichungssystem (äquivalent) formulieren.*
- *Neville-Verfahren und dividierte Differenzen lassen sich auch für das Hermite-Interpolationsproblem formulieren.*
- *Analog zu Lagrange (dieselbe Basis) kann man Fehlerdarstellung und Feherabschätzung beweisen, z.B.*

$$|f(x) - p(x)| \leq C_{\mathbb{K}} \frac{||f^{(N+1)}||_{L^\infty(a,b)}}{(N+1)!} \prod_{j=0}^{n} |x - x_j|^{n_j+1}$$

$C_{\mathbb{C}=\sqrt{2}}$ *(vorher 2)*

2.6 Spline-Interpolation

Die Polynominterpolation erfordert hohe Glätte an f, um Fehlerabschätzung zu kriegen. Alternativ kann man deshalb stückweise Polynome betrachten (sog. Splines), um Verfahren und Fehlerkontrolle zu haben, falls f nicht so glatt ist.

Beispiel 9 (affiner Interpolationsspline)**.** *Zu Stützstellen $a = x_0 < x_1 < ... < x_n = b$ und $f \in \mathcal{C}[a, b]$ ist $s \in \mathcal{C}[a, b]$ mit*

- $s|_{[x_{j-1}, x_j]} \in \mathbb{P}_1 \forall j = 1, ..., n$
- $s(x_j) = f(x_j) \forall j = 0, ..., n$

$\implies$ *Offensichtlich eindeutig $s(x) = f(x_{j-1})\frac{x - x_j}{x_{j-1} - x_j} + f(x_j)\frac{x - x_{j-1}}{x_j - x_{j-1}} \forall j \forall x \in [x_j - 1, x_j]$*

Lemma 6. *Zu $f \in \mathcal{C}[a,b] \cap \mathcal{C}^2[x_{j-1},x_j] \forall j = 1,...,n$ sei $s \in \mathcal{C}[a,b]$ der affine Interpolationsspline.*
Def $h : [a,b] \to \mathbb{R}_{>0}, h|_{[x_{j-1},x_j]} := x_j - x_{j-1}$ ***lokale Netzweite***
$\implies ||f-s||_{L^\infty(a,b)} \leq \frac{C_\mathbb{K}}{8}||h^2 f''||_{L^\infty(a,b)}$

Proof. Sei $x \in [x_{j-1},x_j]$

$$\implies |f(x)-s(x)| \leq C_\mathbb{K}\frac{||f''||_{L^\infty(x_{j-1},x_j)}}{2}\underbrace{|(x-x_{j-1})(x-x_j)}_{\text{maximal für } x=\frac{x_{j-1}+x_j}{2}}$$

$$\implies ||f-s||_{L^\infty(x_{j-1},x_j)} \leq C_\mathbb{K}\frac{||f''||_{L^\infty(x_{j-1},x_j)}}{2}\cdot\frac{(x_j-x_{j-1})^2}{9} = \frac{C_\mathbb{K}}{8}||h^2 f''||_{L^\infty(x_{j-1},x_j)}$$

□

Lemma 7. *Zu $f \in \mathcal{C}[a,b]$ und $s \in \mathcal{C}[a,b]$ affiner Int.spline*

$$\implies (i)||f-s||_{L^2(a,b)} \leq ||hf'||_{L^2(a,b)}, \text{ sofern } f \in \mathcal{C}^1[x_{j-1},x_j] \text{ für alle } j$$
$$(ii)||f-s||_{L^2(a,b)} \leq ||h^2 f||_{L^2(a,b)}, \text{ sofern } f \in \mathcal{C}^2[x_{j-1},x_j] \text{ für alle } j$$

Proof. zz: (i) elementweise für $I_j = [x_{j-1},x_j]$

$$F := f - s \in \mathcal{C}^1[x_{j-1},x_j], h_j := x_j - x_{j-1}$$
$$F(x_{j-1}) = 0 \implies \int_{I_j}|F(x)|^2 = \int_{I_j}\left|\int_{x_{j-1}}^{x_j}F'dx\right|^2 \leq h_j^2||F'||^2_{L^2(I_j)}$$
$$\implies ||F||_{L^2(I_j)} \leq h_j||F'||_{L^2(I_j)}$$
$$F(x_{j-1}) = 0 = F(x_j) \implies \int_{I_j}F'dx = 0$$

$s'|_{I_j}$ konstant

$$\implies s'|_{I_j} = \frac{1}{h_j}\int_{I_j}f'dx \implies < f', s' >_{L^2(I_j)} = ||s'||^2_{L^2(I_j)}$$
$$\implies ||F'||^2_{L^2(I_j)} = ||f'||^2_{L^2(I_j)} - 2Re < f', s' >_{L^2(I_j)} + ||s'||^2_{L^2(I_j)} = ||f'||^2_{L^2(I_j)} - ||s'||^2_{L^2(I_j)}$$
$$\implies ||f-s||_{L^2(I_j)} = ||F||_{L^2(I_j)} \leq h_j||F'||_{L^2(I_j)} \leq h_j||f'||_{L^2(I_j)} = ||hf'||_{L^2(I_j)}$$

zz: (ii) elementweise

$$ReF(x_{j-1}) = 0 = ReF(x_j) \implies ReF'(\zeta) = 0 \text{ für } x_{j-1} < \zeta < x_j$$
$$|ReF'(x)| = |\int_J ReF''dt| \leq h_j^{\frac{1}{2}}||ReF''||_{L^2(I_j)}$$

analog für ImF'

$$\implies \int_{I_j}|F'(x)|^2 \leq h_j^2||F''||^2_{L^2(I_j)} \implies ||F'||_{L^2(I_j)} \leq h_j||F''||_{L^2(I_j)}$$
$$\implies ||f-s||_{L^2(I_j)} \leq h_j||F'||_{L^2(I_j)} \leq h_j^2||\underbrace{F''}_{=f'' \text{ auf } I_j}||_{L^2(I_j)} = ||h^2 f''||_{L^2(I_j)}$$

(3) $||f-s||^2_{L^2(a,b)} = \sum_{j=1}^n ||f-s||^2_{L^2(I_j)} \leq \sum_{j=1}^n ||h^2 f''||^2_{L^2(I_j)} = ||h^2 f''||^2_{L^2(a,b)}$ □

Definition 7. *Es sei $\Delta = (x_0,...,x_n)$ eine* ***Zerlegung von*** *$[a,b]$, d.h. $a = x_0 < ... < x_n = b$. Zu gegebenen $p, q \in \mathbb{N}_0$ heißt $s : [a,b] \to \mathbb{K}$* ***Spline vom Grad*** *p* ***mit Glattheit*** *q, gdw. $s \in \mathcal{C}^q[a,b]$ mit $s|_{[x_{j-1},x_j]} \in \mathbb{P}_p \forall j = 1,...,n$.*
Schreibweise $s \in \mathbb{S}^p_q(\Delta)$ bzw. $s \in \mathbb{S}^p(\Delta)$, falls $q = p-1$.

Bemerkung 17. *Die wichtigsten Beispiele sind* ***affine Splines*** *$\mathbb{S}^1(\Delta)$,* ***quadratische Splines*** *$\mathbb{S}^2(\Delta)$,* ***kubische Splines*** *$\mathbb{S}^3(\Delta)$.*

Bemerkung 18. *Für Wohlgestelltheit von Interpolationsaufgaben muss man $\dim \mathbb{S}^p_q(\Delta)$ bestimmen.*

Beispiel 10. $\dim \mathbb{S}_1^p(\Delta) = \underbrace{n}_{Anz.\ Intervalle} \underbrace{(p+1)}_{=\dim \mathbb{P}_p} \hat{=}$ *global unstetige stückweise Polynome*

$\dim \mathbb{S}_0^p(\Delta) = n(p+1) - (n-1) \hat{=}$ *global stetige stw. Polynome*

$\implies$ Bestimmung von Basen (und Dimension) von $\mathbb{S}_q^p(\Delta)$ ist komplizierter als bei Polynomräumen, lässt sich aber über sogn. B-Splines bewerkstelligen.

Bemerkung 19. *Bei Splines hat man mehrere Möglichkeiten, um den Fehler $||f - s||$ zu verringern*

- ***h-Methode**, d.h. die Zerlegung wird verfeinert*
- ***p-Methode**, d.h. man erhöht den Polynomgrad*
- ***r-Methode**, d.h. man erhält $\#\Delta = n$, aber man verschiebt die Stützstellen.*

Zusätzlich kann man alles lokal kombinieren, z.B.

- ***hp-Methode**: Verfeinere Δ, wo f unglatt ist, und erhöhe p, wo f glatt ist.*

Satz 8. *Sei $\Delta = (x_0, ..., x_n)$ Zerlegung von $[a, b]$*
$\implies$ $\dim \mathbb{S}^p(\Delta) = n + p$ *und* $\mathcal{B} := \{x^j, \max\{x - x_k, 0\}^p | j = 0, ..., p, k = 1, ..., n-1\}$ *ist eine Basis.*

Proof. klar: $\mathcal{B} \subseteq \mathbb{S}^p(\Delta)$
(1) zz. $\mathcal{B}$ ist lin. unabh.
Seien $\lambda_j, \mu_j \in \mathbb{K}$ mit $0 = \sum_{j=0}^p \lambda_j x^j + \sum_{k=1}^{n-1} \mu_k q_k$
klar: Auf $[a, b]$ gilt $q_k|_{[x_0,x_1]} = 0 \forall k = 1, ..., n-1 \implies \lambda_j = 0 \forall j = 0, ..., p$, da Monome lin. unabh. auf $[x_0, x_1]$.
klar: Auf $[x_1, x_2]$ gilt $q_k|_{[x_1,x_2]} = 0 \forall k = 2, ..., n-1 \implies \mu_1 q_1 = 0$ auf $[x_1, x_2]$, $q_1(x_2) \neq 0 \implies \mu_1 = 0$
Induktives Vorgehen zeigt $\mu_k = 0 \forall k = 1, ..., n-1$
(2) zz: $\mathbb{S}^p(\Delta) \subseteq [\mathcal{B}]$
Sei $s \in \mathbb{S}^p(\Delta)$. Sei $z_1 \in \mathbb{P}_p$ mit $s|_{[x_0,x_1]} = z_1|_{[x_0,x_1]}$.
Für $k = 2, ..., n$ definiere sukzessive

$$z_k := z_{k-1} + \frac{s(x_k) - z_{k-1}(x_k)}{q_{k-1}(x_n)} q_{n-1} \in span(B)$$

zz. $z_n = s$
Betrache Residuum $r := s - z_n$
klar: $r(x_k) = 0 \forall k = 0, ..., n$
Auf $[x_0, x_1]$ gilt $q_k|_{[x_0,x_1]} = 0 \forall k = 1, ..., n-1 \implies z_n|_{[x_0,x_1]} = z_1|_{[x_0,x_1]} = s|_{[x_0,x_1]} \implies r|_{[x_0,x_1]} = 0$.
Auf $[x_1, x_2]$ gilt $r \in \mathbb{P}_p$ mit $r(x_1) = 0 = r(x_2)$. $r \in \mathcal{C}^{p-1}[a, b]$, insb. $(p-1)$-mal stetig diffbar in x_1, d.h. x_1 ist p-fache Nst. von r. $\implies p + 1$ viele Nst. auf $[x_1, x_2] \implies r = 0$ auf $[x_1, x_2]$.
Dasselbe Vorgegen auf allen Intervallen zeigt $r = 0$ auf $[a, b]$. □

Bemerkung 20. *Sei $\Delta = (x_0, ..., x_1)$ Zerlegung von $[a, b]$ und $s \in \mathbb{S}^p(\Delta)$ mit $s(x_j) = y_j \forall j = 0, ..., n$. Dann sind nur $n+1$ Bedingungen fixiert, aber $\dim \mathbb{S}^p(\Delta) = n+p$, d.h. es fehlen noch $p-1$ Bedingungen für Wohlgestelltheit.*

Diese werden in der Regel als Randbedingungen formuliert, d.h. an Bedingungen an Ableitungen von s in a und b. Um diese symmetrisch zu stellen, müssen wir annehmen, dass $p - 1 = 2r$ gerade ist.

- *(H) **Hermite-Randbedingungen** (oder **vollständige Randbedingungen**): Es werden $s^(j)(a), s^(j)(b)$ für $j = 1, ..., r$ zusätzlich vorgegeben.*
- *(N) **Natürliche Randbedingungen**: $r \leq n$ und $s^{(j)}(a) = 0 = s^{(j)}(b)$ für $j = r + 1, ..., 2r$*
- *(P) **Periodische Randbedingungen**: $s^{(j)}(a) = s^{(b)}$ für $j = 1, ..., 2r$*

Satz 9. *1. Zu $p - 1 = 2r$ und Randbedingungen (H), (P), (N) existiert (jeweils) ein eindeutiger Interpolationsspline $s \in \mathbb{S}^p(\Delta)$, der diese Randbedingung und $s(x_j) = y_j \forall j = 0, ..., n$ wobei $\Delta = (x_0, ..., x_n)$ und $y_j \in \mathbb{K}$ gegeben.*

2. Erfüllt $g \in \mathcal{C}^{(r+1)}[a, b]$ dieselben Interpolations- und Randbedingungen wie s, so gilt

$$||g^{(r+1)} - s^{(r+1)}||^2_{L^2(a,b)} = ||g^{(r+1)}||^2_{L^2(a,b)} - ||s^{(r+1)}||^2_{L^2(a,b)}$$

d.h. der Spline erfüllt die Minimaleigenschaft

$$||s^{(r+1)}||_{L^2(a,b)} \leq ||g^{(r+1)}||_{L^2(a,b)}$$

3. *Falls s (N) erfüllt, so muss g nur die Interpolationsbedingungen erfüllen!*

Bemerkung 21. *Für $p = 2$ minimieren kubische Interpolationssplines also die Krümmungsenergie $||s''||^2_{L^2(a,b)}$. Daher kommt auch der Name Spline (dt: Biegestab)*

Proof. (1) zz. (i) unter der Veraussetzung, dass (ii) gilt.

Seien $s, \tilde{s} \in \mathbb{S}^p(\Delta)$ mit derselben Interpolations- und Randbedingungen. Aus (ii) folgt dann

$$||\tilde{s}^{(r+1)} - s^{(r+1)}||_{L^2(a,b)} = 0 \implies \rho := \tilde{s} - s \in \mathbb{P}_r$$

(H) $\rho(a) = 0$ und $\rho^{(j)}(a) = 0$ für $j = 1, ..., r \implies a$ ist $(r+1)$-fache Nullstelle $\implies \rho = 0$
(N) $\rho(x_j) = 0$ für $j = 0, ..., n$ und $r \leq n \implies$ mehr als $r+1$ Nullstellen $\implies \rho = 0$
(P) $\rho^{(j)}(a) = \rho^{(j)}(b)$ für $j = 1, ..., 2r$. $\rho \in \mathbb{P}_r \implies \rho^{(r-1)} \in \mathbb{P}_1$ und $\rho^{(r-1)}(a) = \rho^{(r-1)}(b) \implies \rho^{(r-1)} \in \mathbb{P}_0 \implies \rho \in \mathbb{P}_{r-1}$. Induktiv $\rho \in \mathbb{P}_0$ mit $\rho(a) = 0 \implies \rho = 0$.

Jetzt folgt mit Standardargumentation, dass der lineare Auswertungsoperator $\mathcal{A} : \mathbb{S}^p(\Delta) \to \mathbb{K}^{n+p}$ bijektiv ist und aus Dimensionsgründen bijektiv. Insbesondere ist der Lösungsoperator $\mathcal{L} = \mathcal{A}^{-1}$ bijektiv.

(ii) Betrachte $|x-y|^2 = |x|^2 - 2Rex\bar{y} + |y|^2 = |x|^2 - |y|^2 - 2Re(x-y)\bar{y}$

$$\implies ||g^{(r+1)} - s^{(r+1)}||^2_{L^2(a,b)} = ||g^{(r+1)}||^2_{L^2(a,b)} - ||s^{(r+1)}||^2_{L^2(a,b)} - 2Re \underbrace{\int_a^b (g^{(r+1)} - s^2(r+1))\overline{s^{(r+1)}}dx}_{\text{zz: } =0}$$

Wh: $\int_a^b F'G + FG'dx = [FG]_a^b$

$$\int_a^b (g^{(r+1)} - s^{(r+1)})\overline{s^{(r+1)}}dx = \left[(g^{(r)} - s^{(r)})\overline{s^{(r+2)}}dx\right]_a^b - \int_a^b (g^{(r)} - s^{(r)})\overline{s^{(r+2)}}dx =$$

$$\left[(g^{(r)} - s^{(r)})\overline{s^{(r+2)}}dx\right]_a^b - \left[(g^{(r-1)} - s^{(r-1)})\overline{s^{(r+2)}}dx\right]_a^b + \int_a^b (g^{(r-1)} - s^{(r-1)})\overline{s^{(r+3)}}dx =$$

$$\left(\sum_{j=0}^{r-1} (-1)^j \left[(g^{(r-j)} - s^{(r-j)})\overline{s^{(r+1+j)}}\right]_a^b\right) + (-1)^j \int_a^b (g' - s') \underbrace{\overline{s^{(2r+1)}}_{=\overline{s^{(p)}} \leftarrow \text{ ist konstant auf } [x_{j-1}, x_j]}} dx$$

$$\int_a^b (g' - s')\overline{s^{(p)}}dx = \sum_{j=1}^n \underbrace{\int_{x_{j-1}}^{x_j} (g' - s') \underbrace{\overline{s^{(p)}}}_{\in \mathbb{K}} dx}_{= \overline{s^{(p)}}_{[x_{j-1}, x_j]} \underbrace{\int_{x_{j-1}}^{x_j} (g' - s')dx}_{=[g-s]_{x_{j-1}}^{x_j} = 0}}$$

(H) $(g^{(r-j)} - s^{(r-j)})(\frac{a}{b}) = 0 \implies \sum(.) = 0$
(N) $\overline{s^{(r+1+j)}}(\frac{a}{b}) = 0 \implies \sum(.) = 0$
(P) $[g^{(r-j)}]_a^b = 0, [s^{(r-j)}]_a^b = 0 \implies \sum(.) = 0$ □

Bemerkung 22. *Man kann mit Hilfe der sog. B-Splines Basen von allgemeineren Spline-Räumen konstruieren: Sei $(t_j)_{j \in \mathbb{Z}}$ monoton steigend mit $\lim_{j \to \pm\infty} t_j = \pm\infty$.*

Def $B_{j,0} := \Xi_{[t_j, t_{j+1})}$ für $p = 0$, $B_{j,p} := w_{j,p}B_{j,p-1} + (1 - w_{j,p})B_{j+1,p-1}$ für $p \geq 1$ mit $w_{j,p}(x) = \frac{x - t_j}{t_{j+p} - t_j}$ für $t_j < t_{j+p}$ und $w_{j,p}(x) = 0$ für $t_j = t_{j+p}$

- *Man kann zeigen, dass für $t_j = x_j$ mit $\Delta = (x_0, ..., x_n)$ Zerlegung von $[a, b]$ gilt $\mathbb{S}^p(\Delta) = span\{B_{j,p}|_{[a,b]} : j = p, ..., n-1\}$ und diese $B_{j,p}$ bilden auch eine Basis (da $n + p$ viele).*
- *Ferner haben die $B_{j,p}$ "gute" Eigenschaften, z.B. $B_{j,p} > 0, supp(B_{j,p}) = [t_j, t_{j+p+1}], \sum_{j \in \mathbb{Z}} B_{j,p} = 1$.*
- *Die Vielfachheit eines Knotens $t_j = ... = t_{j+k}$ reduziert die Differenzierbarkeit der Splines bei t_k um k, d.h. $\mathcal{C}^{p-(k+1)}$-Differenzierbarkeit.*

Übung: Sei $s \in \mathbb{S}^p(\Delta)$ interpolierend $s(x_j) = y_j$ mit natürlichen Randbedingungen ($\rightsquigarrow$ "naives" Gleichungssystem mit Basis aus Satz wobei $(n+3) \times (n+3)$ Gleichungssysteme mit vollbesetzter Vandermonde-Matrix). Auf $[x_{j-1}, x_j]$ machen wir den Ansatz $s|_{[x_{j-1}, x_j]} = a_0^{(j)} + a_1^{(j)}(x - x_j) + a_2^{(j)}(x - x_j)^2 + a_3^{(j)}(x - x_j)^3$ mit unbekannten Koeffizienten $a_0^{(j)}, a_1^{(j)}, a_2^{(j)}, a_3^{(j)} \forall j = 1, ..., n$.

$$\implies a_0^{(j)} = y_j, a_1^{(j)} = \frac{y_j - y_{j-1}}{h_j} + \frac{h_j}{3}(2a_2^{(j)} + a_2^{(j-1)}), a_3^{(j)} = \frac{a_2^{(j)} - a_2^{(j-1)}}{3h_j}, h_j := x_j - x_{j-1} \text{ und } a_2^{(j)} = 0$$

TODO Matrix 07 27:32

Bemerkung 23. *Für interpolierende Splines kann man Fehlerabschätzungen der Form* $||f-s||_{L^\infty(a,b)} \leq Ch^{p+1}||f^{(p+1)}||_{L^\infty(a,b)}$ *wobei* $h := \max_j(x_j - x_{j-1})$. *Aber* $C > 0$ *hängt an der Glattheit der Splines. Das ist technisch zu beweisen und hängt (natürlich) auch an den Randbdg.*

2.7 Diskrete und schnelle Fourier-Transformation

Das Ziel dieses Abschnitts ist die Präsentation und Analyse des FFT-Algorithmus ("fast Fourier transform"), der in der Numerik einer der wichtigsten und vielfaltigsten Algorithmen ist. Unsere "Motivation" ist die einfachste Anwendung.

Definition 8. *Wir bezeichnen* $\mathbb{T}_n := \{\sum_{j=0}^{n-1} \lambda_j \exp(ijx) | \lambda_0, ..., \lambda_{n-1} \in \mathbb{C}\}$ *als Raum der* ***trigonometrischen Polynome****.*

Satz 10. *1.* $\dim \mathbb{T}_{n-1} = n$

2. Zu Stützstellen $0 \leq x_0 < ... < x_{n-1} \lneq 2\pi$ *und Funktionswerten* $y_j \in \mathbb{C}$ *ex. eind.* $p \in \mathbb{T}_{n-1}$ *mit* $p(x_j) = y_j \forall j = 0, ..., n-1$

Proof. klar: $\dim \mathbb{T}_{n-1} \leq n$
Sei $p(x) = \sum_{j=0}^{n-1} \lambda_j \exp(ijx) \in \mathbb{T}_n$ mit $p(x_j) = 0 \forall j = 0, ..., n-1$
zz: $\lambda_j = 0 \forall j = 0, ..., n-1$ (dann lin. unabh. und Eind. des Int.pol.)
Def $z_k := \exp(ix_k) \in \mathbb{C}, \exp(ijx_k) = z_k^j$
$\implies 0 = p(x_k) = \sum_{j=0}^{n-1} \lambda_j z_k^j =: \tilde{p}(z_n), \tilde{p} \in \mathbb{P}_{n-1} \implies \tilde{p} = 0$, da $\tilde{p} n$ Nullstellen hat $\implies \lambda_j = 0 \forall j = 0, ..., n-1$. □

Im restlichen Abschnitt betrachten wir **äquidistante Stützstellen** (oder **uniforme Stützstellen**) $x_k = \frac{2\pi k}{n}$ für $k = 0, ..., n-1$. Dies führt auf zusätzliche Struktur der Vandermonde-Matrix, die durch FFT genutzt werden kann.

Satz 11. *Seien* $x_k = \frac{2\pi k}{n}, y_k \in \mathbb{C}$ *gegeben für* $k = 0, ..., n-1$. *Sei* $p \in \mathbb{T}_n$ *das eind. trig. Int.pol. mit* $p(x_j) = y_j \forall k$.
Sei $p(x) = \sum_{j=0}^{n-1} \lambda_j \exp(ijx)$ *mit Koeff.* $\lambda_j \in \mathbb{C}$.
Def $w_n := \exp(-\frac{2\pi i}{n})$ ***n-te Einheitswurzel*** *und* $V_n \in \mathbb{C}^{n \times n}, V_n := (w_n^{jk})_{j,k=0}^{n-1}$ ***Fourier-Matrix*** *(oder:* ***DFT-Matrix****)*
$\implies$

1. $\frac{1}{n} V_n y = \lambda$ *mit* $y = (y_0, ..., y_{n-1})$, *d.h.* $\lambda_j = \frac{1}{n} \sum_{k=0}^{n-1} w_n^{jk} y_k$

2. $\frac{1}{\sqrt{n}} V_n$ *ist symmetrisch und orthogonal, d.h.* $\left(\frac{1}{\sqrt{n}} V_n\right)^{-1} = \frac{1}{\sqrt{n}} \overline{V_n}$.

3. *Insb. ist* $W = \overline{V_n}$ *die Vandermonde-Matrix der trigonometrischen Interpolation*

Proof. (iii) λ ist Lösung des Vandermonde-Systems $W\lambda = y$ mit

$$W = \begin{pmatrix} p_0(x_0) & \cdots & p_{n-1}(x_0) \\ \vdots & & \vdots \\ p_0(x_{n-1}) & \cdots & p_{n-1}(x_{n-1}) \end{pmatrix} \text{ mit } p_j = \exp(ijx)$$

$W_{jk} = p_k(x_j) = \exp(ikx_j) = \exp\left(\frac{2\pi i}{n} jk\right) = w_n^{-jk} = (\overline{w_n})^{jk} \implies W = \overline{V_n} = W$ symmetrisch.
Beweis (i), (ii): Sei $W^{(k)} = (W_{jk})_{j=0}^{n-1}$ k-te Spalte von W.

$$W^{(k)} \cdot W^{(k)} = \sum_{j=0}^{n-1} W_{jk} \overline{W}_{jk} = \sum_{j=0}^{n-1} \underbrace{|W_{jk}|^2}_{=|w_n|^{2jk}=1} = n$$

Für $k + l$ gilt

$$W^{(k)} \cdot W^{(l)} = \sum_{j=0}^{n-1} W_{jk} \overline{W_{jl}} = \sum_{j=0}^{n-1} \underbrace{w_n^{j(l-k)}}_{=(w_n^{(l-k)})^j=0} = \frac{1 - (w_n^{(l-k)})^n}{1 - w_n^{(l-k)}}$$

$$(w_n^{(l-k)})^n = (w_n^n)^{l-k} = 1$$

$\implies \frac{1}{\sqrt{n}} W$ ist symmetrisch und orthogonal und $W^{-1} = \frac{1}{\sqrt{n}} \left(\frac{1}{\sqrt{n}} W\right)^{-1} = \frac{1}{n} \bar{W} = \frac{1}{n} V_n$. □

Definition 9. *Die Abbildung $\mathcal{F}: \mathbb{C}^n \to \mathbb{C}^n, \mathcal{F}_n(x) = V_n x$ bezeichnet man als* ***diskrete Fourier-Transformation (DFT) der Länge*** *n. Die Inverse $\mathcal{F}_n^{-1}\mathbb{C}^n \to \mathbb{C}^n, \mathcal{F}_n^{-1}(x) = \frac{1}{n}\bar{V}_n x$ heißt* ***diskrete Fourier-Rücktransformation****.*

Bemerkung 24. *Die Skalierung von $\mathcal{F}_n$ ist in der Literatur uneinheitlich. Manchmal $\frac{1}{n}V_n, \frac{1}{\sqrt{n}}V_n$,* ***hier*** *$\mathcal{F}_n(x) = V_n x$.*

Satz 12. *Für $p \in \mathbb{N}$ und $n = 2^p$ und $m = \frac{n}{2}$ sei $w_n := \exp(-\frac{2\pi i}{n})$. Betrachte Permutation $\sigma_n : \mathbb{C}^n \to \mathbb{C}^n, \sigma_n(x) = (\underbrace{x_1, x_3, ..., x_{n-1}}_{\text{ungerade Indizes}}, \underbrace{x_2, x_4, ..., x_n}_{\text{gerade Indizes}})$.*

Für $x \in \mathbb{C}^n$ definiere $a, b \in \mathbb{C}^m$ durch $a_j := x_j + x_{j+m}, b_j = (x_j - x_{j+m})w_n^{j-1}$ für $j = 1, ..., m = \frac{n}{2}$.

$\implies \sigma_n(\mathcal{F}_n(x)) = (\mathcal{F}_m(a), \mathcal{F}_m(b))$, d.h. Auswertung von $\mathcal{F}_n(x)$ wird auf Fourier-Trafo halber Länge $\mathcal{F}_m(a), \mathcal{F}_m(b)$ zurückgeführt (+ Vertauschen).

Proof.

$$\mathcal{F}_n(y_0, ..., y_{n-1}) = \left(\sum_{l=0}^{n-1} w_n^{jl} y_l | j = 0, ..., n-1\right)$$
$$\mathcal{F}_n(x_1, ..., x_n) = \left(\sum_{l=0}^{n-1} w_n^{(j-1)l} x_{l+1} | j = 1, ..., n\right)$$

(1) zz: $(\mathcal{F}_n(x))_{2j-1} = (\mathcal{F}_m(a))_j$
klar: $w_n^2 = w_m, w_m^m = 1$

$$\implies (\mathcal{F}_n(x))_{2j-1} = \sum_{l=0}^{n-1} w_n^{(2j-2)l} x_{l+1} = \sum_{l=0}^{m-1} (\underbrace{w_n^{2(j-1)l}}_{=w_m^{(j-1)l}} x_{l+1} + \underbrace{w_n^{2(j-1)(l+m)}}_{=w_m^{(j-1)l}\underbrace{w_m^{2(j-1)m}}_{=1}} x_{l+m+1}) =$$
$$\sum_{l=0}^{m-1} w_m^{(j-1)l} \underbrace{(x_{l+1} + x_{l+m+1})}_{=a_{l+1}} = (\mathcal{F}_m(a))_j$$

(2) zz: $(\mathcal{F}_n(x))_{2j} = (\mathcal{F}_m(b))_j$
klar: $w_n^{(2j-1)l} = w_n^{2(j-1)l} w_n^l = w_m^{(j-1)l} w_n^l$, $w_n^{(2j-1)(l+m)} = w_n^{2(j-1)l} w_n^l \underbrace{w_n^{(2j-1)m}}_{=(w_n^{n/2})^{(2j+1)}=-1} = -w_m^{(j-1)l} w_n^l$

$$\implies (\mathcal{F}_n(x))_{2j} = \sum_{l=0}^{n-1} w_n^{(2j-1)} x_{l+1} = \sum_{l=0}^{n-1} \underbrace{(w_n^{(2j-1)l} x_{l+1} + w_n^{(2j-1)(l+m)} x_{l+m+1})}_{=w_m^{(j-1)l}} \underbrace{(x_{l+1} - x_{l+m+1}) w_n^l}_{=b_{l+1}} = (\mathcal{F}_m(b))_j$$

□

Bemerkung 25. *In Matrixform kann man die FFT als Faktorisierung schreiben:*

$$V_n = P_n \begin{pmatrix} V_{n/2} & 0 \\ 0 & V_{n/2} \end{pmatrix} \begin{pmatrix} I_{n/2} & I_{n/2} \\ D_{n/2} & -D_{n/2} \end{pmatrix}$$

mit $P_n = (e_1, e_3, ..., e_{n-1}, e_2, e_4, ..., e_n) \in \mathbb{R}^{n \times n}$, $I_{n/2}$ Identität, $D_{n/2} = diag(w_n^0, ..., w_n^{n-1}) \in \mathbb{C}^{n \times n}$

Korollar 2. *Als* ***Fast-Fourier-Transformation (FFT)*** *bezeichnet man die rekursive Berechnung von $\mathcal{F}_n(x)$ mit $n = 2^p$ mittels der Rekursion aus dem Satz. Die gesamte Rekursion behötigt weniger als $\frac{3}{2} n \log_2 n$ arithmetische Operationen plus die Berechnung von w_n^l für $l = 0, ..., n-1$.*

Proof. (1) Wegen $w_{n/2}^l = w_n^{2l}$ reicht es, alle w_n^l für $l = 0, ..., n-1$ und das maximale n zu berechnen.
(2) Beweis des arithm. Aufwands durch Induktion nach p.

- $A(p) =$ Anzahl Additionen/Subtraktionen
- $M(p) =$ Anzahl Multiplikationen

Beh: $A(p) = p2^p, M(p) \leq \frac{1}{2}p2^p$ (dann fertig, da Aufwand$(\mathcal{F}_n(x)) = A(p) + M(p) \leq \frac{2}{3}p2^p = \frac{3}{2}n\log_2 n$)
Ind.anf. $p = 1, n = 2, ..., m-1 \implies A(p) = 2, M(p) = 0$ ✓
Ind.schritt: Die Aussage gelte für p, dann $A(p+1) = 2\underbrace{A(p)}_{=p2^p} + 2\cdot 2^p = (p+1)2^{p+1}$ ✓. $M(p+1) = 2\underbrace{M(p)}_{\leq\frac{1}{2}p2^p} + 2^p \leq p2^p + 2^p = \frac{1}{2}(p+1)2^{p+1}$ ✓ □

Bemerkung 26. *Die FFT ist eine schnelle Matrix-Vektor-Multiplikation für vollbesetzte Matrizen, die eine gewisse Struktur haben, ohne die Matrix jeweils voll aufzubauen, d.h. Speicherbedarf $\mathcal{O}(n)$.*

UE: $A \in \mathbb{K}^{n\times n}$ **zirkulant**, d.h.

$$A = \begin{pmatrix} a_0 & a_{n-1} & \cdot & a_1 \\ a_1 & a_0 & \ddots & \vdots \\ \vdots & \ddots & \vdots & \\ a_{n-1} & a_{n-2} & \cdots & a_0 \end{pmatrix}$$

$\implies V_nAV_n^{-1} = D = diag(p(1), p(w_n), ..., p(w_n^{n-1}))$ mit $p(x) = \sum_{j=0}^n a_jx^j$
$\implies$ klar: A ist regulär, gdw. $p(w_n^j) \neq 0 \forall j = 0, ..., n-1$
Sei A zusätzlich regulär, $b \in \mathbb{K}^k$, Ziel: Löse $Ax = b$
$\implies A = V_n^{-1}DV_n, A^{-1} = V_n^{-1}D^{-1}V_n \implies x = V_n^{-1}D^{-1}V_nb = \frac{1}{n}\overline{V_n}D^{-1}V_nb$, $\overline{V_n\overline{y}}$
Frage: Wie berechnet man die Diagonale von D effizient?
naiv: $\mathcal{O}(n^2)$, clever: $V_nA = DV_n$, insb. $V_na = D(1, ..., 1)^T = diag(D) \implies$ FFT liefert $diag(D)$.

Bemerkung 27.
- *Die FFT verdankt ihren Namen dem Zusammenhang mit der Fourier-Transformation: Für $f \in L^2[0, 2\pi]$ definiere $\hat{f}(k) := \frac{1}{2\pi}\int_0^{2\pi} f(t)e^{-ikt}dt$ und $S_n(t) := \sum_{k=-n}^n \hat{f}(k)e^{ikt}$*
  $\implies \lim_n ||f - S_n||_{L^2(0,2\pi)} = 0$ *und* $\frac{1}{2\pi}||f||^2_{L^2(0,2\pi)} = \sum_{k=-\infty}^{\infty} |\hat{f}(k)|^2$
- *Die trig. Interpolation ist auch für $p(x) := \sum_{k=-n}^n c_k \exp(ikt)$ mit $p(x_j) = y_j \forall j = -n, ..., n$ eind. lösbar, denn $p(x) = \sum_{l=0}^{2n} c_{l-n} \underbrace{\exp(i(l-n)x)}_{=\exp(ilk)\exp(-inx)\neq 0}$*

3 Extrapolation

Bei einem Interpolationsproblem versucht man einen Funktionswert $\Phi(x)$ zu approximieren mit $x \in [x_0, x_n]$ und $x_0 < ... < x_n$ Stützstellen. Bei der Extrapolation ist der einzige Unterschied, dass $x \notin [x_0, x_n]$.

3.1 Richardson-Extrapolation

Problemstellung: $\Phi : [0, 1] \to \mathbb{K}$ stetig und $\Phi(h)$ berechenbar für $h > 0$, Ziel: Approximiere $\Phi(0)$

Algorithmus 3. *Input: Die stetige Funktion Φ habe eine asympotische Entwicklung*

$$\Phi(h) = \Phi(0) + \sum_{j=1}^n a_jh^{\alpha j} + \mathcal{O}(h^{\alpha(n+1)})$$

für $h > 0$ und $n \in \mathbb{N}$, wobei $\alpha > 0$ bekannt, aber $\Phi(0)$ und $a_1, ..., a_n$ unbekannt.

Prozedur: Berechne $\Phi(h_j)$ für $h_0 > ... > h_n$ und werte das eindeutige $p_n \in \mathbb{P}_n$ mit $p_n(h_j^\alpha) = \Phi(h_j)\forall j = 0, ..., n$ mittels Neville-Verfahren bei $h = 0$ aus, d.h. $\Phi(0) \approx p_n(0)$.

Bemerkung 28. *Meist versucht man kein fixes n, sondern iteriert, bis $p_{n+1}(0) \approx p_n(0)$. Dabei nutzt die Implementierung, dass man das Neville-Verfahren "leicht" um den Knoten $(h_{n+1}^\alpha, \Phi(h_{n+1}))$ erweitern kann.*

Bemerkung 29. *Aus der asymp. Entwicklung folgt, dass $|\Phi(0) - \Phi(h)| = \mathcal{O}(h^\alpha)$, sofern $a_1 \neq 0$, und $\alpha > 0$ ist nicht verbesserbar, d.h. maximale Konvergenzordnung.*

Beispiel 11 (einseitiger Diff.quotient)**.** $\Phi(h) = \frac{f(x+h)-f(x)}{h}$ *für $h > 0$ und $\Phi(0) = f'(x)$.*

Taylor $\implies f(x+h) = f(x) + \sum_{j=1}^{n+1} \frac{f^{(j)}(x)}{j!} h^j + \mathcal{O}(h^{n+2})$

$$\implies \underbrace{\frac{f(x+h)-f(x)}{h}}_{=\Phi(h)} = \underbrace{f'(x)}_{=\Phi(0)} + \underbrace{\sum_{j=2}^{n+1} \frac{f^{(j)}(x)}{j!} h^{j-1}}_{=\sum_{k=1}^{n} \underbrace{\frac{f^{(k+1)}(x)}{(k+1)!}}_{=a_k} h^k} + \mathcal{O}(h^{n+1})$$

und $\alpha = 1$.

Beispiel 12 (zentraler Diff.quotient)**.** $\Phi(h) = \frac{f(x+h)-f(x-h)}{2h}$ *für* $h > 0$ *und* $\Phi(0) = f'(x)$.

Taylor: $f(x \pm h) = \sum_{j=0}^{2n+2} \frac{f^{(j)}(x)}{j!} (\pm h)^j + \mathcal{O}(h^{2n+3})$

$$\implies \Phi(h) = \frac{1}{2n} \left(\sum_{j=0}^{2n+2} \frac{f^{(j)}(x)}{j!} \underbrace{(h^j - (-h)^j)}_{=0 \text{ für } j \text{ gerade}, =2h^j \text{ für } j \text{ ungerade}} + \mathcal{O}(h^{2n+3}) \right) =$$

$$\frac{1}{2n} \left(\sum_{l=1}^{n+1} \frac{f^{(2l-1)}(x)}{(2l-1)!} h^{2(l-1)} + \mathcal{O}(h^{2n+3}) \right) = \underbrace{f'(x)}_{=\Phi(0)} + \underbrace{\sum_{l=2}^{n+1} \frac{f^{(2l-1)}(x)}{(2l-1)!} h^{2(l-1)}}_{=\sum_{k=1}^{n} \frac{f^{(2(k+1)-1)}(x)}{(2(k+1)-1)!} h^{2k} + \mathcal{O}(h^{2(n+1)})} + \mathcal{O}(h^{2n+2})$$

$\implies$ *asymp. Entwicklung mit* $\alpha = 2$.

Algorithmus 4 (Richardson Extrapolation)**.** *Die stetige Funktion* $\Phi : [0,1] \to \mathbb{K}$ *erfülle* $\Phi(h) = \Phi(0) + \sum_{j=1}^{n} a_j h^{\alpha j} + \mathcal{O}(h^{\alpha(n+1)})$ *für* $h > 0$, *wobei* $\alpha > 0$ *bekannt,* $\Phi(0), a_1, ..., a_n$ *unbekannt. Sind für* $h_0 > ... > h_n$ *die Funktionswerte* $\Phi(h_j)$ *bekannt, so sei* $p \in \mathbb{P}_n$ *mit* $p(h_j^\alpha) = \Phi(h_j) \forall j = 0, ..., n$. *Verwende Neville-Verfahren, um* $p(0) \approx \Phi(0)$.

Beispiel 13.
- *einseitiger Diff.quotient* $\alpha = 1$
- *zentraler Diff.quotient* $\alpha = 2$

Beispiel 14 (Romberg-Verfahren)**.** *Sei* $f \in \mathcal{C}[a,b], \Phi(0) = \int_a^b f dx$. $\Phi(h) = \frac{h}{2} \left(f(a) + 2\sum_{j=1}^{n-1} f(x_j) + f(b) \right)$ *mit* $h = \frac{b-a}{n}, x_j := a + jh$, *sog.* ***summierte Trapezregel***

$\implies \Phi(h) = \Phi(0) + \sum_{j=1}^{n} a_j h^{2j} + \mathcal{O}(h^{2(n+1)})$ *gilt, sog.* ***Euler-Maclaurin'sche Summenformel***.

Romberg-Verfahren = *Anwendung von Richardson-Extrapol. auf summierte Trapezregel.*

Satz 13. *Mit* $C > 0, \alpha > 0$ *erfülle* $\Phi : [0,1] \to \mathbb{K}$ *die Entwicklung* $\Phi(h) = \Phi(0) + \sum_{j=1}^{n} a_j h^{\alpha j} + a_{n+1}(h)$ *mit* $|a_{n+1}(h)| \le C h^{\alpha(n+1)} \forall h > 0$.

Es sei $0 < q < 1$ *und* $h_k = q^k$ *für* $k = 0, ..., n$. *Sei* $p \in \mathbb{P}_n$ *mit* $p(h_j^\alpha) = \Phi(h_j) \forall j = 0, ..., n$.

$$\implies |\Phi(0) - p(0)| \le M q^{\alpha \frac{n(n+1)}{2}} \text{ mit } M = C \underbrace{\frac{1}{1-q^\alpha} \exp\left(\frac{2q^\alpha}{(1-q^\alpha)^2} \right)}_{\text{unabhängig vom Restglied}}$$

Bemerkung 30.
- *Unter der Vorraussetzung des Satzes gilt* $|\Phi(0) - \Phi(h_n)| = \mathcal{O}(h_n^\alpha) = \mathcal{O}(q^{\alpha n})$ *bei naiver Realisierung;* $|\Phi(0) - p(0)| = \mathcal{O}(q^{\alpha n^2})$ *durch Extrapolation.* $\implies$ *wesentlich kleineres n nötig, um dieselbe Genauigkeit zu erhalten. Man sagt "Extrapolation mindert Auslöschung".*
- *Achtung: Die Richardson-Extrapolation ist nur sinnvoll, wenn man* $\alpha > 0$ *kennt!*

Proof. (1) zz: $|\Phi(0) - p(0)| \le C \sum_{l=0}^{n} q^{\alpha l(n+1)} |L_l(0)|$ mit $L_l(x) = \prod_{k=0,k\ne l}^{n} \frac{x - x_l}{x_k - x_l}, x_k = h_k^\alpha = q^{\alpha k}$
klar:

$$p(x) = \sum_{l=0}^{n} \underbrace{\Phi(h_l)}_{\text{asymp. Entwicklung}} L_l(x) = \Phi(0) \underbrace{\sum_{l=0}^{n} L_l(x)}_{=1} + \sum_{j=1}^{n} a_j \underbrace{\sum_{l=0}^{n} \overbrace{(h_l^{\alpha j})}^{=x_l^j}}_{=x^j} + \sum_{l=0}^{n} a_{n+1}(h_l) L_l(x) =$$

$$\Phi(0) + \sum_{j=1}^{n} a_j x^j + \sum_{l=0}^{n} a_{n+1}(h) L_l(x)$$

$\implies |p(0)-\Phi(0)| \le \sum_{l=0}^n \underbrace{|a_{n+1}(h_l)|}_{\le Ch_l^{\alpha(n+1)}=Cq^{\alpha l(n+1)}} |L_l(0)|$

(2) zz: $|L_l(0)| = q^{-l\alpha(n+1)} q^{+\frac{\alpha n(n+1)}{2}} \prod_{k=0,k\neq l}^n \frac{1}{|1-q^{(k-l)\alpha}|}$

$$|L_l(0)| = \prod_{k=0,k\neq l}^n \frac{|x_k|}{|x_l-x_k|} = \frac{\prod_{k\neq l} q^{\alpha k}}{\prod_{k\neq l}|q^{\alpha l}-q^{\alpha k}|}$$

- $$\prod_{k\neq l} q^{\alpha k} = q^{-lk}\prod_{k=0}^n q^{\alpha k} = q^{-lk} q^{\alpha \sum_{k=0}^n k} = q^{-lk} q^{\alpha\frac{n(n+1)}{2}}$$

- $$\prod_{k\neq l}|q^{\alpha l}-q^{\alpha k}| = \prod_{k\neq l}(q^{\alpha l}|1-q^{\alpha(k-l)}|) = \underbrace{\left(\prod_{k\neq l} q^{\alpha l}\right)}_{=q^{\alpha l n}} \left(\prod_{k\neq l}|1-q^{\alpha(k-l)}\right)$$

(3) zz: $\prod_{k=0,k\neq l}^n \frac{1}{|1-q^{(k-l)\alpha}|} \le q^{\alpha\frac{l(l+1)}{2}} \prod_{k=1}^n \frac{1}{(1-q^{k\alpha})^2}$
Betrache $\{k-l | k=0,...,n \text{ mit } k\neq l\} = \{-l,...,-1\}\cup\{1,...,n-l\}$

$$\implies \prod_{k=0,k\neq l}^n |1-q^{(k-l)\alpha}| = \underbrace{\left(\prod_{k=1}^{n-l}\underbrace{|1-q^{k\alpha}|}_{=(1-q^{k\alpha})}\right)}_{\ge\prod_{k=1}^n(1-q^{k\alpha})} \left(\prod_{k=1}^{l} \underbrace{|1-q^{-k\alpha}|}_{=|1-\frac{1}{q^{k\alpha}}|=\frac{1-q^{k\alpha}}{q^{k\alpha}}}\right) \ge$$

$$\prod_{k=1}^n(1-q^{k\alpha}) \prod_{k=1}^l q^{-k\alpha} \prod_{k=1}^l (1-q^{k\alpha}) = \prod_{k=1}^n (1-q^{k\alpha})^2 q^{-\alpha\frac{l(l+1)}{2}}$$

(4) zz: $\prod_{k=1}^n \frac{1}{(1-q^{k\alpha})^2} \le \exp\left(\frac{2q^\alpha}{(1-q^\alpha)^2}\right)$

$$\log\left(\prod_{k=1}^n \frac{1}{(1-q^{k\alpha})^2}\right) = 2\sum_{k=1}^n \underbrace{\underbrace{\log\left(\frac{1}{1-q^{k\alpha}}\right)}_{=1+\frac{q^{k\alpha}}{1-q^{k\alpha}}}}_{\le\frac{q^{k\alpha}}{1-q^{k\alpha}}} \le 2\sum_{k=1}^n \frac{q^{k\alpha}}{1-q^{k\alpha}} \le 2\frac{1}{1-q^\alpha}\underbrace{\sum_{k=1}^n q^{k\alpha}}_{\le\frac{q^\alpha}{1-q^\alpha}} \le 2\frac{q^\alpha}{(1-q^\alpha)^2}$$

(5)

$$|p(0)-\Phi(0)| \overset{(1)}{\le} C\sum_{l=0}^n q^{l\alpha(n+1)} \underbrace{|L_l(0)|}_{\overset{(2)}{\le} q^{-l\alpha(n+1)} q^{\alpha\frac{n(n+1)}{2}} \prod_{k\neq l}\frac{1}{|1-q^{(k-l)\alpha}|}} \le$$

$$Cq^{\alpha\frac{n(n+1)}{2}} \sum_{l=0}^n \underbrace{\prod_{k\neq l}\frac{1}{|1-q^{(k-l)\alpha}|}}_{\overset{(3)}{\le} q^{\alpha\frac{l(l+1)}{2}} \underbrace{\prod_{k=1}^n \frac{1}{(1-q^{k\alpha})^2}}_{\overset{(4)}{\le}\exp\left(\frac{2q^\alpha}{(1-q^\alpha)^2}\right)}} \le \left[C\exp\left(\frac{2q^\alpha}{(1-q^\alpha)^2}\right)\underbrace{\sum_{l=0}^n q^{\alpha\frac{l(l+1)}{2}}}_{\le\frac{1}{1-q^\alpha}}\right] q^{\alpha\frac{n(n+1)}{2}}$$

$\square$

3.2 Aitken'sches Δ^2-Verfahren

Beim Δ^2-Verfahren handelt es sich um ein Verfahren zur **Konvergenzbeschleunigung**, d.h. sei $(x_n)_{n\in\mathbb{N}\subset\mathbb{K}}$ eine bekannte, konvergente Folge mit unbekanntem Limes $x = \lim\limits_{n\to\infty} x_n$

Ziel: Konstruiere eine Folge $(y_n)_{n\in\mathbb{N}}$ mit $\lim\limits_{n\to\infty} \frac{x-y_n}{x-x_n} = 0$, d.h. (y_n) konvergiert schneller gegen x.

Bemerkung 31. *Sei $(x_n)_{n\in\mathbb{N}}$ eine geometrisch konvergente Folge mit Limes x, d.h. ex. $q \in \mathbb{K}$ mit $|q| < 1$ und $x - x_{n+1} = q(x - x_n)\forall n \in \mathbb{N}$*

klar: $\lim\limits_{n\to\infty} x_n = x$

Betrachte Differenzenoperator $\Delta y_n := y_{n+1} - y_n$

zz: $x = x_n - \frac{(\Delta x_n)^2}{\Delta^2 x_n} = x_n - \frac{(x_{n+1}-x_n)^2}{\Delta(x_{n+1}-x_n)} = x_n - \frac{(x_{n+1}-x_n)^2}{x_{n+2}-2x_{n+1}+x_n}$, d.h. Limes x kann aus 3 Folgegliedern x_n, x_{n+1}, x_{n+2} exakt berechnet werden.

Proof.

$$x_{n+2} - 2x_{n+1} + x_n = (x_{n+2} - x) - 2(x_{n+1} - x) + (x_n - x) = \underbrace{[q^2 - 2q + 1]}_{\neq 0} \underbrace{(x_n - x)}_{\neq 0 \text{ o.B.d.A}}$$

$$\left(x_n - \frac{(\Delta x_n)^2}{\Delta^2 x_n}\right) - x = (x_n - x) - \frac{(q-1)^2(x_n - x)^2}{(q-1)^2(x_n - x)} = 0$$

$\square$

Satz 14. *Sei $(x_n)_{n\in\mathbb{N}} \subset \mathbb{K}$ mit $x_n \neq x \in \mathbb{K}$ und $x_{n+1} - x = (q + \delta_n)(x_n - x)\forall n \in \mathbb{N}$ mit $q \in \mathbb{K}, |q| < 1, (\delta_n)_{n\in\mathbb{N}} \subset \mathbb{K}$ mit $\lim_n \delta_n = 0$.*

$\Longrightarrow$

1. $\lim_n x_n = x$
2. *Ex. $n_0 \in \mathbb{N}$, sodass $y_n := x_n - \frac{(\Delta x_n)^2}{\Delta^2 x_n} \in \mathbb{K}$ wohldef. für $n \geq n_0$*
3. $\lim_n \frac{x-y_n}{x-x_n} = 0$

Proof. (i) Sei $0 < |y| < \kappa < 1$. Wegen $\lim_n \delta_n = 0$, ex. $\tilde{n}_0 \in \mathbb{N}$ mit $|q + \delta_n| \leq \kappa \forall n \geq \tilde{n}_0$

$\Longrightarrow |x_{n+1} - x| \leq \underbrace{|q + \delta_n|}_{\leq\kappa} |x_n - x| \Longrightarrow \lim_n |x_{n+1} - x| = 0 \Longrightarrow x = \lim x_n$

(ii)

$$x_{n+2} - 2x_{n+1} + x_n = (x_n - x) \quad \underbrace{[(q + \delta_{n+1})(q + \delta_n) - 2(q + \delta_n) + 1]}_{=\underbrace{(q-1)^2}_{\neq 0} + \underbrace{(\delta_n\delta_{n+1} + q(\delta_n + \delta_{n+1} - 2\delta_n))}_{=:\epsilon_n \to 0}}$$

(iii)

$$y_n - x = (x_n - x) - \frac{(x_{n+1} - x_n)^2}{(x_n - x)[(q-1)^2 + \epsilon_n]} =$$

$$(x_n - x) - \frac{(q + \delta_n - 1)^2(x_n - x)^2}{(x_n - x)[(q-1)^2 + \epsilon_n]} = (x_n - x)\left(1 - \frac{(q - 1 + \delta_n)^2}{(q-1)^2 + \epsilon_n}\right)$$

$$\Longrightarrow \frac{y_n - x}{x_n - x} = 1 - \frac{(q - 1 + \delta_n)}{(q-1)^2 + \epsilon_n} \to 0$$

$\square$

Bemerkung 32. *Die Vorraussetzungen von Aitken sind defakto für jedes numerische Verfahren erfüllt, d.h. bevor man nichts macht und naiv (x_n) betrachtet, macht man immer Aitken.*

Beispiel 15 (einseitiger Diff.quot.)**.** $f \in \mathcal{C}^2(\mathbb{R}), z \in \mathbb{R}, h_n > 0$

Taylor $\Longrightarrow f(z + h) = f(z) + h_n f'(z) + \frac{h_n^2}{2} f''(\zeta_n)$ *mit* $z < \zeta_n < z + h_n$

$$\Longrightarrow \underbrace{f'(z)}_{=:x} - \underbrace{\frac{f(z+h) - f(z)}{h_n}}_{=:x_n} = h_n\left(-\frac{f''(\zeta_n)}{2}\right) = h_n\left(-\frac{f''(z)}{2} + \underbrace{\frac{f''(z) - f''(\zeta_n)}{2}}_{=:\epsilon_n}\right)$$

$\epsilon_n \to 0$ *für* $h_n \to 0$

In der Praxis $h_n := 2^{-n} h_0$.

Für $f''(x) \neq 0$

$$\implies x - x_{n+1} = h_{n+1}\left(-\frac{f''(z)}{2} + \epsilon_{n+1}\right) = \underbrace{\frac{1}{2}}_{=:q} \underbrace{\frac{-\frac{f''(z)}{2} + \epsilon_{n+1}}{-\frac{f''(z)}{2} + \epsilon_n}}_{=1+\underbrace{\frac{\epsilon_{n+1} - \epsilon_n}{-\frac{f''(z)}{2} + \epsilon_n}}_{=:\frac{\delta_n}{q}}} (x - x_n)$$

4 Numerische Integration

Im ganzen Kapitel seien $a, b \in \mathbb{R}$ mit $a < b$. Ferner sei $w \in L^1(a,b)$ eine Gewichtsfunktion mit $w(x) > 0$ fast überall.

Ziel: Approximiere $Qf := \int_a^b fwdx$ für $f \in \mathcal{C}[a,b]$

Bemerkung 33. *Soll eine Funktion* $g \in L^1(a,b)$ *numerisch integriert werden, so zerlegt man* $g = fw$ *mit* f *dem glatten Anteil von* g *und* w *dem singulären Anteil, z.B.* $g(x) = \frac{\sin x}{\sqrt{1-x^2}}$*, dann* $f(x) = \sin(x), w(x) = \frac{1}{\sqrt{1-x^2}}$ *analog* $g(x) = x \log x$*, dann* $f(x) = x, w(x) = \log x$.

4.1 Quadraturformeln

Definition 10. *Gegeben seinen* ***Stützstellen*** *(oder:* ***Quadraturknoten****)* $a \leq x_0 < ... < x_n \leq b$ *und* ***Gewichte*** $w_0, ..., w_n \in \mathbb{K}$. *Dann bezeichnet man* $Q_n f = \sum_{j=0}^n w_j f(x_j)$ *als* ***Quadraturformel (der Länge*** n***).*** Q_n *hat* ***Exaktheitsgrad*** $m \in \mathbb{N}_0$*, gdw.* $Q_n p = Qp$ *für alle* $p \in \mathbb{P}_m$*, d.h. Polynome vom Grad* m *werden exakt integriert.*

Lemma 8.

1. Q, Q_n *sind linear und stetig auf* $\mathcal{C}[a,b]$ *mit Operatornorm* $||Q|| := \sup_{f \in \mathcal{C}[a,b], f \neq 0} \frac{|Qf|}{||f||_{L^\infty(a,b)}} = ||w||_{L^1(a,b)}$, $||Q_n|| = \sum_{j=0}^n |w_j|$
2. *Der Exaktheitsgrad von* $Q_n \leq 2n+1$
3. *Ist* Q_n *exakt auf* $\mathbb{P}_{2n+1}$*, so gibt es kein* $p \in \mathbb{P}_{2n+2} \setminus \mathbb{P}_{2n+1}$ *mit* $Q_n p = Qp$
4. *Ist* Q_n *exakt auf* $\mathbb{P}_0$*, so gilt* $\sum_{j=0}^n w_j = ||w||_{L^1(a,b)}$
5. *Für* $w(x) = 1$ *und* Q_n *exakt auf* $\mathbb{P}_1$*, so gilt* $\sum_{j=0}^n w_j = b - a, \sum_{j=1}^n w_j x_j = \frac{b^2 - a^2}{2}$

Bemerkung 34. *Oft verwendet man banale Identitäten wie (iv), (v) um zu testen, ob eine Quadratur korrekt implementiert ist.*

Proof. (i) klar: $|Qf| \leq ||f||_{L^\infty(a,b)} ||w||_{L^1(a,b)}$

$\implies ||Q|| \leq ||w||_{L^1(a,b)}$

klar: $|Q_n f| \leq \sum_{j=0}^n |w_j| |f(x_j)| \leq ||f||_{L^\infty(a,b)} \sum_{j=0}^n |w_j|$

$\implies ||Q_n|| \leq \sum_{j=0}^n |w_j|$

Wähle einen Polygonzug $f \in \mathcal{C}[a,b]$ mit $||f||_{L^\infty(a,b)} \leq 1$ mit $w_j f(x_j) = |w_j|$, d.h. $f(x_j) = sign w_j$ also $|f(x_j)| \leq 1$

$\implies Q_n f = \sum_{j=0}^n w_j f(x_j) \implies ||Q_n|| = \sum_{j=0}^n |w_j|$

(ii) Wähle $p(x) = \prod_{j=0}^n (x - x_j)^2$. $p \in \mathbb{P}_{2n+2}$ und $p > 0$ fü

$\implies Qf = \int_a^b \underbrace{pw}_{>0 \text{ fü}} dx > 0 = Q_n p$

$\implies p$ wird nicht exakt integriert $\implies$ Exaktheit$(Q_n) \leq 2n+1$

(iii) $R := Q - Q_n$ linear. Sei $\{p_0, ..., p_{2n+1}\} \subseteq \mathbb{P}_{2n+1}$ Basis. Falls $p_{2n+2} \in \mathbb{P}_{2n+2} \setminus \mathbb{P}_{2n+1}$, so ist $\{p_0, ..., p_{2n+2}\} \subseteq \mathbb{P}_{2n+2}$ Basis.

Es gilt $R = 0$ auf $\mathbb{P}_{2n+2}$ gdw. $R(p_j) = 0 \forall j = 0, ..., 2n+2$

(iv) $||w||_{L^1(a,b)} = Q1 = Q_n 1 = \sum_{j=0}^n w_j$

(v) Für $w = 1$ gilt $||w||_{L^1(a,b)} = b - a$, $\underbrace{Qx}_{=\int_a^b xdx = \frac{b^2}{2} - \frac{a^2}{2}} = Q_n x = \sum_{j=0}^n w_j x_j$, da Q_n exakt auf $\mathbb{P}_1$ und x ein Monom $\in \mathbb{P}_1$. □

Achtung: Meistens verwendet Quadratur den Laufindex $j = 0, ..., n$ (d.h. $n+1$ Stützstellen, Exaktheit $\leq 2n+1$). Manchmal wird aber $j = 1, ..., n$ betrachtet, d.h. n Stützstellen, Exaktheit $\leq 2n-1$.

Bemerkung 35. *In der Literatur (z.B. Abramowitz oder Secrest-Strand) sind Quadraturformeln auf Standardintervallen tabelliert, z.B.* $[0,1], [-1,1]$. *Um Quadraturformeln auf* $[a,b]$ *zu erhalten, verwendet man in der Regel eine affine Transformation:*

$$\Phi : [-1,1] \to [a,b], \Phi(t) = \frac{1}{2}\{a+b+t(b-a)\}$$

$$\implies \int_a^b fwdx = \int_{-1}^1 f(\Phi(x)) \underbrace{w(\Phi(x))}_{=:\tilde{w}(x)} \underbrace{|\det D\Phi(x)|}_{\frac{b-a}{2}} dx =$$

$$\frac{b-a}{2}\tilde{Q}(f\circ\Phi) \approx \frac{b-a}{2}\tilde{Q}_n(f\circ\Phi) =$$

$$\sum_{j=0}^n \underbrace{\frac{b-a}{2}\tilde{w}_j}_{=:w_j} f(\underbrace{\Phi(\tilde{x}_j)}_{=:x_j}) = \sum_{j=0}^n w_j f(x_j) =: Q_n f$$

analog für $\Phi : [0,1] \to [a,b], \Phi(t) = a + t(b-a)$
klar: Falls $f \in \mathbb{P}_m$, *dann* $f\circ\Phi \in \mathbb{P}_m \implies$ *Exaktheit(*$\tilde{Q}_n$*) = Exaktheit(*Q_n*)*

Satz 15 (Fehlerabschätzung + Konvergenz)**.** *Sei* $Q_n f = \sum_{j=0}^n w_j^{(n)} f(x_j^{(n)})$ *eine Quadraturformel der Länge* n *und Exaktheit* m.

$$\implies |Qf - Q_n f| \le (||w||_{L^1(a,b)} + \sum_{j=0}^n |w_j^{(n)}|) \min_{p\in\mathbb{P}_m} ||f-p||_{L^\infty(a,b)}$$

Ferner sind äquivalent:

- $Qf = \lim\limits_{n\to\infty} Q_n f \forall f \in \mathcal{C}[a,b]$
- $Qp = \lim\limits_{n\to\infty} Q_n p \forall p \in \mathbb{P} := \bigcup_{n\in\mathbb{N}} \mathbb{P}_n$ *und* $\sup_{n\in\mathbb{N}} \sum_{j=0}^n |w_j^{(n)}| < \infty$.

Proof. Sei $p \in \mathbb{P}_m$ mit $||f-p||_{L^\infty(a,b)} = \min_{\tilde{p}\in\mathbb{P}_m} ||f-\tilde{p}||_{L^\infty(a,b)}$

$$|Qf - Q_n f| < |Qf - Qp| + |Q_n p - Qf| \le ||Q|| ||f-p||_{L^\infty(a,b)} + ||Q_n|| ||f-p||_{L^\infty(a,b)}$$

zz: (ii $\implies$ i)
Sei $\epsilon > 0$. Nach Weierstrass ex. $m \in \mathbb{N}$ und $p \in \mathbb{P}_m$ mit $||f-p||_{L^\infty(a,b)} \le \epsilon$

$$\implies |Qf - Q_n f| \le \underbrace{|Qf - Qp|}_{\le ||Q|| \underbrace{||f-p||_{L^\infty(a,b)}}_{\le\epsilon}} + \underbrace{|Qp - Q_n p|}_{\to 0 \text{ für } n\to\infty} + \underbrace{|Q_n p - Q_n f|}_{\le \underbrace{||Q_n||}_{\le M} \underbrace{||f-p||_{L^\infty(a,b)}}_{\le\epsilon}} \le (||w||_{L^1(a,b)} + M)\epsilon + |Qp - Q_n p|$$

$$\implies \limsup_n |Qf - Q_n f| \le (||w||_{L^1(a,b)} + M)\epsilon \forall \epsilon > 0$$

$$\implies 0 \le \liminf_n |Qf - Q_n f| \le \limsup_n |Qf - Q_n f| = 0$$

(i $\implies$ ii) mittels **Satz von Banach-Steinhaus**:
X, Y Banach-Räume, $T_n \in L(X,Y)$
Dann sind äquivalent:

- $sup_{n\in\mathbb{N}} ||T_n||_{L(X,Y)} < \infty$ glm. Beschränktheit
- $\forall x \in X : \sup_{n\in\mathbb{N}} ||T_n x||_Y < \infty$ pktw. Beschränktheit

jetzt $X \in \mathcal{C}[a,b], Y = \mathbb{K}, T_n = Q_n$
d.h. $(Q_n f)_{n\in\mathbb{N}}$ konvergent nach (i), also punktweise Beschränktheit $\implies \infty > \sup_n ||Q_n|| = \sup_n \sum_{j=0}^n |w_j^{(n)}|$. □

Bemerkung 36.

1. *Die Implikation (i* $\implies$ *ii) ist nur mathematisch interessant. Der Satz von Banach-Steinhaus ist einer der Fundamentalsätze der Funktionalanalysis.*
2. *Die Implikation (ii* $\implies$ *i) ist praktisch relevant. Die Eigenschaft* $\lim_n Q_n p = Qp \forall p \in \mathbb{P}$ *gilt für alle interpolatorischen Quadraturformeln.*
3. *Die zentrale Bedingung* $\sup_n \sum_{j=0}^n |w_j^{(n)}| < \infty$ *ist kritisch, aber für alle Gauss-Quadraturen erfüllt.* $\implies$ *Konvergenz gilt immer für Gauss-Quadratur.*

4.2 Interpolatorische Quadraturformeln

Definition 11. *Zu gegebenen Stützstellen $a \le x_0 < ... < x_n \le b$ bezeichnet man $Q_n f := \sum_{j=0}^n \underbrace{\left(\int_a^b L_j w dx\right)}_{=:w_j = Q(L_j)} f(x_j)$ als* ***interpolatorische Quadraturformel*** *(oder* ***Interpolationsquadratur****), wobei $L_j x := \prod_{k=0}^n \frac{x-x_k}{x_j - x_k}$ die Lagrange-Polynome sind*

Satz 16. *Für $Q_n f = \sum_{j=0}^n w_j f(x_j)$ sind äquivalent:*

- *Q_n ist interpolatorisch*
- *Für $f \in \mathcal{C}[a,b]$ mit Lagrange-Interpolationspolynom $p \in \mathbb{P}_n$ (d.h. $p(x_j) = f(x_j) \forall j = 0, ..., n$) gilt $Q_n f = Qp$*
- *Exaktheitsgrad(Q_n)$\ge n$.*

Proof. (i) $\iff$ (ii).
Betrachte $p = \sum_{j=0}^n f(x_j) L_j \implies Qp = \sum_{j=0}^n f(x_j) \int_a^b L_j w dx$
(i $\implies$ iii) klar $Q_n(L_j) = Q(L_j)$, da $L_j(x_k) = \delta_{jk}$
$\implies R = Q - Q_n$ ist Null auf Basis $\{L_0, ..., L_n\} \implies R = 0$ auf $\mathbb{P}_n \implies$ Exaktheit(Q_n) $\ge n$.
(iii $\implies$ i) ✓ □

Bemerkung 37. *Die Gewichte einer Interpolationsquadratur kann man durch Lösen eines lin. GLS berechnen. Sei $\{p_0, ..., p_n\} \subseteq \mathbb{P}_n$ Basis, sei $Q_n f := \sum_{j=0}^n w_j f(x_j)$ eine Int.quadratur*

$$\implies \underbrace{\begin{pmatrix} p_0(x_0) & ... & p_0(x_n) \\ \vdots & & \vdots \\ p_n(x_0) & ... & p_n(x_n) \end{pmatrix}}_{\text{Transponierte der Vandermonde-Matrix aus der Interpolation, also regulär}} \begin{pmatrix} w_0 \\ \vdots \\ w_n \end{pmatrix} = \begin{pmatrix} \int_a^b p_0 w dx \\ \vdots \\ \int_a^b p_n w dx \end{pmatrix}$$

Beispiel 16.

- ***abgeschlossene Newton-Cotes-Formeln*** *$x_j := a + j\frac{b-a}{n}$ für $j = 0, ..., n$*
- ***affine Newton-Cotes-Formeln*** *$x_j := a + (j+1)\frac{b-a}{n+2}$ für $j = 0, ..., n$*
- ***Moolavrin-Formeln*** *$x_j := a + (j + \frac{1}{2})\frac{b-a}{n+1}$*
- ***Clemshaus-Curtis-Formeln*** *Verwende die Extrema oder die Nullstellen des Cebysev-Polynoms.*

Bemerkung 38. *Ein Satz von Kusmin besagt, dass für äquidistante Stützstellen immer gilt $\sum_{j=0}^n |w_j^{(n)}| \to \infty$, d.h. es gibt stetige Funktionen $f \in \mathcal{C}[a,b]$ mit $Q_n f \nrightarrow Qf$ für $n \to \infty$.*
Bei Clemshaus-Curtis kann man $w_j^{(n)} \ge 0$ zeigen, also $\sum_{j=0}^n |w_j^{(n)}| \overset{!}{=} \sum_{j=0}^n w_j^{(n)} = b - a$
$\implies$ Konvergenz $Q_n f \to Qf \forall f \in \mathcal{C}[a,b]$

Bemerkung 39. *Einige der abg. Newton-Cotes-Formeln haben auch eigene Namen.*

- *$n = 1$ Trapezregel,*
- *$n = 2$ Simpson-Regel,*
- *$n = 3$ Newton'sche $\frac{3}{8}$-Regel,*
- *$n = 4$ Milne-Regel*

Bemerkung 40. *Man verwendet die NC-Formeln in der Parix nur für $n \le 8$, da für $n \ge 9$ negative Gewichte auftreten, was zur Auslöschung führt.*

Bemerkung 41. *Sei $\tilde{Q}_n$ eine Quadratur auf $[0,1]$ zu $w = 1$. Sei Q_n^j die induzierte Quadratur auf $[a_j, b_j]$ mit $a = a_0 < b_0 = a_1 < b_1 = ... < b_n = b$. Mit der Zerlegung*

$$Qf = \int_a^b f dx = \sum_{j=0}^N \underbrace{\int_{a_j}^{b_j} f dx}_{=\approx Q_n^j f}$$

erhalte eine sogenannte ***summierte Quadraturformel*** *$Q_{nN} f := \sum_{j=0}^N Q_n^j f$.*

Falls Exaktheitsgrad($\tilde{Q}_n$)≥ 0 und $\max_{j=0,...,N}(b_j^N - a_j^N) \to 0$, so folgt $Q_{nN} f \to Qf$ für $N \to \infty, \forall f \in \mathcal{C}[a,b]$ und sie kriegen sogar a-priori Fehlerabschätzungen!

Satz 17. *Sie Stützstellen seien symmetrisch, d.h.* $x_j = a+b-x_{n-j} \forall j = 0, ..., n$*. Das Gewicht sei symmetrisch, d.h.* $w(x) = w(a+b-x) \forall x \in [a,b]$
$\implies$

1. *Die Gewichte der zugehörigen Interpolationsquad.* Q_n *sind symmetrisch, d.h.* $w_j = w_{n-j} \forall j = 0, ..., n$
2. *Falls* n *gerade, so gilt Exaktheit(*Q_n*)*$\geq n+1$

Proof. (i) Betrachte $\tilde{Q}_n f := \sum_{j=0}^n w_{n-j} f(x_j)$
zz: Exaktheit($\tilde{Q}_n$)$\geq n$ (dann $\tilde{Q}_n$ interpolatorisch, d.h. $\{w_j\}$ eindeutig durch $\{x_j\}$)
Betrachte $p_k(x) := \left(x - \frac{a+b}{2}\right)^k$, $\tilde{p}_k(x) := p_k(a+b-x) = \left(\frac{a+b}{2} - x\right)^k = (-1)^k p_k(x)$

$$\tilde{Q}_n p_k = \sum_{j=0}^n w_{n-j} p_k(x_j) = \sum_{l=0}^n w_l \overbrace{p_k(\underbrace{x_{n-l}}_{=a+b-x_l})}^{=\tilde{p}_k(x_l)} = Q_n \tilde{p}_k = Q\tilde{p}_k \forall k = 0, ..., n$$

$$Q\tilde{p}_k = \int_a^b p_k(a+b-x) \underbrace{w(x)}_{=w(a+b-x)} dx = \int_a^b p_k(y) w(y) dy = Qp_k$$

$\implies \tilde{Q}_n p_k = Qp_k \forall k = 0, ..., n \implies \tilde{Q}_n$ interpolatorisch.
(ii) zz: $Q_n p_{n+1} = Qp_{n+1} = 0$, $x_{\frac{1}{2}} = \frac{a+b}{2}$

$$Q_n p_{n+1} = \sum_{j=0}^n w_j p_{n+1}(x_j) = \sum_{j=1}^{\frac{n}{2}} w_{\frac{n}{2}-j} (x_{\frac{n}{2}-j} - \frac{a+b}{2})^{n+1} + \sum_{j=1}^{\frac{n}{2}} \underbrace{w_{\frac{n}{2}+j}}_{w_{\frac{n}{2}-j}} \left(\underbrace{x_{\frac{n}{2}+j} - \frac{a+b}{2}}_{=\frac{a+b}{2} - x_{\frac{n}{2}-j}} \right)^{\overbrace{n+1}^{\text{ungerade}}} = 0$$

$$Qp_{n+1} = \int_a^b p_{n+1} w dx = -\int_a^b \underbrace{\tilde{p}_{n+1}(x)}_{p_{n+1}(a+b-x)} \underbrace{w(x)}_{=w(a+b-x)} dx = -\int_a^b p_{n+1}(y) w(y) dy = -Qp_{n+1}$$

$\implies Qp_{n+1} = 0$ ☐

Korollar 3 (Konkrete Fehlerabschätzungen)**.** *Sei* $w(x) = 1$, $C_{\mathbb{K}} = 1$ *für* f *reellwertig,* $C_{\mathbb{K}} = \sqrt{2}$ *für* f *komplexwertig.*

1. *Die Trapezregel* $Q_1 f = \frac{b-a}{2}(f(a) + f(b))$ *erfüllt* $|Qf - Q_1 f| \leq C_{\mathbb{K}} \frac{(b-a)^3}{12} ||f''||_{L^\infty(a,b)} \forall f \in \mathcal{C}^2[a,b]$
2. *Die Simpson-Regel* $Q_2 f = \frac{b-a}{6}\left(f(a) + 2f\left(\frac{a+b}{2}\right) + f(b)\right)$ *erfüllt* $|Qf - Q_2 f| \leq C_{\mathbb{K}} \frac{(b-a)^5}{2880} ||f^{(4)}||_{L^\infty(a,b)} \forall f \in \mathcal{C}^4[a,b]$

Proof. Q_1, Q_2 sind abg. Newton-Cotens-Formeln, also interpolatorisch.

1. Sei $p \in \mathbb{P}_1$ mit $p(a) = f(a), p(b) = f(b)$

$$\implies |Qf - Q_1 f| = |Qf - Qp| \leq \int_a^b \underbrace{|f(x) - p(x)|}_{\leq C_{\mathbb{K}} \frac{||f'||_{L^\infty(a,b)}}{2!} |x-a||x-b|} \leq C_{\mathbb{K}} \frac{||f'||_{L^\infty(a,b)}}{2!} \underbrace{\int_a^b (x-a)(b-x) dx}_{=\frac{(b-a)^3}{6}}$$

2. Sei $p \in \mathbb{P}_3$ mit $p(a) = f(a), p(b) = f(b), p\left(\frac{a+b}{2}\right) = f\left(\frac{a+b}{2}\right)$ und $p'\left(\frac{a+b}{2}\right) = f'\left(\frac{a+b}{2}\right)$

$$\implies |Qf - Q_2 f| = |Qf - \underbrace{Q_2}_{=Qp}| = |Qf - Qp| \leq \int_a^b |f(x) - p(x)| dx \leq$$

$$C_{\mathbb{K}} \frac{||f^{(4)}||_{L^\infty(a,b)}}{4!} \overbrace{\int_a^b (x-a)(b-x) \left(\frac{a+b}{2} - x\right)^2 dx}^{=\frac{(b-a)^5}{120}}$$

☐

Bemerkung 42. *Auch mit Fehlerabschätzung aus Konvergenzsatz + Abschätzung des Bestapprox.fehlers durch Interpolationsfehler bekommt man konkrete Fehlerabschätzungen, allerdings schlechtere Konstanten -¿ Trapezregel* $\frac{1}{4}$ *statt* $\frac{1}{12}$*.*

4.3 Gauss-Quadratur

Ziel: Konstruiere (eindeutige) Quadraturformel Q_n mit Exaktheit(Q_n)= $2n+1$
klar: Q_n muss interpolatorisch sein.

Bemerkung 43. *Die Analysis in diesem Abschnitt geht auch für ein unbeschränktes Intervall, z.B.* $(0,\infty),(-\infty,\infty)$ *sofern* $\int_a^b |x|^n w(x) < \infty \forall n \in \mathbb{N}_0$.

Betrachte Innenproduktraum $H := \{f : (a,b) \to \mathbb{R}$ *integrierbar* $|||f|| < \infty\}$ *mit* $||f|| =< f,f >^{\frac{1}{2}}, < f,g >= \int_a^b fgwdx$

klar: $< .,. >$ *Skalarprodukt (-¿* $L^2(a,b;wdx) = H$*)*

Lemma 9 (Gram-Schmidt-Orthogonalisierung)**.** *Sei* $(x^n)_{n\in\mathbb{N}}$ *die Folge der Monome in* H*. Definiere induktiv* $p_0 := x^0 = 1, p_n := x^n - \sum_{k=0}^{n-1} \frac{<x^n,p_k>}{||p_n||^2} p_k$

$\Longrightarrow$ $(p_n)_{n\in\mathbb{N}}$ *sind orthogonal bzgl.* $< .,. >$ *und insb.* $\{p_0,...,p_n\} \subseteq \mathbb{P}_n$ *Basis und alle* p_n *haben Leitkoeffizient 1. Die Polynome* p_n *heißen* ***Orthogonalpolynome****.*

Bemerkung 44. *Für* $(a,b) = (-1,1), w = 1$ *erhält man die* ***Legendre-Polynome****. Für* $(a,b) = (-1,1), w(x) = \frac{1}{\sqrt{1-x^2}}$ *erhält man* ***Cebysev-Polynome****. Für* $(a,b) = (0,\infty)$ *und* $w(x) = e^{-x}$ *erhält man* ***Lagrange-Polynome****.*

Lemma 10. *Es seinen* $x_0,...,x_n \in \mathbb{C}$ *die gemäß Vielfachheit gezählten Nullstellen des Orth.pl.* $p_{n+1} \in \mathbb{P}_{n+1}$ $\Longrightarrow$

1. *alle Nullstellen sind einfach und liegen in* (a,b)
2. *Mit den Lagrange-Polynomen* $L_j(x) = \prod_{k=0}^n \frac{x-x_k}{x_j-x_k}$ *gilt* $x_j = \frac{<xL_j,L_j>}{||L_j||^2}$

Proof. 1. Seien $x_0,...,x_k \in (a,b)$ alle Nullstellen von p_{n+1}, die ungeraden Vielfachheit haben und in (a,b) liegen, bzw. $k=-1$, falls keine solchen existieren.

$q(x) := \prod_{j=0}^k (x-x_j), q \in \mathbb{P}_{k+1}$

$\Longrightarrow r := qp_{n+1} \neq 0$ hat nur Nst. gerade Vielfachheit in $(a,b) \Longrightarrow r \geq 0$ in (a,b) oder $r \leq 0$ in (a,b)

Annahme: $k < n \Longrightarrow 0 =< q,p_{n+1} >= \int_a^b rwdx \Longrightarrow rw = 0$ f.ü. Widerspruch zu $r \neq 0 \neq w$ fast überall. Also $k = n$.

2. Polynomdivision $p_{n+1} = (x-x_j)q$ mit $q \in \mathbb{P}_n$

$\Longrightarrow \underbrace{< p_{n+1},q >}_{=0} =< xq,q > -x_j < q,q > \Longrightarrow x_j = \frac{<xq,q>}{<q,q>}$ und $q = \prod_{k=0,k\neq j}^n (x-x_k) = cL_j$

□

Satz 18 (Existenz + Eindeutigkeit der Gauss-Quadratur)**.**

1. *Ex. eind Quadraturformel* $Q_nf = \sum_{j=0}^n w_jf(x_j)$ *mit Exaktheitsgrad(*Q_n*)=* $2n+1$
2. *Die Knoten sind die Nullstellen von Orth.pol.* $p_{n+1} \in \mathbb{P}_{n+1}$
3. *Die Gewichte erfüllen* $w_j = \int_a^b wL_jdx = \int_a^b wL_j^2dx > 0$
4. $|Qf - Q_nf| \leq C_{\mathbb{K}} \frac{||f^{(2n+2)}||_{L^\infty(a,b)}}{(2n+2)!} \int_a^b w(x) \prod_{j=0}^n (x-w_j)^2 dx \forall f \in \mathcal{C}^{2n+2}(a,b)$

Proof. 1. Existenz: Wähle Nullstellen $x_0,...,x_n$ von p_{n+1}, definiere $w_j = \int_a^b wL_jdx$

$\Longrightarrow$ Exaktheitsgrad(Q_n)$\geq n$

Sei $q \in \mathbb{P}_{2n+1}$. zz: $Q_nq = Qq$

Polynomdivision $\Longrightarrow q = p_{n+1}\alpha + \beta$ mit $\alpha,\beta \in \mathbb{P}_n$

$$Q_nq = Q_n\beta = Q\beta =< \beta,1 > + \underbrace{< p_{n+1},\alpha >}_{=0} =< \underbrace{\beta + \alpha p_{n+1}}_{=q}, 1 >= Qq$$

2. Eindeutigkeit: Sei $\tilde{Q}_nf = \sum_{j=0}^n \tilde{w}_jf(\tilde{x}_j)$ eine weitere Quadraturformel mit Exaktheit($\tilde{Q}_n$)= $2n+1$

zz: $\tilde{x}_j \in \{x_0,...,x_n\} \forall j = 0,...,n$

(dann folgt $\{\tilde{x}_0,...,\tilde{x}_n\} = \{x_0,...,x_n\}$ und damit $\tilde{Q}_n = Q_n$)

Sei $j \in \{0, ..., n\}$.

$$q(x) := \left(\prod_{k=0}^{n}(x - x_n)\right)\left(\prod_{k=0,k\neq j}^{n}(x - \tilde{x}_k)\right) \in \mathbb{P}_{2n+1}$$

$$\implies 0 = Q_n q = Qq = \tilde{Q}_n q = \underbrace{\tilde{w}_j}_{\neq 0} q(\tilde{x}_j)$$

$$q(\tilde{x}_j) = \left(\prod_{k=0}^{n}(\tilde{x}_j - x_k)\right)\underbrace{\left(\prod_{k=0,k\neq j}^{n}(\tilde{x}_j - \tilde{x}_k)\right)}_{\neq 0}$$

$$\implies \tilde{x}_j \in \{x_0, ..., x_n\}.$$

3. $w_j = \int_a^b wL_j^2 dx > 0$

$$w_j = \int_a^b L_j w dx = Q(L_j) = Q_n(L_j) = \sum_{k=0}^{n} w_k \underbrace{L_j(x_k)^2}_{=\delta_{jk}} = Q_n(L_j^2) = Q(L_j^2) = \int_a^b L_j^2 w dx$$

4. zz. Fehlerabschätzung

Wähle $q \in \mathbb{P}_{2n+1}$ mit $\underbrace{q(x_j) = f(x_j)}_{\text{um interpolatorisch}}$, $\underbrace{q'(x_j) = f'(x_j)}_{\text{zusätzliche Freiheit für Verbesserung}}$ $\forall j = 0, ..., n$

$$\implies |f(x) - q(x)| \leq C_{\mathbb{K}} \frac{||f^{(2n+2)}||_{L^\infty(a,b)}}{(2n+2)!} \prod_{j=0}^{n}(x - x_j)^2$$

$$\implies |Qf - \underbrace{Q_n f}_{=Q_n q = Qq}| = |Q(f - q)| \leq \int_a^b |f(x) - q(x)| w(x) dx$$

□

Lemma 11 (3-Term-Rekursion). *Die Orth.pol. $(p_n)_{n\in\mathbb{N}_0}$ erfüllen $p_0(x) = 1, p_1(x) = x - \beta_0, p_{n+1}(x) = (x - \beta_n)p_n(x) - \gamma_n^2 p_{n-1}(x) \forall n \geq 1$ mit reellen Koeff. $\beta_n = \frac{<xp_n,p_n>}{||p_n||^2}, \gamma_n = \frac{||p_n||}{||p_{n-1}||}$*

Proof. durch Induktion nach n.
Ind.anf. $n = 0, 1$:
Erinnerung $p_0(x) = x^0 = 1, p_n(x) = x^k - \sum_{j=0}^{k-1} \frac{<x^k,p_j>}{||p_j||^2} p_j = x - \frac{<x^1,p_0>}{||p_0||^2} p_0 = x - \frac{<x,1>}{||1||^2} 1$
Def $q_{n+1}(x) = (x - \beta_n)p_n(x) - \gamma_n^2 p_{n-1}(x) \in \mathbb{P}_{n+1}$, Leitkoeff($q_{n+1}$)= 1 =Leitkoeff($q_{n-1}$)
$\implies p_{n+1} - q_{n+1} \in \mathbb{P}_n, < p_{n+1}, q >= 0 \forall q \in \mathbb{P}_n$
zz: $< q_{n+1}, q >= 0 \forall q \in \mathbb{P}_n$ (dann $< p_{n+1} - q_{n+1}, \underbrace{p_{n+1} - q_{n+1}}_{\in \mathbb{P}_n} >= 0$ also $p_{n+1} - q_{n+1} = 0$)

zz: $< q_{n+1}, p_j >= 0 \forall j = 0, ..., n$
Sei $j \in \{0, ..., n-2\}$:

$$< q_{n+1}, p_j >= \underbrace{< p_n, \underbrace{(x - \beta_n)p_j}_{\in \mathbb{P}_{n-1}} >}_{=0} - \gamma_n^2 \underbrace{< p_{n-1}, \underbrace{p_j}_{\in \mathbb{P}_{n-2}} >}_{=0} = 0$$

Sei $j = n - 1$:

$$< q_{n+1}, p_{n-1} >=< p_1, xp_{n-1} > -\beta_n \underbrace{< p_n, p_{n-1} >}_{=0} - \underbrace{\gamma_n^2 < p_{n-1}, p_{n-1} >}_{=<p_n,p_n> \text{ Def.}\gamma_n} =< p_n, \underbrace{xp_{n-1} - p_n}_{\in \mathbb{P}_{n-1}} >= 0$$

Sei $j = n$:

$$< q_{n+1}, p_n >=< xp_n, p_n > - \underbrace{\beta_n < p_n, p_n >}_{=<xp_n,p_n> \text{ Def. } \beta_n} -\gamma_n^2 \underbrace{< p_{n-1}, p_n >}_{=0} = 0$$

□

Übung: Mit den Konstanten γ_n, β_n der 3-Term-Rekursion betrachte

$$A = \begin{pmatrix} \beta_0 & -\gamma_1 & 0 & ... & 0 \\ -\gamma_1 & \beta_1 & \gamma_2 & ... & 0 \\ \vdots & & & & \vdots \\ 0 & ... & ... & -\gamma_n & \beta_n \end{pmatrix} \in \mathbb{R}^{(n+1)\times(n+1)}_{\text{sym}} \qquad v^{(k)} = \begin{pmatrix} \tau_0 p_0(x_k) \\ \vdots \\ \tau_n p_n(x_k) \end{pmatrix} \in \mathbb{R}^{n+1}$$

mit $x_0, ..., x_n$ Nullstellen von $p_{n+1}, \tau_j = (-1)^j \left(\prod_{k=1}^j \gamma_n\right)^{-1}$

$\implies Av^{(k)} = x_k v^{(k)}$, d.h. die x_n sind genau die $w_k := \int_a^b L_j w dx = \frac{||w||_{L^1(a,b)}}{||v^{(k)}|_2^2}$

DH: Um eine Gauss-Quadratur zu berechnen, muss man alle EW und alle EV der Matix A bestimmen (d.h. das EW-Problem vollständig lösen).

5 Iterative Lösung von GLS

Ziel:

- Wenn man nichtlineare GLS lösen will, so muss regelmäßig eine Folge von linearen GLS lösen (z.B. Newton).
- Man kann lineare GLS lösen, indem man iterativ Matrix-Vektor-Produkte ausrechnet, insb. muss man die Matrix nicht speichern (z.B. FFT, dividierte Diff.)

5.1 Fixpunktprobleme

Definition 12. *Ein **Iterationsverfahren** ist ein Tripel $(X, \Phi, x*)$ mit X metrischer Raum, $\Phi : X \to X, \Phi(x*) = x*$, d.h. $x*$ ist ein **Fixpunkt** von Φ. Zu einem **Startwert** $x_0 \in X$, sei $x_{k+1} := \Phi(x_k) \forall x \in \mathbb{N}_0$ die erzeugte **Iteriertenfolge** $(x_n)_{n\in\mathbb{N}_0}$.*

Bemerkung 45. *1. Existiert $x := \lim_{n\to\infty} x_n$ und ist Φ stetig bei x, so ist $x = \Phi(x)$, dann $x = \lim_n x_{n+1} = \lim_n \Phi(x_n) = \Phi(x)$.*

2. Ist X normiert und die Lösung von $F(x*) = 0$ gesucht mit $F : X \to X$, so formuliert man dies i.d.R. als Fixpunktproblem, z.B. $x* = \Phi(x*) := x * \pm F(x*)$.

Satz 19 (Banachscher Fixpunktsatz)**.** *X vollständig metrischer Raum, $0 < q < 1$ und $\Phi : X \to X$ mit $d(\Phi(x), \Phi(y)) \leq qd(x, y)$*

$\implies$

1. *Ex. eind. $x* \in X$ mit $\Phi(x*) = x*$*
2. *Für alle $x_0 \in X$ und $x_{k+1} := \Phi(x_k) \forall k \in \mathbb{N}_0$ gilt $\lim_n x_k = x*$*
3. *Für alle $k \in \mathbb{N}_0$ gilt:*
   - $d(x_k, x*) \leq qd(x_{k-1}, x*)$
   - $d(x_k, x*) \leq \frac{q}{1-q} d(x_k, x_{k-1}) \leq \frac{q^k}{1-q} d(x_1, x_0)$
   - $d(x_k, x_{k-1}) \leq (1 + q)d(x_{k-1}, x*)$

Proof. 1. Eindeutigkeit Fixpunkt: Seien $x*, y* \in X$ mit $\Phi(x*) = x*, \Phi(y*) = y*$

$\implies d(x*, y*) = d(\Phi(x*), \Phi(y*)) \leq qd(x*, y*) \implies d(x*, y*) = 0 \implies x* = y*$

2. gezeigt: Falls $(x_k)_{k\in\mathbb{N}}$ konvergiert, ist $x* = \lim_n x_k$ ein Fixpunkt.

3. zz: $(x_k)_{k\in\mathbb{N}}$ für alle Startwerte $x_0 \in X$ eine Cauchy-Folge ist.

Für $m \leq n$ gilt

$$d(x_m, x_n) \leq \sum_{k=m}^{n-1} \underbrace{\overbrace{d(x_k, x_{k+1})}^{=d(\Phi(x_{k-1}),\Phi(x_k))\leq qd(x_{n-1},x_k)\leq q^{k-1}d(x_0,x_1)}}_{\leq q^k d(x_0,x_1)} \leq \left(\sum_{k=m}^{n-1} q^k\right) d(x_0, x_1) \leq$$

$$q^m \frac{1}{1-q} d(x_0, x_1) \to 0, m \to \infty.$$

4. Abschätzungen:

$$d(x_k, x*) = d(\Phi(x_{k-1}), \Phi(x*)) \leq q \underbrace{d(x_{k-1}, x*)}_{\leq d(x_{k-1},x_k)+d(x_k+x*)}$$
$$\implies d(x_k, x*)(1-q) \leq q \underbrace{d(x_{n-1}, x_n)}_{\leq q^{k-1} d(x_0,x_1)} \leq q^k d(x_0, x_1)$$

und $d(x_k, x_{k-1}) \leq \underbrace{d(x_k, x*)}_{qd(x_{n-1},x*)} + d(x_{k-1}, x*) \leq (1+q) d(x_{k-1}, x*)$

□

Definition 13. *Ein Iterationsverfahren $(X, \Phi, x*)$ heißt*

- ***global konvergent**, gdw.* $\forall x_0 \in X : x* = \lim_{n\to\infty} x_n$ *mit* $(x_n)_{n\in\mathbb{N}_0}$ *der Iteriertenfolge* $x_{n+1} := \Phi(x_n) \forall n$
- ***lokal konvergent**, gdw.* $\exists \epsilon > 0 \forall x_0 \in \underbrace{U_\epsilon(x*)}_{:=\{y\in X | d(x,y)<\epsilon\}} : x* = \lim_n x_n$
- ***linear konvergent** (oder: mit Konvergenzordnung $p = 1$), gdw.* $\exists q \in (0,1) \exists \epsilon > 0 \forall x_0 \in U_\epsilon(x*) \forall n \in \mathbb{N}_0 : d(x*, x_{n+1}) \leq q d(x*, x_n)$
- ***von Konvergenzordnung** $p > 1$, gdw.* $\exists C > 0 \forall \epsilon > 0 \forall x_0 \in U_\epsilon(x*) \forall n \in \mathbb{N}_0 : d(x*, x_{n+1}) \leq C d(x*, x_n)^p$

*Die Menge $U_\epsilon(x*)$ nennt man auch **Konvergenzbereich**.*

Beispiel 17. *Ist $\Phi : X \to X$ eine (strikte) Kontraktion auf einem vollständig metrischen Raum mit Fixpunkt $x* \in X$, so ist $(X, \Phi, x*)$ global linear konv.*

Lemma 12. *Sei $(X, \Phi, x*)$ ein Iterationsverfahren mit Konvergenzordnung $p \geq 1$. Dann ist $(X, \Phi, x*)$ lokal konvergent und in jeder Konvergenzordnung $1 \leq \tilde{p} \leq p$.*

Proof.

- $p = 1 \implies$ lokale konvergenz

  Wähle $0 < q < 1$ und $\epsilon > 0$ gemäß Definition. Sei $x_0 \in U_\epsilon(x*)$. Dann $d(x*, x_n) \leq q^k \underbrace{d(x*, x_0)}_{\in\mathbb{R}}$
- Konvergenzordnung $p > 1 \implies$ lineare konvergenz mit $q = \frac{1}{2}$. Seien $C > 0, \epsilon > 0$ gemäß Def. gewählt. Wähle $\delta := \min\{\epsilon, \left(\frac{1}{2C}\right)^{1/(p-1)}\}$. Sei $x_0 \in U_\delta(x*)$

  Beh. $d(x_n, x*) \leq \underbrace{2^{-n}}_{\leq 1} \underbrace{d(x_0, x*)}_{<\delta} < \delta \forall n \in \mathbb{N}_0$ (und $d(x_n, x*) \leq C \underbrace{d(x_{n-1}, x*)^{p-1}}_{\leq \delta^{p-1} \leq 1/(2C)} d(x_{n-1}, x*) \leq \frac{1}{2} d(x_{n-1}, x*) \forall n \in \mathbb{N}$)

  Beweis der Beh. durch Induktion, klar $n = 0$

$$TODO1321 : 31$$

□