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

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

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

Analytische oder numerische Lösung Wenn sich durch algebraische Umformungen die Lösung einer Gleichung explizit, also in der Form $x = \ldots$, anschreiben lässt (im obigen Beispiel: $x = d/k$, allgemein ein Term, in dem nur die üblichen Standard-Rechenoperationen und -Funktionen auftreten), 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 verlässlich 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

$$x^2 - 12345678x + 9 = 0$$

berechnen. Der (sechzehnstellig) genaue Wert ist $x_1 = 7{,}290\,000\,597\,780\,479 \times 10^{-7}$. 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

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

Algebraische und transzendente Gleichungen 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, G. Cardano[1], N. Tartaglia[2], L. Ferrari[3], um 1540) sind so unhandlich, dass sie praktisch kaum verwendet werden. Numerische Verfahren sind in diesen Fällen meist sinnvoller. Sie liefern Näherungen, die schrittweise, mit immer besserer Genauigkeit, die Lösungen anstreben. Ab Polynomgrad fünf gibt es ohnehin keine allgemeinen Lösungsformeln mehr.

> 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, Wurzelziehen) lösen.

Um die Vorstellung der verschiedenen Gleichungstypen zum Abschluss zu bringen: Gleichungen, in denen auch noch Bruchterme, Wurzeln oder rationale Exponenten vorkommen, lassen sich (möglicher Weise nur mit hohem Aufwand und komplizierten Umformungen) auf Systeme polynomialer (man sagt auch: algebraischer) Gleichungen zurückführen. Terme oder Funktionen, die sich nicht mittels endlich vieler elementarer Rechenoperationen formulieren lassen, sind etwas, das die Kräfte der Algebra übersteigt (*„quod vires algebrae transcendit“*, sagte Leibnitz) und heißen deswegen *transzendent*.

Beispielsweise sind die trigonometrischen Funktionen, die Exponentialfunktion und die entsprechenden Umkehrfunktionen transzendente Funktionen. Treten solche Funktionen in Gleichungen auf, ist normaler Weise nur numerische Lösung möglich.

> Explizite Lösungsformeln gibt es nur für polynomiale Gleichungen niedrigen Grades und die allereinfachsten transzendenten Glaichungen. In allen anderen Fällen können nur numerische Methoden eine Lösung finden.

1.2 Begriffe, Probleme, Lösungen

Hier behandelte Aufgabentypen:

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

> Eine Lösung der Gleichung $f(x) = 0$ heißt *Nullstelle* der Funktion f.
> Eine Lösung der Gleichung $x = \phi(x)$ heißt *Fixpunkt* der Funktion ϕ.

> Eine Fixpunkt-Gleichung $x = \phi(x)$ lässt sich natürlich sofort umformen auf $\phi(x) - x = 0$. Jeder Fixpunkt von ϕ ist also zugleich Nullstelle von $f(x) = \phi(x) - x$.
> Selbstverständlich muss der Funktionsterm in einer Nullstellen-Aufgabe nicht automatisch f heißen, ebensowenig wie in Fixpunkt-Gleichungen die Funktion mit ϕ bezeichnet sein muss. Die Vorlesungsunterlagen schreiben aber in der Regel $x = \phi(x)$, wenn diese Gleichung durch Umformen aus $f(x) = 0$ entstanden ist.

[1] auch bekannt durch Kardanwelle und kardanische Aufhängung, die er ebenfalls nicht erfunden hat
[2] Niccolò Fontana Tartaglia verriet Cardano die Lösung unter dem Siegel der Verschwiegenheit; war stinksauer, als der sie trotzdem veröffentlichte.
[3] Das Rennen um die Lösung für Gleichungen vierten Grades, sozusagen die Formel Vier, wurde damals von Ferrari gewonnen.

Nullstellen von Polynomen nennt man auch *Wurzeln*.[4]

Eine *analytische Lösung* ist ein expliziter Ausdruck, 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.

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

Mehrfache Nullstellen: Eine Funktion f hat an der Stelle x eine genau n-fache Nullstelle, wenn zugleich $f(x) = 0, f'(x) = 0, f''(x) = 0, \ldots, f^{(n-1)}(x) = 0$ und $f^{(n)}(x) \neq 0$. (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 für Polynome und Funktionen (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 der folgenden Abschnitte.

- 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

1.3 Beispiele zum Aufwärmen

In den Übungen und in der Vorlesung diskutieren wir Beispiele der folgenden Art. Auch die folgenden Abschnitte 1.5 und 1.6 bringen weitere Erklärungen.

[4] 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.

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 $1\,\mathrm{bar} = 1 \times 10^5\,\mathrm{Pa}$ 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}:

$$\left(100\,000 + \frac{0{,}129}{V_{mol}^2}\right)(V_{mol} - 0{,}000\,038\,6) = 2437{,}4 \tag{2}$$

Widerstände in Rohrleitungen

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

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

abgeleitet, die bis $\mathrm{Re} = 3{,}4 \times 10^6$ mit Versuchen übereinstimmt. Wie groß ist λ bei $\mathrm{Re} = 1 \times 10^6$?

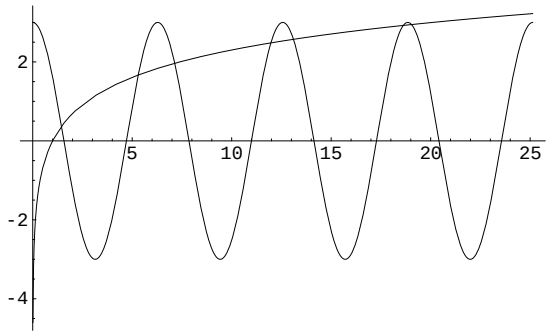

Abbildung 1: Schaubild zur Gleichung $3\cos x = \log x$. Den x-Werten der Schnittpunkte der Funktionsgraphen entsprechen die Lösungen der Gleichung.

Ohne tiefere Bedeutung

Es ist gut, wenn die bisherigen Beispiele den Eindruck einer gewissen Praxisnähe vermittelt haben. Der technischen Hintergrund und damit verbundene 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

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

Wichtiger Hinweis: hier meint log natürlich den natürlichen Logarithmus[5]. Argumente in Winkelfunktionen sind immer im Bogenmaß einzusetzen!

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. Diese Gleichung 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“).

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

[5] Für den dekadischen Logarithmus sprechen außer der evolutionsbedingten Zufälligkeit, dass Menschen zehn Finger haben, keine Argumente. Für Leute, die nicht bis drei zählen können, ist die Basis $e = 2{,}718\,281\,8\ldots$ ohnedies natürlicher.

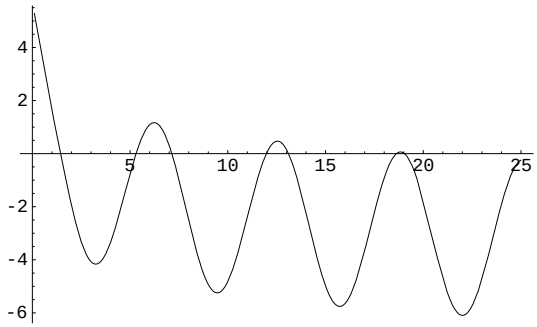

Abbildung 2: Schaubild zur Funktion $f(x) = 3\cos x - \log x$. Die Nullstellen von f entsprechen den x-Werten der Schnittpunkte in der Abbildung 1

und zeigt zum Beispiel: In der Nähe von $x = 5$, jedenfalls im Bereich $4 < x < 6$, muss eine der Nullstellen von f liegen.

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 $\phi(x) = \arccos((\log x)/3)$.

Fixpunkt-Iteration

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

$$1; \quad 1{,}5708; \quad 1{,}419\,69; \quad 1{,}453\,72; \quad 1{,}445\,76; \quad 1{,}447\,61; \quad 1{,}447\,18\ldots$$

Die Folge konvergiert gegen $\xi = 1{,}447\,258\,6$, 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
Gegeben eine Gleichung $x = \phi(x)$.

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

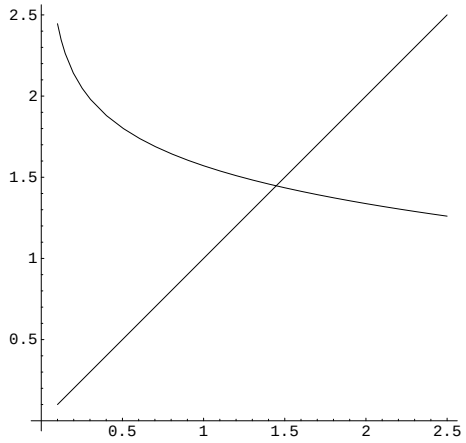

Abbildung 3: Schaubild zur Fixpunktaufgabe mit der Funktion $\phi(x) = \arccos((\log x)/3)$. Der Fixpunkt von ϕ entspricht der Nullstelle von f in der Nähe von 1,4. Weitere Fixpunkte von ϕ gibt es nicht. Durch die Umformulierung sind alle anderen Lösungen der ursprünglichen Gleichung verlorengegangen!

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 0,567 14 (Fixpunkt von $f(x) = 1/\exp x$).
- Berechnen von Quadratwurzeln war schon in der griechischen Antike ein wichtiges Problem und (für rationale Zahlen) gelöst. Die Wurzel aus a ist definiert als Lösung von $x^2 = a$; eine für $x \neq 0$ äquivalente Umformung dieser Gleichung ist

  $$x = \frac{1}{2}\left(x + \frac{a}{x}\right) \ .$$

  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

$$x = \exp(3\cos x) \ .$$

Wenn Sie hier $x = 1$ rechts einsetzen und das für die Ergebnisse jeweils wiederholen, erhalten Sie die Folge

$$1; \quad 5{,}057\,68; \quad 2{,}760\,46; \quad 0{,}061\,745\,5; \quad 19{,}971; \quad 3{,}6805\ldots$$

Ihre Werte wechseln unregelmäßig und konvergieren nicht.

Zusammenfassung

Nicht jede Fixpunkt-Iteration konvergiert. Passende Umformungen sind nicht immer leicht zu finden. Anderseits sind viele numerische Verfahren vom Typ einer Fixpunkt-Iteration. Das rechtfertigt eine ausführliche theoretische Untersuchung solcher Verfahren im Kapitel 1.12.

1.6 Diskussion der Beispiele: Wichtige und unwichtige Terme

Hier werden die in Kapitel 1.1 vorgestellten Beispiele ausführlich besprochen.

1.6.1 Eine fast lineare Gleichung

Die am Anfang von Kapitel 1.1 erwähnte Gleichung

$$x^2 - 12345678x + 9 = 0$$

ist, wenn es um die betragskleinere der beiden Lösungen geht, eigentlich keine quadratische Gleichung! Begründung: Die gesuchte Lösung ist von der Größenordnung 10^{-6} bis 10^{-7}; der Term x^2 in der Gleichung ist also gegenüber dem linearen Term $12345678x$ 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:

$$x = \frac{1}{12345678}(x^2 + 9) \; .$$

Der Startwert $x^{(0)} = 0$ liefert selbst auf den billigsten Taschenrechnern ohne Wurzeltaste bereits ein bessere Näherung $x^{(1)} = 7{,}290\,000\,597\,78 \times 10^{-7}$ 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. Wenn eine Gleichung dadurch leichter lösbar wird, lassen sich solche Terme in erster Näherung vernachlässigen. In weiteren Schritten korrigiert man das Ergebnis, indem man Näherungswerte in den anfangs vernachlässigten Termen einsetzt.

1.6.2 Van der Waals-Gleichung

Die 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, das der Gleichung

$$pV_{mol} = RT$$

gehorcht. In der Van der Waals-Gleichung

$$\left(p + \frac{a}{V_{mol}^2}\right)(V_{mol} - b) = RT \tag{7}$$

ist der Term a/V_{mol}^2 eine Korrektur der idealen Gasgleichung und für die im Beispiel gegebenen Parameter gegenüber p vernachlässigbar klein. Dem umgeformten Polynom (6) sieht man es

nicht an, aber die ursprüngliche Gleichung (7) entspricht – im Bereich der gegebenen Daten – nicht einer „richtigen“ kubischen Gleichung, sondern vielmehr einer lineare Gleichung in V_{mol} plus einem kleinen Korrekturterm a/V_{mol}^2.

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

$$V_{mol} = \frac{RT}{p + a/V_{mol}^2} + b = \frac{2437{,}4}{100000 + 0{,}129/V_{mol}^2} + 0{,}000\,038\,6$$

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

$$V_0 = \frac{2437{,}4}{100000} + 0{,}000\,038\,6 = 0{,}024\,413\ .$$

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

$$V_1 = \frac{2437{,}4}{100000 + 0{,}129/0{,}024\,413^2} + 0{,}000\,038\,6 = 0{,}024\,360$$

zu berechnen. Wiederholtes Einsetzen liefert keine weitere Verbesserung:

$$V_2 = \frac{2437{,}4}{100000 + 0{,}129/0{,}024\,360^2} + 0{,}000\,038\,6 = 0{,}024\,360\ .$$

Damit haben wir (jedenfalls auf fünf Dezimalstellen genau) den Wert $V_{mol} = 0{,}024\,360\,\mathrm{m}^3$ 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.

1.6.3 Finanzmathematik

In Gleichung 1 erwarten wir für den Aufzinsungsfaktor q einen Wert knapp über 1. Den Term q^{-180} im Nenner wird vermutlich $\ll 1$ 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

$$q_1 = 1 + \frac{900}{100000}(1 - 1{,}009^{-180}) = 1{,}007\,206$$

zu berechnen. Wiederholtes Einsetzen liefert

$$q_2 = 1{,}006\,529 \quad q_3 = 1{,}006\,210 \quad q_4 = 1{,}006\,047\ldots$$

Es braucht aber hier insgesamt 14 Iterationen, bis sich die Werte bei $q = 1{,}005\,851$ 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 $a \leq x \leq b$ 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 $a \leq x \leq b$, 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 $a \leq x \leq c$ oder $c \leq x \leq b$ 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, zwei Werte a und b mit $f(a) \cdot f(b) < 0$, eine Fehlerschranke $\epsilon > 0$. Ist f im Intervall $a \leq x \leq b$ stetig, dann findet dieser Algorithmus die Näherung c an eine Nullstelle ξ von f mit Fehler $|c - \xi| < \epsilon$.

Wiederhole
  setze $c \leftarrow (a + b)/2$
  falls $f(a) \cdot f(c) < 0$
    setze $b \leftarrow c$
  sonst
    setze $a \leftarrow c$
bis $|b - a| < \epsilon$ oder $f(c) = 0$

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: Wenn bei einem Verfahren für die Fehlerschranken aufeinanderfolgender Iterationsschritte gilt

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

spricht man von *linearer* Konvergenz.

Vor- und Nachteile

Vorteile der Intervallhalbierung: einfach zu verstehen, leicht zu programmieren. 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.

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

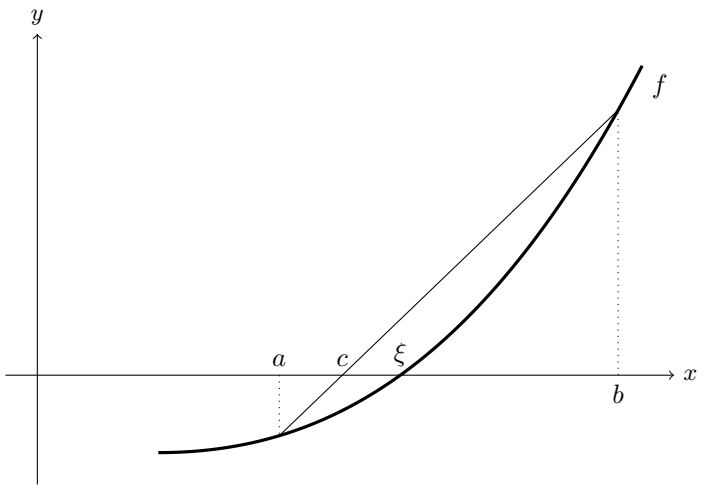

Abbildung 4: Die Regula Falsi berechnet c, die Nullstelle der Verbindungsgeraden, als Näherungswert an die Nullstelle ξ der Funktion f.

Regula Falsi (lineares Eingabeln)
Gegeben eine Funktion f, 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[a] eine Näherung c an eine Nullstelle ξ von f mit Genauigkeit $|c - \xi| < \epsilon$.

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 $f(c) = 0$

[a] mit dem hier gegebenen Abbruchkriterium allerdings nur die beiden Varianten

Allerdings bringt die Regula Falsi in der Standard-Version im Vergleich zur Intervallhalbierung kein wesentlich besseres Konvergenzverhalten. Typischer Weise bleibt nach einigen Iterationen die Intervallgrenze a fix, die andere Grenze b konvergiert zwar zur Nullstelle, aber die Abbruchbedingung $|b - a| < \epsilon$ wird nicht erreicht. 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.

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 können diese 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 kleine, aber nicht unwesentliche Verbesserungen der Rechenregel (Pegasus-, Illinois-Variante). Sogar noch kürzlich, 2020, veröffentlichten Oliveira und Takahashi eine weitere Verbesserung (`https://en.wikipedia.org/wiki/ITP_method`).

1.9 Sekantenmethode

Die Sekantenmethode berechnet gleich wie die Regula Falsi eine neue Näherung durch lineare Interpolation, verlangt aber nicht, dass die Werte a und b die Nullstelle einschließen, siehe Abbildung 5.

Die formale Beschreibung des Verfahrens bezeichnet hier die Startwerte a und b mit $x^{(0)}$ und $x^{(1)}$ und die weiteren iterativ berechneten Näherungswerte mit $x^{(k)}, x^{(k+1)}, \ldots$

Sekantenmethode
Gegeben eine Funktion f, 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 ξ von f mit Genauigkeit $|x^{(k)} - \xi| \approx \epsilon$ 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 $k \geq k_{max}$

Superlineare Konvergenz

Die Sekantenmethode zeigt *superlineare* Konvergenz. (Notwendige technische Details: f zweimal stetig differenzierbar, keine mehrfache Nullstelle.) Das heißt, für die Fehlerschranken $|x^{(k+1)} - \xi|$ und $|x^{(k)} - \xi|$ aufeinanderfolgender Schritte gilt, sofern $|x^{(k)} - \xi|$ schon hinreichend klein ist:

$$|x^{(k+1)} - \xi| \leq C|x^{(k)} - \xi|^p \quad \text{mit } p > 1 \ .$$

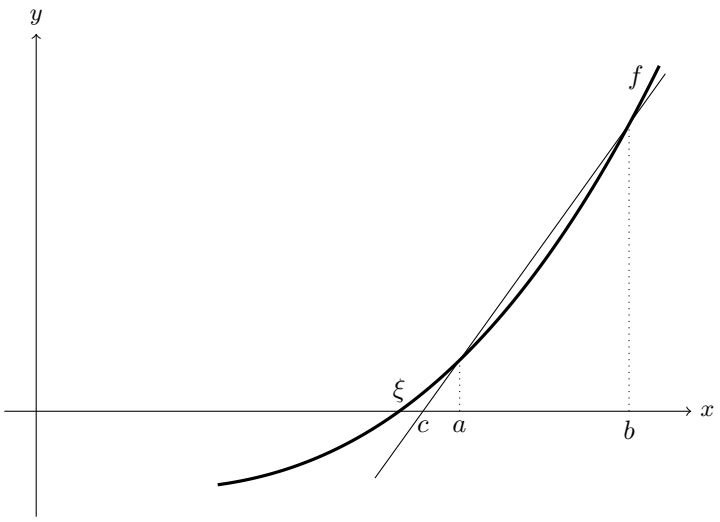

Abbildung 5: Die Sekantenmethode berechnet den nächsten Näherungswert c mittels einer Schnittgeraden (Sekante) durch zwei Punkte des Funktionsgraphen. Die beiden Werte a und b schließen die Nullstelle ξ jedoch nicht unbedingt ein.

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 $|x^{(k)} - \xi| = 0{,}01$. Überlegen Sie sich, was den Fehler stärker verringert: 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. 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

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

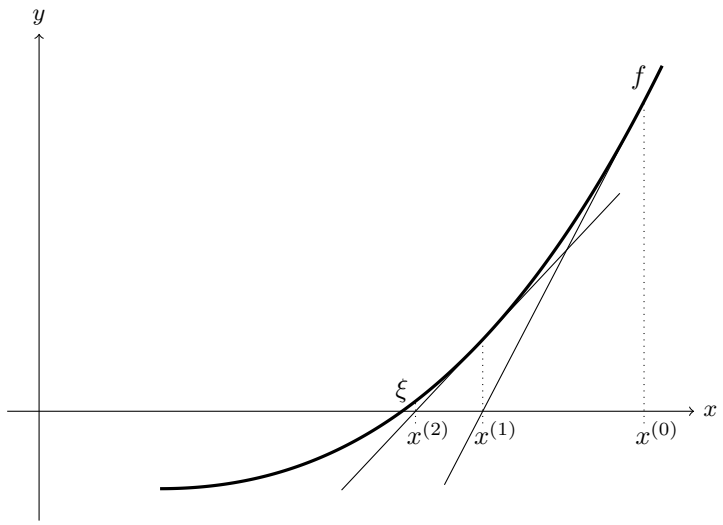

Abbildung 6: Graphische Deutung des Newton-Verfahrens: Die Tangente an f im Punkt $(x^{(0)}, f(x^{(0)}))$ schneidet die x-Achse in $x^{(1)}$. Der Wert $x^{(2)}$ im nächsten Schritt liegt schon nahe an der Nullstelle ξ.

aus der sich x ausdrücken lässt:

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

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

Iterationsvorschrift

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

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 $\epsilon_k = 1 \times 10^{-4}$. Das heißt, der Fehler beträgt eine Einheit in der vierten Nachkommastelle. Dann gilt bei quadratischer Konvergenz $\epsilon_{k+1} = C \cdot 1 \times 10^{-8}$. 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

Rechner haben nur eine fixe Zahl von Binärstellen zur Verfügung, um Gleitkommazahlen zu speichern. Möglicherweise erreicht $f(x)$ für kein Gleitkomma-Argument x exakt den Wert Null. Wenn die Nullstelle ξ in der Gegend von 1 liegt, können Sie leicht eine Näherung x mit absolutem Fehler $|x - \xi| < 10^{-6}$ finden. Liegt die Nullstelle um $\xi \approx 10^{22}$, 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 ϕ und ein Startwert $x^{(0)}$.
Gesucht ein Fixpunkt ξ von ϕ.

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

Fixpunkt-Iteration konvergiert für kontrahierende Abbildungen
Die Funktion ϕ besitze einen Fixpunkt ξ Sei ferner I ein offenes Intervall der Form $(\xi - r, \xi + r)$ um den Fixpunkt ξ, in dem ϕ als *kontrahierende Abbildung* wirkt, d. h.

$$|\phi(x) - \phi(y)| \leq C|x - y| \text{ gilt mit } C < 1 \text{ für alle } x, y \in I \ .$$

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

Beweis: Zuerst zeigt man durch Induktion: $x^{(k)} \in I$ für alle $k = 0, 1, 2 \ldots$ Die Aussage ist laut Voraussetzung richtig für $k = 0$. Angenommen, es liegt bereits $x^{(k)} \in I$, also weniger als r von ξ entfernt: $|x^{(k)} - \xi| < r$. Dann können wir die Kontraktionsbedingung und Fixpunkt-Eigenschaft für $x^{(k)}$ und ξ anwenden und erhalten

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

Da $C < 1$, ist also auch

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

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

$$\epsilon^{(k+1)} \leq C\epsilon^{(k)} \leq C^k \epsilon_0, \text{ somit } \epsilon^{(k+1)} \to 0 \text{ für } k \to \infty \; .$$

So wie der Satz hier formuliert ist, setzt er die Existenz eines Fixpunktes voraus. Dadurch wird der Konvergenz-Beweis kurz und schmerzlos. Eine etwas allgemeinere Formulierung und ein technisch aufwändigerer Beweis zeigen, dass aus der Kontraktions-Eigenschaft auch schon die Existenz und Eindeutigkeit eines Fixpunktes folgen. Das ist der berühmte Fixpunktsatz von Banach.

Zusammenhang kontrahierende Abbildung–Steigung der Funktion

Die Eigenschaft $|\phi(x) - \phi(y)| \leq C|x - y|$ bedeutet für $C < 1$ anschaulich: Funktionswerte unterscheiden sich weniger als die Eingabewerte. Wie stark sich Funktionswerte im Verhältnis zu Eingabewerten ändern, ist (im Grenzwert für kleine Änderungen) durch die Steigung der Funktion bestimmt.

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

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

Damit ist auch

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

Eine Kurzfassung dieser Aussage:

Abbildung 7 illustriert das Konvergenzverhalten der Fixpunkt-Iteration für verschiedene ϕ.

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 ϕ, 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 Konvergenzverhalten 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, wobei $C \approx |\phi'(\xi)|$. Ganz besonders rasche, nämlich superlineare Konvergenz tritt auf, wenn $|\phi'(\xi)| = 0$.

Mit Hilfe der Taylorentwicklung lässt sich zeigen: 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, angewandt auf die Funktion f, 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.

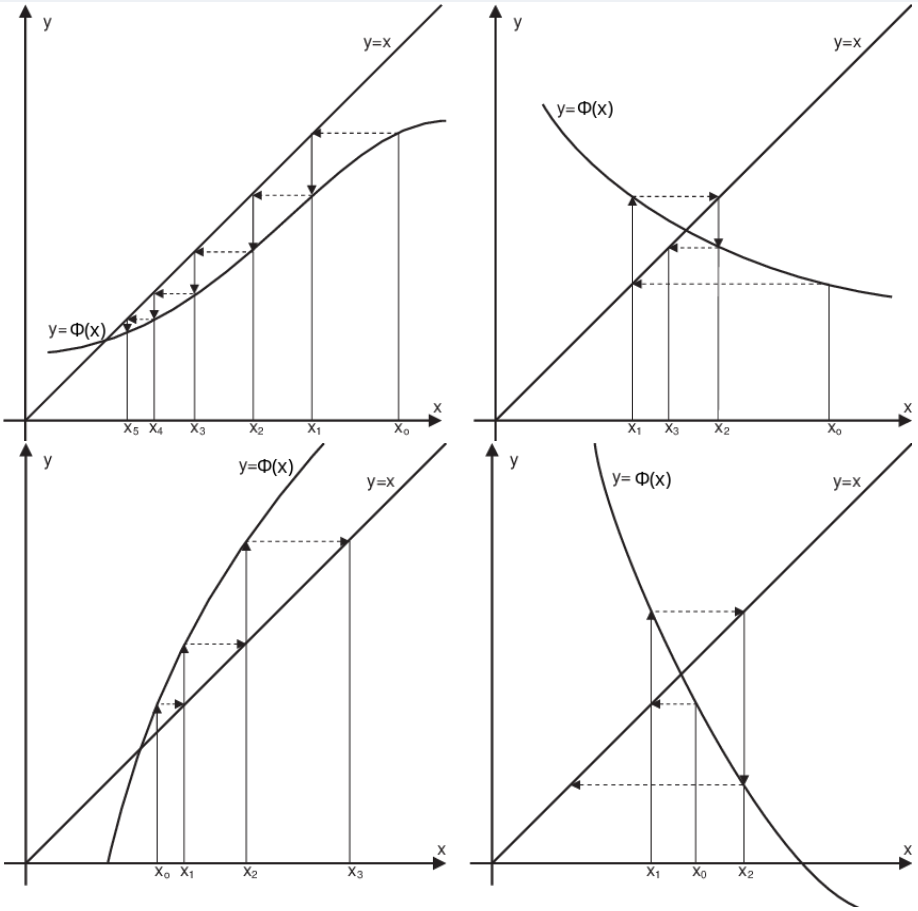

Abbildung 7: Fixpunkt-Iteration in graphischer Darstellung für verschiedene Funktionen ϕ. Mögliche Fälle: Einseitige Annäherung an den Fixpunkt, falls in einer Umgebung des Fixpunktes $0 < \phi' < 1$; alternierende Konvergenz, falls $-1 < \phi' < 0$, Divergenz falls $\phi' > 1$ oder $\phi' < -1$.

2 Systeme nichtlinearer Gleichungen

Abschnitt 1.2 definiert die Begriffe *Lösung, Nullstelle, Fixpunkt* für skalare Funktionen $\mathbb{R} \to \mathbb{R}$. Diese Begriffe lassen sich problemlos auf vektorwertige Funktionen $\mathbb{R}^n \to \mathbb{R}^n$ übertragen. Auch hier lassen sich Gleichungen auf verschiedene Weise formulieren.

Schreibweise für Vektoren und vektorwertige Funktionen: Fettdruck

Reellwertige Funktionen, Skalare: $f : \mathbb{R} \to \mathbb{R}$, $y = f(x)$
Vektorwertige Funktionen, Vektoren: $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$, $\mathbf{y} = \mathbf{f}(\mathbf{x})$

Komponenten eines Vektors $\mathbf{x} \in \mathbb{R}^n$:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad \text{oder } \mathbf{x}^T = [x_1, x_2, \ldots, x_n]$$

Normalerweise ist mit $\mathbf{x}$ ein Spalten-, mit $\mathbf{x}^T$ ein Zeilenvektor gemeint.

Iterationsindizes werden hier (um sie von Vektorkomponenten zu unterscheiden) hochgestellt und in Klammern gesetzt: $\mathbf{x}^{(k)}, k = 0, 1, 2 \ldots$

2.1 Lösung, Nullstelle und Fixpunkt: mehrdimensionaler Fall

Aufgabentypen im $\mathbb{R}^n$

Es seien $\mathbf{f}, \mathbf{g}, \mathbf{h}, \mathbf{\Phi}$ Funktionen $\mathbb{R}^n \to \mathbb{R}^n$ und $\mathbf{x} \in \mathbb{R}^n$

Problemstellung: gesucht ist ein $\mathbf{x}$, für das gilt...

$\mathbf{g}(\mathbf{x}) = \mathbf{h}(\mathbf{x})$, (Finden einer *Lösung* des Gleichungssystems)
$\mathbf{f}(\mathbf{x}) = 0$, (Finden einer *Nullstelle* der Funktion $\mathbf{f}$)
$\mathbf{x} = \mathbf{\Phi}(\mathbf{x})$, (Finden eines *Fixpunktes* der Funktion $\mathbf{\Phi}$)

Im Vergleich zu den Definitionen von Abschnitt 1.2 hat fast nichts geändert außer der Schreibweise.

Beispiel: ein *nichtlineares Gleichungssystem* mit zwei Unbekannten

$$\begin{aligned} 4x_1 - x_2 + x_1x_2 &= 1 \\ -x_1 + 6x_2 &= 2 - \log(x_1x_2) \end{aligned}$$

hat die Form $\mathbf{g}(\mathbf{x}) = \mathbf{h}(\mathbf{x})$ mit

$$\mathbf{g}(\mathbf{x}) = \begin{bmatrix} g_1(x_1, x_2) \\ g_2(x_1, x_2) \end{bmatrix} = \begin{bmatrix} 4x_1 - x_2 + x_1x_2 \\ -x_1 + 6x_2 \end{bmatrix}, \qquad \mathbf{h}(\mathbf{x}) = \begin{bmatrix} h_1(x_1, x_2) \\ h_2(x_1, x_2) \end{bmatrix} = \begin{bmatrix} 1 \\ 2 - \log(x_1x_2) \end{bmatrix}$$

Das Gleichungssystem lässt sich umformulieren:

$$\begin{aligned} 4x_1 - x_2 + x_1x_2 - 1 &= 0 \\ -x_1 + 6x_2 + \log(x_1x_2) - 2 &= 0 \end{aligned}$$

In dieser Form lautet die Aufgabe: gesucht sind *Nullstellen der vektorwertigen Funktion* $\mathbf{f} : \mathbb{R}^2 \to \mathbb{R}^2$, also Lösungen von $\mathbf{f}(\mathbf{x}) = 0$ mit

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} f_1(x_1, x_2) \\ f_2(x_1, x_2) \end{bmatrix} = \begin{bmatrix} 4x_1 - x_2 + x_1x_2 - 1 \\ -x_1 + 6x_2 + \log(x_1x_2) - 2 \end{bmatrix}$$

Eine andere, äquivalente Umformung liefert

$$\begin{aligned} x_1 &= \frac{1}{4}(x_2 - x_1x_2 + 1) \\ x_2 &= \frac{1}{6}(x_1 - \log(x_1x_2) + 2) \end{aligned}$$

Hier sind *Fixpunkte der vektorwertigen Funktion* $\mathbf{\Phi} : \mathbb{R}^2 \to \mathbb{R}^2$ gesucht, also Lösungen von $\mathbf{x} = \mathbf{\Phi}(\mathbf{x})$ mit

$$\mathbf{\Phi}(\mathbf{x}) = \begin{bmatrix} \phi_1(x_1, x_2) \\ \phi_2(x_1, x_2) \end{bmatrix} = \begin{bmatrix} \frac{1}{4}(x_2 - x_1x_2 + 1) \\ \frac{1}{6}(x_1 - \log(x_1x_2) + 2) \end{bmatrix}$$

Noch ein Hinweis zur Schreibweise: Wenn wir einen speziellen Fixpunkt gefunden haben, dann bezeichnen wir den im Folgenden mit $\boldsymbol{\xi}$, um ihn von anderen, allgemeinen Werten $\mathbf{x}$ zu unterscheiden.

2.2 Mehrdimensionale Fixpunkt-Iteration

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. Noch einmal: 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 ist ein Fixpunkt $\boldsymbol{\xi}$ von $\mathbf{\Phi}$.

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

Beachte die Konvergenzbedingungen (Abschnitt 2.4)!

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 $(1;1)$, $(1/4;1/2)$, $(0{,}343\,75; 0{,}721\,574)$, $(0{,}368\,383; 0{,}622\,985)$, …, die gegen den Fixpunkt $(0{,}353\,443\,88; 0{,}639\,968\,47)$ konvergiert.

2.3 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 $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$

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

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

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

Erinnern Sie sich an die Definition einer Norm aus Mathematik 2?

Eine Norm im $\mathbb{R}^n$ ist eine Funktion, die jedem Vektor $\mathbf{x} \in \mathbb{R}^n$ eine nichtnegative reelle Zahl $\|\mathbf{x}\| \in \mathbb{R}_0^+$ zuordnet, wobei $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $\forall \alpha \in \mathbb{R}$ drei Bedingungen gelten müssen:

- Nur der Nullvektor hat Norm 0

$$\|\mathbf{x}\| = 0 \quad \Rightarrow \quad \mathbf{x} = 0$$

- Skalar α läst sich als Betrag herausheben

$$\|\alpha \cdot \mathbf{x}\| = |\alpha| \cdot \|\mathbf{x}\|$$

- Die Dreiecksungleichung gilt

$$\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$$

Norm und Distanz

Eine Norm kann auch die *Distanz* zwischen zwei Punkten $\mathbf{x}$ und $\mathbf{y}$ messen:

$$\text{dist}(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$$

- Taxis in Manhattan messen Strecken in der 1-Norm.

  Deswegen heißt die 1-Norm auch Taxi- oder Manhattan-Norm

- Abstand in der Luftlinie entspricht der 2-Norm.
- Größter Unterschied in den Komponenten: ∞-Norm.

Matrixnormen

- Der Hauptberuf von Matrizen ist, Vektoren zu multiplizieren.
- Das Ergebnis einer Matrix-Vektor-Multiplikation ist wieder ein Vektor; der ist gewöhnlich länger oder kürzer und verdreht gegenüber Ausgangsvektor.
- Eine *Matrixnorm* misst als „Größe“ einer Matrix, wie „stark“ sie auf Vektoren wirkt.
- Eine gegebene Matrix kann Vektoren nicht beliebig stark verlängern. Es gibt für jede Matrix einen „Maximal-Verlängerungs-Faktor“

Der „Maximal-Verlängerungs-Faktor“ ist eine Matrixnorm

Verschiedene Matrixnormen

Die 1-, 2- und ∞-Normen lassen sich von den entsprechenden Vektornormen ableiten: Sie geben für die Rechenoperation $\mathbf{y} = A \cdot \mathbf{x}$ an, um wieviel $\mathbf{y}$ gegenüber $\mathbf{x}$ *maximal vergrößert* wird. Eins- und ∞-Norm lassen sich aus den Matrixelementen einfach berechnen:

$\|A\|_1$	*Einsnorm* : maximale Spaltenbetragssumme
$\|A\|_\infty$	*Unendlich-Norm* : maximale Zeilenbetragssumme

Für die Matrix-Zweinorm lässt sich keine so einfache Rechenvorschrift angeben, obwohl gerade sie häufig verwendet wird.

MATLAB kann alle Normen leicht berechnen: $\|A\|_1$ = `norm(A,1)`, $\|A\|_2$ = `norm(A)`, $\|A\|_\infty$ = `norm(A,Inf)`.

Matrixnorm, allgemeine Definition

Matrizen lassen sich addieren und mit Skalaren multiplizieren. In diesem Sinn verhalten sie sich genauso wie Vektoren des $\mathbb{R}^n$. Alles, was sich wie ein Vektor verhält, können wir als „Vektor“ interpretieren: Die $m \times n$-Matrizen bilden einen *Vektorraum*. Der Begriff „Norm“ wird genau so definiert wie die Norm von Vektoren des $\mathbb{R}^n$. Vergleichen Sie die Definition einer Norm im $\mathbb{R}^n$ auf Seite 22 – sie wird hier nahezu wörtlich übernommen.

Eine *Norm* im $\mathbb{R}^m \times \mathbb{R}^n$ ist eine Funktion, die jeder $m \times n$-Matrix A eine nichtnegative reelle Zahl $\|A\| \in \mathbb{R}_0^+$ zuordnet, wobei $\forall A, B \in \mathbb{R}^m \times \mathbb{R}^n$, $\forall \alpha \in \mathbb{R}$ drei Bedingungen gelten müssen:

- Nur die Nullmatrix hat Norm 0:

$$\|A\| = 0 \quad \Rightarrow \quad A = 0$$

- Skalar α läst sich als Betrag herausheben:

$$\|\alpha \cdot A\| = |\alpha| \cdot \|A\|$$

- Die Dreiecksungleichung gilt:

$$\|A + B\| \leq \|A\| + \|B\|$$

Diese drei Grundregeln muss jede Norm erfüllen. Aber es gibt für die 1-, 2- oder ∞-Norm noch Bonus-Features. Für diese Matrix-Normen gelten nämlich noch folgende Rechenregeln:

$$\|A \cdot B\| \leq \|A\| \cdot \|B\| \tag{8}$$

$$\|A \cdot \mathbf{x}\| \leq \|A\| \cdot \|\mathbf{x}\| \tag{9}$$

Vergleiche Absolutbetrag: $|a \cdot b| = |a| \cdot |b|$

Frobeniusnorm:

Noch eine weitere Norm; Die Frobenius-Norm $\|A\|_F$ wird so ähnlich berechnet wie die Vektor-Zweinorm: *Quadrieren, summieren, Wurzel ziehen*

Frobenius-Norm: $\qquad \|A\|_F = \sqrt{\sum a_{ij}^2}$

Die Frobeniusnorm lässt sich leichter berechnen als die Matrix-Zweinorm und dient zu deren Abschätzung:

$$\|A\|_2 \leq \|A\|_F$$

Auch für $\|A\|_F$ gelten neben den Norm-Axiome noch die Rechenregeln

$$\|A \cdot B\|_F \leq \|A\|_F \cdot \|B\|_F \quad , \quad \|A \cdot \mathbf{x}\|_2 \leq \|A\|_F \|\mathbf{x}\|_2$$

MATLAB: $\|A\|_F =$ `norm(A,'fro')`.

Matrixnormen – das Kleingedruckte[6]

[6] Was hier dasteht, ist nicht wichtig,
wenn 's nicht dastünd', wär's nicht richtig.

Die lockere Erklärung *„Matrixnorm ist maximaler Verlängerungsfaktor“* ist mathematisch korrekt für 1-, 2- und ∞-Norm, wenn Vektorlängen in den jeweiligen Normen gemessen werden. Die Frobeniusnorm überschätzt aber gewöhnlich den maximal auftretenden Verlängerungsfaktor, wenn Vektorlängen in der 2-Norm gemessen werden. Immerhin liefert sie eine obere Schranke für den Verlängerungsfaktor.
Auch die Vorschrift $\|A\| = \max_{i,j} |a_{ij}|$ erfüllt die drei Bedingungen einer Norm, ist aber nicht immer eine obere Schranke für den Verlängerungsfaktor.

2.4 Konvergenz

Die Konvergenz der mehrdimensionalen Fixpunkt-Iteration hängt wie im eindimensionalen Fall mit dem Begriff der kontrahierenden Abbildung zusammen.

Konvergenz der Fixpunkt-Iteration im $\mathbb{R}^n$
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 in der Form $B = \{\mathbf{x} : \|\boldsymbol{\xi} - \mathbf{x}\| < r\}, \quad r > 0$. Wenn $\mathbf{\Phi}$ in B eine *kontrahierende Abbildung* in (irgend-) einer Norm $\|\cdot\|$ ist, d. h.,

$$\|\mathbf{\Phi}(\mathbf{x}) - \mathbf{\Phi}(\mathbf{y})\| \leq C\|\mathbf{x} - \mathbf{y}\| \,, \quad C < 1 \text{ für alle } \mathbf{x}, \mathbf{y} \in B \,,$$

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. Auch der Begriff der Konvergenzordnung läßt sich unter Verwendung von Normen geradewegs auf den mehrdimensionalen Fall übertragen.

Kontraktion und Jacobi-Matrix

Das Konvergenzkriterium $|\phi'(\xi)| < 1$ im eindimensionalen Fall (vergleiche Seite 17) lässt sich auf den mehrdimensionalen Fall übertragen. Dazu fasst man die partiellen Ableitungen von $\mathbf{\Phi}$ in einer Matrix D_ϕ, genannt die *Jacobi-Matrix*, zusammen.

Jacobi-Matrix D_ϕ einer Funktion $\mathbf{\Phi} : \mathbb{R}^n \to \mathbb{R}^n$

$$D_\phi = \begin{bmatrix} \frac{\partial \phi_1}{\partial x_1} & \frac{\partial \phi_1}{\partial x_2} & \cdots & \frac{\partial \phi_1}{\partial x_n} \\ \frac{\partial \phi_2}{\partial x_1} & \frac{\partial \phi_2}{\partial x_2} & \cdots & \frac{\partial \phi_2}{\partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial \phi_n}{\partial x_1} & \frac{\partial \phi_n}{\partial x_2} & \cdots & \frac{\partial \phi_n}{\partial x_n} \end{bmatrix}$$

Dann lässt sich ganz ähnlich wie im eindimensionalen Fall aussagen:

Das Fixpunktverfahren konvergiert lokal,
falls in der 1-,2-, Frobenius- oder ∞-Matrixnorm gilt

$$\|D_\phi\| < 1$$

2.5 Newton-Verfahren für Systeme

Gegeben sei eine vektorwertige Funktion $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$. Gesucht sei eine Nullstelle von $\mathbf{f}$. Das ist 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. Und noch einmal sei darauf hingewiesen: wir setzen 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 mit

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

Das Newton-Verfahren für Systeme führt die Lösung eines nichtlinearen Systems auf die Lösung einer Folge von linearen Gleichungssystemen zurück. Die Lösung von Systemen linearer Gleichungen ist vergleichsweise einfach gegenüber nichtlinearen Gleichungssystemen. Wir behandeln lineare Gleichungssysteme später noch ausführlich, aber einstweilen nehmen wir an, dass Sie aus der Mittelschule damit hinreichend vertraut sind.

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

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

Angenommen, ein Punkt $\mathbf{x}^{(0)}$ ist als Startwert in der Nähe einer Nullstelle gegeben. Dann lässt sich $\mathbf{f}$ in der Umgebung von $\mathbf{x}^{(0)}$ in linearisierter Näherung schreiben als (Taylorscher Lehrsatz für Funktionen mehrerer Veränderlicher)

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

mit einem Restglied $\mathbf{R}$, das im Limes $\Delta\mathbf{x} \to 0$ mit höherer Ordnung verschwindet. Wir vernachlässigen das Restglied und fordern $\mathbf{f}(\mathbf{x}^{(0)} + \Delta\mathbf{x}) = 0$. Aus der daraus entstandenen Gleichung

$$0 = \mathbf{f}(\mathbf{x}^{(0)}) + D_f(\mathbf{x}^{(0)}) \cdot \Delta\mathbf{x}$$

lässt sich der Korrekturvektor $\Delta\mathbf{x}$ bestimmen und damit eine verbesserte Näherung $\mathbf{x}^{(1)} = \mathbf{x}^{(0)} + \Delta\mathbf{x}$.

Das Newton-Verfahrens für Systeme lässt sich so formulieren:

Newton-Verfahren für Systeme
Gegeben eine differenzierbare vektorwertige Funktion $\mathbf{f}$ 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.

Man kann zeigen: Sofern D_f^{-1} an der Nullstelle existiert, konvergiert das Verfahren für genügend genaue Startwerte 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. Dafür muss $x^{(0)}$ 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: nichtlineares Gleichungssystem aus Abschnitt 2.2

Die Funktion $\mathbf{f}$ und ihre 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 Korrekturvektor 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!