Numerische Methoden II, 170 024, und Übungen, 170 025
Vorlesungs-Notizen C. Brand, 7. März 2018

1 Finite-Differenzen-Methode

Unterlagen, Informationsquellen Für die ersten Einheiten zum Thema Finite-Differenzen-Methode verwenden wir:

Kapitel 1 und Teile von Kapitel 2 aus
Randall J. LeVeque, *Finite Difference Methods for Ordinary an Partial Differential Equations*, SIAM 2007.

Zu diesem Buch gibt es eine Webseite mit Aufgaben und Matlab-Code:
`http://faculty.washington.edu/rjl/fdmbook/`

Und ergänzend Kapitel 9.1 und 9.2 aus
Robert Plato, *Numerische Mathematik kompakt, Grundlagenwissen für Studium und Praxis* 3., aktualisierte und verbesserte Auflage, Vieweg, Wiesbaden 2006

Als erstes sieht man heutzutage in Wikipedia nach. Dort steht zu diesem Stichwort:

> Finite-Differenzen-Methoden sind eine Klasse numerischer Verfahren zur Lösung gewöhnlicher und partieller Differentialgleichungen.
>
> Zunächst wird das Gebiet, für das die Gleichung gelten soll, durch eine endliche Zahl von Gitterpunkten diskretisiert. Eindimensionale Intervalle werden dazu in gleich lange Teilintervalle zerlegt, mehrdimensionale Gebiete in Rechteckgitter. Die Ableitungen der gesuchten Funktion an den Gitterpunkten werden dann durch Differenzenquotienten approximiert [...]. Die Differentialgleichung wird auf diese Weise durch ein System von Differenzengleichungen angenähert, die mittels verschiedener Algorithmen zur numerischen Lösung von Gleichungssystemen gelöst werden können.

Einleitung, Kontext

- Physikalische Gesetze sind zumeist als differentielle Beziehungen (Kraft=Masse ×Beschleunigung, Maxwell-Gleichungen, Navier-Stokes-Gleichungen...) formuliert.

  Aber: näher an der physikalischen Realität sind oft Integral-Formulierungen: Gesamtimpuls für einen Volumsbereich bleibt erhalten,... oder Extremal-Prinzipien: Potentielle Energie wird minimal,... Finite Volums- beziehungsweise Finite-Elemente-Methoden nehmen diese Formulierungen als Ausgangspunkt.

- Analytische Lösungen gibt es nur in den einfachsten Fällen. Ansonsten ist man auf *numerische Lösungen* angewiesen.

Aufgabenstellung

Gegeben

- Eine *Differentialgleichung* (gewöhnlich oder partiell, erster oder höherer Ordnung, in einer oder mehreren unabhängigen Variablen; oder auch ein System solcher Differentialgleichungen)
- Ein *Gebiet* (räumlich ein- oder mehrdimensional, Zeitintervall)
- *Rand-* und/oder *Anfangsbedingungen*

Gesucht Eine Funktion, die

- im gesamten Rechengebiet die Differentialgleichung und
- am Rand die Rand- und/oder Anfangsbedingungen erfüllt.

Aber: Diese Forderung ist zu streng, es gibt sinnvolle Aufgabenstellungen, wo eine solche Lösung (eine *Lösung im klassischen Sinn*) nicht existiert, obwohl Integral- oder Extremal-Formulierungen sogenannte *schwache Lösungen* zulassen.

Grundidee, Lösungsansatz: *Diskretisierung* des kontinuierlichen Problems

- Überziehe das Gebiet mit einem Netz aus *Gitterpunkten*
- Ersetze *Differentialoperatoren* durch *Differenzen* von Gitterwerten
- Aus (nicht-)linearen Differentialgleichungen werden Systeme (nicht-)linearer Gleichungen.

Vorteile

- einfache Herleitung, leichte Implementierung für geometrisch einfache Gebiete.

Nachteile

- auf unregelmäßigen Geometrieen. Nicht so mächtiger Theorie-Unterbau wie bei Finiten Elementen.

1.1 Differenzenformeln

approximieren Differentialoperatoren auf einem Rechengitter

Elementare Formeln für erste und zweite Ableitung

lassen sich an einem Funktionsgraph geometrisch anschaulich interpretieren: Sekanten-Steigung als Näherung an Tangenten-Steigung. Siehe Figure 1.1 in R.LeVq.

Erste Ableitung

$$u'(x) \approx \frac{u(x+h) - u(x)}{h} \qquad \text{Vorwärtsdifferenz} \qquad (1)$$

$$u'(x) \approx \frac{u(x) - u(x-h)}{h} \qquad \text{Rückwärtsdifferenz} \qquad (2)$$

$$u'(x) \approx \frac{u(x+h) - u(x-h)}{2h} \qquad \text{Zentrale Differenz} \qquad (3)$$

Zweite Ableitung

$$u''(x) \approx \frac{u(x-h) - 2u(x) + u(x+h)}{h^2} \qquad \text{Zentrale Differenz} \qquad (4)$$

Es gibt auch noch andere Differenzenformeln, zum Beispiel die unsymmetrische Vierpunkt-Formel

$$u'(x) \approx \frac{1}{6h}\left[2u(x+h) + 3u(x) - 6u(x-h) + u(x-2h)\right] \qquad (5)$$

Beispiel 1: Für $u(x) = \sin(x)$ ist $u'(x)$ an der Stelle $x = 1$ gleich $\cos(1)$. Je nach gewählter Schrittweite h liefern die Differenzenformeln unterschiedlich genaue Näherungen. Die Fehler (Näherung minus korrekter Wert) sind hier für einige Schrittweiten tabellarisch gelistet.

h	vorwärts	rückwärts	zentral	vierpunkt
$1{,}0 \times 10^{-1}$	$-4{,}2939 \times 10^{-2}$	$4{,}1138 \times 10^{-2}$	$-9{,}0005 \times 10^{-4}$	$6{,}8207 \times 10^{-5}$
$5{,}0 \times 10^{-2}$	$-2{,}1257 \times 10^{-2}$	$2{,}0807 \times 10^{-2}$	$-2{,}2510 \times 10^{-4}$	$8{,}6491 \times 10^{-6}$
$1{,}0 \times 10^{-2}$	$-4{,}2163 \times 10^{-3}$	$4{,}1983 \times 10^{-3}$	$-9{,}0050 \times 10^{-6}$	$6{,}9941 \times 10^{-8}$
$5{,}0 \times 10^{-3}$	$-2{,}1059 \times 10^{-3}$	$2{,}1014 \times 10^{-3}$	$-2{,}2513 \times 10^{-6}$	$8{,}7540 \times 10^{-9}$
$1{,}0 \times 10^{-3}$	$-4{,}2083 \times 10^{-4}$	$4{,}2065 \times 10^{-4}$	$-9{,}0050 \times 10^{-8}$	$6{,}9959 \times 10^{-11}$

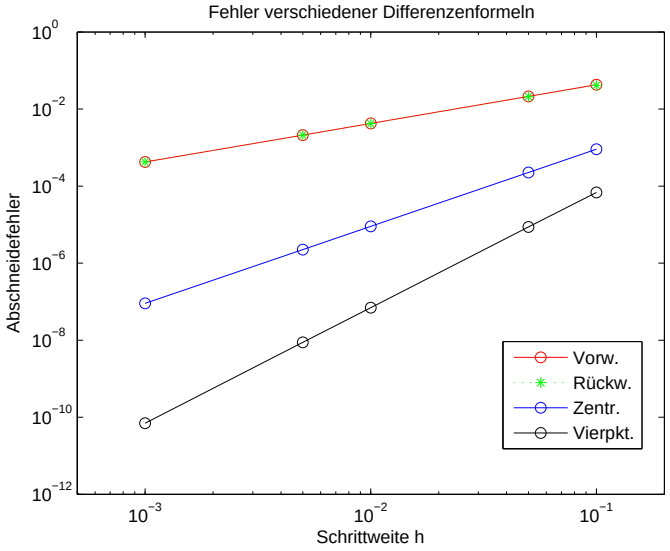
Interessant ist, wie rasch der Fehler für $h \to 0$ abnimmt. Man erkennt:

- Vorwärts- und Rückwärtsdifferenz haben absolut etwa gleich große Fehler, aber mit unterschiedlichen Vorzeichen.
- Bei Vorwärts- und Rückwärtsdifferenz nimmt der Fehler pro Zehnerpotenz in h ebenfalls um eine Zehnerpotenz ab.
- Bei der zentralen Differenz nimmt der Fehler pro Zehnerpotenz in h *zwei* Zehnerpotenzen ab.
- Bei der Vierpunkt-Formel: pro Zehnerpotenz in h *drei* Zehnerpotenzen!

Ein log-log-Diagramm zeigt besonders gut, dass die Fehler je nach Verfahren proportional zu unterschiedlichen Potenzen von h sind.

Bemerkung: In doppelt logarithmischer Darstellung werden Funktionen der Art $y = cx^p$ zu Geraden. Die Geradensteigung entspricht der Hochzahl.

Das MATLAB-Programm `chap1ex1.m`[1] berechnet die Tabelle und zeichnet das Diagramm.



Groß-O Schreibweise verwenden wir, um den Zusammenhang Gitterweite–Fehler einfach zu beschreiben.

Für eine Funktion $g = g(h)$ bedeutet der Ausdruck $g = O(h^p)$, sprich „g ist groß-Oh von h hoch p“: Der Betrag von g ist kleiner als eine Konstante mal $|h|^p$, sofern h genügend nahe bei 0 liegt.

Formaler geschrieben: Es gilt $g = O(h^p)$ für $h \to 0$, wenn es Konstante C und h_0 gibt, so dass $|g(h)| < C|h|^p \quad \forall h$ mit $|h| < h_0$ gilt.

Abschneidefehler Der Fehler einer Differenzenformel lässt sich durch Taylorreihenentwicklung untersuchen. Wir nehmen an, dass $u(x)$ hinreichend oft differnzierbar ist. Dann gilt

$$u(x+h) = u(x) + hu'(x) + \frac{1}{2}h^2u''(x) + \frac{1}{6}h^3u'''(x) + \frac{1}{24}h^4u''''(x) + O(h^5) \qquad (6)$$

$$u(x-h) = u(x) - hu'(x) + \frac{1}{2}h^2u''(x) - \frac{1}{6}h^3u'''(x) + \frac{1}{24}h^4u''''(x) + O(h^5) \qquad (7)$$

[1] Aus dem Begleitmaterial zu R. LeVeque

Dann ergeben passende Umformungen

$$\frac{u(x+h)-u(x)}{h} = u'(x) + \frac{1}{2}hu''(x) + O(h^2) \qquad \text{Vorwärts-Differenz} \qquad (8)$$

$$\frac{u(x)-u(x-h)}{h} = u'(x) - \frac{1}{2}hu''(x) + O(h^2) \qquad \text{Rückwärts-Differenz} \qquad (9)$$

$$\frac{u(x+h)-u(x-h)}{2h} = u'(x) + \frac{1}{6}h^2u'''(x) + O(h^4) \qquad \text{Zentrale Differenz} \qquad (10)$$

Man spricht von Abschneidefehler (*truncation error*), weil er durch „Abschneiden“ der höheren Terme in der Taylorreihe entsteht.

Bei der zentralen Formel verschwinden übrigens alle ungeraden Potenzen von h, eine typische Eigenschaft zentraler Differenzen. Für die Vierpunktformel (5) brauchen wir noch die Reihenentwicklung von $u(x-2h)$. Tatsächlich verschwinden in Formel (5) auch noch die h^2-Terme. Der Differenzenausdruck ist gleich

$$u'(x) + \frac{1}{12}h^3u'''(x) + O(h^4)$$

Verwendet man die *Taylorformel mit Restglied*, lassen sich die Abschneidefehler auch ohne $O(h^p)$-Term angeben, zum Beispiel ist bei der Vorwärtsdifferenz (8)

$$\frac{u(x+h)-u(x)}{h} = u'(x) + \frac{1}{2}hu''(\xi)$$

wobei ξ ein Wert $x \le \xi \le x+h$ ist. Der Fehler ist also proportional der Ableitung u'', ausgewertet an einer Stelle in der Nähe von x. Für eine Abschätzung reicht es oft, den Maximalwert $\max_x |u''(x)|$ zu verwenden.

In unserem Beispiel 1 ist $\frac{1}{2}u''(1) = -\frac{1}{2}\sin(1) = -0,42073$, und das passt (wenn man mit dem entsprechendem h multipliziert) gut zu den Zahlenwerten in der Tabelle.

Aufgaben Vergleichen Sie die theoretischen Abschneidefehler der anderen Differenzenformeln mit den Tabellenwerten.

Verwenden Sie $u(x) = \exp(x)$ und bestimmen Sie für $u''(1)$ die Abschneidefehler für verschiedene h und Differenzenformeln. Verwenden Sie dazu aus der folgenden Tabelle

- symmetrische Drei-, Fünf- und Siebenpunkt-Formel
- unsymmetrische Vier-, Fünf- und Siebenpunkt-Formel

Eine Tabelle mit Differenzenformeln

Eine sehr ausführliche Zusammenstellung verschiedener Differenzenformel, (auch so genannter *kompakter* Formeln) aus: Lothar Collatz, *Numerische Behandlung von Differentialgleichungen*, Springer 1955.

Weil es mittels Computer-Algebra-Systemen inzwischen einfach ist, solche Formeln je nach Bedarf herzuleiten, sind solche Tabellen nicht mehr so wichtig. Herleiten von Differenzenformeln wird im nächsten Abschnitt erklärt.

Tafel III.

Ausdrücke des Differenzenverfahrens bei gewöhnlichen Differentialgleichungen.

		Formel Abkürzungen: $y_j = y(jh)$, $y'_j = y'(jh)$ usw.	Das nächste nichtverschwindende Glied der TAYLOR-Entwicklung
Formeln für y'	symmetrisch	$y'_0 = \frac{1}{2h}(-y_{-1} + y_1) +$	$-\frac{1}{6}h^2 y'''_0 - \cdots$
		$y'_0 = \frac{1}{12h}(y_{-2} - 8y_{-1} + 8y_1 - y_2) +$	$+\frac{1}{30}h^4 y^{\mathrm{V}}_0 + \cdots$
		$y'_0 = \frac{1}{60h}(-y_{-3} + 9y_{-2} - 45y_{-1} + 45y_1 - 9y_2 + y_3) +$	$-\frac{1}{140}h^6 y^{\mathrm{VII}}_0 + \cdots$
		$y'_{-1} + 4y'_0 + y'_1 + \frac{3}{h}(y_{-1} - y_1) = 0 +$	$+\frac{1}{30}h^4 y^{\mathrm{V}}_0 + \cdots$
		$y'_{-1} + 3y'_0 + y'_1 + \frac{1}{12h}(y_{-2} + 28y_{-1} - 28y_1 - y_2) = 0 +$	$-\frac{1}{420}h^6 y^{\mathrm{VII}}_0 - \cdots$
		$y'_{-2} + 16y'_{-1} + 36y'_0 + 16y'_1 + y'_2 + \frac{5}{6h}(5y_{-2} + 32y_{-1} - 32y_1 - 5y_2) = 0 +$	$+\frac{1}{630}h^8 y^{\mathrm{IX}}_0 + \cdots$
		$7y'_{-2} + 32y'_{-1} + 12y'_0 + 32y'_1 + 7y'_2 + \frac{45}{2h}(y_{-2} - y_2) = 0 +$	$+\frac{4}{21}h^6 y^{\mathrm{VII}}_0 + \cdots$
		$y'_{-2} + 4y'_{-1} + 4y'_1 + y'_2 + \frac{1}{6h}(19y_{-2} - 8y_{-1} + 8y_1 - 19y_2) = 0 +$	$+\frac{1}{35}h^6 y^{\mathrm{VII}}_0 + \cdots$
	unsymmetrisch	$y'_0 = \frac{1}{h}(-y_0 + y_1) +$	$-\frac{1}{2}h y''_0 - \cdots$
		$y'_0 = \frac{1}{2h}(-3y_0 + 4y_1 - y_2) +$	$+\frac{1}{3}h^2 y'''_0 + \cdots$
		$y'_0 = \frac{1}{12h}(-3y_{-1} - 10y_0 + 18y_1 - 6y_2 + y_3) +$	$-\frac{1}{20}h^4 y^{\mathrm{V}}_0 + \cdots$
		$y'_0 = \frac{1}{60h}(2y_{-2} - 24y_{-1} - 35y_0 + 80y_1 - 30y_2 + 8y_3 - y_4) +$	$+\frac{1}{105}h^6 y^{\mathrm{VII}}_0 + \cdots$
		$y'_0 + y'_1 + \frac{2}{h}(y_0 - y_1) = 0 +$	$+\frac{1}{6}h^2 y'''_0 + \cdots$
		$y'_{-1} + 9y'_0 + 9y'_1 + y'_2 + \frac{1}{3h}(11y_{-1} + 27y_0 - 27y_1 - 11y_2) = 0 +$	$+\frac{1}{140}h^6 y^{\mathrm{VII}}_0 + \cdots$
Formeln für y''	symmetrisch	$y''_0 = \frac{1}{h^2}(y_{-1} - 2y_0 + y_1) +$	$-\frac{1}{12}h^2 y^{\mathrm{IV}}_0 + \cdots$
		$y''_0 = \frac{1}{12h^2}(-y_{-2} + 16y_{-1} - 30y_0 + 16y_1 - y_2) +$	$+\frac{1}{90}h^4 y^{\mathrm{VI}}_0 + \cdots$
		$y''_0 = \frac{1}{180h^2}(2y_{-3} - 27y_{-2} + 270y_{-1} - 490y_0 + 270y_1 - 27y_2 + 2y_3) +$	$-\frac{1}{560}h^6 y^{\mathrm{VIII}}_0 + \cdots$
		$y''_{-1} + 10y''_0 + y''_1 - \frac{12}{h^2}(y_{-1} - 2y_0 + y_1) = 0 +$	$+\frac{1}{20}h^4 y^{\mathrm{VI}}_0 + \cdots$
		$2y''_{-1} + 11y''_0 + 2y''_1 - \frac{3}{4h^2}(y_{-2} + 16y_{-1} - 34y_0 + 16y_1 + y_2) = 0 +$	$-\frac{23}{5040}h^6 y^{\mathrm{VIII}}_0 + \cdots$

Tafel III. (Fortsetzung.)

		Formel Abkürzungen: $y_j = y(jh)$, $y'_j = y'(jh)$ usw.	Das nächste nichtverschwindende Glied der TAYLOR-Entwicklung
Formeln für y''	symmetrisch	$23y''_{-2} + 688y''_{-1} + 2358y''_0 + 688y''_1 + 23y''_2 -$ $-\frac{15}{h^2}(31y_{-2} + 128y_{-1} - 318y_0 + 128y_1 + 31y_2) = 0 +$	$+\frac{79}{1260}h^8 y^{\mathrm{X}}_0 + \cdots$
		$y''_{-1} - 8y''_0 + y''_1 + \frac{9}{h}(y'_{-1} - y'_1) + \frac{24}{h^2}(y_{-1} - 2y_0 + y_1) = 0 +$	$+\frac{1}{2520}h^6 y^{\mathrm{VIII}}_0 + \cdots$
		$y''_{-1} - y''_1 + \frac{1}{h}(7y'_{-1} + 16y'_0 + 7y'_1) + \frac{15}{h^2}(y_{-1} - y_1) = 0 +$	$-\frac{1}{315}h^5 y^{\mathrm{VII}}_0 + \cdots$
	unsymmetrisch	$y''_0 = \frac{1}{h^2}(2y_0 - 5y_1 + 4y_2 - y_3) +$	$+\frac{11}{12}h^2 y^{\mathrm{IV}}_0 + \cdots$
		$y''_0 = \frac{1}{12h^2}(11y_{-1} - 20y_0 + 6y_1 + 4y_2 - y_3) +$	$+\frac{1}{12}h^3 y^{\mathrm{V}}_0 + \cdots$
		$y''_0 = \frac{1}{180h^2}(-13y_{-2} + 228y_{-1} - 420y_0 + 200y_1 + 15y_2 - 12y_3 + 2y_4) +$	$-\frac{1}{90}h^5 y^{\mathrm{VII}}_0 + \cdots$
Formeln für y'''	symmetrisch	$y'''_0 = \frac{1}{2h^3}(-y_{-2} + 2y_{-1} - 2y_1 + y_2) +$	$-\frac{1}{4}h^2 y^{\mathrm{V}}_0 + \cdots$
		$y'''_0 = \frac{1}{8h^3}(y_{-3} - 8y_{-2} + 13y_{-1} - 13y_1 + 8y_2 - y_3) +$	$+\frac{7}{120}h^4 y^{\mathrm{VII}}_0 + \cdots$
		$y'''_{-1} + 2y'''_0 + y'''_1 + \frac{2}{h^3}(y_{-2} - 2y_{-1} + 2y_1 - y_2) = 0 +$	$-\frac{1}{60}h^4 y^{\mathrm{VII}}_0 + \cdots$
		$y'''_{-2} + 56y'''_{-1} + 126y'''_0 + 56y'''_1 + y'''_2 + \frac{120}{h^3}(y_{-2} - 2y_{-1} + 2y_1 - y_2) = 0 +$	$-\frac{1}{252}h^6 y^{\mathrm{IX}}_0 + \cdots$
	unsym.	$y'''_0 = \frac{1}{2h^3}(-3y_{-1} + 10y_0 - 12y_1 + 6y_2 - y_3) +$	$+\frac{1}{4}h^2 y^{\mathrm{V}}_0 + \cdots$
		$y'''_0 = \frac{1}{8h^3}(-y_{-2} - 8y_{-1} + 35y_0 - 48y_1 + 29y_2 - 8y_3 + y_4) +$	$-\frac{1}{15}h^4 y^{\mathrm{VII}}_0 + \cdots$
Formeln für y^{IV}	symmetrisch	$y^{\mathrm{IV}}_0 = \frac{1}{h^4}(y_{-2} - 4y_{-1} + 6y_0 - 4y_1 + y_2) +$	$-\frac{1}{6}h^2 y^{\mathrm{VI}}_0 + \cdots$
		$y^{\mathrm{IV}}_0 = \frac{1}{6h^4}(-y_{-3} + 12y_{-2} - 39y_{-1} + 56y_0 - 39y_1 + 12y_2 - y_3) +$	$+\frac{7}{240}h^4 y^{\mathrm{VIII}}_0 + \cdots$
		$y^{\mathrm{IV}}_{-1} + 4y^{\mathrm{IV}}_0 + y^{\mathrm{IV}}_1 - \frac{6}{h^4}(y_{-2} - 4y_{-1} + 6y_0 - 4y_1 + y_2) = 0 +$	$+\frac{1}{10}h^4 y^{\mathrm{VIII}}_0 + \cdots$
		$y^{\mathrm{IV}}_{-2} - 124y^{\mathrm{IV}}_{-1} - 474y^{\mathrm{IV}}_0 - 124y^{\mathrm{IV}}_1 + y^{\mathrm{IV}}_2 + \frac{720}{h^4}(y_{-2} - 4y_{-1} + 6y_0 - 4y_1 + y_2) = 0 +$	$+\frac{5}{21}h^6 y^{\mathrm{X}}_0 + \cdots$

1.2 Herleiten von Differenzenformeln

über Taylor-Reihe Ansatz mit unbestimmten Koeffizienten.

Beispiel 2: Einseitige Dreipunkt-Formel für u', siehe Ausarbeitung Example 1.2 in R. LeVeque.

über Interpolationspolynom Grundidee: Finde ein Interpolationspolynom $p(x)$, differenziere p und verwende $p'(x)$ als Näherung für $u'(x)$.

Beispiel 3: Einseitige Dreipunkt-Formel: Wir leiten sie für die drei Wertepaare

$$(0; u(0)), \quad (-h; u(-h)), \quad (-2h; u(-2h))$$

her. (Eine Verschiebung des Koordinatenursprungs ändert nicht die Koeffizienten in den Differenzenformeln. Wir berechnen daher der Einfachheit halber die Differenzenformel für $u'(0)$.) In der Lagrange-Form lässt sich das Polynom direkt hinschreiben:

$$p(x) = u(-2h)L_{-2}(x) + u(-h)L_{-1}(x) + u(0)L_0(x)$$

mit

$$L_{-2}(x) = \frac{(x+h)x}{(-2h+h)(-2h)} \qquad L_{-1}(x) = \frac{(x+2h)x}{(-h+2h)(-h)} \qquad L_0(x) = \frac{(x+2h)(x+h)}{(2h)h}$$

Differenzieren der $L_i(x)$ und Auswerten für $x = 0$ liefert

$$L_{-2}(0)' = \frac{1}{2h} \qquad L_{-1}(0)' = -\frac{2}{h} \qquad L_0(0') = \frac{3}{2h}$$

Vergleiche mit Formel 1.11 in LeVeque!

über Vandermonde-System in MATLAB Kapitel 1.5 in LeVeque beschreibt, wie sich die Koeffizienten als Lösung eines Gleichungssystems mit einer Vandermonde-Matrix finden lassen und wie MATLAB dieses System erstellt (Skript `fdcoeffV.m`).

1.3 Die 1-D Poisson-Gleichung

Das übliche einführende Standard-Beispiel für Randwertprobleme. Siehe Plato 9.1, RLeVq 2.3 und 2.4 und das Beispiel beim Stichwort Finite-Differenzen-Methode in Wikipedia.

Die 1-D Randwertaufgabe lässt sich einfach und direkt durch Integration lösen; dafür sind Finite Differenzen-Verfahren gar nicht notwendig. Aber die hier erklärten Grundideen und Lösungsansätze lassen sich direkt auf 2-D und 3-D Problemen übertragen.

Randwert-Aufgabe für die Poisson-Gleichung, klassische Formulierung Gesucht ist eine Funktion $u : [0,1] \to \mathbb{R}$, die für gegebenen *Quellterm* $f : (0,1) \to \mathbb{R}$ erfüllt:

$$\begin{aligned} u'' &= f \quad \text{in } \Omega = (0,1) \\ u(0) &= u(1) = 0 \end{aligned} \tag{11}$$

Die Aufgabe ist (in dieser Schreibweise) nur sinnvoll , wenn u in $(0,1)$ zweimal differenzierbar ist und bis zum Rand hin, also in $[0,1]$, stetig ist.

Für eine mathematisch präzise Formulierung muss festgelegt sein, welche Funktionenklassen für f und u betrachtet werden. Für eine *klassische Lösung* der Randwertaufgabe 11 fordert man $f \in C^0(\Omega)$, $u \in C^2(\Omega) \cap u \in C^0(\bar{\Omega})$. Das ist aber für die praktische Anwendung viel zu einschränkend: Als Quellterme f können auch *Punktquellen* oder unstetige Stufenfunktionen sinnvoll sein. Andererseits ist in 2-D oder 3-D die Forderung $f \in C^0(\Omega)$ nicht einmal hinreichend für die Existenz einer Lösung u.

1.4 Diskretisierung der Randwertaufgabe

Äquidistantes Gitter für das Rechengebiet Ω der Randwertaufgabe 11: n innere Punkte x_j im Intervall $0 < x < 1$

$$x_j = jh, \quad j = 1, \ldots, n, \quad \text{Gitterweite } h = \frac{1}{n+1}$$

führt auf Gleichungssystem

$$A\mathbf{u} = \mathbf{f} \quad \text{mit } A = -\frac{1}{h^2}\text{tridiag}(-1, 2, -1) \quad \text{und } \mathbf{f} = [f(x_j)] \tag{12}$$

1.5 Lokaler Diskretisierungsfehler

Es sei $\mathbf{u}_{ex}$ der Vektor der exakten Lösung an den Gitterpunkten: $\mathbf{u}_{ex} = [u(x_j)]$. Einsetzen von $\mathbf{u}_{ex}$ erfüllt das Gleichungssystem nicht exakt, es bleibt ein Restvektor δ, der *lokale Diskretisierungsfehler*.

$$A\mathbf{u}_{ex} - \mathbf{f} = \boldsymbol{\delta} \tag{13}$$

Aus Taylorreihenentwicklung folgt für j-te Zeile von 13

$$\begin{aligned} \delta_j =& \frac{1}{h^2}\left(u(x_{j-1}) - 2u(x_j) + u(x_{j+1})\right) - f(x_j) = \\ & u''(x_j) + \frac{1}{12}h^2 u''''(\xi) - f(x_j) = \\ & \frac{1}{12}h^2 u''''(\xi) \quad \text{mit } x_{j-1} \le \xi \le x_{j+1} \end{aligned}$$

Norm des lokalen Diskretisierungsfehler-Vektors

$$||\boldsymbol{\delta}||_\infty \leq \frac{1}{12} h^2 ||u''''||_\infty$$

Maximal-Betrag in $\boldsymbol{\delta}$ ist beschränkt durch Maximalwert der vierten Ableitung mal $\frac{h^2}{12}$. Beachte: links Vektornorm im $\mathbb{R}^n$, rechts Norm im Raum der viermal stetig diff'baren Funktionen.

1.6 Fehler der Lösung

$$\boldsymbol{\epsilon} = \mathbf{u}_{ex} - \mathbf{u}$$

heißt *globaler Fehler.*

Der lokale Diskretisierungsfehler $\boldsymbol{\delta}$ lässt sich abschätzen, uns aber interessiert der globale Fehler $\boldsymbol{\epsilon}$. Es gilt

$$A\boldsymbol{\epsilon} = A\mathbf{u}_{ex} - A\mathbf{u} = A\mathbf{u}_{ex} - \mathbf{f} = \boldsymbol{\delta}$$

also

$$A\boldsymbol{\epsilon} = \boldsymbol{\delta} \tag{14}$$

Der globale Fehler ϵ erfüllt ein Gleichungssystem mit derselben Finite-Differenzen-Matrix wie im Originalproblem, und mit Quellterm δ auf der rechten Seite.

Man kann auch so formulieren

$A\mathbf{u} = \mathbf{f}$	Näherung $\mathbf{u}$ erfüllt dieses System
$A\mathbf{u}_{ex} = \mathbf{f} + \boldsymbol{\delta}$	Exakte Lösung erfüllt System mit zusätzlichem Quellterm

Der lokale Diskretisierungsfehler wirkt wie ein zusätzlicher Quellterm!

Entscheidende Frage: folgt aus kleiner Störung $\boldsymbol{\delta}$ im Quellterm auch ein kleiner Fehler $\boldsymbol{\epsilon}$ in der Lösung?

Aus Gleichung 14 folgt

$$\boldsymbol{\epsilon} = A^{-1}\boldsymbol{\delta}$$

und daraus die Forderung, dass A^{-1} keine „zu starke“ Vergrößerung des $\boldsymbol{\delta}$-Vektors bewirken darf.

1.7 Stabilität

Angenommen, eine Finite-Differenzen-Diskretisierung auf zunehmend feineren Gittern führt zu einer Folge von Matrix-Gleichungen

$$A^{(h)}\mathbf{u}^{(h)} = \mathbf{f}^{(h)} \quad \text{mit } h \to 0 \ .$$

Wenn (jedenfalls für genügend kleine h) die Inverse $\left(A^{(h)}\right)^{-1}$ existiert und deren Norm durch eine von h unabhängige Konstante C beschränkt ist,

$$|| \left(A^{(h)}\right)^{-1} || \leq C \quad \forall h \leq h_0 \ ,$$

dann heißt die Diskretisierung *stabil*.

(genauer: stabil in der entsprechenden Norm; wir arbeiten hier mit Stabilität in der Maximumsnorm)

1.8 Konsistenz

Eine Diskretisierung einer Randwertaufgabe heißt *konsistent* (in der entsprechenden Norm), wenn $||\boldsymbol{\delta}^{(h)}|| \to 0$ für $h \to 0$.

Speziell: *konsistent von p-ter Ordnung*, wenn $||\boldsymbol{\delta}^{(h)}|| = O(h^p)$.

Beachte: Auch die Diskretisierung der Randbedingungen ist mit zu berücksichtigen. Die homogenen Dirichlet-Randbedingungen in unserem Musterbeispiel lassen sich problemlos einbinden. In allgemeineren Fällen kann die konsistente Diskretisierung der Randbedingungen nicht trivial sein.

1.9 Konvergenz

Eine Diskretisierung einer Randwertaufgabe heißt *konvergent*, wenn $||\boldsymbol{\epsilon}^{(h)}|| \to 0$ für $h \to 0$.

Es gilt der wichtige Zusammenhang bei linearen Problemen:

$$\text{Konsistenz + Stabilität} \Rightarrow \text{Konvergenz}$$

Beweis ist einfach:

$$||\boldsymbol{\epsilon}^{(h)}|| = || \left(A^{(h)}\right)^{-1} \boldsymbol{\delta}^{(h)}|| \leq || \left(A^{(h)}\right)^{-1} || \cdot ||\boldsymbol{\delta}^{(h)}|| \leq C||\boldsymbol{\delta}^{(h)}|| \to 0$$

Ebenso folgt: Konvergenz-Ordnung = Konsistenz-Ordnung!

1.10 Eigenschaften der $(-1, 2, -1)$-Tridiagonalmatrix

Um nachzuweisen, dass die Diskretisierung (12) der Randwertaufgabe (11) stabil ist, und um den globalen Diskretisierungsfehler abzuschätzen, sind Eigenschaften der Matrix $A = \text{tridiag}(-1, 2, -1)$ wesentlich: Nichtsingularität und Norm.

Für diese spezielle Matrix gibt es einfache Formeln und exakte Ergebnisse: Sei A die $n \times n$ Tridiagonalmatrix A= tridiag(-1,2,-1), dann gilt

$$\det A = n + 1$$

$$||A^{-1}||_\infty = \begin{cases} \frac{1}{8}n(n+2) \\ \frac{1}{8}(n+1)^2 \end{cases} \text{ für } n \begin{cases} \text{gerade} \\ \text{ungerade} \end{cases}$$

Diese Matrix ist aber nur das prototypische Beispiel für eine allgemeinere Klasse von Matrizen: irreduzibel diagonaldominante Matrizen. Wir wollen die grundsätzlichen Ideen vorstellen, ohne zu viel technischen Aufwand zu treiben. Die hier am einfachen Beispiel der tridiag$(-1, 2, -1)$-Matrix vorgestellten Argumente und Beweismethoden lassen sich allgemein für irreduzibel diagonaldominante Matrizen formulieren.

Diagonaldominanz

$$|a_{ii}| \geq \sum_{j \neq i} |a_{ij}| \quad \forall i = 1, \ldots, n$$

und mindestens in eine Zeile i gilt echt $>$ statt $\geq$.

Irreduzibilität vereinfacht: „das Rechengitter ist zusammenhängend“

Definition: wenn für $i \neq j$ der Matrixeintrag $a_{ij} \neq 0$ ist, dann heißen die Indizes i und j *direkt verbunden*. Die Matrix A ist *irreduzibel*, wenn es ausgehend von jedem beliebigen i eine Folge direkter Verbindungen $i \to j \to k \to \cdots \to \ell$ zu jedem anderen ℓ gibt.

Zusammenhang Matrix–Rechengitter: Ein Differenzenoperator greift auf benachbarte Punkte im Rechengitter zu. Sind die Gitterpunkte i und j betroffen, dann entsteht in A ein Eintrag $a_{ij} \neq 0$. In der Regel ist dann auch $a_{ji} \neq 0$. Man sagt: Das Besetzungsmuster der Matrix ist *symmetrisch*.

In unserem Musterbeispiel Gleichung 12 gilt sogar $a_{ij} = a_{ji}$, die Matrix A ist *symmetrisch*.

Aufgaben zur Illustration: Gibt es in der Matrix

$$B = \begin{bmatrix} 2 & 0 & 0 & -1 & -1 \\ 0 & 2 & 0 & -1 & 0 \\ 0 & 0 & 2 & 0 & -1 \\ -1 & -1 & 0 & 2 & 0 \\ -1 & 0 & -1 & 0 & 2 \end{bmatrix}$$

eine Verbindung von $i = 2$ nach $j = 3$? Wie sieht der zugeordnete Graph aus? Ist A irreduzibel?

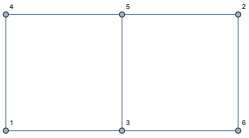
Rechengitter und Matrixstruktur Gibt es in den Matrizen

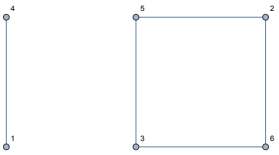
$$C = \begin{bmatrix} 4 & 0 & -1 & -1 & 0 & 0 \\ 0 & 4 & 0 & 0 & -1 & -1 \\ -1 & 0 & 4 & 0 & -1 & -1 \\ -1 & 0 & 0 & 4 & -1 & 0 \\ 0 & -1 & -1 & -1 & 4 & 0 \\ 0 & -1 & -1 & 0 & 0 & 4 \end{bmatrix} \quad \text{und } D = \begin{bmatrix} 4 & 0 & 0 & -1 & 0 & 0 \\ 0 & 4 & 0 & 0 & -1 & -1 \\ 0 & 0 & 4 & 0 & -1 & -1 \\ -1 & 0 & 0 & 4 & 0 & 0 \\ 0 & -1 & -1 & 0 & 4 & 0 \\ 0 & -1 & -1 & 0 & 0 & 4 \end{bmatrix}$$

jeweils eine Verbindung von $i = 1$ nach $j = 2$? Wie sehen hier die zugeordneten Graphen aus? Sind B und C irreduzibel?

Es ist gar nicht so einfach, ausgehend von der Matrix die Form des zu Grunde liegende Rechengitters zu rekonstruieren. Die Beispielmatrizen A, B und C stammen von einfachen Gittern, deren Punkte sehr „durcheinander“ nummeriert sind.

B 

C 

D 

Bei „vernünftiger“ Nummerierung sieht die Matrixstruktur regelmäßiger aus. Jedenfalls illustrieren diese Aufgaben auch:

Die Matrixstruktur hängt von der Nummerierung der Gitterpunkte ab. Umnummerieren bedeutet gleichzeitiges Vertauschen entsprechender Zeilen und Spalten in der Matrix.

$A =$ **tridiag**$(-1, 2, -1)$ **ist nicht singulär** Für diese Matrix lässt sich das am einfachsten durch Gauß-Elimination nachrechnen. Aber wir zeigen das in der Form: aus $A\mathbf{x} = \mathbf{0}$ folgt $\mathbf{x} = \mathbf{0}$. Das ist nämlich eine Beweismethode, die auch allgemein für irreduzible diagonaldominante Matrizen gilt.

Wenn (indirekt angenommen) ein Vektor $\mathbf{x} \neq \mathbf{0}$ die Gleichung $A\mathbf{x} = \mathbf{0}$ erfüllt, dann gibt es mindestens eine Komponente $x_i \neq 0$. Wir wählen ein x_i mit maximalem Betrag. Das kann nicht x_1 sein, weil es gilt (erste Zeile von $A\mathbf{x} = \mathbf{0}$ umgeformt)

$$2x_1 = x_2 \ .$$

Das wäre ein Widerspruch zu $|x_1| \geq |x_2|$. Und es kann auch nicht (gleiches Argument) x_n sein. Also ist es ein x_i mit $i \in \{2, \ldots, n-1\}$. Dann gilt (Gleichung in Zeile i von $A\mathbf{x} = \mathbf{0}$ umgeformt)

$$x_i = \frac{1}{2}(x_{i-1} + x_{i+1})$$

Diese Gleichung sagt: x_i ist der Mittelwert seiner beiden Nachbarn. Es können also nicht beide Nachbarn kleiner als x_i sein. Ebenso wenig können beide Nachbarn größer sein. Es kann aber schon gar nicht ein Nachbar größer und der andere kleiner sein, weil wir $|x_i| \geq |x_j| \quad \forall j$ angenommen haben.

Daher müssen $x_{i+1} = x_i = x_{i-1}$ *alle gleich* sein. Dann gilt aber auch (in gleicher Weise argumentierend)

$$x_i = x_{i-1} = x_{i-2} = \cdots = x_2 = x_1 \ .$$

(Hier habern wir die offensichtliche Tatsache verwendet, dass im Rechengitter x_i mit x_{i-1} und weiter bis schließlich x_1 verbunden ist. Im allgemeinen Beweis geht hier die Irreduzibilität ein.)

Gleichzeitig gilt auch $2x_1 = x_2$ (siehe oben). Daraus folgt schließlich:

$$x_i = 0 \quad \forall i$$

Bemerkung: Dieser Beweis lässt sich mit ganz ähnlicher Argumentation, aber etwas mehr technischem Aufwand allgemein für irreduzible, diagonaldominante Matrizen führen.

Deswegen sind Diagonaldominanz und Irreduzibilität wichtige Eigenschaften: sie garantieren die Lösbarkeit des Finite-Differenzen-Gleichungssystems.

A ist invers-positiv Eine Matrix X heißt *invers-positiv*, wenn X^{-1} existiert und (elementweise) $X^{-1} \geq 0$ gilt.

Für die (-1;2;-1)-Tridiagonalmatrix A gilt sogar: $A^{-1} > 0$.

Dass A^{-1} existiert, haben wir gerade gezeigt. Der Beweis von $A^{-1} > 0$ argumentiert ähnlich. Für die k-te Spalte der Inversen gilt:

$$A\mathbf{x} = \mathbf{e}_k \qquad (k\text{-ter Einheitsvektor auf der rechten Seite}) \tag{15}$$

Der k-te Einheitsvektor hat Komponenten

$$\mathbf{e}_k = [\delta_{ki}],$$

(mit Kronecker-δ-Symbol; nicht zu verwechseln mit lokalem Diskretisierungsfehler $\boldsymbol{\delta}$)

Zeilenweise ausgeschrieben:

$$x_1 = \frac{1}{2}x_2 + \delta_{k1} \tag{16}$$

$$x_j = \frac{1}{2}(x_{j-1} + x_{j+1}) + \delta_{kj} \qquad (j = 2, \ldots, n-1)$$

$$x_n = \frac{1}{2}x_{n-1} + \delta_{kn} \tag{17}$$

Weglassen des δ_{kj}-Terms macht daraus die Ungleichungen

$$x_j \geq \frac{1}{2}(x_{j-1} + x_{j+1}) \qquad \text{für } j = 2, \ldots, n-1$$

Es können nicht alle Komponenten in $\mathbf{x}$ gleich sein, denn dann würde aus einer der beiden Randgleichungen 16,17 folgen: $x_j = 0 \; \forall j = 1, \ldots, n$. Dann kann aber Gleichung 15 nicht erfüllt sein; Widerspruch.

Sei nun x_i die kleinste Komponente in $\mathbf{x}$ (oder eine davon, falls es mehrere gibt), und wenigstens einer der beiden Nachbarn sei echt größer. Angenomen, $i \in 2, \ldots, n-1$, dann gilt für so ein kleinstes x_i

$$x_i \geq \frac{1}{2}(\underbrace{x_{i-1}}_{\geq x_i} + \underbrace{x_{i+1}}_{\geq x_i}) > \frac{1}{2}(x_i + x_i)$$

Das hieße aber $x_i > x_i$; Widerspruch. Also ist $i = 1$ oder $i = n$. Wir nehmen an $i = 1$; der andere Fall lässt sich analog behandeln.

Wäre $x_1 < 0$, dann wäre laut Randgleichung 16 $x_2 = 2x_1 - \delta_{k1} < x_1$, wiederum ein Widerspruch. Ebenso ergibt $x_1 = 0$ einen Widerspruch zu den Voraussetzungen.

Es bleibt schließlich als einzige Möglichkeit: die kleinste Komponente x_i muss > 0 sein.

Monotonie vereinfacht: „Quellterm ≥ 0 bewirkt Lösung ≥ 0“. Genauer: Eine Matrix M heißt *monoton*, wenn aus $M\mathbf{u} \geq 0$ folgt $\mathbf{u} \geq 0$ (jeweils zeilenweise).

Für die Monotonie der (-1;2;-1)-Tridiagonalmatrix A gilt sogar: Aus $A\mathbf{u} \geq 0$ und $A\mathbf{u} \neq 0$ folgt $\mathbf{u} > 0$. In Worten: selbst wenn nur eine einzige Komponente des Quellterms positiv ist und alle anderen gleich Null, sind alle Komponenten der Lösung positiv.

Diese Aussage folgt direkt aus derEigenschaft $A^{-1} > 0$.

Fehlerabschätzung Der Vektor $\mathbf{v} = [v_j]$ mit $v_j = \frac{1}{2}j(n-j+1)$ erfüllt

$$A\mathbf{v} = \mathbf{1} \text{ (der lauter-Einsen-Vektor)} \tag{18}$$

Die maximale Komponente von $\mathbf{v}$ liegt (bei ungeradem n) „in der Mitte“, für $j = (n+1)/2$, Wert $v_j = \frac{1}{8}(n+1)^2$.

Die Idee der nun folgenden Argumentation ist: Wenn auf der rechten Seite betragsmäßig nicht mehr als $\mathbf{1}$ steht, ist (wegen Monotonie oder, gleichbedeutend, $A^{-1} \geq 0$) auch die Lösung betragsmäßig nicht größer als $\mathbf{v}$. Und weiter: Wenn gilt

$$A\mathbf{x} = \mathbf{b}$$

mit rechter Seite $\|\mathbf{b}\|_\infty \leq 1$, dann gilt auch für die Lösung $\|\mathbf{x}\|_\infty \leq \frac{1}{8}(n+1)^2$

Gleichbedeutend mit

$$\|A^{-1}\|_\infty \leq \frac{1}{8}(n+1)^2$$

für die (-1;2;-1)-Tridiagonalmatrix A. (Bei ungeradem n gilt Gleichheit.) Für die Matrix $A = -\frac{1}{h^2}\text{tridiag}(-1,2,-1)$ in Gleichung (12) kürzt sich wegen $h = \frac{1}{n+1}$ der $(n+1)^2$-Term. Für diese skalierte Matrix A gilt dann

$$\|A^{-1}\|_\infty \leq \frac{1}{8}$$

Entschuldigung, das ist nicht klug, sowohl die mit h^2 skalierte als auch die unskalierte Tridiagonalmatrix mit demselben Symbol A zu bezeichnen...

Es gilt dann (siehe Abschnitt Konvergenz) die Abschätzung

$$\|\epsilon\|_\infty \leq \|A^{-1}\|_\infty \|\delta\|_\infty \leq \frac{1}{8} \cdot h^2 \frac{1}{12}\|u''''\|_\infty = h^2 \frac{1}{96}\|u''''\|_\infty$$

2 Deriving Difference Approximations for Derivatives

Consider a smooth function $f(x)$. ("Smooth" here means that as many derivatives of f exixt as we may need.)

We will illustrate, by way of example, a method to find approximations to derivatives. It is sometimes called the *method of undetermined coefficients.*

Example: Derive an approximation for f'' at gridpoint x_0 using values at gridpoints x_{-1}, x_0 and x_1 of an equidistant grid with spacing h.

Expand f_{-1} and f_1 in Taylor series around x_0

$$\begin{aligned} f_{-1} &= f_0 - hf_0' + \frac{h^2}{2}f_0'' - \frac{h^3}{6}f_0''' + \frac{h^4}{24}f_0^{\text{iv}} + \mathcal{O}(h^5) \\ f_1 &= f_0 + hf_0' + \frac{h^2}{2}f_0'' + \frac{h^3}{6}f_0''' + \frac{h^4}{24}f_0^{\text{iv}} + \mathcal{O}(h^5) \end{aligned}$$

We now seek a linear combination of the three values f_{-1}, f_0 and f_1 approximating f_0'' in the form

$$f_0'' \approx af_{-1} + bf_0 + cf_1.$$

To determine the coefficients a, b and c, we insert the Taylor series and collect terms in $f_0, f_0', \ldots$. We find

$$\begin{aligned} f_0'' \approx\ & (a+b+c)f_0 + h(-a+c)f_0' + \frac{h^2}{2}(a+c)f_0'' + \\ & \frac{h^3}{6}(-a+c)f_0''' + \frac{h^4}{24}(a+c)f_0^{\text{iv}} + \mathcal{O}(h^5) \end{aligned}$$

Since we have three degrees of freedom (three undetermined coefficients), we can plan on setting the coefficients of f_0 and f_0' equal to zero and the coefficient of f_0'' equal to one. We are left with the following system of equations.

$$\begin{aligned} a+b+c &= 0 \\ -a+c &= 0 \\ a+c &= \frac{2}{h^2} \end{aligned}$$

Solving the above system of equations gives

$$a = \frac{1}{h^2}, \quad b = -\frac{2}{h^2}, \quad c = \frac{1}{h^2}.$$

As an additional bonus, this solution makes the coefficient of f_0''' equal to zero too. The coefficient of f_0^{iv} turns out to be $h^2/12$. Hence, we get the approximation

$$\frac{1}{h^2} f_{-1} - \frac{2}{h^2} f_0 + \frac{1}{h^2} f_1 = f_0'' + \frac{h^2}{12} f_0^{\mathrm{iv}} + \mathcal{O}(h^3).$$

This is the well-known centered difference approximation that we have used already.

As an additional exercise, you may derive a five-point centered approximation to f_0'''. The system of equations for five coefficients a, b, c, d and e in this case will be

$$\begin{aligned} a+b+c+d+e &= 0 \\ -2a-b+d+2e &= 0 \\ 4a+b+d+4e &= 0 \\ -8a-b+d+8e &= 6/h^3 \\ 16a+b+d+16e &= 0 \end{aligned}$$

(Using an algebraic manipulator such as Maple or Mathematica makes solving such systems much easier!) The solution is

$$a = -\frac{1}{2h^3}, \quad b = \frac{1}{h^3}, \quad c = 0, \quad d = -\frac{1}{h^3}, \quad e = \frac{1}{2h^3}.$$

An alternative method: find an interpolating polynomial $p(x)$ for the data points (x_i, f_i). Differentiate the polynomial with respect to x and evaluate at x_0.

If you do this by hand, you should use the *Lagrange interpolation formula*:

The polynom $p(x)$ that interpolates the $n+1$ pairs of values

$$(x_0, f_0),\ (x_1, f_1), \ldots, (x_n, f_n)$$

is given by

$$p(x) = f_0 L_0(x) + f_1 L_1(x) + \cdots + f_n L_n(x)\ ,$$

where

$$L_k(x) = \frac{(x-x_0)(x-x_1)\cdots(x-x_{k-1})(x-x_{k+1})\cdots(x-x_n)}{(x_k-x_0)(x_k-x_1)\cdots(x_k-x_{k-1})(x_k-x_{k+1})\cdots(x_k-x_n)}$$

for each $k = 0, 1, \ldots, n$.

In Mathematica, this method conveniently finds difference approximations for higher derivatives.

The commands

```
data={{-2h,f[-2h]},{-h,f[-h]},{0,f[0]},{h,f[h]},{2h,f[2h]}};
poly=InterpolatingPolynomial[data,x];
appr=D[poly,{x,3}]/.x->0 //Together
```

find an interpolating polynomial and evaluate its third derivative at $x = 0$. Mathematica gives

$$\text{appr} = \frac{-f(-2\,h) + 2\,f(-h) - 2\,f(h) + f(2\,h)}{2\,h^3}.$$

We may find the truncation error by the command

```
Series[appr,{h,0,3}]
```

which produces the output

$$f^{(3)}(0) + \frac{f^{(5)}(0)\,h^2}{4} + O(h)^4.$$

Since many tables provide difference approximations for $f', f'', \ldots$, it is normally not necessary to derive basic formulae. However, for irregular grids, complicated equations or some specific boundary conditions, these methods are of great value.

3 Non-Equidistant Grid

Write

$$f^- = f(x - h^-), \quad f^0 = f(x), \quad f^+ = f(x + h^+),$$
$$h^- = r_i - r_{i-1}, \quad h^+ = r_{i+1} - r_i.$$

Then

$$f'(x) = \frac{(f^+ - f^0)\frac{h^-}{h^+} + (f^0 - f^-)\frac{h^+}{h^-}}{h^- + h^+} - \frac{h^+h^-}{6} f'''$$

$$f''(x) = \frac{2}{h^- + h^+}\left(\frac{f^+ - f^0}{h^+} - \frac{f^0 - f^-}{h^-}\right) - \frac{h^+ - h^-}{3} f''' - \frac{(h^+)^2 - h^+h^- + (h^-)^2}{12} f^{\text{iv}}$$

Note that in the second approximation, for $h^+ \neq h^-$ now there is a low-order truncation error term involving f'''. Equidistant grids usually provide approximations with higher-order truncation errors. So you have to balance the gain by finer meshsize against possibly higher truncation error.

4 Linear solvers

4.1 The Thomas Algorithm for tridiagonal Systems

The so-called Thomas Algorithm is just a form of elimination for solving tridiagonal systems of linear equations. Llewellyn H. Thomas used it around 1950 to solve elliptic partial differential equations. The attribution to Thomas seems to be more common in some engineering disciplines than it is in numerical analysis.

Let A be a tridiagonal matrix and b the right-hand side of a system $Ax = b$.

$$A = \begin{bmatrix} d_1 & e_1 & & & & & \\ c_2 & d_2 & e_2 & & & & \\ & c_3 & d_3 & e_3 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & c_{n-1} & d_{n-1} & e_{n-1} \\ & & & & c_n & d_n \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_{n-1} \\ b_n \end{bmatrix}.$$

Transform it to upper triangular form, by Gaussian elimination. Use backsubstitution to solve the resulting equivalent system $Ax' = b'$

$$A' = \begin{bmatrix} 1 & e'_1 & & & & & \\ & 1 & e'_2 & & & & \\ & & 1 & e'_3 & & & \\ & & & \ddots & \ddots & & \\ & & & & 1 & e'_{n-1} \\ & & & & & 1 \end{bmatrix}, \quad b' = \begin{bmatrix} b'_1 \\ b'_2 \\ b'_3 \\ \vdots \\ b'_{n-1} \\ b'_n \end{bmatrix}.$$

The expressions for the entries in A' and b' follow from a step-by-step application of the usual elimination procedure.

$$e'_1 = \frac{e_1}{d_1}, \quad b'_1 = \frac{b_1}{d_1};$$

for $i = 2, \ldots, n$:

$$e'_i = \frac{e_i}{d_i - c_i e'_{i-1}}, \quad b'_i = \frac{b_i - c_i b'_{i-1}}{d_i - c_i e'_{i-1}};$$

Backsubstitution: $x_n = b'_n$, und für $i = n-1, \ldots, 1$:

$$x_i = b'_i - e'_i x_{i+1}.$$

This algorithm is stable if

$$\begin{cases} d_i > 0, & i = 1, 2, \ldots, n, \\ d_1 > |e_1|, & \\ d_i \geq |c_i| + |e_i| \text{ und } c_i \neq 0, e_i \neq 0, & i = 2, \ldots, n-1, \\ d_n \geq |c_n|. & \end{cases}$$

Essentially these conditions require A to be (weakly) diagonally dominant. They are sufficient but not necessary.