nichts mit Computerprogrammen zu tun.) Das wichtigste Verfahren zur Lösung solcher Aufgaben, der Simplex-Algorithmus, wird in der Vorlesung *Diskrete Mathematik A* 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 Minimum 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

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

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 $a \in D$ ein Extremum unter den Nebenbedingungen

$$g_1(a) \ge 0, \quad g_2(a) \ge 0, \quad \ldots, \quad g_r(a) \ge 0\,.$$

Dann sind die $r+1$ Vektoren

$$\nabla f(a), \quad \nabla g_1(a), \quad \nabla g_2(a), \quad \ldots, \quad \nabla g_r(a)$$

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 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

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

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

$$\nabla f(a), \quad \nabla g_1(a), \quad \nabla g_2(a), \quad \ldots, \quad \nabla g_r(a)$$

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

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

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.

g) Ausblick: Numerische Methoden

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.

Eine Möglichkeit dazu ist die sogenannte *Gradientenmethode:*

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

$$f(x+h) = f(x) + \langle \operatorname{grad} f(x), h\rangle + o(\|h\|)\,;$$

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

Nach der CAUCHY-SCHWARZschen Ungleichung ist, wenn wir mit der EUKLIDischen Norm arbeiten,

$$|\langle \operatorname{grad} f(x), h\rangle| \le \|\operatorname{grad} f(x)\| \cdot \|h\|\ ;$$

wir erhalten also die maximal mögliche Veränderung bei vorgegebener Länge von h genau dann, wenn h parallel zum Gradienten ist.

Damit bietet sich folgende Strategie an: Wir wählen irgendeinen Ausgangspunkt x_0 und berechnen dort den Gradienten $\nabla f(x_0)$. Falls er der Nullvektor ist. haben wir einen Kandidaten für ein Extremum gefunden, den wir mit noch zu entwickelnden Methoden weiter untersuchen müssen.

Andernfalls geben wir uns eine Länge ℓ_0 für den Vektor h vor, die von der Länge des Gradienten abhängen kann oder auch nicht, und setzen wir bei der Suche nach einem Maximum

$$h_0 = \frac{\ell_0}{\|\nabla f(x_0)\|}\nabla f(x_0)\,;$$

bei der Suche nach Minima nehmen wir das Negative davon.

Als nächstes betrachten wir den Punkt

$$x_1 \underset{\text{def}}{=} x_0 + h_0\,,$$

berechnen dort den Gradienten $\nabla f(x_1)$, setzen – so er nicht verschwindet – mit einer geeigneten Länge ℓ_1

$$h_1 = \pm\frac{\ell_1}{\|\nabla f(x_1)\|}\nabla f(x_1)$$

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

$$x_2 \underset{\text{def}}{=} x_1 + h_1$$

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 herum oszillieren.

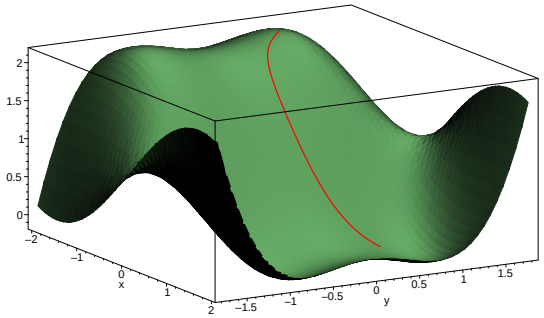
Theoretisch ist auch möglich, daß wir in einem Sattelpunkt landen, aber wenn man sich überlegt, wie die Gradienten in der Umgebung

eines Sattelpunktes aussehen, wird schnell klar, daß dies nur sehr selten passiert.

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

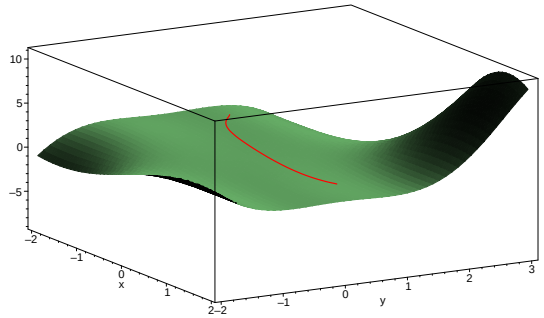
$$\begin{pmatrix} h_i \\ k_i \end{pmatrix} = 0{,}1 \cdot \nabla f(x_i, y_i)$$

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



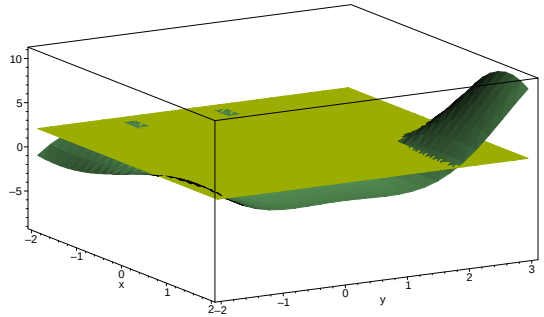
Eine Anwendung der Gradientenmethode

Die darauffolgende Abbildung allerdings 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 wir in der Nähe des (absoluten) Optimums starten, führen sie schnell und zuverlässig zum Ziel, ansonsten aber ist die Gefahr sehr groß, daß wir in einem nur lokalen Optimum steckenbleibt.



Der Weg aus der vorigen Abbildung aus einem weiteren Blickwinkel

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 x_M geführt, von dem aus es nicht mehr weiterkommt. Um zu sehen, ob $z_M = f(x_M)$ wirklich der kleinste Wert ist, den f im betrachteten Bereich annehmen kann, versuchen wir, eine weitere Lösung der Gleichung $f(x) = z_M$ zu finden.



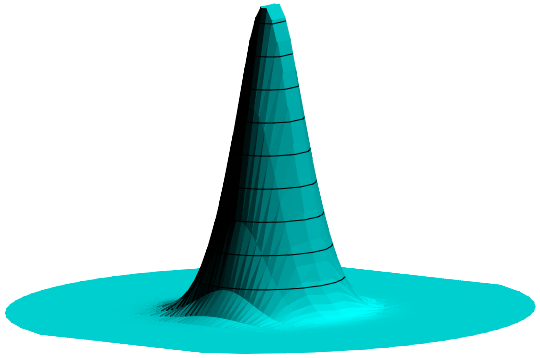
„Tunnelung“ für Maxima

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 x_M sein; deshalb sucht man tatsächlich nicht nach Lösungen der Gleichung

$f(x) = z_M$, sondern nach Lösungen einer leicht abgewandelten Gleichung der Form $\widetilde{f}(x) = z_M$, wobei $\widetilde{f}$ dadurch aus f entsteht, daß man die Funktionswerte in der unmittelbaren Umgebung von (x_M, y_M) stark anhebt, um so das dortige Minimum zum Verschwinden zu bringen. 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 der nächsten Abbildung zu sehen ist, und $\widetilde{f}(x,y) = f(x,y) + G(x,y)$ setzen.



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

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

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

so es eine gibt, liegt also nicht in der unmittelbaren Umgebung von 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.

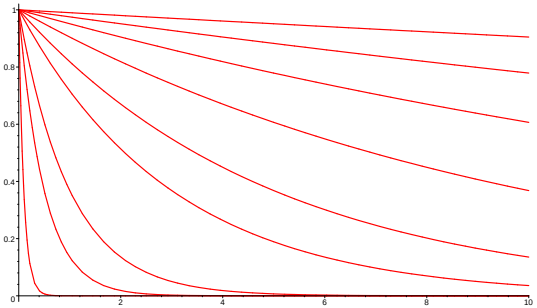
Im obigen Beispiel geht es nicht um ein Minimum, sondern um ein Maximum, da die Suche danach graphisch besser darstellbar ist. Also graben

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.)

Eine weitere Idee zur Vermeidung von Zwischenhochs kommt aus der Physik: 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. Die folgende Abbildung zeigt für verschiedene Werte von kT die Graphen der Funktion $e^{-E/kT}$; wie man sieht; erwartungsgemäß sind diese für große Werte von kT sehr flach, während sie für kleine Temperaturen rechts schnell gegen Null gehen. Bei der *simulierten Abkühlung* ahmt man dies nach, indem man mit einer hohen Temperatur startet und der Richtung, in der man weitergeht, einer dieser Temperatur entsprechende Freiheit läßt.



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

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.

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 h_j soll dabei analog zur BOLTZMANN-Verteilung festgelegt werden, d.h. wir ordnen ihm eine „Energie“ $E_j = \pm\big(f(x+h_j) - f(x,y)\big)$ zu (positiv bei der Suche nach einem Minimum, negativ bei der Suche nach einem Maximum) und die Wahrscheinlichkeit dafür, daß wir in Richtung 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 eine Zufallszahl $Z \in [0,\,1]$ und gehen in Richtung h_j, wenn

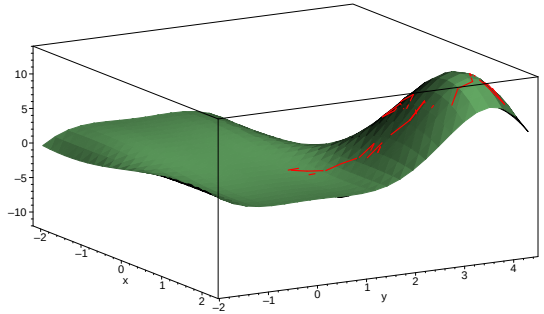
$$\sum_{\ell=1}^{j-1} p_\ell < Z \le \sum_{\ell=1}^{j} p_\ell$$

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

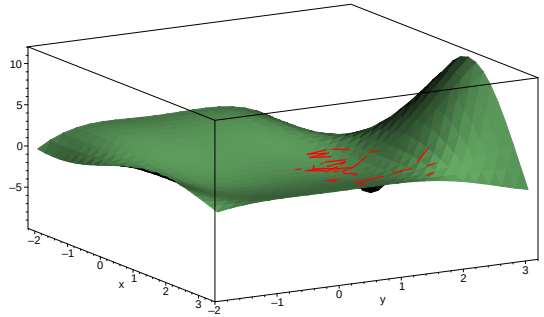
Bei hohen Temperaturen ist damit die Richtung fast vollständig zufallsbedingt gewählt, während in der Nähe des absoluten Nullpunkts praktisch nur noch die optimale Richtung eine Chance hat. Falls wir bei hoher Temperatur in einem Zwischenextremum landen, sorgt dies mit sehr hoher Wahrscheinlichkeit dafür,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

Am Ende wollen wir allerdings zum absoluten Optimum kommen, 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.

Die beiden folgenden Abbildungen zeigen, wie sich der Algorithmus bei zwei verschiedenen Folgen von Zufallszahlen verhalten bei einer Abkühlungsregel, die im r-ten Schritt mit Temperatur T_1/r arbeitet:



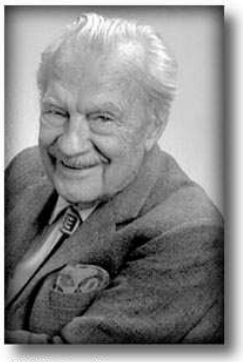
Der METROPOLIS-Algorithmus für obiges Problem



Ditto mit anderen Zufallszahlen

Im ersten Beispiel funktioniert alles sehr gut, im zweiten dagegen bleibt

die Kurve ziemlich lange im Tal hängen, kommt aber immerhin in eine gute Startposition für weitere Iterationen. Oft wird es ohnehin am besten sein, nach hinreichend vielen METROPOLIS-Schritten ein gewöhnliches Gradientenverfahren zu starten.


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 mathematischer Probleme; 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.

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, haben aber den Nachteil, daß sie kein Optimum garantieren können: Selbst wenn man eines erreicht hat, kann die Methode dies nicht erkennen. (Es gibt alternative numerische Methoden, die das können.)

Wie schon diese sehr kleine Auswahl von Optimierungsverfahren zeigt, ist nichtlineare Optimierung ein sehr weites Feld, von dem eine Grundvorlesung wie die *Analysis* nur einen winzigen 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 *Diskreten Mathematik* 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 Spezialvorlesungen verwiesen.

§3: Höhere Ableitungen

Im Eindimensionalen ist die Ableitung einer Funktion $f\colon D \to \mathbb{R}$ wieder eine Funktion $D \to R$; falls sie differenzierbar ist, bezeichnen wir ihre Ableitung als die zweite Ableitung von f, und so weiter. Für eine Funktion $f\colon D \to \mathbb{R}$ von n Veränderlichen ist die Ableitung jedoch eine Funktion $D \to \mathbb{R}^n$, also etwas komplizierteres als die Ausgangsfunktion. Dies macht den Umgang mit höheren Ableitungen im Mehrdimensionalen schwieriger. Wir beschränken uns daher zunächst auf die zweite Ableitung einer reellwertigen Funktion.

a) Die Hesse-Matrix

Wir lassen uns vom eindimensionalen Fall leiten: Die zweite Ableitung ist die Ableitung der Ableitung.

Die Ableitung einer differenzierbaren Funktion $f\colon D \to \mathbb{R}$ mit $D \subseteq \mathbb{R}^n$ ist der Gradient von f, also die Abbildung

$$\nabla f\colon D \to \mathbb{R}^n; \quad x \mapsto \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n} \right) .$$

Im vorigen Paragraphen haben wir auch solche Funktionen differenziert; ihre Ableitung war eine Matrix.

Definition: Wenn der Gradient von $f\colon D \to \mathbb{R}$ differenzierbar ist, bezeichnen wir seine JACOBI-Matrix als die HESSE-Matrix

$$H_f(x) = J_{\nabla f}(x)$$

von f im Punkt x.

Die HESSE-Matrix ist somit stets quadratisch, denn die Anzahl der Komponenten des Gradienten ist gleich der Anzahl der Variablen.

LUDWIG OTTO HESSE (1811–1874) wurde in Königsberg geboren und unterrichtete zunächst Physik und Chemie am dortigen Gymnasium. 1840 bekam er eine Stelle als Mathematiker an der dortigen Universität, von 1856 bis 1868 war er Professor in Heidelberg, danach in München. Aus der Schule ist er wohl vor allem durch die HESSEsche Normalenform der Ebenengleichung bekannt; der Schwerpunkt seiner Forschungen lag allerdings auf dem Gebiet der Invariantentheorie und der algebraischen Funktionen. Auch die HESSE-Matrix führte er 1842 in einer Arbeit über Invarianten von kubischen und biquadratischen Kurven ein.

Genau wie der Gradient bei gutartigen Funktionen durch die partiellen Ableitungen berechnet werden kann, sollte auch sie durch Differentiationsverfahren aus der Analysis einer Veränderlichen berechenbar sein. Das Hilfsmittel dazu sind die zweiten partiellen Ableitungen:

Für eine in ganz D partiell differenzierbare Funktion $f: D \to \mathbb{R}$ ist auch jede partielle Ableitung f_{x_i} wieder eine Funktion von D nach $\mathbb{R}$, und auch diese kann wieder partiell differenzierbar sein. Falls ja, bezeichnen wir die partielle Ableitung von f_{x_i} nach x_j als zweite partielle Ableitung

$$f_{x_i x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i} \underset{\text{def}}{=} \frac{\partial}{\partial x_j}\left(\frac{\partial f}{\partial x_i}\right)$$

von f nach x_i und x_j. Im Fall $i = j$ schreiben wir kurz

$$f_{x_i x_i} = \frac{\partial^2 f}{\partial x_i^2}\,.$$

Analog lassen sich auch höhere partielle Ableitungen einführen durch die Definition

$$f_{x_{i_1} x_{i_2} \ldots x_{i_k}} \underset{\text{def}}{=} \frac{\partial^k f}{\partial x_{i_k} \ldots \partial x_{i_2} \partial x_{i_1}} \underset{\text{def}}{=} \frac{\partial}{\partial x_{i_k}} \frac{\partial}{\partial x_{i_{k-1}}} \cdots \frac{\partial}{\partial x_{i_2}} \frac{\partial f}{\partial x_{i_1}}\,.$$

Wie wir oben gesehen haben, ist Gradient einer Funktion f gleich dem Vektor der partiellen Ableitungen, falls diese allesamt existieren und stetig sind; die JACOBI-Matrix ist der entsprechende Zeilenvektor. Falls auch die zweiten partiellen Ableitungen allesamt existieren und stetig

sind, zeigt dasselbe Lemma, daß deren Ableitungen die Zeilenvektoren

$$\left(\frac{\partial^2 f}{\partial x_i \partial x_1}, \ldots, \frac{\partial^2 f}{\partial x_i \partial x_n}\right)$$

sind, d.h.

Lemma: Falls alle ersten und zweiten partiellen Ableitungen von $f: D \to \mathbb{R}$ existieren und stetig sind, ist die HESSE-Matrix von f gleich der $n \times n$-Matrix mit Einträgen $\frac{\partial^2 f}{\partial x_i \partial x_j}$.

∎

Als erstes Beispiel können wir etwa die zweiten partiellen Ableitungen der Funktion

$$f(x, y) = x^4 + 2x^3y + 3x^2y^2 + 4xy^3 + 5y^4$$

berechnen: Die partielle Ableitung nach x ist

$$f_x(x, y) = 4x^3 + 6x^2y + 6xy^2 + 4y^3\,,$$

also ist

$$f_{xx}(x, y) = \frac{\partial^2 f}{\partial x^2}(x, y) = 12x^2 + 12xy + 6y^2 \quad \text{und}$$

$$f_{xy}(x, y) = \frac{\partial^2 f}{\partial y \partial x}(x, y) = 6x^2 + 12xy + 12y^2\,.$$

Entsprechend ist

$$f_y(x, y) = 2x^3 + 6x^2y + 12xy^2 + 20y^3\,,$$

also

$$f_{yx}(x, y) = \frac{\partial^2 f}{\partial x \partial y}(x, y) = 6x^2 + 12xy + 12y^2 \quad \text{und}$$

$$f_{yy}(x, y) = \frac{\partial^2 f}{\partial y^2}(x, y) = 6x^2 + 24xy + 60y^2\,.$$

Damit ist

$$H_f(x, y) = \begin{pmatrix} 12x^2 + 12xy + 6y^2 & 6x^2 + 12xy + 12y^2 \\ 6x^2 + 12xy + 12y^2 & 6x^2 + 24xy + 60y^2 \end{pmatrix}\,.$$

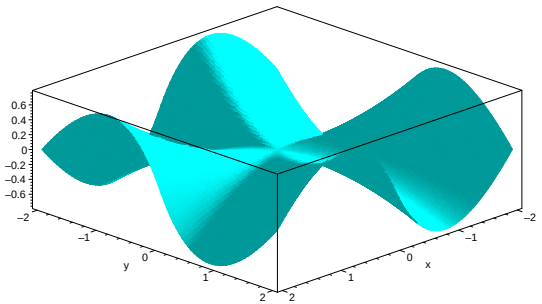
Zumindest in diesem Fall ist dies eine symmetrische Matrix, d.h.

$$f_{xy} = f_{yx}\,.$$

Diese Formel gilt, wie wir gleich sehen werden, *fast* immer; in der Tat galt sie für die Mathematiker des 18. Jahrhunderts wie NICOLAUS I. BERNOULLI, der 1719 darüber schrieb, LEONARD EULER (1730), JOSEPH-LOUIS LAGRANGE (1772) und viele andere als selbstverständlich. Erst im 19. Jahrhundert, als sich ein präziser Funktionsbegriff durchzusetzen begann, wurde erkannt, daß Voraussetzungen notwendig sind. Diese waren zu Beginn des Jahrhunderts zunächst unnötig stark; erst 1873 fand HERMANN AMANDUS SCHWARZ in seiner Arbeit *Über ein System voneinander unabhängiger Voraussetzungen zum Beweis des Satzes* $\frac{\partial}{\partial y}\left(\frac{\partial f(x,y)}{\partial x}\right) = \frac{\partial}{\partial x}\left(\frac{\partial f(x,y)}{\partial y}\right)$, was wirklich notwendig ist.

Als Gegenbeispiel betrachtet er die in der folgenden Abbildung dargestellte Funktion

$$f(x,y) = \begin{cases} y^2 \arctan\frac{x}{y} - x^2 \arctan\frac{y}{x} & \text{für } (x,y) \neq (0,0) \\ 0 & \text{für } (x,y = (0,0) \end{cases} .$$



Ein Gegenbeispiel zum Vertauschungssatz

Auch wenn es auf den ersten Blick nicht so aussieht, ist diese Funktion auch für $y = 0$ und $x \neq 0$ definiert: Der Bruch x/y ist dann zwar nicht definiert, aber da der Arkustangens nur Werte zwischen $-\pi/2$ und $\pi/2$ annimmt, existiert für jedes $x \neq 0$ der Grenzwert $\lim_{y\to 0} y^2 \arctan\frac{x}{y}$ und verschwindet. Entsprechend ist auch $\lim_{x\to 0} x^2 \arctan\frac{y}{x} = 0$ für alle $y \neq 0$, und wir wollen die obige Formel so interpretieren, daß $f(0,y) = f(x,0) = 0$ sein soll für alle $x, y \neq 0$ und, nach Definition, natürlich auch für $x = y = 0$. Man überlegt sich leicht, daß die so definierte Funktion stetig ist auf ganz $\mathbb{R}^2$.

Die Berechnung der ersten partiellen Ableitungen ist etwas umfangreich, jedoch läßt sich das Ergebnis deutlich vereinfachen: Wir erhalten

$$f_x(x,y) = \begin{cases} y - 2x \arctan\frac{y}{x} & \text{für } (x,y) \neq (0,0) \\ 0 & \text{für } (x,y) = (0,0) \end{cases}$$

und

$$f_y(x,y) = \begin{cases} -x + 2y \arctan\frac{x}{y} & \text{für } (x,y) \neq (0,0) \\ 0 & \text{für } (x,y) = (0,0) \end{cases},$$

wobei die Division durch Null beim Argument des Arkustangens wegen des Faktors vor dem Arkustangens wieder wie oben interpretiert werden soll, wir haben also

$$f_x(0,y) = y \quad \text{und} \quad f_y(x,0) = -x\,.$$

Diese Funktionen können wir problemlos differenzieren; wir erhalten

$$f_{xy}(0,y) = +1 \quad \text{und} \quad f_{yx}(x,0) = -1\,.$$

Im Nullpunkt ist somit $f_{xy}(0,0) = +1 \neq -1 = f_{yx}(0,0)$.

Für Punkte $(x,y) \neq (0,0)$ rechnet man leicht nach, daß

$$f_{xy}(x,y) = f_{yx}(x,y) = \frac{y^2 - x^2}{y^2 + x^2}$$

ist. Insbesondere ist daher für $x, y \neq 0$

$$f_{xy}(x,0) = -1, \quad f_{xy}(0,y) = +1 \quad \text{und} \quad f_{xy}(x,x) = 0\,;$$

f_{xy} nimmt also in jeder noch so kleinen Umgebung des Nullpunkts jeden der drei Werte $0, 1$ und -1 (und viele andere) an. Damit kann f_{xy} in $(0,0)$ nicht stetig sein, genauso wenig wie f_{yx}. Wie SCHWARZ erkannte, ist genau das die fehlende Voraussetzung für die Vertauschbarkeit der partiellen Ableitungen:

Schwarzsches Lemma: $f\colon D \to \mathbb{R}$ sei auf der offenen Menge $D \subseteq \mathbb{R}^n$ erklärt und sowohl die ersten partiellen Ableitungen f_{x_i} als auch die gemischten partiellen Ableitungen $f_{x_i x_j}$ seien stetig auf D. Dann ist

$$f_{x_i x_j}(x) = f_{x_j x_i}(x)$$

für alle $x \in D$ und alle i, j mit $1 \leq i, j \leq n$.

Beweis: Da bei der partiellen Differentiation alle Variablen außer einer als konstant betrachtet werden, können wir uns auf den Fall $n = 2$ beschränken: Wir interessieren uns nur für die beiden Variablen x_i und x_j,

die wir als x und y bezeichnen (wenn sie verschieden sind – andernfalls gibt es aber ohnehin nichts zu beweisen), und betrachten alle sonstigen x_k als konstant.

Für den Punkt (x, y) aus D wählen wir dann $h, k \in \mathbb{R}$ so, daß das Quadrat mit den vier Ecken

$$(x, y),\ (x+h, y),\ (x, y+k) \quad \text{und} \quad (x+h, y+k)$$

vollständig in D liegt; dies ist möglich, da wir D als offene Menge vorausgesetzt haben. Nach Voraussetzung existieren die partiellen Ableitungen f_x, f_y, f_{xy} sowie f_{yx} und sind stetig.

Nach Definition ist

$$f_{xy}(x,y) = \lim_{k\to 0} \frac{f_x(x,y+k) - f_x(x,y)}{k}$$

$$= \lim_{k\to 0} \frac{\lim\limits_{h\to 0} \dfrac{f(x+h,y+k) - f(x,y+k)}{h} - \lim\limits_{h\to 0} \dfrac{f(x+h,y) - f(x,y)}{h}}{k}\,.$$

Falls alle Grenzübergänge miteinander vertauschbar sind (was, wie wir im obigen Gegenbeispiel gesehen haben, keineswegs selbstverständlich ist), ist das ein Limes über den Ausdruck

$$\frac{f(x+h,y+k) - f(x,y+k) - f(x+h,y) + f(x,y)}{hk}$$

für $h, k \to 0$; es liegt also nahe, sich diesen Bruch genauer anzuschauen.

Für den Beweis wird es genügen, wenn wir uns auf den Fall $h = k$ beschränken; wir wollen den Ausdruck

$$D(h) = \frac{f(x+h,y+h) - f(x,y+h) - f(x+h,y) + f(x,y)}{h^2}$$

auf zwei Arten ausrechnen:

Zunächst fassen wir, wie oben, die beiden ersten und die beiden letzten Summanden zusammen: Mit der Abkürzung

$$g(y) = \frac{f(x+h,y) - f(x,y)}{h}$$

ist dann

$$D(h) = \frac{g(y+h) - g(y)}{h}\,.$$

Nach dem Mittelwertsatz der Differentialrechnung ist dieser Differenzenquotient gleich dem Differentialquotient $g'(\eta)$ für eine (von h abhängige) Zahl η zwischen y und $y+h$. Somit ist

$$D(h) = g'(\eta) = \frac{f_y(x+h,\eta) - f_y(x,\eta)}{h} = f_{yx}(\xi,\eta)$$

für ein η zwischen x und $x+h$, denn natürlich können wir auch auf diesen Differenzenquotienten den Mittelwertsatz anwenden.

Für die zweite Berechnung fassen wir in $D(h)$ den ersten und den dritten sowie den zweiten und den vierten Term zusammen. Mit der Abkürzung

$$\widetilde{g}(x) = \frac{f(x,y+h) - f(x,y)}{h}$$

ist dann dieses Mal

$$D(h) = \frac{\widetilde{g}(x+h) - \widetilde{g}(y)}{h}\,,$$

und nach dem Mittelwertsatz der Differentialrechnung gibt es dazu ein $\widetilde{\xi}$ zwischen x und $x+h$, so daß dies gleich $\widetilde{g}'(\widetilde{\xi})$ ist. Also ist

$$D(h) = \widetilde{g}'(\widetilde{\xi}) = \frac{f_x(\widetilde{\xi},y+h) - f_x(\widetilde{\xi},y)}{h} = f_{xy}(\widetilde{\xi},\widetilde{\eta})$$

für eine Zahl $\widetilde{\eta}$ zwischen y und $y+h$. Somit ist

$$D(h) = f_{yx}(\xi,\eta) = f_{xy}(\widetilde{\xi},\widetilde{\eta})\,.$$

Lassen wir nun h gegen Null gehen, konvergieren ξ und $\widetilde{\xi}$ gegen x und η wie auch $\widetilde{\eta}$ gegen y. Wegen der vorausgesetzten Stetigkeit der zweiten partiellen Ableitungen konvergiert daher $f_{yx}(\xi,\eta)$ gegen $f_{yx}(x,y)$ und $f_{xy}(\widetilde{\xi},\widetilde{\eta})$ gegen $f_{xy}(x,y)$, d.h. der Grenzwert existiert und

$$D(0) = f_{yx}(x,y) = f_{xy}(x,y)\,.$$

Damit ist das Lemma bewiesen. ∎

Der deutsche Mathematiker KARL HERMAN AMANDUS SCHWARZ (1843–1921) beschäftigte sich hauptsächlich mit konformen Abbildungen und mit sogenannten Minimalflächen, d.h. Flächen mit vorgegebenen Eigenschaften, deren Flächeninhalt minimal ist. Im Rahmen einer entsprechenden Arbeit für die WEIERSTRASS-Festschrift von 1885 (im Falle eines durch Doppelintegrale definierten Skalarprodukts) bewies er die CAUCHY-SCHWARZsche Ungleichung, die CAUCHY bereits 1821 für endlichdimensionale Vektorräume bewiesen hatte. SCHWARZ lehrte nacheinander in Halle, Zürich, Göttingen und Berlin.

Tatsächlich bewies SCHWARZ das obige Lemma (für $n = 2$) unter einer etwas schwächeren Voraussetzung: Es reicht, wenn *eine* der partiellen Ableitungen f_{xy} oder f_{yx} existiert und stetig ist. Am Beweis ändert sich wenig; falls etwa über die Ableitung f_{yx} nichts vorausgesetzt ist, muß man die Existenz aller damit zusammenhängenden Grenzwerte explizit durch Abschätzungen nachweisen und daraus nachträglich die Existenz und Stetigkeit von f_{yx} folgern. Ein Leser, der seine *Analysis I* noch nicht ganz vergessen hat, sollte dies auf etwa einer Seite tun können. Für Anwendungen ist die SCHWARZsche Formulierung etwas nützlicher als die obige, denn wenn man beispielsweise f_{xy} berechnet und seine Stetigkeit nachgewiesen hat, folgt automatisch, daß auch f_{yx} existiert und gleich f_{xy} ist. Für uns wird das keine sehr große Rolle spielen, denn bei den meisten uns interessierenden Funktionen wird die Existenz und Stetigkeit der partiellen Ableitungen klar sein; lediglich ihre Berechnung wird im allgemeinen mit Arbeit verbunden sein.

Ein analoger Satz zum SCHWARZschen Lemma gilt auch für höhere partielle Ableitungen; für k-fache Ableitungen müssen wir natürlich voraussetzen, daß alle partiellen Ableitungen bis zu den k-fachen existieren und stetig sind (wobei diese Voraussetzung wieder streng genommen nicht für alle k-fachen wirklich notwendig ist).

Definition: Für eine offene Teilmenge $D \subseteq \mathbb{R}^n$ bezeichne $\mathcal{C}^k(D, \mathbb{R})$ die Menge aller Funktionen $f\colon D \to \mathbb{R}$, deren sämtliche partielle Ableitungen bis zu den k-ten existieren und stetig sind. Für $k = 0$ bezeichnen wir mit $\mathcal{C}^0(D, \mathbb{R})$ einfach die Menge aller stetiger Funktionen $D \to \mathbb{R}$.

Man überlegt sich sofort, daß $\mathcal{C}^k(D, \mathbb{R})$ ein $\mathbb{R}$-Vektorraum ist, und es ist auch nicht schwer einzusehen, daß die Funktionen aus $\mathcal{C}^k(D, \mathbb{R})$ alle die Eigenschaften haben, die man sich bei der Betrachtung von k-ten Ableitungen wünscht:

Erstens ist die Berechnung einer k-ten partiellen Ableitung von der

Reihenfolge der partiellen Differentiationen unabhängig: Wie aus der *Linearen Algebra* bekannt sein sollte, kann jede Permutation als Produkt von Transpositionen geschrieben werden; es genügt also zu zeigen, daß man die Reihenfolge zweiter partieller Differentiationen vertauschen kann. Eine Transposition $(i_r\ i_{r+k})$ wiederum läßt sich gemäß

$$(i_r\ i_{r+k}) = (i_r\ i_{r+1}) \cdots (i_{r+k-1}\ i_{r+k})(i_{r+k-2}\ i_{r+k-1}) \cdots (i_r\ i_{r+1})$$

als Produkt von Transpositionen benachbarter Elemente schreiben, und für eine solche Transposition ist

$$\frac{\partial}{\partial x_{i_1}} \cdots \frac{\partial}{\partial x_{i_r}} \frac{\partial}{\partial x_{i_{r+1}}} \cdots \frac{\partial f}{\partial x_{i_k}} = \frac{\partial}{\partial x_{i_1}} \cdots \frac{\partial}{\partial x_{i_{r+1}}} \frac{\partial}{\partial x_{i_r}} \cdots \frac{\partial f}{\partial x_{i_k}}\,.$$

Für $f \in \mathcal{C}^k(D, \mathbb{R})$ ist sichergestellt, daß $\frac{\partial}{\partial x_{i_{r+2}}} \cdots \frac{\partial f}{\partial x_{i_k}}$ in $\mathcal{C}^2(D, \mathbb{R})$ liegt; nach obigem Lemma ist daher

$$\frac{\partial}{\partial x_{i_r}} \frac{\partial}{\partial x_{i_{r+1}}} \left(\frac{\partial}{\partial x_{i_{r+2}}} \cdots \frac{\partial f}{\partial x_{i_k}} \right) = \frac{\partial}{\partial x_{i_{r+1}}} \frac{\partial}{\partial x_{i_r}} \left(\frac{\partial}{\partial x_{i_{r+2}}} \cdots \frac{\partial f}{\partial x_{i_k}} \right)\,.$$

Differenziert man hier beide Seiten noch partiell nach x_{i_1} bis $x_{i_{r-1}}$, ändert dies natürlich nichts an der Gleichheit.

Zweitens besagt das vorletzte Lemma, daß eine Funktion $f \in \mathcal{C}^1(D, \mathbb{R})$ differenzierbar ist. Eine nahe liegende Verallgemeinerung des dortigen Beweises, bei der man anstelle von linearen Approximationen solche höherer Ordnung betrachtet, zeigt, daß eine Funktion $f \in \mathcal{C}^2(D, \mathbb{R})$ zweifach differenzierbar ist, und daß entsprechend eine Funktion f aus $\mathcal{C}^k(D, \mathbb{R})$ bis auf einen Fehler der Größenordnung $o\big(|h|^k\big)$ durch ein Polynom k-ten Grades approximiert werden kann. Wie das im einzelnen aussieht, wollen wir uns im nächsten Abschnitt genauer anschauen.

b) Taylor-Polynome

Im letzten Semester haben wir die höheren Ableitungen einer Funktion dazu benutzt, um sie nicht nur durch eine lineare Funktion, sondern durch ein Polynom höheren Grades anzunähern. Der wesentliche Satz über TAYLOR-Polynome war der folgende:

Satz: $f\colon (a, b) \to \mathbb{R}$ sei stetig und mindestens $(k + 1)$-fach stetig differenzierbar auf dem Intervall $(a, b) \subseteq \mathbb{R}$. Dann gilt für jedes x aus (a, b)

und jedes $h \in \mathbb{R}$ mit $x + h \in (a, b)$ die Formel

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2}f''(x) + \cdots + \frac{h^k}{k!}f^{(k)}(x) + R_{k+1}(x,h)$$
$$= \sum_{i=0}^{k} \frac{h^i}{i!}f^{(i)}(x) + R_{k+1}(x,h)$$

mit einem Restglied $R_{k+1} = O(h^{k+1})$. Dieses kann beispielsweise dargestellt werden als

$$R_{k+1}(x,h) = \frac{h^{k+1}}{(k+1)!}f^{(k+1)}(x+\eta h)$$

mit einer reellen Zahl η zwischen 0 und 1.

Um daraus einen auch für Funktionen mehrerer Veränderlichen nützlichen Satz zu machen, verwenden wir wieder Richtungsableitungen. Da wir eine Funktion $f\colon \mathbb{R}^n \to \mathbb{R}^m$ problemlos komponentenweise behandeln können, genügt es, den Fall $m = 1$ zu betrachten.

Die Richtungsableitung einer Funktion $f\colon \mathbb{R}^n \to \mathbb{R}$ in Richtung v ist nach Definition die Ableitung der Funktion einer Veränderlichen

$$g\colon \begin{cases} (-a,a) & \to \mathbb{R} \\ t & \mapsto f(x+tv) \end{cases}$$

mit geeignet gewähltem $a \in \mathbb{R}_+$ nach t für $t = 0$; wir können sie berechnen als Skalarprodukt des Gradienten mit dem Vektor v.

Natürlich können wir g nicht nur einmal ableiten; für $f \in C^{k+1}(D, \mathbb{R})$ existiert das TAYLOR-Polynom k-ten Grades und

$$g(t) = g(0) + tg'(0) + \frac{t^2}{2}g''(0) + \cdots + \frac{t^k}{k!}g^{(k)}(0) + O(t^{k+1}) .$$

Schreiben wir die Richtungsableitung in Richtung v wieder als

$$\partial_v f(x) = \langle \nabla f(x), v \rangle = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x) \cdot v_i ,$$

so wird dies zu

$$f(x+tv) = f(x) + t\partial_v f(x) + \cdots + \frac{t^k}{k!}\partial_v^k f(x) + O(t^{k+1}) .$$

Da wir $\|v\|$ hier als eine Konstante betrachten, können wir $O(t^{k+1}$ auch schreiben als $O\big((t\|v\|)^{k+1}$ und damit insbesondere $p\big((t\|v\|)^k$; speziell

für $t = 1$ erhalten wir die kompakte Schreibweise der TAYLOR-Formel, nämlich

$$f(x+v) = f(x) + \partial_v f(x) + \cdots + \frac{1}{k!}\partial_v^k f(x) + o\big(\|v\|^k\big) .$$

∂_v^k steht hierbei natürlich für die k-fache Anwendung des Operators ∂_v.

Wenn wir das konkret ausrechnen müssen, verschwindet die Kompaktheit allerdings schnell: Mit

$$v = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}$$

ist $\partial_v f(x) = \langle \nabla f(x), v \rangle = \sum_{i=1}^{n} v_i f_{x_i}(x)$ und dementsprechend

$$\partial_v^2 f(x) = \partial_v\big(\partial_v f(x)\big) = \partial_v \left(\sum_{i=1}^{n} v_i f_{x_i} \right) = \sum_{j=1}^{n} v_j \frac{\partial}{\partial x_j} \left(\sum_{i=1}^{n} v_i f_{x_i} \right)$$
$$= \sum_{i=1}^{n}\sum_{j=1}^{n} v_i v_j f_{x_i x_j} ,$$

wir haben also schon eine Summe mit n^2 Termen.

Mit Hilfe der linearen Algebra können wir diese Summe noch relativ kurz schreiben: Da der (i,j)-Eintrag der HESSE-Matrix $H_f(x)$ gerade gleich die partielle Ableitung $f_{x_i x_j}$ ist, rechnet man leicht nach, daß

$$v^T H_f(x) v = (v_1, \ldots, v_n) H_f(x) \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} = \sum_{i=1}^{n}\sum_{j=1}^{n} v_i v_j f_{x_i x_j}$$

ist. Für höhere Ableitungen geht so etwas nicht mehr: Völlig analog zur obigen Rechnung überzeugt man sich leicht davon, daß

$$\partial_v^3 f(x) = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} v_i v_j v_k f_{x_i x_j x_k}$$

ist, und diese Summe aus n^3 Summanden läßt sich nicht mehr mit Matrix-Vektor-Produkten darstellen: Die dritte Ableitung ist gegeben durch einen sogenannten *Tensor dritter Stufe,* d.h. ein dreidimensionales würfelförmiges Zahlenschema, und mit jeder weiteren Ableitung steigt die Dimension um eins an.

Für die Diskussion im nächsten Abschnitt reicht uns glücklicherweise die Formel

$$f(x+h) = f(x) + \langle \nabla f(x), h\rangle + \frac{1}{2} h^T H_f(x) h + o(\|h\|^2)\,.$$

c) Höhere Ableitungen und lokale Extrema

Wenn grad $f(x_0)$ verschwindeti und f mindestens zweimal stetig differenzierbar ist, hängt also das Verhalten von f in der Umgebung von x_0 ab von der quadratischen Form $h \mapsto h^T H_f(x_0)\, h$, wobei H_f nach dem SCHWARZschem Lemma eine symmetrische Matrix ist.

Definition: *a)* Eine symmetrische Matrix $A \in \mathbb{R}^{n\times n}$ heißt *positiv definit*, wenn für alle Vektoren $v \neq 0$ aus $\mathbb{R}^n$ gilt: $v^T A v > 0$.
b) A heißt *negativ definit*, wenn für alle $v \neq 0$ aus $\mathbb{R}^n$ gilt: $v^T A v < 0$.
c) A heißt *indefinit*, wenn es Vektoren $v, w \in \mathbb{R}^n$ gibt, so daß gilt: $v^T A v > 0$, aber $w^T A w < 0$.

Mit dieser Terminologie ist das folgende Lemma klar:

Satz: Wenn die differenzierbare Funktion $f \in \mathcal{C}^1(D, \mathbb{R})$ im Punkt $x_0 \in D$ 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 $x \in D$ verschwindet, gilt:
a) Falls die HESSE-Matrix $H_f(x_0)$ positiv definit ist, hat f im Punkt x_0 ein Minimum.
b) Falls $H_f(x_0)$ negativ definit ist, hat f im Punkt x_0 ein Maximum.
c) Falls $H_f(x_0)$ indefinit ist, hat f im Punkt x_0 einen Sattelpunkt.
∎

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 für Funktionen aus $\mathcal{C}^2(D, \mathbb{R})$ nach dem SCHWARZschen Lemma symmetrisch ist, und wie wir aus der Linearen Algebra wissen, ist jede symmetrische Matrix diagonalisierbar.

Für eine Diagonalmatrix A mit Einträgen $\lambda_1, \ldots, \lambda_n$ und einen Vektor v

mit Komponenten $v_1, \ldots, v_n$ wird obige quadratische Form zu

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

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. Als Zusammenfassung der obigen Diskussion können wir daher festhalten

Satz: Eine symmetrische reelle 2×2-Matrix A ist 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.

d) Lineare Regression

Als etwas umfangreicheres Beispiel für die Anwendung des obigen Satzes wollen wir ein klassisches Problem aus der Statistik betrachten, die Suche nach einer sogenannten *Ausgleichskurve* durch eine gegebene Punktmenge. Einige werden vielleicht aus dem Physikunterricht mit Ausgleichsgeraden vertraut sein: Wenn zwei physikalische Größen in einem linearen Zusammenhang stehen, werden trotzdem die zugehörigen Meßgrößen auf Grund die lineare Gleichung im allgemeinen nicht erfüllen: Messungen sind praktisch immer mit Fehlern behaftet. Bei wirtschafts- und sozialwissenschaftlichen Daten sind exakte Gesetze ohnehin die Ausnahme; hier kann man nur auf näherungsweise Gültigkeit hoffen.

Das Problem, zu vorgegebenen Daten einen möglichst guten näherungsweisen Zusammenhang zu finden, bezeichnet man als *Regression;* falls sich die Koeffizienten der gesuchten Funktion durch Lösen eines linearen Gleichungssystems aus den Daten berechnen lassen, redet man von *linearer Regression.* (Der Zusammenhang zwischen den Daten muß also *nicht* linear sein.)

Ausgangspunkt für die Lösung solcher Probleme ist die näherungsweise Lösung von linearen Gleichungssystemen: Wenn wir ein System aus N Gleichungen in m Unbekannten

$$\sum_{j=1}^{m} a_{ij} x_j = b_i\,, \qquad i = 1, \ldots, N$$

haben mit N wesentlich größer als m, können wir nur in seltenen Ausnahmefällen erwarten, daß es eine Lösung gibt. Wenn wir allerdings davon ausgehen, daß unsere Daten x_i ohnehin fehlerbehaftet sind, sollten auch gar nicht erst nach einer exakten Lösung suchen, sondern uns begnügen mit einem Tupel $(x_1, \ldots, x_m)$, für das die Gleichungen „einigermaßen“ erfüllt sind. Diese schwammige Formulierung läßt sich auf viele, nicht äquivalente Weisen präzisieren; die einfachste geht auf CARL FRIEDRICH GAUSS zurück: Wenn wir auf den linken Seiten aller Gleichungen Werte für die x_j einsetzen, erhalten wir einen Vektor aus $\mathbb{R}^N$; genauso bilden auch die rechten Seiten b_i einen Vektor aus $\mathbb{R}^N$. Falls das Gleichungssystem lösbar ist und wir eine Lösung einsetzen, stimmen die beiden Vektoren überein. Wenn es keine exakte Lösung gibt, können wir stattdessen fordern, daß die (EUKLIDische) Länge des Differenzvektors möglichst klein sein soll. Das ist äquivalent zur Forderung, daß das Quadrat der Länge möglichst klein sein soll, d.h.

$$\sum_{i=1}^{N} \left(\sum_{j=1}^{m} a_{ij} x_j - b_i \right)^2$$

soll im Punkt $(x_1, \ldots, x_m)$ ein Minimum annehmen. Da es sich hier um eine Summe von Quadraten handelt, wird dieser Ansatz auch als *Methode der kleinsten Quadrate* bezeichnet.

Als quadratische Funktion in den x_j ist die obige Summe natürlich beliebig oft differenzierbar; nach dem Satz aus dem vorigen Abschnitt muß