8. Mehrdimensionale Funktionen

Wer Grenzen überschreitet, versucht, in eine neue Dimension vorzustoßen.
*[Daniel Mühlemann, (*1959), Übersetzer und Aphoristiker]*

Einige Leute sollten nicht dünn werden, denn dadurch riskieren sie den Verlust ihrer einzigen Dimension.
*[Pavel Kosorin, (*1964), tschechischer Schriftsteller]*

8.1. Einleitung

8.1.1. Worum geht es?

Bisher hatten wir bei der Differentiation nur Funktionen *einer* Veränderlichen betrachtet. Bei den meisten Problemen der realen Welt treten aber mehrere Veränderliche auf:

- Eine **Fläche in der Computergrafik** kann durch $z = f(x,y)$ beschrieben werden
- **Zustandsvektor einer Wii** als Funktion der Zeit: Gestenerkennung, MCI
  - Masterprojekt Kristine Hein
- **Zustandsgleichung Gas**: Der Druck p ist Funktion von Temperatur T und Volumen V

$$p = p(T,V) = \frac{r \cdot T}{V}$$

- Der **Gewinn eines Unternehmens** ist eine Funktion der Umsätze aller seiner n Produkte und m Kostenstellen: $G = G(u_1,u_2, \dots ,u_n, k_1,\dots,k_m)$

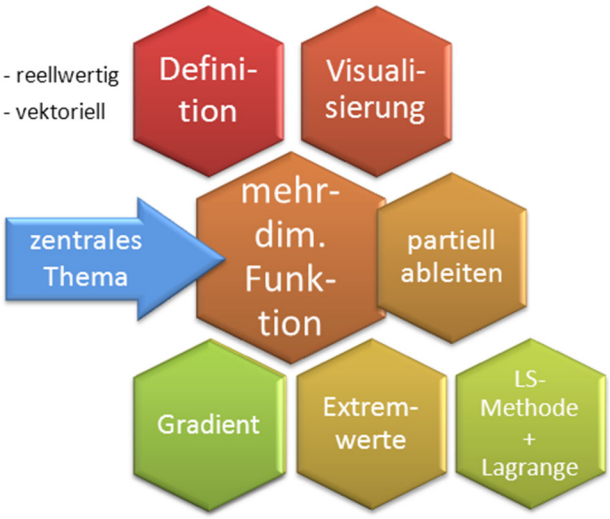
8.1.2. Warum InformatikerInnen mehrdimensionale Funktionen brauchen

Wir werden uns in diesem Kapitel mit der Definition und der Differentiation solcher Funktionen beschäftigen. Damit können wir dann folgende Probleme und Anwendungen lösen:

- Flächen und Trajektorien in **Computergraphik** und **Game Physics** darstellen.
- Wie differenziert man mehrdimensionale Funktionen? → **partielle** Differentiation
- **Modelloptimierung**: Wie findet man **Extremwerte**? Anwendungsfall: Welches ist die beste **Regressionsgerade** $y = ax+b$ für eine Menge von Punkten?
- Optimierung mit Nebenbedingungen: Die Methode der **Lagrange-Multiplikatoren**.

Da man bei den meisten Realwelt-Optimierungsaufgaben an mehreren (vielen) "Stellschrauben" drehen kann, sind solche Probleme von großer praktischer Bedeutung.

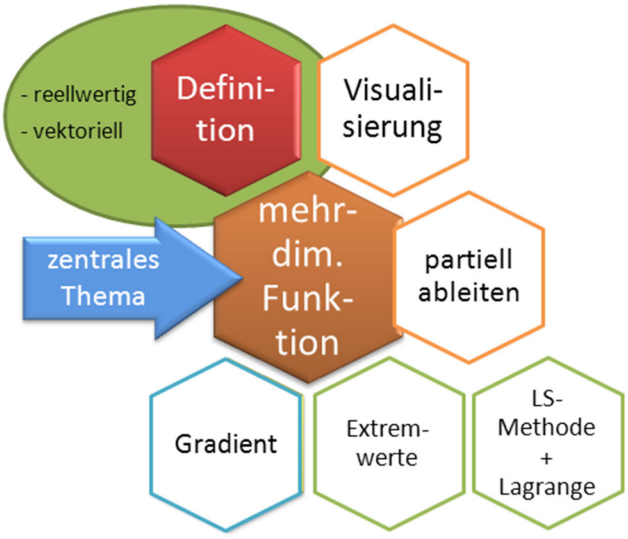
8.1.3. Welche Kompetenzen Sie erwerben



Nach Abschluss dieses Kapitels werden Sie wissen

- … , wie man mehrdimensionale Funktionen definiert
- …, wie man sie visualisiert (im Kopf, auf dem Papier und am Rechner)
- …, wie man durch (partielles) Ableiten Optimalwerte findet
- …, wie man ein Modell mit mehreren Parametern an Daten anpasst
- …, wozu ein Gradient gut ist
- …, wie man optimiert und dabei gleichzeitig Nebenbedingungen einhält (Lagrange)

8.2. Definition einer Funktion mehrerer Veränderlicher



Eine Funktion mehrerer Veränderlicher können wir uns gut als Java-Methode mit mehreren Parametern klarmachen. Nehmen wir die Zustandsgleichung für ein Gas:

$$p\colon \mathbb{R}\times\mathbb{R}\to\mathbb{R}, \qquad p(T,V)=\frac{rT}{V}$$

die als Java-Methode lauten würde

```
public double pressure(double temp, double volume) {
      static double r = 8.13;
      return r*temp/volume;
}
```

Allgemeiner können wir die Parameter einer Funktion mehrerer Veränderlicher in einem Vektor zusammenfassen, hier z.B.:

$$\vec{x} = \begin{pmatrix} T \\ V \end{pmatrix}$$

Es macht also mathematisch durchaus Sinn, sich mit Vektoren mit beliebig vielen Komponenten zu beschäftigen, auch wenn unsere Anschauung auf 3-dimensionale Räume beschränkt ist. Wir definieren den n-dimensionalen Raum

$$\mathbb{R}^n = \underbrace{\mathbb{R} \times \mathbb{R} \times \ldots \times \mathbb{R}}_{n-mal}$$

wie in Mathe 1 (Kap. 7.4 „Vektoren“):

Def D 8-1 n-dimensionaler Raum

Jedes Element der Menge $\mathbb{R}^n$wird als Punkt eines ndimensionalen Vektorraumes $\mathbb{R}^n$ bezeichnet. In der Regel wird ein solcher Punkt durch den Vektor $\vec{x}$ bezeichnet.

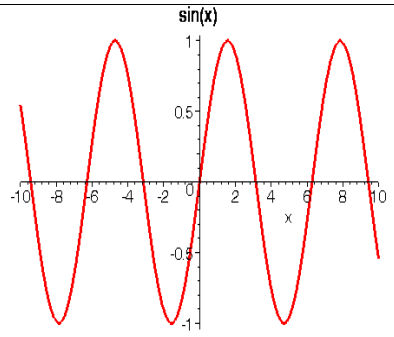
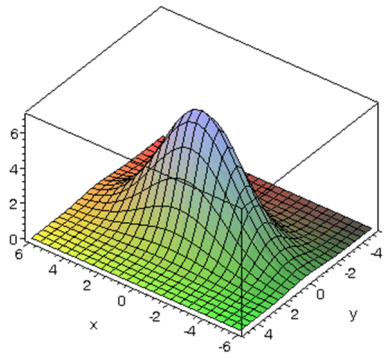
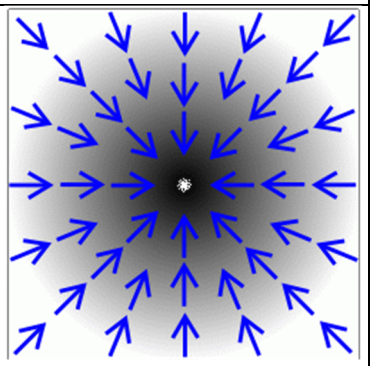
Def D 8-2 reellwertige Funktion mehrerer Veränderlicher

Eine reellwertige Funktion f ordnet jedem Punkt $x_1,..,x_n$ (bzw. Vektor $\vec{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$) aus einer zusammenhängenden Teilmenge **D** des $\mathbb{R}^n$ eindeutig einen reellen Wert $y \in \mathbb{R}$ zu, und man schreibt:

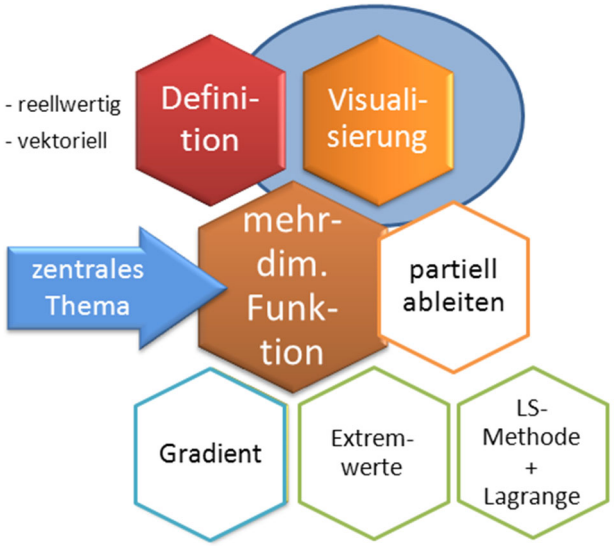
$$f: \; D \subseteq \mathbb{R}^n \to \mathbb{R} \qquad \text{mit} \quad y = f(x_1, x_2, \ldots, x_n)$$

Beispiel:
Die Temperatur auf der Erde ist eine Funktion der Längen- und Breitenkoordinate sowie der Höhe über dem Erdboden.

ANMERKUNG: Wir beschäftigen uns hier also mit **reell**wertigen Funktionen $f: \; \mathbb{R}^n \to \mathbb{R}$. In Kapitel 8.7 werden wir noch kurz auf **vektor**wertige Funktionen $\vec{f}: \; \mathbb{R}^n \to \mathbb{R}^m$ eingehen, die einen n-dim. Vektor auf einen m-dim. Vektor abbilden. Beispiele:

„normale“ Funktion	reellwertige Funktion	vektorwertige Funktion
$f: \mathbb{R} \to \mathbb{R}$	$f: \mathbb{R}^2 \to \mathbb{R}$	$f: \mathbb{R}^2 \to \mathbb{R}^2$
		
„Kurve“	(Hyper-) „Fläche“	„Pfeile“ (Windkarte)

8.3. Visualisierung einer Funktion mehrerer Veränderlicher



[Papula, Bd. 2, S. 272-286]

Ziel: Sich ein Bild von einer Funktion machen. Verbessern Sie Ihre Fertigkeiten zum „Bild machen“.

Wir fragen uns hier nur, welche Darstellungsformen grundsätzlich in Frage kommen und gehen auf „**Fläche im Raum**“ kurz ein. Wie kann man sich einen Überblick verschaffen, wie eine Funktion z=f(x,y) aussieht? [Methoden sammeln]

Mehr zu diesem Gebiet, der sog. **Visualisierung** (von Funktionen), können Sie auch im WPF „Computergrafik und Visualistik“ von Horst Stenzel erfahren.

8.3.1. Analytische Darstellung

Darstellung in Form einer Gleichung

		Eigenschaft	Vorteil
explizite Form	z = f(x,y)	nach z aufgelöst, nur ein z-Wert je (x,y)	leichter zu analysieren

implizite Form	F(x,y,z) = 0	nicht nach z aufgelöst	kann komplexere Flächen (mehrere z-Werte, Kugel)

Beispiele in Vorlesung.

Man verwendet die implizite Form, wenn eine Auflösung nach einer Variablen nicht möglich ist, oder, wenn sie zwar prinzipiell möglich, aber zu aufwendig oder mit unnötigen Schwierigkeiten verbunden ist. Die implizite Form kann komplizierte Flächen im R^3 darstellen, die explizite Form „kann" nur solche Flächen, die jedem (x,y) höchstens ein z zuordnen. [1]

Anmerkung: Jede explizite Form läßt sich mit

$$F(x,y,z) = f(x,y)-z$$

in die "kanonische" implizite Form bringen. Die umgekehrte Richtung kann dagegen schwierig sein.

Zum Spielen und für „schöne Forme(l)n" ist der ZEIT.de-Skulpturenwettbewerb **wärmstens** empfohlen !!

[Programme – Surfer zeigen, z.B. mit (x^2+y^2+z^2-1)*(x^3+y^3+z^3-1)]

8.3.2. Tabellarische Darstellung

Bevorzugte Darstellung für Tabellenkalkulationsprogramme

$z = f(x,y)$

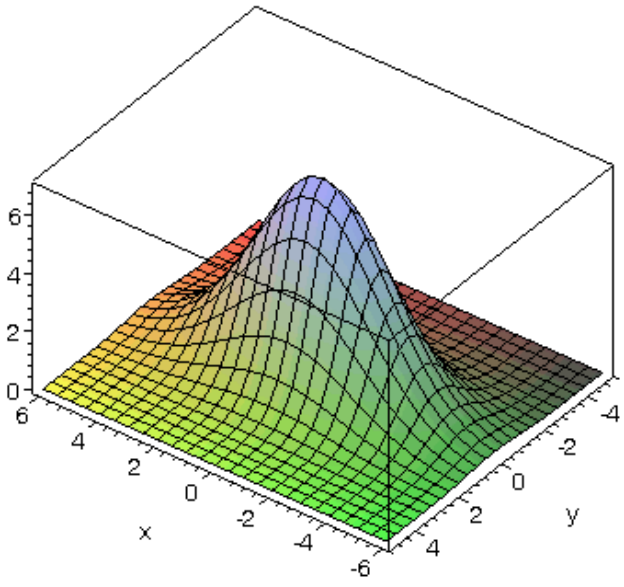
	y_1	y_2	....	y_k	...	y_n
x_1	z_{11}	z_{12}	...	z_{1k}	...	z_{1n}
....	...	...	...	...	...	...
x_m	z_{m1}	z_{m2}	...	z_{mk}	...	z_{mn}

8.3.3. Fläche im Raum

Bevorzugte Darstellung in Maple (**plot3d**)

Beispiel "Gaußglocke":

$$f(x,y) = z = 7\exp\left(-\frac{x^2+4y^2}{10}\right)$$



8.3.4. Schnittkurven: Höhenlinien, Kennlinienfeld

Eine wichtige alternative Darstellung kennt man aus Wanderkarten: Die 3. Dimension (Höhe) wird durch Höhenlinien abgebildet. Dort, wo die Höhenlinien dicht

[1] Beispiel zu implicitplot3d mit Maple-Befehl:

```
implicitplot3d((x/2)^2+y^2+z^2-10,x=-5..5,y=-5..5,z=-5..5);
```

zusammenliegen, herrscht eine hohe Steigung.

- **Höhenliniendiagramm** (engl: **contour plot**):
  - Horizontalschnitte („Baum fällen“): schneide das Funktionsgebirge in fester Höhe z=konstant auf und zeichne die Schnittkante „x gegen y“
- **Kennlinienfeld**:
  - Vertikalschnitte („Brotlaib“): für festes y=konstant zeichne „x gegen z“
  - (oder auch vertauscht: für festes x=konstant zeichne „y gegen z“)

Darstellung mit Maple:

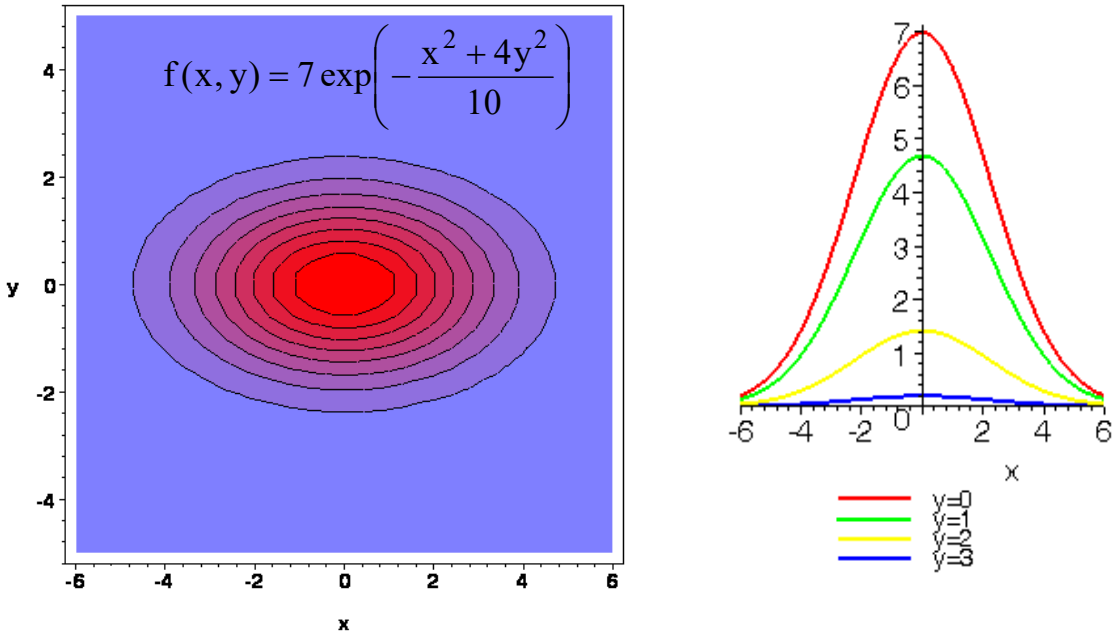


Abbildung 8-1: (a) Höhenliniendiagramm, (b) Kennlinienfeld[2]

Wie findet man die Höhenlinien für eine explizite Form? – Indem man die linke Seite als konstant festsetzt und nach y auflöst. Im Beispiel: [3]

$$f(x,y) = z = 7\,exp\left(-\frac{x^2+4y^2}{10}\right)$$

$$\Leftrightarrow \quad ln\frac{z}{7} = -\frac{x^2+4y^2}{10} \quad \Leftrightarrow \quad y = \pm\frac{1}{2}\sqrt{-10\,ln\frac{z}{7} - x^2}$$

Wenn sich die Gleichung nicht analytisch nach y auflösen läßt, geht es nur mühsamer: Numerisch ein Raster vieler Funktionswerte bestimmen und Punkte mit gleichen Werten verbinden. Oder durch numerische Nullstellenbestimmung.

Ein Kennlinienfeld läßt sich dagegen für die explizite Form immer leicht zeichnen: einfach verschiedene feste Werte für y einsetzen.

[2] Erzeugt durch folgende Maple-Befehle:

(a)
```
g:=(x,y)->7*exp(-(x^2+4*y^2)/10);
contourplot(g(x,y),x=-6..6,y=-5..5,filled=true,axes=boxed,
coloring=[COLOR(RGB,0.5,0.5,1),red],font=[HELVETICA,BOLD,12]);
```

(b)
```
plot([seq(g(x,y),y=0..3)],x=-6..6, legend=["y=0","y=1","y=2",
"y=3"],font=[HELVETICA,12],thickness=2);
```

[3] Unter der Wurzel steht tatsächlich nichts Negatives: ln(z/7)<0 ⇒ -10ln(z/7)>0. Weiter x^2<-10ln(z/7).

Übung: Leider ist gerade Ihr Laptop kaputt und Sie haben kein Maple zur Hand. Machen Sie sich trotzdem ein Bild von der Funktion $f(x,y) = x^2 e^y$, indem Sie handschriftlich ein Höhenliniendiagramm im Bereich 1,2,4,8 und ein Kennlinienfeld für y=0.5,1,2 erstellen.

Weitere Beispiele in Übungen!

8.3.5. Mehr als zwei Veränderliche

Die Anschauung versagt, die Funktion läßt sich nicht mehr als Ganzes zu erfassen. Zahlreiche Techniken sind entwickelt worden, um sich dennoch ein Bild von der Lage zu machen; Stichwort "Visualisierung von Daten". Basis-Methoden:

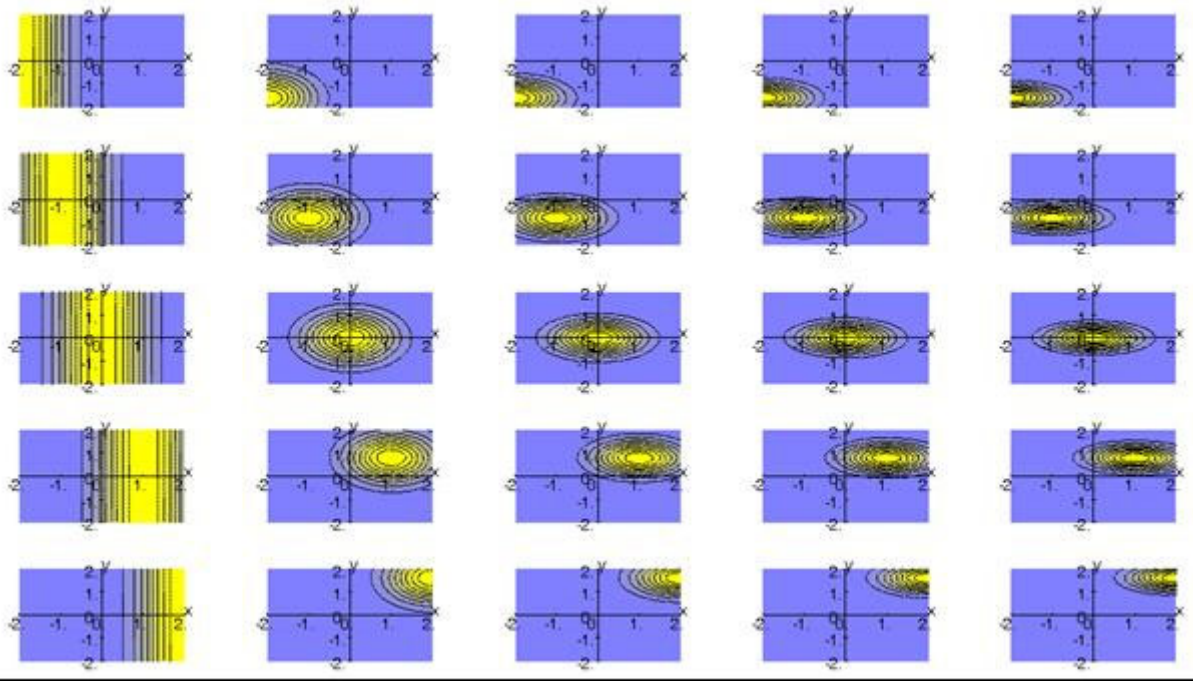
- Festhalten von n-2 Parametern und Betrachtung eines Schnitts, z.B. Höhenliniendiagramm in den restlichen beiden Parametern
- Anordnung vieler solcher Schnitte in rechteckigem Plot-Feld
- Animation, d.h. einer oder mehreren Variablen wird ein zeitlicher Verlauf zugeordnet, und man beobachtet die Änderung, die sich im Bild der anderen Variablen als Funktion der Zeit ergibt.
- u.v.a.m.

Beispiel 1: Anordnung in rechteckigem Plot-Feld:

Sei f: $\mathbb{R}^4 \rightarrow \mathbb{R}$ eine Funktion von 4 Veränderlichen x, y, v, w:

$$f(x,y,v,w) = exp(-(x-v)^2 - (w+2)(y-0.8v)^2)$$

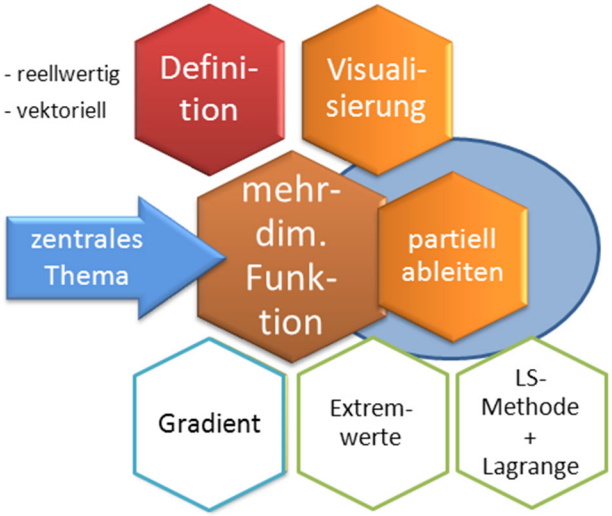
Wir stellen f durch ein Array von x-y-Höhenliniendiagrammen dar, in den Reihen läuft v von -2 bis 2, in den Spalten läuft w von -2 bis 2:

w = -2	w = -1	w = 0	w = 1	w = 2
				

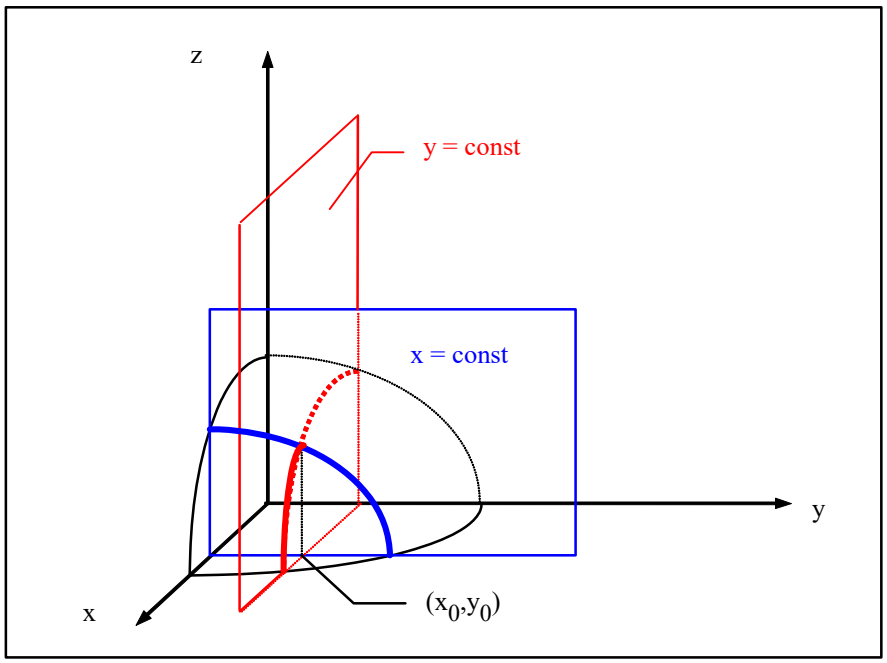
Welche Wirkung hat also der Parameter w, welche der Parameter v?

Beispiel 2: Wir stellen die gleiche Funktion $f(x, y, v, w)$ als Animation dar, wobei der Animationspfad längs der Diagonalen im v, w -Raum läuft, also von $v = w = -2$ bis $v = w = 0.5$.
Überlegen Sie: Wie wird die Animation in etwa aussehen? [Fragend entwickeln]
Lösung: s. plot3d.mws, Animation in Abschnitt "Mehr als zwei Veränderliche".

8.4. Partielle Ableitungen



Wie schon bei Funktionen einer Veränderlichen liefert der Begriff der Ableitung auch bei Funktionen mehrerer Veränderlichen den Schlüssel zur Analyse von Zusammenhängen. Die Ableitung einer Funktion mehrerer Veränderlicher wird mittels partieller Ableitungen auf den Fall eindimensionaler Funktionen zurückgeführt. Betrachten wir die Situation zunächst bei Funktionen zweier Veränderlicher (Skizze).



Im Punkt (x_0,y_0) sind die Schnittebenen x = const und y = const eingezeichnet. Innerhalb der jeweiligen Schnittebene liegt dann nur noch eine Funktion z = f(x) (für y = const) bzw. z = g(y) (für x = const) vor. Insbesondere bereitet die Bildung der Ableitung in diesen Fällen keine Schwierigkeiten. Dies führt uns zum Begriff der partiellen Ableitung.

Def D 8-3 Partielle Ableitung

Die partielle Ableitung 1. Ordnung der Funktion

$$y = f(x_1, x_2, \ldots, x_n)$$

nach der Variablen $\mathbf{X}_i$ ist durch den folgenden Grenzwert definiert:

$$\frac{\partial y}{\partial x_i}(\vec{x}) = \lim_{h \to 0} \frac{f(x_1, \ldots, x_{i-1}, x_i + h, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n)}{h}$$

Umgangssprachlich bedeutet dieser Grenzwert: Betrachte alle Variablen mit Ausnahme von $\mathbf{X}_i$ als Konstanten und bilde die übliche Ableitung nach der Variablen $\mathbf{X}_i$.

Anschaulich: Setze n-1 Variablen fest, dann passt die verbleibende Variable in eine „Schautafel“ (rotes oder blaues Rechteck in obiger Zeichnung), d.h. einen Graphen für eine „normale“ Funktion, den wir wie üblich ableiten können).

Weitere, allgemein übliche Symbole für partielle Ableitungen sind

$$\frac{\partial y}{\partial x_i}(\vec{x}) = y_{x_i}(\vec{x}) = \frac{\partial f}{\partial x_i}(\vec{x}) = f_{x_i}(\vec{x})$$

Wir werden im Folgenden meist die Schreibweise $f_{x_i}(\vec{x})$ benutzen, wenn keine Verwechslung mit dem Index (einer Vektorfunktion) zu befürchten ist.

Beispiel:

Die Zustandsgleichung eines idealen Gases lautet:

$$p(V,T) = \frac{RT}{V}$$
$$\frac{\partial p}{\partial V} = p_V = -\frac{RT}{V^2}$$
$$\frac{\partial p}{\partial T} = p_T = \frac{R}{V}$$

Anschaulich: Wenn ich das Volumen um einen kleinen Wert ∂V ändere, dann ändert sich der Druck um $\partial p = -\frac{RT}{V^2}\partial V$. D.h. bei Volumenvergrößerung sinkt der Druck, weil $-\frac{RT}{V^2} < 0$ (wenn man bei einer geschlossenen Luftpumpe den Kolben nach aussen zieht, gibt es eine rückziehende Kraft nach innen, weil der Druck innen niedriger ist als aussen), bei Temperaturerhöhung steigt der Druck.

Übung: Für $z(x,y) = 5xy - 2y + 3$ bestimme man z_x und z_y

Für $y(x_1,x_2,x_3) = x_1^2 x_2\, lnx_3 + \sqrt{x_1}\, sinx_2 + \frac{e^{x_3}}{x_1}$ bestimme man y_{x_1}, y_{x_2} und y_{x_3}

Wie diese Beispiele zeigen, sind die partiellen Ableitungen im Allgemeinen selbst wieder Funktionen sämtlicher, in der Ausgangsfunktion auftretender, Veränderlicher.

Sind alle partiellen Ableitungen stetig, so heißt die Funktion stetig differenzierbar.

Def D 8-4 Stetig differenzierbar

Ist eine Funktion an allen Stellen eines Gebietes G (einmal) differenzierbar und sind die partiellen Ableitungen stetig, so heißt die Funktion im Gebiet (einmal) stetig differenzierbar.

Analog: n-mal stetig differenzierbare Funktionen.

Die besondere Bedeutung dieser Definition liegt darin, dass stetig differenzierbare Funktionen in einer (kleinen) Umgebung eines Punktes durch den Funktionswert in diesem Punkt und sämtliche partiellen Ableitungen angenähert (approximiert) werden können (s. Kap. **Fehler! Verweisquelle konnte nicht gefunden werden.** "Linearisierung einer Funktion").

Def D 8-5 Partielle Ableitungen 2. Ordnung

Ist eine Funktion 2mal stetig differenzierbar, so kann jede partielle Ableitung 1. Ordnung selbst wieder nach allen Variablen differenziert werden. Hierdurch entstehen partielle Ableitungen 2. Ordnung.

Beispiel: Zu $y(x_1,x_2,\dots)$ ist *eine* Ableitung 2. Ordnung $y_{x_1x_2} = \left(y_{x_1}\right)_{x_2}$

Analog: Partielle Ableitungen n. Ordnung.

Übung: Bilden Sie $y(x_1,x_2,x_3) = x_1^2 x_2\, lnx_3 + \sqrt{x_1}\, sinx_2 + \frac{e^{x_3}}{x_1}$ (unter Verwendung der Ergebniss y_{x_1}, y_{x_2} und y_{x_3} aus voriger Übung) die 2. Ableitungen $y_{x_1x_2}$ und $y_{x_2x_1}$

Satz S 8-1 Satz von Schwarz

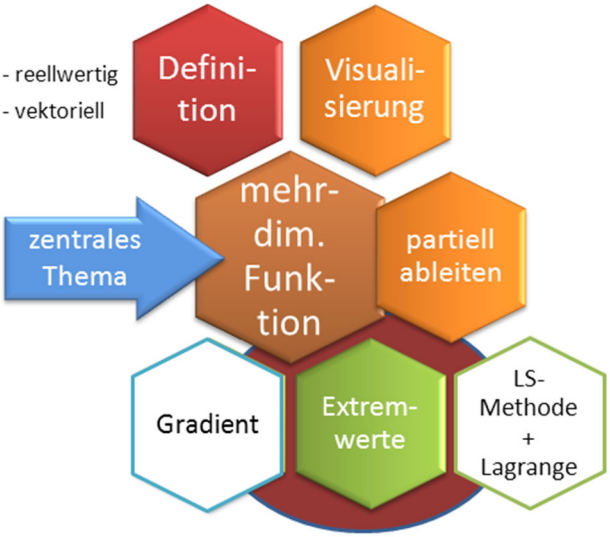
Ist eine Funktion von mehreren Veränderlichen k-mal stetig differenzierbar, so sind die gemischten Ableitungen k-ter Ordnung unabhängig von der Reihenfolge des Differenzierens.

Wie wir gerade gesehen haben, gilt für k = 2 für die Funktion $u(x, y, \dots)$:

$$u_{xy} = (u_x)_y = \left(u_y\right)_x = u_{yx}$$

Übung: Überprüfen Sie an der Funktion $f(x, y, z) = \frac{e^{ax}\, cos\, by}{zx}$ durch explizites Nachrechnen, dass gilt: $f_{xz} = f_{zx}$. Ist eine der Reihenfolgen ökonomischer?

8.5. Extremwerte



8.5.1. Lokale und globale Extremwerte

[Stingl, S. 361]

Analog zur Situation bei Funktionen mit einer Veränderlichen, lassen sich auch bei Funktionen mehrerer Veränderlicher die Begriffe lokales Minimum oder Maximum definieren. Notwendige Bedingungen ergeben sich aus den partiellen Ableitungen.

Def D 8-6 Relatives Minimum, relatives Maximum

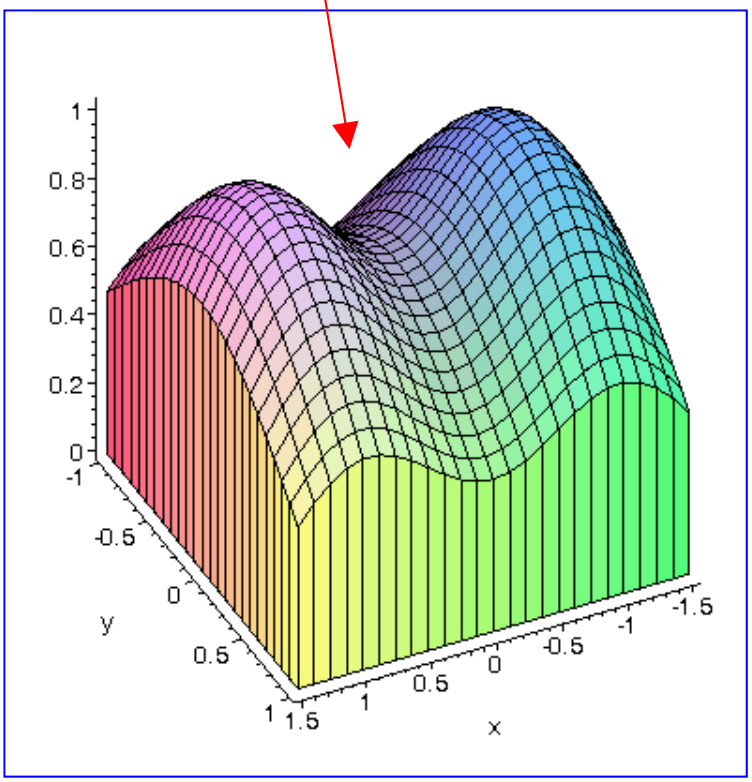
Eine Funktion $y = f(x_1, x_2, .., x_n)$ besitzt im Punkt $\vec{x}_0 = (x_{01}, x_{02}, \ldots, x_{0n})$ ein relatives Minimum, wenn in einer Umgebung von $\vec{x}_0$ stets:

$$f(x_1, \ldots, x_n) > f(x_{01}, \ldots, x_{0n}) \quad \text{für alle } \vec{\mathbf{x}} \neq \vec{x}_0$$

gilt. Ein relatives Maximum liegt vor, falls in einer Umgebung stets:

$$f(x_1, \ldots, x_n) < f(x_{01}, \ldots, x_{0n}) \quad \text{für alle } \vec{\mathbf{x}} \neq \vec{x}_0 \text{ gilt.}$$

Sattelpunkt



Ein Kriterium für das Vorliegen von Extremwerten liefert der nächste Satz:

Satz S 8-2 Stationärer Punkt

Ein Punkt $\vec{x}_0$ in dem sämtliche partiellen Ableitungen 1. Ordnung zu Null werden, $f_{x_1}(\vec{x}_0) = f_{x_2}(\vec{x}_0) = \cdots = f_{x_n}(\vec{x}_0) = 0$ heißt **stationärer Punkt**. Eine notwendige, aber im Allgemeinen nicht hinreichende Bedingung für einen Extremstelle ist, dass sie ein stationärer Punkt ist.

Bemerkungen:

1. Bei zwei Veränderlichen folgt der Satz aus der Forderung, dass ein Extremwert eine waagerechte Tangentialebene haben muß.
2. Wie bei Funktionen einer Veränderlichen ist die Bedingung aus **Satz S 8-2** nicht hinreichend, auch **Sattelpunkte** können waagerechte Tangentialebenen haben. (Wie jeder weiß, der schon mal Bergsteigen war, muss es zwischen zwei Gipfeln eines stetigen Gebirges sogar Sattelpunkte geben.)
   Beispiel (s. nebenstehendes Bild):
   $$z = f(x, y) = e^{-(x-1)^2 - y^2/2} + e^{-(x+1)^2 - y^2/2}$$
3. Die Angabe hinreichender Kriterien ist bei mehr als zwei Variablen schwierig. Für zwei Variablen erhält man als hinreichendes Kriterium:

Satz S 8-3 Hinreichendes Kriterium für lokale Extrema (2 Veränderliche)

Es sei $\Delta(x, y) = f_{xx}(x, y) f_{yy}(x, y) - \left[f_{xy}(x, y)\right]^2$ die Determinante der sog. Hesse-Matrix.

Eine Funktion $f(x, y) : D \to R$ besitzt an der Stelle (x_0, y_0) mit Sicherheit ein lokales Extremum, wenn die folgenden Bedingungen zugleich erfüllt sind:

1. $f_x(x_0, y_0) = 0$ und $f_y(x_0, y_0) = 0$ stationärer Punkt, notwendige Bedingung

und

2. $\Delta(x_0, y_0) > 0$

Im Fall $f_{xx}(x_0, y_0) < 0$ liegt ein lokales Maximum, im Fall $f_{xx}(x_0, y_0) > 0$ ein lokales Minimum vor.
Ist $\Delta(x_0, y_0) < 0$, so liegt kein Extremwert, sondern ein **Sattelpunkt** vor.

Satz S 8-4 Hinreichendes Kriterium für globale Extrema (2 Veränderliche)

Eine Funktion $f(x, y) : D \to \mathbb{R}$ besitzt an einem **stationären Punkt** (x_0, y_0) mit Sicherheit ein globales Extremum, wenn gilt

1. $\Delta(x, y) > 0$ und $f_{xx}(x, y) < 0$ für **alle (x,y)**$\in$D (globales Maximum)
– oder –
2. $\Delta(x, y) > 0$ und $f_{xx}(x, y) > 0$ für **alle (x,y)**$\in$D (globales Minimum)

Beispiele und Übungen in Vorlesung!

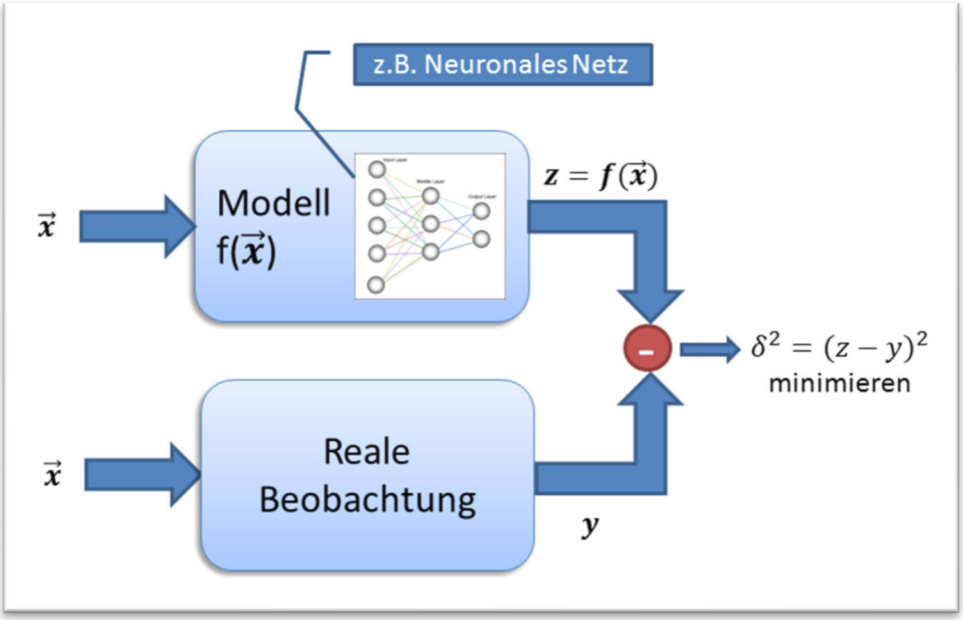
Ü Übung 1: Bestimmen Sie die lokalen Extrema von

$$W(x, y) = 6x + 3y^2 - 0.1x^2 - \frac{1}{4}y^4$$

Ü Übung 2: Gegeben sind n Punkte im zweidimensionalen Raum mit den Koordinaten $P_i = (x_i, y_i), i = 1, \ldots, n$. Für welchen Punkt $P = (x, y)$ ist die Summe der Abstandsquadrate zu den gegebenen Punkten P_i minimal?

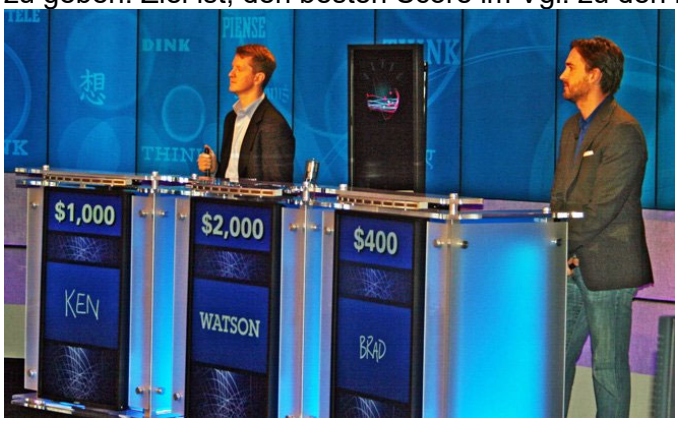
8.6. LS-Methode (Methode der kleinsten Quadrate)

8.6.1. Anwendungsfall: Modelle in der Informatik



Fast alle Computerprogramme sind Modelle der realen Welt. Die Modelle sollen (in bestimmten Aspekten) der realen Welt entsprechen. Beispiele:

- Crash-Test-Simulation in Automobilindustrie
- Zeitreihenvorhersage, Data Mining
- Neuronale Netze, Entscheidungsbäume (Lernen von Beispielen)
- Modell = NPC (Non-Person Character) in Computerspielen
- IBM Watson: !Jeopardy (lokale Kopie hier)): Hier modelliert der Computer Sprachwissen und Weltwissen, um auf möglichst viele Quizfragen die richtige Antwort zu geben. Ziel ist, den besten Score im Vgl. zu den Mitspielern zu erzielen.



Oft müssen die Modelle vor (oder während) der Inbetriebnahme optimiert (angepasst) werden, damit sie möglichst gut mit der realen Welt übereinstimmen. Diese **Modellanpassung** kann oft schwierig sein, weil ein Modell für verschiedene Fälle passen soll. Man spricht auch von **Parameter-Tuning**, Gegenstand unserer Forschungsprojekte FIWA/SOMA → www.gociop.de.

Ziel (s. Graphik):

Minimiere δ^2 = (Modell-Output – realer Output)^2 = $(f(\vec{x}) - y)^2$

Wenn mehrere Input-Output-Paare $\{(\vec{x}_i, y_i) |\ i = 1, \dots, n\}$ gegeben sind:

Minimiere $\delta^2 = \sum_{i=1}^{n} (f(\vec{x}_i) - y_i)^2$

Da der quadratische Fehler minimiert werden soll (wieso eigentlich quadratisch?), spricht man von der Methode der „kleinsten Quadrate“, engl. „least square“. Gebräuchliche Abkürzungen sind daher **KQ-Methode** oder **LS-Methode**.

Die LS-Methode ist eine der wichtigsten und gebräuchlichsten Methoden der mathematischen Optimierung.

8.6.2. Die LS-Methode für Geraden und die GLS-Methode

Wir werden in dieser Vorlesung nicht das IBM-Watson-Modell optimieren können (wer mehr über diese faszinierende KI-Challenge lesen will, s. http://www.stanford.edu/class/cs124/AIMagzine-DeepQA.pdf)

Wir nehmen uns als viel bescheideneres Modell zunächst „nur“ eine Gerade vor. Aber was Sie hier lernen, können Sie genausogut auf komplexere Modelle übertragen.

Im Praktikum werden Sie sich mit einem vereinfachten Neuronalen Netz beschäftigen.

Nun geht es also los mit der Geraden:

Gegeben seien n Meßpunkte (x_i,y_i), die nicht unbedingt auf einer Geraden liegen (Meßfehler, systematische Abweichungen). Wie findet man die Gerade, die am besten zu den Meßpunkten passt?

Anwendung: Praktikum Physik bei Prof. Koch, z.B. Messungen zu Hall-Effekt oder Kondensator.

Modell = Ausgleichsgerade (Regressionsgerade): $y = a + bx$

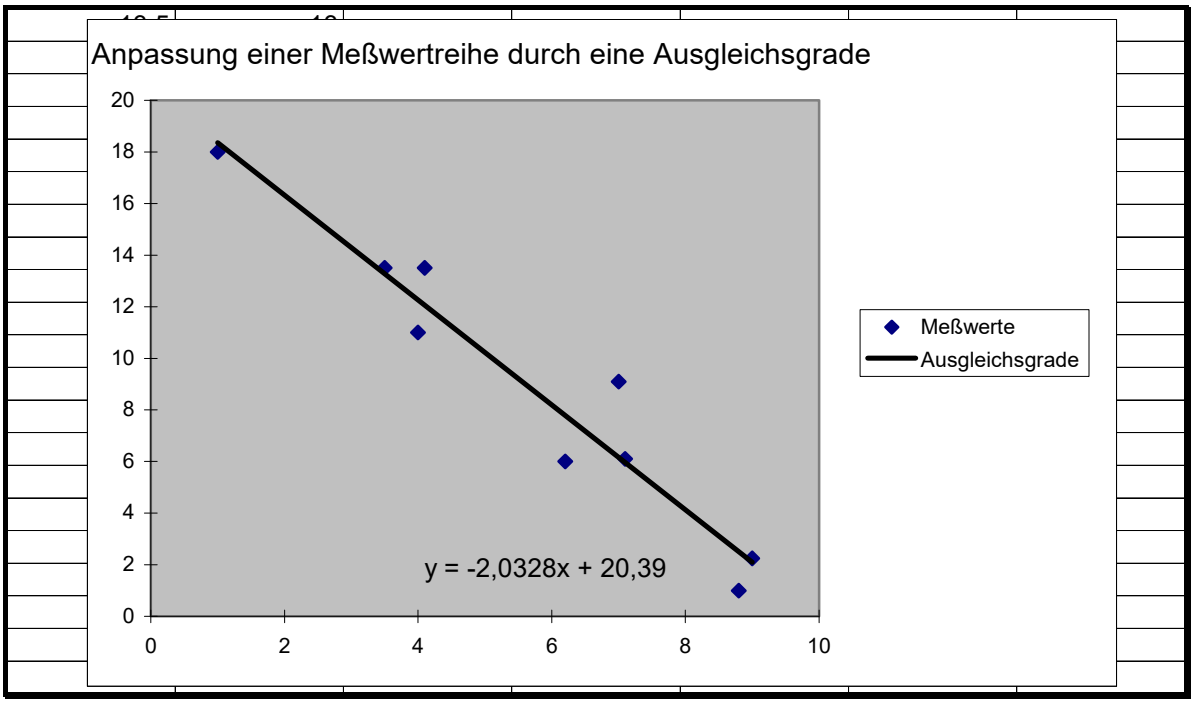
Welche Parameter sind zu optimieren?

Abweichung der Ausgleichsgeraden beim i-ten Datenpunkt: $\delta_i = a + bx_i - y_i$

Wir setzen voraus, dass **nicht alle x_i identisch** sind, denn dann hätten wir eine senkrechte Gerade, die wir nicht als Funktion beschreiben können.

Zu minimierende Funktion:

$$Z(a,b) = \sum_{i=1}^{n} \delta_i^2 = \sum_{i=1}^{n} (a + bx_i - y_i)^2$$



Wir setzen die partiellen Ableitungen gleich Null:

$$Z_a = 2\sum_{i=1}^{n}(a + bx_i - y_i) = 0$$
$$Z_b = 2\sum_{i=1}^{n}(a + bx_i - y_i)x_i = 0$$

Es ergibt sich ein lineares Gleichungssystem von zwei Gleichungen für die beiden Unbekannten a und b:

$$an + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$
$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

Wenn wir mit S_x, S_y, ... geeignete Abkürzungen für die Summen einführen, können wir das kürzer schreiben:

$$an + bS_x = S_y$$
$$aS_x + bS_{xx} = S_{xy}$$

Man multipliziert nun die 1. Gleichung mit S_x und die 2. Gleichung mit n durch, zieht voneinander ab und erhält:

$$b = \frac{nS_{xy} - S_x S_y}{nS_{xx} - (S_x)^2}$$

$$a = \frac{S_{xx}S_y - S_x S_{xy}}{nS_{xx} - (S_x)^2}$$

Übung: (a) Theoretisch könnte ja der Nenner in den obigen Formeln für "pathologische" Kombinationen der x_i auch mal Null werden. Können Sie zeigen, dass der Nenner immer ungleich Null ist? Hinweis: Es gilt die nützliche Identität

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)^2 \quad \text{mit Mittelwert} \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i$$

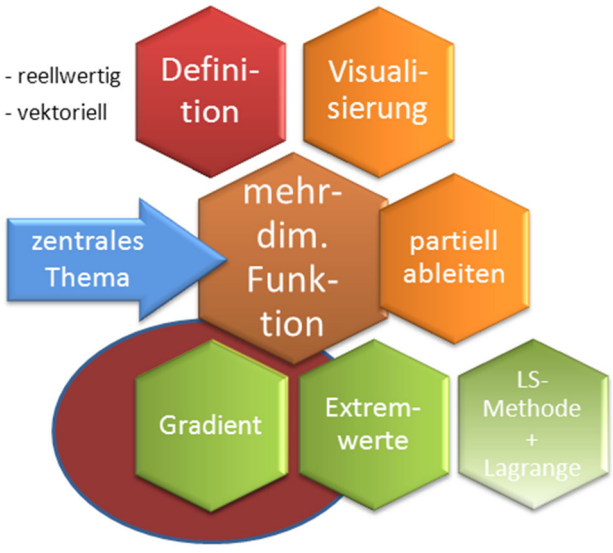
(b) Weisen Sie nach, dass es sich bei der Lösung {a,b} tatsächlich um ein Minimum handelt (s. **Satz S 8-3**)

Übung: Es muss nicht immer eine Gerade sein! Kombinationen von anderen "Basisfunktionen" gehen genauso gut.[4] Beispiel: In einem Behälter sind radioaktive Stoffe vom Typ A, der proportianal e^{-x} zerfällt und vom Typ B, der proportional e^{-2x} zerfällt. Durch Messungen soll ermittelt werden, wieviel vom Typ A, wieviel vom Typ B. Gegeben seien die Messpunkte:

x_i	0	1	2	3
y_i	4.1	1.3	0.4	0.3

Welches Modell $y = f(a,b) = ae^{-x} + be^{-2x}$ passt am besten zu diesen Daten? D.h. welche Parameter a, b minimieren die Summe der Abweichungsquadrate? Zeichnen Sie Ihr Modell und die Messpunkte in ein Diagramm!

8.7. Der Gradient



8.7.1. Vektorfunktionen

Die Königsetappe: Synthese von Linearer Algebra und Analysis: Wie kann ich einen Vektor ableiten?

[4] Den allgemeinen Fall beliebiger Basisfunktionen nennt man **GLS = "generalized least square"**.

Def D 8-7 Vektorfunktion

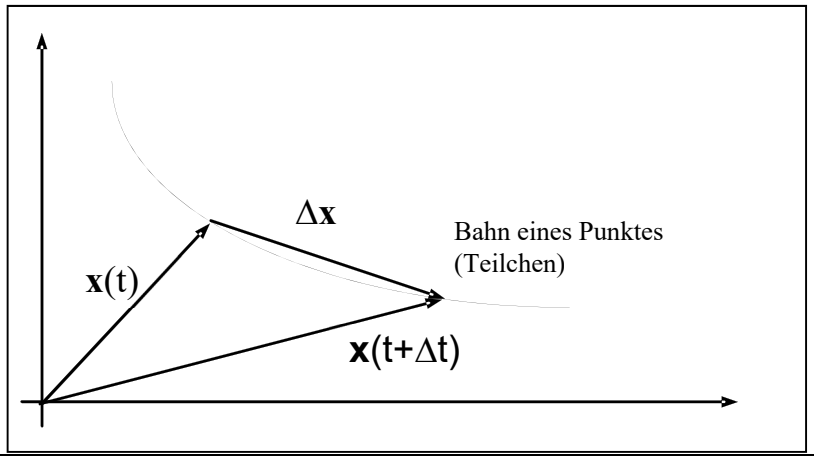
Sind die Koordinaten eines Vektors $\vec{x}$ als Funktionen einer skalaren Größe t (z.B. Zeit) gegeben, so liegt eine Vektorfunktion $\vec{x}: \mathbb{R} \to \mathbb{R}^3$ vor. In den Komponenten erhält man:

$$\vec{x}(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{pmatrix}$$

Bezeichnet t die Zeit und x_1, x_2, x_3 die Raumkoordinaten, so heißt $\vec{x}$ der Ortsvektor des Punktes $P(x_1, x_2, x_3)$.

Ist zusätzlich für den Parameter t ein Intervall $t_1 \le t \le t_2$ vorgegeben, so beschreibt die Menge aller Punkte $\{\vec{x}(t) | t_1 \le t \le t_2\}$ eine räumliche Kurve.

In Vorlesung: Raumkurve, mittlere Geschwindigkeit, Momentangeschwindigkeit.



Def D 8-8 Ableitung einer Vektorfunktion

Die 1. Ableitung der Vektorfunktion **x**(t) ist der Grenzwert:

$$\vec{v}(t) = \lim_{\Delta t \to 0} \frac{\vec{x}(t + \Delta t) - \vec{x}(t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{\Delta \vec{x}}{\Delta t} = \frac{d\vec{x}}{dt} \equiv \dot{\vec{x}}(t)$$

Der Vektor $\dot{\vec{x}}(t)$ ist der Tangentenvektor der Bahnkurve an der Stelle $\vec{x}(t)$.

Satz S 8-5

Die Koordinaten der Ableitung eines Vektors erhält man durch Differenzieren der Koordinaten des Vektors.

ANMERKUNGEN:

1. Die Definitionen gelten sinngemäß auch für m statt für 3 Koordinaten.
2. Die Koordinatenfunktionen eines Vektors können genauso gut Funktionen von n Veränderlichen sein (statt nur Funktionen von t). Dann haben wir die allgemeine vektorwertige Funktion $f: \mathbb{R}^n \to \mathbb{R}^m$ vor uns. Jede einzelne Koordinate ist eine Funktion von n Veränderlichen.

Wie man Funktionen von n Veränderlichen abzuleiten hat, ist Gegenstand des nächsten Kapitels.

8.7.2. Der Gradient: Wo bitte geht's nach oben?

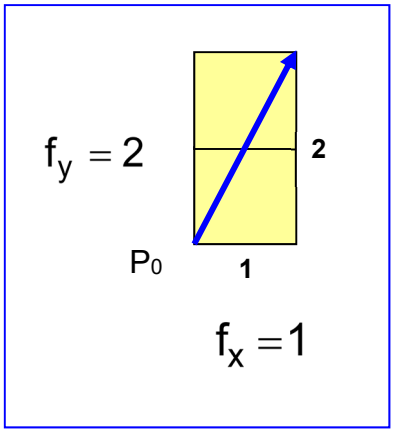
Stellen Sie sich vor, Sie stehen an einer Stelle $P_0=(x_0,y_0)$ im Funktionengebirge $f(x,y)$ und wollen wissen, wo geht es nach oben? Genauer: Wo geht's möglichst steil nach oben?

Mathematischer: Wenn ich einen (kleinen) Schritt der Länge ds mache, welche Richtung wähle ich? Das Problem: Es gibt unendlich viele Richtungen! Alle ausprobieren??

Zum Glück gibt es ein wesentlich einfacheres Rezept, das mit nur zwei (!) Messungen auskommt:

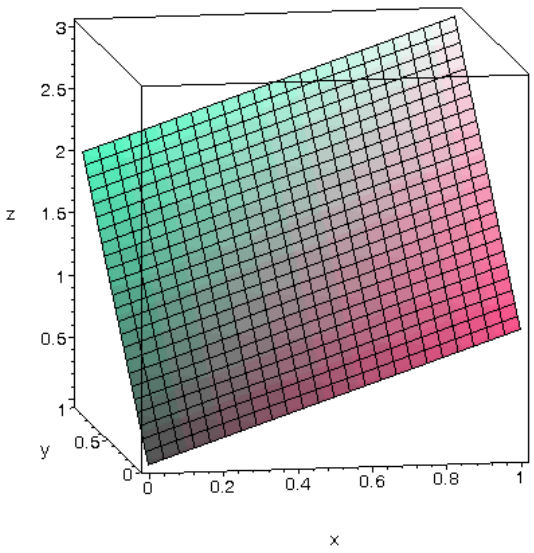
Rezept:

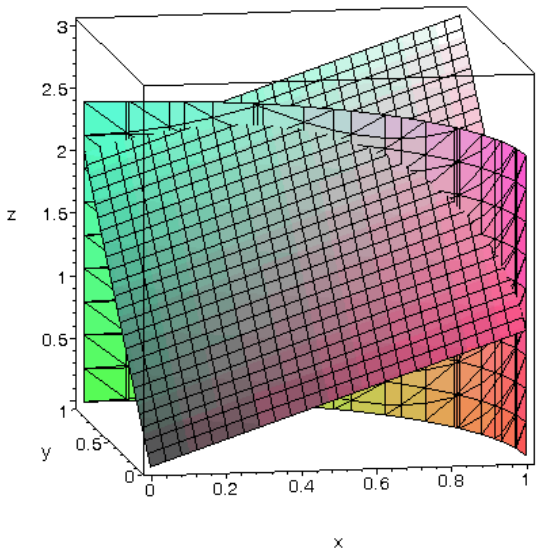
- Bilde die partiellen Ableitungen an der Stelle (x_0,y_0). Nehmen wir an, es sei $f_x(x_0,y_0)=1$ und $f_y(x_0,y_0)=2$. (Die Ableitungen sind die Steigungen, d.h. in der Nähe von (x_0,y_0) ist der Zuwachs in f je waagerechter Kästchenkante 1, der Zuwachs je senkrechter Kästchenkante ist 2.)



- Stecke die Zahlen in einen Vektor und marschiere in die Richtung, die der Vektor angibt. Also hier: 1mm in x-Richtung und 2 mm in y-Richtung.
  - Vektor $\binom{1}{2}$, Strecke: $\sqrt{1^2+2^2}=\sqrt{5}$ mm.
  - Zuwachs: 1 + 2 +2 = 5, also
  - Zuwachs/mm = $\frac{5}{\sqrt{5}}=\sqrt{5}=2.23$
  - Das ist ein höherer Zuwachs/mm als in x-Richtung alleine (1) oder in y-Richtung alleine (2)
  - Keine andere Richtung bringt einen höheren Zuwachs/mm. Probieren Sie's aus!
  - Der Vektor $\binom{1}{2}$ heißt Gradient an der Stelle (x_0,y_0).

Ausführlich kommentiertes Beispiel: plotGrad.mws. Hier 2 Abbildungen daraus:



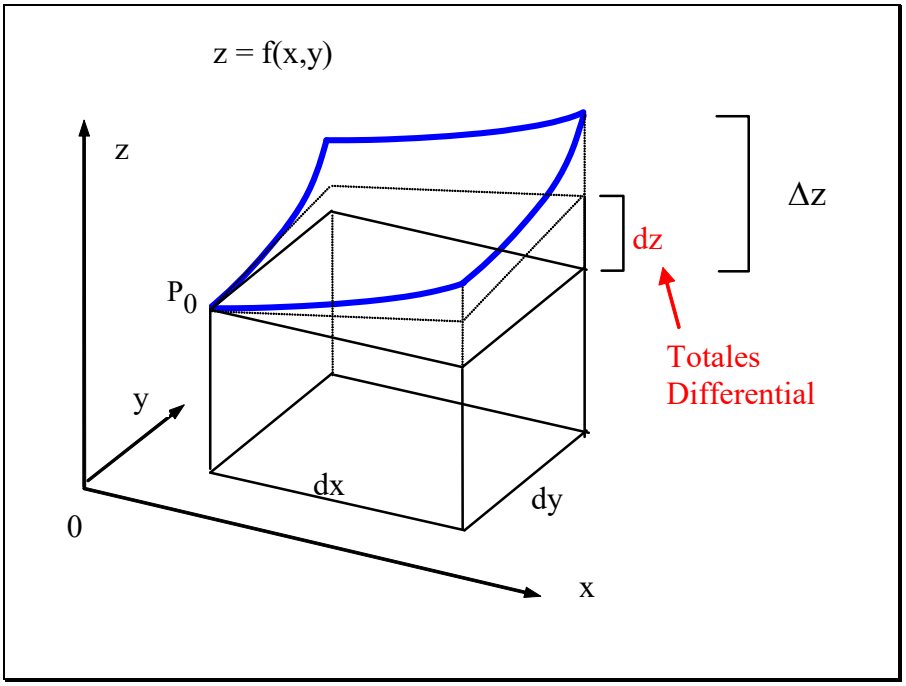


Wer's genauer verstehen will: Totales Differential, Gradient.

8.7.3. Totales Differential

[evtl. nur Def. bringen, Rest im Selbststudium]

Betrachten wir eine Funktion $f(x,y)$ in zwei Veränderlichen an der Stelle $P_0=(x_0,y_0)$:



Wenn ich von P_0 ein Stück (dx,dy) weitergehe, dann ist:

Totales Differential dz	**=**	**Zuwachs der Tangentialebene in P_0, wenn in allen Koordinaten um (dx,dy) weitergegangen wird**
Funktionsänderung Δz	=	Zuwachs der Funktion, wenn man um denselben Vektor (dx,dy,…) weitergeht

Als Formel:

$$\mathrm{dz} = f_x(x_0,y_0)\mathrm{dx} + f_y(x_0,y_0)\mathrm{dy}$$
$$\Delta z = f(x_0+dx, y_0+dy) - f(x_0,y_0)$$

Def D 8-9 Totales Differential (2 Veränderliche)

Das totale Differential dz einer Funktion $z = f(x,y)$ im Punkt (x_0,y_0) ist definiert durch:

$$dz = f_x(x_0,y_0)\mathrm{dx} + f_y(x_0,y_0)\mathrm{dy}$$

Es gilt: $dz \approx \Delta z$ wenn dx, dy hinreichend klein sind (s. Zeichnung).

Die Tangentialebene im Punkt (x_0,y_0) ist gegeben durch:

$$Z(x,y) = f(x_0,y_0) + f_x(x_0,y_0)(\mathrm{x}\text{-}\mathrm{x}_0) + f_y(x_0,y_0)(\mathrm{y}\text{-}\mathrm{y}_0)$$

Zum Beweis der Tangentialebenen-Gleichung setzt man in allgemeiner Form

$$Z = a + b(x-x_0) + c(y-y_0)$$

an und führt einen Koeffizientenvergleich durch.

Bei Funktionen von n Variablen erweitert man dies ganz analog:

Def D 8-10 Totales Differential (n Veränderliche)

Das totale Differential dz einer Funktion $z = f(x_1, x_2, \ldots, x_n) = f(\vec{x})$ wird definiert durch:

$$dz = f_{x_1} dx_1 + f_{x_2} dx_2 + \ldots + f_{x_n} dx_n$$

dabei sind alle partiellen Ableitungen im betreffenden Punkt zu nehmen.

Es gilt auch hier: $dz \approx \Delta z = f(\vec{x} + \boldsymbol{d\vec{x}}) - f(\vec{x})$, wenn $\boldsymbol{d\vec{x}} = (dx_1, dx_2, \ldots, dx_n)$ hinreichend klein ist.

Beispiel:

$1)\, z = 2x + y^2 \quad x = 3, y = 5, dx = 0.3, dy = 0.2$

$z_1 = f(x,y) = f(3,5) = 31$

$z_2 = f(x + dx, y + dy) = f(3.3, 5.2) = 33.64$

$\Delta z = 2.64$

$dz = 2dx + 2ydy = 2 \cdot 0.3 + 2 \cdot 5 \cdot 0.2 = 2.6$

also gilt tatsächlich: $\Delta z \approx dz$

8.7.4. Der Gradient: Woher weht der Wind?

[Stingl, S. 343 und 353]

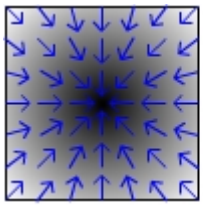
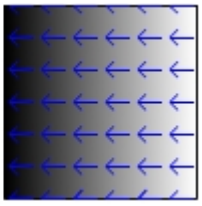
*lat. Verb: **gradior**, gressus sum = **schreiten***
*lat. Substantiv **gradus** = **Schritt**, Standpunkt, Stufe (vgl. graduell)*
(hängt also eng mit unserem Bild vom Ausschreiten im Funktionengebirge zusammen)

Def D 8-11 Gradient

Der **Gradient grad f** einer Funktion z= f(x_1, x_2,..., x_n) ist eine **Vektorfunktion** (s. **Def D 8-7**), die aus den partiellen Ableitungen besteht. Wertet man den Gradient an einer bestimmten Stelle P_0 = (x_{10}, x_{20},..., x_{n0}) aus, so entsteht **(grad f)(P_0)**, ein einfacher **Vektor**:

$$grad\, f = \begin{pmatrix} f_{x_1} \\ \vdots \\ \vdots \\ f_{x_n} \end{pmatrix} \qquad (grad\, f)(P_0) = \begin{pmatrix} f_{x_1}(P_0) \\ \vdots \\ \vdots \\ f_{x_n}(P_0) \end{pmatrix}$$

In den beiden folgenden Bildern stellen die Grauschattierungen die Funktion f dar, wobei schwarz den höchsten Funktionswert darstellt, und die Pfeile symbolisieren den zugehörigen Gradienten:

[http://de.wikipedia.org/wiki/Gradient_(Mathematik)]

Man beachte: Der Gradient "lebt" im Raum (x,y), in dem die Funktion f definiert ist, NICHT im Raum (x,y,z), den man braucht, um sich die Funktion vorzustellen.
[in Vorlesung: wieso der Gradient die Windrichtung angibt]

Anwendungsbeispiel Gradient: Bildverarbeitung, s. Bilder in Burger_2005\Bilder\ch07 und in Lehrmaterial\ch07.

Beispiel: Der Gradient der Funktion $f(x,y) = 3xy + y^2$ lautet $grad\ f = \begin{pmatrix} 3y \\ 3x + 2y \end{pmatrix}$, an der Stelle (x,y)=(2,1) wird er zum Vektor $(grad\ f)(2,1) = \begin{pmatrix} 3 \cdot 1 \\ 3 \cdot 2 + 2 \cdot 1 \end{pmatrix} = \begin{pmatrix} 3 \\ 8 \end{pmatrix}$, an der Stelle (x,y)=(2,0) wird er zum Vektor $(grad\ f)(2,0) = \begin{pmatrix} 0 \\ 6 \end{pmatrix}$.

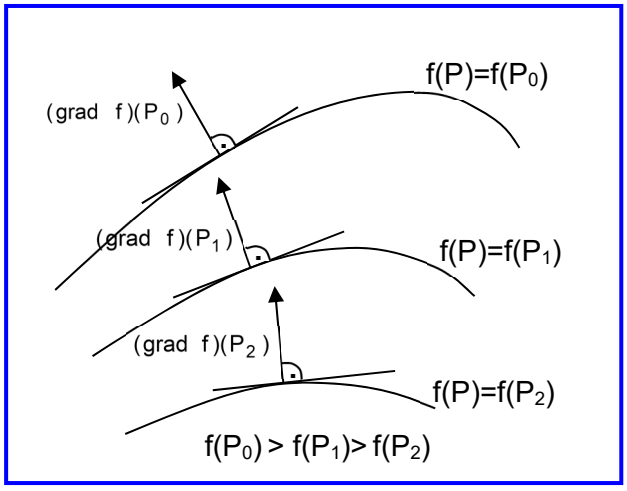
Die Ableitung einer Funktion mehrerer Veränderlicher $f(\mathbf{x}) = f(x_1, x_2, \ldots, x_n)$ nach der Zeit läßt sich mit dem Gradienten sehr kompakt schreiben:

$$\frac{df(\vec{x})}{dt} = \frac{df(x_1, \ldots, x_n)}{dt} = grad\ f \cdot \frac{d\vec{x}}{dt}$$

Satz S 8-6 Eigenschaften des Gradienten

1. Der Gradient $(grad\ f)(P_0)$ steht senkrecht auf der durch P_0 verlaufenden Äquipotentiallinie- oder fläche, also der Punktmenge $\{ P \in \mathbf{R}^n \mid f(P) = f(P_0) \}$.
2. Der Gradient weist in die Richtung des steilsten Anstiegs. D. h. die Änderung von f an der Stelle P_0 hat in Richtung von $(grad\ f)(P_0)$ ihren Maximalwert, nämlich den Betrag $|(grad\ f)(P_0)|$.

Der Gradient hat also eine sehr anschauliche Bedeutung im "Funktionengebirge".



Beispiele und Beweis von **Satz S 8-6** in Vorlesung

Übung: Wir befinden uns im Punkt P=(x,y,z)=(1,2,-1). In welcher Richtung hat die Funktion

$$f = f(x,y,z) = exp(x^2 + y^2 - 2z^2)$$

ihren steilsten Anstieg?

Übung: Gegeben sei Punkt P=(x,y)=(2,1) und die Funktion

$$g = g(x,y) = e^{x^2-2y^2}$$

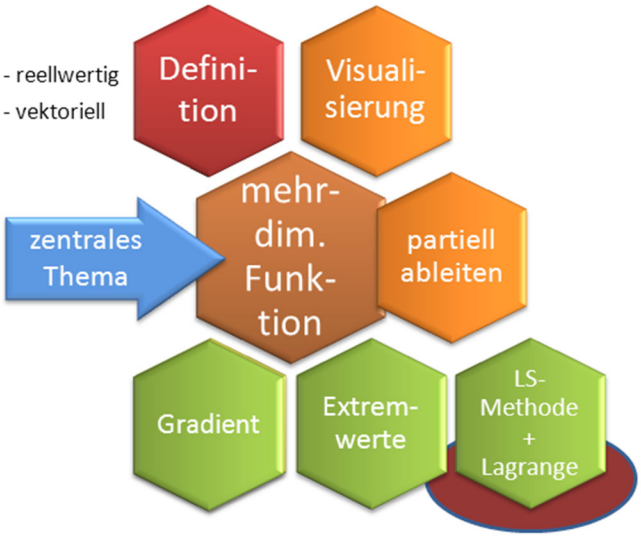
a) Welche (Tangential-)Richtung hat die Höhenlinie $g(x,y) = e^2$ im Punkt P und in anderen Punkten, die sie durchläuft?
[Hinweis: Ortsvektor $\vec{r} = \vec{r}(y)$ als Funktion von y parametrieren und Ableitung bilden]
b) Wie lautet der Gradient im Punkt P und in anderen Punkten dieser Höhenlinie?
c) Zeigen Sie, dass Gradient und Tangentialvektor der Höhenlinie im Punkt P und in jedem anderen Punkt der Höhenlinie aufeinander senkrecht stehen.

Der Gradient spielt eine große Rolle in der Optimierung, bei der man oft ein bestimmtes Fehlersignal zu minimieren hat. Statt unzählige (unendlich viele) Funktionsdifferenzen auszuprobieren, reicht es für „glatte" Funktionen, an der Stelle P_0 den Gradienten auszurechnen (einen Vektor aus lauter Zahlen!) und ein Stückchen in die Gegenrichtung zu marschieren. Man spricht vom **Gradienten-Abstiegsverfahren** (*engl.* ***gradient descent***), einer wichtigen Methode der Optimierung.

Große Bedeutung für die **praktische Optimierung**: Wenn ich ein Modell mit 5 oder 10 oder 50 Dimensionen habe (Parameter-Tuning für Simulationsmodell), dann bin ich in diesem hochdimensionalen Raum „blind wie ein Maulwurf"! Nur der Gradient gibt mir die Information, wie ich an den Steuerknöpfen drehen muss, um meinen Output zu verbessern.

Gilt natürlich nur, wenn es im Funktionengebirge nicht „auf und ab" geht (was leider in der Praxis häufiger zutrifft, als einem lieb ist). Hierfür haben die Wissenschaftler aber auch pfiffige Rezepte entwickelt: Ein Applet zu PSO (Particle Swarm Optimization) von http://gecco.org.chemie.uni-frankfurt.de/PsoVis/index.html zeigt ein Beispiel für eine komplexere Optimierungsstrategie. „Ein Schwarm ist intelligenter als seine Individuen" (→ WPF Spiele, Simulation u. Dynamische Systeme, Kapitel Partikel- und Schwarmsysteme).

8.8. Optimierung mit Lagrange-Multiplikatoren



[Papula, Bd. 2, S. 333-340],
http://www.slimy.com/~steuard/teaching/tutorials/Lagrange.html

Die meisten realen Optimierungsprobleme haben Nebenbedingungen:

- Maximiere den Gewinn, wobei die Summe der Maschinen-Stunden konstant ist
- Minimiere die Freistunden in einem Stundenplan, wobei jeder Raum in jeder Stunde nur durch eine Klasse belegt sein darf
- usw.

Beispiel: Wo liegen die Extrema von $Z(x,y) = x+2y$, wenn die Nebenbedingung $x^2+y^2=5^2$ einzuhalten ist?
[Lösung in den Übungen]

Der simple Ansatz: Nebenbedingung nach einer Variablen auflösen, z.B. y=y(x), in Z(x,y) einsetzen, dann Extrema von F(x) = Z(x,y(x)) suchen.

Dies geht jedoch nicht immer: Sei $Z(x,y)$ eine zu optimierende Zielfunktion und $\varphi(x,y)=0$ die Nebenbedingung. Die obige Methode funktioniert nicht (gut),

- wenn die Auflösung von $\varphi(x,y)=0$ nach x oder y nicht möglich oder aber zu aufwendig ist;
- wenn die Auflösung $y=y(x)$ zwar gelingt, aber $Z(x,y(x)) = F(x)$ zu unnötig komplizierten Ableitungen $F'(x)$ oder $F''(x)$ führt.

Die Methode der Lagrange-Multiplikatoren bietet hier ein elegantes anderes Verfahren:

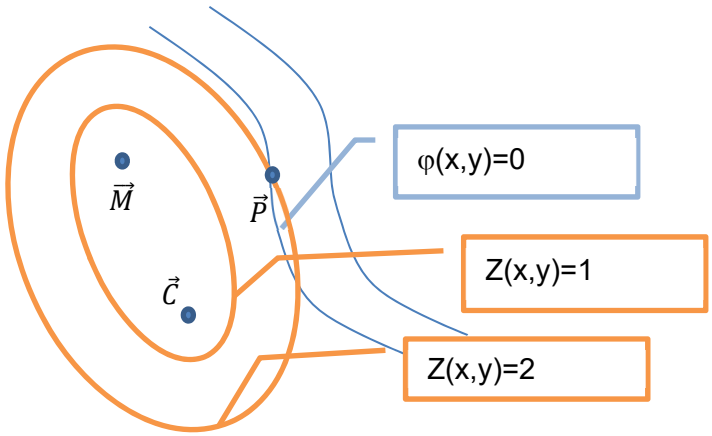
Satz S 8-7 Lagrange-Multiplikator

Gegeben sei eine zu optimierende Zielfunktion $Z(x,y)$ und eine Nebenbedingung $\phi(x,y) = 0$, die gleichzeitig einzuhalten ist. Dieses Problem wird in folgenden Schritten gelöst:

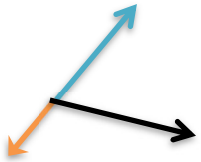
1. Bilde die Hilfsfunktion
$$F(x,y,\lambda) = Z(x,y) + \lambda\phi(x,y)$$
Der (noch unbekannte) Parameter λ heißt Lagrange-Multiplikator
2. Setze die partiellen Ableitungen gleich Null:
$$F_x = Z_x(x,y) + \lambda\phi_x(x,y) = 0$$
$$F_y = Z_y(x,y) + \lambda\phi_y(x,y) = 0$$
$$F_\lambda = \phi(x,y) = 0$$
Aus diesen 3 Gleichungen lassen sich die 3 Unbekannten x, y und λ bestimmen.
3. Gibt es mehrere Lösungen, so kann man durch Einsetzen in Z(x,y) herausfinden, welche der Lösungen ein Maximum (bzw. Minimum) sein *kann*. (Einen hinreichenden Nachweis hat man damit allerdings nicht)

Die Sache mutet wie ein Taschenspielertrick an: Erst ergänzen wir ein $\lambda\cdot 0$, erhalten so eine neue Funktion $F(x,y,\lambda)$, eliminieren dann λ wieder und haben angeblich eine Lösung von $Z(x,y)$, die die Nebenbedingung einhält? Wieso?

In Vorlesung erklären wir, wieso dieser Trick funktioniert („Milchmädchenproblem“).



Kollinearität: Zwei Vektoren $\vec{a}, \vec{b}$sind genau dann **kollinear**, wenn es eine Konstante λ gibt mit $\vec{a} = -\lambda\vec{b}$ Beispiel:



Der blaue und der orange Vektor sind kollinear zueinander, der schwarze nicht.

Anmerkung:

- Das Verfahren der Lagrange-Multiplikatoren lässt sich ohne Schwierigkeiten auch auf Funktionen von n Variablen mit m Nebenbedingungen (m<n) verallgemeinern. Die Hilfsfunktion lautet dann:

$$F(x_1, \dots, x_n, \lambda_1, \dots, \lambda_m) = Z(x_1, \dots, x_n) + \sum_{i=1}^{m} \lambda_i \phi_i(x_1, \dots, x_n)$$

  und die (n+m) partiellen Ableitungen und damit Gleichungen ergeben sich analog.
- Die Nebenbedingungen müssen in **Gleichungsform** vorliegen. Bei Nebenbedingungen in Ungleichungsform helfen die Lagrange Multiplikatoren nicht weiter, hier braucht man andere Optimierungsmethoden (Simplex oder Interior Points). Das wollen wir aber hier nicht weiterverfolgen.

Anwendungsbeispiel Informatik:

8.8.1. Shannon's Informationsmaß und Kodierungstheorie

Aus der Theoretischen Informatik ist nach Shannon bekannt: Wenn über einen Kommunikationskanal Zeichen aus dem Alphabet $\{a_i \mid i=1,\dots,N\}$ mit relativer Häufigkeit p_i geschickt werden, dann ist der mittlere Informationsgewinn, wenn das nächste Zeichen bekannt wird

$$I_{sh} = -\sum_{i=1}^{N} p_i ld(p_i)$$

Man rechnet häufig auch mit

$$I = -\sum_{i=1}^{N} p_i \, ln(p_i)$$

das unterscheidet sich nur durch einen konstanten Faktor und ist leichter zu differenzieren.

Problemstellung: Wenn man die relativen Häufigkeiten p_i frei wählen kann (unter Einhaltung der Nebenbedingung $\sum_{i=1}^{N} p_i = 1$, die immer erfüllt sein muss), welche p_i maximieren dann den mittleren Informationsgewinn?

Lösung:

$$\phi(p_1, \dots, p_N) = p_1 + \cdots + p_N - 1$$

$$F(p_1, \dots, p_N, \lambda) = -\sum_{i=1}^{N} p_i \, ln(p_i) + \lambda(p_1 + \cdots + p_N - 1)$$

$$F_{p_1} = -p_1 \frac{1}{p_1} - ln(p_1) + \lambda = -1 - ln(p_1) + \lambda = 0,$$

$$\vdots$$

$$F_{p_N} = -p_N \frac{1}{p_N} - ln(p_N) + \lambda = -1 - ln(p_N) + \lambda = 0$$

Setzt man 1. und 2. Gleichungen gleich, so folgt $\ln(p_1)=\ln(p_2) \Rightarrow p_1= p_2$,
setzt man 2. und 3. Gleichung gleich, so folgt $\ln(p_2)=\ln(p_3) \Rightarrow p_2= p_3$, usw. Insgesamt folgt also $p_1= p_2= \dots = p_N$ und mit der Nebenbedingung $\sum_{i=1}^{N} p_i = 1$ wird daraus $p_i = \frac{1}{N}$.

Antwort: Der Kommunikationskanal überträgt genau dann die maximale Informationsmenge pro Zeichen, wenn alle Zeichen aus dem Alphabet gleichwahrscheinlich sind. Bei N=4 ist der maximale mittlere Informationsgewinn

$$I_{sh} = -\sum_{i=1}^{4} p_i \mathrm{ld}(p_i) = -\sum_{i=1}^{4} \frac{1}{4} \mathrm{ld}\left(2^{-2}\right) = -(-2)\sum_{i=1}^{4} \frac{1}{4} \cdot 1$$

$$= 2 \text{ [bit]}$$

Kleiner Exkurs: **Shannon-Fano-Kodierung**

Teile die Buchstaben in 2 Gruppen, dass die Summe der Häufigkeiten in jeder Gruppe möglichst gleich ist:

Buchstabe	e	g	a
rel. Häufigkeit	50%	25%	25%
Code	0	10	11

Dann kommen alle zweistelligen Zeichenfolgen gleichhäufig vor:

Zeichenfolge	Buchstabenkette	Wahrscheinlichkeit
00	„ee“	50%*50% = 25%
01…	„eg“ oder „ea“	2*50%*25% = 25%
10	„g“	25%
11	„a“	25%

Der Kommunikationskanal überträgt also die maximale Informationsmenge.

(evtl. Übung 2 vor Übung 1 machen)

Übung 1: Wir erweitern das obige Beispiel: Gegeben sei ein Alphabet mit 4 Zeichen mit Wahrscheinlichkeiten p_1, p_2, p_3, p_4 sowie den **zwei** Nebenbedingungen

(1) $p_1+ p_2+ p_3+ p_4=1$

(2) $p_1= 2p_2$

Welche Wahrscheinlichkeiten p_i maximieren unter diesen beiden Nebenbedingungen den mittleren Informationsgewinn $I = -\sum_{i=1}^{N} p_i \, ln(p_i)$?

Übung 2: Ein Zufallsexperiment habe 4 mögliche Ergebnisse, die mit den Wahrscheinlichkeiten $p_1,...,p_4$ auftreten. Weil *eines* dieser Ergebnisse immer herauskommen muss, gilt offensichtlich $p_1+ p_2+ p_3+ p_4=1$. Bei welchen Wahrscheinlichkeiten wird das Produkt

$$Z(p_1,...,p_4)=p_1p_2p_3p_4$$

maximal?

Zeigen Sie mit Lagrange-Multiplikatoren, dass die Lösung $p_1=...=p_4=0.25$ ist!

Anmerkung: Weil die p_i Wahrscheinlichkeiten sind, gilt $p_i \in [0,1]$ $\forall i=1,\ldots,4$.

8.9. Fazit

Wichtige Begriffe und Ergebnisse aus diesem Kapitel waren:

reelle Funktion mehrerer Veränderlicher	$f:\ \mathbb{R}^n \to \mathbb{R}$: n Veränderliche, 1 abhängige Größe
Vektorfunktion	$\vec{\boldsymbol{x}}:\ \mathbb{R}^n \to \mathbb{R}^m$: n Veränderliche, m abhängige Größen
Tangentialebene	Ebene im Raum $\boldsymbol{R}^{n+1}$ durch den Punkt $(\vec{\boldsymbol{x}}, f(\vec{\boldsymbol{x}}))$, die in allen Richtungen die Steigung der (stetigen) Funktion f in $\vec{\boldsymbol{x}}$ hat.
Äquipotentialflächen	Flächen mit $f(\vec{\boldsymbol{x}}) = const.$ im $\vec{\boldsymbol{x}}$-Raum. Für $\vec{\boldsymbol{x}} \in \boldsymbol{R}^2$ werden die Flächen zu Linien, den Höhenlinien.
partielle Ableitung nach x_i	alle Veränderlichen außer x_i als konstant festsetzen, dann "normal" nach x_i ableiten
totales Differential	Zuwachs in der Tangentialebene bei Verrückung um $d\vec{\boldsymbol{x}}$
Gradient von f	Vektorfunktion im Raum $\mathbb{R}^n$, die i. Komponente ist f_{x_i}.

Wichtige Ergebnisse:

- Funktionen mehrerer Veränderlicher lassen sich über Flächen im Raum, über Höhenliniendiagramme oder über Kennlinienfelder visualisieren (Kap. 8.3).
  - Höhenlinien: z = f(x,y) nach y auflösen
  - Kennlinien: alle Veränderliche bis auf eine konstant festsetzen.
- Die Differentialrechnung einer Veränderlichen läßt sich auf Funktionen mehrerer Veränderlicher übertragen (Kap.8.4)
  - **partielle Ableitung**: alle Veränderliche bis auf eine konstant, dann ableiten.
- **Extremwerte** (Kap. 8.5): Hinreichende Kriterien sind für mehr als 2 Variablen schwierig, für 2 Variablen aber gut angebbar (Satz S 8-3).
- **Modelle in der Informatik**: Mit der Methode der **kleinsten Quadrate (LS-Methode)** (Kap. 8.6) lassen sich Parameter von Modellen optimieren. Unser Beispiel: Ausgleichsgerade (Regression).
- Der **Gradient** (Kap. 8.7) ist der Vektor aller 1. partiellen Ableitungen. Er steht an jeder Stelle senkrecht auf den Äquipotentialflächen und weist in Richtung des steilsten Anstiegs.
- Viele reale Optimierungsprobleme mit mehreren Veränderlichen haben neben einem Maximierungsziel auch weitere Nebenbedingungen zwischen den Veränderlichen in Gleichungsform. Hier hilft die Methode der **Lagrange-Multiplikatoren** (Kap. 8.8) entscheidend weiter.