Kapitel 5
Optimierung, Fehlerrechnung und Statistik

In der Schule werden Ableitungen hauptsächlich benutzt, um die Extremwerte einer Funktion zu bestimmen; ein Gesichtspunkt, der im letzten Semester bei der Differentialrechnung mehrerer Veränderlicher keine Rolle spielte. In diesem letzten Kapitel der Vorlesung soll dies nachgeholt werden, wobei insbesondere die Anwendungen auf die Fehler- und Ausgleichsrechnung wichtige Beispiele liefern. Zu deren besseren Verständnis sollen auch einige Grundbegriffe der Statistik erörtert werden.

§1: Extrema von Funktionen mehrerer Veränderlicher

a) Der eindimensionale Fall

Erinnern wir uns an die Schule: Wenn die stetig differenzierbare Funktion $f\colon (a,\, b) \to \mathbb{R}$ im Punkt $x_0 \in (a,\, b)$ ein Extremum annimmt, verschwindet dort die Ableitung $f'(x_0)$. Der Grund ist klar: Nach Definition der Differenzierbarkeit ist

$$f(x_0 + h) = f(x_0) + h f'(x_0) + o(h)\,;$$

falls $f'(x_0)$ nicht verschwindet, ist $f(x_0 + h)$ für kleine h mit demselben Vorzeichen wie $f'(x_0)$ größer und für solche mit entgegengesetztem Vorzeichen kleiner als $f(x_0)$. In x_0 kann f somit weder ein Maximum noch ein Minimum annehmen.

Die Umkehrung gilt nicht: Standardbeispiel ist die Funktion $f(x) = x^3$, für die $f'(0)$ verschwindet, ohne daß im Nullpunkt ein Maximum oder Minimum wäre.

Für zweimal stetig differenzierbare Funktionen gibt es bekanntlich auch eine hinreichende Bedingung sowie die Möglichkeit, Maxima und Minima voneinander zu unterscheiden: Falls $f'(x_0)$ verschwindet und $f''(x_0)$ negativ ist, hat f im Punkt x_0 ein Maximum; bei positivem $f''(x_0)$ liegt ein Minimum vor. Auch hier folgt alles sofort aus der Definition der zweimaligen Differenzierbarkeit: Wegen

$$\begin{aligned} f(x_0 + h) &= f(x_0) + h f'(x_0) + \frac{h^2}{2} f''(x_0) + o(h^2) \\ &= f(x_0) + \frac{h^2}{2} f''(x_0) + o(h^2) \end{aligned}$$

sieht der Graph von f in diesen Fällen in der unmittelbaren Umgebung von x_0 aus wie eine nach unten *bzw.* oben geöffnete Parabel.

b) Verallgemeinerung aufs Mehrdimensionale

Nun betrachten wir eine stetig differenzierbare Funktion $f\colon D \to \mathbb{R}$ auf einer offenen Teilmenge $D \subset \mathbb{R}^n$. Dann bedeutet Differenzierbarkeit bekanntlich, daß es in jedem Punkt $\boldsymbol{x}_0 \in D$ einen Vektor

$$\nabla f(\boldsymbol{x}_0) = \operatorname{grad} f(\boldsymbol{x}_0) \in \mathbb{R}^n$$

gibt, den Gradienten, so daß für hinreichend kleine Vektoren $\vec{h} \in \mathbb{R}^n$ gilt

$$f(\boldsymbol{x}_0 + \vec{h}) = f(\boldsymbol{x}) + \operatorname{grad} f(\boldsymbol{x}_0) \cdot \vec{h} + o\big(|\vec{h}|\big)\,.$$

Hier muß also für jeden Extremwert $\operatorname{grad} f(\boldsymbol{x}_0)$ gleich dem Nullvektor sein, denn setzt man für $\vec{h}$ ein kleines Vielfaches $t \cdot \operatorname{grad} f(\boldsymbol{x}_0)$ des Gradienten ein, wäre sonst

$$f(\boldsymbol{x}_0 + \vec{h}) = f(\boldsymbol{x}) + t\big(\operatorname{grad} f(\boldsymbol{x}_0) \cdot \operatorname{grad} f(\boldsymbol{x}_0)\big) + o\big(|\vec{h}|\big)$$

für kleine positive t größer als $f(x_0)$ und für kleine negative t kleiner.

Die Frage, welche Nullstellen des Gradienten wirklich Extremwerten entsprechen, ist schwieriger; in der Praxis wird es oft am einfachsten sein, sich die Umgebung des betreffenden Punktes mit irgendwelchen *ad hoc*-Methoden genauer anzusehen und dann zu entscheiden.

Klassisches Beispiel eines Punkts, in dem der Gradient verschwindet, ohne daß ein Extremwert vorliegt, ist der in Abbildung 54 gezeigte

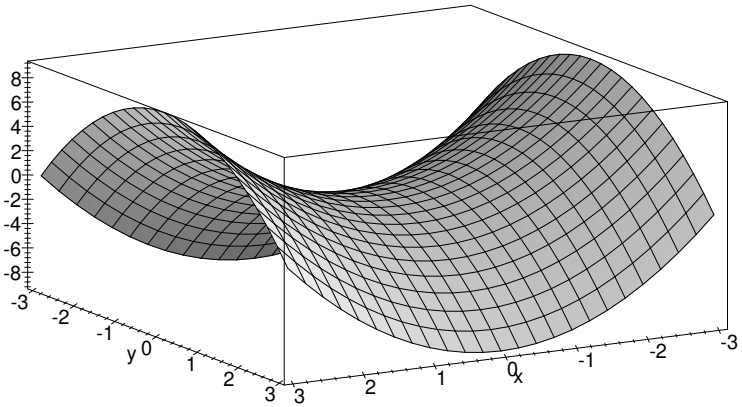


Abb. 54: Graph der Funktion $f(x,y) = x^2 - y^2$

Sattelpunkt, hier dargestellt als Funktionswert über dem Punkt $(0,0)$ für die Funktion $f(x,y) = x^2 - y^2$.

Für zweifach stetig differenzierbare Funktionen kann man genau wie im eindimensionalen Fall ein hinreichendes Kriterium finden, das nur von der zweiten Ableitung im Punkt $\boldsymbol{x}_0$ abhängt:

Die zweite Ableitung von $f \in \mathcal{C}^2(D,\mathbb{R})$ im Punkt $\boldsymbol{x}_0 \in D$ ist bekanntlich gegeben durch die HESSE-Matrix

$$H_f(\boldsymbol{x}_0) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_1 \partial x_2} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n} & \frac{\partial^2 f}{\partial x_2 \partial x_n} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}$$

und zweimalige Differenzierbarkeit bedeutet, daß

$$f(\boldsymbol{x}_0 + \vec{h}) = f(\boldsymbol{x}_0) + \operatorname{grad} f(\boldsymbol{x}_0) \cdot \vec{h} + \tfrac{1}{2}\,{}^t\vec{h}\, H_f(\boldsymbol{x}_0)\, \vec{h} + o\big(\big|\vec{h}^2\big|\big)$$

ist für kleine $\vec{h}$.

Wenn $\operatorname{grad} f(\boldsymbol{x}_0)$ verschwindet, hängt also das Verhalten von f in der Umgebung von $\boldsymbol{x}_0$ ab von der quadratischen Form

$$\vec{h} \mapsto {}^t\vec{h}\, H_f(\boldsymbol{x}_0)\, \vec{h}\,.$$

Definition: *a)* Eine Matrix $A \in \mathbb{R}^{n\times n}$ heißt *positiv definit,* wenn für alle Vektoren $\vec{v} \neq \vec{0}$ aus $\mathbb{R}^n$ gilt:

$${}^t\vec{v} A(\boldsymbol{x}_0)\vec{v} > 0\,.$$

b) A heißt *negativ definit,* wenn für alle $\vec{v} \neq \vec{0}$ aus $\mathbb{R}^n$ gilt:

$${}^t\vec{v} A(\boldsymbol{x}_0)\vec{v} < 0\,.$$

c) A heißt *indefinit,* wenn es Vektoren $\vec{v}, \vec{w} \in \mathbb{R}^n$ gibt mit

$${}^t\vec{v} A(\boldsymbol{x}_0)\vec{v} > 0 \quad \text{und} \quad {}^t\vec{w} A(\boldsymbol{x}_0)\vec{w} < 0\,.$$

Mit dieser Terminologie ist das folgende Lemma klar:

Lemma: Wenn die differenzierbare Funktion $f \in \mathcal{C}^1(D,\mathbb{R})$ im Punkt $\boldsymbol{x}_0 \in D$ ein lokales Extremum hat, ist dort ihr Gradient gleich dem Nullvektor.
Falls umgekehrt für $f \in \mathcal{C}^2(D,\mathbb{R})$ der Gradient im Punkt $\boldsymbol{x} \in D$ verschwindet, gilt:
a) Falls die HESSE-Matrix $H_f(\boldsymbol{x}_0)$ positiv definit ist, hat f im Punkt $\boldsymbol{x}_0$ ein Minimum.
b) Falls $H_f(\boldsymbol{x}_0)$ negativ definit ist, hat f im Punkt $\boldsymbol{x}_0$ ein Maximum.
c) Falls $H_f(\boldsymbol{x}_0)$ indefinit ist, hat f im Punkt $\boldsymbol{x}_0$ kein Extremum. ■

Damit uns das etwas nützt, brauchen wir jetzt nur noch ein Kriterium, mit dem wir feststellen können, welche Definitheitseigenschaften die HESSE-Matrix hat. Dazu erinnern wir uns daran, daß die HESSE-Matrix symmetrisch ist, und daß nach Kapitel 4, §2*d)* jede symmetrische Matrix diagonalisierbar ist.

Für eine Diagonalmatrix A mit Einträgen $\lambda_1, \ldots, \lambda_n$ und einen Vektor $\vec{v}$ mit Komponenten $v_1, \ldots, v_n$ wird obige quadratische Form zu

$$(v_1, v_2, \ldots, v_n) \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_n \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} = \lambda_1 v_1^2 + \cdots + \lambda_n v_n^2\,;$$

eine Diagonalmatrix ist also genau dann positiv definit, wenn alle Diagonaleinträge positiv sind und genau dann negativ definit, wenn sie alle

negativ sind. Falls es sowohl positive als auch negative Diagonaleinträge gibt, ist die Matrix indefinit.

Nun ist es für den Wertebereich einer Funktion irrelevant, bezüglich welches Koordinatensystems wir die Argumente ausdrücken; wir können eine symmetrische Matrix also bezüglich einer Basis aus Eigenvektoren betrachten, wo sie zur Diagonalmatrix wird mit den Eigenwerten als Einträgen. Daher gilt:

Lemma: Eine symmetrische Matrix ist genau dann positiv definit, wenn alle ihre Eigenwerte positiv sind und genau dann negativ definit, wenn alle ihre Eigenwerte negativ sind. Falls es sowohl positive als auch negative Eigenwerte gibt, ist sie indefinit. ■

Da die Determinante einer Matrix gleich dem Produkt ihrer Eigenwerte ist, folgt, daß eine Matrix nur dann positiv definit sein kann, wenn ihre Determinante positiv ist; für negativ definite $n \times n$-Matrizen muß die Determinante bei geradem n ebenfalls positiv sein, bei ungeradem negativ.

Für symmetrische 2×2-Matrizen läßt sich daraus leicht ein notwendiges und hinreichendes Kriterium machen: Das charakteristische Polynom von

$$A = \begin{pmatrix} a & b \\ b & d \end{pmatrix}$$

mit Eigenwerten λ_1 und λ_2 ist

$$\lambda^2 - (a+d)\lambda + (ad - b^2) = (\lambda - \lambda_1)(\lambda - \lambda_2)\,;$$

daher ist

$$\lambda_1 + \lambda_2 = a + d\,.$$

(In der Tat rechnet man auf genau die gleiche Weise leicht nach, daß für jede $n \times n$-Matrix die Summe der n Eigenwerte gleich der Summe der n Diagonaleinträge ist, die sogenannte *Spur* der Matrix.)

Wenn $\det A = ad - b^2$ positiv ist, haben nicht nur λ_1 und λ_2, sondern auch a und d dasselbe Vorzeichen, das somit gleich dem von $a + d = \lambda_1 + \lambda_2$ ist. Also ist A genau dann positiv definit, wenn $\det A > 0$ und $a > 0$ ist, negativ definit, wenn $\det A > 0$ und $a < 0$ ist, und indefinit wenn $\det A < 0$ ist. (Anstelle von a könnte hier natürlich überall auch d stehen.)

Beispielsweise ist die Matrix $\begin{pmatrix} 1 & 2 \\ 2 & 5 \end{pmatrix}$ positiv definit, denn sie hat Determinante eins und positive Diagonaleinträge. Im obigen Beispiel des Sattelpunkts mit $f(x,y) = x^2 - y^2$ ist

$$H_f(0,0) = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix}$$

offensichtlich indefinit, was man nicht nur an der negativen Determinanten sieht.

§2: Maxima und Minima unter Nebenbedingungen

Bei einem realen physikalischen oder technischen Prozeß können sich die Variablen selten frei im gesamten $\mathbb{R}^n$ bewegen: Physikalisch sinnvoll ist meist nur eine beschränkte Teilmenge. Im Gegensatz zur Dimension eins, wo diese Teilmenge praktisch immer ein Intervall ist, gibt es aber im Mehrdimensionalen keinen Grund, warum diese Teilmenge offen oder zumindest der Abschluß einer offenen Teilmenge sein sollte: Im $\mathbb{R}^3$ kann man sich beispielsweise auch interessieren für das Maximum oder Minimum der Ladungsdichte auf einer Kugeloberfläche oder die elektrische Feldstärke oder Temperaturverteilung auf der Innenhaut eines Reaktordruckbehälters.

Diese Maxima oder Minima sind im allgemeinen keine lokalen Maxima oder Minima der betrachteten Funktion: Wenn man die jeweilige Fläche verläßt, läßt sich der Funktionswert selbst für einen solchen Extremwert meist noch – je nach Richtung – sowohl vergrößern als auch verkleinern. Dementsprechend können die Methoden, die wir in §1 diskutiert haben, solche Extremwerte üblicherweise nicht finden; wir brauchen weitere Werkzeuge, die in diesem Paragraphen bereitgestellt werden sollen.

Die Situation, um die es hier geht, ist typischerweise die folgende: Gegeben ist eine Funktion $f\colon \mathbb{R}^n \to \mathbb{R}$, möglicherweise auch nur auf einer Teilmenge $D \subset \mathbb{R}^n$ definiert, deren Extremwerte nicht auf $\mathbb{R}^n$ oder D

gesucht werden, sondern nur auf einer Teilmenge, die beispielsweise durch das Verschwinden einer weiteren Funktion $g\colon \mathbb{R}^n \to \mathbb{R}$ gegeben ist. Falls wir uns für Extremwerte auf einer Kugel vom Radius r um den Nullpunkt interessieren, wäre dies etwa die Funktion

$$g\colon \begin{cases} \mathbb{R}^3 & \to \mathbb{R} \\ (x,y,z) & \mapsto x^2+y^2+z^2-r^2 \end{cases}.$$

Eine mögliche Strategie zur Lösung solcher Probleme besteht darin, die Gleichung $g = 0$ nach einer der Variablen aufzulösen, diese dann in f einzusetzen und sodann eine gewöhnliche Extremwertaufgabe zu lösen. Diese Auflösung ist *explizit* nur in sehr einfachen Fällen möglich, aber selbst wenn wir nur wissen, daß eine solche Auflösung *existiert,* können wir doch damit argumentieren und Kriterien ableiten.

Unter Maxima und Minima sollen hier *lokale* Extrema verstanden werden, so daß wir die üblichen Kriterien anwenden können:

Definition: Wir sagen, die Funktion $f\colon D \to \mathbb{R}$ auf einer Teilmenge $D \subseteq \mathbb{R}^n$ habe im Punkt $\boldsymbol{a} \in D$ ein lokales $\left\{\begin{matrix}\text{Maximum}\\ \text{Minimum}\end{matrix}\right\}$ unter der Nebenbedinung $g = 0$, wobei $g\colon D \to \mathbb{R}$ eine weitere Funktion ist, wenn $g(\boldsymbol{a}) = 0$ ist und es eine Umgebung U von $\boldsymbol{a}$ gibt, so daß für alle $\boldsymbol{x} \in U$ gilt: Ist $g(\boldsymbol{x}) = 0$, so ist $f(\boldsymbol{x}) \left\{\begin{matrix}\le\\ \ge\end{matrix}\right\} f(\boldsymbol{a})$.

Als Einstiegsbeispiel betrachten wir eine beliebte Schulbuchaufgabe zur Minimumsbestimmung: Eine Konservendose soll bei einem vorgegebenen Volumen von 100 cm^3 möglichst wenig Blech benötigen, d.h. ihre Oberfläche soll minimal sein.

Die Oberfläche eines Zylinders der Höhe h mit einer Grundfläche vom Radius r ist

$$f(r,h) = 2\pi r^2 + 2\pi r \cdot h\,;$$

die Nebenbedingung für das Volumen $V = \pi r^2 h$ besagt, daß

$$g(r,h) = \pi r^2 h - 100 = 0$$

sein soll.

Hier läßt sich natürlich die Nebenbedingung sofort nach h auflösen:

$$h = \frac{100}{\pi r^2}\,,$$

und wir müssen nur noch die Funktion

$$F(r) = f\left(r, \frac{100}{\pi r^2}\right) = 2\pi r^2 + \frac{200}{r}$$

minimieren. Für diese ist

$$F'(r) = 4\pi r - \frac{200}{r^2}\,,$$

und dies verschwindet genau dann, wenn

$$4\pi r^3 = 200 \quad \text{oder} \quad r = \sqrt[3]{\frac{50}{\pi}}$$

ist.

In diesem einfachen Fall kann man solche Aufgaben also zurückführen auf gewöhnliche Extremwertaufgaben, indem man die Nebenbedingung nach einer der Variablen auflöst und diese dann in f einsetzt; in anderen Fällen kann man gelegentlich die Nebendedingung durch geeignete Parameterwahl oder Wahl eines angepaßten Koordinatensystems berücksichtigen. Im allgemeinen wird aber beides nicht möglich sein, so daß wir andere Methoden brauchen.

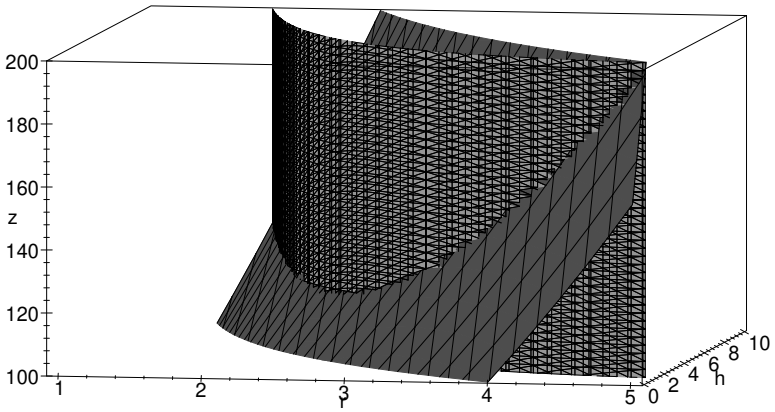


Abb. 55: Oberfläche einer Konservendose mit festem Volumen

Unser bisherige Theorie für lokale Extrema ist in dieser Situation nicht anwendbar, denn die lokalen Extrema von f werden nur in den seltensten Fällen die Nebenbedingung $g = 0$ erfüllen; im obigen Beispiel zeigt Abbildung 55 die Nebenbedingung als eng schraffierte Fläche dargestellt und der Graph von f als weiter schraffierte; wie man sieht, läßt sich der Wert von f problemlos verkleinern, wenn man nur die Fläche $g = 0$ verläßt, und in der Tat ist auch ohne jede Mathematik sofort klar, daß man mit weniger Blech auskommt, wenn man die Konservendose einfach schmäler oder kürzer macht.

Die Grundidee für ein alternatives Verfahren wird klar bei der Betrachtung der Niveaulinien in Abbildung 56: Die Niveaulinie für $g = 0$ ist gestrichelt eingezeichnet, verschiedene Niveaulinien von f als durchgezogene Kurven.

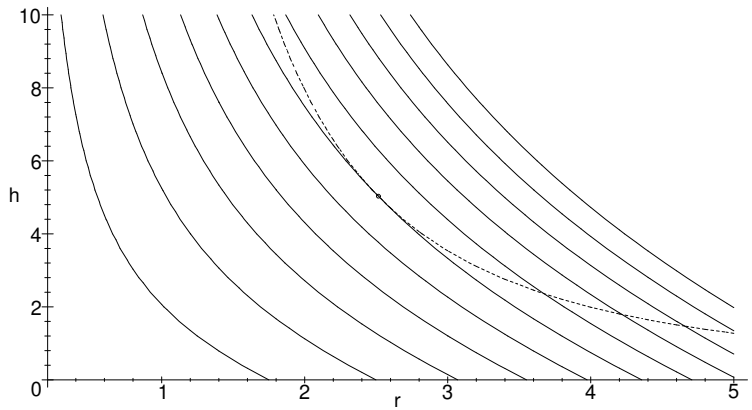


Abb. 56: Niveaulinien für Oberfläche und Volumen

Wie man sieht, schneiden einige dieser Niveaulinien die gestrichelte Kurve überhaupt nicht: Wenn man zu wenig Blech hat, kann man keine Dose mit 100 cm^3 Inhalt zusammenlöten. Wenn es dagegen genug Blech gibt, gibt es gleich zwei Schnittpunkte: Die Dose kann entweder eher höher oder eher breiter gemacht werden. In einem solchen Fall kann man die Niveaulinie durch eine zu einem etwas niedrigeren Niveau ersetzen,

die im allgemeinen auch wieder Schnittpunkte haben wird, so daß das Niveau noch nicht minimal sein kann. Erst wenn man im Minimum ist, fallen die beiden Schnittpunkte zusammen; wenn man nun das Niveau noch weiter erniedrigt, gibt es keine Schnittpunkte mehr.

Da somit im Minimum zwei Schnittpunkte zusammenfallen, berühren sich dort die Niveaulinien von f und von g, d.h. sie haben eine gemeinsame Tangente. Da der Gradient, wie wir wissen, senkrecht auf der Tangenten der Niveaulinien steht (die Richtungsableitung entlang einer Niveaulinie ist schließlich null), sind somit die Gradienten von f und g im Minimum zueinander parallel, d.h. der eine ist ein Vielfaches des anderen.

Dies gilt nicht nur im vorliegenden Beispiel, sondern allgemein:

Satz: $D \subseteq \mathbb{R}^n$ sei eine offene Menge und $f, g \in \mathcal{C}^1(D, \mathbb{R})$ seien stetig differenzierbare Funktionen auf D. Falls f im Punkt $\boldsymbol{a} \in D$ ein Extremum hat unter der Nebenbedinung $g(\boldsymbol{x}) = 0$, so sind grad $f(\boldsymbol{a})$ und grad $g(\boldsymbol{a})$ linear abhängig.

Beweis: Die Grundidee ist einfach: Auch wenn wir die Nebenbedingung nicht *explizit* nach einer der Variablen auflösen können, sagt uns der Satz über implizite Funktionen in vielen Fällen dennoch, daß zumindest lokal eine Auflösung existiert. Diese Auflösung kennen wir zwar nicht, aber wir können mit ihr argumentieren und, zumindest formal, auch rechnen.

Falls grad $g(\boldsymbol{a})$ der Nullvektor ist, gibt es nichts mehr zu beweisen, denn jede Menge, die den Nullvektor enthält, ist linear abhängig.

Wir können daher annehmen, daß grad $g(\boldsymbol{a})$ mindestens eine von Null verschiedene Komponente hat, und durch Umnummerieren der Koordinaten können wir o.B.d.A. annehmen, daß dies die n-te Komponente ist, d.h. $g_{x_n}(\boldsymbol{a}) \neq 0$.

Dann gibt es nach dem Satz über implizite Funktionen (**[HM I]**, *Kap. 2, §3d*) eine Umgebung U von $(a_1, \ldots, a_{n-1})$ und eine Funktion $h\colon U \to \mathbb{R}$ mit $h(a_1, \ldots, a_{n-1}) = a_n$, so daß

$$g\big(x_1, \ldots, x_{n-1}, h(x_1, \ldots, x_{n-1})\big) = 0 \quad \text{für alle } (x_1, \ldots, x_{n-1}) \in U\,.$$

Nachdem f in $\boldsymbol{a}$ ein lokales Extremum unter der Nebenbedinung $g = 0$ hat, nimmt die Funktion

$$F(x_1, \ldots, x_{n-1}) \underset{\text{def}}{=} f\big(x_1, \ldots, x_{n-1}, h(x_1, \ldots, x_{n-1})\big)$$

in $(a_1, \ldots, a_{n-1})$ ein lokales Extremum im üblichen Sinne an, d.h. der Gradient von F verschwindet dort.

Nach der Kettenregel ist für $i = 1, \ldots, n-1$

$$F_{x_i}(a_1, \ldots, a_{n-1}) = f_{x_i}(\boldsymbol{a}) + f_{x_n}(\boldsymbol{a}) \cdot h_{x_i}(a_1, \ldots, a_{n-1}),$$

und nach dem Satz über implizite Funktionen ist $h_{x_i} = -g_{x_i}/g_{x_n}$, d.h.

$$F_{x_i}(a_1, \ldots, a_{n-1}) = f_{x_i}(\boldsymbol{a}) - f_{x_n}(\boldsymbol{a})\frac{g_{x_i}(\boldsymbol{a})}{g_{x_n}(\boldsymbol{a})}.$$

Da die linke Seite verschwindet, gilt dasselbe auch für die rechte. Die rechte Seite ist im Gegensatz zur linken auch für $i = n$ definiert und verschwindet aus trivialen Gründen; also ist für alle i

$$f_{x_i}(\boldsymbol{a}) - \frac{f_{x_n}(\boldsymbol{a})}{g_{x_n}(\boldsymbol{a})} g_{x_i}(\boldsymbol{a}) = 0$$

oder, anders ausgedrückt,

$$\operatorname{grad} f(\boldsymbol{a}) - \frac{f_{x_n}(\boldsymbol{a})}{g_{x_n}(\boldsymbol{a})} \operatorname{grad} g(\boldsymbol{a}) = \vec{0}.$$

Damit sind die beiden Gradienten in der Tat linear abhängig. ■

Falls der Gradient von g im Punkt $\boldsymbol{a}$ nicht verschwindet, gibt es somit eine Zahl $\lambda \in \mathbb{R}$, so daß

$$\operatorname{grad} f(\boldsymbol{a}) - \lambda \operatorname{grad} g(\boldsymbol{a}) = \vec{0}$$

ist, nämlich

$$\lambda = \frac{f_{x_n}(\boldsymbol{a})}{g_{x_n}(\boldsymbol{a})}.$$

Diese Zahl bezeichnet man als LAGRANGEschen Multiplikator; mit seiner inhaltlichen Interpretation werden wir uns in Kürze beschäftigen.

JOSEPH-LOUIS LAGRANGE (1736–1813) wurde als GIUSEPPE LODOVICO LAGRANGIA in Turin geboren und studierte dort zunächst Latein. Erst eine alte Arbeit von HALLEY über algebraische Methoden in der Optik weckte sein Interesse an der Mathematik, woraus ein ausgedehnter Briefwechsel mit EULER entstand. In einem Brief vom 12. August 1755 berichtete er diesem unter anderem über seine Methode zur Berechnung von Maxima und Minima; 1756 wurde er, auf EULERs Vorschlag, Mitglied der Berliner Akademie; zehn Jahre später zog er nach Berlin und wurde dort EULERs Nachfolger als mathemascher Direktor der Akademie. 1787 wechselte er an die Pariser Académie des Sciences, wo er bis zu seinem Tod blieb und unter anderem an der Einführung des metrischen Systems beteiligt war. Seine Arbeiten umspannen weite Teile der Analysis, Algebra und Geometrie.

Zur praktischen Bestimmung von Extremwerten unter Nebenbedingungen geht man wie folgt vor: Über die Punkte, in denen der Gradient von g verschwindet, macht obiger Satz keine verwertbare Aussage; diese Punkte müssen also vorab berechnet und untersucht werden.

Danach müssen die Punkte gefunden werden, in denen es ein $\lambda \in \mathbb{R}$ gibt, so daß

$$\begin{aligned} f_{x_1}(\boldsymbol{x}) - \lambda g_{x_1}(\boldsymbol{x}) &= 0 \\ &\vdots \\ f_{x_n}(\boldsymbol{x}) - \lambda g_{x_n}(\boldsymbol{x}) &= 0 \\ g(\boldsymbol{x}) &= 0 \end{aligned}$$

ist. Dies ist ein System von $n+1$ Gleichungen für die $n+1$ Unbekannten, allerdings ist dieses Gleichungssystem nur selten linear und damit oft nicht mit bekannten Methoden lösbar. Manchmal kann man das Gleichungssystem durch geeignete Umformungen und Fallunterscheidungen vollständig lösen, in anderen Fällen helfen nur die aus der Numerik bekannten Näherungsverfahren wie etwa die Methode von NEWTON-RAPHSON.

Falls alle Gleichungen Polynomgleichungen sind (oder durch Einführung geeigneter zusätzlicher Variablen auf Polynomgleichungen zurückgeführt werden können), kann man im Falle einer endlichen Lösungsmenge diese auch exakt bestimmen: Genau wie der GAUSS-Algorithmus zur Lösung eines linearen Gleichungssystems dieses auf eine

Treppengestalt bringt, aus der man die Lösungen einfach ermitteln kann, gibt es in der Computeralgebra einen Algorithmus, der dasselbe für beliebige Systeme von Polynomgleichungen versucht; die Gleichungen, die dieser Algorithmus liefert, bezeichnet man als GRÖBNER-Basis oder Standardbasis. Zum Verständnis dieses Algorithmus, den man als eine Art Synthese aus EUKLIDischen Algorithmus und GAUSS-Algorithmus ansehen kann, sind Kenntnisse der kommutativen Algebra erforderlich, für die die Zeit in dieser Vorlesung nicht ausreicht; bei einigen Implementierungen werden zusätzlich auch noch Algorithmen aus der Informatik eingesetzt, die typischerweise nicht in Grundvorlesungen behandelt werden. Deshalb sei hier nur darauf hingewiesen, daß die gängigen universellen Computeralgebrasysteme wie Maple, Mathematica, MuPad allesamt entsprechende Routinen enthalten, mit denen man auch dann experimentieren kann, wenn man die dahinterstehende Theorie nicht versteht.

Als Beispiel, wie gelegentlich auch ein nichtlineares Gleichungssystem elementar gelöst werden kann, betrachten wir eine Anwendung aus den Wirtschaftswissenschaften: Die Gesamtproduktion eines Unternehmens oder eines Staats in Abhängigkeit von n eingesetzten Ressourcen $x_1, \ldots, x_n$ wird oft modelliert durch eine sogenannte COBB-DOUGLAS-Funktion der Form

$$P(x_1, \ldots, x_n) = \alpha x_1^{e_1} \ldots x_n^{e_n} ,$$

benannt nach den beiden Wissenschaftlern, die dieses Modell 1928 für die amerikanische Gesamtproduktion in Abhängigkeit von Kapital und Arbeit in den Jahren 1899 bis 1922 entwickelten. (Sie fanden $P \approx 1{,}01 A^{3/4} K^{1/4}$ mit A = Anzahl der Beschäftigten und K = Kapitaleinsatz.)

Betrachten wir stattdessen die Produktion eines Wirtschaftsguts aus zwei Resourcen x, y gemäß der Funktion

$$f(x, y) = P(x, y) = x^{1/2} y^{1/4} .$$

Falls wir der Einfachheit halber annehmen, daß die Kosten pro Einheit für x und y gleich sind und die Gesamtkosten höchstens gleich zwölf sein dürfen, müssen wir f maximieren unter der Nebenbedingung

$$x + y \leq 12 .$$

Nun ist aber f eine monoton wachsende Funktion sowohl von x als auch von y, d.h. die maximale Produktion wird sicherlich erreicht in einem Punkt, für den $x + y = 12$ ist, denn für jeden anderen Punkt

(x, y) mit $x + y < 12$ ist $f(x, y) < f(x, 12 - x)$. Daher können wir die Nebenbedingung in der gewohnten Form

$$g(x, y) = x + y - 12 = 0$$

schreiben. Diese Nebenbedingung sowie die zu maximierende Funktion sind in Abbildung 57 dargestellt.

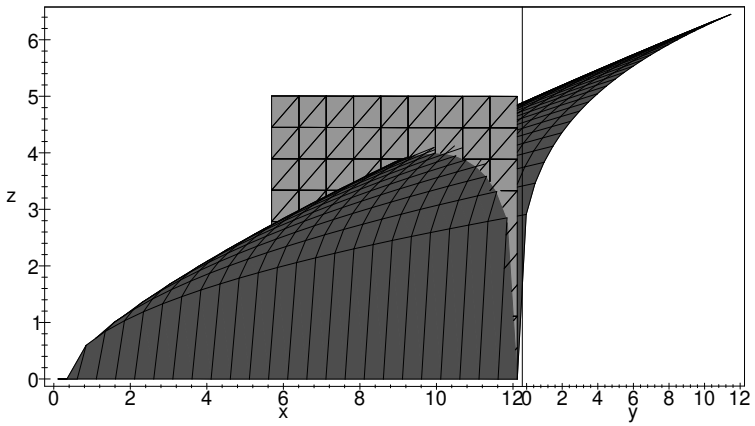


Abb. 57: Maximierung einer Produktionsfunktion bei festem Kapitaleinsatz

Ableitung beider Funktionen zeigt, daß

$$\operatorname{grad} g = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \text{und} \quad \operatorname{grad} f = \begin{pmatrix} y^{1/4}/2x^{1/2} \\ x^{1/2}/4y^{3/4} \end{pmatrix}$$

ist; das zu lösende Gleichungssystem wird also zu

$$\begin{aligned} \frac{y^{1/4}}{2x^{1/2}} - \lambda &= 0 \\ \frac{x^{1/2}}{4y^{3/4}} - \lambda &= 0 \\ x + y - 12 &= 0 \end{aligned} .$$

(Die Nenner brauchen uns nicht zu stören, denn da $f(0, y) = f(x, 0) = 0$ ist, kommen Lösungen mit $x = 0$ oder $y = 0$ für das Maximum ohnehin nicht in Frage; wir können sie also getrost ausschließen.)

Als Ansatz zu einer möglichen Lösung können wir ausnutzen, daß λ in den beiden ersten Gleichungen isoliert steht; wenn wir danach auflösen und gleichsetzen, erhalten wir die Gleichung

$$\frac{y^{1/4}}{2x^{1/2}} = \frac{x^{1/2}}{4y^{3/4}} .$$

Multiplikation mit dem Hauptnenner macht daraus

$$4y^{1/4}y^{3/4} = 2x^{1/2}x^{1/2} \quad \text{oder} \quad 2y = x .$$

Einsetzen in die dritte Gleichung ergibt $3y = 12$, also ist

$$y = 4 \quad \text{und} \quad x = 8 ;$$

der Maximalwert von f ist

$$f(8,4) = 8^{1/2} \cdot 4^{1/4} = 2\sqrt{2} \cdot \sqrt{2} = 4 .$$

Auch den LAGRANGEschen Multiplikator λ können wir noch ausrechnen:

$$\lambda = \frac{y^{1/4}}{2x^{1/2}} = \frac{4^{1/4}}{2 \cdot 8^{1/2}} = \frac{\sqrt{2}}{2 \cdot 2\sqrt{2}} = \frac{1}{4} .$$

Die Berechnung von λ war für die Bestimmung des Optimums eigentlich überflüssig; λ ist nur eine Hilfsgröße zur Berechnung des Extremums. Wir wollen uns als nächstes überlegen, daß wir λ auch inhaltlich interpretieren können: Dazu betrachten wir eine Nebenbedingung

$$g(x_1, \ldots, x_n) = c$$

mit *variabler* rechter Seite c und ein Extremum der Funktion

$$f(x_1, \ldots, x_n) .$$

Dieses Extremum wird natürlich von c abhängen; wir schreiben es in der Form

$$\big(x_1(c), \ldots, x_n(c)\big)$$

und nehmen an, daß die Funktionen $x_i(c)$ stetig differenzierbar seien. (Ein interessierter Leser kann sich anhand des Satzes über implizite Funktionen überlegen, welche Bedingungen f und g erfüllen müssen, damit dies garantiert ist.) Der Optimalwert von f in Abhängigkeit von c ist dann

$$F(c) \underset{\text{def}}{=} f\big(x_1(c), \ldots, x_n(c)\big) .$$

Nach der Kettenregel aus **[HM I]**, *Kapitel 2,* §3*c)* ist

$$F'(c) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \frac{dx_i(c)}{dc} .$$

Genauso können wir

$$G(c) \underset{\text{def}}{=} g\big(x_1(c), \ldots, x_n(c)\big)$$

betrachten und erhalten

$$G'(c) = \sum_{i=1}^{n} \frac{\partial g}{\partial x_i} \frac{dx_i(c)}{dc} .$$

Da $\big(x_1(c), \ldots, x_n(c)\big)$ ein Optimum ist, sind dort die Gradienten von f und $g - c$ proportional mit Proportionalitätsfaktor λ. Da wir bei der Gradientenbildung nur nach den x_i ableiten, von denen die rechte Seite c nicht abhängt, ist der Gradient von $g - c$ gleich dem von g selbst, d.h.

$$\frac{\partial f}{\partial x_i} = \lambda \frac{\partial g}{\partial x_i} \qquad \text{für alle } i .$$

Somit ist $F'(c) = \lambda G'(c)$. Da der Punkt $\big(x_1(c), \ldots, x_n(c)\big)$ die Nebenbedingung mit rechter Seite c erfüllt, ist aber $G(c) = c$ und damit $G'(c) \equiv 1$. Also ist $\lambda = F'(c)$ die Wachstumsrate für das Optimum bei Änderung der rechten Seite der Nebenbedingung.

Im obigen Beispiel steigt also die Maximalmenge $f(x, y)$, die mit Kapitaleinsatz 12 produziert werden kann, für kleines h ungefähr um $h/4$, wenn wir den Kapitaleinsatz auf $12 + h$ erhöhen. Die Erhöhung des Kapitaleinsatzes lohnt sich, wenn für das fertige Produkt ein Preis pro Einheit erzielt werden kann, der größer ist als vier.

Als letztes wollen wir uns noch überlegen, was passiert, wenn wir nicht nur eine, sondern mehrere Nebenbedingungen erfüllen müssen. Es geht

also wieder darum, eine Funktion $f(x_1, \ldots, x_n)$ zu optimieren, jetzt aber unter den Nebenbedingungen

$$g_1(x_1, \ldots, x_n) \geq 0, \quad \ldots \quad g_r(x_1, \ldots, x_n) \geq 0\,.$$

(Es genügt, Bedingungen mit $\geq$ zu betrachten, denn durch Multiplikation mit minus Eins kann man jede Ungleichung mit $\leq$ in eine mit $\geq$ überführen. Auch Gleichungen $g_i = 0$ kann man zumindest formal durch die beiden Ungleichungen $g_i \geq 0$ und $-g_i \geq 0$ ausdrücken.)

Die wichtigsten Beispiele solcher Optimierungsaufgaben sind die Fälle mit linearen Funktionen f und g_i; hier redet man von *linearen Programmen.* (Das Wort *Programm* in diesem Zusammenhang hat natürlich nichts mit Computerprogrammen zu tun.) Das wichtigste Verfahren zur Lösung solcher Aufgaben, der Simplex-Algorithmus, wird in der Vorlesung *Numerik I* behandelt, so daß wir uns hier auf die *nichtlineare Programmierung* beschränken können.

Man überlegt sich leicht, daß im linearen Fall die Nebenbedingungen ein (endliches oder unendliches) Polyeder im $\mathbb{R}^n$ definieren und eine lineare Funktion, so sie ein endliches Maximum oder Minumum hat, dieses auf dem Rand dieses Polyeders annimmt, und dort sogar in einer Ecke. Man muß daher „nur" die Ecken dieses Polyeders untersuchen – deren Anzahl allerdings wächst exponentiell mit der Anzahl der Variablen. Trotzdem führt der Simplex-Algorithmus selbst im Fall von Zehntausenden von Variablen in der Regel fast immer sehr schnell ans Ziel; das theoretische Problem der exponentiellen Komplexität im schlimmsten Fall hat also für praktische Anwendungen keine Bedeutung.

Bei nichtlinearen Funktionen ist die Situation komplizierter, denn nun kann es auch im Innern Extrema geben: Die Funktion

$$f(x, y) = e^{-x^2-y^2} \quad \text{mit der Nebenbedingung} \quad x^2 + y^2 \leq 1$$

etwa nimmt ihr Maximum im Punkt $(0, 0)$ an; auf dem Rand des Einheitskreises liegen nur die Minima. Im allgemeinen Fall eines nichtlinearen Programms kann ein Optimum also entweder ganz im Innern liegen oder aber eine beliebige Teilmenge der Nebenbedingungen exakt erfüllen.

Falls wir es mit inneren Punkte zu tun haben, sind diese lokale Maxima oder Minima ohne Nebenbedingungen, und wir haben uns bereits in §1

überlegt, wie man diese bestimmt: In jedem solchen Punkt verschwindet der Gradient der zu optimierenden Funktion.

Im Falle einer einzigen *Gleichung* als Nebenbedingung ist der Gradient von f linear abhängig vom Gradienten der Nebenbedingung; da der Nullvektor von jedem anderen Vektor linear abhängig ist, schließt dies auch den Fall der Optima bei inneren Punkten mit ein. Die naheliegende Verallgemeinerung auf den Fall mehrerer Nebenbedingungen ist der

Satz: Die Funktion $f\colon D \to \mathbb{R}$ auf $D \subseteq \mathbb{R}^n$ habe im Punkt $\boldsymbol{a} \in D$ ein Extremum unter den Nebenbedingungen

$$g_1(\boldsymbol{a}) \geq 0, \quad g_2(\boldsymbol{a}) \geq 0, \quad \ldots, \quad g_r(\boldsymbol{a}) \geq 0\,.$$

Dann sind die $r + 1$ Vektoren

$$\operatorname{grad} f(\boldsymbol{a}), \quad \operatorname{grad} g_1(\boldsymbol{a}), \quad \operatorname{grad} g_2(\boldsymbol{a}), \quad \ldots, \quad \operatorname{grad} g_r(\boldsymbol{a})$$

linear abhängig.

Der *Beweis* erfordert keine wesentlich neuen Ideen gegenüber dem Fall einer einzigen Nebenbedingung und sei daher nur kurz skizziert: Falls die Gradienten der g_i im Punkt $\boldsymbol{a}$ bereits untereinander linear abhängig sind, gibt es nichts mehr zu beweisen; nehmen wir also an, sie seien linear unabhängig. Dann gibt es (mindestens) r verschiedene Variablen x_{j_1} bis x_{j_r}, so daß

$$\frac{\partial g_i}{\partial x_{j_i}}(\boldsymbol{a}) \neq 0$$

ist. Also kann nach dem Satz über implizite Funktionen jede Nebenbedingung zur Elimination einer anderen Variablen benutzt werden, und im wesentlichen dieselbe Rechnung wie im Fall einer Nebenbedingung zeigt die Behauptung. ∎

Die lineare Abhängigkeit der Vektoren

$$\operatorname{grad} f(\boldsymbol{a}), \quad \operatorname{grad} g_1(\boldsymbol{a}), \quad \operatorname{grad} g_2(\boldsymbol{a}), \quad \ldots, \quad \operatorname{grad} g_r(\boldsymbol{a})$$

bezeichnet man als KUHN-TUCKER-Bedingung; sie ist eine offensichtliche Verallgemeinerung der Bedingung von LAGRANGE, ist allerdings deutlich jünger: Sie erschien 1951 in einer gemeinsamen Arbeit von

H.W. KUHN und A.W. TUCKER, vier Jahre, nachdem G. DANTZIG den Simplex-Algorithmus entwickelt hatte, und fast zweihundert Jahre, nachdem LAGRANGE seine Multiplikatoren zur Bestimmung von Extrema unter einer Nebenbedingung eingeführt hatte.

Das Problem bei der praktischen Anwendung des Satzes von KUHN und TUCKER besteht darin, daß in einem Optimum manche Nebenbedingungen als Gleichungen, andere als echte Ungleichungen erfüllt sind; man muß also jede der möglichen Kombinationen untersuchen.

Eine mögliche Abhilfe sind sogenannte *barrier*-Methoden: Man läßt die Nebenbedingungen eine Barriere errichten, indem man (bei der Suche nach einem Maximum) Maxima *ohne* Nebenbedingung der Funktion

$$f(x_1,\dots,x_n)+\sum_{i=1}^{r}\varepsilon_i\log g_i(x_1,\dots,x_n)$$

sucht, wobei die ε_i positive Konstanten sind. Da die Logarithmen am Rand gegen $-\infty$ gehen, liegen diese Maxima stets im Innern. Falls man nun alle ε_i in geeigneter Weise gegen Null gehen läßt, kann man in manchen Fällen zeigen, daß diese Maxima gegen Maxima der Funktion *mit* Nebenbedingung konvergiert.

Ein Beispiel dafür ist der 1984 gefundene Algorithmus von KARMAKAR für den Fall linearer Funktionen f, g_i. Er ist eine Alternative zum Simplex-Algorithmus, die stets in polynomialer Zeit zu einer Lösung führt, und war der erste mathematische Algorithmus, der patentiert wurde. In der Praxis ist er jedoch bei fast allen Problemen dem Simplex-Algorithmus unterlegen; lediglich bei einigen wenigen Spezialfällen, bei denen bekannt ist, daß der Simplex-Algorithmus schlecht funktioniert, führt KARMAKAR schneller zu einer Lösung.

§3: Numerische Verfahren

Wie wir gesehen haben, führt die Methode der LAGRANGEschen Multiplikatoren im allgemeinen auf nichtlineare Gleichungssysteme, die nur in einfachen Fällen explizit lösbar sind. In allen anderen Fällen muß man mit numerischen Methoden arbeiten, und da bietet sich an, das Problem

von vornherein ohne den Umweg über LAGRANGEsche Multiplikatoren Extrema numerisch zu bearbeiten.

a) Die Gradientenmethode

Für eine differenzierbare Funktion f auf $D\subseteq\mathbb{R}^n$ ist

$$f(\boldsymbol{x}+\vec{h})=f(\boldsymbol{x})+\operatorname{grad}f(\boldsymbol{x})\cdot\vec{h}+o\big(\|\vec{h}\|\big)\ ;$$

wenn wir ein Maximum (oder Minimum) von f ansteuern wollen, liegt es daher nahe, $\vec{h}$ so zu wählen, daß sich der Funktionswert möglichst stark vergrößert (oder verkleinert).

Nach der CAUCHY-SCHWARZschen Ungleichung ist

$$\left|\operatorname{grad}f(\boldsymbol{x})\cdot\vec{h}\right|\le\|\operatorname{grad}f(\boldsymbol{x})\|\cdot\left\|\vec{h}\right\|\ ;$$

wir erhalten also die maximalmögliche Veränderung bei vorgegebener Länge von $\vec{h}$ genau dann, wenn $\vec{h}$ parallel zum Gradienten ist.

Damit bietet sich folgende Strategie an: Wir wählen irgendeinen Ausgangspunkt $\boldsymbol{x}_0$ und berechnen dort den Gradienten $\nabla f(\boldsymbol{x}_0)$. Weiter geben uns eine Länge ℓ_0 für den Vektor $\vec{h}$ vor, die von der Länge des Gradienten abhängen kann oder auch nicht. Dann setzen wir bei der Suche nach einem Maximum

$$\vec{h}_0=\frac{\ell_0}{\|\nabla f(\boldsymbol{x}_0)\|}\nabla f(\boldsymbol{x}_0)\ ;$$

bei der Suche nach Minima nehmen wir das Negative davon.

Als nächstes betrachten wir den Punkt

$$\boldsymbol{x}_1\underset{\text{def}}{=}\boldsymbol{x}_0+\vec{h}_0\ ,$$

berechnen dort den Gradienten $\nabla f(\boldsymbol{x}_0)$, setzen mit geeignetem ℓ_1

$$\vec{h}_1=\pm\frac{\ell_1}{\|\nabla f(\boldsymbol{x}_1)\|}\nabla f(\boldsymbol{x}_1)$$

(+ für Maxima, − für Minima) zur Definition des nächsten Punkts

$$\boldsymbol{x}_2\underset{\text{def}}{=}\boldsymbol{x}_1+\vec{h}_1$$

und so weiter. In jedem Schritt erhöhen (oder erniedrigen) wir den Funktionswert soweit wie es mit der vorgegebenen Länge ℓ_i nur möglich ist in der Hoffnung, so irgendwann auf ein Maximum (oder Minimum) zu stoßen. Dieses können wir erreichen, wenn wir am Rand des Definitionsbereichs von f angelangt sind, oder aber wenn wir in einem Punkt sind, in dem der Gradient verschwindet: Von dort aus geht es mit diesem Verfahren nicht mehr weiter.

Da wir mit einem numerischen Verfahren nur ein verschwindend geringe Chance haben, exakt in einem Extremum zu enden, zeigt sich hier auch die Notwendigkeit einer intelligenten Wahl der Schrittweiten ℓ_i: Wenn diese zu groß sind, kann es passieren, daß wir endlos um ein Extremum herumoszillieren.

Theoretisch ist auch möglich, daß wir in einem Sattelpunkt landen, aber wenn man sich überlegt, wie die Gradienten in der Umgebung eines Sattelpunktes aussehen, wird schnell klar, daß dies nur sehr selten passiert.

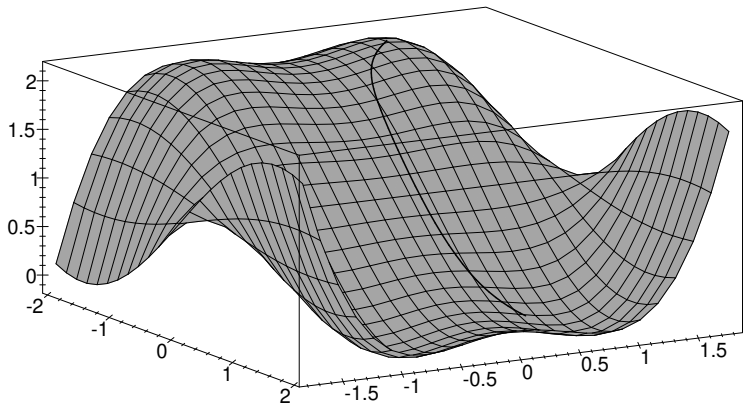


Abb. 58: Eine Anwendung der Gradientenmethode

Abbildung 58 zeigt ein einfaches Beispiel für einen mit der Gradientenmethode zurückgelegten Weg; hier wurde in jedem Schritt

$$\begin{pmatrix} h_i \\ k_i \end{pmatrix} = 0{,}1 \cdot \nabla f(x_i, y_i)$$

gesetzt. Der Weg geht offensichtlich recht zielstrebig auf das Maximum zu.

Abbildung 59 zeigt dasselbe Bild in einen etwas größeren Zusammenhang; hier sehen wir, daß unser Streben nach kurzfristigen Gewinnen langfristig wohl doch nicht so erfolgreich war: Wenn wir vom Startpunkt aus nach rechts in die kleine Mulde abgestiegen wären, hätten wir auf dem gegenüberliegenden Hang deutlich größere Funktionswerte erreicht als im lokalen Maximum, in dem wir schließlich gelandet sind.

Dies ist ein grundsätzliches Problem von Gradientenverfahren: Falls man sie in der Nähe des (absoluten) Optimums starten läßt, führen sie schnell und zuverlässig ans Ziel, ansonsten aber ist die Gefahr sehr groß, daß man in einem nur lokalen Optimum steckenbleibt.

Um von dort wieder weiterzukommen, gibt es verschiedene Strategien. Eine anschaulich recht klare ist die sogenannte „Tunnelung“. Der Name entstand aus der Betrachtung von Minimisierungsproblemen; nehmen wir also an, wir wollen das Minimum der Funktion $f(x, y)$ in einem gewissen Bereich finden und ein Gradientenverfahren hat uns in einen Punkt $\boldsymbol{x}_M$ geführt, von dem aus es nicht mehr weiterkommt. Um zu sehen, ob $z_M = f(\boldsymbol{x}_M)$ wirklich der kleinste Wert ist, den f im betrachteten Bereich annehmen kann, versuchen wir, eine weitere Lösung der Gleichung

$$f(\boldsymbol{x}) = z_M$$

zu finden. Dafür gibt es eine ganze Reihe numerischer Verfahren, z.B. das Verfahren von NEWTON-RAPHSON, mit denen sich zumindest ein solcher Punkt leicht finden läßt. Leider könnte dieser Punkt unser Ausgangspunkt $\boldsymbol{x}_M$ sein; deshalb sucht man tatsächlich nicht nach Lösungen der Gleichung $f(\boldsymbol{x}) = z_M$, sondern nach Lösungen einer leicht abgewandelten Gleichung der Form

$$\widetilde{f}(\boldsymbol{x}) = z_M \ ,$$

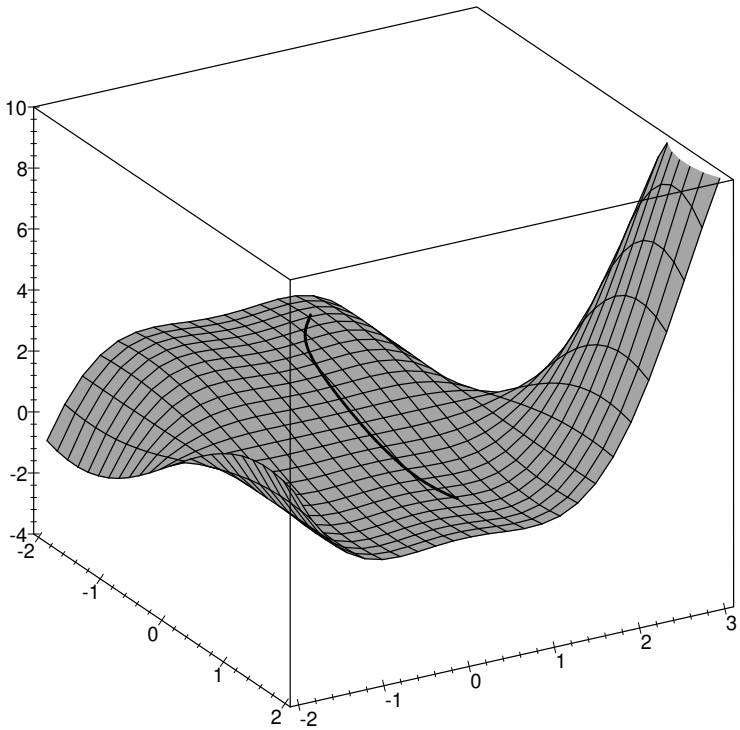


Abb. 59: Der Weg aus Abb. 58 aus einem weiteren Blickwinkel

wobei $\widetilde{f}$ dadurch aus f entsteht, daß man die Funktionswerte in der unmittelbaren Umgebung von (x_M, y_M) starkt anhebt, so daß das dortige Minimum verschwindet. Dazu kann man beispielsweise eine Funktion der Form

$$G(x,y) = ae^{\frac{(x-x_M)^2+(y-y_M)^2}{b}}$$

mit geeigneten Parametern a, b wählen, wie sie in Abbildung 60 zu sehen ist, und

$$\widetilde{f}(x,y) = f(x,y) + G(x,y)$$

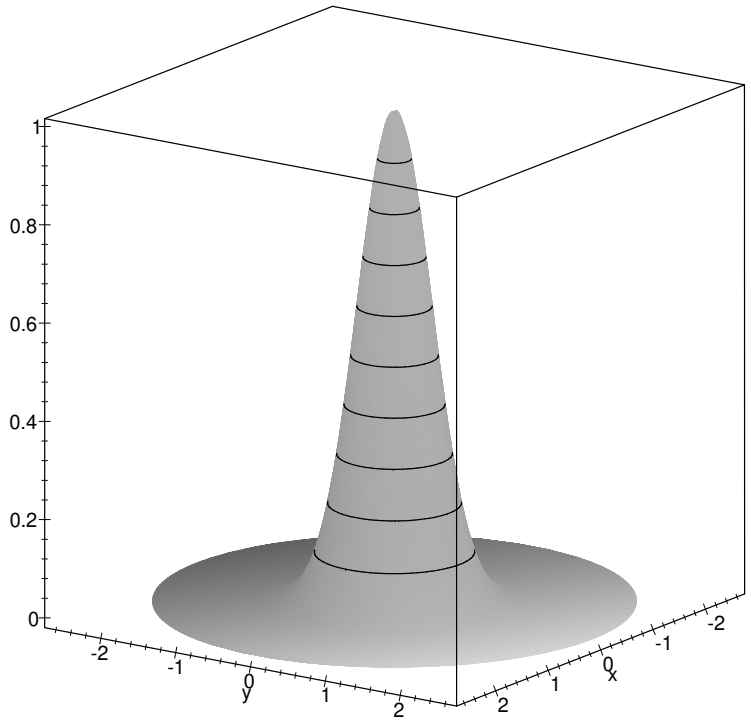


Abb. 60: $G(x,y) = e^{-3(x^2+y^2)}$

setzen.

Dies bringt das Minimum im Punkt $\boldsymbol{x}_M$ zum Verschwinden und verändert die Funktion praktisch nicht, wenn man nur hinreichend weit entfernt ist von $\boldsymbol{x}_M$. (Je kleiner b ist, umso lokalisierter ist die Veränderung.) Eine Lösung der Gleichung

$$\widetilde{f}(\boldsymbol{x}) = z_M \ ,$$

so es eine gibt, liegt also nicht in der unmittelbaren Umgebung von $\boldsymbol{x}_M$ und ist daher ein guter Ausgangspunkt, um dort die Gradientenmethode

noch einmal zu starten bis zum nächsten lokalen Minimum und so weiter. Sobald die Gleichung nicht mehr lösbar ist, können wir ziemlich sicher sein, daß z_M das globale Minimum ist – es sei denn, wir hätten die Parameter a und b sehr dumm gewählt.

Tunnelung ist auch ein wichtiges Konzept in der Physik: Dort versucht ein System bekanntlich stets, sein Energieminimum zu erreichen. Dies kann jedoch daran scheitern, daß es sich in einem lokalen Minimum befindet und nicht genügend Energie aufbringen kann, um den Energiewall zu überwinden, der es vom absoluten Minimum trennt. Zumindest im Bereich der Quantentheorie gibt es dann auch den sogenannten *Tunneleffekt,* der es einzelnen Teilchen erlaubt, diesen Wall zu tunneln und auf diese Weise einen Zustand niedrigerer Energie zu erreichen.

Im obigen Beispiel geht es nicht um ein Minimum, sondern um ein Maximum, da die Suche danach graphisch besser darstellbar ist. Also graben wir auch keinen Tunnel, sondern spannen ein Hochseil, das irgendwo auf der eingezeichneten Ebenen liegt und uns vom erreichten Zwischenhoch zur Startposition für einen weiteren Anstieg bringt. (Tatsächlich ist die Ebene etwas zu tief eingezeichnet, damit man das alte Maximum noch erkennen kann; das Seil muß also etwas höher hängen.)

b) Der Metropolis-Algorithmus

Eine weitere Idee zur Vermeidung von Zwischenhochs hat ebenfalls viel mit Physik zu tun: Ein Gas erreicht seinen Zustand minimaler Energie dann, wenn die Bewegungsenergie $\frac{1}{2}mv^2$ eines jeden Teilchens gleich null ist, wenn sich also nichts mehr bewegt. Dies geschieht aber höchstens am absoluten Nullpunkt; bei positiven Temperaturen werden die meisten Teilchen positive kinetische Energie haben. Nach LUDWIG BOLTZMANN ist dabei die Wahrscheinlichkeit dafür, daß ein Teilchen die Energie $E = \frac{1}{2}mv^2$ hat, bei Temperatur T proportional zu

$$e^{-\frac{E}{kT}},$$

mit einer Konstanten $k \approx 1{,}38066 \cdot 10^{-23}\, J/K$, die heute als BOLTZMANN-Konstante bezeichnet wird.

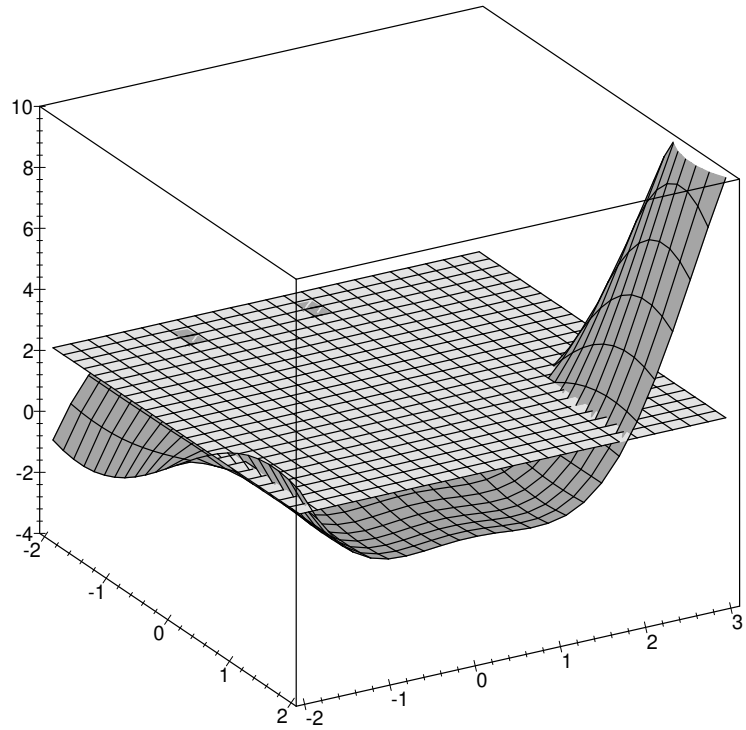


Abb. 61: „Tunnelung“ für Maxima

LUDWIG BOLTZMANN (1844–1906) wuchs auf und studierte in Wien; danach lehrte er in Graz, Heidelberg, Berlin, Graz, Wien, Graz, Wien, Leipzig und Wien. Er war Professor für Theoretische Physik, für Mathematik und für Experimentalphysik. Auf seiner letzten Stelle in Wien hielt er eine so erfolgreiche Philosophievorlesung, daß ihn Kaiser Franz Josef in den Palast einlud. Am bekanntesten ist er für die Begründung der statistischen Mechanik, einer damals sehr umstrittene Theorie. Ob die damit verbundenen Anfeindungen zu seinem Selbstmord führten, ist unbekannt.

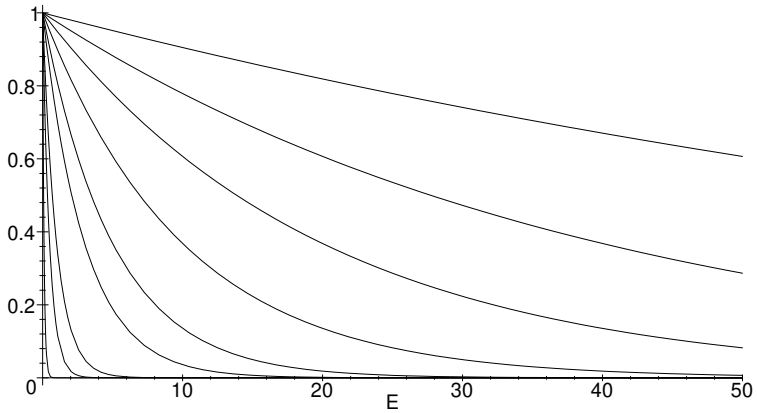


Abb. 62: $e^{-E/kT}$ für $kT = 0{,}1, 0{,}5, 1, 3, 5, 10, 20, 40, 100$

Bei der *simulierten Abkühlung* oder BOLTZMANN-*Maschine* ahmt man dies nach, indem man mit einer hohen Temperatur startet und der Richtung, in der man weitergeht, einer dieser Temperatur entsprechende Freiheit läßt. Man geht also nicht mehr unbedingt in Richtung des Gradienten, sondern geht zufällig in eine von endlich vielen vorgegebenen Richtungen. Die Wahrscheinlichkeit für den Richtungsvektor $\vec{h}_j$ soll dabei analog zur BOLTZMANN-Verteilung festgelegt werden, d.h. wir ordnen ihm eine „Energie“

$$E_j = \pm\big(f(\boldsymbol{x} + \vec{h}_j) - f(x, y)\big)$$

zu (positiv bei der Suche nach einem Minimum, negativ bei der Suche nach einem Maximum) und die Wahrscheinlichkeit dafür, daß wir in Richtung $\vec{h}_j$ gehen, soll proportional sein zu $e^{-E_j/kT}$. Sie ist also, falls N Richtungen zur Verfügung stehen, gleich

$$p_j \underset{\text{def}}{=} \frac{e^{-E_j/kT}}{\sum_{\ell=1}^{N} e^{-E_\ell/kT}}.$$

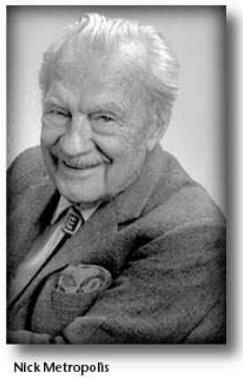
Zur Wahl einer Richtung erzeugen wir uns somit eine Zufallszahl Z zwischen null und eins und gehen in Richtung $\vec{h}_j$, wenn

$$\sum_{\ell=1}^{j-1} p_\ell < Z \le \sum_{\ell=1}^{j} p_\ell$$

ist. (Die Frage, wie lang die Richtungsvektoren im wievielten Schritt sein sollen, wollen wir hier ausklammern.)

Bie hohen Temperaturen ist damit die Richtung fast vollständig zufallsbedingt gewählt, während in der Nähe des absoluten Nullpunkts praktisch nur noch die optimale Richtung eine Chance hat. Falls wir bei hoher Temperatur in einem Zwischenextremum landen, sorgt dies also mit recht hoher Wahrscheinlichkeit dafür, daß wir dort nicht steckenbleiben.

Am Ende wollen wir allerdings im absoluten Optimum steckenbleiben, d.h. wir müssen die Temperatur im Verlauf der Rechnung immer weiter senken – daher der Name *simulated annealing* = simulierte Abkühlung. Bei der Anwendung auf Optimierungsprobleme bezeichnet man diese Vorgehensweise als den METROPOLIS-Algorithmus. In welcher Weise man die Temperatur am besten senkt, ist immer noch ein Gebiet aktiver Forschung. Man kann zeigen, daß man statistisch betrachtet praktisch immer im Optimum landet, wenn man mit einer hinreichend hohen Ausgangstemperatur T_1 startet und im r-ten Schritt mit Temperatur $T_1/\log(r+1)$ arbeitet, aber bei einer derart langsamen Abkühlung braucht der Algorithmus viel zu lange, um ans Ziel zu kommen.



Nick Metropolis

NICHOLAS METROPOLIS (1915–1999) wuchs auf in Chicago, wo er Physik studierte und 1941 promovierte. Seit 1943 arbeitete er, unterbrochen durch Professuren an der Universität Chicago von 1945–1948 und 1957–1965, in den Los Alamos Labaratories, die ihn im Nachruf als *giant of mathematics and one of the founders of the Information Age* bezeichneten. Sein Ruhm als Mathematiker beruht vor allem auf den von ihm entwickelten Anwendungen statistischer Verfahren auf eine Vielzahl von mathematischen Problemen; zum Pionier des Informationszeitalter macht ihn u. a., daß er einer der ersten Anwender des ersten elektronischen Computers ENIAC war, dessen Nachfolger MANIAC baute und an der Universität Chicago das Institute for Computer Research gründete und bis 1965 leitete.

In Abbildung 63 sieht man, wie sich der Algorithmus bei einer Abkühlungsregel verhält, die im r-ten Schritt mit Temperatur T_1/r arbeitet: Zumindest im gezeigten Fall funktioniert das recht gut.

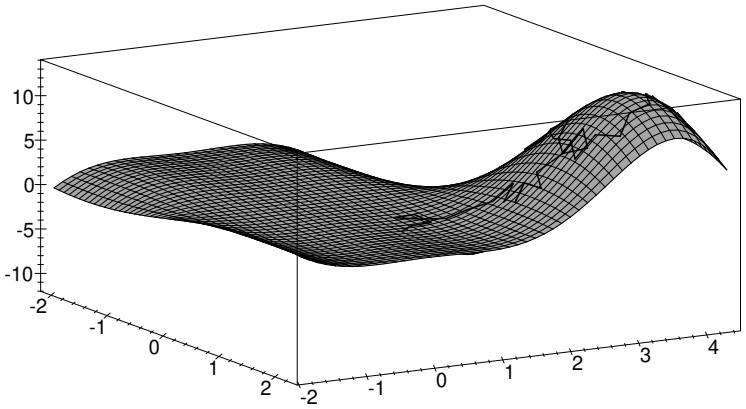


Abb. 63: Der METROPOLIS-Algorithmus für obiges Problem

In anderen Fällen (d.h., wenn andere Zufallszahlen gezogen werden) bleibt man damit aber auch gelegentlich ziemlich lange im Tal hängen; ein Beipspiel dafür zeigt Abbildung 64.

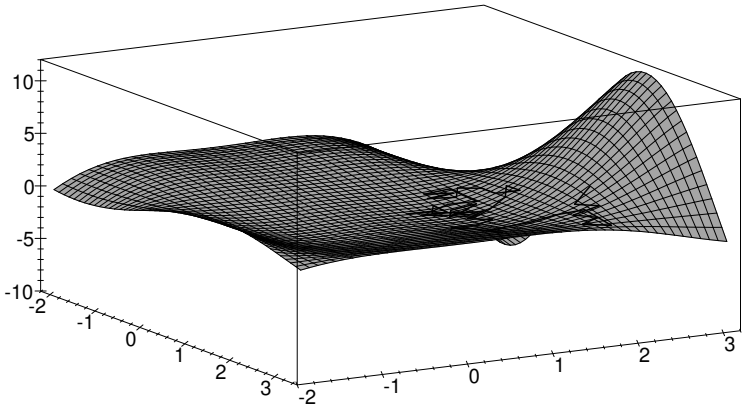


Abb. 64: Ditto mit anderen Zufallszahlen

Auch hier kommt man aber immerhin in eine gute Startposition, und oft wird es am besten sein, nach hinreichend vielen METROPOLIS-Schritten

einfach ein gewöhnliches Gradientenverfahren zu starten.

Zusammenfassend läßt sich sagen, daß der METROPOLIS-Algorithmus und verwandte Verfahren (die sogenannten Monte-Carlo-Methoden) sehr nützliche Hilfsmittel zur Optimierung sind, falls man so gut wie nichts über die zu optimierende Funktion weiß. Sie funktionieren nicht nur bei kontinuierlichen Problemen, wie den hier betrachteten, sondern auch für diskrete und kombinatorische Optimierungsprobleme.

Sie haben allerdings den Nachteil, daß man nie garantieren kann, daß man ein Optimum erreichen wird, und selbst wenn man eines erreicht, kann die Methode dies nicht erkennen. (Es gibt alternative numerische Methoden, die das können.)

c) Zusammenfassung

Die nichtlineare Optimierung ist ein sehr weites Feld, von dem eine Grundvorlesung wie die *Höhere Mathematik* nur einen kleinen Ausschnitt behandeln kann. Dieser Ausschnitt besteht nicht aus den für die Praxis wichtigsten Verfahren, sondern aus denen, die sich am besten in den Stoff der Vorlesung einordnen. Sie sind zwar (in Kombination mit dem aus der *Numerik* bekannten Simplex-Verfahren) die Grundbausteine, aus denen die meisten praktisch relevanten Verfahren zusammengesetzt sind, aber für die vielen kleinen Abwandlungen, die dazu führen, daß man ein Problem wirklich effizient lösen kann, müßte man deutlich mehr Zeit aufwenden, als hier zur Verfügung steht. Interessenten seien auf entsprechende Spezialvorlesung aus dem Bereich der Mathematik oder Operations Research verwiesen.

§4: Grundzüge der Fehler- und Ausgleichsrechnung

Physikalische Gesetze machen meist nur dann eine Aussage über ein reales System, wenn alle Umgebungsbedingungen exakt kontrolliert werden können. Das ist in der Praxis natürlich nie möglich. Insbesondere hat man bei der Anwendung physikalischer Prinzipen zur Messung von Daten keine Chance, den exakten Wert der zu messenden Größe

zu bestimmen; der gemessene Wert wird immer von zahlreichen kleineren Störungen beeinflußt sein, die man bei einem gut durchgeführten Experiment für alle praktischen Zwecke als zufällig betrachten kann.

Zusätzlich kann die Messung noch durch mehr oder weniger große *systematische* Fehler verfälscht sein; diese können hervorgerufen werden durch ein falsch kalibriertes Meßgerät, Ablesen auf der falschen Skala eines Meßinstruments, durch falsche Anwendung von Meßvorschriften *usw.* Mit diesen systematischen Fehlern wollen wir uns hier nicht beschäftigen; in diesem Paragraphen soll es nur um *Zufallsfehler* gehen.

a) Das Laplacesche Fehlermodell

Der französische Mathematiker PIERRE SIMON, MARQUIS DE LAPLACE (1749–1827), dem wir in dieser Vorlesung bereits mehrfach begegnet sind, entwickelte ein extrem vereinfachtes Modell für das Zustandkommen zufälliger Meßfehler. Trotz seiner unrealistischen Annahmen ist es als Einstieg in die Fehlerrechnung noch immer interessant, denn wie wir bald sehen werden, gelten seine Schlußfolgerungen viel allgemeiner und zumindest näherungsweise auch in vielen praktischen Situationen.

Die Grundannahme des LAPLACEschen Fehlermodells können wir uns so vorstellen, daß eine große Anzahl von „Dämonen“ (oder Fehlerquellen) unsere Meßergebnisse verfälschen; jeder einzelne dieser „Dämonen“ verursacht einen Fehler derselben Größe ε in positiver oder negativer Richtung, wobei die Wahrscheinlichkeit für $+\varepsilon$ *bzw.* $-\varepsilon$ für jeden der „Dämonen“ jeweils 50% sein soll und die einzelnen „Dämonen“ unabhängig voneinander handeln sollen.

Im Falle eines einzigen „Dämonen“ wäre der Fehler also mit gleicher Wahrscheinlichkeit $+\varepsilon$ oder $-\varepsilon$, bei zwei „Dämonen“ wäre er in jeweils 25% aller Fälle $+2\varepsilon$ oder -2ε, während sich in 50% der Fälle die beiden Fehler aufheben würden.

Allgemein gibt es bei n „Dämonen“ 2^n gleichwahrscheinliche Möglichkeiten für deren Verhalten; die folgende Tabelle zeigt für $n \leq 5$ jeweils die Anzahl der Fälle, die zu dem in der Kopfzeile angegebenen Gesamtfehler führen:

	-5ε	-4ε	-3ε	-2ε	$-\varepsilon$	0	$+\varepsilon$	$+2\varepsilon$	$+3\varepsilon$	$+4\varepsilon$	$+5\varepsilon$
$n = 0$						1					
$n = 1$					1		1				
$n = 2$				1		2		1			
$n = 3$			1		3		3		1		
$n = 4$		1		4		6		4		1	
$n = 5$	1		5		10		10		5		1

Diese dreiecksförmige Anordnung von Zahlen bezeichnet man als PASCAL*sches Dreieck.* Offenbar kann man es dadurch rekursiv zeilenweise berechnen, daß man an jede Stelle die Summe der beiden links und rechts darüberstehenden Zahlen schreibt: Die n-te Störung bringt den Fehler genau dann auf $i \cdot \varepsilon$, wenn sie entweder gleich $+\varepsilon$ ist und die ersten $n-1$ Störungen einen Fehler $(i-1) \cdot \varepsilon$ produziert haben, oder aber wenn sie gleich $-\varepsilon$ ist und die ersten $n-1$ Störungen einen Fehler $(i+1) \cdot \varepsilon$ produziert haben. Entsprechend ist auch klar, daß die Summe aller Zahlen in der n-ten Zeile gleich 2^n ist, denn in der nullten Zeile haben wir Summe eins, und da die jeweils neu hinzukommende Störung genau zwei Möglichkeiten hat, verdoppelt sich die Summe von Zeile zu Zeile. Die Wahrscheinlichkeit dafür, daß sich n Störungen zu $i \cdot \epsilon$ aufsummieren, ist also gerade gleich der Zahl, die in der n-ten Spalte unter $i \cdot \varepsilon$ steht (beziehungsweise Null, wenn dort keine Zahl steht), dividiert durch 2^n.

Bekanntlich kann man die Zahlen in diesem Dreieck auch explizit berechnen: An der n-ten Zeile stehen die $n+1$ Zahlen

$$\binom{n}{i} = \frac{n!}{i!(n-i)!} \quad \text{für} \quad i = 0, \ldots, n\,.$$

Wer diese Formel nicht kennt, kann sie leicht durch vollständige Induktion beweisen: Für $n = 1$ sowie allgemein für $i = 0$ oder $i = n$ ist alles klar; für $n > 1$ und $0 < i < n$ stehen über $\binom{n}{i}$ die beiden Zahlen $\binom{n-1}{i-1}$ und $\binom{n-1}{i}$, für die in der Tat gilt

$$\binom{n-1}{i-1} + \binom{n-1}{i} = \frac{(n-1)!}{(i-1)!(n-i)!} + \frac{(n-1)!}{i!(n-i-1)!} = \frac{(n-1)!}{i!(n-i)!} \cdot \Big(i + (n-i)\Big)$$
$$= \frac{n!}{i!(n-i)!} = \binom{n}{i}\,.$$

Betrachten wir als nächstes die Größe des Gesamtfehlers. Falls n gerade ist, treten nur Vielfache von 2ε auf und alle diese Vielfachen zwischen $-n\varepsilon$ und $n\varepsilon$ kommen tatsächlich vor; ensprechend sind für ungerades n nur ungeradzahlige Vielfache von ε möglich, und auch hier werden wieder alle solchen Werte zwischen $-n\varepsilon$ und $n\varepsilon$ angenommen. Wir können dies dadurch zusammenfassen, daß in beiden Fällen genau die Werte $(n-2k)\varepsilon$ mit $k = 0, \ldots, n$ angenommen werden, und das PASCALsche Dreieck zeigt, daß der Fehler $(n-2k)\varepsilon$ in

$$\binom{n}{n-k} = \binom{n}{k}$$

Fällen auftritt. Da n „Dämonen“ insgesamt 2^n Möglichkeiten zur Fehlererzeugung haben, ist die Wahrscheinlichkeit für den Gesamtfehler $(n-2k)\varepsilon$ also

$$\binom{n}{k} \cdot 2^{-n}.$$

Diese Wahrscheinlichkeit sollte für einen festen Fehlerbetrag im wesentlichen unabhängig von n sein: Da wir nicht wirklich an Dämonen glauben, können wir deren Anzahl schließlich nicht in ein realistisches Fehlermodell einfließen lassen.

Der Formel können wir dies allerdings nicht ansehen, und die Berechnung der Wahrscheinlichkeiten wird für große n auch schnell sehr aufwendig, da die Binomialkoeffizienten schnell sehr groß werden. Um trotzdem einen Eindruck davon zu bekommen, was für größere n passiert, sind auf der nächsten Doppelseite die Wahrscheinlichketen für $n = 5, 10, 50, 100, 500$ und 1000 graphisch dargestellt. (Die Tatsache, daß ab $n = 50$ deutlich weniger als $n + 1$ Balken zu sehen sind, erklärt sich daraus, daß die restlichen Wahrscheinlichkeiten zu klein sind, um noch sichtbar zu sein.)

Die Balkendiagramme zeigen, daß sich die Verteilung der Fehlerwahrscheinlichkeiten für große n einer festen Kurve annähern sollte, der in Abbildung 65 (und früher auch auf jedem Zehnmarkschein) zu findenden *Glockenkurve* oder GAUSS-Kurve.

Wenn sich Fehler oder auch beliebige Daten so veteilen, wie es dieser Kurve entspricht, redet man von *normalverteilten* Daten. Damit haben

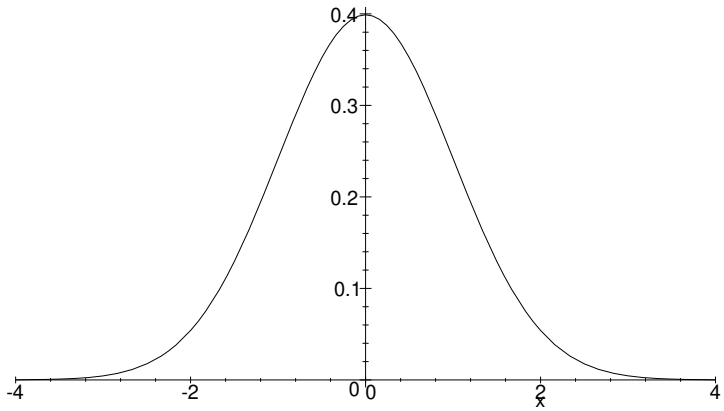


Abb. 65: Die „Glockenkurve“ $y = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$

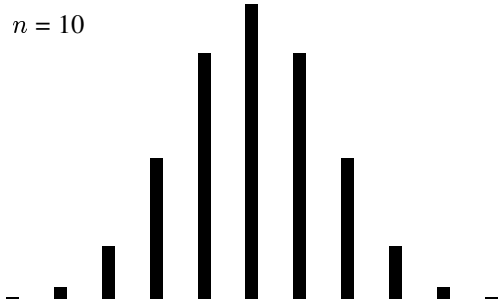
wir also zumindest graphisch gesehen, daß Meßfehler nach dem LAPLACEschen Fehlermodell normalverteilt sind. Das sagt noch nicht unbedingt etwas über die Verteilung realer Meßfehler, da das LAPLACEsche Fehlermodell von unrealistisch einfachen Annahmen ausgeht; nach einem der fundamentale Gesetze der Statistik, dem *zentralen Grenzwertsatz,* führen aber auch realistischere Annahmen zu genau derselben Verteilung: Sind $u_1, \ldots, u_n$ beliebige Quellen von Zufallsfehlern, über deren Verteilung wir (fast) nichts voraussetzen müssen, so ist ihre Summe für hinreichend großes n annähernd normalverteilt; siehe §5. Das eingeklammerte Wort „fast“ ist dabei für praktische Zwecke bedeutungslos, und als „groß“ kann man sich ein n ab etwa dreißig oder vierzig vorstellen.

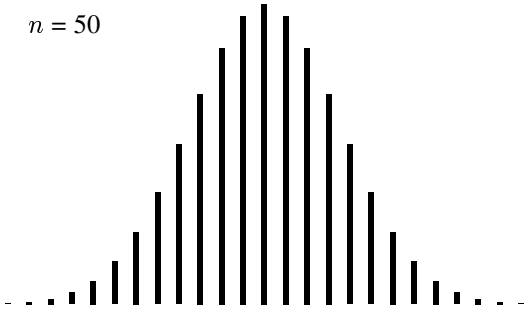
b) Statistische Kenngrößen

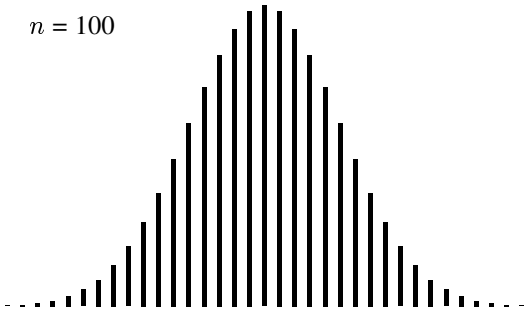
Die übliche Strategie zum Umgang mit Zufallsfehlern ist wohlbekannt: Man begnügt sich nicht mit einer einzigen Messung, sondern mißt mehrfach, so daß man eine ganze Meßreihe $x_1, x_2, \ldots, x_N$ erhält. Dann bildet man das *arithmetische Mittel*

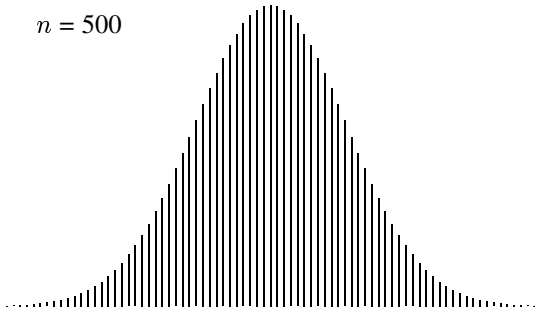
$$\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

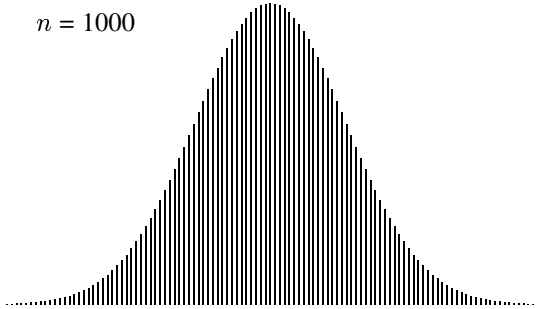
$n = 5$


$n = 10$


$n = 50$


$n = 100$


$n = 500$


$n = 1000$

der Meßreihe in der Hoffnung, daß sich hierbei die Fehler „ausmitteln“, so daß $\overline{x}$ dem theoretisch korrekten Wert $\hat{x}$ nahekommt.

Die Wahl des arithmetischen Mittels läßt sich auch geometrisch begründen: Eine Meßreihe $x_1, \ldots, x_N$ für eine Meßgröße mit exaktem Wert $\hat{x}$ definiert einen Vektor im $\mathbb{R}^n$. Falls es keine Meßfehler gäbe, hätte dieser lauter identische Komponenten $\hat{x}$. Tatsächlich ist dies natürlich nicht der Fall; wir können aber nach einem Vektor mit identischen Komponenten suchen, der möglichst nahe am Vektor der Meßwerte liegt. Für einen Vektor, dessen sämtliche Komponenten gleich x sind, ist der EUKLIDische Abstands zum Vektor der Meßwerte gleich

$$d(x) = \sqrt{\sum_{i=1}^{N}(x - x_i)^2} = \sqrt{Nx^2 - 2x\sum_{i=1}^{N} x_i + \sum_{i=1}^{N} x_i^2}\,.$$

Die quadratische Funktion $d(x)^2$ hat ein eindeutig bestimmtes Minimum bei der Nullstelle ihrer Ableitung

$$2Nx - 2\sum_{i=1}^{N} x_i\,,$$

also beim arithmetischen Mittel $\overline{x}$, und dieses ist auch das einzige Minimum von $d(x)$. Wir nehmen daher das arithmetische Mittel $\overline{x}$ als besten verfügbaren Schätzwert für den unbekannten korrekten Wert $\hat{x}$.

Als Maß für die Schwankungen innerhalb der Meßreihe und damit für die Meßfehler könnte man versucht sein, den Abstand $d(\hat{x})$ zu nehmen; er hat aber den Nachteil, daß er mit steigendem N immer größer wird, d.h. die Schwankungen würden umso größer, je mehr man mißt. Das ist natürlich absurd; daher dividieren wir das Abstandsquadrat noch durch N und definieren die *mittlere quadratische Abweichung* oder *Varianz* der Meßreihe als

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(\hat{x} - x_i)^2\,.$$

Die (nichtnegative) Quadratwurzel σ hieraus heißt *Standardabweichung* der Meßreihe.

Das Ergebnis einer Messung wird meist angegeben in der Form

$$x = \overline{x} \pm \sigma\,,$$

man betrachtet also die Standardabweichung der Meßreihe als Maß für den Meßfehler. Da deren Definition allerdings vom (im allgemeinen unbekannten) korrekten Wert $\hat{x}$ abhängt, können wir sie nicht berechnen, sondern müssen im folgenden sehen, wie wir sie zumindest schätzen können.

Als einfachste Möglichkeit bietet sich an, σ^2 durch

$$\frac{1}{N}\sum_{i=1}^{N}(\overline{x} - x_i)^2$$

zu schätzen, aber das führt sicherlich zu einem zu kleinen Ergebnis: Schließlich ist $d(\overline{x})$ das eindeutig bestimmte Minimum der Abstandsfunktion d, so daß der korrekte Wert $d(\hat{x})$ für $\hat{x} \neq \overline{x}$ notwendigerweise größer sein muß.

In Abschnitt *d)* werden wir aus dem Fehlerfortpflanzungsgesetz einen besseren Schätzwert für σ herleiten.

Warum betrachten wir eigentlich quadratische Abweichungen und nicht die einfacheren linearen Abweichungen? Nun, der Mittelwert aller Abweichungen ist

$$\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x}) = \frac{1}{N}\sum_{i=1}^{N} x_i - \frac{1}{N}N\overline{x} = \overline{x} - \overline{x} = 0\,,$$

also ist dies keine geeignete Maßzahl. Möglich wäre die mittlere *betragsmäßige* Abweichung

$$\frac{1}{N}\sum_{i=1}^{N}|x_i - \overline{x}|\,,$$

allerdings wird die im allgemeinen nicht für das arithmetische Mittel $\bar{x}$ minimal, sondern, wie man sich leicht überlegen kann, für jede Zahl $\overline{x}$ mit der Eigenschaft, daß gleichviele Meßwerte größer und kleiner als $\overline{x}$ sind; einen solchen Wert $\overline{x}$ bezeichnet man als *Median* der Meßreihe. Für die Beschreibung wirtschafts- und sozialwissenschaftlicher Daten ist dieser Median meist eine aussagekräftigere Kennzahl als das arithmetische Mittel; in den Naturwissenschaften und der Technik spielt er allerdings keine große Rolle. Im nächsten Paragraphen werden wir sehen, daß auch das LAPLACEsche Fehlermodell in natürlicher Weise auf quadratische Abweichungen führt.

c) Das Fehlerfortpflanzungsgesetz

Gegeben seien zwei Größen

$$x = \hat{x} \pm \sigma_x \quad \text{und} \quad y = \hat{y} \pm \sigma_y$$

(die Verallgemeinerung auf mehr als zwei Größen erfordert, wie man sich bei der folgenden Rechnung leicht klarmacht, nur etwas mehr Schreibaufwand; sie ist nicht prinzipiell schwieriger), und eine Größe

$$w = f(x, y)\,,$$

die von diesen beiden abhängt. Um vernünftige Aussagen machen zu können, setzen wir dabei f als stetig differenzierbar voraus.

Für x seien N Meßwerte $x_1, \ldots, x_N$ gegeben, und für y entsprechend M Werte $y_1, \ldots, y_M$. Wenn wir echte Zufallsfehler haben, können wir davon ausgehen, daß die Fehler der x-Werte und die der y-Werte voneinander unabhängig sind, und das wollen wir im folgenden auch annehmen.

Für w haben wir dann NM Werte $w_{ij} = f(x_i, y_j)$, deren Mittelwert die beste Schätzung für den „wahren" Wert $\widehat{w} = f(\hat{x}, \hat{y})$ ist. Dieser Mittelwert ist für komplizierte Funktionen f und/oder große Werte von n und m umständlich auszurechnen; günstiger wäre es, einfach den Mittelwert $\bar{x}$ der x_i und den Mittelwert $\bar{y}$ der y_j zu berechnen, um dann $f(\bar{x}, \bar{y})$ als Schätzung für $\widehat{w}$ zu benutzen. Zur Abschätzung des dadurch bedingten Fehler setzen wir

$$x_i = \bar{x} + h_i \quad \text{und} \quad y_j = \bar{y}_j + k_j\,;$$

dann ist wegen der Differenzierbarkeit von f

$$\begin{aligned} w_{ij} &= f(x_i, y_j) = f(\bar{x} + h_i, \bar{y} + k_j) \\ &= f(\bar{x}, \bar{y}) + f_x(\bar{x}, \bar{y})h_i + f_y(\bar{x}, \bar{y})k_j + o\left(\sqrt{h_i^2 + k_j^2}\right)\,, \end{aligned}$$

wobei

$$f_x = \frac{\partial f}{\partial x} \quad \text{und} \quad f_y = \frac{\partial f}{\partial y}$$

die partiellen Ableitungen von f bezeichnen. Da die h_i und die k_j als Abweichungen vom Mittelwert die Summe null haben, ist also der Mittelwert der w_{ij} bis auf einen Fehler der Größenordnung $o\left(\sqrt{h^2 + k^2}\right)$ gleich $f(\bar{x}, \bar{y})$, wobei h, k die Batragsmaxima der h_i, k_j sind.

Als nächstes müssen wir den Fehler von $\widehat{w}$ berechnen, also den Mittelwert der $(w_{ij} - \widehat{w})^2$. Dazu schreiben wir zunächst

$$x_i = \hat{x} + u_i \quad \text{und} \quad y_j = \hat{y}_j + v_j\,,$$

betrachten also anstelle der Abweichungen vom Mittelwert die echten Meßfehler, und erhalten genau wie eben

$$w_{ij} - \widehat{w} = f(x_i, y_j) - f(\hat{x}, \hat{y}) \approx u_i \cdot f_x(\hat{x}, \hat{y}) + v_j \cdot f_y(\hat{x}, \hat{y})$$

mit Quadrat

$$\begin{aligned} (w_{ij} - \widehat{w})^2 &\approx \left(u_i \cdot f_x(\hat{x}, \hat{y}) + v_j \cdot f_y(\hat{x}, \hat{y})\right)^2 \\ &= u_i^2 \cdot f_x(\hat{x}, \hat{y})^2 + v_j^2 \cdot f_y(\hat{x}, \hat{y})^2 + 2u_i \cdot v_j \cdot f_x(\hat{x}, \hat{y}) \cdot f_y(\hat{x}, \hat{y})\,. \end{aligned}$$

Hier sind die Werte u_i^2, v_j^2 und $u_i v_j$ jeweils Zufallsgrößen, über deren Werte wir nichts sagen können. Wir haben aber gewisse Erwartungen darüber, wie sie sich *im Mittel* verhalten: u_i^2 sollte, da σ_x^2 die mittlere quadratische Abweichung von $\hat{x}$ ist, im Mittel gleich σ_x^2 sein und v_j^2 entsprechend σ_y^2. Genauso sollten u_i und v_j im Mittel gleich null sein, und wenn wir annehmen, daß die Fehler u_i und v_j voneinander unabhängig sind, sollte auch ihr Produkt im Mittel verschwinden. Diese sogenannten *Erwartungswerte* sind offensichtlich die bestmöglichen Schätzwerte für die jeweiligen Größen; als beste Schätzung für σ_w^2 erhalten wir damit das GAUSS*sche Fehlerfortpflanzungsgesetz*

$$\sigma_w^2 = f_x(\hat{x}, \hat{y})^2 \cdot \sigma_x^2 + f_y(\hat{x}, \hat{y})^2 \cdot \sigma_y^2$$

oder

$$\sigma_w = \sqrt{f_x(\hat{x}, \hat{y})^2 \cdot \sigma_x^2 + f_y(\hat{x}, \hat{y})^2 \cdot \sigma_y^2}\,.$$

Genauso gilt dieses Gesetz auch für Funktionen von mehr als zwei Größen; für $w = f(x_1, \ldots, x_n)$ ist

$$\sigma_w = \sqrt{f_{x_1}(\hat{x}_1, \ldots \hat{x}_n)^2 \cdot \sigma_{x_1}^2 + \cdots + f_{x_n}(\hat{x}_1, \ldots \hat{x}_n)^2 \cdot \sigma_{x_n}^2}\,.$$

d) Die Standardabweichung des Mittelwerts und die Schätzung der Varianz

Als einfache Anwendung des Fehlerfortpflanzungsgesetzes betrachten wir die Funktion

$$\bar{x} = f(x_1, \ldots, x_N) = \frac{x_1 + \cdots + x_N}{N},$$

also den Mittelwert der x_i. Jede Messung x_i sei mit demselben erwarteten Fehler σ behaftet; da alle partiellen Ableitungen von f gleich $1/N$ sind, folgt für den Fehler des Mittelwerts

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}.$$

Dies bestätigt die implizit stets angewandte Regel, daß man durch mehrfaches Messen ein zuverlässigeres Ergebnis erhält; durch 25 Messungen beispielsweise läßt sich der Fehler auf ein Fünftel reduzieren, und für $N \to \infty$ geht er gegen Null *(Gesetz der großen Zahl).*

Damit wissen wir, wie man aus den Meßwerten auf den Fehler des Mittelwerts schließen kann – sofern man die Fehler der Meßwerte kennt. Wie lassen sich diese schätzen?

Zunächst ist

$$\begin{aligned}\sigma^2 &= \frac{1}{N}\sum_{i=1}^{N}(\hat{x} - x_i)^2 = \frac{1}{N}\sum_{i=1}^{N}\big((\hat{x} - \bar{x}) + (\bar{x} - x_i)\big)^2 \\ &= \frac{1}{N}\sum_{i=1}^{N}(\hat{x} - \bar{x})^2 + \frac{1}{N}\sum_{i=1}^{N}(\bar{x} - x_i)^2 + \frac{2}{N}\sum_{i=1}^{N}(\hat{x} - \bar{x})\cdot(\bar{x} - x_i).\end{aligned}$$

Die letzte dieser drei Summen ist

$$2 \cdot \frac{(\hat{x} - \bar{x})}{N}\sum_{i=1}^{N}(\bar{x} - x_i) = 0,$$

da $\bar{x}$ der Mittelwert der x_i ist. Die zweite Summe ist der Mittelwert der $(\bar{x} - x_i)^2$, also die Varianz der Meßreihe, und von der ersten schließlich wissen wir, daß $(\hat{x} - \bar{x})^2$, das Quadrat des Fehlers des Mittelwerts, den Erwartungswert σ^2/N hat. Die gesamte erste Summe ist somit

$$\frac{1}{N} \cdot N \cdot \frac{\sigma^2}{N} = \frac{\sigma^2}{N},$$

und obige Formel wird zu

$$\sigma^2 = \frac{\sigma^2}{N} + \frac{1}{N}\sum_{i=1}^{N}(\bar{x} - x_i)^2.$$

Bringt man hier noch den Term σ^2/N auf die linke Seite, so folgt

$$\frac{N-1}{N} \cdot \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(\bar{x} - x_i)^2$$

oder

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\bar{x} - x_i)^2,$$

also

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(\bar{x} - x_i)^2}{N-1}}.$$

Somit läßt sich auch σ aus den Meßdaten berechnen, der Meßfehler kann also ohne Kenntnis des „wahren“ Werts anhand der gemessenen Werte geschätzt werden.

§5: Zufallsvariablen und ihre Verteilungen

Zufallsvariablen sind, anschaulich ausgedrückt, Funktionen, die bei jeder Anwendung einen zufälligen Wert liefern derart, daß die Verteilung dieser Werte gewissen Gesetzmäßigkeiten unterliegt.

Eine exakte Definition müßte mit Wahrscheinlichkeitsräumen und meßbaren Funktionen arbeiten; dafür fehlt im Rahmen dieser Vorlesung erstens die Zeit, und zweitens wäre der dafür notwendige Aufwand bei den wenigen hier betrachteten anwendungen auch übertrieben.

Wir begnügen uns daher im nächsten Abschnitt mit nicht wirklich präzisen *ad hoc* Definitionen der beiden wichtigsten Arten von Zufallsvariablen, den diskreten mit endlichem Wertebereich und den kontinuierlichen mit stetiger Verteilungsfunktion.

a) Zufallsvariablen

Definition: Eine *diskrete Zufallsvariablen* ist ein Prozeß, der zufällig einen Wert aus einer vorgegebenen endlichen Menge

$$\{x_0, \ldots, x_m\}$$

liefert; eine *kontinuierliche Zufallsvariablen* liefert entsprechend einen zufälligen Wert aus $\mathbb{R}$.

Dieser „Zufall“ muß natürlich, falls er mathematisch faßbar sein soll, irgendwelchen Regeln genügen.

Im diskreten Fall nehmen wir dazu an, daß für jeden der möglichen Werte x_i feststeht, mit welcher *Wahrscheinlichkeit* p_i er angenommen wird. Diese „Wahrscheinlichkeit“ definieren wir informell so, daß bei einer großen Anzahl m von Versuchen *ungefähr* $p_i m$-mal der Wert x_i geliefert wird. Das „Gesetz der großen Zahlen“, das wir aus dem letzten Paragraphen kennen, sagt, daß diese Definition sinnvoll ist und man die Wahrscheinlichkeiten p_i damit in wohldefinierter Weise mit beliebiger Genauigkeit bestimmen kann.

Für eine Zufallsvariable, die kontinuierliche Werte annimmt, ist die Wahrscheinlichkeit dafür, daß ein konkreter Wert angenommen wird, praktisch immer gleich Null; hier können wir sinnvollerweise nur fragen, mit welcher Wahrscheinlichkeit die Werte in einem gegebenen Intervall $[a, b]$ liegen. Wie sagen, die Zufallsvariable X habe die *Wahrscheinlichkeitsdichte* f, wenn diese Wahrscheinlichkeit gleich

$$\int_a^b f(x)\,dx$$

ist.

Zwei formale Konsequenzen der Definition sind offensichtlich: Im diskreten Fall ist

$$0 \le p_i \le 1 \quad \text{für alle } i \text{ und} \quad \sum_{i=0}^{m} p_i = 1\,,$$

denn bei m Versuchen muß die Anzahl der auftretenden x_i für jedes i zwischen null und m liegen, und die Summe aller dieser Anzahlen ist m. Im kontinuierliche Fall ist entsprechend

$$f(x) \ge 0 \quad \text{für alle } x \in \mathbb{R} \quad \text{und} \quad \int_{-\infty}^{\infty} f(x)\,dx = 1\,.$$

(Da $f(x)$ eine Wahrscheinlichkeits*dichte* ist, können wir dafür keine allgemein gültige obere Schranke angeben.)

b) Statistische Kenngrößen von Zufallsvariablen

Statistische Kenngrößen sind Zahlen, die Informationen über Zufallsvariablen liefern. Die wichtigste davon ist der *Erwartungswert;* anschaulich betrachtet ist das der erwartete Durchschnitt aus einer großen Anzahl von Werten.

Definition: Der *Erwartungswert* $\mathbb{E}(X)$ einer diskreten Zufallsvariablen X ist

$$\mathbb{E}(X) = \sum_{i=0}^{m} p_i x_i\,;$$

der einer kontinuierlichen Zufallsvariablen ist

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x)\,dx\,.$$

Manche Zufallsvariablen nehmen nur Werte an, die in der Nähe dieses Erwartungswerts liegen; bei anderen streuen die Werte in einem weiten Bereich. Ein erstes Maß dafür ist die Varianz, die wir analog zur mittleren quadratischen Abweichung bei Meßreihen als Erwartungswert der quadratischen Abweichung definieren:

Definition: Die *Varianz* einer Zufallsvariablen X mit Erwartungswert $\mathbb{E}(X)$ ist $\sigma_X^2 = \mathbb{E}\big((X - \mathbb{E}(X))^2\big)$; im diskreten Fall ist also

$$\sigma_X^2 = \sum_{i=0}^{m} p_i \big(x_i - \mathbb{E}(X)\big)^2 = \sum_{i=0}^{m} p_i x_i^2 - \mathbb{E}(X)^2\,;$$

im kontinuierlichen Fall ist

$$\sigma_X^2 = \int_{-\infty}^{\infty} \big(x - \mathbb{E}(X)\big)^2 \cdot f(x)\,dx = \int_{-\infty}^{\infty} x^2 f(x)\,dx - \mathbb{E}(X)^2 .$$

Die Quadratwurzel $\sigma_X = \sqrt{\sigma_X^2}$ heißt *Standardabweichung* von X.

Als erstes Beispiel wollen wir eine diskrete Zufallsvariable betrachten, die das Würfeln beschreibt. Bei einem idealen Würfel wird jede Augenzahl i mit derselben Wahrscheinlichkeit $p_i = \frac{1}{6}$ angenommen; der Erwartungswert ist also

$$\mathbb{E}(X) = \frac{1+2+3+4+5+6}{6} = 3\frac{1}{2} .$$

Damit läßt sich auch die Varianz einfach berechnen:

$$\sigma_X^2 = \frac{\left(-2\frac{1}{2}\right)^2 + \left(-1\frac{1}{2}\right)^2 + \left(-\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 + \left(1\frac{1}{2}\right)^2 + \left(2\frac{1}{2}\right)^2}{6} = \frac{35}{12}$$

und

$$\sigma_X = \sqrt{\frac{35}{12}} \approx 1{,}7078 .$$

§6: Erste Beispiele von Verteilungen

a) Die Gleichverteilung

Das einfachste Beispiel einer kontinuierlichen Verteilung ist die *Gleichverteilung:* Hier kann die Zufallsvariable X nur Werte x annehmen, die zwischen zwei vorgegebenen Werten a und b liegen, und jeder dieser Werte ist gleich wahrscheinlich, d.h., exakt ausgedrückt, die *Wahrscheinlichkeitsdichte* dieser Verteilung ist gleich einer Konstanten γ im Intervall zwischen a und b und ist null außerhalb dieses Intervalls. Da

$$p(a \le x \le b) = \int_a^b f(x)\,dx = \int_a^b \gamma\,dx = \gamma \cdot (b-a) = 1$$

sein muß, ist also

$$f(x) = \gamma = \frac{1}{b-a} \quad \text{für } a \le x \le b .$$

Der *Erwartungswert* einer gleichverteilten Zufallsvariablen ist

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x \cdot f(x)\,dx = \int_a^b x \cdot f(x)\,dx \\ &= \int_a^b x \cdot \gamma\,dx = \gamma \int_a^b x\,dx \\ &= \gamma \cdot \left(\frac{1}{2}b^2 - \frac{1}{2}a^2\right) = \frac{\frac{1}{2}b^2 - \frac{1}{2}a^2}{b-a} \\ &= \frac{a+b}{2} ; \end{aligned}$$

wie zu erwarten war, ist es also der Mittelpunkt des Intervalls.

Die Varianz errechnet sich demnach als

$$\sigma^2 = \int_a^b \gamma \cdot \left(x - \frac{a+b}{2}\right)^2 dx ;$$

da $\frac{1}{3}(x-c)^3$ eine Stammfunktion von $(x-c)^2$ ist, folgt

$$\begin{aligned} \sigma^2 &= \frac{\gamma}{3} \cdot \left(\left(b - \frac{a+b}{2}\right)^3 - \left(a - \frac{a+b}{2}\right)^3\right) \\ &= \frac{\gamma}{3} \cdot \left(\left(\frac{b-a}{2}\right)^3 - \left(\frac{a-b}{2}\right)^3\right) \\ &= \frac{\gamma}{3 \cdot 8} \cdot 2 \cdot (b-a)^3 = \frac{2 \cdot (b-a)^3}{3 \cdot 8 \cdot (b-a)} \\ &= \frac{a^2 + ab + b^2}{12} . \end{aligned}$$

b) Die Binomialverteilung

Als nächstes Beispiel einer konkreten Verteilung wollen wir eine Verteilung für *diskrete* Zufallsvariablen betrachten, die einerseits für sich

selbst interessant ist, andererseits aber auch zwei der wichtigsten kontinuierlichen Verteilungen als Grenzfälle liefert.

Ausgangspunkt ist das n-malige Werfen einer Münze; diese falle jeweils mit Wahrscheinlichkeit p so, daß *Kopf* oben liegt und dementsprechend mit Wahrscheinlichkeit $q = 1 - p$ so, daß die *Zahl* zu sehen ist. Wir wollen die Wahrscheinlichkeit dafür berechnen, daß beim n-maligen Werfen k-mal *Kopf* und somit $(n - k)$-mal *Zahl* erscheint.

Anstelle des Münzwurfs kann man sich natürlich genausogut jedes andere Ereignis vorstellen, das genau zwei mögliche Ausgänge hat; lediglich der Anschaulichkeit halber soll vorläufig vom Werfen einer Münze die Rede sei – auch wenn für ein stark von $\frac{1}{2}$ abweichendes p die Anschaulichkeit vielleicht nicht allzu groß erscheint.

Wir definieren eine Zufallsvariable X durch

$$X = \text{Anzahl der Würfe mit Ergebnis } \textit{Kopf}$$

und suchen die Wahrscheinlichkeiten

$$p_k \underset{\text{def}}{=} p(X = k)$$

für $k = 0, \ldots, n$. Die Verteilung dieser Zufallsvariablen heißt *Binomialverteilung* oder BERNOULLI-*Verteilung* mit Parametern n und p. (Der Parameter $q = 1 - p$ muß nicht eigens erwähnt werden, da er durch p eindeutig bestimmt ist.)

Im Fall $n = 1$ ist alles klar: Die Wahrscheinlichkeit für *Kopf* ist p, die für *Zahl* entsprechend $q = 1 - p$, d.h.

$$p_0 = q \quad \text{und} \quad p_1 = p\,.$$

Auch für $n > 1$ wird den meisten klar sein, wie man die Wahrscheinlichkeiten berechnet: Es gibt $\binom{n}{k}$ Möglichkeiten, aus den n Versuchen diejenigen k auszuwählen, die das Ergebnis *Kopf* haben, und für jede einzelne dieser Möglichkeiten haben wir die Wahrscheinlichkeit $p^k q^{n-k}$ für k-mal *Kopf* und $(n - k)$-mal *Zahl;* somit ist

$$p_k = \binom{n}{k} p^k q^{n-k}\,.$$

Für diejenigen, die das nicht aus der Schule wissen, sei diese Formel kurz hergeleitet.

Für $n = 2$ gibt es drei mögliche Ausgänge des Zufallsexperiments: zweimal *Kopf,* zweimal *Zahl* oder je einmal *Kopf* und *Zahl.* Zweimal *Kopf* ist nur möglich, wenn beim ersten wie beim zweiten Wurf *Kopf* oben liegt. Beide Ereignisse haben, jeweils für sich betrachtet, die Wahrscheinlichkeit p; da sie voneinander unabhängig sind, ist die Wahrscheinlichkeit ihres gemeinsamen Auftretens gleich dem Produkt der beiden Einzelwahrsheinlichkeiten, also $p \cdot p = p^2$. Entsprechend ist die Wahrscheinlichkeit für zweimal *Zahl* natürlich q^2.

Das Ereignis „einmal *Kopf* und einmal *Zahl*" kann auf zweierlei Weise zustande kommen: Entweder fällt beim ersten Wurf *Kopf* und beim zweiten Wurf *Zahl,* oder umgekehrt. Die Wahrscheinlichkeit für die erste Möglichkeit berechnet sich analog zu oben als $p \cdot q$, die für die zweite als $q \cdot p$, was natürlich genau der gleiche Wert ist. Da beide Möglichkeiten offensichtlich nie gleichzeitig auftreten können, ist die Wahrscheinlichkeit für das Auftreten von einer der beiden gerade die Summe der beiden Einzelwahrscheinlichkeiten, also $2pq$. Somit ist hier

$$p_0 = q^2,\ p_1 = 2pq \quad \text{und} \quad p_2 = p^2\,.$$

Die Wahrscheinlichkeiten für die drei möglichen Ausgänge des Zufallsexperiments addieren sich, wie es sich gehört, zu *eins,* denn

$$p^2 + q^2 + 2pq = (p + q)^2 = 1\,,$$

da sich p und q zu eins ergänzen.

Bei *dreimaligem* Werfen der Münze gibt es *vier* Möglichkeiten: Dreimal *Kopf*, dreimal *Zahl* sowie zweimal *Kopf* und einmal *Zahl* oder einmal *Kopf* und zweimal *Zahl.*

Anstatt diese Wahrscheinlichkeiten wieder einzeln zu berechnen, können wir vom Fall der zwei Würfe ausgehen und für jeden möglichen Ausgang den Effekt des dritten Wurfs betrachten: Für das Ereignis „zweimal *Kopf* und einmal *Zahl*" müssen dann nur zwei Fälle betrachtet werden: Entweder war der dritte Wurf *Zahl* und damit die ersten beiden Würfe „zweimal *Kopf*", oder der dritte Wurf lieferte *Kopf,* und die ersten beiden hatten das Ergebnis „einmal *Kopf* und einmal *Zahl*". Die Wahrscheinlichkeit ist also

$$q \cdot p^2 + p \cdot pq = 3p^2 q\,.$$

Entsprechend ist die Wahrscheinlichkeit des Ereignisses „zweimal *Zahl* und einmal *Kopf*" gleich $3pq^2$, und die Wahrscheinlichkeiten von „dreimal *Kopf*" beziehungsweise „dreimal *Zahl*" sind natürlich p^3 beziehungsweise q^3. Somit ist

$$p_0 = q^3,\ p_1 = 3pq^2,\ p_2 = 3p^2 q \quad \text{und} \quad p_3 = p^3\,,$$

und wieder ist die Summe aller Wahrscheinlichkeiten eins, denn

$$1 = (p + q)^3 = p^3 + 3p^2 q + 3pq^2 + q^3\,.$$

Entsprechend können wir argumentieren bei einer beliebigen Anzahl n von Würfen: X kann die $n + 1$ Werte $0 \le k \le n$ annehmen. Wie oben setzt sich das Ereignis $X = k$ zusammen aus verschiedenen Einzelereignissen wie etwa „zuerst k mal *Kopf,* dann $(n - k)$

mal Zahl“ und entsprechend den Ereignissen, bei denen k andere Wurfnummern für *Kopf* verlangt werden. Jedes dieser Ereignisse hat die Wahrscheinlichkeit $p^k q^{n-k}$, so daß p_k ein ganzzahliges Vielfaches von $p^k q^{n-k}$ ist. Wir schreiben dieses Vielfache als

$$p_k = \binom{n}{k} p^k q^{n-k}$$

und nennen $\binom{n}{k}$ den *Binomialkoeffizienten „n über k“*.

Zu seiner Berechnung können wir vorgehen wie im Fall $n = 3$, indem wir alles auf das Werfen von $n - 1$ Münzen zurückführen und somit versuchen, $\binom{n}{k}$ durch geeignete Binomialkoeffizienten der Form $\binom{n-1}{j}$ auszudrücken:

Das Ereignis $X = k$ kann zustandekommen

- *entweder* dadurch, daß bei den ersten $n - 1$ Würfen schon k mal *Kopf* gefallen ist und daß dann beim n-ten Wurf *Zahl* fällt
- *oder* dadurch, daß bei den ersten $n - 1$ Würfen nur $(k - 1)$ mal *Kopf* gefallen ist und daß dann beim n-ten Wurf noch einmal *Kopf* fällt. Definieren wir also eine neue Zufallsvariable Y dadurch, daß Y die Häufigkeit des Ereignisses *Kopf* bei den ersten $n - 1$ Würfen zählt, und eine Zufallsvariable Z als 0 oder 1, je nachdem, ob im n-ten Wurf *Zahl* oder *Kopf* fällt, so ist

$$\begin{aligned} p(X = k) &= p(Y = k-1) \cdot p(Z = 1) + p(Y = k) \cdot p(Z = 0) \\ &= \binom{n-1}{k} p^k q^{n-1-k} \cdot q \\ &\quad + \binom{n-1}{k-1} p^{k-1} q^{(n-1)-(k-1)} \cdot p \\ &= \binom{n-1}{k} p^k q^{n-k} + \binom{n-1}{k-1} p^k q^{n-k} . \end{aligned}$$

Andererseits ist

$$p(X = k) = \binom{n}{k} p^k q^{n-k} ,$$

also erhalten wir die Beziehung

$$\binom{n}{k} = \binom{n-k}{k-1} + \binom{n-1}{k} ,$$

aus der sich $\binom{n}{k}$ rekursiv berechnen läßt.

Dies geschieht am besten über das PASCALsche Zahlendreieck: Wenn wir die Binomialkoeffizienten in der Form

$$\begin{array}{lccccccc} n = 0 & & & & \binom{0}{0} & & & \\ n = 1 & & & \binom{1}{0} & & \binom{1}{1} & & \\ n = 2 & & \binom{2}{0} & & \binom{2}{1} & & \binom{2}{2} & \\ n = 3 & \binom{3}{0} & & \binom{3}{1} & & \binom{3}{2} & & \binom{3}{3} \end{array}$$

$$\begin{array}{lccccccccccc} n = 4 & & \binom{4}{0} & & \binom{4}{1} & & \binom{4}{2} & & \binom{4}{3} & & \binom{4}{4} & \\ n = 5 & \binom{5}{0} & & \binom{5}{1} & & \binom{5}{2} & & \binom{5}{3} & & \binom{5}{4} & & \binom{5}{5} \end{array}$$

anordnen, besagt die gerade hergeleitete Beziehung, daß jeder Binomialkoeffizient die *Summe* der beiden rechts und links über ihm stehenden ist. Außen muß offenbar immer 1 stehen, denn die Wahrscheinlichkeit für 0 mal *Kopf,* d.h. n mal *Zahl,* ist $1 \cdot q^n$, und die für n mal *Kopf* ist $1 \cdot p^n$. Damit läßt sich das Dreieck leicht hinschreiben:

$$\begin{array}{lccccccccccc} n = 0 & & & & & & 0 & & & & & \\ n = 1 & & & & & 1 & & 1 & & & & \\ n = 2 & & & & 1 & & 2 & & 1 & & & \\ n = 3 & & & 1 & & 3 & & 3 & & 1 & & \\ n = 4 & & 1 & & 4 & & 6 & & 4 & & 1 & \\ n = 5 & 1 & & 5 & & 10 & & 10 & & 5 & & 1 \end{array}$$

Für kleine Werte von n läßt sich $\binom{n}{k}$ auf diese Weise einfach bestimmen; für großes n ist dieses Verfahren allerdings zu umständlich. Hier verwendet man besser die explizite Formel

$$\binom{n}{k} = \frac{n!}{k! \cdot (n-k)!} ,$$

in der $n!$ (gesprochen n *Fakultät*) für das Produkt

$$n! = 1 \cdot 2 \cdot 3 \cdot \cdots \cdot (n-1) \cdot n$$

der Zahlen von 1 bis n steht und verabredungsgemäß das „leere“ Produkt $0! = 1$ sein soll. (Wer sich noch an seine Schulmathematik erinnert, kann diese Formel leicht mit elementarer Bruchrechnung durch Induktion nach n beweisen.) Für das praktische Rechnen ist es oft von Vorteil, durch $(n - k)!$ zu kürzen; dadurch wird obige Formel zu

$$\binom{n}{k} = \frac{n \cdot (n-1) \cdots (n-k+1)}{k!} .$$

Für große Werte von n wird die Binomialverteilung nur selten benutzt, da man sie dann meist durch handlichere Verteilungen annähern kann. Trotzdem sei der Vollständigkeit halber auf die in §7*c)* bewiesene STIRLINGsche Formel verwiesen, über die man $n!$ für große Werte von n näherungsweise berechnen kann:

$$n! \approx \left(\frac{n}{e}\right)^n \sqrt{2\pi n} ,$$

wobei

$$e \approx 2{,}71828\,18284\,59045\,23536$$

die EULERsche Zahl ist und π die Kreiszahl. Der Schätzwert ist stets zu klein mit einem Fehler von höchstens $1/12n$.

Die Berechnung von Erwartungswert und Standardabweichung unmittelbar aus der Definition ist wegen der dazu auszuwertenden Summen mit vielen Binomialkoeffizienten eher unangenehm; glücklicherweise kann man darauf aber verzichten und alles auf den Fall $n = 1$ zurückführen: Für $n = 1$ ist der Erwartungswert

$$E(X) = p \cdot 1 + 1 \cdot 0 = p$$

und die Varianz

$$\begin{aligned} E\big((X-p)^2\big) &= p \cdot (1-p)^2 + q \cdot (0-p)^2 \\ &= p \cdot q^2 + q \cdot p^2 = pq(q+p) = pq\,, \end{aligned}$$

die *Standardabweichung* ist also $\sigma = \sqrt{pq}$.

Um daraus die entsprechenden Werte für beliebiges n zu berechnen, können wir uns zunutze machen, daß die verschiedenen Würfe der Münze voneinander *unabhängig* sind, so daß sich die Erwartungswerte einfach addieren, d.h.

$$E(X) = n \cdot p \quad \text{bei } n \text{ Würfen.}$$

Nicht ganz so klar ist, daß sich auch die Varianzen addieren; hierzu müssen wir uns an die Rechnung erinnern, in der wir den Fehler des arithmetischen Mittels aus n Werten einer Zufallsvariablen berechnet haben, indem wir aus dem Verschwinden der Erwartungswerte der gemischten Produkte genau dies geschlossen haben. Somit ist die Varianz im Falle von n Würfen gleich npq und die Standardabweichung entsprechend $\sqrt{npq}$.

c) Die Poisson-Verteilung

Auch hier handelt es sich um eine diskrete Verteilung, nämlich den Grenzfall der Binomialverteilung für großes n und kleines p, wobei der Erwartungswert $\lambda = np$ konstant gehalten wird. Dieses Produkt λ ist demgemäß der einzige Parameter der POISSON-Verteilung.

In Alltagssprache übersetzt bedeuten großes n und kleines p, daß wir die Häufigkeit eines seltenen Ereignisses betrachten; insbesondere interessieren also nur die kleinen Werte von k, da die großen ohnehin extrem unwahrscheinlich sind.

Für die Binomialverteilung mit Parametern n und p ist

$$\begin{aligned} p_k &= \binom{n}{k} p^k q^{n-k} = \frac{n \cdot (n-1) \cdot \cdots \cdot (n-k+1)}{k!} p^k q^{n-k} \\ &= \frac{n \cdot (n-1) \cdot \cdots \cdot (n-k+1)}{k!} \cdot \left(\frac{\lambda}{n}\right)^k \cdot (1-p)^{n-k} \\ &= \frac{n \cdot (n-1) \cdot \cdots \cdot (n-k+1)}{n^k} \cdot \frac{\lambda^k}{k!} \cdot \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{n}{n} \cdot \frac{n-1}{n} \cdot \cdots \cdot \frac{n-k}{n} \cdot \frac{\lambda^k}{k!} \cdot \frac{(1-\frac{\lambda}{n})^n}{\left(1-\frac{\lambda}{n}\right)^k}\,, \end{aligned}$$

da $q = 1 - p$ und $p = \lambda/n$ ist. Wie aus der Schule bekannt (sein sollte), ist

$$\lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^n = e^{-\lambda}\,,$$

und da k klein, n aber groß sein soll, machen wir keinen großen Fehler, wenn wir

$$\frac{n}{n} \approx \frac{n-1}{n} \approx \cdots \approx \frac{n-k}{n} \approx 1 \quad \text{und} \quad \left(1 - \frac{\lambda}{n}\right)^k \approx 1$$

setzen. Damit ist

$$p_k \approx \frac{e^{-\lambda} \cdot \lambda^k}{k!}\,.$$

Führt man den Grenzübergang explizit aus, so erhält man die Wahrscheinlichkeiten

$$p_k = \frac{e^{-\lambda} \cdot \lambda^k}{k!}\,;$$

wir bezeichnen eine Zufallsvariable demnach als POISSON-verteilt, wenn sie diese Wahrscheinlichkeiten hat.

Die Kennzahlen der POISSON-Verteilung bestimmt man am einfachsten über die der Binomialverteilung: Da die POISSON-Verteilung Grenzwert von Binomialverteilungen mit Erwartungswert $\lambda = np$ ist, hat sie

natürlich auch selbst den Erwartungswert λ. Die *Varianz* einer Binomialverteilung mit Parametern n und p ist

$$n \cdot p \cdot q = n \cdot p \cdot (1-p) = np - np^2 = \lambda - p \cdot \lambda\,.$$

Da die POISSON-Verteilung der Grenzwert für $p \to 0$ ist, hat sie also die Varianz λ und damit die Standardabweichung $\sqrt{\lambda}$.

Praktisch kann man davon ausgehen, daß man für Werte von n ab etwa $n = 100$ und für Wahrscheinlichkeiten bis etwa 5%, d.h. $p \leq 0{,}05$, die Binomialverteilung durch die erheblich einfacher zu berechnende POISSON-Verteilung ersetzen kann, ohne einen nennenswerten Fehler zu machen. Auch für etwas kleinere Werte von n ist die Übereinstimmung im allgemeinen schon recht gut.

Beispiele für POISSON-verteilte Zufallsvariablen gibt es viele: Häufigkeiten von Naturkatastrophen wie Erdbeben, Überschwemmungen *usw.,* die Anzahl radioaktiver Zerfälle pro Minute eines schwach radioaktiven Materials wie etwa C^{14} oder Co^{60} und viele mehr.

Da die POISSON-Verteilung nur einen Parameter hat, ist ihre Standardabweichung durch ihren Erwartungswert bestimmt; in der Tat ist sie einfach die Quadratwurzel davon. Daher steigt die Standardabweichung mit zunehmendem λ nur sehr viel langsamer als der Erwartungswert. Interpretiert man die Standardabweichung als Abweichung vom Mittelwert, so wird daher die *relative* Abweichung für große λ immer geringer und für sehr große λ praktisch vernachlässigbar. Dies ist einer der Gründe dafür, daß zahlreiche Vorgänge, die auf der mikroskopischen Ebene eigentlich nur statistisch beschrieben werden können, auf der makroskopischen Ebene im Rahmen der erzielbaren Meßgenauigkeit ein deterministisches Verhalten zeigen.

Ein zwar etwas exotisches, wegen der ungewöhnlich guten Übereinstimmung von Theorie und Praxis aber häufig in Statistiklehrbüchern anzutreffendes Beispiel für POISSON-verteilte Daten sind die Anzahlen der jährlich durch Hufschlag getöteten Offiziere preußischer Kavallerieregimenter: Eine Untersuchung von zehn Regimentern über zwanzig Jahre hinweg ergab insgesamt 122 Todesfälle, also im Mittel

$$\lambda = \frac{122}{10 \cdot 20} = 0{,}61$$

pro Jahr und Regiment.

Die Wahrschenlichkeit dafür, daß es in einem Regiment genau k Todesfälle in einem Jahr gab, sollte damit unter der Annahme einer POISSON-Verteilung mit dem Parameter (= Erwartungswert) 0,61 ungefähr gleich

$$\frac{e^{-0{,}61} \cdot 0{,}61^k}{k!} \approx \frac{0{,}331444 \cdot 0{,}61^k}{k!}$$

sein; wir erwarten also, daß dies in etwa

$$200 \cdot \frac{0{,}331444 \cdot 0{,}61^k}{k!} = \frac{66{,}2888 \cdot 0{,}61^k}{k!}$$

der zweihundert Fälle vorkommt. Die folgende Tabelle zeigt die tatsächlichen und die (gerundeten) berechneten Werte:

k	*tatsächliche Fallzahl*	*berechnete Fallzahl*
0	109	108,67
1	65	66,29
2	22	20,22
3	3	4,11
4	1	0,63
≥ 5	0	0,08

Wie man sieht, ist die Übereinstimmung in der Tat erstaunlich gut.

§7: Die Normalverteilung

a) Die Normalverteilung als Grenzfall der Binomialverteilung

Hinter dem LAPLACEschen Fehlermodell steckt offenbar eine Bionomialverteilung mit Parametern $p = \frac{1}{2}$ und n = Anzahl der „Dämonen“. Wenn wir hier einen vernünftigen Grenzübergang $n \to \infty$ machen wollen, müssen wir den Parameter ε offensichtlich so von n abhängen lassen, daß der Erwartungswert und die Varianz des Fehlers von n unabhängig sind.

Beim Erwartungswert ist das kein Problem: Er ist ohnehin immer null.

Zur Berechnung der Varianz müssen wir die Fehlerquadrate über die 2^n möglichen Verhaltensweisen der „Dämonen“ summieren, d.h. wir müssen

$$\sigma^2 = \frac{1}{2^n} \sum (\varepsilon_1 + \ldots + \varepsilon_n)^2$$

berechnen, wobei sich die Summation über alle n-tupel

$$(\varepsilon_1, \cdots, \varepsilon_n) \quad \text{mit} \quad \varepsilon_i = \pm\varepsilon$$

erstreckt. Beim Ausmultiplizieren heben sich alle gemischten Terme der Form $\varepsilon_i \varepsilon_j$ gegenseitig weg, denn das Tupel, bei dem nur an der i-ten Stelle das Vorzeichen geändert wurde, liefert einen Summanden $-\varepsilon_i \varepsilon_j$. Also bleiben nur die Quadrate; diese sind alle gleich ε^2, und es sind pro Summand n Stück. Da die Anzahl der Summanden gleich dem Nenner des Vorfaktors ist, berechnet sich die Varianz daher zu

$$\sigma^2 = n\varepsilon^2$$

Somit müssen wir für $n \to \infty$ den Einfluß ε jedes einzelnen „Dämonen“ so gegen null gehen lassen, daß $n\varepsilon^2$ konstant bleibt, d.h. wir setzen

$$\varepsilon = \frac{\sigma}{\sqrt{n}}$$

für eine geeignet zu wählende Konstante $\sigma > 0$, die Standardabweichung.

Für festes n kann der Fehler einen der $n + 1$ Werte

$$-n\varepsilon, \quad -(n-2)\varepsilon, \quad \ldots, \quad (n-2)\varepsilon, \quad n\varepsilon$$

annehmen, was wir in der Form

$$u = (n - 2k)\varepsilon = \frac{(n-2k)\sigma}{\sqrt{n}} \quad \text{mit} \quad k = -n, \ldots, n$$

schreiben wollen. Dieser Fehler tritt genau dann auf, wenn k der Dämonen den Fehler ε erzeugen und die restlichen $n - k$ den Fehler $-\varepsilon$. Dies geschieht in $\binom{n}{k}$ der 2^n möglichen Fälle; die Wahrscheinlichkeit dafür ist also $\binom{n}{k} 2^{-n}$.

Für $n \to \infty$ geht dieser Ausdruck gegen null, denn mit n geht schließlich auch die Anzahl der zu betrachtenden Fälle gegen unendlich. Falls es ein Intervall gäbe, in dem die Wahrscheinlichkeit für jeden darin liegenden Fehler größer als irgendein $\alpha > 0$ wäre, ginge allein schon die Summe der Wahrscheinlichkeiten für Fehler aus diesem Teilintervall mit n gegen unendlich, da die Anzahl der dort liegenden möglichen u-Werte wegen der $\sqrt{n}$ im Nenner von u gegen unendlich geht. Da die Summe aller Wahrscheinlichkeiten aber nicht größer als eins werden kann, muß die Wahrscheinlichkeit also in jedem einzelnen Punkt für $n \to \infty$ gegen null gehen.

Wenn wir n variieren lassen, ist es allerdings ohnehin sinnlos, einen genauen Wert des Fehlers zu betrachten: Für jedes n gibt es nur $n + 1$ mögliche Werte, und bei den meisten größeren Werten von n werden diese Zahlen – abgesehen von der Null – nicht auftreten. Wenn wir etwas von n unabhängiges definieren möchten (und sofern wir nicht an Dämonen glauben, bleibt uns kaum etwas anderes übrig), dürfen wir also nicht den genauen *Wert* des Fehlers festlegen, sondern müssen ein *Fehlerintervall* betrachten.

Nun wollen wir aber natürlich als Ergebnis keine Funktion, die von einem Intervall abhängt, sondern eine gewöhnliche Funktion von u, d.h. wir wollen, wie bei kontinuierlichen Verteilungen üblich, eine *Wahrscheinlichkeitsdichte* $\varphi(u)$.

Da die Wahrscheinlichkeit dafür, daß u im Intervall $[a, b]$ liegt, gleich dem Integral von f über dieses Intervall ist, ist für jeden Punkt u_0

$$\varphi(u_0) = \lim_{\varepsilon \to 0} \frac{\text{Wahrscheinlichkeit für } u_0 - \varepsilon \le u \le u_0 - \varepsilon}{2\varepsilon},$$

d.h. also als Wahrscheinlichkeit dividiert durch die Intervallbreite.

Für festes n haben die möglichen Fehlerwerte den Abstand 2ε, wir betrachten daher Intervalle der Länge 2ε mit den Werten $(n - 2k)\varepsilon$ als Mittelpunkten; diese überdecken den möglichen Fehlerbereich lückenlos, und die Wahrscheinlichkeit für einen Fehler in diesem Intervall ist $\binom{n}{k} 2^{-n}$.

Wir interessieren uns daher für $\lim\limits_{n \to \infty} \dfrac{\binom{n}{k} 2^{-n}}{2\varepsilon} = \dfrac{\binom{n}{k} 2^{-n}}{2\sigma/\sqrt{n}} = \dbinom{n}{k} 2^{-n-1} \dfrac{\sqrt{n}}{\sigma}$.

b) Die Eulersche Summenformel

Das Problem bei der Berechnung dieses Grenzwerts ist der Binomialkoeffizient

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}\,;$$

um diesen abzuschätzen brauchen wir einen handhabbaren Ausdruck für $n!$.

Dazu schreiben wir

$$\ln n! = \sum_{k=1}^{n} \ln k$$

und berechnen dies nach einer Methode von EULER, die nicht nur für Summen von Logarithmen anwendbar ist.

Wir betrachten allgemeiner irgendeine reellwertige differenzierbare Funktion f, deren Definitionsbereich das Intervall $[1,\ n]$ enthält.

Für eine reelle Zahl x bezeichnen wir wie üblich mit $[x]$ die größte ganze Zahl kleiner oder gleich x und mit $\{x\} \underset{\text{def}}{=} x-[x]$ den gebrochenen Anteil von x; ist k eine ganze Zahl, ist somit $\{x\} = x-k$ für $x \in [k,\ k+1)$.

Partielle Integration führt auf die Gleichung

$$\int_k^{k+1} \left(\{x\} - \tfrac{1}{2}\right) f'(x)\,dx = \left(x-k-\tfrac{1}{2}\right) f(x)\,\Big|_k^{k+1} - \int_k^{k+1} f(x)\,dx$$

$$= \frac{f(k+1)+f(k)}{2} - \int_k^{k+1} f(x)\,dx\,.$$

Addition aller solcher Gleichungen von $k=1$ bis $k=n-1$ liefert

$$\int_1^n \left(\{x\} - \tfrac{1}{2}\right) f'(x)\,dx = \frac{f(1)}{2} + \sum_{k=2}^{n-1} f(k) + \frac{f(n)}{2} - \int_1^n f(x)\,dx\,,$$

womit man die Summe der $f(k)$ berechnen kann:

Satz (EULERsche Summenformel): Für eine differenzierbare Funktion $f\colon D \to \mathbb{R}$, deren Definitionsbereich das Intervall $[1,\ n]$ umfaßt, ist

$$\sum_{k=1}^{n} f(k) = \int_1^n f(x)\,dx + \frac{f(1)+f(n)}{2} + \int_1^n \left(\{x\} - \tfrac{1}{2}\right) f'(x)\,dx\,.$$

■

Für die Abschätzung der Binomialkoeffizienten und Fakultäten interessiert uns speziell der Fall $f(x) = \ln x$; hierfür wird die EULERsche Summenformel zu

$$\ln n! = \int_1^n \ln x\,dx + \frac{\ln n}{2} + \int_1^n \frac{\{x\} - \frac{1}{2}}{x}\,dx$$

$$= x(\ln x - 1)\,\Big|_1^n + \frac{\ln n}{2} + \int_1^n \frac{\{x\} - \frac{1}{2}}{x}\,dx$$

$$= n(\ln n - 1) + 1 + \frac{\ln n}{2} + \int_1^n \frac{\{x\} - \frac{1}{2}}{x}\,dx\,.$$

In dieser Formel stört noch das rechte Integral; dieses können wir wie folgt abschätzen: Für eine natürliche Zahl k ist

$$\int_k^{k+1} \frac{\{x\} - \frac{1}{2}}{x}\,dx = \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{x}{k+\frac{1}{2}+x}\,dx$$

$$= \int_0^{\frac{1}{2}} \left(\frac{x}{k+\frac{1}{2}+x} - \frac{x}{k+\frac{1}{2}-x}\right) dx = \int_0^{\frac{1}{2}} \frac{-2x^2}{\left(k+\frac{1}{2}\right)^2 - x^2}\,dx\,.$$

Im Intervall von 0 bis $\frac{1}{2}$ ist der Integrand monoton fallend, d.h.

$$0 \geq \frac{-2x^2}{\left(k+\frac{1}{2}\right)^2 - x^2} \geq \frac{-\frac{1}{2}}{\left(k+\frac{1}{2}\right)^2 - \frac{1}{4}} = \frac{-2}{(2k+1)^2 - 1} \geq -\frac{1}{4k^2}\,,$$

und damit ist

$$0 \geq \int_k^{k+1} \frac{\{x\} - \frac{1}{2}}{x}\, dx = \int_0^{\frac{1}{2}} \frac{-2x^2}{\left(k + \frac{1}{2}\right)^2 - x^2}\, dx \geq -\frac{1}{8k^2}\,,$$

denn wir können das Integral abschätzen durch das Produkt aus der Länge des Integrationsintervalls und dem Minimum des Integranden. Summation von $k = 1$ bis $n - 1$ schließlich gibt die Abschätzung

$$0 \geq \int_1^n \frac{\{x\} - \frac{1}{2}}{x}\, dx \geq -\sum_{k=1}^{n-1} \frac{1}{4k^2}$$

für das störende Integral aus der obigen Formel.

Wie wohl jeder schon einmal in einer Analysis I Übungsaufgabe zeigen mußte, konvergiert die rechtsstehende Summe (egal ob mit oder ohne acht im Nenner) für $n \to \infty$; aus Kapitel III, §3*f)* wissen wir sogar, daß der Grenzwert $\pi^2/48$ ist. Auf jeden Fall können wir folgern, daß das uneigentliche Integral

$$\int_1^\infty \frac{\{x\} - \frac{1}{2}}{x}\, dx$$

konvergiert; den uns bislang noch unbekannten Grenzwert wollen wir mit I bezeichnen. Damit ist

$$\ln n! = n(\ln n - 1) + \frac{\ln n}{2} + C + o(1) \quad \text{mit} \quad C = I + 1$$

oder $n! \approx e^C \cdot n^n e^{-n} \sqrt{n}$.

Für Binomialkoeffizienten folgt, daß

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \approx \frac{e^C n^n e^{-n} \sqrt{n}}{e^C k^k e^{-k} \sqrt{k}\, e^C (n-k)^{n-k} e^{-(n-k)} \sqrt{(n-k)}}$$
$$= \frac{1}{e^C} \frac{n^n}{k^k\,(n-k)^{n-k}} \sqrt{\frac{n}{k \cdot (n-k)}}\,.$$

c) Die Stirlingsche Formel und die Normalverteilung

Ausgedrückt durch $u = (n - 2k)\varepsilon = (n - 2k)\sigma/\sqrt{n}$ ist

$$k = \frac{n}{2} - \frac{u\sqrt{n}}{2\sigma}\,,$$

und setzen wir zur Vereinfachung der Schreibweise

$$m = \frac{n}{2}, \quad v = \frac{u}{2\sigma} \quad \text{und} \quad \ell = \sqrt{2m} \cdot v\,,$$

so ergeben sich die Formeln

$$k = m - \sqrt{2m} \cdot v = m - \ell \quad \text{und} \quad n - k = m + \sqrt{2m} \cdot v = m + \ell\,.$$

Setzen wir dies alles in die Formel für die Wahrscheinlichkeitsdichte ein, erhalten wir

$$\binom{n}{k} 2^{-n-1} \frac{\sqrt{n}}{\sigma}$$
$$\approx \frac{1}{e^C} \frac{n^n}{k^k\,(n-k)^{n-k}} \sqrt{\frac{n}{k \cdot (n-k)}}\, 2^{-n-1} \frac{\sqrt{n}}{\sigma}$$
$$= \frac{1}{e^C \sigma} \frac{n^n \cdot 2^{-n}}{k^k\,(n-k)^{n-k}} \frac{n \cdot 2^{-1}}{\sqrt{k \cdot (n-k)}}$$
$$= \frac{1}{e^C \sigma} \frac{(2m)^{2m} \cdot 2^{-2m}}{(m-\ell)^{m-\ell}(m+\ell)^{m+\ell}} \frac{2m \cdot 2^{-1}}{\sqrt{(m-\ell)(m+\ell)}}$$
$$= \frac{1}{e^C \sigma} \frac{m^{2m}}{(m-\ell)^m (m+\ell)^m} \left(\frac{m-\ell}{m+\ell}\right)^\ell \frac{m}{\sqrt{m^2 - \ell^2}}$$
$$= \frac{1}{e^C \sigma} \frac{m^{2m}}{(m^2 - \ell^2)^m} \left(\frac{m-\ell}{m+\ell}\right)^\ell \frac{m}{\sqrt{m^2 - \ell^2}}$$
$$= \frac{1}{e^C \sigma} \frac{1}{\left(1 - \frac{\ell^2}{m^2}\right)^m} \left(\frac{1 - \ell/m}{1 + \ell/m}\right)^\ell \frac{1}{\sqrt{1 - \ell^2/m^2}}$$
$$= \frac{1}{e^C \sigma} \frac{1}{\left(1 - \frac{2v^2}{m}\right)^m} \left(\frac{1 - \sqrt{2/m} \cdot v}{1 + \sqrt{2/m} \cdot v}\right)^{\sqrt{2m} \cdot v} \frac{1}{\sqrt{1 - 2v^2/m}}\,.$$

Nun können wir langsam daran denken, n (und damit auch m) gegen unendlich gehen zu lassen; wir verwenden dazu die aus der Analysis I und wahrscheinlich auch aus der Schule bekannte Beziehung

$$e^x = \lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n .$$

Danach ist insbesondere

$$\lim_{m\to\infty}\left(1-\frac{2v^2}{m}\right)^m = e^{-2v^2}$$

und

$$\lim_{m\to\infty}\left(1\pm\sqrt{\frac{2}{m}}\cdot v\right)^{\sqrt{2m}\cdot v} = \lim_{m\to\infty}\left(1\pm\frac{\sqrt{2}\cdot v}{\sqrt{m}}\right)^{\sqrt{m}\cdot\sqrt{2}\cdot v}$$

$$= \lim_{q\to\infty}\left(1\pm\frac{\sqrt{2}\cdot v}{q}\right)^{q\cdot\sqrt{2}\cdot v} = \left(e^{\pm\sqrt{2}\cdot v}\right)^{\sqrt{2}\cdot v} = e^{\pm 2v^2} ,$$

denn es bleibt sich natürlich gleich, ob m oder $q = \sqrt{m}$ gegen unendlich geht. Da der Term v^2/m gegen null geht, erhalten wir somit als Grenzwert des gesamten obigen Ausdrucks

$$\frac{1}{e^C\sigma}\cdot\frac{1}{e^{-2v^2}}\cdot\frac{e^{-2v^2}}{e^{+2v^2}}\cdot 1 = \frac{1}{e^C\sigma}e^{-2v^2} .$$

Beachten wir nun noch, daß $v = u/2\sigma$ war, erhalten wir

$$\frac{1}{e^C\sigma}e^{-u^2/2\sigma^2} .$$

Damit sind wir fast am Ziel; das einzige, was noch fehlt, ist die Konstante C. Diese können wir bestimmen, indem wir ausnutzen, daß jeder Fehler mit Wahrscheinlichkeit eins zwischen $-\infty$ und ∞ liegt, d.h.

$$\frac{1}{e^C\sigma}\int_{-\infty}^{\infty} e^{-u^2/2\sigma^2}\,du = 1 .$$

Aus **[HM1]**, *Kap.* 2, §6*c),* wissen wir, daß

$$\int_{-\infty}^{\infty} e^{-x^2}\,dx = \sqrt{\pi}$$

ist; mit der Substitution $x = u/\sqrt{2}\sigma$ folgt, daß dann

$$\int_{-\infty}^{\infty} e^{-u^2/2\sigma^2}\,du = \sqrt{2\pi}\sigma$$

ist und

$$e^C = \sqrt{2\pi} \quad \text{oder} \quad C = \tfrac{1}{2}\ln(2\pi) .$$

Damit haben wir die Wahrscheinlichkeitsdichte endlich vollständig berechnet; das Endergebnis ist

$$\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-u^2}{2\sigma^2}} .$$

Auch die Formel für $n!$ können wir nach der Bestimmung von C nun vollständig hinschreiben:

$$\ln n! = n(\ln n - 1) + \ln\sqrt{2\pi n} + o(1) \quad \text{oder} \quad n! \approx \left(\frac{n}{e}\right)^n\sqrt{2\pi n} .$$

Beides bezeichnet man als STIRLINGsche Formel.

Der schottische Mathematiker JAMES STIRLING (1692–1770) war Anhänger des gestürzten Königs Jakob II Stuart und hatte deshalb große politische Probleme bei seinem Studium; unter anderem wurde er deshalb von der Universität Oxford ausgeschlossen. 1717–1722 lebte er in Venedig und hatte auch gute Kontakte zu NICOLAUS BERNOULLI an der Universität von Padua; außerdem brachte er aus Venedig die Produktionsgeheimnisse der dortigen Glasbläser mit. Ab 1724 arbeitete er zehn Jahre lang als Mathematiklehrer in London, wo er viel mit NEWTON zusammentraf; 1735 wurde er Direktor einer schottischen Bergbaugesellschaft. In seine Londoner Zeit fällt die Veröffentlichung seines bedeutendsten Werks *Methodus Differentialis sive Tractatus de Summatione er Interpolatione Serierum Infinitarum* im Jahre 1730, das die obige Formel als Beispiel zwei zu Proposition 28 enthält. Ebenfalls ziemlich bekannt wurde seine 1735 veröffentlichte Arbeit über die Gestalt der Erde.

d) Der zentrale Grenzwertsatz

Betrachten wir die Summe zweier unabhängiger Zufallsvariablen X und Y mit Verteilungsfunktionen f und g, so nimmt $X+Y$ offensichtlich genau dann einen Wert zwischen a und b an, wenn Y einen Wert aus dem Intervall $[a-x, b-x]$ annimmt, wobei x der von der Zufallsvariablen X

gelieferte Wert ist. Diese Wahrscheinlichkeit ist für einen festen Wert von x gleich

$$\int_{a-x}^{b-x} g(y)\,dy = \int_a^b g(u)\,du \quad \text{mit} \quad u = x + y\,.$$

Die Wahrscheinlichkeit dafür, daß $X+Y$ einen Wert aus $[a,b]$ annimmt ist daher nach dem Satz von FUBBINI und wegen der Unabhängigkeit der beiden Variablen

$$\int_{-\infty}^{\infty} f(x)\left(\int_a^b g(u-x)\,du\right)\,dx = \int_a^b \left(\int_{-\infty}^{\infty} f(x)g(u-x)\,dx\right)\,du$$
$$= \int_a^b (f \star g)(u)\,du\,.$$

Die Wahrscheinlichkeitsdichte der Summe $X+Y$ ist daher gleich der Faltung $f \star g$ der Wahrscheinlichkeitsdichten der Summanden.

Falls wir eine Summe aus N unabhängigen Zufallsvariablen X_i mit Verteilungsfunktionen f_i betrachten, ist deren Verteilungsfunktion somit gleich der Faltung $s = f_1 \star f_2 \star \cdots \star f_N$.

Die Verteilungsfunktion f des Mittelwerts der X_i genügt offensichtlich der Bedingung

$$\int_a^b f(x)\,dx = \int_{Na}^{Nb} s(x)\,dx\,.$$

Die Substitution $u = x/N$ macht das rechte Integral zu

$$\int_{Na}^{Nb} s(x)\,dx = \int_a^b s(Nu)\,N\,du = N\int_a^b s(Nu)\,du\,,$$

also ist $f(x) = N \cdot s(Nx)$.

Faltungen, insbesondere solche mit vielen Faktoren, sind eher unangenehm zu berechnen; durch FOURIER-Transformation werden sie zu

harmlosen Produkten. Also nehmen wir an, daß für alle betrachten Verteilungsfunktionen FOURIER-Transformierten existieren (was bei den üblicherweise vorkommenden kontinuierlichen Verteilungsfunktionen keine nennenswerte Einschränkung bedeutet), und rechnen mit diesen:

$$\widehat{f}(\omega) = \int_{-\infty}^{\infty} f(x)e^{-i\omega x}\,dx = N \cdot \int_{-\infty}^{\infty} s(Nx)e^{-i\omega x}\,dx$$
$$= N \cdot \int_{-\infty}^{\infty} s(u)e^{-\frac{i\omega u}{N}}\,\frac{du}{N} = \int_{-\infty}^{\infty} s(u)e^{-i\frac{\omega}{N}u}\,du = \widehat{s}\left(\frac{\omega}{N}\right).$$

Da außerdem $\widehat{s}(\omega) = \widehat{f}_1(\omega)\cdots\widehat{f}_N(\omega)$ ist, folgt

$$\widehat{f}(\omega) = \prod_{k=1}^{N} \widehat{f}_i\left(\frac{\omega}{N}\right).$$

Um zu sehen, was die FOURIER-Transformierte einer Verteilungsfunktion ist, schreiben wir den Exponentialfaktor im FOURIER-Integral als Potenzreihe:

$$\widehat{f}(\omega) = \int_{-\infty}^{\infty} f(x)e^{-i\omega x}\,dx = \int_{-\infty}^{\infty} f(x)\sum_{k=0}^{\infty}\frac{(-i\omega x)^n}{n!}\,dx$$
$$= \sum_{k=0}^{\infty}\frac{(-i\omega)^n}{n!}\int_{-\infty}^{\infty} f(x)x^n\,dx = \sum_{k=0}^{\infty}\frac{(-i\omega)^n}{n!}\mathbb{E}(X^n)\,.$$

$\mathbb{E}(X^0)$ ist natürlich die Konstante Eins, und $\mathbb{E}(X^1) = \mathbb{E}(X)$ ist der Erwartungswert von X. Falls dieser verschwindet, ist $\mathbb{E}(X^2)$ der Erwartungswert der mittleren quadratischen Abweichung vom Mittelwert, also die Varianz.

Betrachten wir der Einfachheit halber zunächst den Fall, daß die Erwartungswerte aller X_i verschwinden. Dann ist $\widehat{f}_k(\omega) = 1 - \frac{\sigma_k^2\omega^2}{2} + \cdots$ und

$$\widehat{f}(\omega) = \prod_{k=1}^{N}\widehat{f}_k\left(\frac{\omega}{N}\right) = \prod_{k=1}^{N}\left(1 - \frac{\sigma_k^2}{2}\left(\frac{\omega}{N}\right)^2 + \cdots\right),$$

wobei σ_k^2 die Varianz von X_k ist. Die weggelassenen und nur durch Punkte angedeuteten Terme enthalten Potenzen von ω/N mit Exponent mindestens drei; für große N können diese Terme gegenüber dem Quadrat von ω/N vernachlässigt werden. Also ist

$$\widehat{f}(\omega) = \prod_{k=1}^{N} \widehat{f}_k\left(\frac{\omega}{N}\right) \approx \prod_{k=1}^{N}\left(1 - \frac{\sigma_k^2}{2}\left(\frac{\omega}{N}\right)^2\right).$$

Falls alle σ_k einen gemeinsamen Wert σ_0 haben, hat der Mittelwert der X_k nach §4*d)* die Varianz $\sigma^2 = \frac{1}{N}\sigma_0^2$ und

$$\widehat{f}(\omega) = \left(1 - \frac{\sigma_0^2}{2}\left(\frac{\omega}{N}\right)^2\right)^N = \left(1 - \frac{\sigma^2\omega^2}{2N}\right)^N,$$

was bekanntlich für $N \to \infty$ gegen $e^{-\frac{\sigma^2\omega^2}{2}}$ konvergiert.

Auch wenn die σ_k verschieden sind, können wir den Grenzwert leicht ausrechnen: Für große Werte von N ist ω/N klein, und für kleine Werte von x ist

$$\ln(1-x) = \ln 1 - \frac{d\,\ln x}{dx}(1)\cdot x + o(x) = -x + o(x),$$

also hier

$$\ln \widehat{f}(\omega) = \sum_{k=1}^{N} \ln \widehat{f}_k\left(\frac{\omega}{N}\right) = \sum_{k=1}^{N} \ln\left(1 - \frac{\sigma_k^2}{2}\left(\frac{\omega}{N}\right)^2 + \cdots\right)$$
$$\approx \sum_{k=1}^{N} \frac{\sigma_k^2}{2}\left(\frac{\omega}{N}\right)^2 = \frac{1}{N^2}\sum_{k=1}^{N} \sigma_k^2 \frac{\omega^2}{2}.$$

Nach dem Fehlerfortpflanzungsgesetz ist die von $\frac{1}{N}\sum X_k$ gleich

$$\sigma^2 = \sum \frac{\sigma_k^2}{N^2}, \quad \text{also ist} \quad \widehat{f}(\omega) \approx e^{\frac{\sigma^2\omega^2}{2}}.$$

Wie wir aus Kapitel 3, §7*b)* wissen, ist dies die FOURIER-Transformierte von

$$f(t) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{t^2}{2\sigma^2}},$$

die Verteilung des Mittelwerts ist also eine Normalverteilung mit Erwartungswert null und Varianz σ^2.

Falls die Zufallsvariablen X_k von Null verschiedene Erwartungswerte haben, etwa $\mathbb{E}(X_k) = \mu_k$, haben die Zufallsvariablen $Y_k \underset{\text{def}}{=} X_k - \mu_k$ Erwartungswert null und dieselbe Varianz σ_k^2 wie die X_k. Das arithmetische Mittel der X_k unterscheidet sich um

$$\mu = \frac{1}{N}\sum_{k=1}^{N} \mu_k$$

vom arithmetischen Mittel null der Y_k, also genügt es einer Normalverteilung mit Mittelwert μ und Varianz σ^2.

e) Eigenschaften der Normalverteilung

Oft interessiert nicht so sehr die Verteilung der Fehler, sondern die der Meßwerte selbst. Ist $\hat{x}$ der korrekte Wert und x_i der i-te Meßwert dafür, der gemäß $x_i = \hat{x} + u_i$ mit dem Fehler u_i behaftet ist, so können wir mit $x = \hat{x} + u$ die obigen Wahrscheinlichkeitsdichte auch als

$$\frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\hat{x})^2/2\sigma^2}$$

schreiben.

Als *Normalverteilung mit Mittelwert a und Standardabweichung σ* bezeichnen wir daher die Verteilung mit Wahrscheinlichkeitsdichte

$$\varphi(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-a)^2}{2\sigma^2}}.$$

Diese Wahrscheinlichkeitsdichte hängt offensichtlich nur von der *normierten* Variablen

$$z = \frac{x-a}{\sigma}$$

ab; diese hat Mittelwert null und Standardabweichung eins. Daher gibt es für die Normalverteilung nicht – wie für viele andere statistische Verteilungen – je nach Parameterwerten verschiedene Tabellen, sondern man findet in allen Tabellenwerken nur die Normalverteilung mit Mittelwert null und Standardabweichung eins, man findet also die Wahrscheinlichkeitsdichte

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

und deren Integral

$$F(z) = \int_{-\infty}^{z} f(u)\,du = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{u^2}{2}}\,du\,.$$

Dieses Integral läßt sich nicht weiter vereinfachen, da sich die Stammfunktion von $e^{-u^2/2}$ nicht durch elementare Funktionen ausdrücken läßt. Für die Bestimmung von $F(z)$ ist man daher auf Tabellen oder Computerprogramme angewiesen; eine graphische Darstellung von $F(z)$ ist in Abbildung 68 zu sehen. Mit dieser Funktion läßt sich die Wahrscheinlichkeit dafür, daß

$$c \le z = \frac{x-a}{\sigma} \le d$$

ist berechnen als $F(d) - F(c)$, und damit läßt sich auch leicht die Wahrscheinlichkeit berechnen, daß x selbst zwischen zwei gegebenen Schranken liegt.

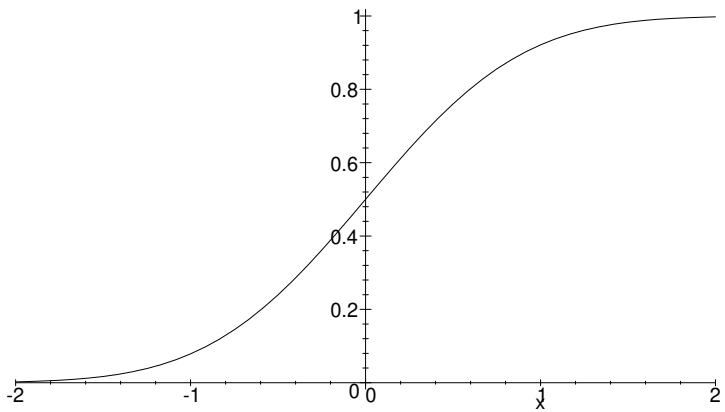


Abb. 68: Das Integral $F(z)$ über die „Glockenkurve“

Mißt man beispielsweise die Temperatur eines Wasserbads eine Viertelstunde lang jede Minute und erhält dabei 15 Meßwerte mit Mittelwert 20,1°C und Standardabweichung 0,2°C, so ist die Standardabweichung des Mittelwerts

$$\sigma_{\overline{\vartheta}} = \frac{0,2^\circ\mathrm{C}}{\sqrt{14}} \approx 0{,}053^\circ\mathrm{C}\,.$$

Wenn wird dann beispielsweise wissen wollen, mit welcher Wahrscheinlichkeit die „tatsächliche“ mittlere Temperatur zwischen 20,0°C und 20,2°C liegt, müssen wir dazu zunächst die normalisierten Werte berechnen:

$$z_1 = \frac{20{,}0 - 20{,}1}{0{,}053} \approx -1{,}89 \quad \text{und} \quad z_2 = \frac{20{,}2 - 20{,}1}{0{,}053} \approx 1{,}89\,.$$

Die Wahrscheinlichkeit ist also

$$F(1{,}89) - F(-1{,}89) \approx 0{,}94\,;$$

oder rund 94%.

Schaut man in einer Tabelle nach, wird man dort allerdings im allgemeinen nur den Wert $F(1{,}89)$ finden, nicht aber $F(-1{,}89)$. Der Grund dafür liegt in der Symmetrie des Graphen von F bezüglich des Punktes $(0, \frac{1}{2})$. Was dahinter steckt, sieht man am besten, wenn man die Dichtefunktion der Normalverteilung betrachtet, also die Glockenkurve: Für $z > 0$ ist $F(-z)$ die in Abbildung 69 links eingezeichnete schraffierte Fläche. Diese Fläche ist wegen der Symmetrie der Glockenkurve zur senkrechten Achse gleich der rechts eingezeichneten schraffierten Fläche, und deren Komplement ist $F(z)$. Also ist

$$F(-z) = 1 - F(z)\,,$$

und es reicht, wenn wir die Werte von F im positiven Bereich kennen.

Oft interessiert auch die Wahrscheinlichkeit dafür, daß der Betrag des Fehlers unterhalb einer bestimmten Schranke liegt, etwa $z \cdot \sigma$; in Abbildung 69 wäre dies der nichtschraffierte Bereich unter der Glockenkurve.

Wie man sich anhand der Abbildung leicht klarmacht, ist diese Wahrscheinlichkeit gleich

$$F(z) - F(-z) = 2F(z) - 1\,;$$

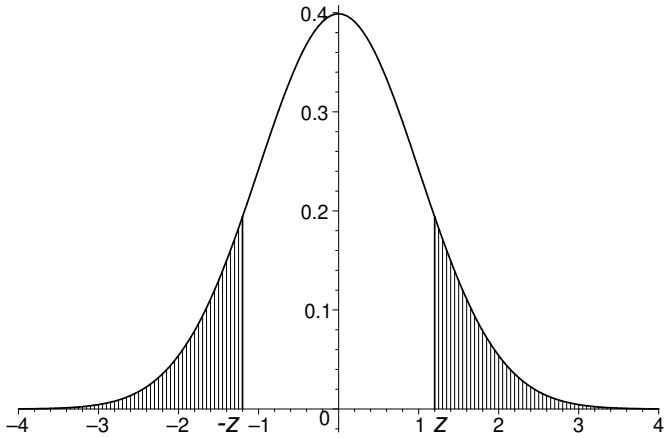


Abb. 69: Zusammenhang zwischen $F(z)$ und $F(-z)$

die Wahrscheinlichkeit, daß wir im obigen Beispiel die mittlere Temperatur mit einem Fehler von höchstens 0,05° gemessen haben, ist also

$$2F\left(\frac{0{,}05}{0{,}053}\right) \approx F(0{,}94) \approx 0{,}83\,.$$

Ein Wasserbad hat üblicherweise den Sinn, ein Experiment unter kontrollierten Temperaturbedingungen durchzuführen; daher interessiert vor allem, inwieweit es gelingt, die Temperatur innerhalb gewisser Schranken zu halten. Die Wahrscheinlichkeit dafür können wir mit denselben Methoden berechnen, allerdings müssen wir dazu mit der Standardabweichung der Meßreihe selbst arbeiten.

Wenn wir etwa wollen, daß die Temperatur immer zwischen 19,5 und 20,5°C liegt, so ist die Wahrscheinlichkeit, daß wir dies mit dem oben ausgemessenen Versuchsaufbau erreichen, gleich

$$F\left(\frac{20{,}5-20{,}1}{0{,}2}\right) - F\left(\frac{29{,}5-20{,}1}{0{,}2}\right) = F(2) - F(-3) \approx 0{,}976\,.$$

In knapp zweieinhalb Prozent aller Fälle, im Schnitt also alle vierzig Minuten, müssen wir also damit rechnen, daß die Toleranzgrenzen überschritten werden.

Wie Abbildung 68 zeigt, liegt $F(-2)$ sehr nahe bei null und $F(2)$ sehr nahe bei eins. In der Tat ist die Wahrscheinlichkeit dafür, daß ein Wert z Betrag größer z hat, nach obiger Diskussion gleich

$$1 - (2F(z) - 1) = 2F(z) - 2\,,$$

was für $z = 2$ zu -0,0455 wird; die Wahrscheinlichkeit ist also kleiner als 5%. Allgemein gilt für eine beliebige Normalverteilung, daß der Wert der Variablen mit folgenden Wahrscheinlichkeiten um höchstens $i\sigma$ vom Mittelwert abweicht:

$i =$	1	2	3	4
Wahrscheinlichkeit:	0,683	0,954	0,9973	0,99994

Damit liegen also etwa zwei Drittel aller Fehler zwischen $-\sigma$ und σ, 95% liegen zwischen -2σ und 2σ und 99,7% zwischen -3σ und 3σ; die Wahrscheinlichkeit dafür, daß der Fehler größer als 3σ ist, beträgt nur etwa 0,27%. Da Ereignisse mit einer so geringen Wahrscheinlichkeit seltener als in einem von 300 Fällen auftreten, betrachtet man Fehler, die außerhalb des 3σ-Bereichs liegen, oft als „Ausreißer", d.h. als grobe Meßfehler, die bei der Bestimmung des Ergebnisses nicht berücksichtigt werden. Sehr vorsichtige Leute reden allerdings erst ab einer Abweichung von 4σ von Ausreißern; solche Fehler treten zufällig weniger als einmal pro 15 000 Messungen auf.

Für Leser, die ihren Computer selbst programmieren und keine spezielle Statistiksoftware haben, sei hier eine Näherungsformel für $F(z)$ angegeben: Mit einem Fehler von höchstens $7{,}5 \cdot 10^{-8}$ ist

$$F(z) = 1 - \varphi(z) \cdot \left(a_1 t + a_2 t^2 + a_3 a^3 + a_4 t^4 + a_5 t^5\right)$$

mit $t = \dfrac{1}{1+pz}$ und $\varphi(z) = \dfrac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$ sowie

$a_1 = 0{,}319\,381\,530$ $\quad a_2 = -0{,}356\,563\,782$ $\quad a_3 = 1{,}781\,477\,973$
$a_4 = -1{,}821\,255\,978$ $\quad a_5 = 1{,}330\,274\,429$ $\quad p = 0{,}231\,641\,9$.

Beim Rechnen mit dem Taschenrechner kann man sich auch mit einer vereinfachten Version begnügen, bei der $a_4 = a_5 = 0$ ist und

$a_1 = 0{,}436\,1836$ $\quad a_2 = -0{,}120\,1676$ $\quad a_3 = 0{,}937\,2980$ $\quad p = 0{,}332\,67$;

hier kann der Fehler bis zu 10^{-5} betragen.

f) Die Maximum Likelihood Methode

GAUSS gab im Laufe seines Lebens mehrere Begründungen für die Methode der kleinsten Quadrate (die er bei sowohl bei seinen astronomischen Arbeiten wie auch bei der von ihm geleiteten Vermessung des Königreichs Hannover zwischen 1818 und 1832 ständig benutzte); die unter dem Gesichtspunkt einer in sich geschlossenen Fehlertheorie interessanteste beruht auf dem LAPLACEschen Fehlermodell:

Danach sollte der Wert u_i für die korrekten Parameterwerte $a, b, \ldots$ aus einer Normalverteilung mit Mittelwert $f(a, b, \ldots; t_i)$ kommen, deren Standardabweichung σ_i von der Genauigkeit abhängt, mit der u_i bestimmt werden kann. Die Wahrscheinlichkeit dafür, daß u_i zwischen zwei Werten a und b liegt, ist damit

$$\int_a^b e^{-\left(u - f(a,b,\ldots;\, t_i)\right)^2/2\sigma_i^2}\,.$$

Von der Wahrscheinlichkeit, daß u_i gleich einem Wert c ist, können wir natürlich nicht reden, da diese nach obiger Formel ein Integral von c nach c wäre, also Null. Aber die Wahrscheinlichkeit dafür, daß u_i in einem kleinen Intervall der Länge ε_i um einen Wert c_i liegt, ist ungefähr proportional zum ε_i-fachen Wert des Integranden an der Stelle c_i, also zu

$$\varepsilon_i \cdot e^{-\left(c_i - f(a,b,\ldots;\, t_i)\right)^2/2\sigma_i^2}\,.$$

Entsprechend ist

$$\varepsilon_j \cdot e^{-\left(c_j - f(a,b,\ldots;\, t_j)\right)^2/2\sigma_j^2}$$

ungefähr gleich der Wahrscheinlichkeit dafür, daß u_j in einem Intervall der Breite ε_j um c_j liegt.

Wenn wir wie üblich davon ausgehen, daß keine systematischen Fehler auftreten, sind die Fehler von u_i und u_j voneinander unabhängig, die Wahrscheinlichkeit dafür, daß (u_i, u_j) in einem Rechteck mit Seiten ε_i und ε_j um (c_i, c_j) liegt, ist also proportional zum Produkt der beiden

obigen Einzelwahrscheinlichkeiten, d.h. zu

$$\varepsilon_i\varepsilon_j\, e^{-\left(c_i - f(a,b,\ldots;\, t_i)\right)^2/2\sigma_i^2 - \left(c_j - f(a,b,\ldots;\, t_j)\right)^2/2\sigma_j^2}\,.$$

Entsprechend kann auch die Wahrscheinlichkeit dafür berechnet werden, daß der Punkt $(u_1, \ldots, u_n)$ in einem kleinen gegebenen Quader mit Kantenlängen $\varepsilon_1, \ldots, \varepsilon_n$ liegt; sie ergibt sich zu

$$L(a, b, \ldots) \cdot \prod_{i=1}^{n} \varepsilon_i$$

mit

$$L(a, b, \ldots) \underset{\text{def}}{=} e^{\sum\limits_{i=1}^{n}\left(u_i - f(a,b,\ldots;\, t_i)\right)^2/2\sigma_i^2}\,.$$

Diese Größe ist selbst keine Wahrscheinlichkeit, sondern der Quotient aus einer Wahrscheinlichkeit und einem Volumen; man spricht daher von einer *Wahrscheinlichkeitsdichte.*

Wenn wir diese Wahrscheinlichkeitsdichte als Funktion von $a, b, \ldots$ betrachten, macht sie eine Aussage über die Güte der Parameter: Schließlich wird man einem Modell, das dem beobachteten Ausgang eines Experiments eine hohe Wahrscheinlichkeit zuweist, eher glauben als einem alternativen Modell, das die beobachteten Daten zu Ausreißern erklärt. Aus diesem Grund kann die Funktion L auch als Maß dafür betrachtet werden, wie „wahrscheinlich“ in irgendeinem umgangssprachlichen (und schwer präzisierbaren) Sinne die Parameter $a, b, \ldots$ sind.

Im englischen gibt es zwei Wörter für Wahrscheinlichkeit: Das romanische Wort *probability* und das germanische Wort *likelihood.* Für den mathematisch exakten Wahrscheinlichkeitsbegriff verwendet man probability, für „Wahrscheinlichkeit“ im Sinne der Funktion L likelihood. Da es im deutschen kein zweites Wort für Wahrscheinlichkeit gibt, spricht man hier in Anlehnung an das Englische von einer *Likelihoodfunktion.*

Die Maximum Likelihood Methode besteht nun genau in dem, was ihr Name besagt: *Man wähle die Parameter $a, b, \ldots$ so, daß die Likelihoodfunktion maximal wird.*

Da $L(a, b, \ldots)$ durch eine Exponentialfunktion beschrieben wird, wird die Likelihoodfunktion genau dann maximal, wenn ihr Exponent maximal wird. Dieser Exponent ist eine negative Zahl, wird also genau dann maximal, wenn sein Betrag *minimal* wird, das heißt, wenn die Quadratsumme

$$\sum_{i=1}^{n} \frac{\big(u_i - f(a, b, \ldots; t_i)\big)^2}{2\sigma_i^2}$$

minimal wird.

In vielen Fällen wird die Zuverlässigkeit der einzelnen Paare (t_i, u_i) miteinander vergleichbar sein, so daß alle σ_i gleich sind; in diesem Fall kann man die σ_i ignorieren und einfach die Quadratsumme

$$\sum_{i=1}^{n} \big(u_i - f(a, b, \ldots; t_i)\big)^2$$

minimieren, d.h. wir kommen wieder zur klassischen Methode der kleinsten Quadrate. Es gibt aber auch Anwendungen, wie etwa oben beim überexponentiellen Bevölkerungswachstum, bei denen die Verschiedenheit des σ_i sehr wesentlich ist: Sicherlich wird man etwa der auf Volkszählungen beruhenden Weltbevölkerungszahl, die die Vereinten Nationen für 1995 veröffentlichten, mehr Vertrauen entgegenbringen als der Schätzung eines Historikers für die Weltbevölkerung des Jahres Null, und selbst bei ein und derselben Meßreihe im Labor kommt es gelegentlich vor, daß (beispielsweise aufgrund unterschiedlicher Genauigkeit eines Meßinstruments in verschiedenen Bereichen) manche Daten zuverlässiger sind als andere.

§8: Kompression von Bild- und Audiodaten

Zum Abschluß der Vorlesung wollen wir wenigstens kurz eine praktische Anwendung kennenlernen, in der mit Eigenwerten und Eigenvektoren symmetrischer Matrizen, FOURIER-Transformationen und Statistik gleich mehrere der Methoden aus diesem Semester gleichzeitig benötigt werden: die Komprimierung von Bild- und Audiodaten.

a) Datenkompression

Ziel der Datenkompression ist es, eine Datei für Zwecke der Speicherung oder Übertragung möglichst stark zu verkleinern, das aber in einer solchen Weise, daß sich die ursprüngliche Datei aus der verkleinerten wieder exakt rekonstruieren läßt.

Es ist klar, daß es keinen universellen Algorithmus zur Datenkompression geben kann: Gäbe es nämlich ein Verfahren, das für beliebige Dateien einen Kompressionsfaktor $\alpha < 1$ garantieren würde, so könnte man dieses Verfahren iterativ anwenden und nach n Anwendungen eine Kompressionsrate von α^n erreichen. Wenn man n nur hinreichend groß wählt, könnte man daher jede Datei auf weniger als ein Bit komprimieren, was natürlich absurd ist.

Ein Kompressionsverfahren kann also nur auf Dateien mit spezieller Struktur erfolgreich angewandt werden und muß die spezielle Redundanz in diesen Dateien ausnutzen. In Textdateien beispielsweise ist dies die Redundanz der Sprache, die schon bei bloßer Beachtung der höchst unterschiedlichen Buchstabenhäufigkeiten Kompressionen von rund 50% gestattet.

Bilddaten werden typischerweise als Matrizen aus ganzen Zahlen zwischen 0 und 255 digitalisiert; bei Audiodaten nimmt man Vektoren von ganzen Zahlen zwischen 0 und $65\,535 = 2^{16} - 1$ oder $16\,777\,215 = 2^{24} - 1$. (Der Unterschied zwischen den Wertenbereichen liegt darin begründet, daß unser Auge selbst bei gedruckten Bildern mit nur 64 Graustufen praktisch keine Artefakte mehr erkennen kann, wohingegen unser Gehör noch auf sehr feine Unterschiede reagiert.)

Bei einer Musik-CD etwa wird das Signal 44 100-mal pro Sekunde abgetastet (dies bedeutet nach dem Abtasttheorem von NYQUIST, daß ein auf den Bereich von 0 bis 22,05kHz bandbegrenztes Signal fehlerfrei rekonstruiert werden kann), und das Ergebnis wird dann so skaliert und quantisiert (d.h. gerundet), daß eine Zahl zwischen 0 und 65535 entsteht. Bei Bilddaten werden je nach Auflösung und Seitenverhältnis zwischen etwa 256×256 und 1024×1024 Bildpunkte abgetastet, für Schwarzweißbilder nur nach Helligkeit, für Farbbildern nach insgesamt

drei Größen, die vom jeweiligen Farbmodell abhängen. Das Ergebnis dieser Abtastungen wird dann entsprechend skaliert und quantisiert.

Typische Komprimierungsverfahren arbeiten daher mit Vektoren oder Matrizen aus Zahlen zwischen 0 und einer geeigneten Zahl M, die aus praktischen Gründen meist von der Form $2^{8r} - 1$ ist, wobei die Zahl r der Empfindlichkeit unserer Sinne angepaßt zwischen eins und drei liegt.

Da man zur eindeutigen Festlegung von N beliebigen Zahlen zwischen 0 und 2^{8r} nicht mit weniger als den $8Nr$ Bit auskommen kann, die man zum Hinschreiben der Zahlen braucht, sehen wir auch hier wieder, daß kein Verfahren *alle* solchen Vektoren komprimieren kann; wir müssen also eine Teilmenge auszeichnen.

Die ideale solche Teilmenge wäre hier natürlich die Menge aller möglicher Bilder (oder Audiosequenzen), aber diese Menge dürfte mathematisch kaum definierbar sein: Schließlich hängt es sehr vom Betrachter ab, welches Pixelmuster er noch als „Bild“ gelten läßt und welches nicht. Sinnvoll läßt sich eine solche Menge daher höchstens definieren, wenn von vornherein feststeht, welche Bilder berücksichtigt werden sollen – und dann ist wohl ein Verfahren, das statt vom Bildinhalt von einer Bildnummer ausgeht, unschlagbar.

Die meisten klassischen Verfahren, die beliebige, aber realistische Bilder komprimieren sollen, gehen aus von einem *statistischen Modell,* das zwar auch viele Matrizen produziert, die niemand als „Bilder“ anerkennen würde, das aber dennoch genügend viele Eigenschaften realer Bilder reproduziert, um eine große Anzahl von „Nichtbildern“ auszuschließen.

Ausgangspunkt ist die Beobachtung, daß es in einem Bild oder Musikstück nur wenige aprupte Übergänge gibt. Zwar gibt es natürlich immer wieder ein plötzliches *fortissimo,* das auf eine leise Stelle folgt, aber da das Signal 44 100-mal pro Sekunde abgetastet wird und solche Übergänge selbst bei der schrägsten Musik deutlich seltener als im Sekundenrythmus erfolgen, sind diese Sprünge innerhalb des zu behandelnden Datenstroms in der Tat sehr seltene Ereignisse. Wir können daher davon ausgehen, daß sich die unmittelbaren Nachbarn eines Datums *im Mittel* nur wenig vom gegebenen Datum unterscheiden.

Dasselbe gilt auch für Bilddaten: Falls das Bild digital hinreichend fein dargestellt wird, so daß keine Rastereffekte erkennbar sind, kommen große Sprünge in den Helligkeitswerten nur selten vor.

Bei diesem engen Zusammenhang zwischen benachbarten Werten setzen viele gängige Komprimierungsalgorithmen an: Wenn zwei Größen typischerweise sehr ähnlich sind, wird bei der Übertragung oder Speicherung *beider* Werte ein großer Teil der Information doppelt betrachtet; die Informationsdichte kann also deutlich erhöht werden, wenn man nur Informationen betrachtet, die weitgehend unabhängig voneinander sind.

Aus dem letzten Semester kennen wir ein Maß für die gegenseitige Abhängigkeit von Daten: Als wir dort untersuchten, wie das Klausurergebnis eines Studenten von seiner Arbeit bei den wöchentlichen Übungen abhängt oder die Korruption eines Staats vom Bruttosozialprodukt pro Einwohner, überprüften wir die Qualität unserer Modelle mit Hilfe des Korrelationskoeffizienten: Dieser lag bei ± 1 bei perfekter Übereinstimmung, und nahe Null, wenn das Modell keinen Zusammenhang zwischen den Daten lieferte.

Dieselbe Technik können wir auch anwenden, um Abhängigkeiten innerhalb einer Folge zu finden; bevor wir diese sogenannte Autokorrelation verstehen können, brauchen wir aber zunächst noch einige Vorbereitungen aus der Stochastik.

b) Korrelation von Zufallsvariablen

Ein guter Komprimierungsalgorithmus muß auch für Bilder funktionieren, die wir erst in ein paar Jahren photographieren. Für den Grundalgorithmus zur Datenkompression müssen wir daher Daten zulassen, über die wir noch nichts konkretes wissen – abgesehen von gewissen vagen Gesetzmäßigkeiten, durch die sich „echte“ Bilddaten von beliebigen Matrizen unterscheiden. Es bietet sich daher an, die Helligkeits- und/oder Farbwerte der einzelnen Pixel *bzw.* die Schalldruckwerte bei Tonaufnahmen durch Zufallsvariablen zu beschreiben. Da wir an digitalen Bild- und Audiodaten interessiert sind, verwenden wir dazu diskrete Zufallsvariablen.

So, wie wir sie bislang definiert haben, ist jede Zufallsvariable ein eigenständiger Prozeß, und zwei verschiedene Zufallsvariablen haben nichts miteinander zu tun. Das ist natürlich nicht das, was wir hier brauchen; wir müssen wir davon ausgehen, daß ein einziger Prozeß gleichzeitig einen ganzen Vektor *bzw.* eine ganze Matrix von Zufallswerten erzeugt, wobei deren einzelne Komponenten dann sehr wohl voneinander abhängig sein können.

Für zwei solche Komponenten X und Y mit jeweiligen Wertebereichen $\{x_0, \ldots, x_m\}$ und $\{y_0, \ldots, y_n\}$ sowie Wahrscheinlichkeiten p_i für x_i und q_j für y_j ist die Wahrscheinlichkeit dafür, daß X den Wert x_i liefert und Y den Wert y_j dann nicht $p_i q_j$, wie das bei unabhängigen Variablen der Fall wäre, sondern irgendeine Wahrscheinlichkeit π_{ij}, von der wir nur wissen, daß aus offensichtlichen Gründen etwa

$$\sum_{j=0}^{m} \pi_{ij} = p_i \quad \text{und} \quad \sum_{i=0}^{m} \pi_{ij} = q_j$$

sein muß. Für so ein Paar definieren wir

Definition: *a)* Die Kovarianz eines solchen Paars (X, Y) ist

$$\begin{aligned}\operatorname{cov}(X,Y) &= \mathbb{E}\big((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\big) \\ &= \sum_{i=0}^{m}\sum_{j=0}^{n} \pi_{ij}\big((x_i - \mathbb{E}(X)\big)\big((y_j - \mathbb{E}(Y)\big)\,.\end{aligned}$$

b) Die Korrelation von (X, Y) ist $\rho(X,Y) = \dfrac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$.

Der Vollständigkeit halber sei auch die entsprechende Definition für koninuierliche Zufallsvariablen angegeben: Haben X und Y die zweidimensionale Wahrscheinlichkeitsdichte f, ist also die Wahrscheinlichkeit dafür, daß das Paar (X, Y) einen Wert in einer Teilmenge $B \subseteq \mathbb{R}^2$ liefert, gleich

$$\iint_B f(x,y)\,dx\,dy\,,$$

so definieren wir die Kovarianz des Paars als

$$\begin{aligned}\operatorname{cov}(X,Y) &= \mathbb{E}\big((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\big) \\ &= \iint_{\mathbb{R}^2} \big(x - \mathbb{E}(X)\big)\big(y - \mathbb{E}(Y)\big) f(x,y)\,dx\,dy\,;\end{aligned}$$

die Korrelation wird dann über dieselbe Formel wie im diskreten Fall definiert.

Wir bezeichnen zwei diskrete Zufallsvariablen X und Y entsprechend der üblichen Definition für Ereignisse als voneinander unabhängig, falls für alle i, j gilt: $\pi_{ij} = p_i q_j$; im kontinuierliche Fall verlangen wir entsprechend, daß die zweidimensionale Wahrscheinlichkeitsdichte f das Produkt der Wahrscheinlichkeitsdichten von X und von Y ist. Alsdann ist im diskreten Fall

$$\begin{aligned}\operatorname{cov}(X,Y) &= \sum_{i=0}^{m}\sum_{j=0}^{n} \pi_{ij}\big(x_i - E(X)\big)\big((y_j - E(Y)\big) \\ &= \sum_{i=0}^{m}\sum_{j=0}^{m} \Big[p_i(x_i - E(X))\big)\Big]\Big[q_j\big(y_j - E(Y)\big)\Big] \\ &= \left(\sum_{i=0}^{m} p_i\big(x_i - E(X)\big)\right)\left(\sum_{j=0}^{n} q_j\big((y_j - E(Y))\big)\right) = 0\,,\end{aligned}$$

denn

$$\sum_{i=0}^{m} p_i\big((x_i - E(X)\big) = \sum_{i=0}^{m} p_i x_i - \sum_{i=0}^{m} p_i E(X) = \sum_{i=0}^{m} p_i x_i - E(X)$$

verschwindet nach Definition des Erwartungswerts.

Damit haben zwei voneinander unabhängige diskrete Zufallsvariablen also Kovarianz und Korrelation null; man rechnet leicht nach, daß dies auch im kontinuierlichen Fall gilt. Solche Zufallsvariablen heißen unkorreliert.

Bei Bilddaten wird das im allgemeinen nicht der Fall sein; hier wird man ims Gegenteil davon ausgehen, daß die Zufallsvariablen zu benachbarten Pixeln sehr stark miteinander korrelieren. Wir können beispielsweise

annehmen, daß $Y = \rho X + Z$ ist mit einer von X unabhängigen Zufallsvariablen Z und einer positiven reellen Zahl $\rho < 1$. Dann ist

$$E(Y) = E(\rho X + Z) = \rho E(X) + E(Z)\,;$$

falls wir annehmen, daß X und Y denselben Erwartungswert haben, ist daher

$$E(Z) = (1-\rho)E(X)\,,$$

was für ein ρ nahe eins deutlich kleiner ist als $E(Z)$.

In dieser Situation ist

$$\begin{aligned}
\operatorname{cov}(X,Y) &= \sum_{i=0}^{m}\sum_{j=0}^{n} \pi_{ij}\big(x_i - E(X)\big)\big((y_j - E(Y)\big) \\
&= \sum_{i=0}^{m}\sum_{j=0}^{n} \pi_{ij}\big(x_i - E(X)\big)\big(\rho x_i + z_j - \rho E(X) - E(Z)\big) \\
&= \sum_{i=0}^{m}\sum_{j=0}^{n} \pi_{ij}\big(x_i - E(X)\big)\big[\rho\big(x_i - E(X)\big) + \big(z_j - E(Z)\big)\big] \\
&= \sum_{i=0}^{m}\sum_{j=0}^{n} \pi_{ij}\rho\big(x_i - E(X)\big)^2 \\
&\qquad + \sum_{i=0}^{m}\sum_{j=0}^{n} \pi_{ij}\big(x_i - E(X)\big)\big(z_j - E(Z)\big) \\
&= \rho\sum_{i=0}^{m} p_i\big(x_i - E(X)\big)^2 + \operatorname{cov}(X,Z) = \rho\sigma_X^2\,,
\end{aligned}$$

da X und Z voneinander unabhängige Zufallsvariablen sind.

Wenn wir jetzt noch annehmen, daß $\sigma_X = \sigma_Y$ ist, folgt

$$\rho(X,Y) = \frac{\rho\sigma_X^2}{\sigma_X\sigma_Y} = \rho\,,$$

wir können auf diese Weise also für beliebiges $\rho \in [0,1]$ ein Paar voneinander abhängiger Zufallsvariablen mir Korrelation ρ erzeugen.

Besser noch: Wann immer zwei Zufallsvariablen mit gleicher Standardabweichung Korrelation ρ haben, sind wir immer im obigen Fall, denn definieren wir eine neue Zufallsvariable Z durch $Z = Y - \rho X$, so ist $E(Z) = E(Y) - \rho E(X)$, und das Paar (X,Z) ist unkorreliert, da

$$\begin{aligned}
\operatorname{cov}(X,Z) &= \sum_{i=0}^{m}\sum_{j=0}^{n} \big(x_i - E(X)\big)\big(z_j - E(Z)\big) \\
&= \sum_{i=0}^{m}\sum_{j=0}^{n} \big(x_i - E(X)\big)\Big[\big(y_j - E(Y)\big) - \rho\big(x_i - E(X)\big)\Big] \\
&= \sum_{i=0}^{m}\sum_{j=0}^{n} \big(x_i - E(X)\big)\big(y_j - E(Y)\big) \\
&\qquad - \rho\sum_{i=0}^{m}\sum_{j=0}^{n} \big(x_i - E(X)\big)^2 \\
&= \operatorname{cov}(X,Y) - \rho\sigma_X^2 = \rho\sigma_X\sigma_Y - \rho\sigma_X^2 = 0\,.
\end{aligned}$$

c) Das Datenmodell

Wir modellieren Bild- und Audiodaten durch eine Folge (X_i) von Zufallsvariablen, die allesamt denselben Erwartungswert μ und dieselbe Varianz σ^2 haben. Außerdem nehmen wir noch an, daß die Korrelation zwischen X_i und X_{i+1} stets denselben Wert κ haben soll, die sogenannte *Autokorrelation* der Folge.

Bei Bildern haben wir es natürlich tatsächlich mit einer zweifach indizierten Folge (X_{ij}) zu tun; hier verlangen wir, daß sowohl die Korrelation zwischen X_{ij} und $X_{i+1,j}$ als auch die zwischen X_{ij} und $X_{i,j+1}$ gleich κ sein soll.

Um die Bedeutung der Kenngrößen μ, σ^2 und κ in der Bildverarbeitung zu veranschaulichen, sind auf der nächsten Doppelseite sechs beliebte Testbilder zusammen mit den Werten dieser Kenngrößen abgedruckt. Die Werte sind entnommen aus

P.M. FARELLE: Recursive Block Coding for Image Data Compression, *Springer,* 1990 ;

Peppers

$\mu =$ 115,6
$\sigma^2 =$ 5632
$\sigma =$ 75,0
$\rho =$ 0,98
$x_{\min} =$ 0
$x_{\max} =$ 237

Lenna

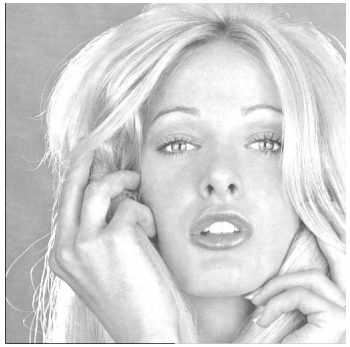
$\mu =$ 99,1
$\sigma^2 =$ 2796
$\sigma =$ 52,9
$\rho =$ 0,97
$x_{\min} =$ 3
$x_{\max} =$ 248

Sailboat

$\mu =$ 124,3
$\sigma^2 =$ 6027
$\sigma =$ 77,6
$\rho =$ 0,97
$x_{\min} =$ 0
$x_{\max} =$ 249

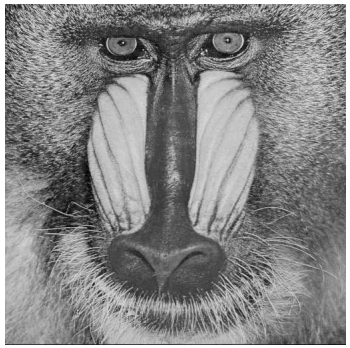
Stream

$\mu =$ 113,8
$\sigma^2 =$ 2996
$\sigma =$ 54,7
$\rho =$ 0,94
$x_{\min} =$ 0
$x_{\max} =$ 255



Tiffany

$\mu =$ 208,6
$\sigma^2 =$ 1126
$\sigma =$ 33,6
$\rho =$ 0,87
$x_{\min} =$ 3
$x_{\max} =$ 255



Baboon

$\mu =$ 128,9
$\sigma^2 =$ 2282
$\sigma =$ 47,8
$\rho =$ 0,86
$x_{\min} =$ 0
$x_{\max} =$ 236

sie beziehen sich natürlich auf die Originalbilder und nicht auf das, was der Druckvorgang hier im Skriptum daraus gemacht hat. Trotzdem sollte der Vergleich von Bildern und Daten einen einigermaßen korrekten Eindruck zumindest der relativen Situation vermitteln, da hoffentlich alle hier abgedruckte Bilder in derselben Weise verunstaltet sind.

Die mittlere Helligkeit eines Bildes, dessen (viele) Pixel durch je eine Zufallsvariable mit Erwartungswert μ produziert werden, sollte ziemlich nahe bei μ liegen; der beste Schätzwert für den gemeinsamen Erwartungswert der Zufallsvariablen ist also die mittlere Helligkeit des Bildes. Typischerweise werden Helligkeiten durch Zahlen zwischen 0 und 255 kodiert, wobei schwarz der Zahl Null entspricht und weiß der 255. Dies sieht man gut an den Beispielbildern, wo das mit Abstand hellste Bild „Tiffany“ auch den mit Abstand größten Mittelwert μ hat; den kleinsten Wert hat das auch visuell dunkelste Bild „Lenna“.

Die nächste wichtige Kenngröße ist die Varianz, welche angibt, wie stark eine Zufallsvariable um ihren Erwartungswert streut. Auch hier wollen wir wieder davon ausgehen, daß alle Zufallsvariablen zu einem gegebenen Bild *bzw.* einer gegebenen Audiosequenz aus N dieselbe Varianz haben. Wir schätzen diese gemeinsame Varianz aufgrund der vorliegenden Daten als

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu)^2 ,$$

wobei $y_1, \ldots, y_N$ die Helligkeits- *bzw.* Lautstärkewerte sind. (Wer sich wundert, daß vor dieser Summe mit N Summanden nur $N-1$ im Nenner steht, sollte zu §4*d)* zurückblättern.)

Was die Varianz und die Standardabweichung bedeuten, sieht man wieder deutlich an den Beispielbildern: Bilder mit geringem Kontrast wie „Tiffany“ oder „Lenna“ haben deutlich geringere Werte als die kontrastreicheren Bilder „Peppers“ und „Sailboat“.

Da der Begriff der Autokorrelation das wohl am schwersten verständliche der hier eingeführten statistischen Konzepte ist, sind die sechs Testbilder in Richtung fallender Autokorrelation geordnet: Die höchste

Autokorrelation hat mit $\rho = 0{,}98$ das Bild „Peppers“, wo die recht homogenen Flächen der Paprikaschoten dafür sorgen, daß sich ein Pixel nur selten von seinen Nachbarn unterscheidet; auch „Lenna“ und „Sailboat“ werden von flächigen Strukturen dominiert. Bei „Stream“ kommen in stärkerem Maße feine Verästelungen von Bäumen und Büschen ins Spiel, so daß die Autokorrelation auf 0,94 absinkt, und bei „Tifffany“ und „Baboon“ schließlich sorgen die vielen Haare für feine Details, die die Autokorrelation auf 0,87 *bzw.* 0,86 herunterdrücken.

d) Komprimierung durch Dekorrelation

Um zu sehen, wie sich die Autokorrelation zur Komprimierung der Daten ausnutzen läßt, betrachten wir der Einfachheit halber zunächst nur eine Folge $X_1, \ldots, X_n$ von Zufallsvariablen; der gemeinsame Erwartungswert sei μ, die gemeinsame Standardabweichung σ, und die Korrelation zwischen X_i und X_{i+1} sei jeweils ρ.

Nach dem oben Gesagten gibt es dann für jede der Zufallsvariablen X_i mit $i < n$ eine davon unabhängige Zufallsvariable Z_i, so daß

$$X_{i+1} = \rho X_i + Z_i$$

ist; entsprechend ist für $i < n-1$

$$X_{i+2} = \rho X_{i+1} + Z_{i+1} = \rho^2 X_i + \rho Z_i + Z_{i+1}$$

usw.; wenn wir zusätzlich annehmen, daß alle Z_i voneinander unabhängig sind, ist also $\rho(X_i, X_{i+2}) = \rho^2$ und allgemein

$$\rho(X_i, X_j) = \rho^{|i-j|} .$$

In der Signalverarbeitung spricht man bei einer solchen Folge von Zufallsvariablen von einem *autoregressive Prozeß;* der hier betrachtete allereinfachste Fall, bei dem alle Korrelationen nur von der Korrelation zwischen zwei benachbarten Zufallsvariablen abhängen, wird als AR(1)-Modell bezeichnet; in der Sprechweise der Wahrscheinlichkeitstheorie handelt es sich hier um spezielle sogenannte MARKOV-Ketten.

Der russische Mathematiker ANDREĬ ANDREEVICH MARKOV (1856–1922) studierte in Sankt Petersburg, wo er später auch Professor wurde. Er beschäftigte sich zunächst hauptsächlich mit Zahlentheorie und Analysis; erst später kommen die wahrscheinlichkeitstheoretischen Arbeiten, für die er heute vor allem bekannt ist. MARKOV-Ketten sind Prozesse ohne Erinnerung, in denen das zukünftige Verhalten nur vom augenblicklichen Zustand abhängt, nicht aber von der Geschichte des Systems. Damit sind sie gerade hier bei Bilddaten nur eine unvollkommene Approximation an die Realität, aber dennoch sehr nützlich.

Falls wir bei einer solchen Folge von Zufallsvariablen die Werte von $X_1, \ldots, X_n$ nacheinander übertragen, übertragen wir zuerst den Wert von X_1, dann mit X_2 noch einmal zu $100 \times \rho$ % denselben Wert, mit X_3 dasselbe noch einmal zu $100 \times \rho^2$ %, *usw.*

Eine offensichtliche Alternative hierzu wäre, nur den Wert von X_1 zu übertragen und ansonsten nur die Werte der Z_i. Eine ähnliche Vorgehensweise wird tatsächlich gelegentlich angewandt, allerdings macht man es sich dann noch einfacher und überträgt nur die *Differenzen,* also

$$X_1, X_2 - X_1, \ldots, X_n - X_{n-1} .$$

Der Nachteil dieses Verfahrens ist, daß sowohl diese Differenzen als auch die Z_i von Zeit zu Zeit sehr groß werden *müssen,* da es in fast jedem Bild oder Musikstück gelegentliche aprupte Veränderungen gibt.

Die Idee hinter allen Komprimierungsverfahren, die auf Transformationen beruhen, ist es, anstelle der Zufallsvariablen X_i geeignete Linearkombinationen

$$Y_i = \sum_{j=1}^{n} \alpha_{ij} X_j$$

zu betrachten, wobei (α_{ij}) eine *invertierbare* $n \times n$-Matrix ist, so daß sich auch umgekehrt die X_i wieder aus den Y_j rekonstruieren lassen. Diese Matrix wird so gewählt, daß die neuen Variablen möglichst unkorreliert sind und daß man die Größe der neuen Variablen möglichst gut abschätzen kann.

Letzterer Aspekt erfordert statistische Betrachtungen, auf die wir hier verzichten wollen; die Dekorrelation der Zufallsvariablen aber führt uns geradewegs zu Eigenvektoren symmetrischer Matrizen:

Wir definieren für eine Folge von Zufallsvariablen deren *Korrelationsmatrix*

$$\mathrm{Kor}(X_1, \ldots, X_n) \in \mathbb{R}^{n \times n}$$

dadurch, daß der Eintrag an der Stelle ij dieser Matrizen jeweils die Korrelation $\rho(X_i, X_j)$ sein soll.

Das Ideal, auf das wir hinarbeiten, sind Zufallsvariablen, deren Korrelationsmatrix eine Diagonalmatrix ist, denn dann sind je zwei verschiedene Variablen unkorreliert.

Da die Korrelationsmatrix eine symmetrische Matrix ist, gibt es eine Orthonormalbasis des $\mathbb{R}^n$ aus reellen Eigenvektoren, bezüglich derer sie Diagonalgestalt hat; die Vektoren dieser Orthonormalbasis seien

$$\vec{b}_1 = \begin{pmatrix} \alpha_{11} \\ \vdots \\ \alpha_{1n} \end{pmatrix}, \ldots, \vec{b}_n = \begin{pmatrix} \alpha_{n1} \\ \vdots \\ \alpha_{nn} \end{pmatrix} .$$

Wir definieren die neuen Zufallsvariablen durch

$$Y_i = \sum_{j=1}^{n} \alpha_{ij} X_j ;$$

dann ist

$$\begin{aligned} \rho(Y_i, Y_k) = \vec{v}_{Y_i} \cdot \vec{v}_{Y_k} &= \left(\sum_{j=1}^{n} \alpha_{ij} \vec{v}_{X_j} \right) \cdot \left(\sum_{\ell=1}^{n} \alpha_{k\ell} \vec{v}_{X_\ell} \right) \\ &= \sum_{j=1}^{n} \sum_{\ell=1}^{n} \alpha_{ij} \; \vec{v}_{X_j} \cdot \vec{v}_{X_\ell} \; \alpha_{k\ell} = \sum_{j=1}^{n} \sum_{\ell=1}^{n} \alpha_{ij} \; \rho(X_j, X_\ell) \; \alpha_{k\ell} \\ &= \sum_{\ell=1}^{n} \left(\sum_{j=1}^{n} \alpha_{ij} \; \rho(X_j, X_\ell) \right) \; \alpha_{k\ell} . \end{aligned}$$

Der Inhalt der großen Klammer ist offensichtlich der Eintrag an der Stelle $i\ell$ der Produktmatrix $A \cdot \mathrm{Kor}(X_1, \ldots, X_n)$, wobei $A = (\alpha_{ij})$ die

Matrix der Koeffizienten α_{ij} ist, und die Summation über ℓ macht daraus den Eintrag an der Stelle ik des Produkts mit ${}^tA$. Insgesamt haben wir also gezeigt, daß

$$\mathrm{Kor}(Y_1, \ldots, Y_n) = A \cdot \mathrm{Kor}(X_1, \ldots, X_n) \cdot {}^tA$$

ist. Nun müssen wir nur noch beachten, daß die Spaltenvektoren der Matrix A als die Vektoren einer Orthonormalbasis des $\mathbb{R}^n$ gewählt waren; der Eintrag an der Stelle ij der Matrix $A\,{}^tA$ ist also das Standardskalarprodukt des i-ten und des j-ten Vektors aus einer Orthonormalbasis und somit null für $i \neq j$ und eins für $i = j$. Daher ist $A \cdot {}^tA = E$, also ${}^tA^{-1}$ und somit auch

$$\mathrm{Kor}(Y_1, \ldots, Y_n) = A \cdot \mathrm{Kor}(X_1, \ldots, X_n) \cdot A^{-1}\,.$$

Damit ist $\mathrm{Kor}(Y_1, \ldots, Y_n)$ eine Diagonalmatrix, denn für jede Matrix $B \in \mathbb{R}^{n\times n}$ ist ABA^{-1} die Matrix B bezüglich der Basis aus den Spaltenvektoren vom A. Diese Basis besteht hier aber aus lauter Eigenvektoren der Korrelationsmatrix, die transformierte Matrix ist also eine Diagonalmatrix.

Unter den Annahmen unseres statistischen Modells können wir also jede Folge von Zufallsvariablen durch eine lineare Transformation in eine Folge unkorrelierter Zufallsvariablen überführen. Diese Transformation bezeichnet man, obwohl sie zuerst von HOTELLING vorgeschlagen wurde, als KARHUNEN-LOÈVE-Transformation.

HAROLD HOTELLING (1895–1973) war ein amerikanischer Statistiker und Ökonom; er lehrte an der Columbia University und der University of North Carolina. In einer 1933 veröffentlichten Arbeit im *Journal of Educational Psychology* schlug er erstmalig diese Transformation vor, die von Statistikern heute in Anlehnung an den Titel seiner Arbeit meist als *Hauptkomponentenanalyse* bezeichnet wird. In Europa erschien die Transformation fast gleichzeitig um 1947 *bzw.* 1948 in wahrscheinlichkeitstheoretischen Arbeiten des Finnen KARI KARHUNEN (* 1915) und des Franzosen MICHEL LOÈVE (1907–1979), nach denen sie in der technischen Literatur benannt wird.

Die Matrix A der linearen Transformation hängt nur von ρ ab und kann daher für gängige Werte von ρ vorberechnet werden; die KARHUNEN-LOÈVE-Transformation ist also einfach die Multiplikation mit einer bekannten Matrix.

e) Die diskrete Cosinus-Transformation

Für die Multiplikation zweier $n \times n$-Matrizen benötigt man allerdings n^3 Multiplikationen und noch einmal $n^2(n-1)$ Additionen; der Aufwand steigt mit großem n also sehr stark an. In der Praxis gibt man sich daher mit einen Kompromiß zufrieden und zerlegt eine Folge von Zufallszahlen in kurze Teilsequenzen, die bei eindimensionalen Folgen typischerweise die Länge 8 haben; dies ist beispielsweise der Standard bei Musik-CDs.

Allerdings wird weder bei Musik-CDs noch sonstwo die KARHUNEN-LOÈVE-Transformation wirklich angewandt. Der Grund liegt an der Struktur der Eigenvektoren der Korrelationsmatrix: Betrachten wir etwa als typisches Beispiel den $n = 8$; dann haben wir die Matrix

$$\mathrm{Cov}(X_1, \ldots, X_8) = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 & \rho^5 & \rho^6 & \rho^7 \\ \rho & 1 & \rho & \rho^2 & \rho^3 & \rho^4 & \rho^5 & \rho^6 \\ \rho^2 & \rho & 1 & \rho & \rho^2 & \rho^3 & \rho^4 & \rho^5 \\ \rho^3 & \rho^2 & \rho & 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^5 & \rho^4 & \rho^3 & \rho^2 & \rho & 1 & \rho & \rho^2 \\ \rho^6 & \rho^5 & \rho^4 & \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^7 & \rho^6 & \rho^5 & \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}.$$

Ihre Eigenwerte für $\rho = 0{,}95$ können zumindest näherungsweise berechnet werden, und auch die Eigenvektoren lassen sich bestimmen. Diese sollen hier jedoch nicht numerisch angegeben werden: Eine Folge von acht reellen Zahlen ist schließlich im allgemeinen eher unanschaulich. Stattdessen sind in den Abbildungen 70 bis 77 die Eigenvektoren *graphisch* dargestellt, wobei einem Vektor

$$(a_1, \ldots, a_8) \in \mathbb{R}^8$$

die acht Striche vom Punkt $(i, 0)$ bis (i, a_i) in der Ebenen entsprechen sollen. Zusätzlich ist in jedes dieser Diagramme noch eine der Kurven

$$y = \cos\left(\frac{(2x-1)(j-1)\pi}{16}\right)$$

für $j = 1, \ldots, 8$ eingezeichnet; wie man sieht, lassen sich die Komponenten der Eigenvektoren sehr gut durch diese Cosinuswerte annähern. Dies gilt nicht nur für den speziellen Wert $\rho = 0{,}95$, sondern für jeden Wert von ρ, der hinreichend nahe bei eins liegt.

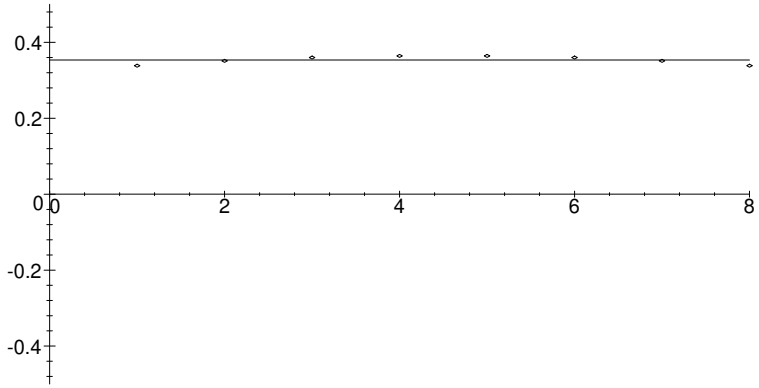


Abb. 70: Der erste Eigenvektor der Korrelationsmatrix

Aus diesem Grund arbeitet man in der Praxis lieber mit den Cosinuswerten; der Basiswechsel hin zur Basis aus den Cosinusvektoren bezeichnet man als *diskrete Cosinustransformation.* Ihr Hauptvorteil gegenüber der KARHUNEN-LOÈVE-Transformation ist, daß sie durch einen schnellen Algorithmus berechnet werden kann, der anstelle des Aufwands n^3 für eine Matrixmultiplikation nur den Aufwand $n^2 \log n$ hat. Für Einzelheiten sei auf die Vorlesung *Numerik I* verwiesen.

Die diskrete Cosinustransformation ist Teil fast aller gängiger Normen zur Bildkomprimierung: Sowohl der JPEG-Standard für Photographien, die Standards MPEG 1 und 2 für digitale (Unterhaltungs-)Videos als

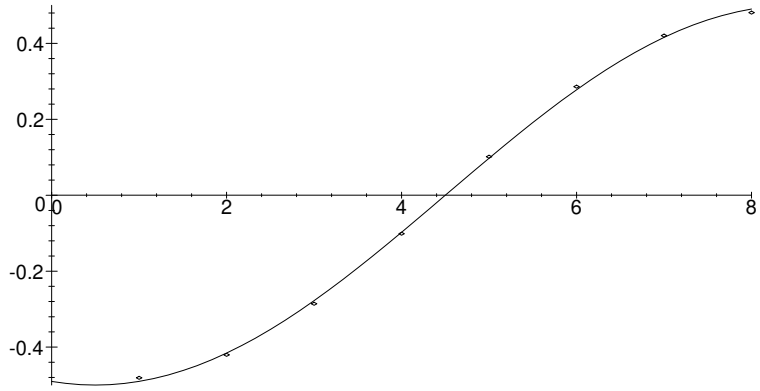


Abb. 71: Der zweite Eigenvektor der Korrelationsmatrix

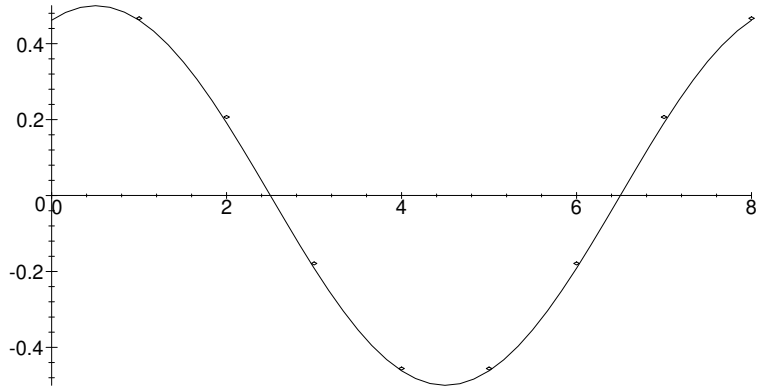


Abb. 72: Der dritte Eigenvektor der Korrelationsmatrix

auch der Standard CCITT H.261 für Videokonferenzen enthalten (neben anderen Bestandteilen) jeweils eine diskrete Cosinustransformation. Auch bei Audio-CDs ist sie ein Teil der Codierung.

Die Transformation allein ist natürlich noch keine Komprimierung: Schließlich haben wir nur einen Vektor in einer anderen Basis hingeschrieben, und die Anzahl der reellen Zahlen, die man zur Beschrei-

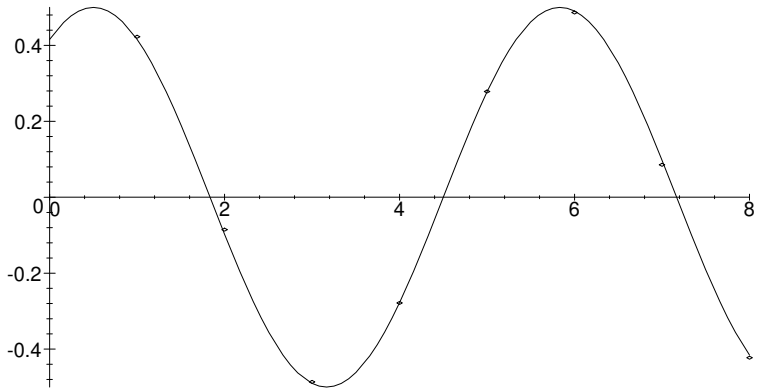


Abb. 73: Der vierte Eigenvektor der Korrelationsmatrix

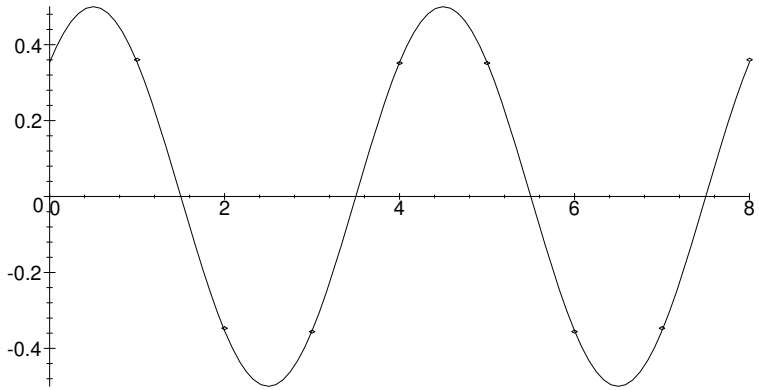


Abb. 74: Der fünfte Eigenvektor der Korrelationsmatrix

bung eines solchen Vektors benötigt, ist unabhängig von der Basis. Der wesentliche Vorteil der neuen Basis ist, daß man statistisch recht gute Aussagen über die Größe der Komponenten machen können. Hier wollen wir auf exakte statistische Berechnungen verzichten und stattdessen informell diskutieren, warum dies der Fall sein könnte.

Wie die Abbildungen der Basisvektoren zur KARHUNEN-LOÈVE-Trans-

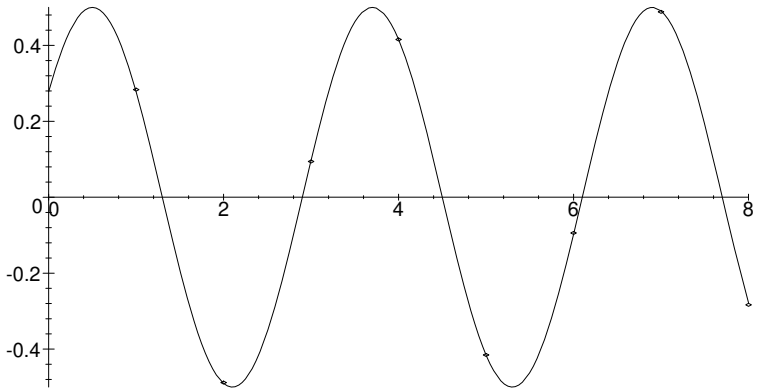


Abb. 75: Der sechste Eigenvektor der Korrelationsmatrix

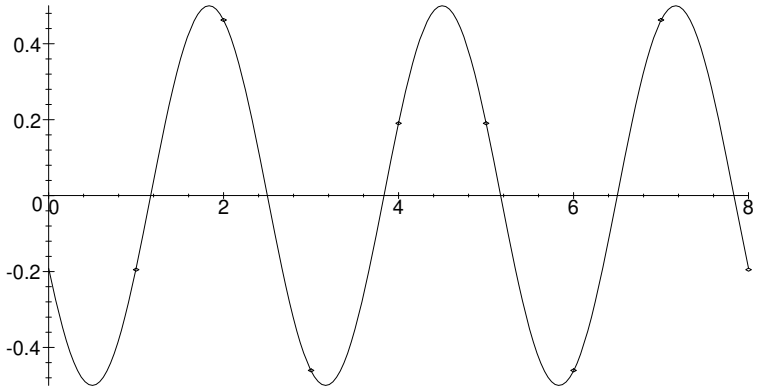


Abb. 76: Der siebte Eigenvektor der Korrelationsmatrix

formation und die Formeln für die Basisvektoren zur diskreten Cosinustransformation zeigen, werden die Basisvektoren, wenn man sie in der hier angegebenen Reihenfolge betrachtet, immer hochfrequenter. In einem hinreichend fein abgetasteten Bild oder Audiosignal erwarten wir, daß hochfrequente Schwankungen keine große Rolle spielen und somit die entsprechenden Basisvektoren nur kleine Koeffizienten haben

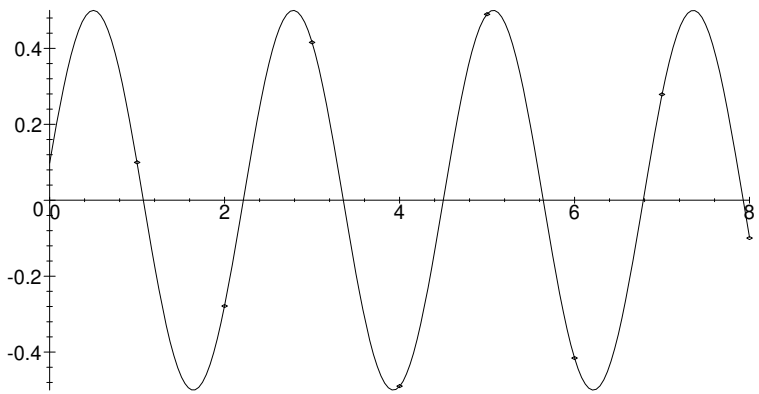


Abb. 77: Der achte Eigenvektor der Korrelationsmatrix

oder in vielen Fällen sogar gleich gar nicht auftreten. Dementsprechend genügt es, für die Übertragung dieser Koeffizienten nur wenige Bits bereitzustellen; bei nur geringen Abstrichen an die Qualität kann man auf gewisse Koeffizienten sogar ganz verzichten.

Ein Kompressionsverfahren wird daher, je nach Anspruch an die Qualität, entweder alle Koeffizienten des Signals in der neuen Basis übertragen und durch eine geeignete Darstellung der Daten dafür sorgen, daß Folgen von Nullen nur wenig Platz benötigen, oder aber es wird nur eine Auswahl der Koeffizienten übertragen und auch für diese jeweils festlegen, wie viele Bit dafür in Anspruch genommen werden. Diese Anzahl wird umso geringer sein, je höher die Frequenz des jeweiligen Basisvektors ist; bei einigen Verfahren wie etwa JPEG können die Anzahlen auch variablel in Abhängigkeit von einer Qualitätszahl gewählt werden.

Zum Schluß sei noch ganz kurz erwähnt, daß die KARHUNEN-LOÈVE-Transformation und damit (mit ganz geringen Abstrichen) auch die diskrete Cosinustransformation zwar die Korrelationsmatrix in optimaler Weise diagonalisieren, daß aber daraus nicht folgt, daß sie auch optimale Kompressionsverfahren liefern: Ausßer der Kovarianz gibt es noch weitere Quellen für Redundanz eines Bildes.

Ein gewisser Nachteil der Cosinustransformation ist außerdem, daß man für aprupte Übergänge, wie sie etwa bei Kanten immer wieder einmal auftauchen, die hochfrequenten Basisvektoren braucht, die dann aber nicht nur die Kante selbst beeinflussen, sondern das gesamte Quadrat, auf das die Transformation angewandt wird.

Eine bessere Möglichkeit wäre es daher, wenn man anstelle von Cosinusfunktionen Funktionen verwenden könnte, die sowohl im Zeit- als auch im Frequenzbereich lokalisiert sind. Solche Funktionen gibt es in der Tat, etwa die sogenannten *Wavelets*. Hierbei handelt es sich um schnell abklingende Wellen, und neuere Arbeiten deuten darauf hin, daß diese für gewisse Bildmodelle (die im Gegensatz zum hier betrachteten nicht mit Wahrscheinlichkeiten arbeiten) nicht zu weit vom Optimum entfernt sein sollten. Im Rahmen dieser Vorlesung ist es jedoch zeitlich weder möglich, auf diese Modelle einzugehen, noch ist an eine genauere Behandlung von Wavelets zu denken.

Einen allgemein verständlichen Überblick über Wavelets findet man etwa bei

BARBARA BURKE HUBBARD: Wavelets: Die Mathematik der kleinen Wellen, *Birkhäuser* 1997;

das zitierte Optimalitätsresultat ist beschrieben im Vortrag

STÉPHANE MALLAT: Applied Mathematics meets signal processing

auf dem Internationalen Mathematikerkongress 1998 in Berlin, nachzulesen in Band I der Proceedings, S. 319–338, oder unter http://www.mathematik.uni-bielefeld.de/documenta/xvol-icm/00/Mallat.MAN.html.

Eine für Technische Informatiker gut geeignete fundierte Einführung in diesen Themenkreis ist etwa

STÉPHANE MALLAT: A wavelet tour of signal processing, *Academic Press, 1998.*

E N D E

S C H Ö N E F E R I E N !