1 Nichtlineare Gleichungen in einer Unbekannten

1.1 Ein kurzer Rundgang im Garten der Gleichungen

Als Einstieg in die Numerische Mathematik behandeln wir numerische Lösungsverfahren für Gleichungen in einer Unbekannten. *Linear* sind solche Gleichungen, wenn sie sich in der Form

$$kx + d = 0, \quad k, d \text{ gegeben}, \quad x \text{ gesucht}$$

schreiben lassen. Offensichtlich gibt es, falls $k \neq 0$, eine eindeutige Lösung. Dieses Thema ist also abgehakt und wir kümmern uns vorläufig nur mehr um *nichtlinearen* Gleichungen. (Die linearen Gleichungen werden uns erst dann intensiver beschäftigen, wenn sie in Massen, als Systemen mit *mehreren* Unbekannten auftreten.)

Wenn sich durch Äquivalenzumformungen die Lösung einer Gleichung explizit, also in der Form $x = \ldots$, anschreiben lässt (im obigen Beispiel: $x = -d/k$), spricht man von einer *analytischen* Lösung.

Analytisch lösbar sind beispielsweise *quadratische* Gleichungen, also solche, die sich als

$$x^2 + px + q = 0 \quad p, q \text{ gegeben}, \quad x \text{ gesucht}$$

schreiben lassen. Sie kennen sicherlich die Lösungsformel

$$x_{1,2} = -\frac{p}{2} \pm \sqrt{\frac{p^2}{4} - q}$$

Es reicht aber nicht, eine Lösungsformel hinschreiben zu können, sie muss auch genaue Ergebnisse liefern. Die scheinbar triviale Lösung einer quadratischen Gleichung nach obiger Formel kann recht ungenau werden. Lassen Sie Ihren Taschenrechner damit die kleinere Lösung der quadratischen Gleichung

$$x^2 - 1234567x + 8 = 0$$

berechnen. Der (zehnstellig) genaue Wert ist $x_1 = 6{,}480004730 \cdot 10^{-6}$. Obwohl übliche Rechner zehn- bis vierzehnstellig genau rechnen, liefern sie nur die ersten paar Stellen richtig. Die numerisch genauere Methode berechnet zuerst die betragsmäßig *größere* Lösung x_1 mit der klassischen Formel und findet dann die betragsmäßig *kleinere* Lösung x_2 mit der alternativen Lösungsformel

$$x_2 = \frac{q}{x_1}\,.$$

Lineare, quadratische und *kubische* Gleichungen sind die einfachsten Beispiele *polynomialer* Gleichungen. Ein Polynom in einer Variablen x ist eine Summe von x-Potenzen, multipliziert mit *Koeffizienten*, also ein Ausdruck der Form

$$a_n x^n + \cdots + a_2 x^2 + a_1 x + a_0.$$

Die höchste auftretende Potenz heißt die *Ordnung* des Polynoms oder der Gleichung.

Kubische Gleichungen und Gleichungen vierter Ordnung sind im Prinzip analytisch lösbar, aber die Formeln (Cardanische Formeln, N. Tartaglia, G. Cardano[1], L. Ferrari, um 1540) sind so unhandlich, dass sie praktisch kaum verwendet werden. Numerische Verfahren für solche Gleichungen sind rechnerisch einfacher und eleganter. Sie liefern Näherungen, die schrittweise, mit immer besserer Genauigkeit, die Lösungen anstreben.

[1] auch bekannt durch Kardanwelle und kardanische Aufhängung, die er ebenfalls nicht erfunden hat

Der junge norwegische Mathematiker Niels Henrik Abel führt 1826 den „Beweis der Unmöglichkeit, algebraische Gleichungen von höheren Graden als dem vierten allgemein aufzulösen“. Ab dem fünften Grad lassen sich Gleichungen also (im Allgemeinen) nicht durch eine *endliche Zahl elementarer Rechenoperationen* (Addition, Subtraktion, Multiplikation, Division, ganzzahliges Wurzelziehen) lösen.

Um die Vorstellung der verschiedenen Gleichungstypen zum Abschluss zu bringen: Gleichungen, in denen nur elementare Rechenoperationen vorkommen, heißen *algebraisch*. Eine Gleichung oder Funktion, die sich nicht mittels endlich vieler elementarer Rechenoperationen formulieren lässt, ist etwas, das die Kräfte der Algebra übersteigt (*„quod vires algebrae transcendit“*) und heißt deswegen *transzendent*. Beispielsweise sind die trigonometrischen Funktionen, die Exponentialfunktion und die entsprechenden Umkehrfunktionen transzendente Funktionen.

1.2 Begriffe, Probleme, Lösungen

Hier behandelte Aufgabentypen:

$$\begin{aligned} g(x) &= h(x), && \text{(Finden einer } \textit{Lösung} \text{ einer Gleichung)} \\ f(x) &= 0, && \text{(Finden einer } \textit{Nullstelle} \text{ der Funktion } f\text{)} \\ x &= f(x), && \text{(Finden eines } \textit{Fixpunktes} \text{ der Funktion } f\text{)} \end{aligned}$$

Unter einer *Nullstelle* der Funktion f versteht man eine Lösung der Gleichung $f(x) = 0$. Unter einem *Fixpunkt* der Funktion f versteht man eine Lösung der Gleichung $x = f(x)$.

Die Nullstellen-Aufgabe $f(x) = 0$ und die Fixpunkt-Aufgabe $x = f(x)$ haben im allgemeinen nicht die gleichen Lösungen. Aber die Gleichung $f(x) = 0$ lässt sich *umformen* und auf Fixpunkt-Form bringen. Dann steht aber nicht $f(x)$, sondern ein anderer Term $\phi(x)$ in der Fixpunkt-Gleichung. Wir schreiben deswegen

$$x = \phi(x) \ ,$$

wenn die Fixpunkt-Gleichung durch Umformen der Gleichung

$$f(x) = 0$$

entstanden ist.

Wichtige Begriffe

Nullstellen von Polynomen nennt man auch *Wurzeln*.[2]

Eine *analytische Lösung* ist ein expliziter Ausdruck für die Lösung, in dem nur bekannte Größen und Funktionen vorkommen.

Welche Funktionen dabei als „bekannt“ vorausgesetzt werden, ist nicht exakt festgelegt. Letztlich lassen sich auch von so geläufigen Funktionen wie Sinus oder Cosinus Werte nur durch numerische Verfahren berechnen – auch wenn Ihnen der Taschenrechner diese Arbeit abnimmt.

[2] Allerdings klingt „Wurzel“ statt „Lösung“ oder „Nullstelle“ im heutigen Fachdeutsch eher veraltet; im Englischen ist *root of a polynomial* der gängige Fachausdruck, und auch *root of a function or an equation* ist neben *zero of a function or solution of an equation* durchaus üblich.

Demgegenüber steht die *numerische Lösung*, eine Rechenvorschrift, die eine schon irgendwie bekannte Näherung schrittweise verbessert.

Mehrfache Nullstellen: Eine Funktion $f(x)$ hat für $x = a$ eine genau n-fache Nullstelle, wenn zugleich $f(a) = 0, f'(a) = 0, f''(a) = 0, \ldots, f^{(n-1)}(a) = 0$ und $f^{(n)} \neq 0$. (Dabei setzen wir die Existenz stetiger Ableitungen mindestens bis zur n-ten Ordnung voraus.)

Die auftretenden Funktionen $f, g, \ldots$ und Variablen $x, y, \ldots$ bezeichnen in dieser Vorlesung in der Regel *reelle* Größen. Die *komplexen* Zahlen sind an sich der natürliche Lebensraum polynomialer Gleichungen (unter anderem deswegen, weil Polynome n-ten Grades dort immer genau n Nullstellen haben, Fundamentalsatz der Algebra). Die meisten Definitionen und Verfahren lassen sich leicht für komplexe Variable und komplexwertige Funktionen verallgemeinern. Trotzdem beschränken wir uns (abgesehen von gelegentlichen Hinweisen) auf Rechenverfahren in den reellen Zahlen.

Checkliste zum Lösen nichtlinearer Gleichungen

Gleichzeitig Inhaltsangabe und Stoffübersicht für diesen Abschnitt.

- Vorarbeiten
  - Überblicken Sie den Verlauf der Funktionen (Wertetabelle, graphische Darstellung).
  - Definitionsbereich? Wo können die Lösung liegen? Wie viele Lösungen gibt es?
  - Lassen sich günstige Umformungen finden?
- Trivialmethoden für Computer oder Taschenrechner
  - Systematisches Einsetzen in Wertetabelle
  - Hineinzoomen im Funktionsgraph
- Klassische Lösungsverfahren
  - Intervallhalbierung
  - Sekantenmethode und Regula Falsi
  - Newton-Verfahren (heißt auch Newton-Raphson-Verfahren)
  - Fixpunkt-Iteration (allgemeine Formulierung, wichtig wegen theoretischer Fundierung)

1.3 Beispiele zum Aufwärmen

In den Übungen und in der Vorlesung diskutieren wir Beispiele der folgenden Art.

Aus der Finanzmathematik

Ein Kredit von 100.000€ soll in 180 Monatsraten zu je 900€ zurückgezahlt werden. Was ist der Zinssatz bei diesen Konditionen?

Die Rentenformel für nachschüssige Zahlung liefert für den (monatlichen) Aufzinsungsfaktor q die Gleichung

$$900 = 100\,000\,\frac{q-1}{1-q^{-180}}. \tag{1}$$

Zustandsgleichung eines realen Gases

Wie groß ist das Molvolumen von Stickstoff bei 20 C und 1 bar = 10^5 Pa nach der Van der Waals-Gleichung?

Die Zustandsgleichung

$$\left(p + \frac{a}{V_{mol}^2}\right)(V_{mol} - b) = RT$$

beschreibt den Zusammenhang zwischen Druck p, Molvolumen V_{mol} und Temperatur T. Die Konstanten a und b haben für Stickstoff die Werte

$$a = 0.129\,\mathrm{Pa\,m^6/mol^2}, \quad b = 38.6 \times 10^{-6}\,\mathrm{m^3/mol}.$$

Die molare Gaskonstante ist $R = 8.3145\,\mathrm{J/molK}$. Nach Einsetzen der Zahlenwerte verbleibt als Gleichung für V_{mol}:

$$\left(100\,000 + \frac{0{,}129}{V_{mol}^2}\right)(V_{mol} - 0{,}0000386) = 2437{,}4 \tag{2}$$

Widerstände in Rohrleitungen und Armaturen

Der sogenannten Widerstandsbeiwert λ hängt von der Reynoldszahl Re ab. Bei laminarer Strömung gilt einfach $\lambda = 64/Re$. Im turbulenten Bereich, ab etwa $Re > 2000$, listen technische Handbücher verschiedene, teilweise empirische Formeln für λ. Auf theoretischem Weg hat Prandtl für ein glattes Rohr die Beziehung

$$\lambda = \frac{1}{(2\log_{10}(Re\sqrt{\lambda}) - 0{,}8)^2} \tag{3}$$

abgeleitet, die bis $Re = 3{,}4 \cdot 10^6$ mit Versuchen übereinstimmt. Wie groß ist λ bei $Re = 10^6$?

Schlicht und ergreifend

Es ist gut, wenn die bisherigen Beispiele Ihnen den Eindruck einer gewissen Praxisnähe vermittelt haben. Der technischen Hintergrund der Gleichungen und die damit verbundenen Verständnisschwierigkeiten verstellen aber den Blick auf die mathematischen Inhalte. Sie lernen hier nicht Physik, sondern numerische Verfahren, und die lassen sich leichter an einfachen Musterbeispielen illustrieren. Deswegen:

Finden Sie die Lösungen der Gleichung

$$3\cos x = \log x \tag{4}$$

1.4 Graphische Lösung: Ein Bild sagt mehr als tausend Formeln

Entsprechend der Checkliste aus Kapitel 1.2 verschaffen wir uns am Beispiel von Gleichung (4) einen ersten Überblick. Die Gleichung (4) läßt nicht unmittelbar erkennen, ob, wo und wieviele Lösungen sie hat. Da sowohl Cosinus als auch Logarithmus geläufige Funktionen sind, bietet sich eine graphische Darstellung an. (Abbildung 1). Aus dem Schaubild lässt sich die Anzahl und ungefähre Lage der Lösungen erkennen. Rechenprogramme, die Wertetabellen berechnen oder in einen Funktionsgraphen hineinzoomen können, liefern rasch brauchbare Werte (die Checkliste nennt diese Vorgangsweisen Trivialmethoden).

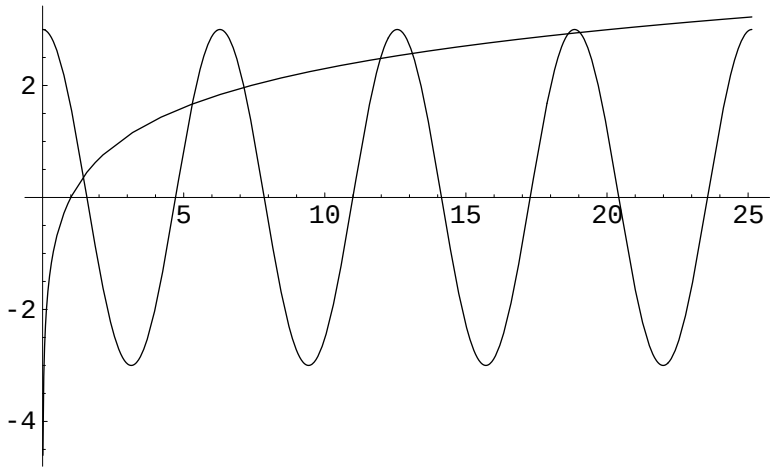

Abbildung 1: Schaubild zur Gleichung $3\cos x = \log x$. Den x-Werten der Schnittpunkte der Funktionsgraphen entsprechen die Lösungen der Gleichung.

Wichtiger Hinweis: hier meint log natürlich den natürlichen Logarithmus[3]. Argumente in Winkelfunktionen sind immer im Bogenmaß einzusetzen!

1.5 Passende Umformungen: Nullstellen und Fixpunkte

Die Lösungen der Gleichung $3\cos x = \log x$ sind genau die Nullstellen der Funktion $f(x) = 3\cos x - \log x$. Ein Vergleich von Abbildung 1 mit Abbildung 2 stellt diesen Sachverhalt klar und zeigt zum Beispiel: In der Nähe von $x = 5$, jedenfalls im Bereich $4 < x < 6$, muss eine der Nullstellen von f liegen.

[3]Für den dekadischen Logarithmus sprechen (außer der evolutionsbedingten Zufälligkeit, dass Menschen zehn Finger haben) kaum Argumente. Für Leute, die nicht bis drei zählen können, ist die Basis $e = 2{,}7182818\ldots$ ohnedies natürlicher.

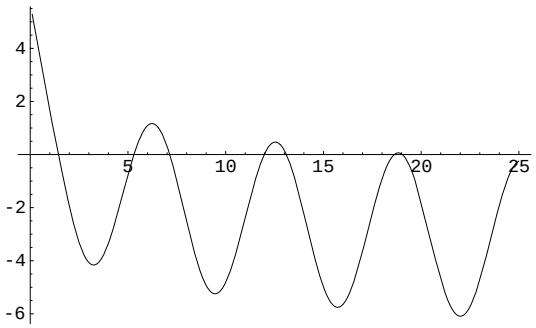

Abbildung 2: Schaubild zur Funktion $f(x) = 3\cos x - \log x$. Die Nullstellen lassen sich direkt ablesen und entsprechen den x-Werten der Schnittpunkte in der Abbildung 1

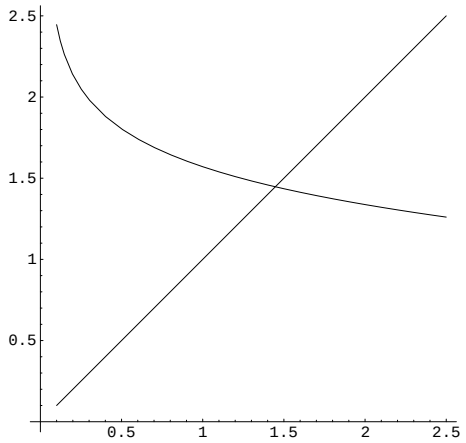

Abbildung 3: Schaubild zur Fixpunktaufgabe mit der Funktion $\phi(x) = \arccos((\log x)/3)$. Der Fixpunkt von ϕ entspricht der Nullstelle von f in der Nähe von 1,4. Weitere Fixpunkte von ϕ gibt es nicht. Durch die Umformulierung sind Lösungen der ursprünglichen Gleichung verlorengegangen!

Welche Form der graphischen Darstellung man günstigerweise wählt, hängt von der gegebenen Gleichung ab. In diesem Beispiel lassen sich cos und log als bekannte Funktionen leicht skizzieren, deswegen ist die Darstellung der Lösung durch die (x-Werte der) Schnittpunkte zweier Kurven übersichtlich. Andererseits lässt die Darstellung von $f(x) = 3\cos x - \log x$ die Nullstellen unmittelbar erkennen. Die klassischen Methoden zum Finden von Nullstellen ab Kapitel 1.7 erfordern ohnedies eine solche Umformung der Gleichung.

Die Gleichung $3\cos x = \log x$ lässt sich aber auch beispielsweise umformen zu

$$x = \arccos \frac{\log x}{3} \quad . \tag{5}$$

In dieser Form liegt eine Fixpunkt-Aufgabe $x = \phi(x)$ vor, mit $\phi(x) = \arccos((\log x)/3)$.

Was passiert, wenn man auf der rechten Seite von Gleichung (5) einen Wert für x einsetzt, den Ausdruck ausrechnet und das Ergebnis wieder in der rechten Seite einsetzt? Beginnend etwa mit $x = 1$ liefert dieses Verfahren die Folge

$$1; \quad 1{,}5708; \quad 1{,}41969; \quad 1{,}45372; \quad 1{,}44576; \quad 1{,}44761; \quad 1{,}44718\ldots$$

Die Folge konvergiert gegen $\xi = 1{,}4472586$, das ist die kleinste Lösung der gegebenen Gleichung und gleichzeitig der einzige Fixpunkt der Funktion

$$\phi(x) = \arccos \frac{\log x}{3}.$$

Sie sehen hier ein Beispiel einer *Fixpunkt-Iteration*.

Fixpunkt-Iteration (locker formuliert)
Gegeben eine Gleichung $x = \phi(x)$.

> Beginne mit einem Startwert
> Setze Wert auf rechter Seite der Formel ein
> Setze das Ergebnis wieder und wieder rechts in die Formel ein, bis sich die Resultate nicht mehr ändern

Weitere Beispiele von Fixpunkt-Iterationen:

- Geben Sie eine Zahl in den Taschenrechner ein und drücken Sie wiederholt auf die Wurzeltaste. Die Ergebnisse konvergieren gegen 1 (Fixpunkt von $f(x) = \sqrt{x}$).
- Geben Sie eine Zahl < 20 in den Taschenrechner ein und drücken Sie abwechselnd wiederholt auf die Tasten exp und $1/x$. Die Ergebnisse (nach dem $1/x$-Schritt) konvergieren gegen 0,56714 (Fixpunkt von $f(x) = 1/\exp x$).
- Die Berechnung der Quadratwurzel einer Zahl a war schon in der griechischen Antike ein wichtiges Problem und (für rationale Zahlen) gelöst. Die dazugehörige nichtlineare Gleichung ist
$$x^2 = a$$
Schon den Babyloniern soll die oft als Heron-Verfahren bezeichnete Iteration
$$x^{(0)} = a; \quad x^{(k+1)} = \frac{1}{2}\left(x^{(k)} + \frac{a}{x^{(k)}}\right) \text{ für } k = 0{,}1{,}2{,}\ldots$$
bekannt gewesen sein.
- Gleichung (3) ist eine Fixpunkt-Gleichung. Mit dem Startwert 0,05 liefern wenige Fixpunkt-Iterationen eine genaue Lösung.

Aber es funktioniert nicht immer: Eine andere mögliche Fixpunkt-Form von Gleichung (4) lautet
$$x = \exp(3\cos x) \quad .$$
Wenn Sie hier $x = 1$ rechts einsetzen und das für die Ergebnisse jeweils wiederholen, erhalten Sie die Folge
$$1; \quad 5{,}05768; \quad 2{,}76046; \quad 0{,}0617455; \quad 19{,}971; \quad 3{,}6805\ldots$$
Ihre Werte wechseln unregelmäßig und konvergieren nicht.

Conclusio: Viele numerische Verfahren sind im Grunde Fixpunkt-Iterationen. Nicht jede Fixpunkt-Iteration konvergiert. Passende Umformungen sind nicht immer leicht zu finden. Das rechtfertigt eine ausführliche theoretische Untersuchung solcher Verfahren im Kapitel 1.12.

1.6 Diskussion der Beispiele: Wichtige und unwichtige Terme

Die eingangs (Kapitel 1.1) erwähnte Gleichung
$$x^2 - 1234567x + 8 = 0$$
ist, wenn es um die betragskleinere der beiden Lösungen geht, eigentlich keine quadratische Gleichung! Begründung: Die gesuchte Lösung ist von der Größenordnung 10^{-5} bis 10^{-6}; der Term x^2 in der Gleichung ist also gegenüber dem linearen Term $1234567x$ um mehr als zehn

Größenordnungen kleiner. Für alle praktischen Zwecke ist eine solche Gleichung linear mit einem kleinen quadratischen Korrekturterm. Lösen Sie daher nach dem linearen Term auf:

$$x = \frac{1}{1234567}(x^2 + 8) \ .$$

Der Startwert $x^{(0)} = 0$ liefert selbst auf den billigsten Taschenrechnern ohne Wurzeltaste bereits ein bessere Näherung $x^{(1)} = 6{,}480004730 \cdot 10^{-6}$ als die meisten Rechner durch Anwendung der Standard-Lösungsformel erreichen können.

> locker formuliert: Viele Gleichungen enthalten Terme, in denen die Unbekannte zwar auftritt, aber im Vergleich zu anderen Termen wenig Einfluss hat. Solche Terme lassen sich in erster Näherung vernachlässigen oder als Korrekturterme auffassen.

Die Van der Waals-Gleichung (2) lässt sich zu einer kubischen Gleichung umformen,

$$-4.9794 \cdot 10^{-6} + 0.129 V_{mol} - 2441.3 V_{mol}^2 + 100000 V_{mol}^3 = 0 \quad , \tag{6}$$

und wäre damit im Prinzip analytisch lösbar. Tun Sie 's nicht! Ein wenig Einsicht in den physikalischen Hintergrund dieser Gleichung legt eine andere Vorgangsweise nahe: Bei Zimmertemperatur ist Stickstoff nahezu ein ideales Gas. Der Term a/V_{mol}^2 in der Van der Waals-Gleichung ist eine Korrektur der idealen Gasgleichung und für die gegebenen Parameter gegenüber p vernachlässigbar klein. Auch wenn man es dem Polynom 6 nicht ansieht: Gleichung (2) ist keine „richtige“ kubische Gleichung, sondern eigentlich eine lineare Gleichung in V_{mol} plus einem kleinen Korrekturterm a/V_{mol}^2.

Auch diese Gleichung lässt sich auflösen, wenn man „unwichtige“ Terme der Unbekannten auf der rechten Seite stehen lässt. Hier formen wir um zu

$$V_{mol} = \frac{RT}{p + a/V_{mol}^2} + b = \frac{2437{,}4}{100000 + 0{,}129/V_{mol}^2} + 0{,}0000386$$

und ignorieren wir erst einmal den Korrekturterm a/V_{mol}^2. Das liefert eine nullte Näherung für das Molvolumen,

$$V_0 = \frac{2437{,}4}{100000} + 0{,}0000386 = 0{,}024413$$

Der Trick ist nun, diese Näherung für V_{mol} in der rechten Seite der Gleichung einzusetzen und daraus eine verbesserte Näherung

$$V_1 = \frac{2437{,}4}{100000 + 0{,}129/0{,}024413^2} + 0{,}0000386 = 0{,}024360$$

zu berechnen. Wiederholtes Einsetzen liefert keine weitere Verbesserung:

$$V_2 = \frac{2437{,}4}{100000 + 0{,}129/0{,}024360^2} + 0{,}0000386 = 0{,}024360$$

Damit haben wir (jedenfalls auf fünf Dezimalstellen genau) den Wert $V_{mol} = 0.024360\,\mathrm{m}^3$ bestimmt.

> Bußübung für die Fastenzeit: Schlagen Sie in Wikipedia die Cardanischen Formeln nach und lösen Sie die Aufgabe damit. Vergleichen Sie den Zeitaufwand mit der obigen Methode.

In Gleichung (1) erwarten wir für den Aufzinsungsfaktor q einen Wert knapp über 1. Den Term q^{-180} im Nenner wird dann $\ll 1$ und nicht so wichtig sein. Das motiviert, die Gleichung nach dem q im Zähler aufzulösen.

$$q = 1 + \frac{900}{100000}(1 - q^{-180})$$

Ignoriert man q^{-180} auf der rechten Seite, dann folgt als nullte Näherung

$$q_0 = 1 + \frac{900}{100000} = 1{,}009$$

Auch hier funktioniert der Trick, q_0 in der rechten Seite einzusetzen und daraus eine verbesserte Näherung

$$q_1 = 1 + \frac{900}{100000}(1 - 1{,}009^{-180}) = 1{,}007206$$

zu berechnen. Wiederholtes Einsetzen liefert

$$q_2 = 1{,}006529 \quad q_3 = 1{,}006210 \quad q_4 = 1{,}006047\ldots$$

Es braucht aber hier insgesamt 14 Iterationen, bis sich die Werte bei $q = 1{,}005851$ stabilisieren.

Bemerkungen zum Abschluss

Ist eine Gleichung in der Form $f(x) = g(x)$ gegeben (Beispiel: Gleichung 4), lässt sich nicht unmittelbar erkennen, welche Terme „wichtig“ oder „unwichtig“ sind. Regel: man löse nach jener Seite der Gleichung auf, welcher den *steileren* Funktionsgraph im Schnittpunkt hat.

Passende Umformungen für Fixpunkt-Iterationen erfordern oft ein tieferes Verständnis der einzelnen Terme in einer Gleichung. Es gibt zum Glück Lösungsverfahren, die mehr nach „Schema F“ ablaufen. Eines davon stellt das nächste Kapitel vor.

1.7 Intervallhalbierung

Kennen Sie die Geschichte von den zwei Möglichkeiten? Sie beginnt mit dem Zwischenwertsatz.

Zwischenwertsatz

Eine Funktion f, die auf einem abgeschlossenen Intervall $[a,b]$ stetig ist, nimmt in diesem Intervall auch jeden Wert zwischen $f(a)$ und $f(b)$ an.

Ist f insbesondere für $x = a$ negativ und für $x = b$ positiv (oder umgekehrt), dann garantiert der Zwischenwertsatz: f hat mindestens eine Nullstelle in diesem Intervall.

Es gibt immer zwei Möglichkeiten...

Angenommen, wir suchen eine Nullstelle einer im Bereich $a \le x \le b$ stetigen Funktion. Es lässt sich rechnerisch sofort prüfen, ob $f(a)$ und $f(b)$ unterschiedliches Vorzeichen haben. Wenn ja, dann garantiert der Zwischenwertsatz die Existenz eine Nullstelle im Bereich $a \le x \le b$, aber wir wissen nicht, wo sie liegt. Nun gibt es zwei Möglichkeiten: Entweder ist $b - a$ klein, dann ist es gut: Wir können sowohl a als auch b als Näherung für eine Nullstelle von f auffassen. Andernfalls berechnen wir den Mittelpunkt c des Intervalls, $c = (a + b)/2$. Nun gibt es wieder zwei Möglichkeiten. Ist $f(c) = 0$, so ist es gut: es liegt dort eine Nullstelle vor. Anderenfalls hat f an den Enden eines der Teilintervalle $a \le x \le c$ oder $c \le x \le b$ verschiedene Vorzeichen (klar? Das ist der springende Punkt!). In einem der beiden Intervalle muss also eine Nullstelle liegen.

Betrachten wir dieses Intervall und nennen wir der Einfachheit die neuen Intervallgrenzen wieder a und b.

Nun gibt es zwei Möglichkeiten: Entweder ist $b-a$ klein, dann ist es gut: Wir können sowohl a als auch b als Näherung für eine Nullstelle von f auffassen. Andernfalls bilden wir $c = (a+b)/2$. Nun gibt es wieder zwei Möglichkeiten...

Sie können nun die Geschichte selber fortsetzen. Beachten Sie aber, dass die Intervalllänge in jedem Erzählschritt halbiert wird. Für jede beliebig klein vorgegebene Genauigkeitsschranke $\epsilon > 0$ erreichen Sie nach einer endlichen Anzahl von Schritten ein Intervall mit Länge $b-a < \epsilon$. Damit endet die Geschichte wie im wirklichen Leben: Es gibt immer zwei Möglichkeiten, aber jede Entscheidung schränkt den Freiraum für weitere Aktionen ein. Irgendwann sind die Alternativen dann doch ausgeschöpft.

Formalisiert angeschrieben, lautet dieses Verfahren

Intervallhalbierung (Bisektionsverfahren)
Gegeben eine Funktion $f(x)$, zwei Werte a und b mit $f(a) \cdot f(b) < 0$, eine Genauigkeitsschranke $\epsilon > 0$. Ist $f(x)$ im Intervall $a \leq x \leq b$ stetig, dann findet dieser Algorithmus die Näherung c an eine Nullstelle c_0 von f mit Genauigkeit $|c - c_0| < \epsilon$.

setze $c = (a + b)/2$
Wiederhole solange $|b - a| \geq \epsilon$ und $f(c) \neq 0$
  falls $f(a) \cdot f(c) < 0$
    ersetze $b \leftarrow c$
  sonst
    ersetze $a \leftarrow c$
  setze $c = (a + b)/2$

Lineare Konvergenz

Die beste Schätzung für den Wert der Nullstelle ist der Mittelpunkt des Intervalls. Der maximale Fehlerbetrag ist dann durch $\epsilon_0 \leq |b-a|/2$ beschränkt; größer als die halbe Intervallbreite kann er nicht sein. Intervallhalbierung reduziert diese Fehlerschranke pro Schritt um den Faktor $1/2$ oder, da

$$\left(\frac{1}{2}\right)^{3{,}3} \approx \frac{1}{10} \quad ,$$

um einen Faktor $1/10$ pro (durchschnittlich) 3,3 Schritten. Man kann sagen: Intervallhalbierung produziert eine korrekte Dezimalstelle pro 3,3 Iterationen. Der maximale Fehler nach dem i-ten Schritt, ϵ_i, ist höchstens halb so groß wie der vorherige maximale Fehler ϵ_{i-1}. Es gilt also

$$\epsilon_i \leq C\epsilon_{i-1} \quad \text{mit } C = \frac{1}{2} \, .$$

Allgemein spricht man, wenn bei einem Verfahren für die Fehlerschranken aufeinanderfolgender Iterationsschritte gilt

$$\epsilon_i \leq C\epsilon_{i-1} \quad \text{mit } C < 1 \,.$$

von *linearer* Konvergenz.

Vorteile der Intervallhalbierung: mathematisch und programmiertechnisch einfach. Wenn die Voraussetzungen erfüllt sind, konvergiert es mit Sicherheit. Es ist ein *Einschlussverfahren*, das heißt, es liefert nicht nur einen Näherungswert, sondern grenzt die Lösung von beiden Seiten her ein.

Nachteile: Man braucht Startwerte–aber das ist ein Problem jedes numerischen Verfahrens. Intervallhalbierung ist langsam; nur lineare Konvergenz – die dafür aber sicher.

1.8 Regula Falsi (lineares Eingabeln)

Funktionen, die in der Umgebung der Nullstelle glatt verlaufen, lassen sich dort durch eine Gerade annähern. Statt, wie bei der Intervallhalbierung, den Wert c genau in der Mitte zwischen a und b anzunehmen, wählen wir c als Nullstelle der Gerade durch $(a,f(a))$ und $(b,f(b))$, Siehe Abbildung 4.

$$c = a - f(a)\frac{a-b}{f(a)-f(b)} = \frac{af(b)-bf(a)}{f(b)-f(a)}$$

Regula Falsi (lineares Eingabeln)
Gegeben eine Funktion $f(x)$, zwei Werte a und b mit $f(a)\cdot f(b) < 0$ und eine Genauigkeitsschranke $\epsilon > 0$. Ist $f(x)$ im Intervall $a \leq x \leq b$ stetig, dann findet dieser Algorithmus die Näherung c an eine Nullstelle c_0 von f mit Genauigkeit $|c - c_0| < \epsilon$.

Wiederhole
  setze $c \leftarrow a - f(a)\dfrac{a-b}{f(a)-f(b)}$
  falls $f(b)\cdot f(c) < 0$
    setze $a \leftarrow b$
  sonst
    (klassische Version) nix
    (Illinois-Variante) reduziere $f(a)$ auf $\frac{1}{2}f(a)$
    (Pegasus-Variante) reduziere $f(a)$ auf $\dfrac{f(a)f(b)}{f(b)+f(c)}$
  setze $b \leftarrow c$
bis $|b-a| < \epsilon$ oder $f(c) = 0$

Für extrem bösartige Funktionen kann die Intervallhalbierung immer noch rascher als die klassische Regula Falsi konvergieren. Es gibt auch keine Garantie, dass sich die Intervalllänge pro Schritt zumindest halbiert. Sorgfältige Programmierer würden im obigen Algorithmus jedenfalls noch eine Notbremse einbauen: zähle die Anzahl der Iterationen mit und brich ab, wenn eine Maximalzahl überschritten wird.

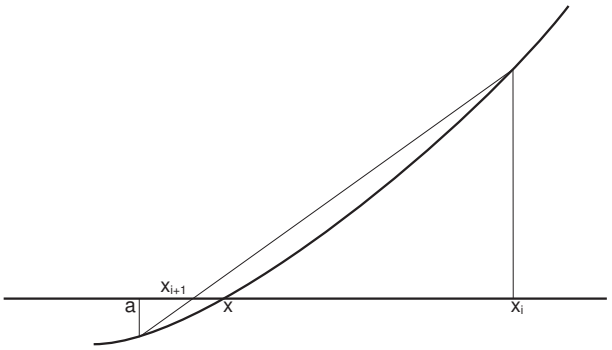

Abbildung 4: Die Regula Falsi gewinnt den nächsten Näherungswert $x^{(i+1)}$ aus der Nullstelle der Verbindungsgeraden.

Die Illinois- oder die Pegasus-Variante verbessern das Konvergenzverhalten im Vergleich zur Intervallhalbierung deutlich; mutige Programmierer würden in diesem Fall auf die Abfrage nach einer maximalen Iterationszahl verzichten.

Intervallhalbierung und die verschiedenen Regula-Falsi-Versionen haben gemeinsam, daß sie die Nullstelle von beiden Seiten her „eingabeln“ — sie sind Einschlussverfahren, das ist gut. Nachteilig ist, dass man zu Beginn des Verfahrens zwei Näherungswerte braucht, und zwar je einen auf jeder Seite der Nullstelle. Das kann sehr schwer zu erreichen sein, wenn man zwei nahe beisammen liegende Nullstellen hat, da dann eine der ursprünglichen Näherungen dazwischen liegen muß. Mehrfache Nullstellen gerader Ordnung kann man mit diesen Verfahren überhaupt nicht finden.

Was ist „falsch“ an der Regula Falsi? Natürlich nicht die Regel selbst, sondern die angenommenen Startwerte a und b. Aus diesen beiden „falschen Lösungen“ berechnet die Regel eine bessere Näherungslösung.
Die Methode ist uralt, die Grundidee war schon Jahrhunderte vor Chr. weltweit bekannt: Babyloniern, Ägypter, Inder und Chinesen lösten damit lineare Gleichungen. Aus arabischen Quellen nach Europa bringt sie um 1200 Leonardo von Pisa, genannt Fibonacci. Er beschreibt mehrere Varianten, darunter die *regula duarum falsarum positionum*, die „Methode vom doppelten falschen Ansatz“. So sollte sie auch richtiger Weise heißen, aber es hat sich schlampig verkürzt „Regula Falsi“ durchgesetzt.
Fibonacci löste damit nur lineare Probleme; da berechnet die Regel aus zwei falschen Startwerten sofort die richtige Lösung. Die Anwendung als iteratives Verfahren für Nullstellen nichtlinearer Funktionen ist dann doch nicht so alt. Mitte des vorigen Jahrhunderts fand man sogar noch kleine, aber nicht unwesentliche Verbesserungen der Rechenregel (Illinois-Variante)

1.9 Sekantenmethode

Die Sekantenmethode berechnet gleich wie die Regula Falsi eine neue Näherung durch lineare Interpolation, verzichtet aber auf den Einschluss der Nullstelle, siehe Abbildung 5.

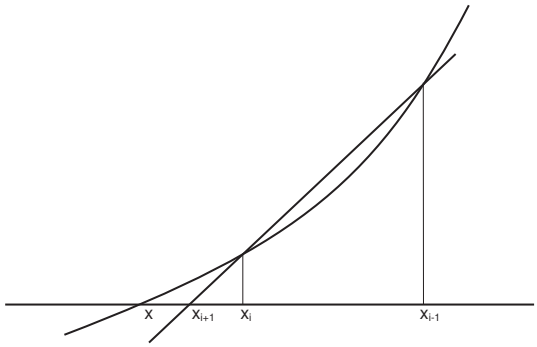

Abbildung 5: Die Sekantenmethode gewinnt den nächsten Näherungswert c mittels einer Schnittgeraden (Sekante) durch zwei Punkte des Funktionsgraphen. Die jeweils letzten beiden Näherungen schließen die Nullstelle jedoch nicht unbedingt ein.

Sekantenmethode

Gegeben eine Funktion $f(x)$, zwei Werte $x^{(0)}$ und $x^{(1)}$, eine Genauigkeitsschranke $\epsilon > 0$ und eine maximale Iterationsanzahl k_{max}. Für hinreichend gute Startwerte $x^{(0)}$ und $x^{(1)}$ findet dieser Algorithmus die Näherung $x^{(k)}$ an eine Nullstelle x von f mit Genauigkeit $|x^{(k)} - x| < \epsilon$ oder bricht nach einer Maximalzahl von k_{max} Schritten ab.

setze $k = 1$
Wiederhole
  setze $x^{(k+1)} = x^{(k)} - f(x^{(k)}) \dfrac{x^{(k)} - x^{(k-1)}}{f(x^{(k)}) - f(x^{(k-1)})}$
  erhöhe $k = k + 1$
bis $|x^{(k+1)} - x^{(k)}| < \epsilon$ oder $k >= k_{max}$

Superlineare Konvergenz

Die Sekantenmethode zeigt *superlineare* Konvergenz. Das heißt, für die Fehlerschranken $\epsilon_{k+1} = |x^{(k+1)} - x|$ und $\epsilon_k = |x^{(k)} - x|$ aufeinanderfolgender Schritte gilt, sofern ϵ_k schon hinreichend klein ist:

$$\epsilon_{k+1} \leq C\epsilon_k^p \quad \text{mit } p > 1.$$

Der Fehler reduziert sich also nicht bloß um einen Faktor C, sondern zusätzlich noch mit der Potenz p. Für die Sekantenmethode lässt sich zeigen

$$p = \frac{1 + \sqrt{5}}{2} \approx 1{,}618.$$

Angenommen, es ist $\epsilon_k = 0{,}01$. Überlegen Sie sich, was mehr bewirkt: Multiplikation mit einem Faktor $C = 1/2$, oder Potenzieren mit $p = 1{,}6$!

1.10 Newton-Verfahren

Heißt auch Newton-Raphson-Verfahren, aber erst einige Jahrzehnte nach Isaac Newton und Joseph Raphson formuliert Thomas Simpson das Verfahren so, wie wir es heute kennen.

Gesucht sei eine Nullstelle der Funktion $f(x)$. Gegeben sei ein Startwert $x^{(0)}$ in der Nähe der Nullstelle. Das Newton-Verfahren versucht, ähnlich der Sekantenmethode, die Funktion f durch eine lineare Funktion anzunähern und verwendet dazu die Tangente an f im Punkt $(x^{(0)},f(x^{(0)}))$. Der Schnittpunkt der Tangente mit der x-Achse ist der nächste Näherungswert, siehe Abbildung 6.

Herleitung aus der Taylorentwicklung von f um den Punkt $x^{(0)}$. Ist f genügend oft differenzierbar, dann gilt:

$$f(x) = f(x^{(0)}) + (x - x^{(0)})f'(x^{(0)}) + \frac{(x - x^{(0)})^2}{2!}f''(x^{(0)}) + \dots$$

Es soll gelten $f(x) = 0$. Vernachlässigen von Gliedern höherer Ordnung liefert die Gleichung

$$0 = f(x^{(0)}) + (x - x^{(0)})f'(x^{(0)})$$

aus der sich x ausdrücken lässt:

$$x = x^{(0)} - \frac{f(x^{(0)})}{f'(x^{(0)})} \ .$$

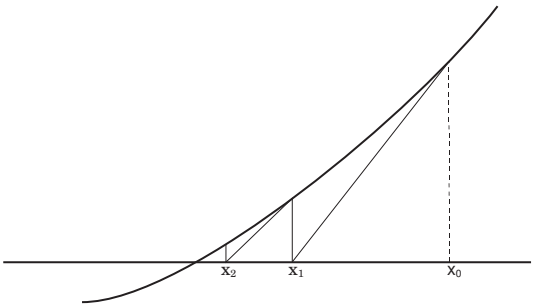

Abbildung 6: Graphische Deutung des Newton-Verfahrens: Der Schnittpunkt der Tangente an f im Punkt $(x^{(0)},f(x^{(0)}))$ mit der x-Achse liefert den verbesserten Näherungswert $x^{(1)}$.

Newton-Verfahren

Gegeben eine differenzierbare Funktion $f(x)$ und ein Startwert $x^{(0)}$. Gesucht eine Nullstelle von f.

Iterationsvorschrift

$$x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})} \qquad \text{für } k = 0,1,2\dots$$

Quadratische Konvergenz

Das Newton-Verfahren zeigt *quadratische* Konvergenz. Das heißt, für die Fehlerschranken $\epsilon_{k+1} = |x^{(k+1)} - x|$ und $\epsilon_k = |x^{(k)} - x|$ aufeinanderfolgender Schritte gilt, sofern ϵ_k schon hinreichend klein ist:

$$\epsilon_{k+1} \leq C\epsilon_k^2$$

Der neue Fehler ist also um einen Faktor C kleiner als das *Quadrat* des alten Fehlers. Der genaue Wert von C ist dabei nicht so wichtig.

Angenommen, es ist $\epsilon_k = 10^{-4}$. Das heisst, der Fehler beträgt eine Einheit in der vierten Nachkommastelle. Dann gilt bei quadratischer Konvergenz $\epsilon_{k+1} = C \cdot 10^{-8}$. Der Fehler beträgt also C Einheiten in der achten Nachkommastelle. Wenn C größenordnungsmäßig im Bereich 1 ist, hat sich die Anzahl der korrekten Stellen ungefähr verdoppelt.

Quadratische Konvergenz: Neuer Fehler $\sim$ Quadrat des alten Fehlers.

Faustregel: Sofern schon einige signifikante Stellen exakt sind, sind im nächsten Näherungswert etwa doppelt so viele signifikante Stellen korrekt.

1.11 Abbruchbedingungen

Vergessen Sie nie, dass Rechner nur eine fixe Zahl von Binärstellen zur Verfügung haben, um Gleitkommazahlen zu speichern. Möglicherweise erreicht $f(x)$ für kein Gleitkomma-Argument x exakt den Wert Null. Wenn die Nullstelle x_0 in der Gegend von 1 liegt, können Sie leicht eine Näherung x mit absolutem Fehler $|x - x_0| < 10^{-6}$ finden. Liegt die Nullstelle um $x \approx 10^{22}$, werden Sie einen absoluten Fehler dieser Güte nicht erreichen können. Eine übliche Wahl der Abbruchschranke ϵ ist $\varepsilon_m(|a| + |b|)/2$, wenn ε_m die Maschinengenauigkeit und a,b die ursprünglichen Intervallgrenzen sind. Wenn a,b und die Nullstelle selber nahe bei Null liegen, ist Vorsicht bei dieser Formel geboten. Die Abbruchschranke darf jedenfalls nicht kleiner als die kleinste positive Maschinenzahl sein (typischerweise um 10^{-38} für 4-Byte-Datentypen, 10^{-308} für 8-Byte-Datentypen).

Maschinengenauigkeit

Die Maschinengenauigkeit ε_m ist die kleinste positive Gleitkommazahl, die, zur Gleitkommazahl 1.0 addiert, eine von 1.0 verschiedene Summe ergibt (typischerweise um 10^{-7} für 4-Byte-Datentypen, 10^{-16} für 8-Byte-Datentypen).

1.12 Fixpunkt-Iteration

Im Abschnitt 1.5 haben wir bereits Fixpunkte von Funktionen durch wiederholtes Einsetzen bestimmt. Viele numerische Verfahren lassen sich als Spezialfälle einer Fixpunkt-Iteration betrachten. Aussagen über die Konvergenz von Fixpunkt-Iterationen sind deswegen von allgemeiner Bedeutung.

Fixpunkt-Iteration

Gegeben eine Funktion $\phi(x)$ und ein Startwert $x^{(0)}$. Gesucht ein Fixpunkt ξ von ϕ.

$x^{(0)}$ als Startwert gegeben.
Iterationsvorschrift
$x^{(k+1)} = \phi(x^{(k)})$ für $k = 0,1,2\ldots$

Fixpunkt-Iteration konvergiert für kontrahierende Abbildungen

Die Funktion $\phi(x)$ besitze einen Fixpunkt ξ: $\phi(\xi) = \xi$. Sei ferner I ein offenes Intervall der Form $(\xi - r,\xi + r)$ um den Fixpunkt ξ, sodass ϕ in I eine *kontrahierende Abbildung* ist, d. h.

$$|\phi(x) - \phi(y)| \leq C|x - y| \,, \quad C < 1$$

gilt für alle $x,y \in I$.
Dann konvergiert die Fixpunkt-Iteration $x^{(k+1)} = \phi(x^{(k)})$ mindestens linear gegen ξ für alle $x^{(0)} \in I$.

Beweis: Zuerst zeigt man durch Induktion: $x^{(k)} \in I$ für alle $k = 0,1,2\ldots$: Die Aussage ist laut Voraussetzung richtig für $k = 0$. Nun ist

$$|x^{(k+1)} - \xi| = |\phi(x^{(k)}) - \phi(\xi)| \leq C|x^{(k)} - \xi| \,.$$

Nach der Induktionsannahme liegt $x^{(k)} \in I$, also weniger als r von ξ entfernt: $|x^{(k)} - \xi| < r$. Da $C < 1$, ist also auch

$$|x^{(k+1)} - \xi| < r \quad \text{und} \quad x^{(k+1)} \in I$$

Aus dem Induktionsbeweis folgt unmittelbar für die Fehler $\epsilon^{(k)} = |x^{(k)} - \xi|$ und $\epsilon^{(k+1)} = |x^{(k+1)} - \xi|$:

$$\epsilon^{(k+1)} \leq C\epsilon^{(k)} \leq C^k\epsilon_0 \,.$$

Bemerkung: Ist ϕ in einer Umgebung von ξ stetig differenzierbar und$|\phi'(\xi)| < 1$, so ist in einer Umgebung von ξ die Kontraktionseigenschaft erfüllt: Wegen der Stetigkeit von ϕ' gibt es ein offenes Intervall I um ξ, in dem $\phi' \leq C < 1$ gilt. Für $x,y \in I$ gilt nach dem Mittelwertsatz der Differentialrechnung

$$\phi(x) - \phi(y) = (x - y)\phi'(\eta) \quad \text{für } \eta \in I \,.$$

Damit ist auch

$$|\phi(x) - \phi(y)| \leq C|x - y| \,, \quad C < 1$$

Eine Kurzfassung dieser Aussage:

Das Fixpunktverfahren konvergiert lokal, falls $|\phi'(\xi)| < 1$.

Das Konvergenzverhalten des Algorithmus für verschiedene f wird in Abbildung 7 graphisch dargestellt.

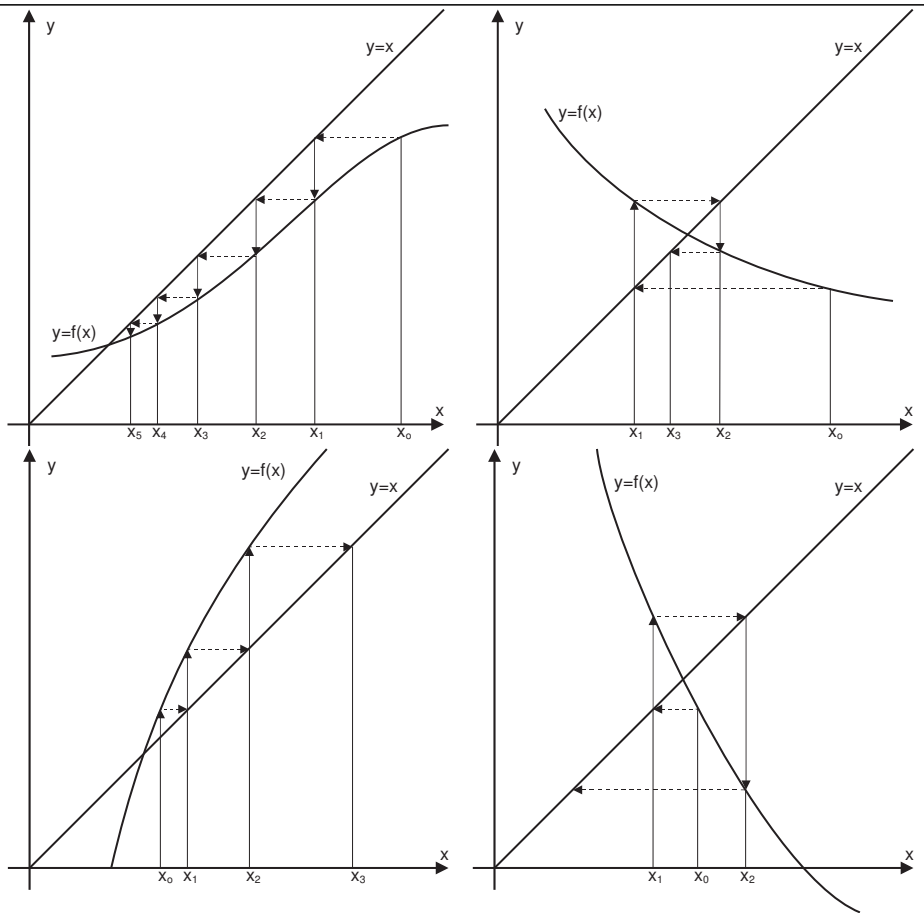

Abbildung 7: Fixpunkt-Iteration in graphischer Darstellung für verschiedene Funktionen f. Mögliche Fälle: Einseitige Annäherung an den Fixpunkt, falls in einer Umgebung des Fixpunktes $0 < f' < 1$; alternierende Konvergenz, falls $-1 < f' < 0$, Divergenz falls $f' > 1$ oder $f' < -1$.

1.13 Konvergenzordnung

Wir haben lineare, superlineare und quadratische Konvergenz bereits erwähnt. Hier fassen wir den Begriff der Konvergenzordnung genauer.

Konvergenzordnung
Sei ξ Fixpunkt von $\phi(x)$, und es gelte für alle Startwerte aus einem Intervall um ξ und die zugehörige Folge $\{x^{(k)}\}$ aus der Vorschrift $x^{(k+1)} = \phi(x^{(k)}), k = 0,1,2\ldots$

$$|x^{(k+1)} - \xi| \leq C|x^{(k)} - \xi|^p$$

mit $p \geq 1$ und $C < 1$, falls $p = 1$.
Das Iterationsverfahren heißt dann ein Verfahren von mindestens p-ter Ordnung

Für das lokale Konvergenz-verhalten einer Fixpunkt-Iteration ist der Wert der ersten Ableitung am Fixpunkt maßgeblich. Für $|\phi'(\xi)| < 1$ ist lineare Konvergenz gesichert; je kleiner der Betrag der Ableitung, desto schneller konvergiert das Verfahren. Ist sogar $|\phi'(\xi)| = 0$, dann können wir superlineare Konvergenz zeigen.

Es gilt: Ist $\phi(x)$ in einer Umgebung von ξ genügend oft differenzierbar und

$$\phi'(\xi) = 0, \phi''(\xi) = 0, \ldots, \phi^{(p-1)}(\xi) = 0, \text{ und } \phi^{(p)}(\xi) \neq 0\ ,$$

dann liegt für $p = 2,3,\ldots$ ein Verfahren p-ter Ordnung vor. Ein Verfahren erster Ordnung liegt vor, wenn zu $p = 1$ gilt: $|\phi'(\xi)| < 1$.

1.14 Konvergenz des Newton-Verfahrens

Das Newtonverfahren entspricht einem Fixpunkt-Verfahren für die Funktion ϕ,

$$\phi(x) = x - \frac{f(x)}{f'(x)}$$

Nun ist

$$\phi'(x) = \frac{f''(x)f(x)}{(f'(x))^2}\ ,$$

und da an einer einfachen Nullstelle $f(x) = 0, f'(x) \neq 0$ gilt, verschwindet $\phi'(x)$ dort. Man überzeugt sich leicht, dass $\phi''(x) \neq 0$ gilt, sofern $f''(x) \neq 0$. Daraus folgt die quadratische Konvergenz des Newtonverfahrens bei einfachen Nullstellen. Bei mehrfachen Nullstellen lässt sich lineare Konvergenz nachweisen.

2 Systeme nichtlinearer Gleichungen

2.1 Fixpunkt-Iteration, mehrdimensionaler Fall

Fixpunkt-Iterationen sind auch im mehrdimensionalen Fall möglich. Ein Fixpunkt einer Abbildung $\mathbf{\Phi} : \mathbb{R}^n \to \mathbb{R}^n$ ist – völlig analog zur eindimensionalen Definition – ein Wert $\boldsymbol{\xi} \in \mathbb{R}^n$, für den gilt:

$$\boldsymbol{\xi} = \mathbf{\Phi}(\boldsymbol{\xi}) \; .$$

Genauso wie im eindimensionalen Fall findet Fixpunkt-Iteration (falls sie konvergiert) einen Fixpunkt. Wir setzen hier Vektoren aus dem $\mathbb{R}^n$ und vektorwertige Funktionen in fetter Schrift ($\mathbf{\Phi}, \boldsymbol{\xi}, \mathbf{x} \ldots$), zum Unterschied von Variablen und reellwertigen Funktionen ($\phi, \xi, x, \ldots$). Sonst ändert sich nichts am Schema der Fixpunkt-Iteration.

Fixpunkt-Iteration, mehrdimensional

Gegeben sei eine Abbildung $\mathbf{\Phi} : \mathbb{R}^n \to \mathbb{R}^n, \quad \mathbf{x} \to \mathbf{\Phi}(\mathbf{x})$. Gesucht ein Fixpunkt $\boldsymbol{\xi}$ von $\mathbf{\Phi}$.

$\mathbf{x}^{(0)}$ als Startwert gegeben.
Iterationsvorschrift
$\mathbf{x}^{(k+1)} = \mathbf{\Phi}(\mathbf{x}^{(k)})$ für $k = 0,1,2 \ldots$

Beispiel: Fixpunkt-Iteration für ein System zweier nichtlinearer Gleichungen

Gegeben sei das nichtlineare Gleichungssystem (log ist natürlich der natürliche Logarithmus)

$$\begin{aligned} 4x - y + xy - 1 &= 0 \\ -x + 6y + \log(xy) - 2 &= 0 \end{aligned}$$

Ausgehend von der Näherungslösung $x_0 = 1$ und $y_0 = 1$ bestimme man durch geeignete Fixpunkt-Iteration verbesserten Näherungen.

In der Nähe des Startwertes hängt die erste Gleichung am stärksten vom Term $4x$ ab; die zweite Gleichung von $6y$. Vorgangsweise: löse die beiden Gleichungen jeweils nach diesen Termen auf.

$$\begin{aligned} x &= \frac{1}{4}(y - xy + 1) \\ y &= \frac{1}{6}(x - \log(xy) + 2) \end{aligned}$$

Die Funktion $\mathbf{\Phi}$ ist hier ein Vektor aus zwei reellwertigen Funktionen ϕ und ψ, der Vektor $\mathbf{x}$ hat zwei Komponenten x und y.

$$\mathbf{\Phi}(\mathbf{x}) = \begin{bmatrix} \phi(x,y) \\ \psi(x,y) \end{bmatrix} = \begin{bmatrix} \frac{1}{4}(y - xy + 1) \\ \frac{1}{6}(x - \log(xy) + 2) \end{bmatrix}$$

Iteration liefert die Folge $(1;1)$, $(1/4;1/2)$, $(0{,}34375;0{,}721574)$, $(0{,}368383;0{,}622985), \ldots$, die gegen den Fixpunkt $(0{,}35344388;0{,}63996847)$ konvergiert.

Normen

Exakte Lösung, Näherungslösung und Fehler sind bei Gleichungssystemen jeweils Vektoren im $\mathbb{R}^n$. Wir brauchen ein Maß für die „Größe“ oder „Länge“ des Fehlervektors, oder für den „Abstand“ der Näherung von der exakten Lösung. Im eindimensionalen Fall messen wir die „Größe“ von x mit dem Absolutbetrag $|x|$, und den Abstand zweier Werte x und y auf der reellen Achse durch $|y - x|$.

Während es aber in $\mathbb{R}$ nur eine sinnvolle Definition für den Absolutbetrag gibt, stehen im $\mathbb{R}^n$ mehrere Möglichkeiten offen. Da ist zunächst einmal die „übliche“ Definition für die Länge eines Vektors, auch *euklidische* Länge oder *2-Norm* genannt. Oft lässt sich aber mit anderen Normen einfacher arbeiten. Wir verwenden noch die *1-Norm* und die *∞-Norm*.

Normen im $\mathbb{R}^n$
Für einen Vektor $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i| \quad , \quad \text{Einsnorm}$$

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} (x_i)^2} \quad , \quad \text{euklidische Norm, Zweinorm}$$

$$\|\mathbf{x}\|_\infty = \max_i |x_i| \quad , \quad \text{Unendlich-Norm, Maximums-Norm}$$

Konvergenz

Die Konvergenz der mehrdimensionalen Fixpunkt-Iteration lässt sich in der gleichen Weise zeigen wie im eindimensionalen Fall, wenn eine Kontraktionseigenschaft vorliegt. Auch der Begriff der Konvergenzordnung läßt sich unter Verwendung von Normen geradewegs auf den mehrdimensionalen Fall übertragen.

Fixpunkt-Iteration konvergiert für kontrahierende Abbildungen $\mathbb{R}^n \to \mathbb{R}^n$
Die Funktion $\mathbf{\Phi}(x)$ besitze einen Fixpunkt $\boldsymbol{\xi}$: $\mathbf{\Phi}(\boldsymbol{\xi}) = \boldsymbol{\xi}$. Sei ferner B eine offene Umgebung um den Fixpunkt $\boldsymbol{\xi}$ in der Form $B = \{\mathbf{x} : \|\boldsymbol{\xi} - \mathbf{x}\| < r\}, \quad r > 0$, sodass $\mathbf{\Phi}$ in B eine *kontrahierende Abbildung* ist, d. h. es gilt

$$\|\mathbf{\Phi}(\mathbf{x}) - \mathbf{\Phi}(\mathbf{y})\| \leq C\|\mathbf{x} - \mathbf{y}\| \, , \quad C < 1$$

für alle $\mathbf{x}$,$\mathbf{y} \in B$ in einer Norm $\|\cdot\|$.
Dann konvergiert die Fixpunkt-Iteration $\mathbf{x}^{(k+1)} = \mathbf{\Phi}(\mathbf{x}^{(k)})$ mindestens linear gegen $\boldsymbol{\xi}$ für alle $\mathbf{x}^{(0)} \in B$.

Der Beweis erfolgt analog zu der eindimensionalen Form des Konvergenzsatzes.

Ob eine Abbildung kontrahierend ist, hängt von den partiellen Ableitungen ab. Man kann zeigen: die mehrdimensionale Fixpunkt-Iteration konvergiert lokal in einer Umgebung des Fixpunktes, wenn dort für die partiellen Ableitungen von $\mathbf{\Phi}$ gilt

$$\sum_{i=1}^{n} \left| \frac{\partial \phi_i}{\partial x_k} \right| \leq C < 1 \qquad \text{für} \qquad k = 1, \ldots, n.$$

2.2 Newton-Verfahren für Systeme

Gegeben sei eine vektorwertige Funktion $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$. Gesucht sei ein Vektor $\mathbf{x} \in \mathbb{R}^n$ als Lösung von

$$\mathbf{f}(\mathbf{x}) = \mathbf{0}$$

Dies ist die allgemeine Formulierung eines Systems von n linearen oder nichtlinearen Gleichungen in n Unbekannten. Wir setzen hier Vektoren aus dem $\mathbb{R}^n$ und vektorwertige Funktionen in fetter Schrift ($\mathbf{x}$,$\mathbf{f}(\mathbf{x}), \ldots$), zum Unterschied von Variablen und reellwertigen Funktionen (x,$f(x), \ldots$). Komponentenweise ausgeschrieben lautet das Gleichungssystem mit

$$\mathbf{f} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{bmatrix} \quad \text{und} \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad : \quad \begin{array}{l} f_1(x_1,x_2, \ \ldots \ ,x_n) = 0 \\ f_2(x_1,x_2, \ \ldots \ ,x_n) = 0 \\ \qquad \ldots \\ f_n(x_1,x_2, \ \ldots \ ,x_n) = 0 \end{array} .$$

Die Lösung von Systemen linearer Gleichungen ist vergleichsweise einfach gegenüber nichtlinearen Gleichungssystemen. Das Newton-Verfahren für Systeme führt die Lösung eines nichtlinearen Systems auf die Lösung einer Folge von linearen Gleichungssystemen zurück.

Sofern die entsprechenden partiellen Ableitungen existieren, definieren wir die *Jacobi-Matrix* D_f von $\mathbf{f}$ durch

$$D_f = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1} & \dfrac{\partial f_1}{\partial x_2} & \cdots & \dfrac{\partial f_1}{\partial x_n} \\ \dfrac{\partial f_2}{\partial x_1} & \dfrac{\partial f_2}{\partial x_2} & \cdots & \dfrac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & & \vdots \\ \dfrac{\partial f_n}{\partial x_1} & \dfrac{\partial f_n}{\partial x_2} & \cdots & \dfrac{\partial f_n}{\partial x_n} \end{bmatrix}$$

Damit lässt sich $\mathbf{f}$ in der Umgebung eines Punktes $\mathbf{x}^{(0)}$ in linearisierter Näherung schreiben (Taylorscher Lehrsatz für Funktionen mehrerer Veränderlicher) als

$$\mathbf{f}(\mathbf{x}^{(1)}) = \mathbf{f}(\mathbf{x}^{(0)}) + D_f(\mathbf{x}^{(0)}) \cdot (\mathbf{x}^{(1)} - \mathbf{x}^{(0)}) + \mathbf{R}$$

mit einem Restglied $\mathbf{R}$, das im Limes $\mathbf{x} \to \mathbf{x}^{(0)}$ mit höherer Ordnung verschwindet. Wir vernachlässigen das Restglied und fordern $\mathbf{f}(\mathbf{x}^{(1)}) = 0$. Es verbleibt die Gleichung

$$0 = \mathbf{f}(\mathbf{x}^{(0)}) + D_f(\mathbf{x}^{(0)}) \cdot (\mathbf{x}^{(1)} - \mathbf{x}^{(0)}) \ ,$$

aus der $\mathbf{x}^{(1)}$ als verbesserte Näherung an die Lösung von $\mathbf{f}(\mathbf{x}) = 0$ bestimmt werden kann.

Setzen wir $\Delta\mathbf{x}^{(k)} = \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}$, so lässt sich der Iterationsschritt des Newton-Verfahrens für Systeme so formulieren:

Newton-Verfahren für Systeme

Gegeben eine differenzierbare vektorwertige Funktion $\mathbf{f}(\mathbf{x})$ und ein Startwert $\mathbf{x}^{(0)}$. Gesucht eine Nullstelle von $\mathbf{f}$.

Iterationsvorschrift

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \Delta\mathbf{x}^{(k)}$$

mit $\Delta\mathbf{x}^{(k)}$ als Lösung von $D_f(\mathbf{x}^{(k)})\Delta\mathbf{x}^{(k)} = -\mathbf{f}(\mathbf{x}^{(k)})$

Auch dieses Verfahren ist ein Fixpunktverfahren, und zwar für die Funktion

$$\mathbf{\Phi}(\mathbf{x}) = \mathbf{x} - D_f^{-1}(\mathbf{x})\mathbf{f}(\mathbf{x}).$$

Notwendig für die Durchführbarkeit ist, dass D_f^{-1} existiert.

Sind die Nullstellen einfach, so konvergiert das Verfahren jedenfalls quadratisch. Da es oft sehr mühsam ist, immer alle Elemente von D_f an jedem Punkt $\mathbf{x}^{(k)}$ zu berechnen, geht man manchmal so vor, daß man D_f an einem einzigen Punkt $\mathbf{x}^{(0)}$ berechnet und für den weiteren Verlauf des Verfahrens fix lässt. Dieses Verfahren heißt vereinfachtes Newton-Verfahren. Dabei sollte $x^{(0)}$ bereits eine brauchbare Näherung sein. Das vereinfachte Newton-Verfahren konvergiert allerdings nur linear.

Das Newton-Verfahren für Systeme erfordert also in jedem Schritt die Lösung eines linearen Gleichungssystems. Das nächste Kapitel bringt die systematische Behandlung linearer Gleichungssysteme.

Beispiel: Gleichungssystem aus Abschnitt 2.1

Die Funktion $\mathbf{f}$ und die Jacobi-Matrix D_f sind hier

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} 4x - y + xy - 1 \\ -x + 6y + \log(xy) - 2 \end{bmatrix}, \qquad D_f = \begin{bmatrix} 4 + y & -1 + x \\ -1 + \frac{1}{x} & 6 + \frac{1}{y} \end{bmatrix}.$$

Startwert $(1; 1)$ eingesetzt liefert

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} 3 \\ 3 \end{bmatrix}, \qquad D_f = \begin{bmatrix} 5 & 0 \\ 0 & 7 \end{bmatrix}.$$

Zu lösen ist also das Gleichungssystem

$$\begin{bmatrix} 5 & 0 \\ 0 & 7 \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} = - \begin{bmatrix} 3 \\ 3 \end{bmatrix}$$

Es liefert den Korrekturterm und die verbesserte Lösung

$$\Delta\mathbf{x}^{(0)} = \begin{bmatrix} -0{,}6 \\ -0{,}428571 \end{bmatrix}, \quad \mathbf{x}^{(1)} = \mathbf{x}^{(0)} + \Delta\mathbf{x}^{(0)} = \begin{bmatrix} 0{,}4 \\ 0{,}571429 \end{bmatrix}.$$

Der nächste Schritt wertet zuerst $\mathbf{f}$ und D_f für die neuen Werte von $\mathbf{x}$, löst das Gleichungssystem für den Korrekturterm $\Delta\mathbf{x}^{(1)}$ und errechnet daraus die verbesserte Näherung $\mathbf{x}^{(2)} = \mathbf{x}^{(1)} + \Delta\mathbf{x}^{(1)}$. Die Matrix D_f hat aber hier nicht mehr so „schöne“ Einträge; das Gleichungssystem ist deswegen nicht so unmittelbar lösbar wie im ersten Schritt. Das vereinfachte Newtonverfahren würde zwar $\mathbf{f}$ neu auswerten, die Matrix D_f des ersten Schrittes beibehalten. Einfacherer Rechengang, aber langsamere (nur lineare statt quadratischer) Konvergenz!