Optimierung

Vorlesung, zuerst gehalten im Wintersemester 2012/13

Tomas Sauer

Version 0.0
Letzte Änderung: 28.1.2013

Statt einer Leerseite . . . | 0

Chaos is found in greatest abundance whereever order is being sought. It always defeats order, because it is better organized.

T. Pratchett, *Interesting times*

Die wahren Analphabeten sind schließlich diejenigen, die zwar lesen können, es aber nicht tun. Weil sie gerade fernsehen.

L. Volkert, *SZ–Online, 11.7.2009*

When the epoch of analogue (which was to say also the richness of language, of *analogy*) was giving way to the digital era, the final victory of the numerate over the literate.

S. Rushdie, *Fury*

The most incredible thing about miracles is that they happen.

G. K. Chesterton, *The Innocence of Father Brown*

And it didn't stop being magic just because you found out how it was done.

T. Pratchett, *Wee Free Men*

Tomas Sauer
Lehrstuhl für Mathematik mit Schwerpunkt Digitale Bildverabeitung
Universität Passau
Innstr. 43
94032 Passau

0 Inhaltsverzeichnis

Erst die natürlichen Betrachtungen gemacht, ehe die subtilen kommen, und immer vor allen Dingen erst versucht, ob etwas ganz simpel und natürlich werden könne.

G. Chr. Lichtenberg

1 Optimierung – Grundlagen und Beispiele

Eigentlich ist Optimierung (in „voller“ Allgemeinheit) ein ziemlich einfaches Problem, nämlich das Auffinden eines Extremums:

Zu einer Funktion $F : D \to \mathbb{R}$ *und* $D' \subset D$ *finde man ein* $x^* \in D'$*, so daß*

$$F(x^*) \leq F(x) \qquad \textit{oder} \qquad F(x^*) \geq F(x)$$

für alle $x \in D'$.

Zuerst einmal sollte man bemerken, daß es egal ist, ob man die **Zielfunktion** F minimiert oder maximiert, denn man kann die Suche nach einem Maximum von F immer auch als Suche nach einem Minimum von $-F$ auffassen. Je nachdem, was uns gerade genehm ist, können wir daher bei einer **Normalform** des Optimierungsproblems annehmen, daß wir eine Zielfunktion nur maximieren oder nur minimieren wollen.
Weitere generelle Vereinfachungen kann man allerdings weder in der Theorie noch in der Praxis machen:

1. Die Funktion F kann, je nach Problemstellung differenzierbar, stetig oder auch unstetig sein; und selbst wenn F differenzierbar sein sollte, ist es noch lange nicht klar, ob und wie man diese Ableitung auch wirklich bestimmen kann.

2. Die **Auswertung** der Funktion F, das heißt, die Berechnung des Wertes $F(x)$ für ein $x \in D$, kann *teuer* oder *billig* sein. Das kann sich auf Rechenzeit beziehen, denn manchmal kann $F(x)$ nur durch aufwendige Simulationen berechnet werden, oder aber auch auf „echte“ Unkosten, wenn die zur Bestimmung von $F(x)$ reale Experimente oder Messungen nötig sind.

3. Der **zulässige Bereich** D' kann ganz D umfassen, insbesondere ist $D' = D = \mathbb{R}^n$ möglich, oder er kann eine echte, „dünne“ oder kompakte Teilmenge von D sein.

4. Insbesondere kann D' **implizit** gegeben sein, das heißt, man kennt lediglich eine Funktion $g : D \to \mathbb{R}^m$, so daß

$$D' = \{x \in D \ : \ g(x) = 0\}.$$

All diese Situationen verlangen natürlich nach unterschiedlichen Methoden, wenn man die Lösung,

- das Optimum,
- ein Optimum,
- einen nahezu optimalen Wert,

praktisch bestimmen will. Mit derartigen Verfahren, die natürlich wesentlich von der Struktur des Optimierungsproblems abhängen, soll sich diese Vorlesung beschäftigen – **das** Black–Box–Verfahren schlechthin, das jedes Optimierungsproblem löst, kann und wird es nicht geben.
Sehen wir uns nun aber zuerst einmal ein paar Beispiele für Optimierungsprobleme an.

1.1 Lineare und Ganzzahlprogrammierung

Der „einfachste“ Fall eines Optimierungsproblems liegt vor, wenn die Zielfunktion $F : \mathbb{R}^n \to \mathbb{R}$ eine **lineare Funktion** der Form

$$F(x) = v^T x, \qquad x \in \mathbb{R}^n,$$

für ein gegebenes $v \in \mathbb{R}^n$ ist. Da eine nichttriviale lineare Funktion auf ganz $\mathbb{R}^n$ alle Werte zwischen $\pm\infty$ annimmt, sind bei derartigen Optimierungsproblemen Nebenbedingungen unvermeidbar, wenn man vernünftige Aussagen machen will. Besonders schön sind dabei natürlich **lineare Nebenbedingungen** der Form

$$a_j^T x \geq b_j, \qquad j = 1, \ldots, N,$$

die man dann schön in der Matrixform $Ax \geq b$ schreiben kann.

Beispiel 1.1 (Einkauf chemischer Rohstoffe ohne Realitätsbezug) *Eine Chemiefirma*[1] *benötigt zwei Chemikalien A und B zur Herstellung ihres Produkts, und zwar mindestens 3t von Stoff A und 4t von Stoff B. Allerdings sind diese beiden Rohstoffe nicht in reiner Form erhältlich, sondern lediglich die beiden Rohstoffe X und Y die A und B enthalten, und zwar wie folgt:*

Rohstoff	*Anteil A*	*Anteil B*	*Kosten*
X	*60 %*	*40 %*	*300*
Y	*30 %*	*50 %*	*200*

[1] Dieses Beispiel stammt (in allgemeinerer Form) aus (Duden, 1969, S. 501), die grafische Lösung dort ist ein sehenswertes Kunstwerk.

Hierbei bezeichnet „Kosten" die Summe aus Einkaufspreis und den Aufwendungen für die Gewinnung der Rohstoffe. Was ist nun die günstigste Einkaufspolitik?

Nun, dieses Problem ist relativ einfach zu lösen! Seien nämlich x, y die gekauften Mengen der Rohstoffe X und Y, dann ergibt sich das Optimierungsproblem

$$\min\ 300x + 200y, \qquad \begin{bmatrix} .6 & .3 \\ .4 & .5 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \geq \begin{bmatrix} 3 \\ 4 \end{bmatrix}. \tag{1.1}$$

Jetzt hätten wir fast was vergessen: Da wir keine negativen Mengen einkaufen können[2], müssen wir auch noch $x, y \geq 0$ fordern. Der **zulässige Bereich** für dieses Optimierungsproblem ist in Abb 1.1 dargestellt. Um das Optimierungsprob-

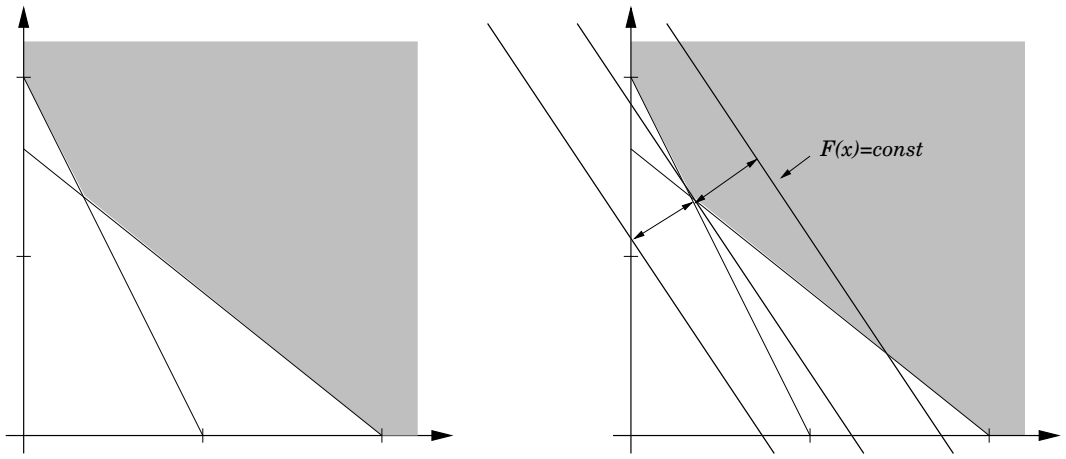

Abbildung 1.1: Der zulässige Bereich (links) und die grafische Lösung (rechts) des Optimierungsproblems aus Beispiel 1.1. Bei der grafischen Lösung verschiebt man die Gerade $F(x) = c$, $c \in \mathbb{R}$, so lange, bis sie den zulässigen Bereich gerade noch berührt – das ist dann offensichtlich der Minimalwert.

lem (grafisch) zu lösen betrachten wir, daß die Kosten jeweils auf der Gerade $F(x) = c$ den konstanten Wert c haben; verschiebt man also die Gerade nach „links unten" so erhält man günstigere Ergebnisse – natürlich nur, solange die Gerade auch den zulässigen Bereich schneidet, denn ansonsten würde man zwar billig einkaufen, könnte aber nicht produzieren[3]. Also „schieben" wir solange, bis die Gerade den zulässigen Bereich gerade noch berührt und haben die optimale Lösung, nämlich denjenigen Punkt $[x, y]^T$, der sich als

$$\begin{bmatrix} .6 & .3 \\ .4 & .5 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \end{bmatrix} \quad \Rightarrow \quad \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1\frac{2}{3} \\ 6\frac{2}{3} \end{bmatrix}$$

[2]Das ist übrigens keine generelle Annahme für Optimierungsprobleme: In der Finanzmathematik sind sogenannte *Leerverkäufe*, also Verkauf von Aktien, die man gar nicht hat, durchaus nicht verboten.

[3]Bemerkungen über moderne Managementkonzepte sollen an dieser Stelle nicht gemacht werden.

ergibt.

Das alles ist schön und gut, solange nur zwei Parameter zu optimieren sind – dann kann man sich sehr einfach auf grafische Lösungsverfahren zurückziehen. Es kann aber natürlich vorkommen, daß die Anzahl der Parameter sehr viel größer wird.

Beispiel 1.2 (Transportproblem) *Ein Konzern*[4] *hat* m *Fabriken, die pro Tag* a_j *Tonnen einer Chemikalie herstellen,* $j = 1, \ldots, m$*, und* n *Verkaufsstellen, die pro Tag einen Mindestbedarf von* b_k *Tonnen der Chemikalie,* $k = 1, \ldots, n$*, haben. Der Transport einer Tonne der Chemikalie von Fabrik* j *zu Verkaufsstelle* k *kostet* c_{jk} *Euro. Wie erhält man eine kostenoptimale Versorgung der Verkaufsstellen?*

Hier haben wir es offenbar mit einer ganzen Menge, genauer gesagt mn, Parametern x_{jk}, $j = 1, \ldots, m$, $k = 1, \ldots, n$, zu tun und das Optimierungsproblem lautet

$$\min \sum_{j=1}^{m} \sum_{k=1}^{n} c_{jk}\, x_{jk}$$

unter den Nebenbedingungen

$$\sum_{j=1}^{m} x_{jk} \geq b_k, \quad k = 1, \ldots, n, \qquad \sum_{k=1}^{n} x_{jk} \leq a_k, \quad j = 1, \ldots, m.$$

Und da tut sich grafisch nicht mehr viel ...

Es war nicht ganz zufällig, daß wir in den beiden vorhergegangenen Beispielen von Chemikalien gesprochen haben – man kann nämlich davon ausgehen, daß man die in beliebigen Bruchteilen hin- und herverschieben kann. Das wird anders, wenn es sich um „quantisierte" Objekte handelt, die nicht beliebig unterteilt werden können, denn dann muß man nach *ganzzahligen* Lösungen suchen und landet bei der **Ganzzahlprogrammierung**, die auch als „**Integer programming**" bekannt ist.

Beispiel 1.3 (Ganzzahliges Transportproblem) *Eine Transportfirma*[5] *transportiert zwei verschiedene Typen, A und B von Paletten, die unterschiedliche Größe und Gewicht haben und unterschiedlich bezahlt werden:*

Typ	*Größe (cbm)*	*Gewicht (kg)*	*Bezahlung*
A	*2*	*400*	*11*
B	*3*	*500*	*15*

Ein Transportfahrzeug hat eine Zuladung von 3700 kg und ein Ladevolumen von 20 cbm. Was ist die optimale Beladung.

Ganz genau wie vorher können wir unser Optimierungsproblem in der Form

$$\max\ 11x + 15y, \qquad \begin{bmatrix} 2 & 3 \\ 400 & 500 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \leq \begin{bmatrix} 20 \\ 3700 \end{bmatrix}, \qquad x, y \geq 0,$$

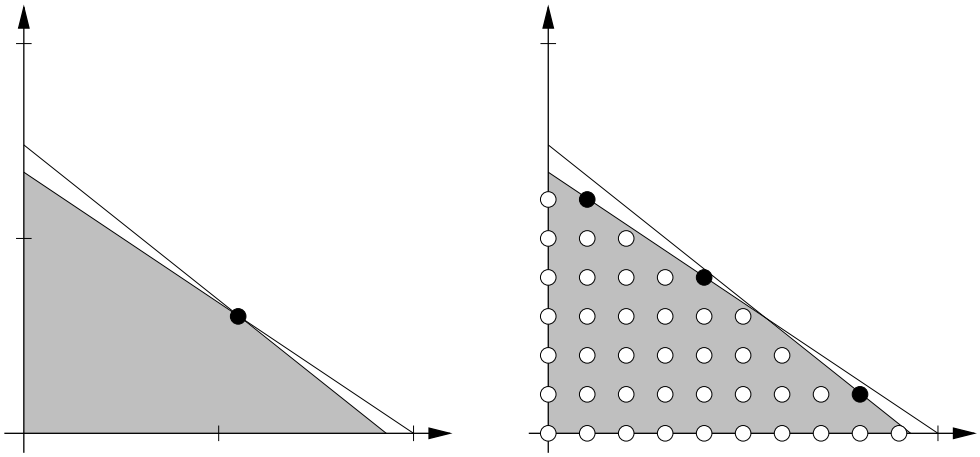

Abbildung 1.2: Links der kontinuierliche zulässige Bereich für Beispiel 1.3 mit der (kontinuierlichen) Optimallösung $\left[5\frac{1}{2}, 3\right]^T$, rechts die ganzzahligen Punkte im zulässigen Bereich, die schwarz markierten liegen gerade so auf dem Rand.

hinschreiben, nur interessieren wir uns jetzt nur noch für die *ganzzahligen* Lösungen und die unter diesen optimale, nämlich $x = y = 4$. Was unseren graphischen Ansatz angeht, wird ja eigentlich alles viel leichter, denn man muß halt nun nur noch die Gerade so verschieben, daß sie durch einen der ganzzahligen Punkte im zulässigen Bereich geht und sich da den größten Wert aussuchen. Aber man sieht schon an diesem einfachen Beispiel, daß das ein extrem genaues Arbeiten nötig ist.

Der zweite Ansatz wäre ein schlichtes Ausprobieren, man schreibt ein kleines Programm, das alle zulässigen ganzen Zahlen durchprobiert und so den Optimalwert ermittelt. Das funktioniert noch bei kleinen Beispielen, wird aber schnell hoffnungslos, wenn man die Anzahl der Parameter erhöht. Tatsächlich verwendet ein methodisches Vorgehen algebraische Geometrie, Gröbnerbasen für torische Ideale und Eliminationsideale, siehe (Cox *et al.*, 1998, Chapter 8) oder (Hoşten & Thomas, 1998).

1.2 Quadratische Programmierung

Ein weiterer einfacher Fall ist die Situation, daß F ein *quadratisches* Polynom ist, das heißt,

$$F(x) = x^T A x + b^T x, \qquad x \in \mathbb{R}^n,$$

[4] Aus (Nocedal & Wright, 1999, S. 4), aber in leicht verallgemeinerter Form.
[5] Aus (Cox *et al.*, 1998, S. 359–360).

am besten noch mit einer *symmetrischen, positiv definiten*[6] Matrix $A \in \mathbb{R}^{n\times n}$. Dann hat man es nämlich mit genau einem Minimum zu tun: da

$$\lim_{|x|\to\infty} F(x) = \infty$$

und, für $j = 1, \ldots, n$,

$$\begin{aligned}\frac{\partial}{\partial x_j} F(x) &= \frac{\partial}{\partial x_j}\left(\sum_{k=1}^{n} a_{kk}x_k^2 + 2\sum_{1\le k<\ell\le n} a_{k\ell}\,x_k x_\ell + \sum_{k=1}^{n} b_k\,x_k\right)\\ &= 2a_{jj}x_j + 2\sum_{k\neq j} a_{jk}x_k + b_j = 2\,(Ax)_j + b_j,\end{aligned}$$

haben wir ein Extremum an x^* wenn

$$0 = \nabla F\,(x^*) =: \left[\frac{\partial F}{\partial x_j}\,(x^*)\ :\ j = 1, \ldots, n\right],$$

also wenn $0 = 2Ax^* + b$ oder

$$x^* = -\frac{1}{2}A^{-1}\,b,$$

vorausgesetzt, A ist invertierbar, was sicher der Fall ist, wenn A positiv definit ist. Mindestens ein Minimum muß es aber geben, also ist x^* das eindeutige Minimum – wie man das berechnet, und zwar effizient, das ist dann wieder eine andere Frage.

Beispiel 1.4 (Portfolio–Optimierung) *Gegeben seien n Investitionsmöglichkeiten*[7] *mit „Return" oder Auszahlung r_j, $j = 1, \ldots, n$. Diese Auszahlungen betrachtet man als Zufallsvariable, von denen jeweils der Erwartungswert $\mu_j = E\,[r_j]$ und die Varianz $\sigma_j^2 = E\left[(r_j - \mu_j)^2\right]$, $j = 1, \ldots, n$, bekannt sind. Ein* Portfolio[8] *besteht nun aus x_j Anteilen der Investition Nummer j, $j = 1, \ldots, n$, der Einfachheit so normiert, daß $x_1 + \cdots + x_n = 1$. Die Auszahlung des Portfolio ist dann*

$$R = \sum_{j=1}^{n} r_j\,x_j$$

und die erwartete Auszahlung

$$E[R] = E\left[\sum_{j=1}^{n} r_j\,x_j\right] = \sum_{j=1}^{n} \mu_j\,x_j = \mu^T x.$$

Mit Hilfe der Kovarianzmatrix

$$K := \left[\frac{E\,[(r_j - \mu_j)\,(r_k - \mu_k)]}{\sigma_j\,\sigma_k}\ :\ j, k = 1, \ldots, n\right]$$

[6] Positiv definit bedeutet hier *strikt* positiv definit, also $x^T Ax > 0$ für $x \neq 0$. *Achtung:* Diese Terminologie ist *nicht* eindeutig in der Literatur.

[7] Aus (Nocedal & Wright, 1999, S. 216).

[8] Der ganz einfachen Form.

der r_j *ist dann*

$$\sigma_R := E\left[(R - E[R])^2\right] = x^T G x,$$

wobei

$$G = \Sigma^T K \Sigma = [E[(r_j - \mu_j)(r_k - \mu_k)] \ : \ j,k = 1,\dots,n]$$

und

$$\Sigma = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{bmatrix}.$$

Nun sieht man σ_R *als das* Risiko *des Portfolios an und wählt, für einen Parameter* $\kappa \in [0,\infty)$*, der die „Risikofreude" des Anlegers widerspiegeln soll*[9]*, die Anteile der Investitionen als Lösung des quadratischen Optimierungsproblems*

$$\max\ \mu^T x - \kappa\, x^T G x, \qquad \sum_{j=1}^{n} x_j = 1, \quad x_j \geq 0.$$

Beispiel 1.5 (Glättende Funktionen und Lerntheorie) *Gegeben seien nun Punkte* x_j*,* $j = 0,\dots,N$*, und vorgeschriebene oder gemessene Werte* y_j*,* $j = 0,\dots,N$*, und man möchte gerne aus diesen diskreten oder „abgetasteten" Werten eine Funktion rekonstruieren, oder auch ein zugrundeliegendes „Bildungsgesetz". Der erste Ansatz, der einem sofort in den Sinn kommt, wäre* Interpolation*, also die Bestimmung einer Funktion* f *mit* $f(x_j) = y_j$*,* $j = 0,\dots,N$*, siehe z.B. (Isaacson & Keller, 1966; Kunz, 1957). Das ist an sich nichts schlimmes, wirft aber sofort die folgenden Probleme auf*

- *Es gibt unendlich viele Funktionen, die das Interpolationsproblem lösen, das Problem ist also* schlechtgestellt*. Man kann dem Problem dadurch begegnen, daß man* f *als*

$$f = \sum_{j=0}^{N} c_j f_j, \qquad \mathbf{c} = (c_0,\dots,c_N) \in \mathbb{R}^{N+1}$$

  ansetzt und den $N+1$*–dimensionalen Funktionenraum der* f_j *passend wählt*[10]*.*

- *Die Werte* y_j *sind oftmals nicht genau, sondern fehlerbehaftet, z.B. bei Meßwerten, und dieses* Rauschen *kann durchaus beträchtliche Ausmaße haben.*

Vor allem das zweite Problem sorgt dafür, daß Interpolanten in vielen Fällen nicht wirklich geeignet sind, um Funktionen zu konstruieren, weswegen man sich mit einem anderen Ansatz behilft: Man wählt nicht $N+1$ *Funktionen, sondern* $M+1$ *Funktionen, wobei* M *mit* N *nichts zu tun haben muß, es kann größer oder kleiner als* N

[9] Je kleiner κ, desto weniger wird das „Risiko", das sich in der Varianz versteckt, in Betracht gezogen und desto mehr zählt die Auszahlung, genauer gesagt, die erwartete Auszahlung.

[10] Möglicherweise in Abhängigkeit von den x_j, aber warum auch nicht? Niemand erwartet an dieser Stelle die ultimative Universallösung.

sein, ja selbst $M = N$ *ist nicht verboten und sucht zuerst mal nach der Lösung des Minimierungsproblems*

$$\min\left\{\sum_{j=0}^{N}|f(x_j)-y_j|^2 \;:\; f=\sum_{j=0}^{M}c_jf_j\right\}. \tag{1.2}$$

Da

$$\begin{bmatrix} f(x_0)\\ \vdots\\ f(x_N)\end{bmatrix}=\begin{bmatrix} c_0f_0(x_0)+\cdots+c_Mf_M(x_0)\\ \vdots\\ c_0f_0(x_N)+\cdots+c_Mf_M(x_N)\end{bmatrix}=\begin{bmatrix} f_0(x_0) & \dots & f_M(x_0)\\ \vdots & \ddots & \vdots\\ f_0(x_N) & \dots & f_M(x_N)\end{bmatrix}\begin{bmatrix} c_0\\ \vdots\\ c_M\end{bmatrix},$$

oder, in Kurzform, $f(\mathbf{x})=\mathbf{Fc}$*, ist die zu minimierende Funktion*

$$\left\|\mathbf{Fc}-\mathbf{y}\right\|_2^2=(\mathbf{Fc}-\mathbf{y})^T(\mathbf{Fc}-\mathbf{y})=\mathbf{c}^T\mathbf{F}^T\mathbf{Fc}-2\mathbf{y}^T\mathbf{Fc}+\mathbf{y}^T\mathbf{y},$$

was bekanntlich[11] *durch die Lösung des Gleichungssystems*

$$\mathbf{F}^T\mathbf{Fc}=\mathbf{F}^T\mathbf{y} \tag{1.3}$$

minimiert wird. OK, wo ist nun das Problem? Die Lösungsfunktion f *wird immer noch versuchen, den vorgegebenen Werten so gut zu folgen, wie es geht, insbesondere wird sie an allen Punkten interpolieren, wenn es ihr mglich ist, denn dann ist der Fehler Null und besser geht es ohnehin nicht. Und damit ist unser Rauschproblem wieder da, Rauschen in den Daten* y_j *kann durchaus wieder in der Funktion landen und zu Oszillationen führen. Um das zu unterbinden,* verbieten *wir einfach der Funktion* f *das Oszillieren, indem wir zum Minimierungsproblem ein* Glättefunktional *hinzufügen, das zu viel Oszillation bestraft, beispielsweise das sehr beliebte*[12]

$$\begin{aligned}\int|f''(x)|^2\,dx &= \int\left|\sum_{j=0}^{M}c_jf_j''(x)\right|^2dx=\int\sum_{j,k=0}^{M}c_jc_kf_j''(x)f_k''(x)\,dx\\ &= \int\mathbf{c}^T\begin{bmatrix} f_0''(x)f_0''(x) & \dots & f_0''(x)f_M''(x)\\ \vdots & \ddots & \vdots\\ f_M''(x)f_0''(x) & \dots & f_M''(x)f_M''(x)\end{bmatrix}\mathbf{c}=\mathbf{c}^T\mathbf{F}'\mathbf{c},\end{aligned}$$

wobei $\mathbf{F}'$ *das Integral über die Matrix ist. Für einen Parameter* $\lambda\geq 0$ *erhalten wir dann das modifizierte Optimierungsproblem*

$$\min_{\mathbf{c}}\left\|\mathbf{Fc}-\mathbf{y}\right\|_2^2+\lambda\mathbf{c}^T\mathbf{F}'\mathbf{c},$$

dessen Lösung über das Gleichungssystem

$$\left(\mathbf{F}^T\mathbf{F}+\lambda\mathbf{F}'\right)\mathbf{c}=\mathbf{F}^T\mathbf{y} \tag{1.4}$$

[11] Das geht wie in der Schule: Ableiten und gleich Null setzen.

[12] Die Integralgrenzen lassen wir hier bewußt weg, das ist Detailkram und uns geht es ja hier schließlich um das Prinzip!

gefunden wird. Diese Lösungsfunktion versucht nun, in Abhängigkeit vom Parameter λ, der die „Prioritäten" fixiert, die vorgegebenen Werte gut zu approximieren, aber eben nicht um jeden Preis, sondern auf möglichst glatte Art und Weise.

Das Verfahren ist alt, um nicht zu sagen klassisch, und beinhaltet den guten alten „smoothing spline" aus Kurvenapproximation und Statistik, aber ganz genauso moderne Ansätze wie Lerntheorie.

1.3 Platinen und Handlungsreisende

Ein klassisches Optimierungsproblem aus dem Bereich der kombinatorischen Optimierung ist das *Travelling Salesman Problem* (TSP), bei dem ein Handlungsreisender eine Tour entwickeln muß, bei der jede von n Städten mindestens einmal angefahren werden muß und bei der der gesamte zurückgelegte Weg minimiert werden muß.

Bezeichnet $d_{jk}, j, k = 1, \ldots, n$, den Abstand[13], zwischen den Städten Nummer j und k, dann besteht eine Formulierung des Travelling Salesman Problem in der Bestimmung eines Vektors $J = (j_1, \ldots, j_N) \in \{1, \ldots, n\}^N$, der den Wert

$$d(I) = \sum_{k=1}^{N-1} d_{j_k, j_{k+1}}$$

unter der Nebenbedingung

$$\{1, \ldots, n\} = \{j_1, \ldots, j_N\},$$

daß jede Stadt *mindestens einmal* besucht wird, minimiert. Diese Nebenbedingung bedeutet, daß $N = n$ ist, daß also jede Stadt einmal besucht wird, oder daß $N = n + 1$ und $j_1 = j_N$ ist, was einer Rundreise entspricht; man könnte auch noch $j_1 = j_N = j^*$ fordern, was heißt, daß der Handlungsreisende genau in der Stadt mit Index j^* – dem Firmensitz zum Beispiel – mit seiner Tour beginnen und enden muß.

Auch wenn der Handlungsreisende ein Klassiker ist, ist er doch nicht wirklich das praktische, realistische Problem. Trotzdem gibt es Varianten davon, die in praktischen Anwendungen ganz natürlich auftreten.

Beispiel 1.6 *(Bohrlöcher in Platinen) In einer Platine sind für die Bestückung mit ICs oder Halbleitern Löcher zu bohren, und zwar nicht 10 oder 20, sondern Größenordnungen von mehreren Hundert oder sogar Tausend Löchern. Man bestimme den kürzesten oder schnellsten*[14] *Weg für den Bohrroboter. Ein ähnliches Problem taucht auch bei der späteren Bestückung der Platine und beim Anbringen der Lötpunkte auf.*

[13] Oder die Reisezeit, z.B. in der „DB–Metrik", oder die Reisekosten, oder eine gewichtete Mischung aus all dem ...

[14] Nicht immer ist der kürzeste Weg auch der schnellste, wenn man Beschleunigung und Abbremsen in Betracht zieht. Man beachte aber, daß die Weglänge eine sehr *einfache* Zielfunktion darstellt, die benötigte Zeit aber auf hochgradig *komplizierte* Art von der Reihenfolge der Punkte abhängen kann.

Ein weiteres Beispiel für Varianten des TSP ist die Produktionsplanung, bei der Prozesse optimal auf verschiedenen Ressourcen verteilt werden sollen – dabei kann es sich um Produktionsprozesse und Maschinen, Flugrouten und Flugzeuge, oder auch um Vorlesungen und Räume, Dozenten und Studenten bei der Stundenplanoptimierung handeln. Eine nette und untechnische Übersicht ist (Dueck *et al.*, 1993), siehe auch (Ablay, 1989).

Der „Reiz" des Travelling Salesman besteht darin, daß die *Komplexität* $K(n)$ des Problems, also die Anzahl der Rechenoperationen, die nötig sind, um die optimale Lösung zu berechnen, nicht polynomial in n beschränkt werden kann, das heißt, es ist

$$\lim_{n\to\infty} \frac{K(n)}{|p(n)|} = \infty$$

für jedes Polynom p. Damit ist es für *realistische* Werte von n ziemlich hoffnungslos *die* exakte Optimallösung berechnen zu wollen; überraschenderweise gibt es aber (heuristische) Verfahren, die gute bis sehr gute Lösungen[15] sehr schnell bestimmen. Man kann vielleicht noch nicht mal beweisen, daß sie immer funktionieren, aber sie tun es trotzdem.

[15] Normalerweise nur um wenige Prozent schlechter als die Optimallösung!

... the calculations, be it remembered,
of the hard-headed, strong handed,
exemplary working men ...

P. Smyth, *The Great Pyramid*

2 Lineare Optimierungsprobleme

Das Ziel von Optimierungsverfahren besteht ja ganz einfach darin, eine Zielfunktion unter vorgegebenen Nebenbedingungen zu maximieren oder zu minimieren. Sind sowohl die Zielfunktion, wie auch die Nebenbedingungen *linear*[16], dann bezeichnet man es – völlig unerwartet – als **lineares Optimierungsproblem**. Ein solches Optimierungsproblem läßt sich immer schreiben als

$$\begin{array}{rcl} c^T x & = & \max, \\ Ax & \geq & b, \end{array} \qquad A \in \mathbb{R}^{m\times n}, \quad c, x \in \mathbb{R}^n, \quad b \in \mathbb{R}^m, \tag{2.1}$$

wobei für zwei Vektoren $x, y \in \mathbb{R}^m$ die Halbordnung[17] $x \leq y$ bedeutet, daß $x_j \leq y_j$, $j = 1, \ldots, m$. Fangen wir mit den Nebenbedingungen an: Eine allgemeine einzelne **lineare Nebenbedingung** an x hätte entweder die Form

$$a^T x \geq b$$

oder aber

$$a^T x \leq b \qquad \Leftrightarrow \qquad (-a)^T x \geq (-b)\,,$$

das heißt, wir können immer von Nebenbedingungen der Form $Ax \geq b$ ausgehen. Ähnliches gilt für die Zielfunktion: Würde man $c^T x$ *minimieren* wollen, so kann man genauso gut $(-c)^T x$ maximieren. Durch Einführung von **Schlupfvariablen** $x_{n+1}, \ldots, x_{n+m}$ kann man die Ungleichungen[18] $a_j^T x \geq b$ auf die äquivalente Form

$$a_j^T x - x_{n+j} = b, \quad x_{n+j} \geq 0, \qquad j = 1, \ldots, m,$$

bringen und erhält so durch passende Erweiterung von A, b und c eine äquivalente **Normalform** des linearen Optimierungsproblems:

$$\begin{array}{rcll} c^T x & = & \max, & \\ Ax & = & b, & I \subset \{1, \ldots, n\}. \\ x_j & \geq & 0, \qquad j \in I, & \end{array} \tag{2.2}$$

[16] Eigentlich natürlich „affin".

[17] Zur Erinnerung: Halbordnung heißt, daß für $x \neq y$ nicht notwendig eine der beiden Beziehungen $x < y$ oder $x > y$ gelten muß.

[18] Hier und im Rest dieses Kapitels seien a_j^T, $j = 1, \ldots, m$, die Zeilenvektoren der Matrix A.

Mit den Normalformen muss man ein bißchen aufpassen! Je nach Literatur und verwendeter Software werden gerne unterschiedliche Normalformen verwendet und man sollte sein Problem dann schon zuerst mal so aufbereiten, daß es auch der geforderten Normalform entspricht, ansonsten braucht man sich nicht zu wundern, wenn die Ergebnisse falsch sind.

Übung 2.1 Wie muß man A, b und c aus (2.1) erweitern, so daß (2.2) eine äquivalente Formulierung ist. ♢

Wie wir sehen werden, haben (lineare) Optimierungsprobleme eine Menge mit *konvexer Analysis*, siehe z.B. (Rockafellar, 1970), zu tun – wobei es durchaus so ist, daß sich die Gebiete gegenseitig beeinflußt und motiviert haben und daß sich Resultate des einen Gebiets auch im anderen Gebiet als hilfreich erwiesen haben.

Bevor wir uns ein bißchen die mathematische Theorie ansehen, befassen wir uns erst einmal mit einem „realistischen" Beispiel aus (Schwarz, 1988, Beispiel 2.1, S. 55).

Beispiel 2.1 *(Produktionsproblem einer Schuhfabrik)*
Eine Schuhfabrik stellt Damen- und Herrenschuhe her, die unterschiedliche Forderungen an Herstellungszeit, Maschinenlaufzeit und Lederbedarf stellen – Ressourcen, die natürlich gewissen Einschränkungen unterliegen. Welche Produktionskombination erzielt den höchsten Gewinn[19]*, wenn die folgenden Parameter vorliegen:*

	Damenschuh	*Herrenschuh*	*Verfügbar*
Herstellungszeit	20	10	8000
Maschinenzeit	4	5	2000
Leder	6	15	4500
Gewinn	16	32	

Die mathematische Formulierung dieses Problems in der Normalform (2.1) ist dann

$$c = \begin{bmatrix} 16 \\ 32 \end{bmatrix}, \qquad A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -20 & -10 \\ -4 & -5 \\ -6 & -15 \end{bmatrix}, \qquad b = \begin{bmatrix} 0 \\ 0 \\ -8000 \\ -2000 \\ -4500 \end{bmatrix},$$

wobei die ersten beiden Zeilen von A und b die Tatsache wiedergeben, daß man keine negative Anzahl von Schuhen produzieren kann.

2.1 Zulässige Punkte

Als erstes sehen wir uns einmal den Bereich an, in dem wir unsere Lösungen suchen, das heißt, die Menge[20]

$$F := F(A, b) := \{x \in \mathbb{R}^n \ : \ Ax \geq b\} \subset \mathbb{R}^n,$$

[19] Unter der (realistischen ?) Annahme, daß alle Schuhe verkauft werden können.
[20] Das "F" steht für die englische Bezeichnung "feasible".

die als **zulässige Menge** oder der **zulässiger Bereich** für das Optimierungsproblem bezeichnet wird. Der zulässige Bereich ist Durchschnitt von Halbräumen[21], wobei ein **Halbraum** eine Menge der Form

$$H(a, b) := \left\{x \ : \ a^T x \geq b\right\} \subseteq \mathbb{R}^n, \qquad a \in \mathbb{R}^n,\ b \in \mathbb{R},$$

ist.

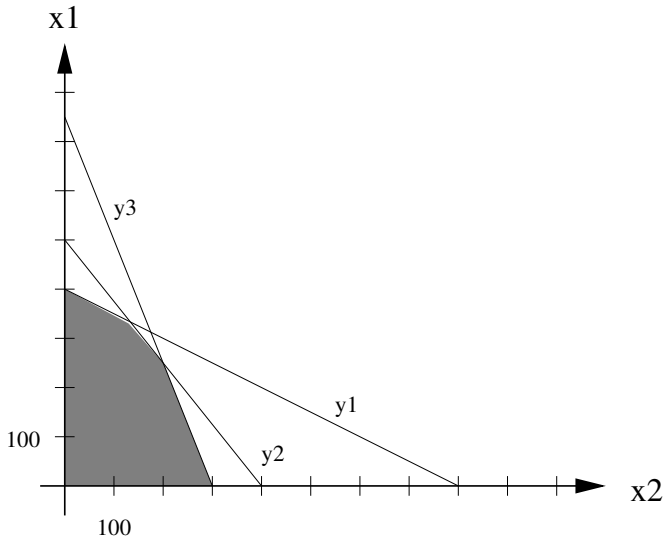

Abbildung 2.1: Zulässige Menge für das Optimierungsproblem aus Beispiel 2.1.

Definition 2.2 *Eine Menge $\Omega \subset \mathbb{R}^n$ heißt konvex, wenn sie mit je zwei Punkten auch ihre Verbindungsstrecke enthält, das heißt, wenn*

$$x, y \in \Omega \qquad \Rightarrow \qquad [x, y] := \{(1-\alpha)\ x + \alpha\, y \ : \ \alpha \in [0, 1]\} \subset \Omega. \tag{2.3}$$

Lemma 2.3 *Für jedes $A \in \mathbb{R}^{m\times n}$ und $b \in \mathbb{R}^m$ ist die Menge $F(A, b)$ konvex.*

Beweis: Der (einfache) Beweis ist eine Konsequenz der beiden folgenden Tatsachen:

1. *Halbräume sind konvex*: Ist $a^T x \geq b$ und $a^T y \geq b$, so gilt auch

$$a^T\left((1-\alpha)x + \alpha y\right) = (1-\alpha)\ \underbrace{a^T x}_{\geq b} + \alpha\ \underbrace{a^T y}_{\geq b} \geq (1-\alpha+\alpha)b = b.$$

[21] Das Übereinanderschreiben von Ungleichungsbedingungen bedeutet ja, daß sie *alle* gleichzeitig erfüllt werden sollen und die Lösungsmenge ist dann eben der Durchschnitt der einzelnen Lösungsmengen.

2. *Der Schnitt zweier konvexer Mengen ist konvex:* Sind Ω, Ω' konvex, dann ist, für $x, y \in \Omega \cap \Omega'$,

$$(1-\alpha)x + \alpha y \in \left\{ \begin{array}{c} \Omega \\ \Omega' \end{array} \right. , \quad \alpha \in [0,1] \qquad \Rightarrow \qquad [x,y] \subset \Omega \cap \Omega'.$$

□

Den Durchschnitt einer endlichen Anzahl von Halbräumen im $\mathbb{R}^n$ bezeichnet man als **konvexes Polyeder**, insbesondere ist also F ein solches konvexes Polyeder. Eine **Ecke** des Polyeders ist ein Punkt, den man *nicht* als Konvexkom-

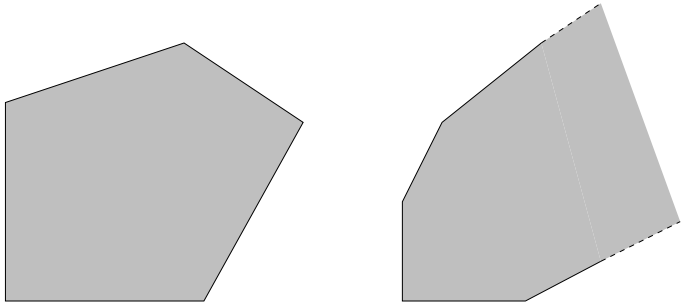

Abbildung 2.2: Ein endliches und ein unendliches konvexes Polyeder.

bination anderer Punkte des Polyeders schreiben kann. Formal heißt das, daß x eine Ecke ist, wenn

$$x \in (y, y'), \quad y, y' \in F \qquad \Leftrightarrow \qquad x = y = y'.$$

Dabei bezeichnet[22]

$$(y, y') = \{\alpha\, y + (1-\alpha)\, y' \ : \ \alpha \in (0,1)\}$$

das *relative Innere* der Strecke[23] $[y, y']$.
Die Ecken von $F(A, b)$ kann man nun sehr elegant mittels Linearer Algebra[24] beschreiben. Dazu wählt man eine Indexmenge $J \subset \{1, \ldots, m\}$, $\#J = n$ aus und betrachtet die **quadratische Teilmatrix**

$$A_J = \left[a_j^T \ : \ j \in J\right] \in \mathbb{R}^{n \times n}$$

und den **Teilvektor**

$$b_J = [b_j \ : \ j \in J] \in \mathbb{R}^n.$$

[22]Das sieht nun sehr stark wie ein *offenes* Intervall, was es aber nur ist, wenn $y \neq y'$ ist, ansonsten ist es einpunktig, abgeschlossen und **nicht** offen. Daß sich "offen" und "abgeschlossen" nicht notwendigerweise gegenseitig ausschließen ist ja hoffentlich bekannt.

[23]Genauer wäre die die **konvexe Hülle** der beiden Punkte.

[24]Für irgendwas muss die ja gut sein.

Lemma 2.4 *Sei* $F = F(A, b)$ *das konvexe Polyeder der zulässigen Punkte für das lineare Optimierungsproblem (2.1). Dann ist ein Punkt* $x \in F$ *genau dann ein Eckpunkt von* F*, wenn es eine Indexmenge* $J \subset \{1, \ldots, m\}$*,* $\#J = n$*, gibt, so daß* $A_J x = b_J$ *und* $\det A_J \neq 0$*.*

Beweis: „⇐": Ist $\det A_J \neq 0$, dann ist $x = A_J^{-1} b_J$ ein **Randpunkt**[25] des Polyeders[26]; wäre außerdem $x = (1-\alpha)y + \alpha y'$ für $y, y' \in F$ und $\alpha \in (0,1)$, dann wäre

$$b_J = A_J x = A_J\left((1-\alpha)y + \alpha y'\right) = (1-\alpha)\underbrace{A_J y}_{\geq b_J} + \alpha \underbrace{A_J y'}_{\geq b_J},$$

weswegen $A_J y = A_J y' = b_J$, also $y = y' = x$ sein muß. Damit ist x aber ein Eckpunkt.
„⇒": Jeder Eckpunkt x des konvexen Polyeders ist ein Randpunkt und liegt damit auf dem Rand mindestens eines Halbraums, erfüllt also $a_j^T x = b_j$ für mindestens ein $j \in \{1, \ldots, m\}$. Setzen wir

$$J := J(x) := \left\{1 \leq j \leq m \;:\; a_j^T x = b_j\right\} \subset \{1, \ldots, m\},$$

dann muß $A_J x = b_J$ sein; hat A_J Rang n, dann ist nach der obigen Argumentation x tatsächlich ein Eckpunkt, ansonsten gibt es einen mindestens eindimensionalen Teilraum $Y \subset \mathbb{R}^n$, so daß $A_J y = 0$, also $A_J(x+y) = b_J$ für alle $y \in Y$. Da alle anderen Ungleichungen strikt gelten, also

$$a_j^T x > b_j, \qquad j \notin J,$$

gilt, gibt es ein $\epsilon > 0$, so daß

$$\{x + y \;:\; y \in Y \| \|y\| \leq \epsilon\} \subset F,$$

und in dieser Menge kann man x konvex kombinieren. □

Allerdings: Diese Charakterisierung von Eckpunkten über Indexmengen J der Mächtigkeit n gilt nur, wenn x zum Polyeder gehört! Diese Information wurde im Beweis ja auch weidlich ausgenutzt.

Bemerkung 2.5 *Die Existenz einer Indexmenge* J*, so daß* $\det A_J \neq 0$ *ist, hat ja eine einfache geometrische Interpretation: Die Hyperebenen*

$$H_j := \left\{x \in \mathbb{R}^n \;:\; a_j^T x = b_j\right\}, \qquad j \in J,$$

schneiden sich in dem eindeutigen Punkt $x_J := A_J^{-1} b_J$*, andernfalls ist der Schnitt, je nach "rechter Seite"* b_J *leer oder ein nichttrivialer linearer Raum. Und im Normalfall liegen die meisten solchen Schnittpunkte eben* nicht *im Polyeder* $F(A, b)$*, siehe Abb. 2.3.*

[25] Also ein Punkt, zu dem es in beliebiger Nähe Punkte gibt, die *nicht* zum Polyeder gehören.
[26] Nach Voraussetzung ist $x \in F$, erfüllt also die anderen Ungleichungen ebenfalls, das heißt $A_K x \leq b_K$, $K = \{1, \ldots, m\} \setminus J$.

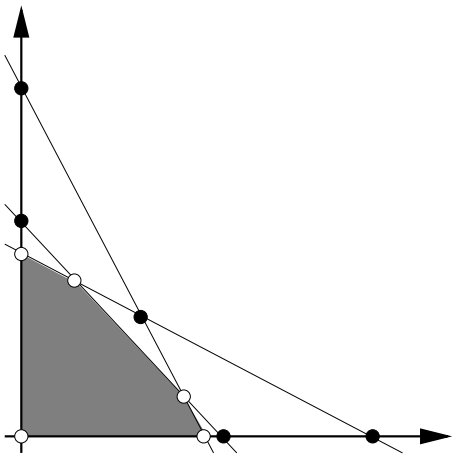

Abbildung 2.3: Alle Schnittpunkte der Nebenbedingungen eines einfachen linearen Optimierungsproblems – einige sind Ecken, einige nicht. Die zugehörige Nebenbedingungsmatrix $A \in \mathbb{R}^{5\times 2}$ hat übrigens die "Standard"–Eigenschaft, daß jede 2×2–Teilmatrix von A invertierbar ist.

Korollar 2.6 *Das konvexe Polyeder* $F(A,b)$ *hat höchstens* $\binom{m}{n}$ *Ecken.*

Beweis: Zu jeder Ecke gehören n linear unabhängige Zeilen von A. Da A insgesamt $m \geq n$ Zeilen hat, ist die Anzahl solcher Konfigurationen höchstens $\binom{m}{n}$. □

Jedes *endliche* konvexe Polyeder ist die konvexe Hülle seiner Eckpunkte[27]; dazu erinnern[28] wir uns, daß die **konvexe Hülle** einer Menge $\Omega \subset \mathbb{R}^n$ definiert ist als die kleinste konvexe Menge, die Ω enthält, also als diejenige Menge, die man erhält, wenn man jede beliebige endliche Konvexkombinationen von Punkten aus Ω bildet. Eine derartige **Konvexkombination** von $x_1,\ldots,x_m \in \mathbb{R}^n$ ist, als Verallgemeinerung von (2.3), ein Ausdruck der Form

$$\sum_{j=1}^{m} \alpha_j x_j, \qquad \alpha_j \geq 0, \quad \sum_{j=1}^{m} \alpha_j = 1,$$

und wir setzen

$$[x_1,\ldots,x_m] = \left\{ \sum_{j=1}^{m} \alpha_j x_j \ : \ \alpha_j \geq 0,\ \sum_{j=1}^{m} \alpha_j = 1 \right\}. \tag{2.4}$$

Die konvexe Hülle von Ω lässt sich dann als

$$[\Omega] := \{[x_1,\ldots,x_m] \ : \ x_1,\ldots,x_m \in \Omega,\ m \in \mathbb{N}\}.$$

[27] Das bedarf natürlich eines Beweises, aber den schenken wir uns hier.
[28] Oder lernen es neu.

beschreiben. Man kann $[\Omega]$ aber auch anders interpretieren, nämlich als Abschluß von Ω unter der Operation "Verbindungslinie" aus (2.3). Dazu betrachtet man die Folge

$$\Omega_0 = \Omega, \qquad \Omega_{j+1} = \Omega_j \cup \{[x, y] \ : \ x, y \in \Omega_j\},$$

und sieht leicht, daß

$$[\Omega] = \lim_{j\to\infty} \Omega_j. \tag{2.5}$$

Übung 2.2 Beweisen Sie (2.5). ◊

Um die konvexe Hülle einer endlichen Menge $X \subset \mathbb{R}^n$ etwas handlicher schreiben zu können, ordnen wir ihre Elemente als Spaltenvektoren einer Matrix X an[29] und erhalten, daß

$$[X] = \{Xu \ : \ u \in \Delta_{\#X}\}, \qquad \Delta_n = \left\{ u \in \mathbb{R}^n \ : \ u_j \geq 0, \ \sum_{j=1}^{n} u_j = 1 \right\}. \tag{2.6}$$

Die Menge Δ_n ist das n–dimensionale **Einheitssimplex** – weswegen man die konvexe Menge $[X]$ auch als **Simplex** bezeichnet.

In der Folge wollen wir nun immer annehmen, daß die zulässigen Punkte $F(A, b)$ ein **endliches Polyeder** bilden, das heißt, es gibt $N > 0$, so daß $F(A, b) \subseteq [-N, N]^n$, denn dann ist auch $F(A, b) = [X]$, wobei X die Eckenmenge von $F(A, b)$ ist. Das sollten wir auch beweisen, schon allein um zu verstehen, was da eigentlich passiert.

Proposition 2.7 *Ist* $F(A, b)$ *ein* endliches *Polyeder und* X *die zugehörige Eckenmenge, dann ist*[30] $F(A, b) \subseteq [X]$.

Beweis: Wir bemerken zuerst, daß jede Seite

$$F_J(A, b) = \{x \in F(A, b) \ : \ A_J x = b_J\}, \qquad \emptyset \neq J \subset \{1, \ldots, m\},$$

ebenfalls ein Polyeder ist und zwar eines von der Dimension $\leq n - \#J$, denn schließlich liegt dieses Polyeder ja im Schnitt von $\#J$ Hyperräumen[31].

Sei nun $x \in F(A, b)$, dann ist x entweder eine Ecke und die Aussage der Proposition ist trivialerweise erfüllt, oder es gibt $y \neq y' \in F(A, b)$ und $\alpha \in (0, 1)$, so daß $x = \alpha y + (1-\alpha)y'$. Die Gerade $\ell : t \mapsto \ell(t) = t\,y + (1-t)\,y'$, $t \in \mathbb{R}$, durch y und y' hat nun die Eigenschaft, daß

$$\ell \cap F(A, b) = \{\ell(t) \ : \ t \in [t_0, t_1]\} = \{\alpha\,\ell\,(t_0) + (1 - \alpha)\ (t_1)\}.$$

[29] Dies ermöglicht die Verwendung von *"Multisets"*, das sind "Mengen", in denen Vielfachheiten der Elemente erfasst und berücksichtigt werden können. Außerdem entspricht es auch sehr schön der `Matlab`–Philosophie.

[30] Die Umkehrung, $[X] \subseteq F(A, b)$ ist klar, denn es ist $X \subset F(A, b)$ und da $F(A, b)$ konvex ist, muss die konvexe Hülle $[X]$ als *kleinste* konvexe Obermenge von X ebenfalls in $F(A, b)$ enthalten sein. Wir behaupten und beweisen also in dieser Proposition nur den "interessanten" Teil der Aussage $F(A, b) = [X]$.

[31] Wobei wir redundante Nebenbedingungen ausschließen wollen – also den Fall, daß A zwei oder mehr identische oder abhängige Zeilen enthält.

Hier haben wir die Beschränktheit des Polyeders verwendet, denn sonst könnte im schlimmsten Fall die Gerade komplett im Inneren des Polyeders verlaufen. Die beiden Punkte $\ell\left(t_{0/1}\right)$ sind Randpunkte von $F(A,b)$, gehören also zu Seiten des Polyeder und sind so – per Induktion[32] über die Dimension der Seite – Konvexkombinationen von Ecken. Da aber x auch eine Konvexkombination dieser beiden Punkte ist, gilt $x \in [X]$. □

2.2 Konvexe Funktionen

Auf konvexen Polyedern besonders einfach zu optimieren sind **konvexe Funktionen**.

Definition 2.8 (Konvexe Funktionen) *Eine Funktion $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}$ heißt **konvex**, wenn für alle $x, y \in \Omega$ und alle $\alpha \in [0,1]$*

$$f\left((1-\alpha)x + \alpha y\right) \leq (1-\alpha)\, f(x) + \alpha\, f(y) \tag{2.7}$$

erfüllt ist.

Bemerkung 2.9 (Konvex & konkav)

1. *Eine Funktion f heißt* konkav, *wenn $-f$ konvex ist. Alle Aussagen über Maxima konvexer Funktionen ergeben sofort auch Aussagen über* Minima *konkaver Funktionen.*

2. *Mit der Notation (2.6) kann man Konvexität einer Funktion auch als*

$$f\left(Xu\right) \leq \sum_{j=1}^{N} u_j f\left(x_j\right), \qquad u \in \Delta_N,\ X = [x_1 \cdots x_N] \in \mathbb{R}^{n\times N}, \tag{2.8}$$

   beschreiben.

3. *Für eine **affine Funktion** der Form $x \mapsto a^T x + b$ gilt in (2.8) Gleichheit, das heißt, affine Funktionen sind gleichzeitig konvex und konkav.*

Das folgende Resultat sagt uns, wo wir nach den Optimallösungen suchen müssen – in den Ecken des konvexen Polyeders.

Satz 2.10 (Konvexe Funktionen auf konvexen Polyedern) *Eine konvexe Funktion nimmt auf einem endlichen konvexen Polyeder ihr Maximum in einer der Ecken an.*

Beweis: Sei $X \in \mathbb{R}^{n\times N}$ die Eckenmenge des konvexen Polyeders, das heißt, jeder Punkt $x \in [X]$ hat die Form $x = Xu$, $u = u(x) \in \Delta_N$. Wegen der Konvexität von f ist dann

$$f(x) = f\left(Xu\right) \leq \sum_{j=1}^{N} u_j f\left(x_j\right) \leq \underbrace{\left(\sum_{j=1}^{N} u_j\right)}_{=1} \max_{j=1,\ldots,N} f\left(x_j\right) \leq \max_{j=1,\ldots,N} f\left(x_j\right).$$

[32] Der Induktionsanfang ist einfach: Eindimensionale Seiten sind Intervalle, also sicherlich Konvexkombinationen ihrer Ecken, der Endpunkte des Intervalls.

□

Dieser Satz gibt eine Möglichkeit, das Optimierungsproblem zu lösen: Man braucht ja „nur“ die Ecken von F zu bestimmen, sich die Zielfunktionen an diesen anzusehen und wird unter diesen *endlich vielen* Werten auch das Maximum finden. Dieser Ansatz ist aber nicht praktikabel, denn das Auffinden der Ecken, das heißt, die Bestimmung von n linear unabhängigen Zeilen in der Matrix A, dauert einige Zeit. Aber viel schlimmer ist, daß dann auch sehr schnell numerische Probleme auftreten, denn es ist eigentlich nicht möglich, diese Gleichheiten numerisch exakt zu bestimmen; nachdem numerisch praktisch jede Matrix invertierbar ist, ist auch die Unterscheidung mit dem $\det A_J \neq 0$ in der Praxis völliger Unsinn, zumal die Determinante einer Matrix kein gutes Maß für deren Invertierbarkeit ist. Außerdem entspricht ja nicht für jede Indexmenge J so daß $\det A_J \neq 0$ die Lösung von $A_J x = b_J$ einer Ecke von F – es kann und wird, wie in Bemerkung 2.5 gezeigt, in vielen Fällen passieren, daß $(Ax)_j < b_j$ für ein $j \in \{1, \ldots, m\} \setminus J$ gilt und damit $x \notin F$ ist.

2.3 Der Simplexalgorithmus

Der Simplexalgorithmus, der auf Dantzig[33] zurückgeht, siehe (Dantzig, 1963), stellt wohl eines der wichtigsten numerischen Verfahren dar[34] So schreibt Laszlo Lovasz[35] 1980:

> *If one would take statistics about which mathematical problem is using up most of the computer time in the world, then ... the answer would probably be linear programming.*

Dantzig selbst bemerkt

> *The tremendous power of the simplex method is a constant surprise to me.*

Das Verfahren nutzt ebenfalls die Tatsache aus, daß die *lineare* und damit sowohl konvexe als auch konkave Zielfunktion ihr Extremum in einer Ecke des konvexen Polyeders F annehmen muß. Anstatt nun alle Ecken des Polyeders der Reihe nach abzusuchen, wird bei diesem Verfahren zu einer bekannten Ecke eine „Nachbarecke“ bestimmt, an der die Zielfunktion einen größeren Wert annimmt – auf diese Weise hofft man, sich relativ schnell und systematisch bis zum Maximum vorzuarbeiten. Das muß natürlich nicht unbedingt rasend schnell funktionieren: Man kann Beispiele angeben, bei denen der Simplexalgorithmus *alle* Ecken ablaufen muß, bevor er das Optimum erreicht und insofern nicht schneller als „systematisches“ Suchen ist.

[33] George Dantzig, 1918–2005, entwickelte dieses „mechanisierte“ Planungsverfahren unter dem Namen **Linear Programming** 1947 für die U.S. Air Force. Später arbeitete er für die bekannte *RAND corporation* und wurde 1966 Professor für Operations Research in Stanford. Ein Nachruf auf ihn erschien 2005 sogar im *Time Magazine*.

[34] Nummer 1 ist fraglos die schnelle Fouriertransformation von Cooley und Tukey (Cooley & Tukey, 1965).

[35] Wer auch immer das ist.

Um richtig konkrete Aussagen machen zu können, müssen wir aber zuerst einmal formal klarstellen, was eine **Nachbarecke** eigentlich ist. Betrachtet man eine Ecke x des konvexen Polyeders F mit Eckenmenge X, dann können wir die dazugehörige Indexmenge

$$J = J(x) = \left\{ j \in \{1, \ldots, m\} \ : \ a_j^T x = b_j \right\}$$

mit $\#J \geq n$ einführen. Geometrisch ist x ja gerade der Schnittpunkt der Hyperebenen, die durch $J(x)$ indiziert sind, also

$$x = \bigcap_{j \in J(x)} \left\{ y \ : \ a_j^T y = b \right\}.$$

Eine **Nachbarecke** y von x ist nun gerade ein Element von X, das mit x durch eine **Kante** verbunden ist; eine Kante ist aber wiederum der Schnitt von $n-1$ Hypereneben und da x und y auf dieser Kante liegen sollen, müssen $n-1$ der definierenden Hyperebenen übereinstimmen. Die Menge der Nachbarecken von x ist also

$$V_x := \{ y \in X \setminus \{x\} \ : \ \#(J(x) \cap J(y)) \geq n-1 \}, \tag{2.9}$$

das heißt, mindestens $n-1$ der Ungleichungen in $A_{J(x)} y \geq b_{J(x)}$ müssen zur Gleichheit werden. Man kann (2.9) dann aber auch anders interpretieren: Man muss aus $J(x)$ einen Index weglassen, um dann den Schnitt zu $J(y)$ erhalten, man kann also die Nachbarecken auch mit $J(x)$ indizieren. Das nutzen wir jetzt aus.

Proposition 2.11 (Nachbarecken) *Sei x^* eine Ecke des konvexen Polyeders F und seien x_j, $j \in J$, die Nachbarecken von x^*. Ist*

$$c^T x_j < c^T x^*, \qquad j \in J,$$

dann ist $c^T x < c^T x^*$ für alle $x \in F$.

Beweis: Die Menge

$$C_{x^*} := x^* + \left\{ \sum_{j \in J} \lambda_j \ (x_j - x^*) \ : \ \lambda_j \geq 0 \right\},$$

ist ein **konvexer Kegel**, der von x und V_x aufgespannt wird und das konvexe Polyeder F enthält. Daher können wir jedes $x \in F \setminus \{x^*\}$ als

$$x = x^* + \sum_{j \in J} \lambda_j \ (x_j - x^*), \qquad \lambda \neq 0,$$

schreiben und erhalten somit, daß

$$c^T x = c^T x^* + \sum_{j \in J} \lambda_j \underbrace{\left(c^T x_j - c^T x^* \right)}_{<0} < c^T x^*,$$

wie behauptet. □

Bemerkung 2.12 *Ersetzen wir die strikte Ungleichung „<“ in Proposition 2.11 durch „≤“, dann gilt die Aussage immer noch! Mit anderen Worten: Wir haben* ein *Maximum erreicht, wenn wir uns beim Übergang zu Nachbarecken nicht verbessern können, und wir haben* ***das*** *Maximum erreicht, wenn wir uns beim Übergang zu Nachbarecken nur verschlechtern können*[36].

Um uns das Leben einfacher zu machen, nehmen wir erst einmal an, daß der Nullpunkt eine **zulässige Ecke** von F ist, und zwar dergestalt, daß

$$A = \begin{bmatrix} I_n \\ \widetilde{A} \end{bmatrix}, \qquad b = \begin{bmatrix} 0 \\ \widetilde{b} \end{bmatrix}, \qquad \widetilde{A} \in \mathbb{R}^{m-n\times n}, \quad 0 \geq \widetilde{b} \in \mathbb{R}^{m-n}. \tag{2.10}$$

Die Nebenbedingungen (2.10) sind äquivalent zu

$$\widetilde{A}x \geq \widetilde{b}, \qquad x \geq 0, \tag{2.11}$$

und unser Ausgangsproblem lässt sich auch immer so transformieren: Ist 0 eine Ecke, so gibt es (unter Umständen nach Permutation der Zeilen) eine invertierbare Matrix B, so daß

$$\begin{bmatrix} 0 \\ \widetilde{b} \end{bmatrix} = b \leq Ax = \begin{bmatrix} B \\ C \end{bmatrix} x = \begin{bmatrix} I \\ CB^{-1} \end{bmatrix} Bx := \begin{bmatrix} I \\ CB^{-1} \end{bmatrix} x'$$

mit dem „Taschenspielertrick“ $x' := Bx$. Die entsprechende Zielfunktion ist dann

$$c^T x = c^T B^{-1} x' = \left(B^{-T} c\right)^T x' =: c'^T x,$$

und so lässt sich jedes Optimierungsproblem mit einer Ecke an $x = 0$ in die Form (2.10) bzw. (2.11) darstellen:

$$\max c'^T x, \qquad A'x \geq 0, \quad x \geq 0, \qquad \text{mit} \qquad c' = B^{-T} c, \quad A' = CB^{-1}. \tag{2.12}$$

Die Vorteile von (2.11) liegen auf der Hand: Wir brauchen jetzt in unserem Optimierungsproblem n Zeilen weniger zu betrachten und haben eine weitere **Normalform** mit der sich sehr einfach rechnen läßt, da die Suche nach Nachbarecken jetzt wegen der Nebenbedingung $x \geq 0$ entlang der Koordinatenachsen verläuft.

> Die einzige Annahme, die wir jetzt also zur Normalisierung unseres linearen Optimierungsproblems gemacht haben, ist daß 0 eine zulässige Ecke ist. Das ist, wie wir sehen werden, nicht immer der Fall, aber immer behebbar.

Es mag banal klingen, aber hier ist eines der ganz wichtigen Konzepte in der angewandten Mathematik: *Bringe ein Problem auf die einfachste Form, die immer noch hinreichend allgemein ist.*

[36] Also fast wie in der Realität.

Die „Startecke“ $x^{(0)} = 0$ ist in dieser Konfiguration dann natürlich durch die Indexmenge $J = \{1, \ldots, n\}$ charakterisiert. Außerdem verwenden wir die rein formale Bezeichnung

$$y := \widetilde{A}x - \widetilde{b},$$

also

$$\begin{bmatrix} x \\ y \end{bmatrix} = Ax - b, \qquad x, y \geq 0. \tag{2.13}$$

Damit haben die Hyperebenen, die $F(A, b)$ beranden, die einfache Form

$$\{ x_j = 0 \}, \quad j = 1, \ldots, n, \qquad \text{und} \qquad \{ y_j = 0 \}, \quad j = n+1, \ldots, m.$$

Bevor wir uns an den formalen Teil machen, wollen wir uns erst einmal die Idee hinter der „Strategie“ ansehen an einem Beispiel ansehen: Wie in Abb. 2.4 zu

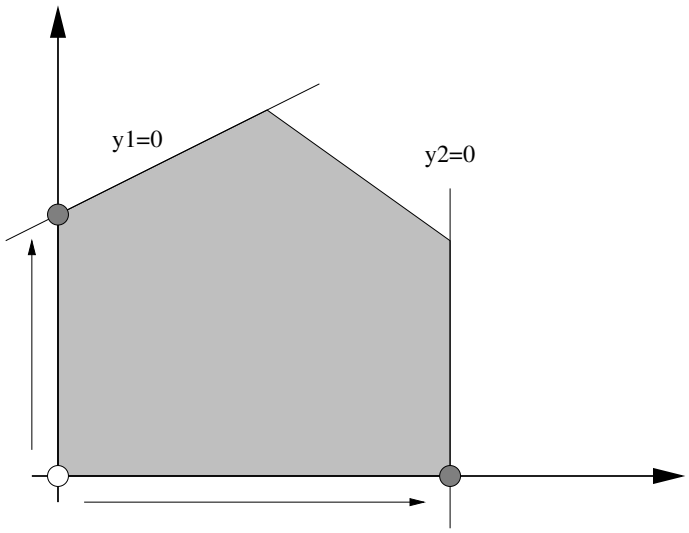

Abbildung 2.4: Benachbarte Ecken des Ursprungs.

sehen ist, hat die zulässige Ecke $x^{(0)} = 0$ zwei Nachbarecken, nämlich

$$\{ x_1 = 0 \} \cap \{ y_1 = 0 \} \qquad \text{und} \qquad \{ x_2 = 0 \} \cap \{ y_2 = 0 \}.$$

Diese beiden Ecken erhalten wir, indem wir entweder x_2 durch y_1 oder x_1 durch y_2 ersetzen, wir vertauschen also die Rollen von zweien der „Variablen“ x_j und y_j und erhalten so die gesuchte Nachbarecke. Unter all diesen Nachbarecken wählen wir dann diejenige als neue „Startecke“ aus, an der die Zielfunktion den größeren[37] Wert annimmt.

Um den **Austauschschritt** durchzuführen, der die *formalen* Variablen x_j und y_k vertauscht, wählen wir $j \in \{1, \ldots, n\}$ und $k \in \{1, \ldots, m - n\}$ und bemerken zuerst einmal, daß wir die Gleichung

$$y_k = a_{n+k}^T x - b_{n+k}$$

[37] Zumindest wenn wir maximieren wollen. Bei Minimierungsproblemen wird der Auswahlprozess und die dazugehörige Regel sehr viel komplexer.

genau dann nach x_j auflösen können, wenn $a_{n+k,j} \neq 0$ ist. In diesem Fall ist

$$x_j = \frac{1}{a_{n+k,j}} \left(y_k - \sum_{\ell \neq j} a_{n+k,\ell}\, x_\ell + b_{n+k} \right), \tag{2.14}$$

was wir in Matrixform als

$$x = \underbrace{\begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ -\frac{a_{n+k,1}}{a_{n+k,j}} & \cdots & -\frac{a_{n+k,j-1}}{a_{n+k,j}} & \frac{1}{a_{n+k,j}} & -\frac{a_{n+k,j+1}}{a_{n+k,j}} & \cdots & -\frac{a_{n+k,n}}{a_{n+k,j}} \\ & & & & 1 & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{bmatrix}}_{=:B} \underbrace{\begin{bmatrix} x_1 \\ \vdots \\ x_{j-1} \\ y_k \\ x_{j+1} \\ \vdots \\ x_n \end{bmatrix}}_{=:z} + \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \frac{b_{n+k}}{a_{n+k,j}} \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

schreiben können, also

$$x = Bz + \frac{b_{n+k}}{a_{n+k,j}}\, e_j. \tag{2.15}$$

Setzt man das in (2.13) ein, dann ergibt sich

$$\begin{aligned} \begin{bmatrix} x \\ y \end{bmatrix} &= Ax - b = \begin{bmatrix} I_n \\ \widetilde{A} \end{bmatrix} \left(Bz + \frac{b_{n+k}}{a_{n+k,j}}\, e_j \right) - \begin{bmatrix} 0 \\ \widetilde{b} \end{bmatrix} \\ &= \underbrace{\begin{bmatrix} B \\ \widetilde{A}\, B \end{bmatrix}}_{=:C} z + \underbrace{\frac{b_{n+k}}{a_{n+k,j}}\, Ae_j}_{=:-\widehat{b}} - b = Cz - \left(b + \widehat{b} \right), \end{aligned} \tag{2.16}$$

wobei[38]

$$\widehat{b} = -\frac{b_{n+k}}{a_{n+k,j}} \begin{bmatrix} a_{1,j} \\ \vdots \\ a_{m,j} \end{bmatrix} = -\frac{b_{n+k}}{a_{n+k,j}} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ a_{n+1,j} \\ \vdots \\ a_{m,j} \end{bmatrix} - \frac{b_{n+k}}{a_{n+k,j}}\, e_j$$

wegen $j \leq n$ für $k \geq 1$ die Eigenschaft

$$\widehat{b}_{n+k} = -\frac{b_{n+k}}{a_{n+k,j}} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ a_{n+1,j} \\ \vdots \\ a_{m,j} \end{bmatrix}_{n+k} - \frac{b_{n+k}}{a_{n+k,j}} \left(e_j \right)_{n+k} = -\frac{b_{n+k}}{a_{n+k,j}}\, a_{n+k,j} = -b_{n+k}$$

[38] Das „e_j“ kommt von der Einheitsmatrix oben in A.

hat. Da daher $y_k = (Cz)_{n+k}$, ist die $(n+k)$–te Zeile von C, also c_{n+k}^T, von besonders einfacher Form:

$$c_{n+k}^T = e_j^T.$$

Vertauschen wir also die Zeilen j und $n+k$ in (2.16), so können wir (2.16) unter Verwendung der Notation

$$x^{(1)} := z = \begin{bmatrix} x_1 \\ \vdots \\ x_{j-1} \\ y_k \\ x_{j+1} \\ \vdots \\ x_n \end{bmatrix} \qquad \text{und} \qquad y^{(1)} := \begin{bmatrix} y_1 \\ \vdots \\ y_{k-1} \\ x_j \\ y_{k+1} \\ \vdots \\ y_{m-n} \end{bmatrix}$$

umschreiben in

$$\begin{bmatrix} x^{(1)} \\ y^{(1)} \end{bmatrix} = A^{(1)}x^{(1)} - b^{(1)} = \begin{bmatrix} I_n \\ \widetilde{A}^{(1)} \end{bmatrix} x^{(1)} - b^{(1)}, \tag{2.17}$$

wobei[39] $b^{(1)} = b + \widehat{b}$. Dann sind natürlich die beiden Nebenbedingungen, das heißt die Ungleichungssysteme $Ax \geq b$ und $A^{(1)}x^{(1)} \geq b^{(1)}$ äquivalent, oder, anders gesagt, die Polyeder $F(A, b)$ und $F\left(A^{(1)}, b^{(1)}\right)$ sind gleich und nur unterschiedlich beschrieben.

Wenn wir es jetzt noch hinbekommen, daß der Nullpunkt wieder eine zulässige Ecke des konvexen Polyeders $F\left(A^{(1)}, b^{(1)}\right)$ ist, das heißt, daß

$$0 = A^{(1)}0 \geq b^{(1)} \qquad \Rightarrow \qquad b^{(1)} \leq 0,$$

dann haben wir, via Austausch, tatsächlich den Schritt von der Nullecke zur einer benachbarten Ecke geschafft, und zwar so, daß diese neue Ecke wieder der Nullpunkt eines umgeschriebenen Systems ist. Dazu bemerken wir zuerst, daß[40]

$$b_j^{(1)} = b_{n+k} + \widehat{b}_{n+k} = b_{n+k} - \frac{b_{n+k}}{a_{n+k,j}}\, a_{n+k,j} = 0$$

und, trivialerweise, $b_\ell^{(1)} = 0$, $\ell \in \{1, \ldots, n\} \setminus \{j\}$, und $x^{(1)} = 0$ ist zumindest schon einmal ein *Kandidat* für eine Ecke des Polyeders – allerdings muß dieser Punkt auch zulässig sein, um wirklich eine Ecke darzustellen. Dies ist nun genau dann der Fall, wenn $b^{(1)} \leq 0$ ist, also wenn

$$0 \geq b_{n+k}^{(1)} = \underbrace{b_j}_{=0} - \frac{b_{n+k}}{a_{n+k,j}} \underbrace{a_{jj}}_{=1} = -\frac{b_{n+k}}{a_{n+k,j}} \tag{2.18}$$

und

$$0 \geq b_{n+\ell}^{(1)} = b_{n+\ell} - \frac{a_{n+\ell,j}}{a_{n+k,j}}\, b_{n+k}, \qquad \ell = 1, \ldots, m-n, \quad \ell \neq k. \tag{2.19}$$

[39] Nach Vertauschung der Komponenenten j und $n+k$ von b!

[40] Nicht vergessen: Die Zeilen j und $n+k$ wurden vertauscht!

Aus (2.18) und der Forderung $b \leq 0$ folgt, daß[41]

$$a_{n+k,j} < 0. \tag{2.20}$$

Die Bedingung (2.19) läßt sich hingegen zuerst einmal in

$$\frac{a_{n+\ell,j}}{a_{n+k,j}}\, b_{n+k} \geq b_{n+\ell}, \qquad \ell = 1, \ldots, m-n,\ \ell \neq k,$$

umformen, was immer erfüllt ist, wenn $a_{n+\ell,j} \geq 0$ ist. Im anderen Fall erhalten wir, daß

$$\frac{b_{n+k}}{a_{n+k,j}} \leq \frac{b_{n+\ell}}{a_{n+\ell,j}}, \qquad \text{falls} \quad a_{n+\ell,j} < 0, \tag{2.21}$$

sein muß. Also haben wir die folgende Regel zur Bestimmung von k (für ein gegebenes $j \in \{1, \ldots, n\}$):

> *Die* ***Pivotzeile*** $n + k$ *ist so zu bestimmen, daß* $a_{n+k,j} < 0$ *und daß der* positive *Quotient*
>
> $$\frac{b_{n+\ell}}{a_{n+\ell,j}}, \qquad a_{n+\ell,j} < 0, \qquad \ell = 1, \ldots, m-n,$$
>
> *minimiert wird, also*
>
> $$\frac{b_{n+k}}{a_{n+k,j}} = \min\left\{ \frac{b_{n+\ell}}{a_{n+\ell,j}} \ :\ a_{n+\ell,j} < 0,\ \ell = 1, \ldots, m-n \right\}. \tag{2.22}$$

Bleibt also noch die Frage, wie man diese ominöse Spalte j wählt, also welche der Variablen man austauscht. Hier kommt jetzt die *Vergrößerung* der Zielfunktion ins Spiel. Zu diesem Zweck setzen wir (2.14) in die Zielfunktion

$$c^T x = \sum_{\ell \neq j} c_\ell\, x_\ell + \frac{c_j}{a_{n+k,j}} \left(y_k - \sum_{\ell \neq j} a_{n+k,\ell}\, x_\ell + b_{n+k} \right),$$

ein, das heißt,

$$c^T x = \frac{c_j}{a_{n+k,j}}\, y_k + \sum_{\ell \neq j} \left(c_\ell - \frac{a_{n+k,\ell}}{a_{n+k,j}}\, c_j \right) x_\ell + \frac{c_j}{a_{n+k,j}}\, b_{n+k} =: {c^{(1)}}^T\, x^{(1)} + d^{(1)},$$

setzen $x^{(1)} = 0$ und erhalten eine Verbesserung[42] gegenüber dem Ausgangswert, falls

$$0 \leq d^{(1)} = \frac{c_j}{a_{n+k,j}}\, b_{n+k}, \tag{2.23}$$

also wenn $c_j \geq 0$ ist, was zu der folgenden Regel führt:

[41] Genaugenommen könnte $a_{n+k,j}$ machen, was es will, wenn $b_{n+k} = 0$ ist; da aber immer durch eine beliebig kleine zulässige Störung $b_{n+k} < 0$ erreicht werden kann, können wir dies hier auch annehmen.

[42] oder zumindest keine Verschlechterung

*Die **Pivotspalte** j ist so zu bestimmen, daß $c_j > 0$ ist.*

Diese beiden Auswahlregeln für j und k können erfüllt werden, solange

1. die Matrix $\widetilde{A}$ **negative Werte** enthält,
2. eine Spalte von $\widetilde{A}$ existiert, die einen negativen Wert enthält und einem nichtnegativen Eintrag von c entspricht.

Ist eine dieser beiden Bedingungen verletzt, so hat dies auch eine Bedeutung:

1. Ist $\widetilde{A} \geq 0$, dann ist $Ax \geq 0$ wann immer $x \geq 0$ und gibt es auch nur ein zulässiges $x^* \geq 0$, dann ist für alle $x \geq 0$ auch
$$A(x^* + x) = \underbrace{Ax^*}_{\geq b} + \underbrace{Ax}_{\geq 0} \geq b,$$
und somit ist $F(A,b) \subseteq x^* + \mathbb{R}^n_+$. Mit anderen Worten: *Das Polyeder $F(A,b)$ ist **unbeschränkt** und ist auch nur ein $c_j > 0$, dann ist auch die Zielfunktion unbeschränkt. Ansonsten ist ihr Maximum ohnehin 0.*
2. Ist $c_j \leq 0$ für alle Spalten von A, die negative Werte enthalten, dann wird beim Übergang zu allen benachbarten Ecken die Zielfunktion nur verkleinert, also: *Wir haben ein **Maximum** gefunden.* Damit kann nach Proposition 2.11 und Bemerkung 2.12 der Algorithmus beendet werden.

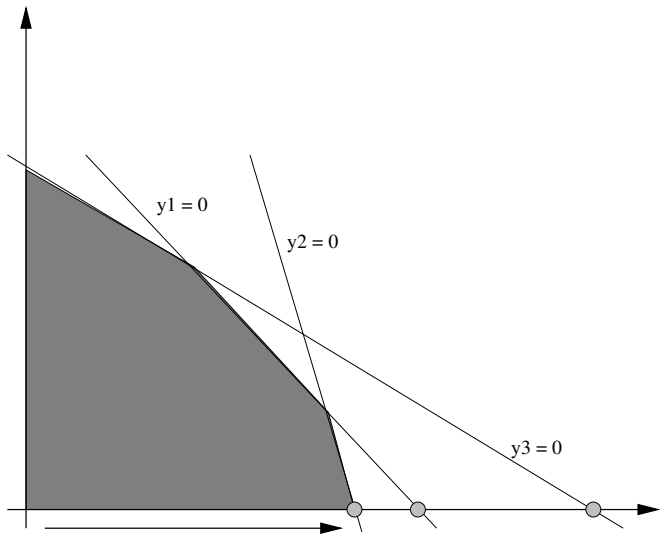

Abbildung 2.5: Geometrische Interpretation der Bedingungen für die Wahl der Austauschparameter. Unter allen „guten“ Schnittpunkten wird derjenige gewählt, der am nächsten bei der Ecke $(0,0)$ liegt.

Noch kurz zur *geometrischen* Interpretation der Bedingungen zur Bestimmung der Pivotzeile[43] $n+k$: Die Bedingung $a_{n+k,j} < 0$ bedeutet, daß man nur

[43] Eigentlich wird natürlich nur k bestimmt, in der Praxis werden wir später die Dauernebenbedingung $x \geq 0$ ohnehin weglassen.

nach Hyperebenen $y_k = 0$ sucht, die „vernünftig“ sind und im Oktanten $x \geq 0$ liegen. Die Minimumsbedingung hingegen ist dafür zuständig, daß unter allen Schnitten der Geraden[44] $x_1 = \cdots = x_{j-1} = x_{j+1} = \cdots = x_n = 0$ mit derartigen Hyperebenen diejenige gewählt wird, die als erste erreicht wird.

Bleibt uns noch die *Bestimmung* der Optimallösung, was aber jetzt einfach ist: Das Problem wurde ja nach r Schritten so modifiziert, daß $x^{(r)} = 0$ die **extremale Ecke** des Problems

$$\max x^T c^{(r)}, \qquad A^{(r)} x \geq b^{(r)}$$

ist. Das heißt aber nach (2.10), daß

$$\begin{bmatrix} x^{(r)} \\ y^{(r)} \end{bmatrix} = A^{(r)} 0 - b^{(r)} = -b^{(r)} = \begin{bmatrix} 0 \\ \widetilde{b}^{(r)} \end{bmatrix},$$

also

$$x^{(r)} = 0, \qquad y^{(r)} = -\widetilde{b}^{(r)}.$$

Um die wirkliche Lösung zu bekommen müssen wir jetzt nur noch nachschauen, in welche Komponenten von $x^{(r)}$ und $y^{(r)}$ die Variablen $x_1, \ldots, x_n$ getauscht wurden.

2.4 Die Implementierung

Als nächstes wollen wir den eben hergeleiteten Simplexalgorithmus in `Matlab` bzw. `octave` implementieren und versuchen, damit unser Beispiel vom Anfang zu lösen. Dazu “vergessen” wir die Einheitsmatrix “oben” in A, schreiben m für $m - n$ und setzen

$$\begin{aligned} c^{(0)} &= (c, 0) = \left(c, d^{(0)}\right) \quad \in \mathbb{R}^{n+1}, \\ A^{(0)} &= \widetilde{A} \quad \in \mathbb{R}^{m \times n}, \\ b^{(0)} &= \widetilde{b} \quad \in \mathbb{R}^{m}. \end{aligned}$$

Ein Austauschschritt für die Matrix $A^{(r)}$ hat nun, für gegebene $j \in \{1 \ldots, n\}$ und $k \in \{1, \ldots, m\}$ die folgende Gestalt:

1. Setze

$$c_\ell^{(r+1)} = \begin{cases} \dfrac{c_j^{(r)}}{a_{kj}^{(r)}}, & \ell = j, \\ c_\ell^{(r)} - \dfrac{a_{k\ell}^{(r)}}{a_{kj}^{(r)}}\, c_j^{(r)}, & \ell \in \{1, \ldots, n\} \setminus \{j\}, \\ c_{n+1}^{(r)} + \dfrac{c_j}{a_{kj}^{(r)}}\, b_k^{(r)}, & \ell = n+1. \end{cases} \tag{2.24}$$

[44] Das ist nichts anderes als die Koordinatenrichtung.

2. Setze

$$b_\ell^{(r+1)} = \begin{cases} -\dfrac{b_k^{(r)}}{a_{kj}^{(r)}}, & \ell = k, \\ b_\ell^{(r)} - \dfrac{a_{\ell j}^{(r)}}{a_{kj}^{(r)}}\, b_k^{(r)}, & \ell \neq k, \end{cases} \qquad \ell = 1, \dots, m. \tag{2.25}$$

3. Setze

$$B = \begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ -\frac{a_{k1}}{a_{kj}} & \cdots & -\frac{a_{k,j-1}}{a_{kj}} & \frac{1}{a_{kj}} & -\frac{a_{k,j+1}}{a_{kj}} & \cdots & -\frac{a_{kn}}{a_{kj}} \\ & & & & 1 & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{bmatrix} \in \mathbb{R}^{n\times n} \tag{2.26}$$

4. Berechne die Matrix $A^{(r)}B$, ersetze deren k–te Zeile durch die k–te Zeile von B und nenne diese Matrix $A^{(r+1)}$.

Diese Operationen sind in `Austausch.m`, Programm 2.1, implementiert. Die Reihenfolge oben ist absichtlich so gewählt, da die Berechnung von $c^{(r+1)}$ sowohl $b^{(r)}$, $c^{(r)}$ wie auch $A^{(r)}$ benötigt, die Bestimmung von $b^{(r+1)}$ lediglich $b^{(r)}$ und $A^{(r)}$ und $A^{(r+1)}$ schließlich aus $A^{(r)}$ berechnet werden kann, das heißt, in dieser Reihenfolge können die Variablen überschrieben werden.

Bemerkung 2.13 *Man kann sich die schematischen Regeln des Simplexalgorithmus recht einfach merken*[45]*:*

1. *Dividiere die Pivotzeile durch das Pivotelement.*
2. *Dividiere die Pivotspalte durch das Negative des Pivotelements.*
3. *Für alle anderen Elemente verwende die „**Rechtecksregel**“:*

	$\dots$	x_j	$\dots$	x_q	$\dots$	
$\vdots$	$\ddots$	$\vdots$		$\vdots$		$\vdots$
y_k	$\dots$	a_{jk}	$\dots$	a_{jq}	$\dots$	b_j
$\vdots$		$\vdots$	$\ddots$	$\vdots$		$\vdots$
y_p	$\dots$	a_{pk}	$\dots$	$\boxed{a_{pq}}$	$\dots$	b_p
$\vdots$		$\vdots$		$\vdots$	$\ddots$	$\vdots$
	$\dots$	c_k	$\dots$	c_q	$\dots$	

$$a_{pq} \longleftarrow a_{pq} - \frac{a_{pk}\, a_{jq}}{a_{jk}}$$

[45] Wenn man den Simplexalgorithmus denn unbedingt manuell durchführen möchte, was in Anbetracht von Programmen wie `Octave` und der Verfügbarkeit exzellenter Implementierungen wie `lpsolve` eigentlich Zeitvergeudung ist.

```
%%
%% Austausch.m
%% Austauschschritt fuer Simplexverfahren
%% Daten:
%%   j  Spaltenindex
%%   k  Zeilenindex
%%   A  Matrix des Problems
%%   b  Nebenbedingungen
%%   c  Zielfunktion

function [AA,bb,cc] = Austausch( j,k,A,b,c )
  [m,n] = size(A);
  a = A(k,j); cj = c( j ); bk = b( k );
  cc = zeros( n+1,1 );

  %% Update von c
  cc(1:n) = c(1:n) - (cj /  a) * A( k,: )';
  cc(j) = cj / a;
  cc(n+1) = c(n+1) + cj / a * b(k);

  %% Update von b
  bb = b - bk / a * A( :,j );
  bb(k) = -bk / a;

  %% Update von A
  B = eye( n );
  B( j,: ) = (-1 / a) * A( k, : );
  B(j,j) = 1 / a;
  AA = A * B;
  AA( k,: ) = B( j,: );

%endfunction
```

Programm 2.1 `Austausch.m`: Ein Austauschschritt.

```
%% auxFindjk.m (Optimierung)
%% Auffinden passender Werte j,k
%% Daten:
%%   A  Matrix des Problems
%%   b  Rechte Seite
%%   c  Zielfunktion

function [j,k] = auxFindjk( A,b,c )
  if min( min( A ) ) > 0                 %% kleinster Eintrag
    disp( "*** Unbeschraenkt ***" );
    j = 0; k = 0;
    return;
  end

  cA = find( min(A) < 0 );               %% Spalten mit neg. Eintrag
  cc = find( c( cA ) > 0 );              %% dort c > 0
  if length( cc != 0 )
    j = cA( cc( 1 ) );                   %% Erste Spalte - warum nicht?
    cA = find( A( :,j ) < 0 );           %% Negative Eintraege in Spalte
    [m,k] = min( b( cA ) ./ A( cA,j ) ); %% Lokalisiere Minimum
    k = cA( k );                         %% k ist entsprechende Spalte
  else
    j = 0; k = 0;
  end
```

Programm 2.2 `auxFindjk.m`: Suche nach den Indizes j, k.

4. *Ersetze das Pivotelement durch seinen Reziprokwert.*

Will man den Algorithmus mit Überschreiben realisieren[46]*, dann muss man natürlich mit Schritt 3 beginnen.*

Zur Bestimmung der Indizes j, k, für die die Vertauschung durchgeführt werden soll, wird die Matrix $A^{(r)}$ spaltenweise durchgegangen. Sobald eine Spalte gefunden wurde, die ein negatives Element enthält[47], wird der zugehörige Eintrag in $c^{(r)}$ geprüft. Ist dieser ≥ 0, so verschlechtert der Übergang zur Nachbarecke das Ergebnis nicht, ist er < 0, so wird die Spalte verworfen. Findet man keine passende Spalte, dann ist die Lösung optimal.

Bleibt noch der Simplexalgorithmus selbst. Um auch die Parameter $x_1, \ldots, x_n$ der Optimallösung angeben zu können, müssen wir Buch führen, welche Gleichungen miteinander ausgetauscht wurden. Das wird durch zwei Vektoren `xVec` und `yVec` erledigt, die angeben, welche der Parameter als Variablen (`xVec`) und welche als affine Funktionen (`yVec`) fungieren; ein positiver Eintrag j bedeutet hierbei die Variable x_j, ein negativer $-k$ die Variable y_k. Da die Optimallösung ja durch $x^{(r)} = 0$ gegeben ist, erhalten diejenigen x_j, die zu affinen

[46] Beispielsweise an einer Tafel mit Auswischen.

[47] Hierbei gehen wir immer davon aus, daß $A^{(0)}$ "vernünftig" gewählt war, also mindestens einen negativen Wert enthalten hat – damit ist der zulässige Bereich nicht total unbeschränkt.

```
%% Simplex.m (Optimierung)
%% Simplexverfahren
%% Daten:
%%   A  Matrix des Problems
%%   b  Nebenbedingungen
%%   c  Zielfunktion

function [x,opt] = Simplex( A,b,c )
  StepNum = 0; c = [ c; 0 ];
  [m,n] = size( A );
  xVec = ( 1:n ); yVec = ( -1:-1:-m );   %% +/- fuer x/y

  disp( [[A,b];c'] );                    %% Zeige Simplextableau

  do
    StepNum = StepNum + 1;
    [j,k] = auxFindjk( A,b,c );          %% Finde Indizes

    if j != 0
      disp( [StepNum,j,k] );             %% Zeige SchrittNr und Austausch
      t = xVec( j ); xVec( j ) = yVec( k ); yVec( k ) = t;
      [A,b,c] = Austausch( j,k,A,b,c );
      disp( [[A,b];c'] );                %% Zeige Simplextableau
    end
  until ( j == 0 )

  x = auxGenx( n,yVec,b );
  opt = c(n+1);
%endfunction
```

Programm 2.3 `Simplex.m`: Der Simplexalgorithmus für ein Problem, das den Nullpunkt als zulässige Ecke hat.

```
%%
%% auxGenx.m
%% Simplexverfahren
%% Daten:
%%   yVec  Vector der Gleichungsnummern; Eintraege > 0 entsprechen Variablen
%%   b     Rechte Seite

function x = auxGenx( n,yVec,b )
  m = length( yVec );
  x = zeros( n,1 );

  for k = 1:m
    if yVec( k ) > 0
      x( yVec(k) ) = -b( k );
    end
  end

%endfunction
```

Programm 2.4 `auxGenx.m`: Bestimmung des Lösungsvektors x für das Optimierungsproblem aus den Austauschinformationen.

Funktionen geworden sind, den Wert der entsprechenden Komponente von $b^{(r)}$, die Variablen geblieben sind, werden hingegen $= 0$ gesetzt.

Beispiel 2.14 *Sehen wir uns nochmals Beispiel 2.1 an. Die für unsere Simplexmethode relevanten Parameter können wir nun wie folgt in einer Tabelle, dem* ***Simplextableau****, darstellen:*

	x_1	$\dots$	x_n	
y_1	a_{11}	$\dots$	a_{1n}	b_1
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
y_n	a_{n1}	$\dots$	a_{nn}	b_n
	c_1	$\dots$	c_n	c_{n+1}

,

in unserem Fall also

	x_1	x_2	
y_1	-20	-10	-8000
y_2	-4	-5	-2000
y_3	-6	-15	-4500
	16	32	0

.

Die einzelnen Schritte des Simplexalgorithmus liefern dann

$(1,1)$ $\Rightarrow$

	y_1	x_2	
x_1	-0.05	-0.5	-400
y_2	0.2	-3	-400
y_3	0.3	-12	-2100
	-0.8	24	6400

$(2,2)$ $\rightarrow$

	y_1	y_2	
x_1	-0.0833	0.1667	-333.33
x_2	0.0667	-0.333	-133.33
y_3	-0.5	4	-500
	0.8	-8	9600

$(1,3)$ $\rightarrow$

	y_3	y_2	
x_1	.1667	-0.5	-250
x_2	-0.133	0.2	-200
y_1	-2	8	-1000
	-1.6	-1.6	10400

$$\Rightarrow \qquad x = \begin{bmatrix} 250 \\ 200 \end{bmatrix}, \max = 10400.$$

Ein wichtiger Implementierungsparameter ist die Bestimmung der **Pivotspalte** j, von der wir bisher nur gefordert haben, daß $c_j^{(r)} > 0$ sein soll. In der Tat gibt es die verschiedensten Strategien, diese Spalte auszuwählen:

1. Man wählt das kleinste j, so daß $c_j > 0$ ist. Diese Methode ist in Programm 2.2 beschrieben.

2. Man sucht sich eine Spalte j aus, wo c_j *maximal* wird, also

   $$c_j \geq c_\ell, \qquad \ell = 1, \ldots, n.$$

   Dies ist wohl auch das "Originalverfahren" bei Dantzig.

3. Zu jeder Spalte j mit $c_j > 0$ bestimmt man die zugehörige Zeile k, bildet den Wert

   $$d_j = \frac{c_j}{a_{kj}} b_k \geq 0,$$

   um den die Zielfunktion verbessert wird und wählt j dann so, daß

   $$d_j = \max\{d_\ell \,:\, \ell = 1, \ldots, n,\ c_\ell > 0\}.$$

   Diese etwas aufwendigere Strategie nennt man **Totalpivotsuche**. Es zeigt sich, daß diese Pivotstrategie wirklich Vorteile haben kann, insbesondere in Extremfällen wie Beispiel 2.22.

Übung 2.3 Implementieren Sie die hier vorgestellten alternativen Pivotstrategien und testen Sie sie. ◇

2.5 Degenerierung und andere Ärgernisse

Unangenehm wird es aber, wenn die **Pivotzeile** k nicht eindeutig ist, das heißt, wenn es mindestens *zwei* Indizes $k \neq k' \in \{1, \ldots, m-n\}$ gibt, so daß

$$\frac{b_k^{(r)}}{a_{k,j}^{(r)}} = \frac{b_{k'}^{(r)}}{a_{k',j}^{(r)}} = \min\left\{ \frac{b_\ell^{(r)}}{a_{\ell,j}^{(r)}} \,:\, a_{\ell,j}^{(r)} < 0,\ \ell = 1, \ldots, m-n \right\}. \tag{2.27}$$

Führt man nun einen Austauschschritt mit dem Paar (j,k) durch, dann folgt aus (2.25), daß

$$b_{k'}^{(r+1)} = b_{k'}^{(r)} - a_{k'j}^{(r)} \frac{b_k^{(r)}}{a_{kj}^{(r)}} = b_{k'}^{(r)} - a_{k'j}^{(r)} \frac{b_{k'}^{(r)}}{a_{k'j}^{(r)}} = 0.$$

Das heißt aber, daß die zur Ecke $x^{(r+1)} = 0$ gehörige Indexmenge $J^{(r+1)}$ mindestens $n+1$ Einträge hat – die ersten, „virtuellen" n Zeilen, die zur Einheitsmatrix gehören und die Zeile k'. Die Ecke $x^{(r+1)} = 0$ ist also der Durchschnitt von mindestens $n+1$ Hyperebenen[48]. Diese Situation bezeichnet man als **Degenerierung**. Degenerierungen können dazu führen, daß der Simplexalgorithmus stationär wird und auf einer nicht optimalen Seite „im Kreis läuft".

Beispiel 2.15 *Wir betrachten das folgende Beispiel aus (Schwarz, 1988, S. 69):*

$$\widetilde{A} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ -1 & 0 & -1 \\ 0 & -1 & -1 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix}, \qquad \widetilde{b} = \begin{bmatrix} -2 \\ -2 \\ -3 \\ -3 \\ -1 \\ -1 \end{bmatrix}, \qquad c = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}.$$

Die geometrische Interpretation des zugehörigen Polyeders ist ein Quader mit aufgeset-

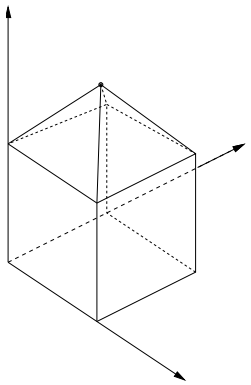

Abbildung 2.6: Das Polyeder $F(A, b)$ zu Beispiel 2.15. Die obere Ecke ist degeneriert, da schneiden sich 4 Seiten des Polyeders.

zter Pyramide, siehe Abb. 2.6. Und in der Tat sorgt die obere Ecke für Schwierigkeiten beim Simplexalgorithmus: Wir erhalten die Folge

$(3,5)$ $\rightarrow$

	x_1	x_2	y_5	
y_1	-1	0	0	-2
y_2	0	-1	0	-2
y_3	-2	0	1	-2
y_4	-1	-1	1	-2
x_3	1	0	-1	-1
y_6	-1	1	1	$\mathbf{0}$
	4	2	-3	3

$(1,6)$ $\rightarrow$

	y_6	x_2	y_5	
y_1	1	-1	-1	-2
y_2	0	-1	0	-2
y_3	2	-2	-1	-2
y_4	1	-2	0	-2
x_3	-1	1	0	-1
x_1	-1	1	1	$\mathbf{0}$
	-4	6	1	3

[48] Der "normale", also generische Fall im $\mathbb{R}^n$ ist, daß sich gerade n Hyperebenen, nicht mehr und nicht weniger, in einem Punkt schneiden.

$(2,3)$ $\rightarrow$

	y_6	y_3	y_5	
y_1	0	$\frac{1}{2}$	$-\frac{1}{2}$	-1
y_2	-1	$\frac{1}{2}$	$\frac{1}{2}$	-1
x_2	1	$-\frac{1}{2}$	$-\frac{1}{2}$	-1
y_4	-1	1	1	**0**
x_3	0	$-\frac{1}{2}$	$-\frac{1}{2}$	-2
x_1	0	$-\frac{1}{2}$	$\frac{1}{2}$	-1
	2	-3	-2	9

$(1,4)$ $\rightarrow$

	y_4	y_3	y_5	
y_1	0	$\frac{1}{2}$	$-\frac{1}{2}$	-1
y_2	1	$-\frac{1}{2}$	$-\frac{1}{2}$	-1
x_2	-1	$\frac{1}{2}$	$\frac{1}{2}$	-1
y_6	-1	1	1	**0**
x_3	0	$-\frac{1}{2}$	$-\frac{1}{2}$	-2
x_1	0	$-\frac{1}{2}$	$\frac{1}{2}$	-1
	-2	-1	**0**	9

und müssen uns nun überlegen, was die "fetten" Nullen bedeuten:

1. *Die Nullen in der* Spalte *auf der rechten Seite zeigen uns an, daß wir uns in einer entarteten, degenerierten, Ecke befinden, was dazu führen kann, daß der Simplexalgrorithmus diese Ecke nicht verläßt. In der Tat nimmt er den Weg*

$$\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \overset{(3,5)}{\rightarrow} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \overset{(1,6)}{\rightarrow} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \overset{(2,3)}{\rightarrow} \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} \overset{(1,4)}{\rightarrow} \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix},$$

   bleibt also mehrmals *in derselben Ecke!*

2. *Die Null in der Zeile unter dem Simplextableau zeigt hingegen an, daß die Optimallösung* nicht *eindeutig ist – es gibt eine Nachbarecke, an der die Zielfunktion denselben Wert annimmt.*

Ein solches Verhalten ist der einfachste Falle eines **Zyklus**, der sich natürlich auch über mehrere Punkte erstrecken kann und der dazu führen kann, daß der Simplexalgorithmus *nie* terminiert; solche Beispiele gibt es, wenngleich sie ziemlich konstruiert sind, aber laut (Nocedal & Wright, 1999) tauchen sie doch recht häufig in der Realität auf, und zwar bei Problemen, die aus der Ganzzahlprogrammierung stammen.

Die Entfernung dieser Zyklen durch geeignete Störungen des Optimierungsproblems ist ein wichtiges Detail bei der praktischen Realisierung des Simplexalgorithmus, siehe (Nocedal & Wright, 1999, S. 389–391).

Beispiel 2.16 *Verwendet man die Totalpivotsuche, dann taucht keiner der Zyklen aus Beispiel 2.15 auf, denn hier hat man die Situation, daß man von jeder Ecke aus die Zielfunktion verbessern kann. Die Folge der Simplextableaus ist dann*

$(2,2)$ $\rightarrow$

	x_1	y_2	x_3	
y_1	-1	0	0	-2
x_2	0	-1	0	-2
y_3	-1	0	-1	-3
y_4	0	1	-1	-1
y_5	1	0	-1	-1
y_6	0	-1	-1	-3
	1	-2	3	4

$(3,4)$ $\rightarrow$

	x_1	y_2	y_4	
y_1	-1	0	0	-2
x_2	0	-1	0	-2
y_3	-1	-1	1	-2
x_3	0	1	-1	-1
y_5	1	-1	1	**0**
y_6	0	-2	1	-2
	1	1	-3	7

		y_1	y_2	y_4	
	x_1	-1	0	0	-2
	x_2	0	-1	0	-2
$(1,1)$	y_3	1	-1	1	$\mathbf{0}$
$\rightarrow$	x_3	0	1	-1	-1
	y_5	-1	-1	1	-2
	y_6	0	-2	1	-2
		-1	1	-3	9

Den Übergang von der Ecke $[2,2,1]^T$ *nach* $[1,1,2]$ *vermeidet diese Version des Simplexalgorithmus, weil* keine *Verbesserung mehr möglich ist, und so terminiert der Simplexalgorithmus, obwohl es eine Spalte gibt, in der* $c_j > 0$ *ist!*

Übung 2.4 Zeigen Sie: Ist eine $n-1$–dimensionale Seitenfläche X von F eine **Niveaufläche** von $c^T\cdot$, d.h. $c^Tx = c^Tx'$, $x, x' \in X$, dann nimmt die Zielfunktion auf X ihr Maximum oder ihr Minimum an. ◊

Ein anderes "Ärgernis" sind **freie Variablen**. Bisher haben wir immer angenommen, daß die automatischen Nebenbedingungen $x \geq 0$ gelten. Solange alle Variablen einseitig beschränkt sind, das heißt, Forderungen der Form

$$x_j \geq \xi_j \qquad \text{oder} \qquad x_j \leq \xi_j \qquad j = 1, \ldots, n,$$

vorliegen, können wir diese immer durch $x \geq 0$ ausdrücken, indem wir x_j durch $\pm(x_j - \xi_j)$ mit passendem Vorzeichen[49] ersetzen. Das resultierende, *äquivalente* Optimierungsproblem hat dann die gewünschte Form (2.1). Hingegen heißt x_j **freie Variable**, wenn es keine solche *direkte* Beschränkung an x_j gibt, sondern sich der zulässige x_j–Bereich nur *indirekt* aus den Nebenbedingungen $Ax \geq b$ ergibt. Solche Variablen stören natürlich, wenn man ein Verfahren verwenden will, bei dem der Nullpunkt eine zulässige Ecke sein soll[50], weswegen man zuerst einmal alle freien Variable austauscht. Die Zeile, die *nach* dem Austausch mit y_k zu der freien Variablen x_j gehört,

$$x_j = \left(A^{(1)}x^{(1)}\right)_{n+k} - b^{(1)}_{n+k}$$

kann man zwar mitführen (z.B. um am Schluß den Wert von x_j zu ermitteln), aber sie muß **redundant**[51] sein, denn ansonsten läge ja plötzlich doch eine Beschränkung an x_j vor – für die *Rechnung* kann man sie aber getrost vergessen.

Damit wollen wir es, was freie Variablen angeht, gut sein lassen, denn die Theorie ist etwas haarig. Allerdings gibt es eine Vielzahl von realistischen Optimierungsproblemen, bei denen freie Variablen auftreten. Hier kann man nur raten, einen guten, fertigen Solver wie `lpsolve` (Berkelaar *et al.*, 2000) oder `glpk` (Makhorin, 2000) zu verweden, es sind so viele Sonderfällle zu berücksichtigen, daß die Selbstprogrammierung keinen Spaß macht. Die *grobe* Theorie ist in (Schwarz, 1997) beschrieben.

[49] Um, wenn nötig, das Ungleichungszeichen umzudrehen.

[50] Denn dazu bräuchten wir ja sowas wie $x_j = 0$.

[51] Das heißt, es kann keine zulässige Ecke x geben, zu deren Bestimmungsmenge $J(x)$ die Gleichung $x_j = 0$ gehören *muß*.

2.6 Auffinden einer Startecke

Bisher haben wir uns in einer Hinsicht das Leben sehr leicht gemacht: Wir haben angenommen, daß der Nullpunkt eine zulässige Ecke war. Nun, eine *Ecke* ist der Nullpunkt immer, wenn das Optimierungsproblem in der Form (2.10) vorliegt[52], wenn also

$$A = \begin{bmatrix} I_n \\ \widetilde{A} \end{bmatrix} \qquad \text{und} \qquad b = \begin{bmatrix} 0 \\ \widetilde{b} \end{bmatrix} \tag{2.28}$$

ist, nur mit der *Zulässigkeit* kann es unter Umständen hapern.

Beispiel 2.17 (Transportproblem) [53]
In den Rangierbahnhöfen A und B stehen 18 bzw. 12 leere Waggons, in den Bahnhöfen X, Y und Z werden 11, 10 und 9 Waggons benötigt. Die Distanzen zwischen den Bahnhöfen betragen

	X	Y	Z
A	5	4	9
B	7	8	10

Welche Verteilung der Waggons minimiert die gefahrene Kilometerzahl[54]*?*
Um dieses Problem mathematisch darzustellen, sei x *die Anzahl der Wagen, die von A nach X fahren und* y *die Anzahl der Wagen, die von A nach* y *fahren. Dann lassen sich alle Wagenbewegungen durch* x *und* y *ausdrücken und zwar*

Strecke	*# Wagen*
$A \to X$	x
$A \to Y$	y
$A \to Z$	$18 - x - y$
$B \to X$	$11 - x$
$B \to Y$	$10 - y$
$B \to Z$	$x + y - 9$

und alle diese Größen müssen selbstverständlich positiv sein. Damit müssen wir den Wert

$$5x + 4y + 9\,(18 - x - y) + 7\,(11 - x) + 8\,(10 - y) + 10\,(x + y - 9) = -x - 3y + 229$$

unter den obigen Nebenbedingungen minimieren, unsere Normalform für das Optimierungsproblem lautet also

$$\max\ x + 3y - 229, \qquad \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \\ -1 & 0 \\ 0 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \geq \begin{bmatrix} 0 \\ 0 \\ -18 \\ -11 \\ -10 \\ 9 \end{bmatrix}$$

[52] Und in diese Form können wir es ja immer durch die Eliminierung eventueller freier Variablen bringen.

[53] Aus (Schwarz, 1988, Beispiel 2.2, S. 57), ein Spezialfall von Beispiel 1.2

[54] Auch hier handelt es sich eigentlich wieder um ein Problem aus der *Ganzzahloptimierung*, aber wieder einmal wird, rein zufällig, die kontinuierliche Optimallösung ganzzahlig sein.

bestimmen; die letzte Zeile der Nebenbedingungen sorgt nun dafür, daß $[x, y] = 0$ *zwar eine Ecke, aber keine* zulässige *Ecke ist – wenn man keine Wagen bewegt, dann kommt halt auch nichts in X, Y oder Z an. Die Nebenbedingungen sind in Abb. 2.7 grafisch dargestellt.*

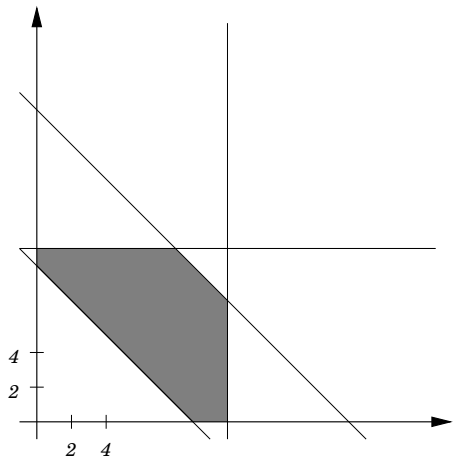

Abbildung 2.7: Der zulässige Bereich für das Transportproblem aus Beispiel 2.17; der Nullpunkt ist offensichtlich "abgeschnitten" worden.

Wenn nun

$$b^* := \max_{j=1,\dots,n} b_j \tag{2.29}$$

positiv ist, dann[55] ist der der Nullpunkt *keine* zulässige Ecke des zulässigen Bereichs mehr und unser bisheriger Simplexalgorithmus ist nicht anwendbar. In diesem Fall behilft man sich mit der sogenannten **Zweiphasenmethode**, bei der man zuerst einmal, und zwar wieder mit dem Simplexalgorithmus, ein Optimierungsproblem löst, um eine Startecke zu finden. Das sehen wir uns nun an.

Dazu setzen wir $1_m = [1, \dots, 1]^T \in \mathbb{R}^m$ und betrachten das erweiterte **Hilfsproblem**

$$\max -(x_0 + b^*), \qquad \underbrace{\begin{bmatrix} 0_n & I_n \\ 1_{m-n} & \widetilde{A} \end{bmatrix}}_{=:\widehat{A}} \underbrace{\begin{bmatrix} x_0 \\ x \end{bmatrix}}_{=:\widehat{x}} \geq \underbrace{b - b^* \begin{bmatrix} 0_n \\ 1_{m-n} \end{bmatrix}}_{=:\widehat{b}}. \tag{2.30}$$

Die Zielfunktion $x_0 + b^*$, beschreibt die „Unzulässigkeit" des Optimierungsproblems: Wäre $b^* < 0$, dann wäre diese Funktion an der Stelle $x_0 = 0$ negativ und alles wäre in Ordnung; ist hingegen $b^* > 0$ – was ja der Fall ist, den wir hier untersuchen wollen – dann hat besteht an $x_0 = 0$ eine echte Unzulässigkeit durch diesen positiven Wert. Da wir diesen Defekt verkleinern, besser noch:

[55] Und nur dann!

minimieren, wollen, müssen wir in (2.30) also die *negative* Unzulässigkeit maximieren, der Normalform wegen. Unser Ziel ist es also, die Unzulässigkeit des Optimierungsproblems zu minimieren und es dadurch zulässig zu bekommen.

Da in (2.30) nun $\widehat{b} \leq 0$ ist, ist der Punkt $\widehat{x} = 0$ eine zulässiger Punkt, allerdings ist x_0 hier eine freie Variable – aber die können (und müssen) wir ja, wie schon gesehen, austauschen.

Nach endlich vielen Schritten mit unserem erweiterten Problem finden wir dann eine Optimallösung $\widehat{x}^*$ von (2.30) mit zugehörigen Werten x_0^* und x^*, die wir aus den entsprechenden Zeilen und Spalten des Simplextableaus bestimmen können. Setzen wir x_0^* und x^* in (2.30) ein, so erhalten wir, daß

$$Ax^* + \begin{bmatrix} 0_n \\ 1_{m-n} \end{bmatrix} x_0^* = \widehat{A} \begin{bmatrix} x_0^* \\ x^* \end{bmatrix} \geq \widehat{b} = b - b^* \begin{bmatrix} 0_n \\ 1_{m-n} \end{bmatrix},$$

also

$$Ax^* \geq b - (x_0^* + b^*) \begin{bmatrix} 0_n \\ 1_{m-n} \end{bmatrix}. \tag{2.31}$$

Die rechte Seite von (2.31) ist nun genau dann $\geq b$, wenn $x_0 + b^* \leq 0$ ist, also wenn $x_0^* \leq -b^*$ ist. In diesem Fall haben wir einen zulässigen Punkt gefunden; erfüllt umgekehrt x die Bedingung $Ax \geq b$, dann erfüllt jeder Punkt der Form $[x_0, x]$ mit $x_0 \leq -b^*$ aber auch (2.30). Das können wir folgendermaßen zusammenfassen.

Lemma 2.18 *Es gibt genau dann einen zulässigen Punkt x mit $Ax \geq b$, wenn es einen Punkt $\widehat{x} = [x_0, x]$ gibt, so daß $\widehat{A}\widehat{x} \geq \widehat{b}$ und $x_0 \leq -b^*$.*

Ist also bei unserer Optimallßung $x_0^* > -b^*$, dann kann es auch keinen zulässigen Punkt des Ausgangsproblems geben und dieses Optimierungsproblem wäre unsinnig, genauer, der dazu gehörende zulässige Bereich wäre leer. Das kann durchaus passieren, beispielsweise bei Transportproblemen im Sinne von Beispiel 2.17, bei denen mehr an den Zielen ankommen soll als in den Ausgangspunkten bereitsteht.

Ein abschliessender Blick auf (2.30) zeigt, daß unsere Bedingung an x_0^*, also $-x_0 \geq b^*$, nichts anderes bedeutet, als daß die Zielfunktion $-x_0 - b^*$ aus (2.30) *nichtnegativ* werden soll. Wir müssen also nicht mal unbedingt bis zum Optimum suchen, es genügt, denn Simplexalgorithmus so lange auf das Hilfsproblem anzuwenden, bis der Wert der (Hilfs–)Zielfunktion nichtnegativ ist. Wir fassen zusammen.

Lemma 2.19 *Ist $x_0^* \leq -b^*$, dann ist der Punkt x^* ist eine zulässige Ecke des Polyeders $F\left(A, \widehat{b}\right)$, wobei*

$$\widehat{b} = b - (x_0^* + b^*) \begin{bmatrix} 0_n \\ 1_{m-n} \end{bmatrix}.$$

Beweis: Daß x^* zulässig ist, wenn $x_0^* \leq -b^*$ ist, das wissen wir ja schon. Da außerdem $\widehat{x}^*$ eine Ecke von F^* ist[56] gibt es eine Menge J, $\#J = n + 1$, so daß

$$[1_m\, A]_J \begin{bmatrix} x_0^* \\ x^* \end{bmatrix} = \widehat{b}_J, \qquad \Rightarrow \qquad A_J\, x^* = b_J^*$$

[56] F^* ist, wie man sich leicht vorstellen kann, der zulässige Bereich des "Hilfsproblems".

und da $A_J \in \mathbb{R}^{n+1\times n}$ Rang n hat gibt es eine Teilmenge $J' \subset J$, $\#J' = n$, so daß

$$\det A_{J'} \neq 0 \qquad \text{und} \qquad A_{J'}\, x^* = b^*_{J'},$$

also ist x^* eine Ecke. □

Korollar 2.20 *Eine zulässige Ecke $\widehat{x} = [x_0, x^T]^T$ von (2.30) enthält genau dann eine Ecke x unseres Ausgangsproblems, wenn $x_0 \leq -b^*$.*

Und damit haben wir unsere Startecke aufgespürt: Verpassen wir nun der *freien* Variablen[57] den Wert $x_0 = -\max b_j$, dann ist x^* eine Ecke von (2.31), insbesondere ist $Ax^* \geq b$, also ist x^* eine zulässige Ecke und damit die Startecke, von der aus wir loslegen können.

Zur praktischen Realisierung betrachten wir wieder die abgeschnittene Form

$$y = \left[1_m\ \widetilde{A}\right] \begin{bmatrix} x_0 \\ x \end{bmatrix} - \left(\widetilde{b} - b^*\, 1_m\right) =: A^{(-1)} \begin{bmatrix} x_0 \\ x \end{bmatrix} - b^{(-1)},$$

mit der Hilfs–Zielfunktion[58]

$$f(x_0, x) = -x_0 - b^*,$$

und der Original–Zielfunktion

$$c(x_0, x) = c^T x$$

tauschen die freie Variable x_0 aus und erhalten, nach eventueller Zeilenvertauschung

$$\begin{bmatrix} x_0 \\ y^{(0)} \end{bmatrix} = A^{(0)} x^{(0)} - b^{(0)}, \qquad y^{(0)} \in \mathbb{R}^{m-n-1},\ A^{(0)} \in \mathbb{R}^{m-n,n+1},\ b^{(0)} \in \mathbb{R}^{m-n}, \quad (2.32)$$

sowie einen Vektor $c^{(0)}$, so daß $c(x) = \left(c^{(0)}\right)^T x^{(0)}$. Beim nun folgende Simplexalgorithmus, bei dem zwar bezüglich f minimiert, aber $c^{(r)}$ stets mitbestimmt werden muß, darf die Zeile, die zu x_0 gehört *nicht* mehr ausgetauscht werden[59], und so erhalten wir nach endlich vielen Schritten das Tableau

$$\begin{bmatrix} x_0 \\ y^{(r)} \end{bmatrix} = A^{(r)} x^{(r)} - b^{(r)},$$

aus dem wir den Wert von x_0 ablesen können; außerdem sagt uns die Zielfunktion, ob unsere Kante so gefundene Kante überhaupt zulässig ist, denn das ist

[57] Und schon tauchen die Dinger wieder auf ...

[58] Weil man es gar nicht oft genug sagen kann: Für die eigentliche Optimierung des Hilfsproblems bräuchten wir den konstanten Term $-b^*$ nicht, aber da die Positivität dieser Zielfunktion bereits ein Abbruchkriterium ist, lohnt es sich.

[59] Schließlich könnte man die Zeilen, die zu freien Variablen gehören, sogar aus dem Simplextableau entfernen.

genau der Fall, wenn $f\left(x_0, x^{(r)}\right) \geq 0$ ist. Nehmen wir an, das wäre der Fall, dann müssen wir nur noch x_0 loswerden, was wir dadurch erreichen, daß wir es mit der zuletzt "eingetauschten" Spalte vertauschen, also einen Austauschschritt mit dem Paar $(1, j_r)$ durchführen, was uns, nach einer **Spaltenvertauschung** die Darstellung

$$\begin{aligned}\left[y^{(r+1)}\right] &= A^{(r+1)}\begin{bmatrix} x_0 \\ x^{(r+1)} \end{bmatrix} - b^{(r+1)} = [a\ B]\begin{bmatrix} x_0 \\ x^{(r+1)} \end{bmatrix} - b^{(r+1)} \\ &= Bx^{(r+1)} + \left(x_0 a - b^{(r+1)}\right), \qquad B \in \mathbb{R}^{m\times n},\ a \in \mathbb{R}^m,\end{aligned}$$

liefert. Setzen wir da nun $x_0 = -\max b_j$ ein, dann haben wir die modifizierte Form unseres "Originalproblems", bei der $x^{(r+1)} = 0$ eine zulässige Ecke ist und können endlich den "normalen" Simplexalgorithmus anwerfen.

Um uns das "Mitführen" der "echten" Zielfunktion c, das heißt die Bestimmung von $c^{(r)}$ etwas leichter zu machen, bemerken wir, daß wir die Austauschregel (2.24) mit der Matrix B aus (2.26) auch als

$$c^{(r+1)} = B\, c^{(r)}$$

schreiben können, ein Update von $c^{(r)}_{n+1}$ ist nicht unbedingt nötig, weil wir die Zielfunktion ja so nur um eine Konstante abändern. Das heißt aber, daß wir die Zielfunktion c lediglich als "tote" Zeile unseres Hilfsproblems mitzuführen brauchen, also als Zeile, die nicht zur Auswahl der Pivotzeilen zugelassen ist.

Beispiel 2.21 *(Transportproblem aus Beispiel 2.17)*
Zuerst transformiert der erste Austauschschritt das erweiterte Tableau folgendermaßen:

	x_0	x_1	x_2	
y_1	1	−1	−1	−27
x_2	1	−1	0	−20
y_3	1	0	−1	−19
y_4	1	1	−1	0
c	0	1	3	−229
	−1	0	0	−9

$(1,4)$
$\rightarrow$

	y_4	x_1	x_2	
y_1	1	−2	−2	−27
y_2	1	−2	−1	−20
y_3	1	−1	−2	−19
x_0	1	−1	−1	0
c	0	1	3	−229
	−1	1	1	−9

Die beiden unteren Zeilen sind nun tabu und das Simplexverfahren liefert jetzt das Tableau

$(2,2)$
$\rightarrow$

	y_4	y_2	x_2	
y_1	0	1	−1	−7
x_1	0.5	−0.5	−0.5	−10
y_3	0.5	0.5	−1.5	−9
x_0	0.5	0.5	−0.5	10
c	0.5	−0.5	2.5	−229
	−0.5	−0.5	0.5	−9

Die Ecke die wir so gefunden haben ist also der Punkt $[10,0]^T$*. Nach Vertauschen von* y_2 *und* x_0 *und den oben beschriebenen Umformungen läuft dann der "normale" Simplexalgorithmus folgendermaßen ab:*

	y_4	x_2	
y_1	-1	0	-9
x_1	1	-1	-9
y_3	0	-1	-10
y_2	-1	1	-2
	1	2	-220

$(2,2)$ $\rightarrow$ * $(1,3)$ $\rightarrow$ * $(2,1)$ $\rightarrow$

	y_3	y_1	
x_1	1	-1	-8
x_2	-1	0	-10
y_4	0	-1	-9
y_2	-1	1	-3
	-2	-1	-191

und die Optimallösung ist also $[x,y] = [8,10]$*.*

2.7 Kleine Komplexitätsbetrachtungen

Wie viele Schritte braucht nun so ein Simplexalgorithmus, um sein Ziel zu erreichen? Laut (Spellucci, 1993) reichen "im Normalfall" so etwa $2m$ bis $3m$ Austauschschritte aus, um das Optimum zu bestimmen. Allerdings sehen die "worst case"–Szenarien etwas übler aus. Hier ein Beispiel aus (Spellucci, 1993), das von Minty und Klee (Minty & Klee, 1972) stammt.

Beispiel 2.22 *Das Optimierungsproblem*

$$\begin{aligned} \sum_{k=1}^{n} 10^{n-k} x_k &= \max, \\ 2\sum_{k=1}^{j-1} 10^{j-k} x_k + x_k &\le 100^{j-1}, \qquad j = 1,\dots,n \\ x &\ge 0, \end{aligned}$$

das in unserer Notation durch die Normalform

$$A = \begin{bmatrix} 1 & & & & & \\ & 1 & & & & \\ & & 1 & & & \\ & & & \ddots & & \\ & & & & 1 & \\ -1 & & & & & \\ -20 & -1 & & & & \\ -200 & -20 & -1 & & & \\ \vdots & \vdots & & \ddots & & \\ -2\times 10^{n-1} & -2\times 10^{n-1} & \dots & -20 & -1 \end{bmatrix} \quad x \ge \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ -1 \\ -100 \\ -10000 \\ \vdots \\ -100^{n-1} \end{bmatrix}$$

dargestellt wird, benötigt $2^n - 1$ *Austauschschritte, wenn man mit der Startecke* $x^{(0)} = 0$ *beginnt.*

Man kann sich dieses Beispiel sehr schön experimentell ansehen[60] und erkennt dabei auch, welche Bedeutung eine gute Pivotstrategie haben. Mit der Methode der "koordinatenweisen" Pivotisierung benötigt das Verfahren immer die vollen $2^n - 1$ Austauschschritte, während ein Verfahren, das auf der „raffinierteren" Pivotstrategie der **Totalpivotsuche** basiert, nach gerade mal *einem* Schritt am Ziel ist.

Wenn nun also die Totalpivotsuche so eine tolle Pivotregel ergibt, warum verwendet man sie dann nicht auch? Ganz einfach: im „normalen" Simplexalgorithmus muß man pro Rechenschritt $O(n)$ Vergleichsoperationen durchführen, um das maximale c_j zu finden und dann $O(m - n)$ Rechenoperationen, um die Pivotzeile zu bestimmen; bei der Totalpivotsuche müßte man im schlimmsten Fall $O(m - n)$ Rechenoperationen für *jede* der n Spalten durchführen, also $O(n(m - n))$ Rechen- und ebensoviele Vergleichsoperationen. Das heißt, daß der Aufwand pro Schritt des Simplexalgorithmus von $O(m)$ auf $O(n(m - n))$, für große Werte von m und n nicht gerade erstrebenswert.

[60] Genau dafür sind `Matlab` und `octave` ja auch da.

To isolate mathematics from the practical demands of the sciences is to invite the sterility of a cow shut away from the bulls.

P. Chebyshev

Lineare Optimierung – Beispiele und Anwendungen

3

Bevor wir uns wieder in das Vergnügen der Theorie stürzen, wollen wir uns zuerst einmal ein paar Beispiele ansehen, in denen wir den Simplexalgorithmus *anwenden* können. Interessant wird hierbei vor allem werden, wie wir die Probleme in einer Form aufbereiten können, daß wir sie in unsere `Octave`–Programme stecken und von diesen lösen lassen können. Hochgestochen gesprochen befassen wir uns also nun mit der **Modellierung** von Optimierungsproblemen. Einige der Beispiele stammen aus (Gass, 1970)[61].

Allerdings, und sei es nur, um Langeweile zu vermeiden, verwenden wir jetzt eine andere Normalform für unsere linearen Optimierungsprobleme, und zwar die „algorithmische Normalform“

$$\min_x c^T x, \qquad Ax \leq b, \qquad x \geq 0. \tag{3.1}$$

Der Grund dafür ist relativ profan: Wir verwenden einfach etwas andere `Octave`–Funktionen[62] für den Simplexalgorithmus, die jetzt nur noch richtige Ausgaben liefern. Aber natürlich lassen sich ja alle Normalformen durch etwas Vorzeichenspielerei ineinander überführen.

3.1 Das Diät–Problem

Beginnen wir erst einmal mit einem ganz typischen, einfachen Problem, ganz änlich[63] zur „Schuhfabrik“.

Beispiel 3.1 (Frühstücksplanung) *Eine Hausfrau versucht für Ihre Familie ein optimales Frühstück zusammenzustellen. Dafür stehen ihr*[64] *zwei verschiedene Typen von Getreideflocken*[65]*, nämlich* Crunchies *und* Krispies *zur Verfügung, die zwei*

[61] Eine sehr empfehlenswerte, anschauliche und auch noch, wie es sich für Dover–Reprints gehört, *preiswerte* Einführung in die lineare Optimierung.

[62] Die wir jetzt aber nicht mehr extra abdrucken wollen.

[63] Und auch ähnlich realistisch

[64] Ja, das Beispiel stammt aus den USA.

[65] Auf gut neudeutsch auch als „Cerealien“ bezeichnet – eine der Wortschöpfungen, die entstehen an Cerebralien mangelt.

Spurenelemente, Thiamin[66] *und* Niacin[67] *in unterschiedlicher Anzahl enthalten, unterschiedlichen Brennwert in Kalorien liefern und natürlich unterschiedlich teuer sind. Das ideale Frühstück versorgt die Familie mit einem gewissen Mindestmaß an „Vitaminen“ und Kalorien und ist dabei natürlich möglichst billig. Die genauen Werte sind in der folgenden Tabelle aufgelistet:*

	Crunchies	*Krispies*	*Benötigt*
Thiamin (in mg)	*0.10*	*0.25*	*1*
Niacin (in mg)	*1.00*	*0.25*	*5*
Kalorien	*110*	*120*	*400*
Preis	*3.8*	*4.2*	

Das Problem ist klar: Was ist die optimate Diät, die diese Randbedingungen erfüllt?

Nun, dieses Problem ist, was die Modellierung angeht, noch richtig einfach, denn wir müssen nur die Bedingungen in Ungleichungsform hinschreiben. Seien dazu x_1 die Menge der verwendeten Crunchies und x_2 die Menge an Krispies, dann erhalten wir die Ungleichungen

$$\begin{aligned} 0.1\,x_1 + .25\,x_2 &\geq 1 \\ x_1 + .25\,x_2 &\geq 5 \\ 110\,x_1 + 120\,x_2 &\geq 400 \end{aligned}$$

und zu minimieren sind die Kosten $3.8\,x_1 + 4.2\,x_2$. Nachdem unsere Normalform aus (3.1) die Ungleichungen als "$\leq$" geschrieben haben will, erhalten wir also das Optimierungsproblem[68]

$$\min 3.8\,x_1 + 4.2\,x_2, \qquad \begin{bmatrix} -0.1 & -0.25 \\ -1 & -0.25 \\ -110 & -120 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \leq \begin{bmatrix} -1 \\ -5 \\ -400 \end{bmatrix},$$

was ein klarer Fall für die Zweiphasenmethode ist. Also ist alles klar, oder? Geben wir das Problem in den Rechner ein, dann erhalten wir mit der Eingabe

```
octave> A = [ -.1 -.25; -1 -.25; -110 -120]; b = [-1 -5 -400]';
octave> c = [ 3.8 4.2]';
octave> [x,opt] = SimSimplex( A,b,c )
```

die etwas überraschende Ausgabe

```
**** Unbeschraenkt ****
x =

  5.00000
  0.00000

opt = 19.000
```

[66] Synonym für Vitamin B_1, siehe (Pschyrembel, 1994).

[67] Synonym für Nicotinsäure, die Nebenwirkungen in (Pschyrembel, 1994) liest man besser nicht.

[68] Von jetzt an schreiben wir die allgegenwärtige Randbedingung $x \geq 0$ **nicht** mehr explizit hin.

die noch nicht einmal zulässig ist, denn die erste Nebenbedingung ist nicht erfüllt. Allerdings sehen wir ja auch an der Ausgabe, wo die Schwierigkeiten herkommen: Das Optimierungsproblem ist **unbeschränkt**, und da funktioniert unser Simplexalgorithmus halt nun einmal nicht[69]. Das sieht man ja auch an der Problemstellung selbst, denn die einfachste Möglichkeit, die Nebenbedingungen zu erfüllen besteht einfach darin, eine Packung von jeder Sorte in sich hineinzustopfen, und wenn's nicht reicht, dann halt noch eine und so weiter. Damit wir unsere Methode anwenden können, müssen wir also das Problem künstlich beschränken. Eine Möglichkeit besteht darin, x_1 und x_2 *individuell* zu beschränken, indem man nachsieht, aus welcher Menge das "kleinste" Frühstück aus Crunchies bestehen muß (nämlich $x_1 = 10$) und wieviele Krispies man mindestens essen muß, um alle "Nährstoffe" aufzunehmen (das ist $x_2 = 20$). Dann können wir die Nebenbedingungen $x_1 \leq 10$ und $x_2 \leq 20$ hinzufügen und erhalten

```
octave> AA = [ A; [ 1 0; 0 1] ]; bb = [ b; 10; 20 ];
octave> [x,opt] = SimSimplex( AA,bb,c )
x =

  4.4444
  2.2222

opt = 26.222
```

das optimale Frühstück kostet also 26.2222 Cent[70]. Eine andere Möglichkeit bestünde darin, den Preis zu beschränken, also beispielsweise nur Früstücke für weniger als einen Dollar:

```
octave> AA = [ A; [ 3.8 4.2 ]  ]; bb = [ b; 100 ];
octave> [x,opt] = SimSimplex( AA,bb,c )
x =

  4.4444
  2.2222

opt = 26.222
```

Und siehe da: Das Ergebnis ist wieder richtig. Wird man hingegen zu knauserig, dann kommt man erneut in Schwierigkeiten:

```
octave> AA = [ A; [ 3.8 4.2 ]  ]; bb = [ b; 15 ];
octave> [x,opt] = SimSimplex( AA,bb,c )
x =
```

[69] Nur um das nochmal klarzustellen: Das Minimum existiert natürlich mit und ohne Beschränkung, nur das **Verfahren** funktioniert nicht!

[70] Und enthält im übrigen rund 756 Kalorien, also fast doppelt so viel wie gewünscht. Vielleicht sollte man dochmal ein grundlegend anderes Frühstück in Betracht ziehen ...

```
  5
  0

opt = 19
```

aber an der Tatsache, daß die zusätzliche Nebenbedingung durch die berechnete Optimallösung verletzt wurde, sieht man schon, daß irgendwas faul sein muß.

3.2 Transportprobleme

Transportprobleme sind immer vom Typ wie in Beispiel 2.17: Ressourcen müssen von Ausgangspunkten zu Zielpunkten transportiert werden, wobei *alle* Ausgangspunkte mit *allen* Zielpunkten als verbunden angenommen werden. Die Entfernungen (oder Kosten) von eine Ausgangs- zu einem Zielpunkt sowie die in den Ausgangspunkten vorrätigen und die in den Zielpunkten benötigten Ressourcen sind typischerweise in einer Matrix aufgelistet:

	Z_1	$\dots$	Z_n	
A_1	a_{11}	$\dots$	a_{1n}	a_1
$\vdots$	$\vdots$	$\ddots$	$\vdots$	
A_m	a_{m1}	$\dots$	a_{mn}	a_m
	z_1	$\dots$	z_n	

Dabei bezeichnet $a_{11}, \dots, a_{mn}$ die **Kostenmatrix**, $a_1, \dots, a_m$ die vorhandenen und $z_1, \dots, z_n$ die benötigten Ressourcen. Damit das Problem überhaupt lösbar ist, muß natürlich

$$a_1 + \cdots + a_m \geq z_1 + \cdots z_n$$

sein.

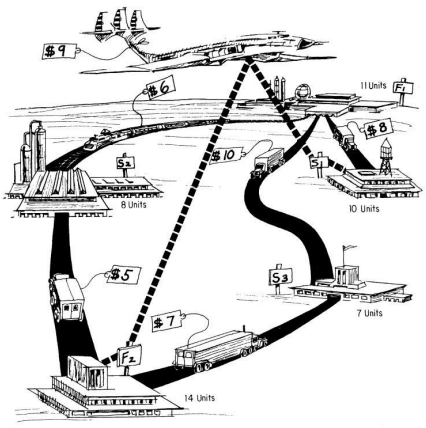

Abbildung 3.1: Kühlschränke und deren Transportwege. Aus (Gass, 1970).

Beispiel 3.2 (Kühlschränke) *Eine Firma stellt in zwei Fabriken,* F_1 *und* F_2*, Kühlschränke her, die in den Läden*[71] S_1, S_3, S_3 *verkauft werden sollten. Die Kosten-/Ressourcen–Matrix ist wie folgt:*

	S_1	S_2	S_3	
F_1	8	6	10	11
F_2	9	5	7	14
	10	8	7	

Wie ist der optimale Transport?

Transportprobleme zeichnen sich dadurch aus, daß man sehr viele Variablen hat, die man am zweckmäßigsten *doppelt* indiziert, nämlich als x_{jk}, wobei x_{jk} die Menge bezeichnet, die vom Ausgangspunkt j zum Zielpunkt k transportiert wird. Die Gesamtkosten sind dann immer

$$\sum_{j=1}^{m} \sum_{k=1}^{n} a_{jk} x_{jk}.$$

In unserem Beispiel haben wir also die Variablen

$$x_{11}, x_{12}, x_{13}, x_{21}, x_{22}, x_{23} \qquad \Rightarrow \qquad x = [x_{11}, x_{12}, x_{13}, x_{21}, x_{22}, x_{23}]^T$$

auch in einen Vektor angeordnet, indem wir die sogenannte **lexikographische Ordnung**[72] verwenden. Unser Beispiel liefert nun die Nebenbedingungen

$$\begin{array}{rrrrrrcl}
x_{11} & +x_{12} & +x_{13} & & & & \leq & 11 \\
 & & & x_{21} & +x_{22} & +x_{23} & \leq & 14 \\
x_{11} & & & +x_{21} & & & \geq & 10 \\
 & x_{12} & & & +x_{22} & & \geq & 8 \\
 & & x_{13} & & & +x_{23} & \geq & 7
\end{array}$$

Die ersten beiden Ungleichungen sind die Beschränkungen an die Ressourcen, die anderen drei betreffen das Minimum, das an den Zielpunkten ankommen soll. Damit können wir uns auch schon wieder ans Modellieren machen: Nachdem wir noch ein paar unpassende Vorzeichen umgedreht haben, erhalten wir die folgenden Parameter:

$$A = \begin{bmatrix} 1 & 1 & 1 & & & \\ & & & 1 & 1 & 1 \\ -1 & & & -1 & & \\ & -1 & & & -1 & \\ & & -1 & & & -1 \end{bmatrix}, \qquad b = \begin{bmatrix} 11 \\ 14 \\ -10 \\ -8 \\ -7 \end{bmatrix}, \qquad c = \begin{bmatrix} 8 \\ 6 \\ 10 \\ 9 \\ 5 \\ 7 \end{bmatrix}$$

Und ab geht's in den Computer:

[71] "Shop".

[72] Indizes werden angeordnet wir im Lexikon: zuerst ordnet man nach dem ersten Eintrag, dann nach dem zweiten und so weiter.

```
octave> A = [ 1 1 1 0 0 0; 0 0 0 1 1 1; -1 0 0 -1 0 0; 0 -1 0 0 -1 0;
    0 0 -1 0 0 -1];
octave> b = [ 11 14 -10 -8 -7 ]'; c = [ 8 6 10 9 5 7 ]';
octave> [x,opt] = SimSimplex(A,b,c)
x =

  10.00000
   1.00000
   0.00000
   0.00000
   7.00000
   7.00000

opt = 170
```

Was man sieht ist, daß bei Transportproblemen zwar die Anzahl der Variablen dramatisch steigt (man hat mn Variablen bei einer $m \times n$ Kostenmatrix), daß man dafür aber sehr einfach strukturierte Matrizen hat: In den ersten n Zeilen stehen n verschobene Zeilen von je m Einsen und darunter n nebeneinandergestellte, mit -1 multiplizierte Einheitsmatrizen.

Beispiel 3.3 (Noch ein Transportproblem) *Ausrüstungsgegenstände sollen von drei Basen auf fünf andere Basen verteilt werden, wobei die zurückgelegte Gesamtdistanz minimiert werden soll. Die Vorgaben, wie in (Gass, 1970, S. 20) sind wie folgt:*

	MacDill	*March*	*Davis-Monthan*	*McConnell*	*Pinecastle*	
Oklahoma City	*938*	*1030*	*824*	*136*	*995*	*8*
Macon	*346*	*1818*	*1416*	*806*	*296*	*5*
Columbus	*905*	*1795*	*1590*	*716*	*854*	*8*
	3	*5*	*5*	*5*	*3*	

Das sind jetzt also solide 15 *Variable und da wird's langsam heftig.*

Es wäre jetzt schon ziemlich eklig, diese Matrix noch von Hand einzugeben, weswegen wir ein kleines `Octave`–Programm namens `TransMat` verwenden, das die Strukturmatrix automatisch generiert. Und dann brauchen wir nur noch unsere Werte einzutippen,

```
octave> A = TransMat( 3,5 ); b = [ 8 5 8 -3 -5 -5 -5 -3 ]';
octave> c = [ 938 1030 824 136 995 346 1818 1416 806 296 905 1795
   1590 716 854 ]';
```

um auf die Optimallösung mit 16384 Kilometern zu kommen.

Übung 3.1 Zeigen Sie, daß Transportprobleme mit ganzzahliger Kostenmatrix immer auch ganzzahlige Lösungen haben. ◇

3.3 Zuordnungsprobleme

Zuordnungsprobleme versuchen Ressourcen und Aufgaben so einander zuzuordnen, daß ein vorgegebener Nutzen maximiert wird. Eine **Zuordnung** zweier Mengen[73] ist eine Funktion, die jedem Element der einen Menge[74] *eindeutig* ein Element der anderen Menge[75] zuordnet. Alternativ ist eine **Zuordnungstabelle** eine quadratische Matrix, die in jeder Spalte und jeder Zeile *genau eine* Eins stehen hat.

Abbildung 3.2: Die drei Mitarbeiter und ihre Fähigkeiten. Aus (Gass, 1970)

Beispiel 3.4 (Personal und Fähigkeiten) *Einer Militäreinheit*[76] *stehen drei neue Mitarbeiter, Able, Baker und Charlie, zur Verfügung, die für drei Aufgaben eingesetzt werden können, und zwar am Schreibtisch, am Funkgerät oder am Computer. In vorhergehenden Tests wurden ihre Fähigkeiten wie folgt ermittelt:*

	Funk	*Computer*	*Schreibtisch*
Able	*5*	*4*	*7*
Baker	*6*	*7*	*3*
Charlie	*8*	*11*	*2*

Wie setzt man die drei Soldaten so ein, daß ein möglichst hoher Wert erreicht wird.

[73] Die gleichviele Elemente enthalten müssen.
[74] Also jeder Ressource.
[75] Also eine Aufgabe.
[76] Nicht meine Erfindung, sondern aus (Gass, 1970, S. 56–61)!

Auch hier beschreibt wieder x_{jk}, in welchem Maße Soldat Nummer j Job Nummer k ausübt[77] und die Nebenbedingungen sind

$$\begin{array}{cccccccccccc}
x_{11} & +x_{12} & +x_{13} & & & & & & & & = & 1 \\
 & & & x_{21} & +x_{22} & +x_{23} & & & & & = & 1 \\
 & & & & & & x_{31} & +x_{32} & +x_{33} & & = & 1 \\
x_{11} & & & +x_{21} & & & +x_{31} & & & & = & 1 \\
 & x_{12} & & & +x_{22} & & & +x_{32} & & & = & 1 \\
 & & x_{13} & & & +x_{23} & & & +x_{32} & & = & 1
\end{array}$$

Kommt uns irgendwie bekannt vor, oder? Wenn wir das nämlich in Ungleichungen umschreiben, dann steht da bis auf die rechte Seite nicht anderes als ein „gedoppeltes“ Transportproblem! Und dafür haben wir ja schon unsere Routinen. Aber nicht vergessen: Da wir *maximieren* wollen, müssen wir die Zielfunktion mit -1 multiplizieren. Also:

```
octave> A = TransMat( 3,3 ); b = [ ones( 3,1 ) ; -ones( 3,1 ) ];
octave> A = [ A; -A ]; b = [ b; -b ];
octave> c = -[ 5 4 7 6 7 3 8 11 2 ]';
octave> [x,opt] = SimSimplex( A,b,c )
x =

   0
   0
   1
   1
   0
   0
   0
   1
   0

opt = -24
```

und die Lösung “Able am Schreibtisch, Baker am Funkgerät und Charlie am Computer” ist ja auch das, worauf man ohne Computer hätte kommen können.

3.4 Fluß in Netzwerken

Die letzte Problemklasse sieht schon richtig fortgeschritten, um nicht zu sagen professionell aus und hat sogar einiges mit Informatik zu tun. Es geht darum, auf verschlungenen Wegen möglichst viel von A nach B zu transportieren. Diese verschlungenen Wege werden in Form eines Netzwerks[78] dargstellt, siehe Abb. 3.3. Wie man sieht, gibt es nun viele verschiedene Möglichkeiten vom Startpunkt „S“ zum Zielpunkt „Z“ zu gelangen, beispielsweise den Weg $S \to 1 \to 2 \to Z$ oder $S \to 1 \to 3 \to Z$ und so weiter.

[77] Man kann zeigen, daß bei der Optimallösung $x_{jk} = 1$ sein muß, das liegt an der Struktur des Problems.

[78] In der Sprache der diskreten Mathematik: ein **gerichteter Graph**.

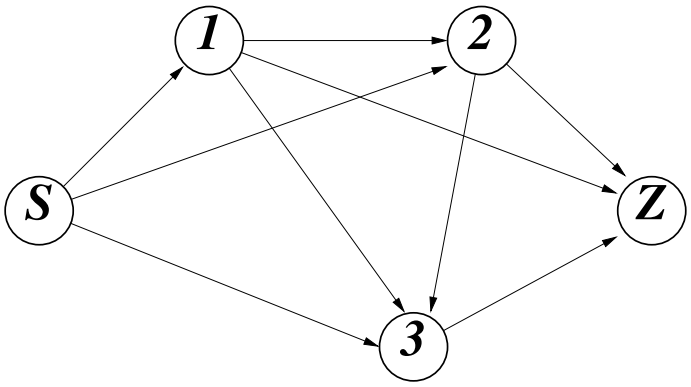

Abbildung 3.3: Die Verbindungen im Netzwerk – wohin kann man von wo aus kommen. Der *gerichtete* Graph bedeutet, daß es keine Schleifen gibt.

Beispiel 3.5 (Maximaler Transport oder Fluß im Netzwerk) *Das Netzwerk aus Abb. 3.3 stelle alle Möglichkeiten dar, mit öffentlichen Verkehrsmitteln von S nach Z zu gelangen, wobei* $1, 2, 3$ *die Umsteigepunkte seien. Wieviele Fahrgäste kann man maximal von S nach Z bringen, wenn die Kapazitäten der Verkehrsmittel*[79] *wie in Abb. 3.4 dargestellt sind, und wie muß man die Fahrgäste auf die einzelnen Verkehrsmittel verteilen?*

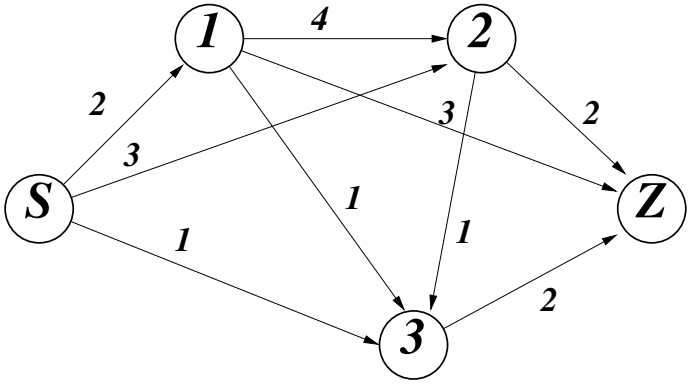

Abbildung 3.4: Die Kapazitäten der einzelnen Kanten des Netzwerk aus Abb. 3.3.

Wieder bezeichnen wir mit x_{jk} die Anzahl der Passagiere, die von Knoten j nach Knoten k fahren, wobei Knoten 0 der Startpunkt und Knoten 4 der Zielpunkt ist. Damit werden die Kapazitätsbeschränkungen, ohne daß wir ir-

[79] Sagen wir in der Einheit „100 Fahrgäste“.

gendwie nachdenken müssen, sofort zu Nebenbedingungen:

$$\begin{array}{rrrrrrrrrcl}
x_{01} & & & & & & & & & \leq & 2 \\
 & x_{02} & & & & & & & & \leq & 3 \\
 & & x_{03} & & & & & & & \leq & 1 \\
 & & & x_{12} & & & & & & \leq & 4 \\
 & & & & x_{13} & & & & & \leq & 1 \\
 & & & & & x_{14} & & & & \leq & 3 \\
 & & & & & & x_{23} & & & \leq & 1 \\
 & & & & & & & x_{24} & & \leq & 2 \\
 & & & & & & & & x_{34} & \leq & 2
\end{array} \tag{3.2}$$

Das war der einfache Teil. Was wir außerdem noch fordern müssen, ist, daß niemand an einem Umsteigepunkt vergessen wird und dort verhungern muß, daß also alles, was in einen Knoten *hineinfließt*, auch wieder *herausfließen* muß, was wir mathematisch als **Erhaltungsprinzip**

$$\sum_j x_{jk} = \sum_j x_{kj}, \qquad \forall k$$

schreiben können: Die Summe über die x_{jk} ist je gerade die Menge, die in den Knoten k hineintransportiert wird und die Summe über die x_{kj} die Menge, die von Knoten k in andere Knoten weitergeleitet wird. In unserem Beispiel entnehmen wir Abb. 3.4 die Nebenbedingungen

$$\begin{array}{rrrrrrrrrcl}
x_{01} & & -x_{12} & -x_{13} & -x_{14} & & & & & = & 0 \\
 & x_{02} & +x_{12} & & & -x_{23} & -x_{24} & & & = & 0 \\
 & & x_{03} & +x_{13} & & +x_{23} & & -x_{34} & & = & 0
\end{array} \tag{3.3}$$

und wie wir die in Ungleichungsbedingungen umwandeln, das wissen wir ja schon. Bleibt noch, daß das was wir in das System reinstecken, also was aus S hinausfließt, auch in Z ankommen muß. Nennen wir diesen Wert t, dann erhalten wir schließlich noch die beiden Nebenbedingungen

$$\begin{array}{rrrrrrrcl}
-x_{01} & -x_{02} & -x_{03} & & & & +t & = & 0 \\
 & & & x_{14} & +x_{24} & +x_{34} & -t & = & 0
\end{array} \tag{3.4}$$

Und was ist unser Ziel? Wir wollen ja den **Gesamtfluß** maximieren, also nichts anderes als den Wert t, der gerade in unserer Nebenbedingung aufgetaucht ist. Dazu müssen wir also t als *zusätzliche* Variable einführen und haben unser Problem fertig modelliert. Jetzt müssen wir es nur noch computergerecht aufbereiten, wobei wir t als zusätzliche, zehnte Variable ansetzen. Das tun wir geschickterweise zuerst für die Nebenbedingungen (3.3) und (3.4), denn die können wir dann mit umgedrehtem Vorzeichen übereinanderstapeln:

```
octave> A = [ 1 0 0 -1 -1 -1 0 0 0 0;
              0 1 0 1 0 0 -1 -1 0 0;
              0 0 1 0 1 0 1 0 -1 0;
              -1 -1 -1 0 0 0 0 0 0 1;
              0 0 0 0 0 1 0 1 1 -1 ];
octave> A = [ A; -A ];
```

Dann fügen wir noch die Nebenbedingungen aus (3.2) hinzu

```
octave> A = [ A; [ eye(9), zeros( 9,1 ) ] ];
```

setzen unsere rechte Seite und die Zielfunktion an, wobei wir beachten müssen, daß wir *maximieren* wollen, also als Zielfunktion $-t$ setzen sollten,

```
octave> b = [ zeros( 1,10 ), [ 2 3 1 4 1 3 1 2 2 ] ]';
octave> c = [ zeros( 1,9 ), -1 ]';
```

und erhalten die Optimallösung als

```
octave> [x,opt] = SimSimplex( A,b,c )
x =

  2
  3
  1
  0
  0
  2
  1
  2
  2
  6

opt = -6
```

Die Lösung ist in Abb. 3.5 dargestellt. Man sieht ihr ein typisches Phänomen

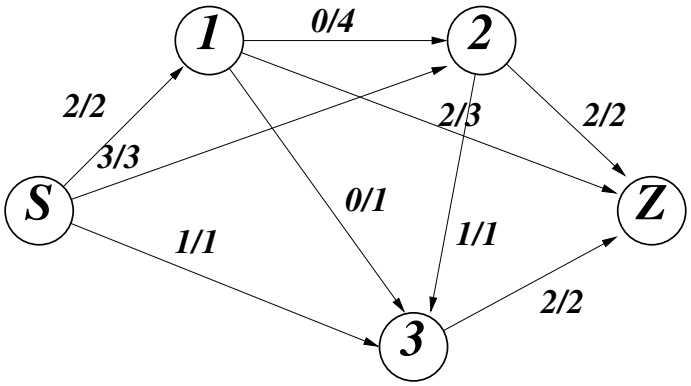

Abbildung 3.5: Fluß der Optimallösung im Vergleich zu deren Kapazität.

von Optimallösungen von Netzwerkproblemen an: Alle Kanten, die aus S herausführen, sind **saturiert**, also voll belegt.
Man kann auch allgemeinere Netzwerkprobleme auf diese Art und Weise angehen, indem man für n Knoten (Start und Ziel mitgezählt!) normalerweise eine **Verbindungsmatrix**

$$V = [v_{jk} \,:\, j,k = 1,\ldots,n]$$

verwendet, in der nur Nullen und Einsen stehen – und zwar $\nu_{jk} = 1$, wenn es zwischen den Knoten j und k eine Verbindung gibt und Null, wenn diese Verbindung nicht existiert. Dabei ist $\nu_{jk} = \nu_{kj} = 1$ zwar möglich, aber nicht zwingend vorgeschrieben[80]. Die Nebenbedingungen ergeben sich dann wieder aus den Kapazitäten der Kanten und aus der Erhaltungseigenschaft, daß alles, was in einen Knoten fließt, auch wieder rausmuss: In formaler Schreibweise heißt das

$$\sum_{j=1}^{n} \nu_{jk}\, x_{jk} = \sum_{j=1}^{n} \nu_{kj}\, x_{kj}, \qquad k = 1, \ldots, n.$$

Übrigens kann man da auch großzügiger sein, indem man lediglich

$$\sum_{j=1}^{n} \nu_{jk}\, x_{jk} \geq \sum_{j=1}^{n} \nu_{kj}\, x_{kj}, \qquad k = 1, \ldots, n$$

fordert – jetzt darf in jedem Knoten auch was versickern. Und dann steckt man nur noch t in das System und minimiert die Zielfunktion $z(x, t) = -t$.

3.5 Ganzzahlprogrammierung

Wir hatten in vielen unserer Anwendungen ja bereits das Problem, daß eigentlich nur Ganzzahllösungen gesucht waren, die sich dann wie durch ein Wunder auch ergeben haben, siehe Beispiel 2.1. Was aber, wenn es nicht funktioniert, wir aber ein Problem der Form

$$\min_x c^T x, \qquad Ax \leq b, \qquad x \in \mathbb{N}_0^n, \tag{3.5}$$

haben, also explizite Ganzzahlbeschränkungen berücksichtigen müssen.

Die naive Idee wäre es, sich damit herauszureden, daß $F(A, b) \cap \mathbb{Z}^s$ ja für beschränkte Polyeder nur endlich viele Punkte enthält und damit eigentlich unproblematisch ist, man müsste nur das Polyeder absuchen und die endlich vielen Werte testen. Praktisch ist das natürlich Nonsens.

Für einen funktionierenden Algorithmis ist es sicherlich keine schlechte Idee, zuerst einmal die **kontinuierliche Lösung** x^* von (3.5) zu bestimmen, also $x \in \mathbb{N}_0^n$ durch $x \geq 0$ zu ersetzen. Ist $x^* \in \mathbb{N}_0^n$, dann haben wir gewonnen, andernfalls gibt es mindestens eine Variable x_j, die einen nichtganzzahligen Wert annimmt, sagen wir

$$x_j = n_j + \alpha_j, \qquad n_j \in \mathbb{N}_0, \qquad \alpha_j \in (0, 1).$$

Die richtige ganzzahlige Lösung erfüllt natürlich nun entweder $x_j \leq n_j$ oder $x_j \geq n_j + 1$. Also bilden wir zwei neue Optimierungsprobleme mit den Nebenbedingungen

$$\begin{bmatrix} A \\ e_j^T \end{bmatrix}, \quad \begin{bmatrix} b \\ n_j \end{bmatrix} \qquad \text{und} \qquad \begin{bmatrix} A \\ -e_j^T \end{bmatrix}, \quad \begin{bmatrix} b \\ -n_j - 1 \end{bmatrix}$$

und minimieren diese separat. Für diese beiden Probleme bestimmen wir

[80] Man denke nur an Einbahnstraßen.

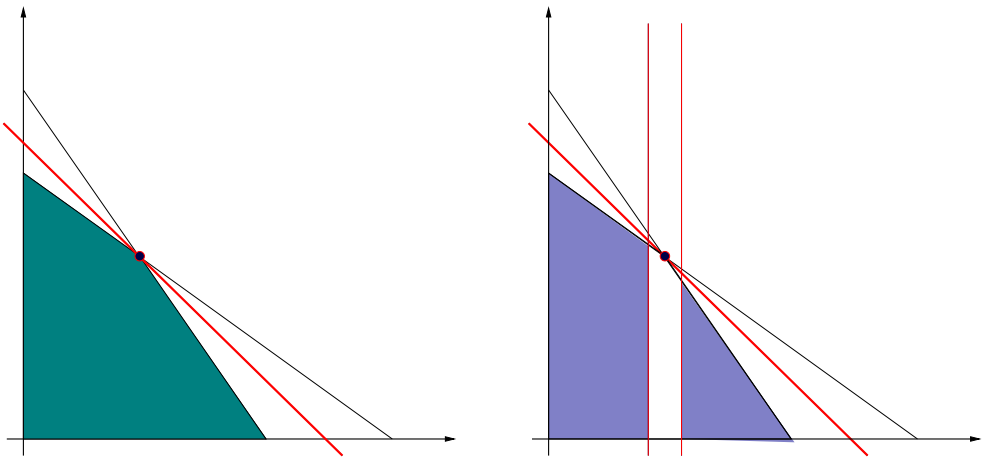

Abbildung 3.6: Die Idee von „Branch & Cut“: Wir separieren das Problem *(links)* in zwei Teilprobleme *(rechts)*, die dann separat gelöst werden sollen.

wieder *kontinuierliche* Optimallösungen und fahren mit dem Prozess solange fort, bis wir Blätter[81] gefunden haben, die nur ganzzahlige Parameterwerte haben[82], oder leere zulässige Menge haben. Sobald wir eine Ganzzahllösung gefunden haben, merken wir uns diesen Wert, denn von da an können wir alle anderen Teilprobleme verwerfen, deren kontinuierliche Lösung einen größeren Wert haben, denn die Restriktionen auf ganzzahlige Lösungen sind ja nur noch „schlechter“.

An diesen Strategien kann man, wenn man will, noch mächtig feilen, um zu wirklich guten Verfahren zu kommen, aber das ist dann schon wieder fast eine eigene Vorlesung für sich.

[81] Also Ecken des Baums.
[82] Und unter diesen bilden wir dann das Minimum.

Der Mensch spielt nur, wo er in voller Bedeutung des Wortes Mensch ist, und er ist nur da ganz Mensch, wo er spielt.

F. Schiller

4 Spieltheorie

Eine wunderschöne Anwendung von Optimierungsmethoden und eine Optimierungstheorie für sich ist die **Spieltheorie**. Und genau deswegen machen wir uns jetzt den Spass, ein wenig in die Grundlagen der Spieltheorie einzutauchen.

4.1 Grundideen der Spieltheorie: Bei–Spiele

Spieltheorie befasst sich mit der Frage, wie man in „Konfliktsituationen" optimale Entscheidungen trifft und ist damit auch wieder eine Form der Optimierung. Um es mit Karlin (Karlin, 1959) zu sagen:

> *The art of making optimal judgments according to various criteria is as old as mankind; it is the essence of every field of endeavor from volleyball to logistics*[83]*. The science of making such judgments, as opposed to the mere art, is a newer development . . .*

Im diesem ersten Kapitel wollen wir uns anhand von ein paar Bei–Spielen eine Übersicht über die wesentlichen Ansätze und Ideen verschaffen, bevor wir uns dann vertieft an die *mathematischen* Grundlagen machen.

Ein klassisches Kinder- und nicht nur Kinderspiel ist „Stein, Schere, Papier", bei dem zwei Spieler, nennen wir sie kreativ S_1 und S_2, gleichzeitig mit jeweils einer Hand entweder einen Stein (Faust), eine Schere (Zeige- und Mittelfinger gespreizt) oder Papier (flache Hand) darstellen. Der Gewinner wird dann nach folgenden Regeln ermittelt:

1. Der Stein macht die Schere stumpf, also gewinnt Stein gegen Schere.
2. Die Schere schneidet Papier, also gewinnt die Schere gegen das Papier.
3. Das Papier wickelt den Stein ein, also gewinnt Papier gegen Stein.

Man sieht, die Situation ist schön symmetrisch und es gibt entweder ein Unentschieden[84] oder ein Spieler gewinnt und der andere verliert. In letzterem Fall

[83] Wir werden uns im Rahmen der Vorlesung allerdings nicht mit Volleyball befassen, so viel ist eigentlich vorhersagbar.

[84] Wenn beide Spieler dieselbe Wahl getroffen haben.

können wir davon ausgehen, daß ein Gewinn[85] vom Verlierer an den Sieger übertragen wird. Diese Situation bezeichnet man als ein **Nullsummenspiel**: *Was einer gewinnt, das muss der andere verlieren*. Normieren wir den Gewinn zu 1, dann können wir den Ablauf des Spiels aus der Sicht der Spieler in Tabellen darstellen:

$S_1 \setminus S_2$	St	Sch	P
St	0	1	-1
Sch	-1	0	1
P	1	-1	0

$S_1 \setminus S_2$	St	Sch	P
St	0	-1	1
Sch	1	0	-1
P	-1	1	0

Das ist jeweils die **Auszahlungstabelle** für S_1 (links) und S_2 (rechts), was sich in Matrixnorm viel schöner darstellen lässt:

$$A_1 = \begin{bmatrix} 0 & 1 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{bmatrix}, \qquad A_2 = \begin{bmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{bmatrix}, \tag{4.1}$$

und die Tatsache, daß wir es mit einem Nullsummenspiel zu tun haben, ist dann schlicht und ergreifend äquivalent zu $A_1 + A_2 = 0$. So können wir natürlich nun jedes Zweipersonenspiel darstellen: Hat Spieler S_1 die Wahlmöglichkeiten $s_{1,1}, \ldots, s_{1,m}$, die man jeweils als **Strategie** bezeichnet und Spieler S_2 entsprechend[86] $s_{2,1}, \ldots, s_{2,n}$, dann stellen wir das Spiel aus der Sicht jedes Spielers durch eine **Auszahlungsmatrix** $A_1 \in \mathbb{R}^{m\times n}$ bzw. $A_2 \in \mathbb{R}^{m\times n}$ dar. Und nochmals langsam: In unserem Beispiel haben wir es mit den symmetrischen Strategien

$$s_{11} = s_{21} = s_1 = \text{Stein}, \qquad s_{21} = s_{22} = s_2 = \text{Schere}, \qquad s_{31} = s_{32} = s_3 = \text{Papier}$$

zu tun.

Beispiel 4.1 (Gefangenendilemma) *Zwei Verbrecher werden festgenommen und getrennt voneinander verhört. Der Polizei ist klar, daß sie beide eine Menge auf dem Kerbholz haben, kann aber leider nur kleinere Vergehen wirklich nachweisen*[87]*. Deswegen erhalten beide das Angebot, als Kronzeuge gegen den anderen auszusagen – in diesem Falle würde der Kronzeuge einen Strafnachlass erhalten, der nicht geständige Verbrecher hingegen die volle Härte der Justiz geniessen dürfen. Wenn natürlich beide geständig sind, dann braucht man keinen Kronzeugen mehr und der „Rabatt" ist dahin. In negativen Jahren Gefängnis sehen also die Auszahlungsmatrizen wie folgt aus*[88]*:*

$$A_1 = A_2^{\mathsf{T}} = \begin{bmatrix} -1 & -5 \\ -3 & -7 \end{bmatrix},$$

wobei s_1 = *„schweigen" und* s_2 = *„aussagen" ist. Ist es besser zu schweigen oder auszusagen? Was ist die optimale Strategie?*

[85] Beispielsweise eine Einheit Geld – der Phantasie sind hier keine Grenzen gesetzt, aber für unsere Zwecke hier ist es völlig irrelevant.

[86] Es hat niemand gesagt, daß Spiele immer symmetrisch sein müssen und daß beide Spieler immer dieselbe Anzahl an Strategien haben.

[87] Das ist nicht so unrealistisch: So wurde beispielsweise Al Capone nur wegen Steuerhinterziehung angeklagt und verurteilt.

[88] Und im Gegensatz zu „Stein, Schere, Papier" liegt das Interesse der Spieler daran, die „Auszahlung" in Jahren zu *minimieren*, deswegen der Vorzeichenwechsel.

Beispiel 4.1 ist kein Nullsummenspiel, denn

$$A_1 + A_2 = \begin{bmatrix} -2 & -8 \\ -8 & -14 \end{bmatrix} \neq 0.$$

Trotzdem kann man es zu einem Nullsummenspiel machen, indem man einen Spieler S_3 einführt, der nur eine Strategie und die Auszahlungsmatrix $A_3 = -A_1 - A_2$ hat – trivialerweise ist dann $A_1 + A_2 + A_3 = 0$. Allerdings haben wir es dann nicht mehr mit einem **Zweipersonenspiel** zu tun, sondern mit einem **Dreipersonenspiel** und wir sehen bereits eine Eigenart von n–Personenspielen mit $n > 2$: Es wird **Kooperation** möglich, bei der mehrere Spieler gegen andere Spiele Koalitionen bilden können.

Beispiel 4.2 (Koalitionen) *Das klassische und aus der Politik bekannte Mehrpersonenspiel heißt "Koalition". Nehmen wir an, ein Parlament habe 9 Sitze, die sich auf die drei Parteien S, C und F wie folgt verteilen:*

Partei	*S*	*C*	*F*
Sitze	*4*	*4*	*1*

Es ist naheliegend, daß die beiden großen Parteien lieber mit F koalieren werden als miteinander, denn sie werden in solch einer Koalition natürlich ein größeres Stück vom Kuchen bekommen[89]*. Was ist also das "optimale" Angebot, das die Parteien einander für Koalitionen machen können?*

4.2 Reine und gemischte Strategien

Beginnen wir einmal mit etwas Begrifflichkeiten, wobei wir uns an den „Standardwerken" (Karlin, 1959) und (Neumann & Morgenstern, 1944) orientieren wollen. Die meisten der Konzepte und Ideen wurden sogar bereits 1928 von John von Neumann[90] (Neumann, 1928) in der anscheinend allerersten Arbeit zum Thema Spieltheorie angegeben.

Definition 4.3 *Ein n–**Personen–Spiel** $\mathscr{S}$, $n \in \mathbb{N}$, besteht für jeden Spieler aus einer **Strategiemenge** S_j, $j = 1, \ldots, n$ und einer **Auszahlungsfunktion** $a : S_1 \times \cdots \times S_n \to \mathbb{R}^n$, die die Auszahlung an die Spieler unter Verwendung der jeweiligen Strategien angibt.*

1. *Das Spiel $\mathscr{S}$ heißt **Zweipersonenspiel**, wenn $n = 2$ ist*[91].
2. *Das Spiel $\mathscr{S}$ heißt **Nullsummenspiel**, wenn*

$$0 \equiv 1^T a = \sum_{j=1}^{n} a_j, \qquad \textit{d.h.} \qquad 0 = \sum_{j=1}^{n} a_j(s_1, \ldots, s_n), \quad s_j \in S_j,\ j = 1, \ldots, n.$$

[89] Und wer an die Sonntagsrede von der Bedeutung einer starken Opposition glaubt ist ohnehin selbst schuld.

[90] Der auch der Vater des *von Neumannschen Universalrechners* ist, einer erstaunlich genauen Beschreibung des Digitalcomputers noch bevor solche Geräte realisiert werden konnten.

[91] Das sollte niemanden so wirklich überraschen.

3. *Sind die Strategiemengen* S_j *endlich, d.h.* $s_j := \#S_j < \infty$, $j = 1, \dots, n$, *dann setzen wir* $S_j = \{1, \dots, s_j\}$, $\sigma = (s_1, \dots, s_n)$, *und erhalten für jeden Spieler eine* Auszahlungsmatrix

$$A_j = [A_{j,\alpha} = a_j(\alpha_1, \dots, \alpha_n) \ : \ \alpha \in \mathbb{N}^n, \ \alpha \le \sigma], \qquad j = 1, \dots, n,$$

wobei

$$\alpha \le \beta \qquad \Leftrightarrow \qquad \alpha_j \le \beta_j, \quad j = 1, \dots, n,$$

eine gebräuchliche Halbordnung für Multiindizes $\alpha, \beta \in \mathbb{N}^n$ *darstellt.*

4. *Im Fall eines Zweipersonen–Nullsummenspiels mit endlicher Strategiemenge ist das Spiel durch eine Matrix* $A \in \mathbb{R}^{S_1 \times S_2}$ *beschrieben, für die* $A_1 = A$ *und* $A_2 = -A$ *gilt.*

Suchen wir nun nach der *besten Strategie*, dann ist „Stein, Schere, Papier" ein gutes Beispiel: Die eigentliche Kunst des Spiels liegt in der Psychologie, also in der Fähigkeit, die Entscheidung des Gegners vorherzusagen und entsprechend dagegen zu handeln – das ist die wirkliche Bedeutung des im Sport so oft mißbrauchten Wortes **antizipieren** . Tatsächlich könnten wir auch vom „Elfmeterdilemma" sprechen. Nur ist Psychologie leider mathematisch nicht wirklich fassbar[92] und deswegen gehen wir bei der Bestimmung der optimalen Lösung immer davon aus, daß wir es mit einem Gegner zu tun haben, der vernünftig agiert[93]. Die Optimallösung versucht nun, die Strategie so zu wählen, daß *bei bester Wahl des Gegners das beste Ergebnis erreicht wird*.

Dabei betrachten wir zuerst das Nullsummenspiel[94]! Hier wird S_1 sich also alle seine Strategien $s_{11}, \dots, s_{1m}$ ansehen und annehmen, daß S_2 clever genug ist, in jedem Fall die **beste Gegenstratgie** zu wählen, daß also für $j = 1, \dots, m$ die Strategie s_{2k} mit von j abhängigem k gewählt wird, so daß

$$(A_2)_{jk} = \max_{k'=1,\dots,n} (A_2)_{jk'} = \max_{k'=1,\dots,n} (-A_1)_{jk'} = \min_{k'=1,\dots,n} (A_1)_{jk'},$$

und dann wird S_1 das beste unter diesen j auswählen:

$$(A_1)_{jk} = \max_{j'=1,\dots,m} \min_{k'=1,\dots,n} (A_1)_{j',k'}. \tag{4.2}$$

Diese Vorgehensweise bestimmt aber nur j! Der Index k der Strategie, die S_2 wählt, ergibt sich als Lösung des **Minimax–Problem**

$$(A_1)_{jk} = \min_{k'=1,\dots,n} \max_{j'=1,\dots,m} (A_1)_{j',k'}, \tag{4.3}$$

und im allgemeinen gilt, daß $\max_j \min_k \neq \min_k \max_j$ – im Falle von Gleichheit spricht man dann von einem **Sattelpunkt**.

[92] Zumindest beim Elfmeter stimmt das nicht so wirklich: Schützen wie Torhüter haben normalerweise Präferenzen, die sich aus der „Händigkeit" ergeben, aber das müsste man dann halt in der Auszahlungsmatrix berücksichtigen.

[93] Was nicht erst seit Loriots klassischem Sketch vom Skatspieler nicht immer und uneingeschränkt zutreffen muss.

[94] Und dabei bleiben wir im Rahmen dieser Vorlesung auch. Mehrpersonenspiele und Nichtnullsummenspiele sind dann doch ein klein wenig komplexer.

Übung 4.1 Geben Sie eine Matrix $A = [a_{jk}]$ an, so daß

$$\max_j \min_k a_{jk} \neq \min_k \max_j a_{jk}$$

ist. ◊

Beispiel 4.4 *S_1 will seine optimale Strategie für „Stein, Schere, Papier" bestimmen und bildet daher für jedes j, also jede Zeile von A_1 die Minima:*

$$\begin{bmatrix} 0 & \boxed{1} & -1 \\ -1 & 0 & \boxed{1} \\ \boxed{1} & -1 & 0 \end{bmatrix} \begin{matrix} 1 \\ 1 \\ 1 \end{matrix},$$

und wenn er das nun maximiert, dann ist er so schlau wie zuvor, denn das Maximum wird für alle drei Strategien angenommen.

Beispiel 4.5 *Natürlich gibt es auch Matrizen mit Sattelpunkt, beispielsweise*

$$\begin{bmatrix} 1 & 2 \\ -1 & -2 \end{bmatrix},$$

denn hier liefert max min *die Auswahl*

$$\begin{bmatrix} \boxed{1} & 2 \\ -1 & \boxed{-2} \end{bmatrix} \begin{matrix} 1 \\ -2 \end{matrix} \qquad \rightarrow \qquad \begin{bmatrix} \boxed{1} & 2 \\ -1 & -2 \end{bmatrix},$$

was dasselbe Resultat liefert wie die min max*–Suche:*

$$\begin{bmatrix} \boxed{1} & \boxed{2} \\ -1 & -2 \\ 1 & 2 \end{bmatrix} \qquad \rightarrow \qquad \begin{bmatrix} \boxed{1} & 2 \\ -1 & -2 \end{bmatrix}.$$

Eine **reine Strategie** zu wählen liefert uns also, wie Beispiel 4.4 zeigt, keine besonders aufregenden Resultate für „Stein, Schere, Papier" – alle Strategien sind absolut gleichwertig und ohnehin sind Spiele mit Sattelpunkt generell nicht besonders interessant und spielenswert sind, denn beide Spieler *müssten* dann auf dieser Strategie beharren. Schach ist übrigens ein Spiel mit Sattelpunkt und daher prinzipiell uninteressant. Daher nehmen wir einen etwas anderen Standpunkt ein und nehmen an, wir würden nicht eine Partie spielen, sondern mehrere und die Spieler wählen jede ihrer Strategien mit einer Wahrscheinlichkeit p_j bzw. q_k, $j = 1, \ldots, m$, $k = 1, \ldots, n$. Eine derartige **gemischte Strategie** kann auch eine reine Strategie werden, indem man ein p_j und ein q_k gleich Eins setzt. Die **erwartete Auszahlung** für ein Nullsummenspiel ist dann, aus Sicht von Spieler 1,

$$E(A, p, q) := \sum_{j=1}^{m} \sum_{k=1}^{n} A_{jk} p_j q_k = p^T A q, \qquad p = \begin{bmatrix} p_1 \\ \vdots \\ p_m \end{bmatrix}, \quad q = \begin{bmatrix} q_1 \\ \vdots \\ q_n \end{bmatrix},$$

und Spieler 1 versucht nun, den Wahrscheinlichkeitsvektor p so zu wählen, daß dieser Ausdruck maximiert wird, wohingegen Spieler 2 sein q so wählt, daß der Ausdruck minimiert wird[95].

[95] Was ja nichts anderes bedeutet, als die erwartete Auszahlung $-E(A, p, q)$ für Spieler 2 zu minimieren.

4.3 Ein ganz einfaches Beispiel

Schauen wir uns einmal den einfachsten Fall eines Spieles an, nämlich ein Zweipersonen–Nullsummenspiel mit nur jeweils zwei Strategien, also der Auszahlungsmatrix

$$A_1 = -A_2 = \begin{bmatrix} a & b \\ c & d \end{bmatrix}.$$

Bei den gemischten Strategien $(p, 1-p)$ und $(q, 1-q)$ ist die zu erwartende Auszahlung

$$\begin{aligned} E(p,q) &= apq + bp(1-q) + c(1-p)q + d(1-p)(1-q) \\ &= d + p(b-d) + q(c-d) + pq(a-b-c+d) \\ &= d + p(b-d) + q\left[(a-b-c+d)p + c - d\right] \\ &= d + q(c-d) + p\left[(a-b-c+d)q + b - d\right]. \end{aligned}$$

Ist also $a + d \neq b + c$, dann können beide Spieler das Spiel *unabhängig* von der Wahl des Gegners machen, indem sie

$$p = \frac{d-c}{a-b+d-c}, \qquad \text{und} \qquad q = \frac{d-b}{a-c+d-b} \tag{4.4}$$

wählen. Nun sollten das natürlich auch Wahrscheinlichkeiten sein, also zwischen 0 und 1 liegen, mit anderen Worten es sollte

$$1 \leq p^{-1} = \frac{a-b+d-c}{d-c} = 1 + \frac{a-b}{d-c} \qquad \Rightarrow \qquad \frac{a-b}{d-c} > 0$$

sein. Wenn das nicht der Fall ist, dann haben $b-a$ und $d-c$ dasselbe Vorzeichen. Sind diese Vorzeichen beide

positiv, das heißt, ist $b > a$ und $d > c$, dann wird Spieler 2 auf jeden Fall Strategie 1 wählen, sind sie beide

negativ, dann ist $b < a$ und $d < c$ und Spieler 2 wird sich auf alle Fälle für Strategie 2 entscheiden[96],

aber in beiden Fällen kann Spieler 1 sich nun die für ihn günstigere Zeile aussuchen und beide Spieler würden sich nur verschlechtern, wenn sie die Strategie wechseln. Mit anderen Worten: *Das Spiel hat einen Sattelpunkt*. Dasselbe Argument trifft natürlich auch für die Wahle von q zu und wir können die folgende Beobachtung machen.

Lemma 4.6 *Die nach (4.4) bestimmten Zahlen* p *und* q *liegen genau dann in* $[0,1]$ *und sind somit Wahrscheinlichkeiten bzw. gemischte Strategien, wenn das Spiel* keinen *Sattelpunkt hat.*

[96] Sofern Spieler 2 vernünftig spielt ...

Damit sind wir im Geschäft: Entweder hat das Spiel einen Sattelpunkt und wird sich auf diesem einpendeln, oder es gibt zwei gemischte Strategien, die den Ausgang des Spieles jeweils von der anderen Strategie unabhängig machen. Nennen wir diese beiden p^* und q^*. Dann ist

$$v := E(p^*, q^*) = \min_{q\in[0,1]} E(p^*, q) = \max_{p\in[0,1]} E(p, q^*)$$

und somit

$$\max_{p\in[0,1]} \min_{q\in[0,1]} E(p, q) \geq v \geq \min_{q\in[0,1]} \max_{p\in[0,1]} E(p, q)\,. \tag{4.5}$$

Umgekehrt gilt aber sogar die folgende allgemeine Tatsache.

Lemma 4.7 *Für jede reellwertige Funktion* $f : X \times Y \to \mathbb{R}$ *gilt*

$$\inf_{y\in Y} \sup_{x\in X} f(x, y) \geq \sup_{x\in X} \inf_{y\in Y} f(x, y). \tag{4.6}$$

Sind X *und* Y *kompakt und ist* f *stetig, dann kann man* sup *und* inf *auch durch* min *und* max *ersetzen.*

Beweis: Für jedes feste $y \in Y$ ist $\sup_{x\in X} f(x, y) \geq f(x, y)$, also ist für $\xi \in [0, 1]$

$$g(\xi, y) := \sup_{x\in X} f(x, y) - f(\xi, y) \geq 0 \qquad \Rightarrow \qquad g(\xi) := \inf_{y\in Y} g(\xi, y) \geq 0,$$

also auch

$$\begin{aligned} 0 \;&\leq\; \sup_{\xi\in X} g(\xi) = \sup_{\xi\in X} \inf_{y\in Y} g(\xi, y) = \sup_{\xi\in X} \inf_{y\in Y} \left(\sup_{x\in X} f(x, y) - f(\xi, y) \right) \\ &=\; \sup_{\xi\in X} \inf_{y\in Y} \sup_{x\in X} f(x, y) - \sup_{\xi\in X} \inf_{y\in Y} f(\xi, y) = \inf_{y\in Y} \sup_{x\in X} f(x, y) - \sup_{x\in X} \inf_{y\in Y} f(x, y), \end{aligned}$$

wie behauptet. Die Kompaktheits- und Stetigkeitsaussagen sollten aus der Analysis bekannt sein, siehe (Heuser, 1984). □

Kombinieren wir also nun (4.6) mit (4.5), dann ist letztendlich

$$v = E(A, p^*, q^*) = \max_{p\in[0,1]} \min_{q\in[0,1]} E(A, p, q) = \min_{q\in[0,1]} \max_{p\in[0,1]} E(A, p, q)$$

und p^* und q^* sind eine[97] *optimale Strategie* für Spieler 1 bzw. Spieler 2. Der Erwartungswert v bei optimaler Spielweise heißt auch **Wert**[98] des Spiels; ist er positiv, dann bevorzugt das Spiel Spieler 1, ist er negativ, dann kommt Spieler 2 besser davon und ist er gleich Null, dann nennt man das Spiel *fair*.

Übung 4.2 [Das Daiquiri–Spiel aus (Williams, 1986)] Zwei Männer, Alex und Olaf, siehe Abb. 4.1, sitzen in einer Bar und vereinbaren folgendes Spiel: Beide legen legen jeweils ein oder zwei Streichhölzer für den anderen unsichtbar auf den Tresen. Stimmen die Zahlen überein, so muß Alex seinem Freund Olaf diese Anzahl an Daiquiris ausgeben (zu je 5.50 Euro), andernfalls kommt Alex mit der Zahlung von einem Euro davon. Welchen Betrag muß Olaf vorher an Alex geben, damit das Spiel fair ist? ◇

Abbildung 4.1: Alex und Olaf aus (Williams, 1986).

Auch „mehrzügige“ Spiele mit Entscheidungen lassen sich in den Formalismus der Auszahlungsmatrix bringen, was anhand eines Spieles aus (Williams, 1986) illustriert[99] werden soll.

Beispiel 4.8 (Russisch Roulette) *Um das „normale“ russische Roulette (ein Trommelrevolver mit 6 Patronen, von denen nur eine scharf ist, die anderen fünf sind blind[100]) etwas interessanter zu machen spielen A und B es mit Einsatz. Das Spiel läuft folgendermaßen ab:*

1. *Jeder Spieler setzt eine Gewinneinheit[101].*
2. *Spieler A kann nun entweder zwei weitere Einheiten setzen und den Revolver an B weitergeben (“Passen”) oder eine Einheit setzen, den Revolver an seine Schläfe halten und abdrücken.*
3. *Hat Spieler A seinen Zug überlebt, dann hat Spieler B dieselben Optionen, darf aber die Trommel nicht verändern.*
4. *Das Spiel ist zuende und die überlebenden Spieler teilen die Einsätze untereinander auf.*

Der Spielbaum zu Beispiel 4.8 findet sich in Abb. 4.2, man sieht, daß es sich offenbar um ein Nullsummenspiel handelt. Spieler A hat zwei Strategien, Passen (P) und Spielen (S), wohingegen Spieler B vier Strategien hat: Immer Spielen (S), immer Passen (P), dieselbe Strategie wie A (A) oder die entgegengesetzte Strategie (E). Gewichten wir die Auszahlungen schließlich mit den jeweiligen Wahrscheinlichkeiten, so erhalten wir die folgende Auszahlungsmatrix, genauer, die folgende Matrix von erwarteten Auszahlungen:

[97] Wir haben nie behauptet, Optimalstrategien seien eindeutig. Und sie sind es im allgemeinen auch nicht!

[98] „*v*alue“

[99] Dieses Spiel ist zur Nachahmung **nicht** empfohlen! Insbesondere nicht für Kinder!

[100] Es hat mit Sicherheit auch schon “Experten” gegeben, die Platzpatronen verwendet haben – wenn man sich die Pistole dann an die Schläfe hält, ist der Unterschied zu scharfen Patronen eher marginal.

[101] Im Originaltext von (Williams, 1986) geht es um Zigarettenpäckchen, so daß den Spielern nur die Wahl zwischen schnellem und langsamem Tod bleibt.

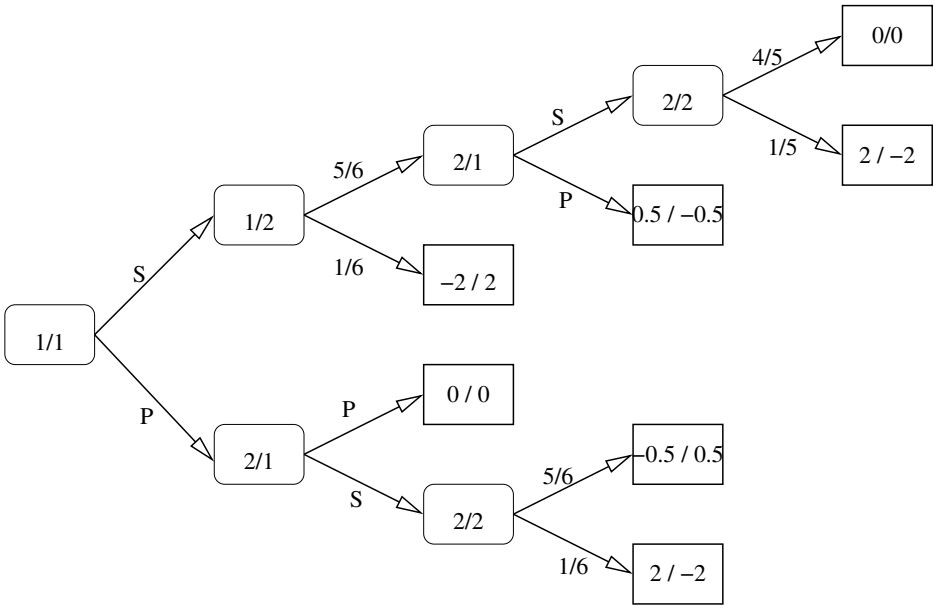

Abbildung 4.2: Der Ablauf des "Russisch Roulette" aus Beispiel 4.8. Bei den Zügen ist angegeben, welcher Spieler am Zug ist und ob es ein Zug erster oder zweiter Art ist, bei den Ergebnissen die Auszahlung.

	S	P	A	E
S	0	1/12	0	1/12
P	−1/12	0	0	−1/12

oder, als Matrix,

$$A = \begin{bmatrix} 0 & \frac{1}{12} & 0 & \frac{1}{12} \\ -\frac{1}{12} & 0 & 0 & -\frac{1}{12} \end{bmatrix} \tag{4.7}$$

Das Spiel ist erstaunlich fair und symmetrisch und die optimale Strategie besteht für beide Spieler darin, zu spielen – der Erwartungswert ist dann Null[102].

Übung 4.3 Bestimmen Sie die Auszahlungsmatrix für den Fall, daß Spieler B die Revolvertrommel nochmals drehen darf. Ist das Spiel dann immer noch fair? ⋄

Doch die Matrix A in (4.7) hat noch eine weitere interessante Eigenschaft: Jeder Eintrag in der ersten Zeile dominiert den zugehörigen Eintrag der zweiten Zeile, auf "mathematisch" $a_{1k} \geq a_{2k}$, $k = 1, 2, 3, 4$. Damit gibt es aber für A gar keinen rationalen Grund, jemals zu passen, denn was auch immer B tun wird, er fährt immer schlechter.

Definition 4.9 (Dominanz von Zeilen und Spalten) *Wir sagen die Zeile* j *der Matrix* $A \in \mathbb{R}^{m\times n}$ *ist* ***dominant*** *gegenüber Zeile* $j' \neq j$*, wenn*

$$a_{jk} \geq a_{j'k}, \qquad k = 1, \ldots, n.$$

[102] Das entspricht der Anschauung: Was hat man bei diesem Spiel schon zu gewinnen?

*Analog definiert sich die **Dominanz** von Spalten. Eine Strategie j für Spieler A heißt **redundant**, wenn die Zeile j von einer anderen Zeile j' dominiert wird, und entsprechend auch für Strategien von B, also Spalten der Matrix A.*

Redundante Strategien können wir also getrost entfernen ohne das Spiel relevant zu verändern, da kein rationaler Spieler sie jemals spielen wird. Tun wir das mit der Matrix aus (4.7), dann wird zuerst die Strategie "P" von Spieler A redundant und dann die beiden Strategien "P" und "E" von Spieler B, der also nur noch spielen oder genauso handeln wird wie A, was, nachdem A ja immer spielt, dasselbe ist. Mit anderen Worten:

> *Die optimale Strategie für beide Spieler besteht darin, zu spielen, nicht zu passen.*

Beispiel 4.10 (Stein, Schere, Papier, Brunnen) *In der Spielpraxis wird "Stein, Schere, Papeier" oftmals noch um die Option* Brunnen *erweitert, wobei der Brunnen Stein und Schere besiegt (beide fallen hinein), aber gegen das Papier verliert, das ihn abdeckt; die Spielmatrix ist also*

$$\begin{bmatrix} 0 & 1 & -1 & -1 \\ -1 & 0 & 1 & -1 \\ 1 & -1 & 0 & 1 \\ 1 & 1 & -1 & 0 \end{bmatrix}$$

Nun dominiert aber die letzte Zeile[103] *dieser Matrix, also die Strategie "Brunnen", die erste Zeile, die zum "Stein" gehört und somit gibt es für Spieler 1 keine Veranlassung, jemals Stein zu spielen – er würde mit Brunnen ja immer besser abschneiden. Entfernt man die erste Zeile, dann sieht man aber auch, daß die erste Spalte (also wieder der Stein) jetzt die letzte Spalte dominiert, daß also auch Spieler 2 vom Stein besser die Finger lässt. Somit erhält man durch die Streichungen*

$$\begin{bmatrix} \boxed{\begin{matrix} 0 & 1 & -1 & -1 \end{matrix}} \\ -1 & 0 & 1 & -1 \\ 1 & -1 & 0 & 1 \\ 1 & 1 & -1 & 0 \end{bmatrix} \to \begin{bmatrix} \boxed{-1} & 0 & 1 & -1 \\ 1 & -1 & 0 & 1 \\ 1 & 1 & -1 & 0 \end{bmatrix} \to \begin{bmatrix} 0 & 1 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{bmatrix},$$

womit wieder genau "Stein, Schere, Papier" übrigbleibt. Der Brunnen bringt dem Spiel also absolut nichts!

4.4 Das Minimax–Theorem

Und schon sind wir also beim ersten Hauptsatz der Spieltheorie, der zuerst in (Neumann, 1928) formuliert und bewiesen wurde.

Satz 4.11 (Minimax) *Für jede Matrix $A \in \mathbb{R}^{m\times n}$ gibt es eine Zahl $v \in \mathbb{R}$, so daß*

$$v = \max_{p\in\Delta_m} \min_{q\in\Delta_n} p^T \mathbf{A} q = \min_{q\in\Delta_n} \max_{p\in\Delta_m} p^T \mathbf{A} q. \tag{4.8}$$

[103] Vielen Dank an Rolf Klaas, der mich auf diese einfache und elegante Lösung hingewiesen hat.

Definition 4.12 *Die Zahl v aus (4.8) heißt **Wert des Spieles** und man nennt ein Spiel **fair**, wenn $v = 0$ ist.*

Beispiel 4.13 („Skin game" aus (Gass, 1970, S. 121–124)) *Die Kunst beim Design von Spielen besteht darin, die Auszahlung so zu wählen, daß das Spiel zwar fair erscheint, es aber nicht ist. Ein schönes Beispiel ist das „Skin Game", bei dem beide Spieler ein As (Wert 1) und eine Zwei (Wert 2) der Farben Karo und Kreuz erhalten, außerdem hat Spieler 1 (der „Carnival Man", der dieses Spiel anderen anbietet) die Karo 2 und sein Gegenspieler die Kreuz 2. Die Regel ist einfach: Bei gleicher Farbe gewinnt Spieler 1, bei unterschiedlichen Farben Spieler 2, die beiden Zweien werden als Unentschieden gewertet. Der Auszahlungsbetrag ist der Wert der Karte die der Sieger gespielt hat, was zur folgenden Auszahlungsmatrix aus Sicht von Spieler 1 führt:*

	♢A	♣A	♣2
♢A	1	−1	−2
♣A	−1	1	1
♢2	2	−1	0

$$\Rightarrow \qquad \mathbf{A} = \begin{bmatrix} 1 & -1 & -2 \\ -1 & 1 & 1 \\ 2 & -1 & 0 \end{bmatrix}.$$

Das sieht doch eigentlich alles sehr fair und symmetrisch aus, aber ist es das auch? Wir werden sehen!

Übung 4.4 Ist das "Skin Game" aus Beispiel 4.13 fair? Bestimmen Sie den Wert des Spiels!
Hinweis: Achten Sie auf Dominanzen. ♢

Korollar 4.14 *Die erwartete Mindestauszahlung für Spieler 1 bei optimaler Spielweise beträgt v, die für Spieler 2 $-v$.*

Korollar 4.15 *Symmetrische Spiele sind fair: Ist A schiefsymmetrisch, dann ist $v = 0$.*

Beweis: Mit $A^T = -A$, also insbesondere $m = n$ folgt, daß $p^T A q = q^T A^T p = -q^T A p$ und daher ist

$$\begin{aligned} v &= \max_{p\in\Delta_m} \min_{q\in\Delta_m} p^T A q = \max_{p\in\Delta_m} \min_{q\in\Delta_m} -q^T A p = \max_{p\in\Delta_m} \left(-\max_{q\in\Delta_m} q^T A p \right) \\ &= -\min_{p\in\Delta_m} \max_{q\in\Delta_m} q^T A p = -\min_{q\in\Delta_n} \max_{p\in\Delta_m} p^T A q = -v, \end{aligned}$$

also $v = 0$. □

Der Beweis von Satz 4.11 ist ein bißchen aufwendig und folgt der Darstellung aus (Neumann & Morgenstern, 1944), der aber **nicht** die "Originalversion" aus (Neumann, 1928) ist. Tatsächlich gibt es mehrere Beweise, einen auf der Basis von Fixpunktsätzen, oder eben auch den, den wir hier sehen werden. Der Einfachheit halber verwenden wir jetzt die *erwartete Auszahlungsfunktion*

$$a(p, q) = p^T A q.$$

Die erste Beobachtung sagt uns, daß wir für die Bestimmung extremaler Strategien nur Eckpunkte des Simplex betrachten müssen.

Lemma 4.16 *Für* $p \in \Delta_m$ *und* $q \in \Delta_n$ *gilt*

$$\min_{q' \in \Delta_n} a(p, q') = \min_{k=1,\dots,n} \sum_{j=1}^{m} a_{jk}\, p_j = \min_{k=1,\dots,n} \left(p^T A\right)_k \tag{4.9}$$

und

$$\max_{p' \in \Delta_m} a(p', q) = \max_{j=1,\dots,m} \sum_{k=1}^{n} a_{jk}\, q_k = \max_{j=1,\dots,m} (Aq)_j\,. \tag{4.10}$$

Beweis: Für beliebige k Strategien $q_1, \dots, q_k \in \Delta_n$, $k \in \mathbb{N}$, sowie $\alpha \in \Delta_k$ und $p \in \Delta_m$ ist

$$a\left(p, \sum_{\ell=1}^{k} \alpha_\ell\, q_\ell\right) = \sum_{\ell=1}^{k} p^T A\,(\alpha_\ell\, q_\ell) = \sum_{\ell=1}^{k} \alpha_\ell\, a(p, q_\ell)\,. \tag{4.11}$$

Für die Einheitsvektoren $e_j \in \Delta_n$, $j = 1, \dots, n$, ist $q = q_1\, e_1 + \cdots + q_m\, e_m$ und daher

$$a(p, q) = \sum_{j=1}^{m} q_j\, a(p, e_j) \geq \underbrace{\sum_{j=1}^{m} q_j}_{=1} \min_{k=1,\dots,n} a(p, e_k) = \min_{k=1,\dots,n} \left(p^T A\right)_k\,.$$

Da die rechte Seite unabängig von q ist, gilt die Abschätzung auch für das Minimum und es ist

$$\min_{q \in \Delta_m} a(p, q) \geq \min_{k=1,\dots,n} \left(p^T A\right)_k = \min_{k=1,\dots,n} a(p, e_k) \geq \min_{q \in \Delta_m} a(p, q)\,,$$

woraus (4.9) folgt. Der Beweis von (4.10) geht ganz analog[104]. □

So, jetzt geht es an die konvexe Analysis. Dazu erinnern wir uns zuerst daran, daß eine Menge $\Omega \subset \mathbb{R}^m$ als *konvex* bezeichnet wird, wenn

$$x, y \in \Omega \qquad \Leftrightarrow \qquad \alpha\, x + (1 - \alpha)\, y \in \Omega, \quad \alpha \in [0, 1]. \tag{4.12}$$

Aus (4.12) folgt dann auch, daß für jede konvexe Menge

$$x_1, \dots, x_n \in \Omega \qquad \Leftrightarrow \qquad \sum_{j=1}^{n} q_j\, x_j \in \Omega, \quad q \in \Delta_n \qquad \Leftrightarrow \qquad [x_1, \dots, x]\, \Delta_n \subseteq \Omega,$$

wobei $[x_1, \dots, x]$ die Matrix mit den Spaltenvektoren $x_1, \dots, x_n$ bezeichnet und

$$[\![x_1, \dots, x_n]\!] := \operatorname{conv}\,(x_1, \dots, x_n) := [x_1, \dots, x]\; \Delta_n$$

die *konvexe Hülle* von $x_1, \dots, x_n$. Eine *Hyperebene* $H \subset \mathbb{R}^m$ ist ein $m-1$–dimensionaler *affiner* Unterraum von $\mathbb{R}^m$, der als

$$H = \left\{x \in \mathbb{R}^m \;:\; v^T x + c = 0\right\}, \qquad v \in \mathbb{R}^m \setminus \{0\}, \quad c \in \mathbb{R} \tag{4.13}$$

gegeben ist. Konvexe Mengen lassen sich immer schön durch Hyperebenen abgrenzen, und das ist das nächste Resultat.

[104] Eine Tatsache, die aber niemanden davon abhalten sollte, diesen Teil des Beweises trotzdem mal übungshalber durchzuspielen!

Proposition 4.17 (Trennhyperebenensatz) *Ist $\Omega \subset \mathbb{R}^m$ abgeschlossen und konvex und $y \in \mathbb{R}^m \setminus \Omega$, dann gibt es $v \in \mathbb{R}^m$ und $c \in \mathbb{R}$, so daß*

$$v^T y + c < 0 < y^T \Omega + c := \left\{ v^T x + c \ : \ x \in \Omega \right\}. \tag{4.14}$$

Mit anderen Worten: y und Ω liegen auf unterschiedlichen Seiten *von H bzw. in unterschiedlichen von H induzierten Halbräumen[105], die zu v und c gehörige Hyperebene ist also ein* Trennhyperebene *für y und Ω.*

Beweis: Wir sammeln ein paar Beobachtungen über konvexe Mengen und Normen auf.

1. Die euklidische Norm $\|\cdot\|_2$ ist *strikt konvex*, d.h. für x, x' und $0 < \alpha < 1$ ist, unter Verwendung der guten alten Cauchy–Schwarz–Ungleichung, siehe[106] z.B. (Fischer, 1984, S. 190–191) oder (Horn & Johnson, 1985, S. 15)

$$\begin{aligned}
\left\|\alpha x + (1-\alpha)x'\right\|_2^2 &= \sum_{j=1}^{m} \left[\alpha^2 x_j^2 + 2\alpha(1-\alpha) x_j x_j' + (1-\alpha)^2 x_j'^2\right] \\
&\leq \alpha^2 \|x\|_2^2 + (1-\alpha)^2 \|x'\|_2^2 + 2\alpha(1-\alpha) \sum_{j=1}^{m} \left|x_j x_j'\right| \\
&\leq \alpha^2 \|x\|_2^2 + (1-\alpha)^2 \|x'\|_2^2 + 2\alpha(1-\alpha) \|x\|_2 \|x'\|_2 \\
&= \left(\alpha \|x\|_2 + (1-\alpha) \|x'\|_2\right)^2,
\end{aligned}$$

   also

$$\left\|\alpha x + (1-\alpha)x'\right\|_2 \leq \alpha \|x\|_2 + (1-\alpha) \|x'\|_2$$

   mit Gleichheit dann und nur dann, wenn $x = x'$ ist[107].

2. Für $y \in \mathbb{R}^m \setminus \Omega$ gibt es *genau ein* $x^* \in \Omega$, so daß[108]

$$\|y - x^*\|_2 < \|y - x\|_2, \qquad x \in \Omega \setminus \{x^*\}.$$

   Die Existenz eines $x \in \Omega$, so daß

$$\|y - x\|_2 = \min_{x' \in \Omega} \|y - x\|_2 \tag{4.15}$$

   folgt aus der Abgeschlossenheit von Ω, was interessant ist, ist die Eindeutigkeit! Gäbe es aber zwei verschiedene "Lösungen" $x_1 \neq x_2$ von (4.15), dann setzen wir $x := \frac{1}{2}x_1 + \frac{1}{2}x_2 \in \Omega$ und erhalten daß,

$$\begin{aligned}
\|y - x\|_2 &= \left\|y - \frac{1}{2}x_1 + \frac{1}{2}x_2\right\|_2 = \left\|\frac{1}{2}(y - x_1) + \frac{1}{2}(y - x_2)\right\|_2 \\
&< \frac{1}{2}\left(\|y - x_1\|_2 + \|y - x_2\|_2\right) = \min_{x' \in \Omega} \|y - x\|_2,
\end{aligned}$$

   und das kann ja nun wirklich nicht sein.

[105] Was natürlich deutlich vornehmer klingt.

[106] Hier sollen zwei Bücher angegeben werden, ein preiswertes und ein gutes Standardwerk.

[107] Dies folgt aus Cauchy–Schwarz!

[108] Das ist nun wieder ein Resultat, das sich auch der Approximationstheorie zuordnen lässt, die Grenzen sind also fließend.

3. Der *Bestapproximant* aus Teil 2 zeichnet sich dadurch aus, daß für alle $x \in \Omega$

$$\begin{aligned} 0 &< \|y-x\|_2^2 - \|y-x^*\|_2^2 \\ &= \|y\|_2^2 - 2y^T x + \|x\|_2^2 - \|y\|_2^2 + 2y^T x^* - \|x^*\|_2^2 \\ &= \|x\|_2^2 - \|x^*\|_2^2 + 2y^T (x^*-x) = (x+x^*)^T (x-x^*) - 2y^T (x-x^*) \\ &= [(x-y)+(x^*-y)]^T (x-x^*) \end{aligned}$$

Da Ω konvex ist, gilt das auch, wenn wir x durch die Konvexkombination $\alpha x + (1-\alpha)x^*$ ersetzen, $0 < \alpha < 1$, was dann

$$0 < [\alpha (x-y) + (2-\alpha)(x^*-y)]^T \alpha (x-x^*)$$

liefert. Dividieren wir diesen Ausdruck durch 2α und lassen dann $\alpha \to 0$ gehen, dann erhalten wir, daß

$$0 \le (x^*-y)^T (x-x^*), \qquad x \in \Omega, \tag{4.16}$$

sein muß. Diese Abschätzung bezeichnet man als *Kolmogoroff–Kriterium*[109] und sie *charakterisiert* sogar den Bestapproximanten, siehe z.B. (Sauer, 2002), aber auch (Neumann & Morgenstern, 1944, 16.3, S. 134–138).

So, wenn man all diese "bekannten" Fakten mal zur Verfügung hat, dann ist der eigentlich Beweis einfach, siehe Abb. 4.3: Zu $y \in \mathbb{R}^m \setminus \Omega$ bestimmen wir *den* Bestapproximanten aus Ω und sehen uns die affine Funktion

$$a(x) = v^T x + c', \qquad v = (x^*-y), \quad c' = -v^T x^*,$$

an, für die nach (4.16) die Ungleichungen

$$a(y) = -\|x^*-y\|_2^2 < 0 = a(x^*) \le a(x), \qquad x \in \Omega$$

gelten. Damit legen v und $c = c' - \frac{1}{2}a(y)$ die gesuchte Trennhyperebene fest. □

Schließlich noch eine Aussage über Matrizen, die auch im Kontext der Optimierung auftaucht[110], siehe z.B. (Spellucci, 1993).

Lemma 4.18 (Alternativensatz für Matrizen) *Zu jeder Matrix $A \in \mathbb{R}^{m\times n}$ gibt es entweder $x \in \Delta_m$, so daß*[111] *$x^T A > 0$ ist, oder ein $y \in \Delta_n$, so daß $Ay \le 0$ ist und diese beiden Möglichkeiten schließen einander aus.*

[109] Man muss natürlich fair sein und berücksichtigen, daß das Kolmogoroff–Kriterium erst 1948 in (Kolmogoroff, 1948) angegeben wurde – dafür gilt es aber auch nicht nur für endlichdimensionale Räume, sondern ebenfalls für Funktionenräume, insbesondere für Polynome. Die Sprechweise hat sich dann erst spter eingebürgert.

[110] Das ist nicht so verwunderlich, denn viele Argumente in der "theoretischen Optimierung" stammen tatsächlich aus der konvexen Analysis.

[111] Was wieder einmal komponentenweise zu verstehen ist: $x \ge y$ bedeutet $x_j \ge y_j$ für alle Indizes j.

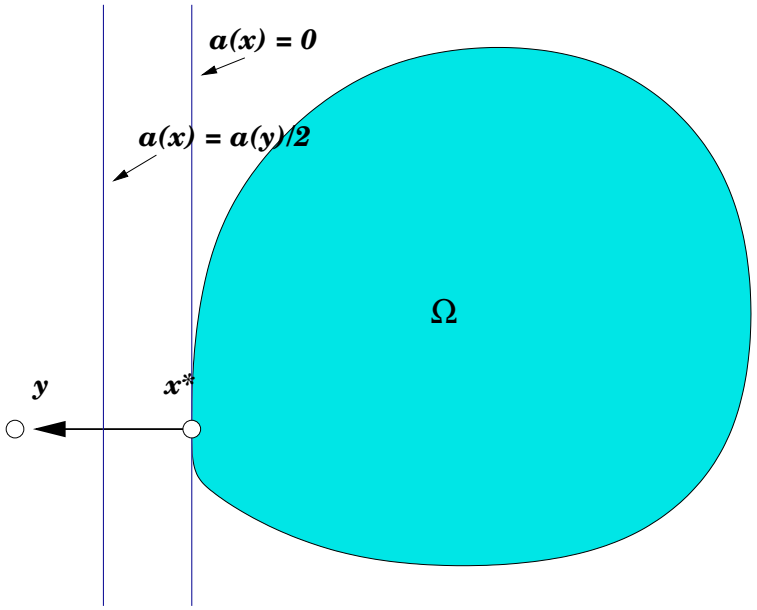

Abbildung 4.3: Konstruktion der Trennhyperebene: Zuerst findet man den Extremalpunkt, dann liefert (4.16) bereits, daß die durch x^* gehende Gerade, die senkrecht auf $y - x^*$ steht, eine "schwache" Trennfunktion hat: Mindestens ein Punkt von Ω, nämlich x^* liegt noch auf dieser Hyperebene – aber der Punkt ist auch eindeutig, wenn die Menge Ω *strikt konvex* ist. Schieben wir sie nun ein bißchen in Richtung y – der Wert $\frac{1}{2}$ war hier total willkürlich – dann sind beide Ungleichungen der Trennung so strikt wie in (4.14) gefordert.

Beweis: Unter Verwendung der *Spaltenvektoren* $a_1, \ldots, a_n \in \mathbb{R}^m$ der Matrix A, d.h. mit $A = [a_1, \ldots, a_n]$, betrachten wir die abgeschlossene konvexe Menge

$$\Omega := [\![a_1, \ldots, a_n, e_1, \ldots, e_m]\!] = [A \mid I]\ \Delta_{m+n} \subseteq \mathbb{R}^m.$$

Diese Menge enthält entweder den Nullpunkt oder sie tut es nicht!

1. Ist $0 \in \Omega$, dann gibt es $u \in \Delta_{m+n}$, so daß

$$0 = \sum_{j=1}^{n} u_j a_j + \sum_{j=1}^{m} u_{n+j} e_j \qquad \Rightarrow \qquad 0 \neq \widehat{u} = (u_j \ : \ j = 1, \ldots, n)\,. \quad (4.17)$$

Da $0 \leq \widehat{u}$ und $\widehat{u} \neq \mathbf{0}$ ist, folgt also auch $0 < u := u_1 + \cdots + u_n$ und mit $y := \widehat{u}/u$ sowie (4.17) erhalten wir, daß

$$Ay = \sum_{j=1}^{n} y_j a_j = \frac{1}{u} \sum_{j=1}^{n} u_j a_j = -\frac{1}{u} \sum_{j=1}^{m} u_{n+j} e_j \leq 0$$

ist, was gerade den zweiten Teil unserer Behauptung darstellt.

2. Ist hingegen $0 \notin \Omega$, dann gibt es nach Proposition 4.17 einen Vektor $v \in \mathbb{R}^m$ und $c \in \mathbb{R}$, so daß

$$c = v^T 0 + c < 0 < v^T a + c, \qquad a = [A \mid I]\, u \in \Omega.$$

Damit gilt für alle $u \in \Delta_{m+n}$ die Ungleichung $0 < v^T [A \mid I]\, u$ und wir erhalten insbesondere für $u = e_j$, $j = 1, \ldots, m+n$, daß

$$0 < v^T [A \mid I]\, e_j = \left[v^T A \mid v^T\right] e_j = \begin{cases} (y^T A)_j\,, & j = 1, \ldots, n \\ v_{j-n}, & j = n+1, \ldots, n+m, \end{cases}$$

also ist $v^T A > 0$ wie auch $v > 0$ und somit ist $x := v/|v| \in \Delta_m$ in diesem Fall der gesuchte Alternativvektor.

Daß sich die beiden Alternativen ausschließen, das sieht man sehr einfach: Gäbe es nämlich x und y, dann erhielten wir

$$0 < x^T A \quad \Rightarrow \quad 0 < x^T A y \qquad \text{und} \qquad Ay \le 0 \Rightarrow \quad x^T A y \le 0,$$

was einen soliden Widerspruch darstellen würde: $0 < 0$! □

So, damit haben wir alle Bausteine beisammen, die wir brauchen, um das Minimax–Theorem zu beweisen, also wollen wir das auch tun.
Beweis von Satz 4.11: Wegen (4.9) ist

$$v_1 = \max_{p \in \Delta_m} \min_{q \in \Delta_n} a(p, q) = \max_{p \in \Delta_m} \min_{k=1,\ldots,n} \left(p^T A\right)_k$$

und nach (4.10) entsprechend

$$v_2 = \min_{q \in \Delta_n} \max_{p \in \Delta_m} a(p, q) = \min_{q \in \Delta_n} \max_{j=1,\ldots,m} (Aq)_j$$

mit $v_1 \le v_2$. Wäre nun $v_1 < v_2$, dann können wir A durch $A' := A - \frac{1}{2}(v_2 - v_1)\, 1 1^T$ ersetzen, was uns die Auszahlungsfunktion

$$a'(p, q) = a(p, q) + \frac{v_1 - v_2}{2} \underbrace{p^T 1}_{=1} \underbrace{1^T q}_{=1} = a(p, q) + \frac{v_1 - v_2}{2}$$

liefert, die ihre Minimaxe immer noch an derselben Stelle wir a hat, für die aber nun $v_1' < 0 < v_2'$ gilt. Doch das kann nicht sein! Denn nach Lemma 4.18 gibt es entweder ein $x \in \Delta_m$ mit $x^T A' > 0$, also

$$v_1' = \max_{p \in \Delta_m} \min_{k=1,\ldots,n} \left(p^T A'\right)_k \ge \min_{k=1,\ldots,n} \left(x^T A'\right)_k > 0,$$

oder aber ein $y \in \Delta_n$ mit

$$v_2' = \min_{q \in \Delta_n} \max_{j=1,\ldots,m} (A'q)_j \le \max_{j=1,\ldots,m} (A'y)_j \le 0,$$

aber keinesfalls kann $v_1' < 0 < v_2'$ sein. □

4.5 Struktur und Berechnung der Optimallösungen

Als erstes leiten wir ein paar Folgerungen aus dem Minimax–Theorem 4.11 her, die uns helfen werden, die Struktur der Optimalstrategien besser zu verstehen.

Korollar 4.19 *Sei $A \in \mathbb{R}^{m\times n}$ eine Auszahlungsmatrix zu einem Spiel mit Wert ν. Dann sind $p \in \Delta_m$ und $q \in \Delta_n$ jeweils genau dann optimale Strategien, wenn*[112]

$$p^T A \geq \nu 1^T, \qquad \text{bzw.} \qquad Aq \leq \nu 1 \tag{4.18}$$

ist.

Beweis: Nach Lemma 4.16 ist für jede Optimalstrategie p

$$\nu = \max_{p'\in\Delta_m} \min_{q\in\Delta_n} a(p', q) = \min_{k=1,\dots,n} \left(p^T A\right)_k \qquad \Rightarrow \qquad \left(p^T A\right)_k \geq \nu,$$

und somit $p^T A \geq \nu 1^T$, und die zweite Ungleichung folgt ganz analog.

Gilt umgekehrt $p^T A \geq \nu 1^T$, dann ist

$$\min_{q\in\Delta_n} \max_{p'\in\Delta_m} a(p', q) \geq \min_{q\in\Delta_n} p^T A q \geq \min_{q\in\Delta_n} \nu \underbrace{1^T q}_{=1} = \nu$$

und p ist eine Optimalstrategie für Spieler 1, denn ganz egal, wie Spieler 2 seine gemischte Strategie wählt ist die erwartete Auszahlung mindestens ν. Analog liefert $Aq \leq \nu 1^T$, daß

$$\max_{p\in\Delta_m} \min_{q'\in\Delta_n} a(p, q') \leq \max_{p\in\Delta_m} p^T A q \leq \max_{p\in\Delta_m} \nu \underbrace{p^T 1}_{=1} = \nu,$$

und jetzt hängt Spieler 1 unterhalb von ν fest. □

Und wieder taucht die ubiquitäre Konvexität auf.

Korollar 4.20 *Die Mengen $\mathscr{P}^* \subseteq \Delta_m$ und $\mathscr{Q}^* \subseteq \Delta_n$ der Optimalstrategien für Spieler 1 bzw. Spieler 2 sind konvex.*

Beweis: Sind $p_1, \dots, p_k \in \mathscr{P}^*$ optimale Strategien und $\alpha \in \Delta_k$, dann setzen wir $p = \alpha_1 p_1 + \cdots + \alpha_k p_k$, erhalten dank Korollar 4.19, daß

$$p^T A = \left(\sum_{j=1}^{k} \alpha_j p_k\right)^T A = \sum_{j=1}^{k} \alpha_j p_k^T A \geq \sum_{j=1}^{k} \alpha_j \left(\nu 1^T\right) = \nu 1^T,$$

und wiederum Korollar 4.19 sagt uns, daß $p \in \mathscr{P}^*$ ist. Einen expliziten Beweis für $\mathscr{Q}^*$ erwartet hoffentlich niemand. □

Was sind nun eigentlich unsere Unbekannten im Minimax–Theorem 4.11? Es sind ja nicht nur die magischen Optimalstrategien p^* und q^*, sondern auch noch der Wert ν des Spiels! Würden wir den Wert kennen, dann müssten wir

[112] Eine kleine Warnung: Die 1–Vektoren, die in (4.18) auftauchen, haben normalerweise unterschiedliche Länge, nämlich n bzw. m.

nur die Ungleichungssysteme $A^T p^* \geq \nu 1$ bzw. $-Aq^* \geq -\nu 1$ lösen[113], um an Optimalstrategien zu kommen. Nun ist aber ν ebenfalls unbekannt, also behandeln wir es wie eine anständige Unbekannte und erhalten, daß p^*, q^* und ν das Ungleichungssystem

$$\begin{aligned} A^T p^* - 1\nu &\geq 0, \\ -Aq^* + 1\nu &\geq 0, \end{aligned}$$

mit $p^* \in \Delta_m$ und $q^* \in \Delta_n$ lösen müssen. Codieren wir die Bedingung $1^T p^* = 1$ in $1^T p^* \geq 1$ und $-1^T p^* \geq -1$, dann suchen wir „nur" noch nach $\nu \in \mathbb{R}$, $p^* \in \mathbb{R}^m$ und $q^* \in \mathbb{R}^n$, die das **Ungleichungssystem**

$$\begin{aligned} A^T p^* - 1\nu &\geq 0, \\ -Aq^* + 1\nu &\geq 0, \\ 1^T p^* &\geq 1, \\ -1^T p^* &\geq -1, \\ 1^T q^* &\geq 1, \\ -1^T q^* &\geq -1, \\ p^* &\geq 0, \\ q^* &\geq 0, \end{aligned}$$

in Matrixschreibweise

$$Bx := \left[\begin{array}{c|c|c} A^T & & -1 \\ & -A & 1 \\ \hline 1^T & & \\ -1^T & & \\ \hline & 1^T & \\ & -1^T & \\ \hline I & & \\ & I & \end{array}\right] \left[\begin{array}{c} p^* \\ q^* \\ \nu \end{array}\right] \geq \left[\begin{array}{c} 0 \\ 0 \\ \hline 1 \\ -1 \\ \hline 1 \\ -1 \\ \hline 0 \\ 0 \end{array}\right], \tag{4.19}$$

erfüllt. Jede Lösung dieses Ungleichungssystems ist eine optimale Strategie und die numerische Bestimmung von optimalen Strategien und damit auch des Wertes eines Spiels besteht also in der numerischen Lösung des Ungleichungssystems. Allerdings ist das Ungleichungssystem hochgradig *überbestimmt*: Den $m + n + 1$ Variablen p^*, q^* und ν stehen insgesamt $2m + 2n + 4$ Ungleichungen gegenüber.

Im Falle eines symmetrischen Spiels ist das alles viel einfacher, denn da ist $\nu = 0$ und die Rollen von p und q vollkommen vertauschbar, so daß sich das Ungleichungssystem auf die viel einfachere Form

$$Bx := \left[\begin{array}{c} A^T \\ 1^T \\ -1^T \\ I \end{array}\right] p^* \geq \left[\begin{array}{c} 0 \\ 1 \\ -1 \\ 0 \end{array}\right] \tag{4.20}$$

[113] Beziehungsweise Halbräume schneiden.

reduziert; die optimale Strategie für Spieler 2 ist dann natürlich $q^* = p^*$.

Aus unseren bisherigen Überlegungen erhalten wir so eine schöne Charakterisierung der Optimallösungen für symmetrische[114] Spiele.

Korollar 4.21 *Eine gemischte Strategie $p \in \Delta_m$ ist genau dann optimal für ein symmetrisches Spiel*[115]*, wenn $A^T p \geq 0$ ist.*

Beispiel 4.22 (Stein, Schere, Papier, mit oder ohne Brunnen) *Die Optimalität der Strategien $p = \left[\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right]^T$ bzw. $p' = \left[0, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right]^T$ für die Variante mit Brunnen sieht man dank Korollar 4.21 nun sofort aus*

$$A^T p = \begin{bmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix} = 0,$$

und

$$A'^T p' = \begin{bmatrix} 0 & -1 & 1 & 1 \\ 1 & 0 & -1 & 1 \\ -1 & 1 & 0 & -1 \\ -1 & -1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1/3 \\ 1/3 \\ 1/3 \end{bmatrix} = \begin{bmatrix} 1/3 \\ 0 \\ 0 \\ 0 \end{bmatrix} \geq 0. \tag{4.21}$$

Der positive Wert $\frac{1}{3}$ in $p'^T A$ bei (4.21) zeigt uns, daß diese Optimalstrategie sogar einen Vorteil gegen einen „unwissenden" Spieler bringt! Wählt Spieler 2 nämlich eine Strategie q „mit Stein", d.h. $q_1 > 0$, dann ist die erwartete Auszahlung aus Sicht von Spieler 1

$$p'^T A q = \left[\frac{1}{3}, 0, 0, 0\right] q = \frac{q_1}{3}$$

und Spieler 2 zahlt auf lange Sicht tatsächlich drauf[116]*.*

Ganz befriedigend ist Korollar 4.21 allerdings noch nicht, denn wir konnten in Beispiel 4.22 ja gar keine Optimalstrategien berechnen, sondern haben nur gezeigt, daß gut geratene Lösungen[117] tatsächlich optimal sind. Der allgemeine Fall ist bei systematischem Zugang eher erschreckend.

Beispiel 4.23 *Für „Stein, Schere, Papier, Brunnen" wird die Optimallösung durch das*

[114] Und damit auch faire!

[115] Zu einer Matrix $A = -A^T$

[116] Was vielleicht ein erster praktischer Lerneffekt dieser Vorlesung sein könnte.

[117] Der Fachmann spricht hier von einem „educated guess", die Fachfrau übrigens auch.

allgemeine Ungleichungssystem

$$\begin{bmatrix} 0 & -1 & 1 & 1 & 0 & 0 & 0 & 0 & -1 \\ 1 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & -1 \\ -1 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & -1 \\ -1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & -1 & 1 & 1 \\ 0 & 0 & 0 & 0 & -1 & 1 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & -1 & -1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ -1 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} p^* \\ q^* \\ v \end{bmatrix} \geq \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ -1 \\ 1 \\ -1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

festgelegt, das man nicht wirklich gerne von Hand lösen möchte.

Erstaunlicherweise haben wir für einen systematischen Lösungsansatz bereits alle Bausteine beisammen, denn wir können (4.19) mit -1 multiplizieren und erhalten dann ein Ungleichungssystem der Form $Ax \leq b$, wobei b positive *und* negative Einträge hat, ganz genau wie im Transportproblem. Und genau wie dort müssen wir nun einen zulässigen Punkt finden, denn ein solcher ist nichts anderes als eine Lösung unseres Ungleichungssystems. Und berechnen können wir ihn mit der Phase I der **Zweiphasenmethode** des Simplexalgorithmus, wobei wir die zweite Phase gar nicht mehr brauchen.

Zur Bestimmung der Optimalstrategie gibt es ein einfaches und recht überschaubares `Octave`–Programm[118] `GameOptStrat.m`, siehe Programm 4.1, das zu einer vorgegebenen Auszahlungsmatrix die Optimalstrategien für beide Spieler und den Wert des Spieles bestimmt.

Beispiel 4.24 (Stein, Schere, Papier, Brunnen) *Für unser Lieblingsspiel „Stein, Schere, Papier, Brunnen" erhalten wir den folgenden Ablauf*

```
octave> A = [ 0 1 -1 -1; -1 0 1 -1; 1 -1 0 1; 1 1 -1 0 ];
octave> [p,q,v] = GameOptStrat ( A )
p =

  0.00000
  0.33333
  0.33333
  0.33333
```

[118] Da herunterzuladen, wo es auch dieses Skript gibt.

```
##  GameOptStrat.m (Spieltheorie)
##  -----------------------------------------------------
##  Optimale gemischte Strategien
##  Eingabe:
##    X    Auszahlungsmatrix des Spiels
##  Ausgabe:
##    p    Strategie fuer Spieler 1
##    q    Strategie fuer Spieler 2
##    v    Wert des Spiels

function [p,q,v] = GameOptStrat( X )
  [m,n] = size( X );

  ## Setup und Phase I
  [A,b] = GameStratBed( X );c = zeros( m+n+1,1 );
  T = SimPhase1( A,b,c );

  ## Extrahiere p,q,v
  p = zeros( m,1 ); q = zeros( n,1 ); v = 0;
  for j = 2:m+n+5
    k = T( j,1 );
    if k <= 0
      continue;
    elseif k <= m
      p( k ) = T( j,m+n+3 );
    elseif k <= m+n
      q( k-m ) = T( j,m+n+3 );
    else
      v = T( j,m+n+3 );
    end
  end
```

Programm 4.1 `GameOptStrat.m`: Berechnung der optimalen Strategien und des Wertes eines Spiels mit Hilfe des Simplexalgorithmus. Die hier aufgerufene Funktion `GameStratBed` bildet auf systematische Weise die Matrix aus (4.19). Ansonsten wird nur die Phase 1 des Simplexalgorithmus verwendet. Die Hauptarbeit besteht wirklich darin, in dem Variablenwust p, q und v zu finden.

```
q =

  0.00000
  0.33333
  0.33333
  0.33333

v = 0
```

und das ist genau das, was wir inzwischen von den optimalen Strategien und dem Wert dieses Spiels erwarten können.

Beispiel 4.25 (Daiquiri–Spiel) *Erinnern wir uns an das Daiquiri–Spiel aus (Williams, 1986), Übung 4.2*[119] *mit der Auszahlungsmatrix*

$$A = \begin{bmatrix} 5.5 & 1 \\ 1 & 11 \end{bmatrix}$$

und sehen wir uns an, was Optimalstrategien und Wert dieses Spiels sind, nämlich

```
octave> A = [ 5.5 1 ; 1 11 ];
octave> [p,q,v] = GameOptStrat ( A )
p =

  0.68966
  0.31034

q =

  0.68966
  0.31034

v = 4.1034
```

und das ist auch genau das, was wir in (Williams, 1986) finden und was wir für dieses 2×2–Spiel auch „zu Fuß" hätten ausrechnen können. Etwas verblüffender wird das Ganze aber, wenn wir uns die Sache aus der Sicht von Olaf ansehen, also die optimale Strategie für $-A^T$ berechenen (lassen), denn dann erhalten wir plötzlich

```
octave> [p,q,v] = GameOptStrat ( -A' )
p =

  -0.22222
   1.22222

q =
```

[119] Hat jemand sich mit dieser Übung beschäftigt?

```
  1
  0

v = 0
```

was ja nun absolut keinen Sinn ergibt! Ist unser Programm defekt?

Beispiel 4.25 zeigt uns, daß wir vor lauter Begeisterung über die Zweiphasenmethode beinahe ein wichtiges Detail übersehen hätten[120]: Die zulässigen Punkte, die wir so bestimmen liegen im *positiven Orthanten*, das ist in den Simplexalgorithmus, auch in seine erste Phase, eingebaut. Mit anderen Worten, es muß

$$\begin{bmatrix} p^* \\ q^* \\ \nu \end{bmatrix} \geq 0 \tag{4.22}$$

sein. Das ist unproblematisch, solange $\nu \geq 0$ ist[121], bricht allerdings zusammen, wenn $\nu < 0$ ist, denn in diesem Fall kann die Phase 1 den Optimalpunkt gar nicht finden.

Beispiel 4.26 *Daß das Spiel zur Matrix*

$$A = \begin{bmatrix} 5.5 & -1 \\ -1 & 11 \end{bmatrix}$$

nicht ganz fair ist, sondern einen positiven Wert, nämlich 3.2162*, hat ist vielleicht nicht so schwer nachzuvollziegen, aber nun liefert uns die* `Octave`*–Routine das Ergebnis*

```
octave> A = [ 5.5 -1; -1 11 ]; [p,q,v] = GameOptStrat ( -A' )
p =

  0.15385
  0.84615

q =

  1
  0

v = 0
```

die auf den ersten Blick recht harmlos und korrekt aussieht!

Allerdings ist das Problem recht leicht gelöst: Wir berechnen einfach die optimalen Strategien zu A *und* $-A^T$. Hat eines dieser beiden Spiele positiven Wert, so müsste das andere negativen Wert haben, was zu einer nicht korrekten Lösung mit Wert Null führt, die wir dann halt verwerfen. Das `Octave`–Programm hierzu findet sich in Programm 4.2.

[120] Selbstverständlich ist es nicht übersehen worden, sondern dieser Aufbau wurde bewusst und gezielt gewählt, um dezidiert einen besonderen didaktischen Spannungsbogen aufzubauen!

[121] Also insbesondere für alle *fairen* Spiele!

```
##  GameSolve.m (Spieltheorie)
##  ----------------------------------------------------------
##  Optimale gemischte Strategien
##  Eingabe:
##    A    Auszahlungsmatrix des Spiels
##  Ausgabe:
##    p    Strategie fuer Spieler 1
##    q    Strategie fuer Spieler 2
##    v    Wert des Spiels

function [p,q,v] = GameSolve( A )
  [p1,q1,v1] = GameOptStrat( A );
  [q2,p2,v2] = GameOptStrat( -A' );

  if ( v2 > 0 )
    p = p2; q = q2; v = v2;
  else
    p = p1; q = q1; v = v1;
  end
```

Programm 4.2 `GameSolve.m`: Berechnung der optimalen Strategien – unter Verwendung von `GameOptStrat.m` eine ganz einfache Geschichte.

Damit sind wir aber auch in der Lage, *alle* optimalen Strategien eines Spielers zu bestimmen, indem wir zuerst mit `GameSolve` *eine* optimale Strategie für Spieler 1, eine für Spieler 2[122] und vor allem den Wert des Spieles bestimmen! Kennen wir einmal den Wert des Spieles, dann sind ja nach Korollar 4.19 die Optimalstrategien von Spieler 1 gerade die Lösungen des Ungleichungssystems

$$\begin{bmatrix} -A^T \\ 1^T \\ -1^T \end{bmatrix} p \le \begin{bmatrix} -v1 \\ 1 \\ -1 \end{bmatrix}, \qquad p \ge 0,$$

das genau die richtige Form hat, um als Nebenbedingung eines linearen Optimierungsproblems aufgefasst zu werden. Außerdem kennen wir mit p^* bereits eine Ecke des zugehörigen Polyeders Ω, die man in einem ersten Schritt mittels Austauschschritten in den Nullpunkt transformiert. In der Praxis würde man das nicht so machen, sondern einfach mit dem Ergebnis von Phase 1 weiterrechnen, indem man die Zeilen und Spalten streicht, die zu q^* und v gehören. Danach kann man mit dem Simplexalgorithmus, bzw. dem **Austauschschritt** systematisch *alle* Ecken des zulässigen Bereichs aufsuchen. Und hat man einmal alle Ecken, dann hat man auch alle Lösungen[123] ...

Beispiel 4.27 (Das „Skin Game“ aus Beispiel 4.13) *Mit dieser Methodik können*

[122] Die werden wir ihm natürlich nicht verraten, er soll gefälligst selbst draufkommen
[123] Siehe Proposition 2.7.

wir auch das „Skin Game“ mit der Auszahlungsmatrix

$$A = \begin{bmatrix} 1 & -1 & -2 \\ -1 & 1 & 1 \\ 2 & -1 & 0 \end{bmatrix}$$

angehen und erhalten

```
octave> [p,q,v] = GameSolve( [ 1 -1 -2; -1 1 1; 2 -1 0 ] )
p =

  0.00000
  0.60000
  0.40000

q =

  0.40000
  0.60000
  0.00000

v = 0.20000
```

das heißt, die optimale Strategie von Spieler 1 ist $\left(0, \frac{3}{5}, \frac{2}{5}\right)$ *und bringt ihm einen erwarteten Gewinn von* $\frac{1}{5}$ *pro Runde! Fair ist offenbar etwas anderes.*

Wir beenden dieses Kapitel mit einem weiteren Bei-Spiel, in dem man die Optimalstrategie nicht so einfach sieht und ohne unsere `Octave`–Prográmmchen auch ganz schön arbeiten muß, oder, wie in (Williams, 1986, S. 164) zu lesen ist:

> *A flash of genius is a useful thing at this point, because straight calculation is wretched.*

Das Spiel selbst wird übrigens nicht nur in (Williams, 1986) diskutiert, sondern auch in den „seriösen“, mathematisch substantiellen Büchern (Karlin, 1959; Neumann & Morgenstern, 1944).

Beispiel 4.28 (Morra) *Jeder Spieler streckt (verdeckt) einen, zwei oder drei Finger aus und rät gleichzeitig, wieviele Finger sein Gegner ausstreckt*[124]*. Rät* ein *Spieler richtig, so wird ihm die* Gesamtzahl *an angezeigten Fingern ausbezahlt*[125]*, andernfalls endet das Spiel unentschieden, was insbesondere der Fall ist, wenn* beide *Spieler richtig raten, ganz egal, wer mehr Finger angezeigt hat.*

Bei Morra ist also ein Strategie ein Paar $(a, r) \in \{1, 2, 3\}^2$ und die Auszahlungstabelle hat die Form

[124] Nachdem die meisten Leute zwei Hände haben, kann man die eine zum Anzeigen, die andere zum Raten verwenden. Alternativ könnte man die Zahlen auf ein Blatt Papier schreiben oder mit zwei Würfeln „einstellen“.

[125] Natürlich nicht in Fingern, sonst ist das ein sehr kurzlebiges, wenn auch vielleicht kurzweiliges Spiel.

	(1,1)	(1,2)	(1,3)	(2,1)	(2,2)	(2,3)	(3,1)	(3,2)	(3,3)
(1,1)	0	2	2	-3	0	0	-4	0	0
(1,2)	-2	0	0	0	3	3	-4	0	0
(1,3)	-2	0	0	-3	0	0	0	4	4
(2,1)	3	0	3	0	-4	0	0	-5	0
(2,2)	0	-3	0	4	0	4	0	-5	0
(2,3)	0	-3	0	0	-4	0	5	0	5
(3,1)	4	4	0	0	0	-5	0	0	-6
(3,2)	0	0	-4	5	5	0	0	0	-6
(3,3)	0	0	-4	0	0	-5	6	6	0

Was man schön sieht, ist die Tatsache, daß das Zeigen vieler Finger den potentiellen Gewinn, aber auch den potentiellen Verlust vergrößert, daß die Spieler so also das Risiko kontrollieren können. Und siehe da – die Bestimmung einer Optimalstrategie ist jetzt ein Klacks, denn mit Hilfe einer kleinen Funktion `MorraMat` zur automatischen Bestimmung der Auszahlungsmatrix erhalten wir

```
octave> [p,q,v] = GameSolve ( MorraMat( 3 ) )
p =

  0.00000
  0.00000
  0.42553
  0.00000
  0.31915
  0.00000
  0.25532
  0.00000
  0.00000

q =

  -0.00000
   0.00000
   0.42553
   0.00000
   0.31915
  -0.00000
   0.25532
  -0.00000
   0.00000

v = 0
```

und das etwas verblüffenden Resultat, daß die Optimalstrategie nur die drei Strategien $(1,3)$, $(2,2)$ und $(3,1)$ verwendet, bei denen vier Finger auf dem Spiel stehen. Und jetzt haben wir auch ein Spiel, bei dem die Optimalstrategie *nicht*

eindeutig ist, denn die Lösung mit den von Null verschiedenen Wahrscheinlichkeiten $\frac{5}{12}$, $\frac{1}{3}$, $\frac{1}{4}$ tut's ganz genauso:

```
octave> p2 = [ 0 0 5 0 4 0 3 0 0 ]' / 12
p2 =

  0.00000
  0.00000
  0.41667
  0.00000
  0.33333
  0.00000
  0.25000
  0.00000
  0.00000

octave> A'*p2
ans =

  0.16667
  0.00000
  0.00000
  0.08333
  0.00000
  0.08333
  0.00000
  0.00000
  0.16667
```

Dieser Vektor ist $\geq \mathbf{0}$ und da Morra ein faires Spiel ist[126] ist auch das eine Optimalstrategie für Spieler 1, siehe Korollar 4.19. Auffällig ist auch, daß die Strategie, die mit dem größten Risiko verbunden ist, mit der geringsten Wahrscheinlichkeit gespielt wird – Feigheit scheint also ein durchaus rationales Verhalten zu sein.

Übung 4.5 Schreiben Sie ein `Octave`–Programm, das *alle* optimalen Strategien eines gegebenen Spiels ermittelt, indem es alle Ecken des zugehörigen konvexen Polyeders bestimmt[127]. ◊

Übung 4.6 Bestimmen Sie alle Optimalstrategien für Morra. ◊

Übung 4.7 Bestimmen Sie alle Optimalstrategien für das Fünf–Finger–Morra. ◊

[126] Dazu hätten wir nicht erst $\nu = 0$ errechnen lassen müssen, jede Morra–Matrix ist schiefsymmetrisch!

[127] Und schicken Sie den Code an mich.

Du glaubst, ich laufe dem Sonderbaren nach, weil ich das Schöne nicht kenne, nein, weil du das Schöne nicht kennst, deswegen suche ich das Sonderbare.

G. Chr. Lichtenberg

5 Lineare Optimierung ganz anders

Einerseits ist die Strategie des Simplex–Algorithmus ja ganz logisch: Nachdem das Maximum in einer Ecke angenommen wird, klappern wir diese eben so lange ab, bis wir an der optimalen Ecke angekommen sind. Andererseits hat das auch einen Nachteil, denn auf diese Art und Weise marschiert man ja immer "außen" am zulässigen Bereich entlang, und erreicht so sein Ziel daher immer auf einem Umweg, denn der "direkte Weg" würde normalerweise quer durch das konvexe Polyeder führen. Deshalb interessiert man sich inzwischen für Verfahren, die auf der Suche nach dem Extremum den zulässigen Bereich durchqueren – solche Verfahren bezeichnet man dann als **innere–Punkte–Methoden** und sie zeichnen sich dadurch aus, daß sie ausgehend von einem inneren Punkt des zulässigen Bereichs[128] eine Folge von inneren Punkten konstruieren, die gegen die Extremalecke konvergiert.

Für dieses Kapitel nehmen wir an, daß ein **lineares Optimierungsproblem** stets in der Normalform

$$\min c^T x, \qquad Ax = b, \quad x \geq 0, \qquad x \in \mathbb{R}^n, \tag{5.1}$$

vorliegt, und zwar mit $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$. Das kann man ja durch Einführung von **Schlupfvariablen** immer erreichen; die Idee dabei ist wieder einmal sehr einfach und basiert auf der fast trivialen Äquivalenz

$$a^T x \geq b \qquad \Leftrightarrow \qquad a^T x = b + t, \quad t \geq 0.$$

Damit erhalten wir

$$Ax \geq b, \quad x \geq 0 \qquad \Leftrightarrow \qquad Ax = b + y, \quad x, y \geq 0,$$

oder eben

$$[A - I]\begin{bmatrix} x \\ y \end{bmatrix} = b, \quad \begin{bmatrix} x \\ y \end{bmatrix} \geq 0, \tag{5.2}$$

und die hinzugefügten Variablen y bezeichnet man als **Schlupfvariablen**. Die Normalform (5.1) ist natürlich nur dann sinnvoll, wenn der Rang von A kleiner als n ist, also wird vernünftigerweise $m \leq n$ sein.

[128] Also gerade *keinem* Randpunkt und insbesondere keiner Ecke!

Um die Verfahren herleiten und verstehen zu können brauchen wir noch ein wenig mehr Theorie, aber das kann ja bekanntlich ohnehin nicht schaden.

5.1 Dualität

Die erste Beobachtung ist, daß dem **Minimierungsproblem** (5.1) immer ein **Maximierungsproblem**, das sogenannte duale Problem, zugeordnet ist, und zwar so, daß die Optimallösungen der beiden Probleme übereinstimmen.

Definition 5.1 *Das **duale Problem** zu (5.1) ist definiert als die Optimierungsaufgabe*

$$\max\ b^T y, \qquad A^T y \le c, \qquad y \in \mathbb{R}^m. \tag{5.3}$$

Dabei bezeichnen b *und* c *in (5.1) und (5.3) tatsächlich* dieselben *Vektoren.*

Bemerkung 5.2 *Mit* $x \ge 0$ *und* $c \ge A^T y$ *ist*

$$c^T x \ge \left(A^T y\right)^T x = y^T A x = y^T b,$$

also ist auch

$$\min_{x \in F_x} c^T x =: c^T x^* \ge b^T y^* := \max_{y \in F_y} b^T y,$$

wobei die zulässigen Bereiche

$$F_x = \{x \in \mathbb{R}^n \ :\ Ax = b,\ x \ge 0\} \qquad \textit{und} \qquad F_y = \left\{y \in \mathbb{R}^m \ :\ A^T y \le c\right\}$$

beide *nichtleer sein sollen. Für* $x \in F_x$ *und* $y \in F_y$ *bezeichnen wir die Größe*

$$0 \le g(x, y) := c^T x - b^T y$$

*als **Dualitätslücke**[129] zwischen* x *und* y*.*

Satz 5.3 (Starker Dualitätssatz) *Sind* $x^* \in F_x \ne \emptyset$ *und* $y^* \in F_y \ne \emptyset$ *Optimallösungen von (5.1) und (5.3), dann ist*

$$c^T x^* = b^T y^*. \tag{5.4}$$

Mit anderen Worten: Bei den Optimallösungen gibt es keine Dualitätslücke.

Bemerkung 5.4 *Eigentlich ist Satz 5.3 nur die "schwache" Version der starken Dualität. Es gilt nämlich außerdem, daß* $F_x = \emptyset$ *genau dann der Fall ist, wenn entweder* $F_y = \emptyset$ *oder* $b^T y$ *auf* $F_y \ne \emptyset$ *nach oben unbeschränkt ist. Umgekehrt bedeutet auch* $F_y = \emptyset$*, daß entweder* $F_x = \emptyset$ *oder* $c^T x$ *nach unten unbeschränkt ist.*
Diese Bemerkung ist durchaus nützlich: Kann man nämlich dem dualen Problem ansehen, daß es keine zulässigen Punkte enthält, dann weiß man auch, daß man sich mit dem primalen Problem gar nicht abzugeben braucht, denn eine Optimallösung kann es ja nicht geben.

[129] *"Duality gap".*

Beweis von Satz 5.3: Sei die Optimallösung x^* eine *nichtentartete*[130] Ecke von F_x, d.h., es gibt

$$J \subset \{1,\ldots,n\}, \qquad \#J = m, \qquad K := \{1,\ldots,n\} \setminus J,$$

so daß die Matrix A'_J, die von den durch J indizierten *Spalten*[131] von A gebildet wird, invertierbar ist und daß

$$A'_J x^*_J = b \quad \Rightarrow \quad x^*_J = \left(A'_J\right)^{-1} b, \qquad x^*_K = 0$$

Da, für beliebiges $x \in F_x$,

$$Ax = A'_J x_J + A'_K x_K = \left[A'_J\, A'_K\right] \begin{bmatrix} x_J \\ x_K \end{bmatrix} = b \qquad \Rightarrow \qquad x_J = \left(A'_J\right)^{-1} (b - A'_K x_K),$$

ist

$$\begin{aligned} c^T x &= c_J^T x_J + c_K^T x_K = c_J^T \underbrace{\left(A'_J\right)^{-1} b}_{=x^*_J} + \underbrace{\left(c_K - (A'_K)^T \left(A'_J\right)^{-T} c_J\right)^T}_{d_K^T} x_K \\ &= c_J^T x^*_J + c_K^T \underbrace{x^*_K}_{=0} + d_K^T x_K = c^T x^* + d_K^T x_K, \end{aligned}$$

daher ist $d_K^T x_K = c^T x - c^T x^* \geq 0$ und da das für alle $x_K \geq 0$ gelten muß, ist auch $d_K \geq 0$.

Nun setzen wir

$$y^* = \left(A'_J\right)^{-T} c_J$$

und erhalten, daß

$$A^T y^* = \begin{bmatrix} \left(A'_J\right)^T \\ \left(A'_K\right)^T \end{bmatrix} \left(A'_J\right)^{-T} c_J = \begin{bmatrix} c_J \\ \left(A'_K\right)^T \left(A'_J\right)^{-T} c_J \end{bmatrix} = \begin{bmatrix} c_J \\ c_K - d_K \end{bmatrix} \leq \begin{bmatrix} c_J \\ c_K \end{bmatrix} = c,$$

also ist y^* eine **zulässige Ecke**[132] mit der Eigenschaft

$$b^T y^* = b^T \left(A'_J\right)^{-T} c_J = c_J^T \underbrace{\left(A'_J\right)^{-1} b}_{=x^*_J} = c_J^T x^*_J = c^T x^*,$$

was gerade zu beweisen war. □

[130] Durch beliebig kleine Störungen von A oder auch von b kann man immer erreichen, daß *alle* Ecken nichtentartet sind, was den Beweis nur um ein ε–Argument bereichern würde: Man zeigt, daß die Aussage für alle hinreichend kleinen $\varepsilon > 0$ gültig ist und lässt dann $\varepsilon \to 0$ gehen ...

[131] Um den Unterschied zur bisherigen Notation, wo A_J ja eine *Zeilenauswahl* darstellte, klarzumachen wird "′" verwendet.

[132] In den "ersten" m Komponenten herrscht ja Gleichheit!

Übung 5.1 ("Kolmogoroff[133]–Kriterium")
Zeigen Sie, daß x^* genau dann Optimallösung von (5.1) ist, wenn

$$0 \geq (Ay - c)^T (x - x^*), \qquad x \in F_x,\ y \in F_y$$

gilt.
Hinweis: x^* ist offensichtlich Optimallösung genau dann, wenn $0 \geq c^T x^* - c^T x$ für alle $x \in F_x$. Formen Sie diesen Ausdruck geeignet um. ◇

Als nächstes eine Aussage, in der die Frage, wann $F_x \neq \emptyset$ ist, über die dualen Nebenbedingungen charakterisiert wird, siehe (Ye, 1997, Theorem 1.9, S. 17) oder (Spellucci, 1993, A2.1.4, S. 40)

Satz 5.5 ("Farkas–Lemma") *Für $A \in \mathbb{R}^{m\times n}$ und $b \in \mathbb{R}^m$ ist $F_x \neq \emptyset$ genau dann, wenn*

$$A^T y \geq 0 \qquad \Rightarrow \qquad b^T y \geq 0. \tag{5.5}$$

Beweis: Die Richtung "⇒" ist einfach: Ist $x \in F_x$, also $Ax = b$ sowie $x \geq 0$, und ist $A^T y \geq 0$, dann ist

$$b^T y = (Ax)^T y = \underbrace{x^T}_{\geq 0} \underbrace{A^T y}_{\geq 0} \geq 0.$$

Für die Umkehrung "⇐" bemerken wir zuerst, daß $F_x = \emptyset$ bedeutet, daß der Vektor $b \in \mathbb{R}^m$ nicht zu dem konvexen Kegel

$$K_A := A\,\mathbb{R}^n_+ = \{Ax \ :\ x \geq 0\} \tag{5.6}$$

gehört. Dieser konvexe Kegel ist der Durchschnitt seiner berandenden Halbräume, das heißt,

$$K_A = \bigcap_{j=1}^{N} \left\{ y \in \mathbb{R}^m \ :\ v_j^T y \geq 0 \right\},$$

wobei die v_j, $j = 1, \ldots, N$, die **Normalenvektoren** auf die *linearen*[134] Halbräume sind, die auf einer Seite der von $m - 1$ Spaltenvektoren von A definierten[135] Hyperebenen liegen, siehe Übung 5.2. Sei also jetzt (5.5) erfüllt und nehmen wir an, es gäbe kein $x \geq 0$, so daß $Ax = b$ ist, da heißt $b \notin K_A$. Nach unserer obigen Bemerkung muß es also ein $v \in \{v_1, \ldots, v_N\}$ geben, so daß $v^T b < 0$ ist, woraus mit (5.5) (rückwärts gelesen) folgt, daß auch $A^T v \not\geq 0$ ist, daß es also mindestens ein $k \in \{1, \ldots, n\}$ gibt, so daß $(A^T v)_k < 0$ ist. Setzen wir nun $x = e_k$, dann ist $Ax \in K_A$ und somit

$$0 \leq v^T Ax = \left(A^T v\right)^T x = e_k^T \left(A^T v\right) = \left(A^T v\right)_k < 0,$$

[133] Andrey Nikolaevich Kolmogoroff (oder "Kolmogorov"), 1903–1987, trug wesentlich zu den Grundlagen der Wahrscheinlichkeitstheorie, aber auch zu Approximationstheorie, Topologie, Funktionalanalysis, Geometrie und so einigem mehr bei. Mit anderen Worten: einer der ganz, ganz großen Mathematiker des 20. Jahrhunderts!

[134] Daher auch, und das ist wichtig, die Beschreibung der Halbräume als $v_j^T y \geq 0$, wir haben hier keinen affinen oder "Verschiebungs"–Anteil!

[135] Aber nicht jede Auswahl von $m - 1$ Spalten liefert natürlich eine Randhyperebene, manche dieser Halbräume könnten ja durchaus redundant sein.

was ein offensichtlicher Widerspruch ist. Also muss $b \in K_A$ gelten. □

Übung 5.2 Zeigen Sie: zu jedem konvexen Kegel K_A, $A \in \mathbb{R}^{m\times n}$, wie in (5.6) gibt es Vektoren $v_1, \ldots, v_N$, so daß

$$K_A = \bigcap_{j=1}^{N} H_j, \qquad H_j = \left\{ y \,:\, v_j^T y \geq 0 \right\}.$$

♢

5.2 Kegel und Multiplikatoren

Das wesentliche Resultat dieses Kapitels wird eine Variante der Lagrangeschen[136] Multiplikatorenformel, die ja bekanntlich Aussagen über Extrema unter Nebenbedingungen macht, was ja im Kontext der Optimierung nicht zu abwegig erscheinen sollte. Um diese Resultate angeben und beweisen zu können brauchen wir aber ein bißchen mehr Terminologie.

Definition 5.6 (Kegel)

1. *Eine Menge* $K \subset \mathbb{R}^n$ *heißt* ***Kegel*** *mit Spitze* x*, wenn*

$$y \in K \qquad \Rightarrow \qquad x + \alpha(y - x) \in K, \quad \alpha \in \mathbb{R}_+.$$

2. *Sei* $M \subset \mathbb{R}^n$ *und* $x \in M$*. Als* ***abgeschlossener Tangentialkegel*** *an* M *in* x *wird die Menge*

$$T(M, x) := \bigcap_{\varepsilon>0} \overline{\{(y - x)\ \mathbb{R}_+ \,:\, y \in M,\ \|y - x\| \leq \varepsilon\}}$$

bezeichnet.

3. *Für* $M \subset \mathbb{R}^n$ *heißt*

$$M' = \left\{ y \in \mathbb{R}^n \,:\, y^T M \geq 0 \right\}$$

positiver Normalenkegel *an* M*.*

Übung 5.3 Zeigen Sie:

1. $T(M, x)$ und M' sind Kegel. Was ist die Spitze der beiden Kegel?
2. Für $y \in \mathbb{R}^n$ gilt $y \in T(M, x)$ genau dann, wenn es Folgen $x_k \in M$ und $\alpha_k \in \mathbb{R}_+$ gibt, so daß

$$\lim_{k\to\infty} x_k = x \qquad \text{und} \qquad \lim_{k\to\infty} \alpha_k (x_k - x) = y.$$

[136] Joseph–Louis Lagrange, 1736–1813, italienisch–französischer Mathematiker (geboren in Turin, das zu dieser Zeit aber zu Sardinien–Piemont gehörte; der Name seines Vaters ist Giuseppe Francesco Lodovico Lagrangia), Beiträge zur Wahrscheinlichkeitstheorie, Variationsrechnung und mathematischen Physik.

◇

Beispiel 5.7 *Sehen wir uns doch einmal den Tangentialkegel von*

$$F_x = \{x \in \mathbb{R}^n \ : \ Ax = b,\ x \geq 0\}$$

für vorgegebene Matrix $A \in \mathbb{R}^{m\times n}$ *und* $b \in \mathbb{R}^m$ *an. Eine* ε*–Umgebung von* x *in* F_x *besteht also aus allen Punkten* $z \in \mathbb{R}^n$*, so daß* $Az = b$*,* $z \geq 0$ *und* $\|z - x\| \leq \varepsilon$*. Der Vektor* $y := z - x$ *hat dann die Eigenschaft, daß*

$$Ay = Az - Ax = b - b = 0$$

und

$$y_j = \underbrace{z_j}_{\geq 0} - x_j \in \begin{cases} \mathbb{R}, & x_j > 0, \\ \mathbb{R}_+, & x_j = 0, \end{cases} \qquad j = 1, \ldots, n.$$

Der Tangentialkegel ist daher

$$T(F_x, x) = \{y \in \mathbb{R}^n \ : \ Ay = 0,\ y_j \geq 0 \text{ falls } x_j = 0\}.$$

Bemerkung 5.8 (Tangentialkegel)

1. *Die Idee des Tangentialkegels besteht darin, sich „beliebig kleine" Umgebungen des Punktes* x *in* M *anzusehen und die Strahlen in alle Richtungen zusammenzufassen, in die man so gehen kann. Den Abschluss verwendet man, um wirklich „auf Nummer sicher" zu gehen – es hat eher mit komplexeren Bereichen als unseren konvexen Polyedern zu tun.*

2. *Ist* $x \in M^\circ$ *ein* innerer *Punkt von* M*, so ist offensichtlich* $T(M, x) = \mathbb{R}^n$*.*

3. *Bei zweidimensionalen konvexen Polyedern, wo es ja nur drei Typen von Punkten, nämlich innere Punkte, Randpunkte und Eckpunkte, gibt, kann man sich die Tangentialkegel einfach vorstellen.*

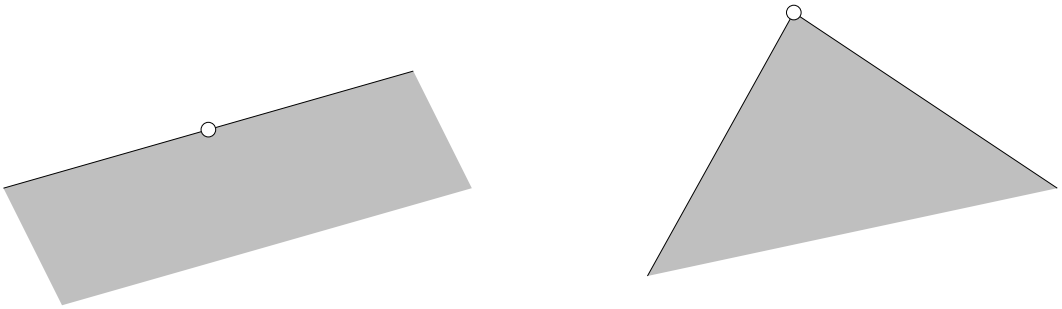

Abbildung 5.1: Die Tangentialkegel an einer Kante und einer Ecke eines konvexen Polyeders.

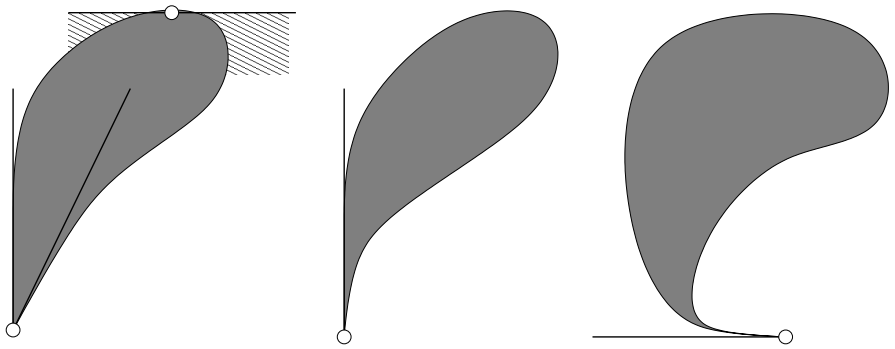

Abbildung 5.2: Beispiele für etwas allgemeinere Tangentialkegel zu Rand- und Eckpunkten. Ist der Rand des Bereichs an einer Stelle differenzierbar, dann besteht der Tangentialkegel aus allen Punkten auf "einer Seite" der Tangente. An Punkten mit Singularitäten der Randkurve kann der Tangentialkegel sogar zu einer Gerade degenerieren (mitte). Im dritten Beispiel erhält man den Tangentialkegel wirklich nur als Abschluß des Grenzwerts.

Definition 5.9 *Sei* $f \in C^1(\mathbb{R}^n)$ *eine stetig differenzierbare*[137] *Funktion, dann ist der* ***Gradient*** $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ *definiert als*

$$\nabla f = \left[\frac{\partial f}{\partial x_j} \ : \ j = 1, \ldots, n \right].$$

Die ***Richtungsableitung*** *von* f *in Richtung* $y \in \mathbb{R}^n$ *ist dann der Ausdruck*

$$D_y f = y^T \nabla f.$$

Mit der so bereitgestellten Notation können wir jetzt die notwendige Bedingung für die Existenz eines lokalen Minimums auf „etwas andere" Art hinschreiben.

Proposition 5.10 *Ist der Punkt* $x \in M \subset \mathbb{R}^n$ *ein lokales Minimum der Funktion* $f \in C^1(M)$*, dann gilt*

$$\nabla f(x) \in T(M, x)'. \tag{5.7}$$

Bemerkung 5.11 *Ist* $T(M, x) = \mathbb{R}^n$*, das heißt, der Punkt* x *ist ein "innerer Punkt", man kann von* x *aus innerhalb von* M *in alle Richtungen gehen, dann ist* $T(M, x)' = \{0\}$ *und wir erhalten in (5.7) das klassische Kriterium aus der Analysis für das Vorliegen eines Extremums: Die Ableitung, also in diesem Fall der Gradient, verschwindet!*

Beweis von Proposition 5.10: Sei $y \in T(M, x)$. Nach Übung 5.3[138] gibt es Folgen $M \ni x_k \to x$ und $\alpha_k \in \mathbb{R}_+$, $k \in \mathbb{N}$, mit $\alpha_k (x_k - x) \to y$. Das heißt insbesondere, daß $\alpha_k \to \infty$ oder $\alpha_k^{-1} \to 0$.

[137] Mit den Details totaler Differenzierbarkeit, stetiger Differenzierbarkeit, partieller Differenzierbarkeit und stetiger partieller Differenzierbarkeit wollen wir uns in dieser Vorlesung nicht herumschlagen. Trotzdem ist klar, daß sie bei hinreichender Allgemeinheit eine wichtige Rolle spielen würden.

[138] Ja, jetzt wird's gemein! Wer volle Gewissheit haben will, muß selbst was tun!

Da $x_k \to x$ und da x ein lokales Minimum ist, gilt also, für hinreichend großes $k \in \mathbb{N}$, daß

$$\begin{aligned} 0 &\leq f(x_k) - f(x) = f(x + (x_k - x)) - f(x) = f\left(x + \alpha_k^{-1}(\alpha_k(x_k - x))\right) - f(x) \\ &=: f\left(x + \alpha_k^{-1} y_k\right) - f(x), \end{aligned}$$

also auch

$$0 \leq \frac{f\left(x + \alpha_k^{-1} y_k\right) - f(x)}{\alpha_k^{-1}} \to D_y f = y^T \nabla f(x),$$

da $y_k \to y$ und da f stetig differenzierbar ist. Wir haben also gezeigt, daß unter der Annahme eines Minimums

$$y \in T(M, x) \qquad \Rightarrow \qquad y^T \nabla f(x) \geq 0 \qquad \Rightarrow \qquad \nabla f(x)\, T(M, x) \geq 0$$

gilt, was nichts anderes als (5.7) ist. □

So noch ein klein wenig Notation und wir können uns an unseren „Hauptsatz" dieses Kapitels machen, in dem wir einen sehr allgemeinen zulässigen Bereich, nämlich

$$\Omega = \Omega(g, h) = \{x \in \mathbb{R}^n \ : \ g(x) = 0,\ h(x) \geq 0\}$$

betrachten, wobei $g : \mathbb{R}^n \to \mathbb{R}^p$ und $h : \mathbb{R}^n \to \mathbb{R}^q$ zumindest mal **differenzierbar**[139] sein sollen.

Definition 5.12 (Aktive Nebenbedingungen und linearisierende Kegel)

1. *Für $x \in \Omega$ nennt man*

$$J(x) = \{1 \leq j \leq q \ : \ h_j(x) = 0\} \subseteq \{1, \ldots, q\}$$

   ***aktive Nebenbedingungen**.*

2. *Als **linearisierender Kegel** der Nebenbedingungen g, h wird die Menge*

$$\begin{aligned} L(x) &= L(x, g, h) \\ &= \left\{y \in \mathbb{R}^n \ : \ y^T \nabla g_j = 0,\ y^T \nabla h_k \geq 0,\ j = 1, \ldots, p,\ k \in J(x)\right\} \end{aligned}$$

   bezeichnet.

Übung 5.4 Zeigen Sie, daß $T(\Omega, x) \subseteq L(x)$. (Hinweis: Taylorformel) ◊

Satz 5.13 *Es sei $x \in \Omega$ eine Minimalstelle von $f \in C^1(\Omega)$ und es sei*

$$L(x)' = T(\Omega, x)'. \tag{5.8}$$

Dann existieren Vektoren $\lambda \in \mathbb{R}^p$ und $\mu \in \mathbb{R}^q_+$, so daß

$$\begin{aligned} \nabla f(x) - \nabla g(x)\, \lambda - \nabla h(x)\, \mu &= 0 \qquad (5.9) \\ \mu^T h(x) &= 0 \qquad (5.10) \end{aligned}$$

[139] Wer weiss noch, wie eine differenzierbare Funktion von $\mathbb{R}^n$ nach $\mathbb{R}^p$ wirklich definiert ist?

Hierbei ist für eine **vektorwertige Funktion** $g = (g_1, \ldots, g_p)$ der **Gradient** als

$$\nabla g = [\nabla g_1 \cdots \nabla g_p] = \begin{bmatrix} \frac{\partial g_1}{\partial x_1} & \cdots & \frac{\partial g_p}{\partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_1}{\partial x_n} & \cdots & \frac{\partial g_p}{\partial x_n} \end{bmatrix}$$

definiert. Das sieht anders aus, als man es zumeist aus Analysisbüchern kennt, ist dafür aber konsistent mit unserer Notation hier, in der die Gradienten skalarer Funktionen als **Spaltenvektoren** geschrieben werden. Bevor wir uns an den (gar nicht mal so schweren) Beweis dieses Satzes machen, erst einmal ein paar Bemerkungen und Konsequenzen daraus.

Bemerkung 5.14 (Multiplikatoren)

1. *Die geometrische Bedingung (5.8) an die Randbedingungen, die die Menge Ω festlegen (und nicht unbedingt an die Menge Ω selbst!) wird laut (Spellucci, 1993) als **Bedingung von Guingard** bezeichnet, die dieser in (Guingard, 1969) angegeben haben soll*[140]*. Was das allerdings für Bereiche sind, die (5.8)* nicht *erfüllen – wer weiß.*

2. *Da $X \subseteq Y \Rightarrow Y' \subseteq X'$ folgt mittels Übung 5.4, daß $L(x)' \subseteq T(\Omega, x)'$; was man also eigentlich in (5.8) fordert, ist daß $L(x)' \supseteq T(\Omega, x)'$. Der einfachste Fall läge sicherlich vor, wenn $L(x) = T(\Omega, x)$, aber dem ist halt leider nicht immer so, siehe Übung 5.5.*

3. *Die Vektoren λ und μ sind die wohlbekannten **Lagrange–Multiplikatoren**, weswegen wir sie auch in Zukunft als **Multiplikatoren** bezeichnen werden. Allerdings hat die spezielle Struktur der Nebenbedingungen auch Auswirkungen auf die Multiplikatoren: man kann λ als Vektor mit* nichtnegativen *Einträgen wählen!*

4. *In Optimiererkreisen werden die Bedingungen (5.9) und (5.10) auch als **Kuhn–Tucker–Bedingungen** bezeichnet.*

5. *Da $\mu \geq 0$ und $h(x) \geq 0$ – schließlich ist ja $x \in \Omega$ – bedeutet (5.10), daß die Träger der beiden Vektoren* disjunkt *sein müssen, das heißt $\mu_j > 0 \Rightarrow h_j(x) = 0$ und $h_j(x) > 0 \Rightarrow \mu_j = 0$.*

Übung 5.5 Bestimmen Sie für $n = 2$, $p = 0$, $q = 3$ und

$$h(x) = \begin{bmatrix} (1-x)^3 - y \\ x \\ y \end{bmatrix}$$

die Kegel $L(e_1)$ und $T(\Omega, e_1)$, wobei $e_1 = [1, 0]^T$ natürlich der erste Einheitsvektor ist. ◊

[140] Es ist ja nur Hörensagen, und es gibt durchaus genug Beispiele, die nahelegen, nicht alles zu glauben, was in Büchern steht.

Korollar 5.15 (Der lineare Fall) *Es sei* $A \in \mathbb{R}^{m\times n}$ *und* $x \in F_x \neq \emptyset$ *eine Minimalstelle von* $f(x) = c^T x$. *Dann gibt es Vektoren* $\lambda \in \mathbb{R}^m$ *und* $\mu \in \mathbb{R}^n_+$, *so daß*

$$\begin{aligned} c - A^T\lambda - \mu &= 0 && (5.11)\\ \mu^T x &= 0. && (5.12)\end{aligned}$$

Beweis: Wir betrachten also den Spezialfall, daß $f(x) = c^T x$, $g(x) = Ax - b$ und $h(x) = x$; damit sind die Gradienten

$$\nabla f = c, \qquad \nabla g = A^T, \qquad \nabla h = I_n \tag{5.13}$$

allesamt konstante Funktionen.
Zuerst weisen wir nach, daß F_x die Bedingung (5.8) erfüllt. Da

$$L(x) = \{z \in \mathbb{R}^n \ : \ Az = 0,\ z_j \geq 0, j \in J(x)\} = T\left(F_x, x\right), \tag{5.14}$$

siehe Beispiel 5.7, ist natürlich auch $L(x)' = T\left(F_x, x\right)'$. Damit können wir Satz 5.13 anwenden und erhalten (5.11) und (5.12), indem wir (5.13) in (5.9) und (5.10) einsetzen. □

Bemerkung 5.16 *Die einfache Beobachtung (5.14) kann man auch als die Tatsache interpretieren, daß die linearisierenden Nebenbedingungen zu linearen Nebenbedingung wieder die linearen Nebenbedingungen sind, was nun wieder nicht allzu überraschend klingt.*

Nun schließlich noch zum Beweis von Satz 5.13, der, im Gegensatz zu den Standard–Beweisen für die Lagrange–Multiplikatoren[141] sogar wohltuend kurz und einfach ist.
Beweis von Satz 5.13: Sei also $x \in \Omega$ eine Minimalstelle von f. Nach Proposition 5.10 und der Annahme (5.8) heißt das, daß

$$\nabla f(x) \in T\left(\Omega, x\right)' = L\left(x\right)', \qquad \Rightarrow \qquad z^T \nabla f(x) \geq 0, \quad z \in L(x).$$

Definieren wir $A \in \mathbb{R}^m$, $m := 2p + \#J(x)$ als

$$\begin{aligned} A \ &:= \ A(x) = [\nabla g(x), -\nabla g(x), [\nabla h_j(x) \ : \ j \in J(x)]] \\ &= \begin{bmatrix} \frac{\partial g_1(x)}{\partial x_1} & \cdots & \frac{\partial g_p(x)}{\partial x_1} & -\frac{\partial g_1(x)}{\partial x_1} & \cdots & -\frac{\partial g_p(x)}{\partial x_1} & \frac{\partial h_k(x)}{\partial x_1} & \cdots & \frac{\partial h_{k'}(x)}{\partial x_1} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_1(x)}{\partial x_n} & \cdots & \frac{\partial g_p(x)}{\partial x_n} & -\frac{\partial g_1(x)}{\partial x_n} & \cdots & -\frac{\partial g_p(x)}{\partial x_n} & \frac{\partial h_k(x)}{\partial x_n} & \cdots & \frac{\partial h_{k'}(x)}{\partial x_n} \end{bmatrix}, \end{aligned}$$

dann ist[142] nach der Definition von $L(x)$

$$z \in L(x) \qquad \Leftrightarrow \qquad A^T z \geq 0,$$

und wir können $\nabla f(x) \in L(x)'$ umschreiben in

$$A^T z \geq 0 \qquad \Rightarrow \qquad z^T \nabla f(x) \geq 0. \tag{5.15}$$

[141] Siehe z.B. (Heuser, 1983, 174.1, S. 320).

[142] Unter Verwendung der tiefliegenden Erkenntnis, daß $x = 0$ genau dann, wenn $[x, -x]^T \geq 0$.

Nach dem **Farkas–Lemma**, Satz 5.5, heißt dies aber, daß die Menge

$$\{\gamma \in \mathbb{R}^m \ : \ A\gamma = \nabla f(x),\ \gamma \geq 0\}$$

nichtleer ist und somit gibt es ein

$$\gamma = \left[\gamma_j^{(1)}, \gamma_j^{(2)}, \gamma_k^{(3)} \ : \ j = 1, \ldots, p,\ k \in J(x)\right] \in \mathbb{R}_+^{2p+\#J(x)},$$

so daß

$$\begin{aligned}
\nabla f(x) &= \sum_{j=1}^{p} \gamma_j^{(1)} \nabla g_j(x) - \sum_{j=1}^{p} \gamma_j^{(2)} \nabla g_j(x) + \sum_{j \in J(x)} \gamma_j^{(3)} \nabla h_j(x) \\
&= \sum_{j=1}^{p} \underbrace{\left(\gamma_j^{(1)} - \gamma_j^{(2)}\right)}_{=:\lambda_j} \nabla g_j(x) + \sum_{j \in J(x)} \underbrace{\gamma_j^{(3)}}_{=:\mu_j} \nabla h_j(x),
\end{aligned}$$

und mit $\mu_j = 0$, $j \in \{1, \ldots, q\} \setminus J(x)$, ergibt sich (5.9). Die andere Folgerung, (5.10), sieht man sofort, wenn man das innere Produkt ausschreibt und sich an die Definition von $J(x)$ erinnert:

$$\mu^T h(x) = \sum_{j \in J(x)} \mu_j \underbrace{h_j(x)}_{=0} + \sum_{j \notin J(x)} \underbrace{\mu_j}_{=0} h_j(x) = 0.$$

□

5.3 Affine Skalierung

Nach diesen Vorarbeiten können wir nun endlich unser erstes Verfahren herleiten, und zwar die **affine Skalierung** von Barnes (Barnes, 1986), die laut (Spellucci, 1993) auf ein wesentlich älteres Verfahren von Dikin (Dikin, 1967) zurückgeht[143]. Die Idee besteht darin, das Optimierungsproblem (5.1) nicht *global* auf F_x, sondern nur in einer Umgebung eines Punktes $x^{(r)}$ zu lösen, das lokale Optimum als $x^{(r+1)}$ zu wählen und sich hoffentlich auf diese Art und Weise dem *globalen* Optimum hinreichend schnell hinreichend genau anzunähern.
Allerdings benötigen wir dazu eine Voraussetzung:

> Das primale wie auch das duale Problem sollen keine *entarteten* Ecken besitzen, d.h.,
>
> $$\#\{j \ : \ x_j > 0\} \geq m, \qquad x \in F_x, \tag{5.16}$$
>
> und
>
> $$\#\left\{j \ : \ \left(A^T y\right)_j = c_j\right\} \leq m, \qquad y \in F_y, \tag{5.17}$$
>
> und A soll Rang m haben.

[143] Die Arbeit (Dikin, 1967) ist in den "Doklady" erschienen, wo Resultate normalerweise nur *vorgestellt*, nicht aber notwendigerweise bewiesen werden.

Auch diese Forderungen sind nur für „singuläre“ Probleme nicht erfüllt und können durch Störungen oder „positive“ Rundungseffekte erreicht werden.

Zur Konstruktion einer Folge von inneren Punkten wählt man einen Parameter $0 < \rho < 1$ und betrachtet zu $x \in F_x$, $x > 0$, das Hilfsproblem

$$\min_y c^T y, \qquad Ay = b, \quad \sum_{j=1}^{n} \frac{(y_j - x_j)^2}{x_j^2} \le \rho^2, \tag{5.18}$$

wobei die nichtlineare Nebenbedingung dafür sorgt, daß wir im Inneren von F_x bleiben, also auch $y > 0$ ist. Wäre nämlich $y_k \le 0$ für ein $k \in \{1, \ldots, n\}$, dann wäre

$$\sum_{j=1}^{n} \frac{(y_j - x_j)^2}{x_j^2} \ge \frac{(y_k - x_k)^2}{x_k^2} \ge 1 > \rho^2.$$

Wir definieren nun die Menge $\Omega \subset \mathbb{R}^n$ durch die Randbedingungen[144]

$$0 = g(y) = Ay - b, \qquad 0 \le h(y) = \rho^2 - \sum_{j=1}^{n} \frac{(y_j - x_j)^2}{x_j^2}.$$

Damit ist Ω der Schnitt der Hyperebene $Ay = b$ mit dem Ellipsoid mit Mittelpunkt x und Halbachsen ρx_j, $j = 1, \ldots, n$, siehe Abb. 5.3.

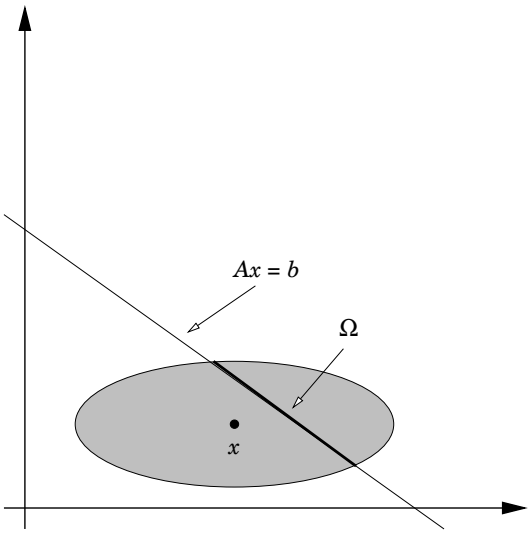

Abbildung 5.3: Der zulässige Bereich Ω als Schnitt der Ellipse um x mit Halbachsen $\rho\, x$ und der Hyperebene $Ax = b$.

Wir wollen jetzt Satz 5.13 anwenden, um das lokale Minimum von $c^T y$ auf Ω zu finden[145], wozu wir erst einmal die Guingard–Bedingung (5.8) nachweisen müssen. Nun hat Ω für y mit $h(y) > 0$ den Tangentialkegel

$$T(\Omega, y) = \left\{ z \,:\, A^T z = 0 \right\},$$

[144] Nicht vergessen: wir halten x fest und optimieren bezüglich y!

[145] Für etwas mehr als pure mathematische Ästhetik sollte der Satz schon gut sein!

ist hingegen $h(y) = 0$, so ist für jedes $y' \in \Omega$

$$\begin{aligned} 0 \;\leq\; h(y') = h(y') - \underbrace{h(y)}_{=0} &= \int_0^1 \left(D_{y'-y} h\right)(y + t(y'-y))\, dt \\ &= \int_0^1 (y'-y)^T \nabla h\,(y + t(y'-y))\, dt, \end{aligned}$$

also auch

$$0 \leq \frac{1}{\|y'-y\|} \int_0^1 (y'-y)^T \nabla h\,(y + t(y'-y))\, dt,$$

und für $\|y'-y\| \to 0$ liefert das die Elemente des Tangentialkegels für $h(y) = 0$ als

$$T(\Omega, y) = \left\{ z \,:\, A^T z = 0,\, z^T \nabla h \geq 0 \right\},$$

was, wie man leicht sieht, wieder nichts anderes als $L(y)$ ist. Also ist die Bedingung (5.8) von Satz 5.13 erfüllt und die gesuchte (und ja auch vorhandene) Minimalstelle y liefert die Existenz von $\lambda \in \mathbb{R}^m$ und $\mu \in \mathbb{R}_+$, so daß

$$\begin{aligned} c - A^T\lambda - 2\mu X^{-2}(x-y) &= 0, & (5.19)\\ \mu\left(\rho^2 - (y-x)^T X^{-2}(y-x)\right) &= 0, & (5.20) \end{aligned}$$

wobei

$$X = \begin{bmatrix} x_1 & & \\ & \ddots & \\ & & x_n \end{bmatrix} \quad \Rightarrow \quad \begin{cases} h(y) = \rho^2 - (y-x)^T X^{-2}(y-x), \\ \nabla h(y) = 2X^{-2}(x-y). \end{cases}$$

Die Lösung $\mu = 0$, die (5.20) so einfach machen würde ist nicht zulässig, denn das ergäbe, eingesetzt in (5.19), daß $A^T\lambda = c$ wäre und somit hätte F_y eine degenerierte Ecke. Also muß $\mu > 0$ sein und wir erhalten aus (5.20) die erste Forderung an y, nämlich, daß

$$\rho^2 = (y-x)^T X^{-2}(y-x) = \left(X^{-1}(y-x)\right)^T \left(X^{-1}(y-x)\right) = \left\|X^{-1}(y-x)\right\|_2^2;$$

anders gesagt: y liegt am *Rand* der Ellipse. Multiplizieren wir außerdem (5.19) von links mit $(x-y)^T$, dann erhalten wir, nach geeigneter Umformung, daß

$$(x-y)^T\left(c - A^T\lambda\right) = 2\mu(x-y)^T X^{-2}(x-y) = 2\mu\rho^2$$

und somit

$$\mu = \frac{1}{2\rho^2}\Big((x-y)^T c - \underbrace{(x-y)^T A^T}_{=b^T - b^T = 0}\lambda\Big) = \frac{c^T(x-y)}{2\rho^2} = \frac{c^T(x-y)}{2\left\|X^{-1}(y-x)\right\|_2^2}.$$

An dieser Stelle sollten wir uns an eine wichtige Forderung aus Satz 5.13 erinnern, nämlich, daß $\mu > 0$ sein muss, was genau dann der Fall ist, wenn

$$c^T(x-y) > 0 \qquad (5.21)$$

ist. Das können wir aber immer gewährleisten, denn wäre y eine zulässige Lösung am Rand der Ellipse, also mit $\left\|X^{-1}(y-x)\right\|_2^2 = \rho^2$ und $Ay = b$, dann können wir y durch $y' = 2x - y$ ersetzen und erhalten

$$\left\|X^{-1}(y'-x)\right\|_2^2 = \left\|X^{-1}(2x-y-x)\right\|_2^2 = \left\|X^{-1}(x-y)\right\|_2^2 = \rho^2$$

sowie

$$c^T(y'-x) = c^T(x-y) = -c^T(y-x).$$

Ausserdem ist y' ebenfalls ein zulässiger Punkt, zumindest wenn x und y zulässig sind, denn dann ist

$$Ay' = 2\underbrace{Ax}_{=b} - \underbrace{Ay}_{=b} = b.$$

Tatsächlich sind die Rollen von y und y' auch geometrisch klar verteilt: Einer der beiden Punkte liefert das *Maximum* des „Hilfsproblems“ (5.18), der andere das *Minimum*. Zwischen den beiden Typen von Extrema können aber Lagrange–Multiplikatoren nicht unterscheiden!

Wir nehmen nun an, wir hätten das „richtige“ y gewählt, dann erhalten wir wirklich

$$c^T y = c^T x - 2\rho^2\,\mu < c^T x, \tag{5.22}$$

der Wert der Zielfunktion wird durch die Lösung unseres Hilfsproblems definitiv *verkleinert*.

Bleibt also noch die Bestimmung von y. Dazu multiplizieren wir (5.19) von links mit AX^2 und erhalten, daß

$$0 = AX^2c - AX^2A^T\lambda - 2\mu\underbrace{A(x-y)}_{=0} = AX^2c - (AX)\,(AX)^T\,\lambda \tag{5.23}$$

Da A den Maximalrang m hat, ist die Matrix $(AX)(AX)^T$ symmetrisch und positiv definit[146] und demnach invertierbar, was es uns ermöglicht, den Multiplikator λ als

$$\lambda = \left(AX^2A^T\right)^{-1} AX^2c \tag{5.24}$$

zu bestimmen. Damit bekommen wir aber auch μ: Formen wir nämlich

$$2\mu X^{-1}(x-y) = X\left(c - A^T\lambda\right)$$

um und nehmen auf beiden Seiten die euklidische Norm, dann ist

$$2\mu = \frac{\left\|X\,(c - A^T\lambda)\right\|_2}{\left\|X^{-1}(y-x)\right\|_2} = \frac{\left\|X\,(c - A^T\lambda)\right\|_2}{\rho}$$

Das setzen wir nun alles nochmal in (5.19) ein und erhalten so, daß

$$y = x + \frac{1}{2\mu}\,X^2\left(c - A^T\lambda\right) = x + \rho\,\frac{X^2\,(c - A^T\lambda)}{\left\|X\,(c - A^T\lambda)\right\|_2}. \tag{5.25}$$

[146] Also *strikt* positiv definit!

Wer will kann jetzt noch (5.24) einsetzen, aber schöner oder gar übersichtlicher wird dadurch auch nichts mehr.

Kehren wir nochmal zum Fall zurück, daß wir y durch $y' = 2x - y$ ersetzen müssen, dann tritt dieser ja ein, wenn

$$0 < c^T(y - x) = \rho \frac{c^T X^2 (c - A^T\lambda)}{\left\| X (c - A^T\lambda) \right\|_2},$$

also wenn $c^T X^2 (c - A^T\lambda) > 0$ und dann ist

$$y' = 2x - \left(x + \rho \frac{X^2 (c - A^T\lambda)}{\left\| X (c - A^T\lambda) \right\|_2} \right) = x - \rho \frac{X^2 (c - A^T\lambda)}{\left\| X (c - A^T\lambda) \right\|_2}.$$

Diese Fallunterscheidung müssen wir also noch in unseren Algorithmus einbauen.

Übung 5.6 Zeigen Sie, daß aus (5.25) und den übrigen Voraussetzungen $Ay = b$ folgt. ♢

Fassen wir also zusammen.

Algorithmus 5.17 (Affine Skalierung)
Gegeben: $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$ *und innerer Punkt* $x^{(0)}$ *mit* $Ax^{(0)} = b$, $x^{(0)} > 0$.

1. *Finde inneren Punkt* $x^{(0)}$ *mit* $Ax^{(0)} = b$, $x^{(0)} > 0$.
2. *Wähle* $\rho \in (0, 1)$
3. *Für* $r = 0, 1, 2, \ldots$
   (a) *Setze*
   $$X = \begin{bmatrix} x_1^{(r)} & & \\ & \ddots & \\ & & x_n^{(r)} \end{bmatrix}.$$
   (b) *Berechne* $B = XA^T$ *und*
   $$\lambda = \left(B^T B \right)^{-1} BXc$$
   (c) *Setze*
   $$z = X\left(c - A^T\lambda \right).$$
   (d) *Berechne*
   $$x^{(r+1)} = \begin{cases} x^{(r)} + \rho \dfrac{Xz}{\|z\|_2}, & c^T Xz < 0, \\ x^{(r)} - \rho \dfrac{Xz}{\|z\|_2}, & c^T Xz > 0. \end{cases}$$

Bemerkung 5.18 *Schreibt man (5.23) als*

$$(AX)(AX)^T\lambda = (AX)Xc$$

um, dann stellt diese Gleichung die sogenannten ***Normalengleichungen*** *zum* ***Least–squares–Problem***

$$\min_y \left\|Xc - (AX)^T y\right\|_2 = \min_y \left\|X\left(c - A^T y\right)\right\|_2$$

und der Vektor $X\left(c - A^T\lambda\right)$ *stellt das* ***Residuum*** *dieses Minimierungsproblems dar; beide Werte, die Lösung wie auch der Fehler, können sehr stabil über ein* QR*–Verfahren bestimmt werden, siehe z.B. (Sauer, 2000), vor allem aber (Golub & van Loan, 1996) oder (Higham, 1996): Dazu bestimmt man eine orthogonale*[147] *Matrix* $Q \in \mathbb{R}^{m\times m}$*, so daß*

$$XA^T = Q\begin{bmatrix} R \\ 0_{n-m,m} \end{bmatrix}, \qquad R = \begin{bmatrix} * & \cdots & * \\ & \ddots & \vdots \\ & & * \end{bmatrix} \in \mathbb{R}^{m\times m},$$

und bestimmt λ *als Lösung des einfachen Dreieckssystems* $R\lambda = Q^T Xc$*. Das führt zu Algorithmus 5.19.*

Algorithmus 5.19 (Affine Skalierung, QR**–Version)**
Gegeben: $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$ *und innerer Punkt* $x^{(0)}$ *mit* $Ax^{(0)} = b$, $x^{(0)} > 0$.

1. *Finde inneren Punkt* $x^{(0)}$ *mit* $Ax^{(0)} = b$, $x^{(0)} > 0$.
2. *Wähle* $\rho \in (0,1)$
3. *Für* $r = 0, 1, 2, \ldots$
   (a) *Setze*
   $$X = \begin{bmatrix} x_1^{(r)} & & \\ & \ddots & \\ & & x_n^{(r)} \end{bmatrix}.$$
   (b) *Berechne* $B = XA^T$ *und* Q, R *als* QR*–Zerlegung*[148] *von* B *und* λ *als Lösung von*
   $$R\lambda = [I_m\, 0]\; Q^T Xc, \qquad I_m = \begin{bmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{bmatrix} \in \mathbb{R}^{m\times m}.$$
   (c) *Setze*
   $$z = X\left(c - A^T\lambda\right)$$

[147] Zur Erinnerung: eine reelle Matrix $Q \in \mathbb{R}^{n\times n}$ heißt *orthogonal*, wenn $Q^T Q = QQ^T = I$ ist.
[148] Die QR–Zerlegung ist ein Standardbestandteil jeder Software für numerische Lineare Algebra, insbesondere von `Matlab` und `Octave`, siehe auch (Anderson *et al.*, 1995).

(d) Berechne

$$x^{(r+1)} = \begin{cases} x^{(r)} + \rho \dfrac{Xz}{\|z\|_2}, & c^T Xz < 0, \\ x^{(r)} - \rho \dfrac{Xz}{\|z\|_2}, & c^T Xz > 0. \end{cases}$$

Man kann nun sogar *beweisen*, daß dieses Verfahren tatsächlich gegen eine Lösung des Optimierungsproblems konvergiert.

Satz 5.20 *Besitzen das lineare Optimierungsproblem und sein duales keine entarteten Ecken, dann konvergiert die Folge*

$$x^{(r+1)} := x^{(r)} + \rho X^{(r)} \frac{X^{(r)}\left(c - A^T\lambda^{(r)}\right)}{\left\|X^{(r)}\left(c - A^T\lambda^{(r)}\right)\right\|_2}, \qquad X^{(r)} = \begin{bmatrix} x_1^{(r)} & & \\ & \ddots & \\ & & x_n^{(r)} \end{bmatrix} \tag{5.26}$$

für jeden Startwert $x^{(0)} \in F_x^\circ$ gegen eine Optimallösung x^* und es gilt für $r \in \mathbb{N}_0$

$$0 \le c^T x^{(r+1)} - c^T x^* \le \beta_r \left(c^T x^{(r)} - c^T x^*\right), \qquad \beta_r = 1 - \frac{\rho}{\sqrt{n - m + \varepsilon_r}}, \quad \varepsilon_r \to 0. \tag{5.27}$$

Bemerkung 5.21 *Die Abschätzung (5.27) liefert eine Aussage über die **Approximationsordnung** oder **Konvergenzordnung** des Verfahrens: Mit $\beta = \max_r \beta_r$ ist ja*

$$\begin{aligned} c^T x^{(r+1)} - c^T x^* &\le \beta\left(c^T x^{(r)} - c^T x^*\right) \le \beta^2\left(c^T x^{(r-1)} - c^T x^*\right) \le \cdots \\ &\le \beta^{r+1}\left(c^T x^{(0)} - c^T x^*\right), \end{aligned}$$

wir haben es also mit exponentieller Konvergenz, aber sogenannter linearer Konvergenzordnung *zu tun*[149]

Beweis: Der Beweis ist etwas länglich und gliedert sich grob in drei Teile:

1. Wir werden zuerst zeigen, daß die Folge $x^{(r)}$ immer gegen einen Grenzwert konvergiert. Dazu brauchen wir die Nichtentartungsbedingungen an F_x und F_y.

2. Dann zeigen wir, daß dieser Grenzwert tatsächlich eine Optimallösung ist, wozu wir ein Dualitätsargument verwenden werden.

3. Schließlich müssen wir noch die Konvergenzordnung (5.27) nachweisen – das ist im wesentlichen eine technische Abschätzung.

Beginnen wir mit der Konvergenz. Da $x^{(r)}$ eine Folge in dem kompakten Polyeder F_x ist, muß zumindest eine Teilfolge konvergieren, oder die Folge einen Häufungspunkt besitzen, den wir einmal x^∞ nennen wollen.

[149] Die Terminologie ist da leider nicht so richtig einheitlich.

Wegen (5.22) ist für $r \in \mathbb{N}_0$

$$\begin{aligned} c^T x^* &\leq c^T x^{(r+1)} \leq c^T x^{(r)} - 2\rho^2 \mu^{(r)} = c^T x^{(r)} - 2\rho \left\| X^{(r)} \left(c - A^T \lambda^{(r)} \right) \right\|_2 \\ &= c^T x^{(r-1)} - 2\rho \left\| X^{(r-1)} \left(c - A^T \lambda^{(r-1)} \right) \right\|_2 - 2\rho \left\| X^{(r)} \left(c - A^T \lambda^{(r)} \right) \right\|_2 \\ &= c^T x^{(0)} - 2\rho \sum_{j=0}^{r} \left\| X^{(j)} \left(c - A^T \lambda^{(j)} \right) \right\|_2 \end{aligned}$$

und daher

$$\sum_{j=0}^{r} \left\| X^{(j)} \left(c - A^T \lambda^{(j)} \right) \right\|_2 \leq \frac{1}{2\rho} \left(c^T x^{(0)} - c^T x^* \right), \qquad r \in \mathbb{N}_0, \tag{5.28}$$

die Reihe auf der linken Seite konvergiert also und es folgt, daß

$$\lim_{r \to \infty} X^{(r)} \left(c - A^T \lambda^{(r)} \right) = 0. \tag{5.29}$$

Da das duale Problem nichtentartet ist gibt es eine Konstante $C > 0$, die nur vom Optimierungsproblem abhängt[150] und zu jedem r eine Indexmenge J mit $\#J \geq n - m$, so daß

$$\left(c - A^T \lambda^{(r)} \right)_J > C \, 1_J;$$

nach eventuellem Übergang zu einer Teilfolge[151], die gegen x^∞ konvergiert, gilt dies dann für *eine* Indexmenge J unabhängig von r. Wegen (5.29) ist dann

$$\lim_{r \to \infty} X_J^{(r)} = \lim_{r \to \infty} x_J^{(r)} = x_J^\infty = 0,$$

die Folge konvergiert also gegen eine Ecke und, wieder wegen der Nichtentartung von F_x, muß $\#J = n - m$ und

$$\lim_{r \to \infty} x_K^{(r)} = x_K^\infty > 0, \qquad K = \{1, \ldots, n\} \setminus J,$$

sein. Insbesondere ist der Häufungspunkt x^∞ eine Ecke. Wir müssen aber noch zeigen, daß auch wirklich *die ganze* Folge gegen x^∞ konvergiert. Dazu schauen wir uns nochmals (5.25) an und formen es in

$$x^{(r+1)} - x^{(r)} = \rho \, X^{(r)} \frac{X^{(r)} \left(c - A^T \lambda^{(r)} \right)}{\left\| X^{(r)} \left(c - A^T \lambda \right)^{(r)} \right\|_2} = \rho X^{(r)} y, \qquad \|y\|_\infty \leq \|y\|_2 = 1,$$

um, woraus

$$\left| x^{(r)} \right| - \left| x^{(r+1)} \right| \leq \rho \left| x^{(r)} \right| \qquad \Rightarrow \qquad \left| x^{(r+1)} \right| \geq (1 - \rho) \left| x^{(r)} \right|$$

[150] Siehe (Spellucci, 1993, S. 269); im wesentlichen hat es damit zu tun, daß zu jedem $y \in F_y$ mindestens $n - m$ Komponenten von $c - A^T y$ *strikt* positiv sein müssen und daß F_x ein *kompaktes* Polyeder ist.

[151] Es gibt nur endlich viele solcher Mengen, also muß mindestens eine für *unendlich viele* Werte von r auftreten, und eine solche greifen wir heraus und gehen zur entsprechenden Teilfolge über.

folgt, was uns liefert, daß $x_J^{(r)} \to 0$ für *alle* r, nicht nur für die Teilfolge, und damit, daß

$$x^\infty = \lim_{r\to\infty} x^{(r)}.$$

Damit ist also Punkt 1), die Konvergenz, erledigt.

Als nächstes weisen wir nach, daß x^∞ Optimallösung ist. Da die Folge der $x^{(r)}$ konvergiert existiert natürlich auch die Diagonalmatrix

$$X^\infty := \lim_{r\to\infty} X^{(r)}, \qquad X_J^\infty = 0, \quad X_K^\infty > 0$$

und hat offensichtlich Rang $\#K = m$. Damit müssen, zumindest für hinreichend große Werte von r, auch die Matrizen $X^{(r)}$ Rang $\geq m$ haben, womit $AX^{(r)}$ Rang m hat[152] und damit die Matrizen

$$\left(\left(AX^{(r)}\right)\left(AX^{(r)}\right)^T\right)^{-1} A\left(X^{(r)}\right)^2$$

existieren und und, da $AX^\infty = A'_K X_K^\infty$, gegen

$$\left(\left(A'_K X_K^\infty\right)\left(A_K X_K^\infty\right)^T\right)^{-1} A'_K \left(X^\infty\right)^2 =: \left[B'_K, B'_J\right], \qquad B'_J = 0_{m,m-n}$$

konvergieren. Damit existiert

$$\begin{aligned}
\lambda^\infty &= \lim_{r\to\infty}\left(\left(AX^{(r)}\right)\left(AX^{(r)}\right)^T\right)^{-1} A\left(X^{(r)}\right)^2 c \\
&= \left(\left(A'_K X_K^\infty\right)\left(A'_K X_K^\infty\right)^T\right)^{-1} A'_K \left(X_K^\infty\right)^2 c_K \\
&= \left(A'_K X_K^\infty X_K^{\infty T} A_K'^T\right)^{-1} A'_K X_K^{\infty 2} c_K = A_K'^{-T}\left(\operatorname{diag} x_K\right)^{-2} A_K'^{-1} A'_K \left(\operatorname{diag} x_K\right)^2 c_K \\
&= A_K'^{-T} c_K.
\end{aligned}$$

Somit ist

$$b^T\lambda^\infty = b^T A_K'^{-T} c_K = c_K^T\left(A_K'^{-1} b\right) = c_K^T x_K^\infty = c^T x^\infty$$

und nach unserem Dualitätssatz, Satz 5.3 ist $x^* = x^\infty$ eine Optimallösung. Den Nachweise der Zulässigkeit von λ^∞ sparen wir uns hier und verweisen nur auf (Spellucci, 1993, S. 271) – und damit ist 2) auch schon erledigt.

Für den Beweis der Konvergenzordnung betrachten wir schließlich

$$\begin{aligned}
0 \leq c^T x^{(r)} - c^T x^* &= c^T x^{(r)} - \underbrace{b^T}_{=Ax^{(r)}}\lambda^{(r)} + \underbrace{b^T}_{=Ax^*}\lambda^{(r)} - c^T x^* = \left(c - A^T\lambda^{(r)}\right)^T\left(x^{(r)} - x^*\right) \\
&= \left(X^{(r)}\left(c - A^T\lambda^{(r)}\right)\right)^T\left(X^{(r)}\right)^{-1}\left(x^{(r)} - x^*\right) \\
&\leq \left\|X^{(r)}\left(c - A^T\lambda^{(r)}\right)\right\|_2 \left\|\left(X^{(r)}\right)^{-1}\left(x^{(r)} - x^*\right)\right\|_2 \\
&\leq \frac{1}{\rho}\left(c^T x^{(r)} - c^T x^{(r+1)}\right)\left\|\left(X^{(r)}\right)^{-1}\left(x^{(r)} - x^*\right)\right\|_2
\end{aligned}$$

[152] Hier verwenden wir nochmal, daß es keine entarteten Ecken gibt! Jeder Rangdefekt würde nämlich eine entartete Ecke liefern.

Da

$$\left\|\left(X^{(r)}\right)^{-1} x^{(r)}\right\|_2^2 = \sum_{j=1}^{n} \frac{\left(x_j^{(r)} - x_j^*\right)^2}{\left(x_j^{(r)}\right)^2} = \sum_{j\in J^*} \frac{\left(x_j^{(r)} - x_j^*\right)^2}{\left(x_j^{(r)}\right)^2} + \sum_{j\in K^*} \frac{\left(x_j^{(r)} - x_j^*\right)^2}{\left(x_j^{(r)}\right)^2},$$

wobei

$$J^* := \left\{ j \,:\, x_j^* = 0 \right\}, \qquad K^* := \left\{ j \,:\, x_j^* \neq 0 \right\},$$

und somit

$$\sum_{j\in J^*} \frac{\left(x_j^{(r)} - x_j^*\right)^2}{\left(x_j^{(r)}\right)^2} = \sum_{j\in J^*} \frac{\left(x_j^{(r)}\right)^2}{\left(x_j^{(r)}\right)^2} = \#J^* \leq n - m,$$

ist also

$$c^T x^{(r)} - c^T x^* \leq \left(c^T x^{(r)} - c^T x^{(r+1)}\right) \underbrace{\frac{1}{\rho}\left(n - m + \sum_{j\in K^*}\left(\frac{x_j^{(r)} - x_j^*}{x_j^{(r)}}\right)^2\right)^{1/2}}_{=\frac{1}{\rho}(n-m+\varepsilon_r)=:\gamma_r}$$

mit $\varepsilon_r \to 0$ wegen der Konvergenz der $x^{(r)} \to x^*$ und weil die positiven Terme im Nenner nach unten beschränkt sind[153]. Also ist

$$\frac{1}{\gamma}\left(c^T x^{(r)} - c^T x^*\right) \leq c^T x^{(r)} - c^T x^* + c^T x^* - c^T x^{(r+1)}$$

und somit

$$c^T x^{(r+1)} - c^T x^* \leq \underbrace{\left(1 - \frac{1}{\gamma_r}\right)}_{=:\beta_r}\left(c^T x^{(r)} - c^T x^*\right),$$

wie behauptet. □

5.4 Primal & Dual I

Eine besondere Familie von innere–Punkte–Metohoden sind die **Primal–Dual–Verfahren**. Dazu greifen wir nochmal auf ein **primales Problem**

$$\min c^T x, \qquad Ax = b, \qquad x \geq 0, \tag{5.30}$$

und ein zugehöriges **duales Problem**

$$\max b^T y, \qquad A^T y \leq c \tag{5.31}$$

zurück und behandeln beide mit den **Lagrange–Multiplikatoren** aus Satz 5.13. Im Fall von (5.30) erhalten wir

$$\begin{aligned} c - A^T\lambda - \mu &= 0, \\ \mu^T x &= 0, \end{aligned} \tag{5.32}$$

[153] Ihr Grenzwert ist ja strikt positiv!

wohingegen (5.31) zu

$$\begin{array}{rcl} b - A\mu & = & 0, \\ \mu^T (A^T y - c) & = & 0, \end{array} \tag{5.33}$$

wird. Und jetzt kommt der „Taschenspielertrick“: Wir ersetzen (λ, μ) in (5.32) durch (y, z) und μ in (5.33) durch x und kommen so zu den **Primal–Dual–Bedingungen**

$$\begin{array}{rcl} 0 & = & c - A^T y - z, \\ 0 & = & b - Ax, \\ 0 & = & x^T \underbrace{\left(A^T y - c\right)}_{=z} = x^T z, \end{array} \tag{5.34}$$

deren Lösung ein *simultanes Extremum* des primalen *und* des dualen Problems ist und damit, nach dem Dualitätssatz 5.3, eine Optimallösung sein muss[154]. Wir suchen damit also eine Nullstelle der Funktion

$$F(x, y, z) = \begin{bmatrix} c - A^T y - z \\ b - Ax \\ x^T z \end{bmatrix} : \mathbb{R}^{2n+m} \to \mathbb{R}^{n+m+1}, \qquad x, z \geq 0. \tag{5.35}$$

Aber: Wie findet man eigentlich Nullstellen von Funktionen?

5.5 Exkurs: Das Newton–Verfahren und seine Freunde

Das **Newton–Verfahren** ist ein Klassiker der Numerischen Mathematik und beschäftigt sich mit dem Lösen nichtlinearer Gleichungen bzw. Gleichungssystemen. Betrachten wir zuerst einmal den Fall einer Gleichung in einer Variablen, also das Problem[155] $f(x) = 0$, das Finden einer **Nullstelle** und sehen uns einen kurzen Abriss der gängigen Verfahren an.

Das **Bisektionsverfahren** beginnt mit zwei Stellen x_-, x_+ an denen $\pm f(x_\pm) > 0$ gilt, an denen also f positives und negatives Vorzeichen hat und betrachtet dann den Mittelwert $x = \frac{1}{2}(x_- + x_+)$. Ist $f(x) = 0$, dann haben wir eine Nullstelle gefunden, ansonsten ersetzen wir x_- durch x falls $f(x) < 0$ ist und x_+ durch x falls $f(x) > 0$ ist. Auf diese Weise erhalten wir ganz schnell und einfach eine **Intervallschachtelung**, die eine Nullstelle enthalten muss[156], solange f wenigstens stetig ist.

Etwas „cleverer“ ist die **Regula Falsi**, die nicht blind die Mitte wählt, sondern den Punkt x in Abhängigkeit von $f(x_-)$ und $f(x_+)$ bestimmt, und zwar als Nullstelle der Verbindungsgeraden. Das liefert die Regel

$$x = \frac{x_+ f(x_-) - x_- f(x_+)}{f(x_-) - f(x_+)}, \tag{5.36}$$

[154] Unter den üblichen Voraussetzungen an Nichtdegeneriertheit natürlich.

[155] Das ist eine **Normalform** auf die man das „allgemeinere“ Problem $f(x) = g(x)$ immer ganz einfach bringen kann.

[156] Zwischenwertsatz

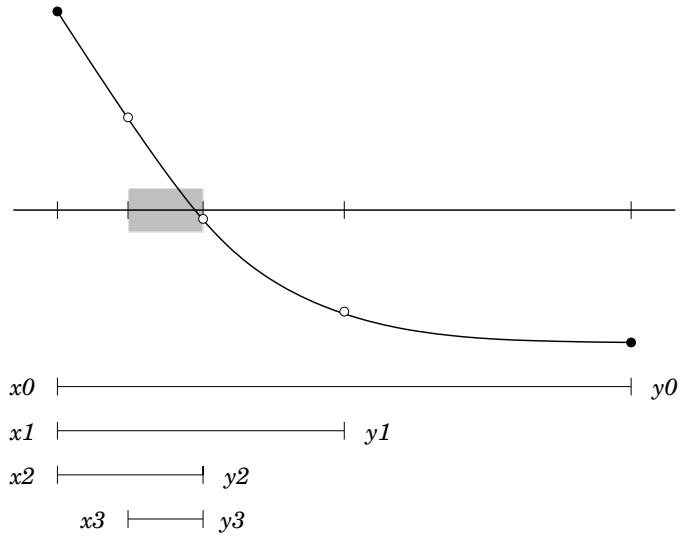

Abbildung 5.4: Die Vorgegehensweise beim Bisektionsverfahren. Ersetzt wird immer der Endpunkt, dessen Vorzeichen mit dem Vorzeichen am Mittelpunkt übereinstimmt. Nach jedem Iterationsschritt dieses Verfahrens besteht – wegen der Stetigkeit der Funktion – nun immer die Gewißheit, daß sich zwischen den Endpunkten eine Nullstelle befindet, nach unseren drei Schritten hier also irgendwo im schraffierten Intervall.

gefolgt von der Zuweisung an x_- bzw. x_+ je nach Vorzeichen von $f(x)$. Die Regula Falsi geht auf Fibonacci[157] zurück, siehe (Sigler, 2002), und wird von Adam Riese in (Riese, 1574) wie folgt beschrieben:

Regula Falsi oder Position.

Wirdt gefaßt von zweyen falschen zahlen / welche der auffgab nach / mit fleiß examinirt sollen werden / in massen das fragstück begeren ist / sagen sie der warheit zu viel / so bezeichne sie mit dem zeichen + plus / wo aber zu wenig / so beschreib sie mit dem zeichen − minus genannt. Als dann nimb ein lügen von der andern / was da bleibt / behalt für den theiler / multiplicir darnach im Creuz ein falsche zahl mit der andern lügen / nimb eins vom andern / vnd das da bleibt theil ab mit fürgemachtem theiler / so kompt berichtung der frag.

Das mag für alle die hilfreich sein, die sich keine Formeln wie (5.36) merken können.

Bisektion und Regula Falsi sind sichere Vefahren in dem Sinne, daß sie **Einschlußverfahren** sind, bei denen mindestens eine Nullstelle garantiert zwischen $x_\pm$ liegen muss. Das ist gut, setzt aber voraus, daß man zwei solche Punkte erst einmal kennen muss. Man kann allerdings auf diese Forderung auch verzichten

[157] Der sie wahrscheinlich auch nicht erfunden, sondern nur von den Arabern aufgeschnappt hat.

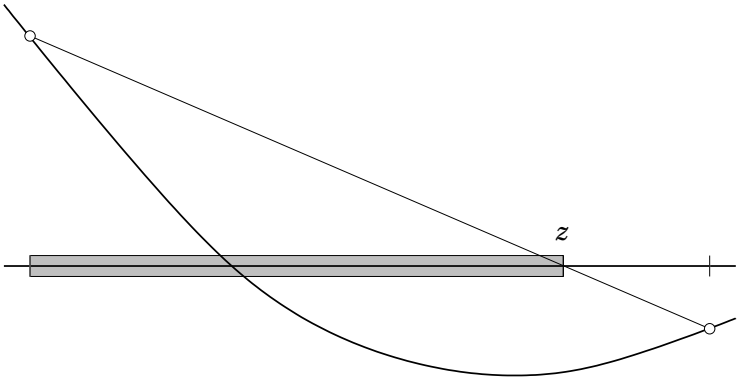

Abbildung 5.5: Ein Iterationsschritt der Regula Falsi. Der "Testpunkt" wird nicht einfach als Intervallmitte gewählt, sondern als (eindeutige!) Nullstelle der Verbindungsstrecke zwischen $f(x_-)$ und $f(x_+)$. Weitergemacht wird also mit dem schraffierten Bereich.

und die Regula Falsi einfach als **iteratives Verfahren** ansehen, das aus zwei Punkten x_{j-1} und x_j einen neuen Punkt x_{j+1} nach der Regel (5.36) bestimmt, also

$$x_{j+1} = \frac{x_j\, f(x_{j-1}) - x_{j-1}\, f(x_j)}{f(x_{j-1}) - f(x_j)}$$

und dann einfach mit dem Paar x_j, x_{j+1} weitermacht. Nun ist x_{j+1} nichts anderes als die Nullstelle der **Sekante** durch die Werte an den Stellen x_{j-1} und x_j, weswegen man[158] hier vom **Sekantenverfahren** spricht. Das Sekantenverfahren gibt die Sicherheit auf, konvergiert nur noch lokal, das heißt, wenn die beiden Startwerte x_0, x_1 bereits hinreichend nahe[159] an einer Nullstelle liegen, dafür aber konvergiert es signifikant schneller als Bisektion und Regula Falsi.

Was ein bisschen seltsam beim Sekantenverfahren ist, ist die Tatsache, daß die Rollen der beiden Startpunkte x_0, x_1 bei der Bildung von x_2 völlig symmetrisch sind, daß man dann aber einen der beiden Punkte verwerfen muss, und zwar den, der „zufällig" x_0 heißt. Einen richtig gute Grund dafür gibt es aber eigentlich nicht. Schreiben wir unsere Rechenvorschrift des Sekantenverfahrens ein wenig um,

$$\begin{aligned} x_{j+1} &= \frac{x_j\, f(x_{j-1}) - x_j\, f(x_j) + x_j\, f(x_j) - x_{j-1}\, f(x_j)}{f(x_{j-1}) - f(x_j)} \\ &= x_j \frac{f(x_{j-1}) - f(x_j)}{f(x_{j-1}) - f(x_j)} - f(x_j)\, \frac{x_{j-1} - x_j}{f(x_{j-1}) - f(x_j)} x_j - f(x_j)\, \frac{x_{j-1} - x_j}{f(x_{j-1}) - f(x_j)}, \end{aligned}$$

dann können wir die beiden Punkte x_{j-1} und x_j zusammenfallen lassen[160] und

[158] Völlig überraschend!
[159] Dafür gibt es keine vernünftigen a–priori–Abschätzungen!
[160] Formal betrachten wir $\lim_{x_{j-1} \to x_j}$ der Interationsvorschrift.

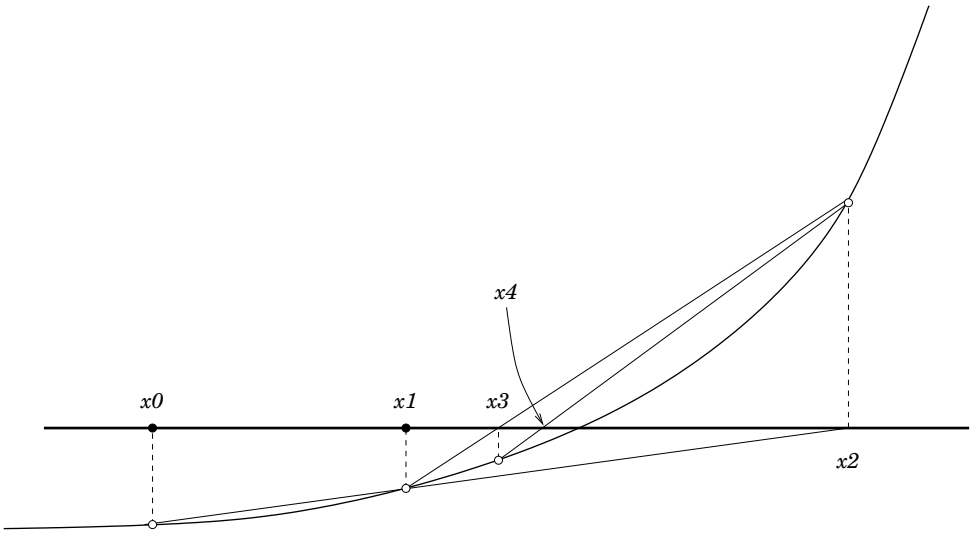

Abbildung 5.6: Die ersten Schritte des Sekantenverfahrens. Die "neuen" Punkte werden ermittelt, indem die beiden letzten "alten" Punkte verbunden werden und die Verbindungslinie mit der Nulllinie geschnitten wird.

erhalten für eine **differenzierbare Funktion** f die Vorschrift

$$x_{j+1} = x_j - \frac{f(x_j)}{f'(x_j)}, \tag{5.37}$$

bei der wir geometrisch die Sekante durch eine **Tangente** ersetzt haben. Und das ist auch schon das **Newton–Verfahren**, das ebenfalls nur **lokal konvergent** ist[161], Differenierbarkeit der Funktion und Berechnung der Ableitung voraussetzt, zur Belohnung aber noch schneller konvergiert als das Sekantenverfahren, und zwar *quadratisch*.

Beispiel 5.22 (Quadratische Konvergenz) *Anstatt jetzt den Begriff der **Konvergenzordnung** formal zu definieren, sehen wir uns ein Beispiel an, das sehr schön illustriert, was quadratische Konvergenz bedeutet, und zwar die Bestimmung des **Reziprokwerts** $1/a$ für gegebenes $a > 0$. Der Wert $1/a$ ist eine Nullstelle der Funktion $f(x) = 1/x - a$, für die sich nach (5.37) die Iterationsvorschrift*

$$x_{j+1} = x_j - \frac{x_j^{-1} - a}{-x_j^{-2}} = x_j + x_j - ax_j^2 = 2x_j - ax_j^2$$

ergibt, die ohne Division auskommt[162]*. Versuchen wir einmal unser Glück mit $a = 3$ und dem gut geratenen Startwert $x_0 = 0.3$, so ergibt sich*

$$x_0 = 0.3$$

[161] Der Startwert muss nahe genug bei der Nullstelle liegen!

[162] So wurde früher auf Computern, die nur multiplizieren konnten, dividiert, heute gibt es deutlich trickreichere Verfahren.

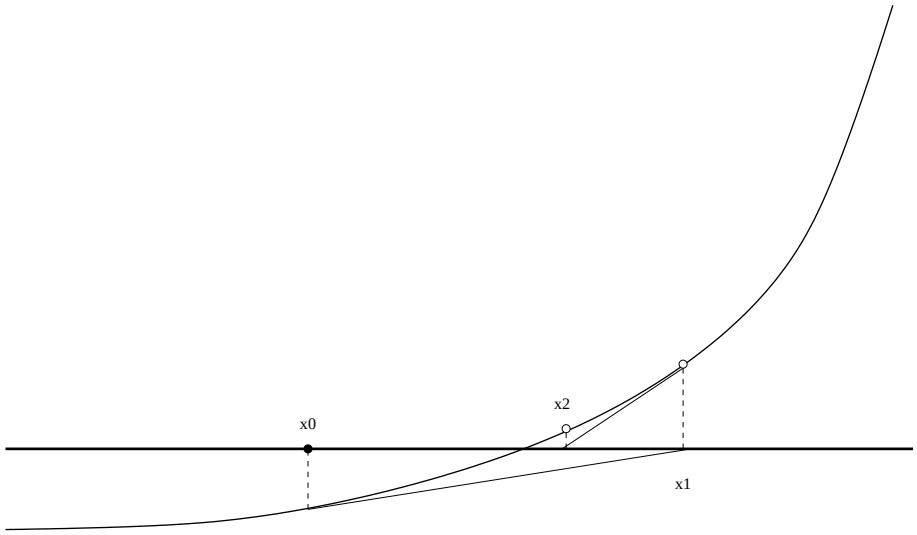

Abbildung 5.7: Die ersten Iterationsschritte des Newton–Verfahrens.

$$\begin{aligned} x_1 &= 0.6 - 3\,(0.3)^2 = 0.6 - 0.27 = 0.33 \\ x_2 &= 0.66 - 3\,(0.33)^2 = 0.3333 \\ &\vdots \\ x_j &= 0.\underbrace{3\ldots3}_{2j}, \end{aligned}$$

die Anzahl der gültigen Ziffern der Lösung verdoppelt sich also mit jedem Schritt.

Jetzt fehlt nur noch das Newton–Verfahren für Funktionen in mehreren Variablen. Das klappt für $f : \mathbb{R}^n \to \mathbb{R}^n$, dann ist die **Jacobi–Matrix** $f' := Df := \left[\frac{\partial f_j}{\partial x_k} : j,k\right]$ eine quadratische Matrix und $1/f' = Df^{-1}$ einfach die Inverse dieser Jacobimatrix. Die resultierende Iterationsvorschrift lautet somit

$$x_{j+1} = x_j - (Df(x_j))^{-1} f(x_j) \tag{5.38}$$

bzw.

$$\Delta x_j := x_{j+1} - x_j = -(Df(x_j))^{-1} f(x_j), \tag{5.39}$$

wobei Δx_j die Iterationsrichtung des Newtonverfahrens oder einfach **Newtonrichtung** bezeichnet: $x_{j+1} = x_j + \Delta x_j$. In der Praxis wird das Newtonverfahren oft **relaxiert**: Man berechnet $x_{j+1} = x_j + \alpha_j \Delta x_j$ für einen Parameter α_j.

5.6 Primal & Dual II

Nun aber zurück zu der Funktion $F(x,y,z)$ aus (5.35), die so, wie sie da steht, erst einmal noch nicht für das Newton–Verfahren geeignet ist. Das Problem ist das innere Produkt $x^T z$, das zu klein ist; wegen $x, z \geq 0$ ist es allerdings äquivalent

zu

$$0 = [x_j z_j : j = 1, \ldots, n] = \underbrace{\begin{bmatrix} x_1 & & \\ & \ddots & \\ & & x_n \end{bmatrix}}_{=:X} \underbrace{\begin{bmatrix} z_1 & & \\ & \ddots & \\ & & z_n \end{bmatrix}}_{=:Z} 1 = XZ1 \in \mathbb{R}^n,$$

und das passt von der Dimension dann auch. Die Funktion ändert sich dann also zu

$$F(x, y, z) = \begin{bmatrix} c - A^T y - z \\ b - Ax \\ XZ1 \end{bmatrix} : \mathbb{R}^{2n+m} \to \mathbb{R}^{2n+m}, \qquad x, z \geq 0,$$

mit

$$DF(x, y, z) = \begin{bmatrix} 0 & A^T & -I \\ -A & 0 & 0 \\ Z & 0 & X \end{bmatrix}.$$

Die Newton–Iteration (5.39) wird damit in ein **lineares Gleichungssystem**[163] der Gestalt

$$\begin{bmatrix} 0 & A^T & -I \\ -A & 0 & 0 \\ Z^{(j)} & 0 & X^{(j)} \end{bmatrix} \begin{bmatrix} \Delta x^{(j)} \\ \Delta y^{(j)} \\ \Delta z^{(j)} \end{bmatrix} = \begin{bmatrix} c - A^T y^{(j)} - z^{(j)} \\ b - Ax^{(j)} \\ X^{(j)} Z^{(j)} 1 \end{bmatrix} \tag{5.40}$$

umgewandelt, und der nächste Iterationswert ergibt sich als

$$\begin{bmatrix} x^{(j+1)} \\ y^{(j+1)} \\ z^{(j+1)} \end{bmatrix} = \begin{bmatrix} x^{(j)} \\ y^{(j)} \\ z^{(j)} \end{bmatrix} + \alpha \begin{bmatrix} \Delta x^{(j)} \\ \Delta y^{(j)} \\ \Delta z^{(j)} \end{bmatrix},$$

wobei α so gewählt werden muss, daß wenigstens

$$0 \leq x^{(j+1)} = x^{(j)} + \alpha \Delta x^{(j)} \qquad \Leftrightarrow \qquad \alpha \leq \min_{\{k : \Delta x_k^{(j)} < 0\}} -\frac{x_k^{(j)}}{\Delta x_k^{(j)}},$$

und entsprechendes für $z^{(j)}$ und $\Delta z^{(j)}$ erfüllt ist. Wenn alles gutgeht, konvergiert dieses Verfahren gegen eine simultane Lösung des primalen und des dualen Problems und die Qualität der Iteration kann über die **Dualitatslücke**

$$\delta_j := c^T x^{(j)} - b^T y^{(j)}$$

kontrolliert werden – der wahre Optimalwert der Zielfunktionen liegt ja irgendwo im Intervall $\left[c^T x^{(j)}, b^T y^{(j)}\right]$. Variationen dieses Verfahrens finden sich in (Nocedal & Wright, 1999).

[163] Kein Numeriker, der eine gewisse berufliche Selbstachtung hat, wird mit inversen Marizen herumrechnen.

Will man weitergehen, so befaßt man beides unter das Sein, dann unter das, was das Sein verleiht. Von hier aus kann man auf analytischem Wege wieder abwärts steigen [...]

Plotin, *Enneaden*, Band 1

6 Abstiegsverfahren für nichtlineare Optimierung

Jetzt ist es langsam an der Zeit, sich mit *beliebigen*[164] Optimierungsproblemen herumzuschlagen, also mit Problemen der Form

$$\min_{x\in\mathbb{R}^n} f(x), \qquad f : \mathbb{R}^n \to \mathbb{R}.$$

Solche Optimierungsprobleme bezeichnet man als **unrestringiert**, denn ein **zulässige Bereich**, auf den das Problem eingeschränkt wird, existiert ja jetzt nicht mehr. Allerdings ist das kein Verlust, sondern höchstens eine Verallgemeinerung. Wäre nämlich $D \subset \mathbb{R}^n$ ein kompakter zulässiger Bereich und wäre f wenigstens stetig[165] auf $\mathbb{R}^n$, dann kann man beispielsweise

$$G := \min_{x\in D} f(x) \qquad \text{und} \qquad g := f\chi_D + (1-\chi_D)\,(|G|+1)$$

setzen, dann nimmt g sein Minimum nur in D an und dort ist es das Minimum von f. Anders gesagt: Ein **restringiertes Optimierungsproblem** läßt sich ganz einfach in ein **unrestringiertes Optimierungsproblem** umschreiben. Allerdings gibt es dafür bessere Methoden.

Einen ganz klaren Vorteil hat die fehlende Nebenbedingung natürlich sofort: Man kann jetzt auf die Bedingungen an den **Tangentialkegel** aus Proposition 5.10 verzichten.

6.1 Notwendige und hinreichende Kriterien für Minima

Wiederholen[166] wir doch nochmal schnell ein paar Kriterien für die Existenz eines (lokalen) Minimums von f.

[164] Naja, so ganz beliebig können sie nicht sein, ohne ein paar Annahmen geht grundsätzlich nichts.

[165] "Prinzipiell" muss man bei den meisten Verfahren sogar die eine oder andere Form von Differenzierbarkeit annehmen.

[166] Sofern sie bekannt sind. Ansonsten kann es ja auch nicht schaden, etwas dazuzulernen.

Definition 6.1 *Eine Funktion* $f \in C(\mathbb{R}^n)$ *heißt* ***richtungsdifferenzierbar*** *in einem Punkt* $x \in \mathbb{R}^n$*, wenn für alle* $y \in \mathbb{R}^n$ *die Grenzwerte*

$$D_y f(x) = \lim_{h \to 0^+} \frac{f(x+hy) - f(x)}{h}$$

existieren.

Richtungsdifferenzierbarkeit ist *schwächer* als Differenzierbarkeit. So ist der Kegel $f(x) = \|x\|_2$ an der Stelle $x = 0$ richtungsdifferenzierbar, weil für jedes y ja $\|0 + hy\| - \|0\| = h\|y\|$ und somit als $D_y f = \|y\|$ ist, aber wegen der Spitze eben nicht differenzierbar. Ist hingegen $f \in C^1(\mathbb{R}^n)$ differenzierbar[167], dann ist natürlich

$$D_y f = (\nabla f)^T \, y. \tag{6.1}$$

Außerdem brauchen wir noch die zweite Ableitung von f, die sogenannte **Hesse–Matrix**, die man in der Optimierungsliteratur meist als

$$\nabla^2 f = \left[\frac{\partial^2 f}{\partial x_j \, \partial x_k} \; : \; j,k = 1,\ldots,n \right]$$

schreibt. Mit ihrer Hilfe kann man nun Minima sehr einfach und klassisch charakterisieren.

Proposition 6.2 *Sei* $f \in C^2(\mathbb{R}^n)$.

1. *Ist* x *ein lokales Minimum von* f*, dann ist* $\nabla f(x) = 0$ *und* $\nabla^2 f(x)$ *positiv semidefinit.*

2. *Ist* $\nabla f(x) = 0$ *und* $\nabla^2 f(x)$ *(strikt) positiv definit, dann ist* x *ein striktes lokales Minimum, d.h. es gibt eine Umgebung* D *von* x*, so daß* $f(x) < f(x')$, $x' \in D$.

Beweis: [168] Man verwendet die Taylor[169]–Entwicklung

$$f(x+ty) = f(x) + t\,(\nabla f(x))^T y + \frac{t^2}{2} y^T \Big(\nabla^2 f(\xi)\Big) y, \qquad \xi \in [x, ty],$$

und die Stetigkeit von ∇f und $\nabla^2 f$. Ist nämlich x ein Minimum, so ist *für alle Richtungen* $y \in \mathbb{R}^n$

$$0 \le \frac{f(x+ty) - f(x)}{t} \quad \to \quad (\nabla f(x))^T y,$$

[167] Im Sinne der Existenz einer (totalen) Ableitung als lineare Form ...

[168] Sozusagen zum "Aufwärmen".

[169] Brook Taylor, 1685–1731, war Mitglied einer 1712 eingesetzten Kommission, die darüber zu entscheiden hatte, ob Newton oder Leibniz, die Analysis "erfunden" hätte. "Seine" berühmten Taylor–Reihen wurden, laut Taylor selbst, durch eine Bemerkung von Machin über "Sir Isaac Newton's series" in *Child's Coffeehouse* motiviert (nicht vergessen, das war etwa 1710 und da gab es weder Starbuck's noch Coffee Bay).

was nur mit $\nabla f(x) = 0$ zu erfüllen ist. Damit ist aber dann

$$\frac{t^2}{2} y^T \left(\nabla f(\xi)\right) y = f(x + ty) - f(x) \geq 0$$

für alle hinreichend kleinen t, das heißt $y^T \left(\nabla f(\xi)\right) y \geq 0$ und mit $t \to 0$ konvergiert ja $\xi \to x$. Die zweite Aussage folgt direkt aus der Taylorentwicklung und der Tatsache, daß die *strikte* positive Definitheit von $\nabla^2 f$ an der Stelle x die *strikte* positive Definitheit von $\nabla^2 f$ in einer ganzen Umgebung D von x impliziert. □

6.2 Nochmals Konvexität

Trotzdem sind Differenzierbarkeit und vor allem zweimalige Differenzierbarkeit von f schon *starke* Forderungen. Und eigentlich braucht man sie ja auch gar nicht immer um Minima zu beschreiben.

Definition 6.3 *Für $x \in \mathbb{R}^n$ und eine in x richtungsdifferenzierbare Funktion f bezeichnen wir mit*

$$G[f, x] \; : \; \mathbb{R}^n \to \mathbb{R}, \qquad G[f, x](y) := D_y f(x)$$

*die **Gateaux–Variation** von f an der Stelle x.*

Übung 6.1 Zeigen Sie: Die Funktion $f(x) = \|x\|_\infty$ ist an der Stelle $x = 0$ richtungsdifferenzierbar, aber ihre Gateaux–Variation ist dort unstetig. Zumindest, wenn $n > 1$ ist. ◊

Proposition 6.4 *Sei $f \in C(\mathbb{R}^n)$ richtungsdifferenzierbar in x. Ist x ein lokales Minimum von f, dann ist*

$$D_y f(x) \geq 0, \qquad y \in \mathbb{R}^n. \tag{6.2}$$

Beweis: Auch noch ganz einfach! Ist x ein lokales Minimum, dann ist $f(x) \leq f(x + ty)$ für alle $y \in \mathbb{R}^n$ und alle hinreichend kleinen[170] $t > 0$. Also ist auch

$$0 \leq \frac{f(x + ty) - f(x)}{t} \quad \to \quad D_y f.$$

□

Zur Erinnerung: Eine Funktion f heißt **konvex**, wenn

$$f\left(\alpha x + (1 - \alpha)x'\right) \leq \alpha f(x) + (1 - \alpha) f\left(x'\right), \qquad x, x' \in \mathbb{R}^n, \quad \alpha \in [0, 1].$$

Lokale Minima x konvexer Funktionen sind immer *globale* Minima! Wäre nämlich $f\left(x'\right) < f(x)$ für irgendein $x' \in \mathbb{R}^n$, dann ist für $\alpha \in (0, 1)$

$$f\left(\alpha x + (1 - \alpha)x'\right) \leq \alpha f(x) + (1 - \alpha) f\left(x'\right) < \alpha f(x) + (1 - \alpha) f(x) = f(x),$$

was für $\alpha \to 1$ gegen $f(x)$ konvergiert. Außerdem sind konvexe Funktionen in gewissem Sinne "differenzierbar".

[170] Zeigen Sie: t hängt von y ab. Naja, eigentlich eher von $\|y\|$.

Proposition 6.5 *Sei* $f \in C(\mathbb{R}^n)$ *konvex. Dann ist* f *an jeder Stelle* $x \in \mathbb{R}^n$ *richtungsdifferenzierbar und die Gateaux–Variation von* f *in* x *ist*

1. ***positiv homogen**, d.h.,*

$$G[f,x](\alpha\cdot) = \alpha\, G[f,x](\cdot), \qquad \alpha > 0.$$

2. ***sublinear**, d.h.*

$$G[f,x](y+y') \le G[f,x](y) + G[f,x](y').$$

Beweis: Zu $x, y \in \mathbb{R}^n$ setzen wir

$$\phi(t) = \frac{f(x+ty)-f(x)}{t}, \qquad t \in (0,1],$$

und betrachten $\lim_{t\to 0^+}\phi(t)$. Mit $x = \frac{x+ty}{t+1} + \frac{t(x-y)}{t+1}$ und daher

$$f(x) \le \frac{1}{t+1} f(x+ty) + \frac{t}{t+1} f(x-y)$$

und

$$\frac{t}{t+1} f(x) \le \underbrace{\frac{1}{t+1}(f(x+ty)-f(x))}_{=\frac{t}{t+1}\phi(t)} + \frac{t}{t+1} f(x-y)$$

erhalten wir zuerst einmal, daß ϕ auf $(0,1]$ unabhängig von t durch $f(x)-f(x-y)$ nach unten beschränkt ist. Ist außerdem $0 < s \le t \le 1$, dann ist

$$\underbrace{f(x+sy)-f(x)}_{=s\,\phi(s)} = f\underbrace{\left(\frac{s}{t}(x+ty)+\frac{t-s}{t}x\right)}_{=x} - f(x) \le \underbrace{\frac{s}{t}f(x+ty)-\frac{s}{t}f(x)}_{s\,\phi(t)},$$

also $\phi(s) \le \phi(t)$ und da ϕ einen monoton steigende Funktion ist, die nach unten beschränkt ist, muß

$$D_y f(x) = \lim_{t\to 0^+}\phi(t)$$

existieren. Außerdem ist[171]

$$D_{\alpha y} f(x) = \lim_{t\to 0^+}\frac{f(x+t(\alpha y))-f(x)}{t} = \lim_{t\to 0^+}\alpha\frac{f(x+(\alpha t)y)-f(x)}{\alpha t} = \alpha\, D_y f(x)$$

sowie

$$\begin{aligned} D_{y+y'} f(x) &= \lim_{t\to 0^+}\frac{f(x+t(y+y'))-f(x)}{t} \\ &\le \lim_{t\to 0^+}\frac{\frac{1}{2}f(x+2ty)+\frac{1}{2}f(x+2ty')-f(x)}{t} \\ &= \lim_{t\to 0^+}\left(\frac{f(x+2ty)-f(x)}{2t}+\frac{f(x+2ty')-f(x)}{2t}\right) = D_y f(x) + D_{y'} f(x). \end{aligned}$$

[171] Und hier braucht man noch nicht einmal Konvexität.

□

Übung 6.2 Zeigen Sie: Ist $f \in C(\mathbb{R}^n)$ konvex, dann ist auch $G[f, x]$ konvex, $x \in \mathbb{R}^n$. ◊

Wir sehen also, daß konvexe Funktionen in der Optimierung wieder einmal eine ganz ausgezeichnete Rolle spielen.

6.3 Abstiegsverfahren – die allgemeine Idee

Wir betrachten jetzt zuerst einmal „glatte" Minimierungsprobleme und nehmen an, daß

1. für vorgegebenes $x \in \mathbb{R}^n$ die **Niveaumenge**

$$F := \{x' \in \mathbb{R}^n \ : \ f(x') \leq f(x)\}$$

   kompakt ist.

2. die **Zielfunktion** f auf einer offenen Umgebung D von F stetig differenzierbar ist.

3. der Gradient ∇f auf F **Lipschitz–stetig** ist, das heißt, es gibt $\gamma_f > 0$, so daß

$$\left\|\nabla f(x) - \nabla f(x')\right\|_2 \leq \gamma_f \|x - x'\|_2$$

Diese Voraussetzungen sind in gewissem Sinne Minimalvoraussetzungen, um „vernünftige" Iterationsverfahren konstruieren zu können. Die Differenzierbarkeit brauchen wir, um eine Richtung zu finden, in der wir uns verbessern können, die Kompaktheit von F_x gibt uns die Hoffnung, irgendwann zumindest bei einem lokalen Minimum anzukommen und die Lipschitz–Stetigkeit sorgt dafür, daß f „praktisch" C^2 ist und so nicht allzuviel Unsinn anstellt.

Definition 6.6 *Ein Punkt $x \in \mathbb{R}^n$ heißt **stationärer Punkt** von f, wenn $\nabla f(x) = 0$ ist.*

Nun zu unserem Verfahren. Ist x kein stationärer Punkt, d.h. ist $\nabla f(x) \neq 0$, dann gibt es Richtungen, so daß $D_y f(x) < 0$ ist. Jede derartige Richtung bezeichnet man als **Abstiegsrichtung** Abstiegsrichtungen haben das Potential, daß $f(x + ty)$ für hinreichend kleine Werte von t kleiner als $f(x)$ wird. Was man tatsächlich auch beweisen kann.

Lemma 6.7 *Sei $x \in F$ und $y \in \mathbb{R}^n$ so gewählt, daß $D_y f(x) < 0$ ist. Dann gibt es einen Wert $t > 0$, so daß*

$$f(x + ty) < f(x).$$

Beweis: Wir geben fast sogar ein quantitatives Resultat, indem wir zeigen, daß

$$f(x+ty) - f(x) \le t\left(D_y f(x) + t\frac{\gamma}{2}\|y\|_2^2\right). \tag{6.3}$$

Das folgt wieder mal aus einer Taylor–Entwicklung

$$\begin{aligned} f(x+ty) - f(x) &= t\,D_y f(x) + \int_0^t \underbrace{(D_y f(x+sy) - D_y f(x))}_{=(\nabla f(x+sy)-\nabla f(x))^T y}\,ds \\ &\le t\,D_y f(x) + \gamma_f \int_0^t \|sy\|_2\,\|y\|_2\,ds = t\,D_y f(x) + t^2\frac{\gamma}{2}\|y\|_2^2 \end{aligned}$$

Wählt man nun in (6.3)

$$0 < t < \frac{2\,|D_y f(x)|}{\gamma\,\|y\|_2^2},$$

dann ist die rechte Seite von (6.3), was die Behauptung beweist. □

Wir haben uns also mit dem folgenden (Doppel-)Problem auseinanderzusetzen:

> Wie wählt man zu gegebenem x
>
> 1. eine **Abstiegsrichtung** y
> 2. eine **Schrittweite** t
>
> so daß
>
> $$f(x+ty) < f(x)$$
>
> ist.

Wie man dann den Prozess auch noch zum Konvergieren bekommt, muss dann noch separat geklärt werden.

6.4 Abstiegsrichtungen – der naive Ansatz

Die erste Idee, die man haben könnte, besteht darin, als Abstiegs*richtung* y, $\|y\|_2 = 1$, diejenige Richtung zu wählen, für die $D_y f(x)$ minimal wird, und das ist natürlich der Wert

$$y = -\frac{\nabla f(x)}{\left\|\nabla f(x)\right\|_2}. \tag{6.4}$$

Und die Schrittweite könnte man ja so wählen, daß man das Minimum auf der *ganzen* Geraden $x+ty$, $t \in \mathbb{R}$, bestimmt. Dieses Minimum zeichnet sich dadurch aus, daß

$$0 = \frac{d}{dt} f(x+ty) = D_y f(x+ty) = (\nabla f(x+ty))^T\, y.$$

Wählt man nun das kleinste positive t mit dieser Eigenschaft, dann muß es zu einem Minimum gehören, da y ja eine Abstiegsrichtung ist. Diese Nullstelle von $\phi(t) := D_y f(x+ty)$, $t \in \mathbb{R}$, könnte man dann mit einem **Newton–Verfahren**, das mit dem Punkt x, also mit $t = 0$, gestartet wird, ermitteln[172]. Das geht gut,

[172] Oder zu ermitteln versuchen.

solange $\phi(0) \geq 0$ ist, da dann, wegen $\phi'(0) < 0$,

$$t_1 = 0 - \frac{\phi(0)}{\phi'(0)} > 0$$

ist und wir zumindest schon mal in der richtigen Richtung anfangen. Diese Schrittweitenwahl bezeichnet man als **exakte Schrittweite**. Und zumindest in einem Fall kann man die exakte Schrittweite auch einfach berechnen.

Beispiel 6.8 *Ist*

$$f(x) = \frac{1}{2}x^T Ax - b^T x + c, \qquad A \in \mathbb{R}^{n\times n},\ b \in \mathbb{R}^n, c \in \mathbb{R}, \tag{6.5}$$

mit einer symmetrischen, positiv definiten Matrix A*, dann ist*

$$\nabla f(x) = Ax - b \qquad \Rightarrow \qquad y = b - Ax.$$

Da

$$f(x+ty) = \frac{1}{2}(x+ty)^T A(x+ty) - b^T(x+ty) = \underbrace{x^T Ax - b^T x}_{=f(x)} + \frac{t^2}{2}\, y^T Ay + t\, x^T Ay - t\, b^T y,$$

erhalten wir somit die Bedingung

$$0 = \frac{d}{dt} f(x+ty) = ty^T Ay + y^T(Ax - b)$$

also

$$t = \frac{y^T(b - Ax)}{y^T Ay} = \frac{\|Ax - b\|_2^2}{(b - Ax)^T A(b - Ax)}.$$

Dieses Beispiel ist nicht ganz unbedeutend, ganz im Gegenteil: Ist $f \in C^2(\mathbb{R}^n)$ und sind wir nahe genug an einem *strikten* lokalen Minimum, dann lässt sich f – Taylor sei Dank – in einer hinreichend kleinen Umgebung immer durch so eine quadratische Parabel annähern. Schreiben wir nämlich

$$f(x) \sim \frac{1}{2}\left(x - x^*\right)^T \nabla^2 f\left(x^*\right)\left(x - x^*\right) + \left(x - x^*\right)^T \nabla f\left(x^*\right) + f\left(x^*\right), \tag{6.6}$$

dann sind wir, zumindest in einer gewissen Umgebung von x^*, in genau der quadratischen Situation. Und ist x^* ein *striktes* lokales Minimum, dann ist ja, nach Proposition 6.2, auch die von Haus aus symmetrische Matrix $\nabla^2 f\left(x^*\right)$ strikt positiv definit. Wir könnten also zu einem Startwert $x^{(0)}$ eine Folge von Näherungslösungen *iterativ* über

$$x^{(j+1)} = x^{(j)} - \frac{\nabla^T f\left(x^{(j)}\right) \nabla f\left(x^{(j)}\right)}{\nabla^T f\left(x^{(j)}\right) \nabla^2 f\left(x^{(j)}\right) \nabla f\left(x^{(j)}\right)} \nabla f\left(x^{(j)}\right), \qquad j \in \mathbb{N}_0,$$

bestimmen.

Allerdings ist aber das Verfahren des steilsten Abstiegs aber leider nicht immer das Mittel der Wahl – naiv geht's eben nicht immer. Denn leider kann es sehr schnell passieren, daß dieses Verfahren beliebig langsam, das heißt numerisch gar nicht, konvergiert. Um das zu verstehen setzen wir

$$f(x) = \frac{1}{2}(Ax-b)^T A^{-1}(Ax-b) = \frac{1}{2}x^TAx - b^Tx + b^TA^{-1}b,$$

sorgen also dafür, daß der Minimalwert gerade Null ist, und erhalten das folgende Resultat.

Lemma 6.9 *Es seien $\lambda_1 \le \ldots \le \lambda_n$ die Eigenwerte von A. Dann ist*

$$f\left(x^{(j+1)}\right) \le \left(1 - \frac{\lambda_1}{\lambda_n}\right) f\left(x^{(j)}\right), \qquad j \in \mathbb{N}_0. \tag{6.7}$$

Beweis: Für $j \in \mathbb{N}_0$ und $y = b - Ax^{(j)}$ ist

$$\begin{aligned}
f\left(x^{(j+1)}\right) &= f\left(x^{(j)}\right) + \frac{1}{2}\left(\frac{y^Ty}{y^TAy}\right)^2 y^TAy + \frac{y^Ty}{y^TAy} y^T \underbrace{\left(Ax^{(j)} - b\right)}_{=-y} \\
&= f\left(x^{(j)}\right) - \frac{1}{2}\frac{(y^Ty)^2}{y^TAy} = f\left(x^{(j)}\right) - \frac{1}{2}\frac{\|y\|_2^4}{y^TAy} \le f\left(x^{(j)}\right) - \frac{1}{2}\frac{\|y\|_2^4}{\lambda_n \|y\|_2^2} \\
&= f\left(x^{(j)}\right) - \frac{1}{2}\frac{y^Ty}{\lambda_n} = \frac{1}{2}y^T\left(A^{-1} - \frac{1}{\lambda_n}I\right)y \le \frac{1}{2}y^T\left(A^{-1} - \frac{\lambda_1}{\lambda_n}A^{-1}\right)y \\
&= \left(1 - \frac{\lambda_1}{\lambda_n}\right) f\left(x^{(j)}\right).
\end{aligned}$$

Der letzte Schritt ist wegen $x^TA^{-1}x \le \lambda_1^{-1}x^Tx$ richtig[173]. □

Übung 6.3 Zeigen Sie: Ist $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit und ist λ der größte Eigenwert von A, dann ist

$$x^TAx \le \lambda \|x\|_2^2, \qquad x \in \mathbb{R}^n.$$

♢

Dieses Verhalten kann man auch praktisch in `Matlab` sehen: man startet mit einer zufälligen $n \times n$–Matrix `A = rand( n )` und startet die Methode des steilsten Abstiegs mit einer Matrix der Form

```
A' * A + t eyes( n )
```

für verschiedene Werte von `t`> 0. Je größer `t` ist, desto besser wird die Methode funktionieren und konvergieren, für `t` $= 0$ hingegen kann man normalerweise nicht mehr von Konvergenz sprechen (die Matrix wird fast singulär). Warum das so ist und was die geometrische Interpretation ist, sieht man einfach am folgenden Beispiel.

[173] Der größte Eigenwert von A^{-1} ist λ_1^{-1}.

Beispiel 6.10 *Wir betrachten*

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 10^{-3} \end{bmatrix} \qquad \textit{und} \qquad b = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Die Lösung ist natürlich $x = [1, 1000]^T$. *Für Vektoren* $x = [x_1, x_2]^T$ *mit moderatem* x_2 *ist dann*

$$y = -\nabla f(x) = b - Ax = \begin{bmatrix} 1 - x_1 \\ 1 - 10^{-3} x_2 \end{bmatrix} \approx \begin{bmatrix} 1 - x_1 \\ 1 \end{bmatrix}$$

und

$$t := \frac{y^T y}{y^T A y} \sim \frac{1 + (1 - x_1)^2}{(1 - x_1)^2}.$$

Ist nun $x_1 \sim 0$ *oder* $x_1 \sim 2$, *dann ist* $\alpha \sim 2$ *und damit ist*

$$x + ty \sim \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + 2 \begin{bmatrix} 1 - x_1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 - x_1 \\ x_2 + 2 \end{bmatrix}$$

Starten wir also mit $x^{(0)} = 0$, *so werden die Werte* $x_1^{(j)}$ *anfangs zwischen* 0 *und* 2 *pendeln*[174] *und die Werte* $x_2^{(j)} \sim 2j$ *sein. Kommt dann das Verfahren richtig "in Fahrt" (also gegen Ende des Verfahrens), dann wird auch die Konvergenz deutlich schneller. Geometrisch bedeutet dies, daß sich das Verfahren an "flachgedrückten" Ellipsen entlanghangelt, siehe Abb. 6.1.*

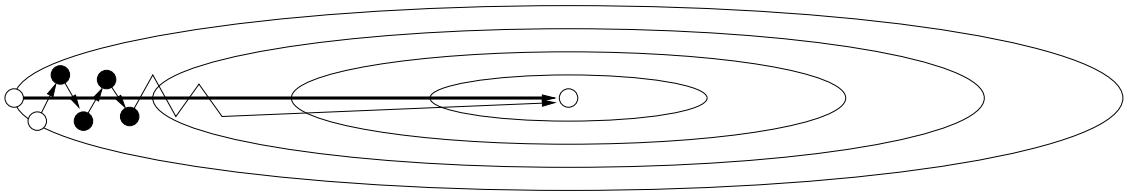

Abbildung 6.1: Verfahren des steilsten Abstiegs für Beispiel 6.10 (nicht maßstabsgetreu). Beachte: mit dem "richtigen" Startwert (nicht ganz zufällig ein Eigenvektor von A) würde das Verfahren nach *einem* Schritt erfolgreich terminieren.

Übung 6.4 Implementieren Sie das Verfahren des steilsten Abstiegs für Probleme der Form (6.5) in `Matlab`. ♢

Es gibt noch ein interessantes Experiment: Verwendet man eine modifizierte Version des Verfahrens des steilsten Abstiegs, bei der der negative Gradient um etwa 5 % *zufällig* gestört wird, dann erhält man Konvergenz in signifikant *weniger* Schritten.

[174] In Wirklichkeit sind sie stets etwas größer als 0 und etwas kleiner als 2 und dieses "etwas" wächst.

6.5 Abstiegsrichtungen – konjugierte Gradienten

Wie das Beispiel mit den „plattgedrückten“ Ellipsen zeigt, sind die *steilsten* Abstiegsrichtungen nicht unbedingt optimal, man muß offensichtlich auch Terme zweiter Ordnung berücksichtigen, die das Problem und die dahinterliegende Metrik verzerren können.

Definition 6.11 *Sei $A \in \mathbb{R}^{n\times n}$ eine symmetrische, positiv definite*[175] *Matrix. Zwei Vektoren x, y heißen* ***konjugiert*** *bezüglich A, wenn $x^T Ay = 0$ ist. Entsprechend heißt eine endliche Menge $X \subset \mathbb{R}^n$ konjugiert zu x,, wenn*

$$x^T AX = 0, \qquad d.h. \qquad x^T Ax' = 0, \quad x' \in X.$$

Übung 6.5 Sei $A \in \mathbb{R}^{n\times n}$ eine symmetrische, positiv definite Matrix. Zeigen Sie, daß

$$\langle x, y\rangle_A := x^T Ay \qquad \text{und} \qquad \|x\|_A := \langle x, x\rangle_A^{1/2}$$

ein Skalarprodukt und eine Norm definieren. ◊

Bemerkung 6.12

1. *Unter Berücksichtigung von Übung 6.5 bedeutet Konjugiertheit von Vektoren also eigentlich nichts anderes als* ***Orthogonalität****, nur eben nicht bezüglich des* Standard–Skalarprodukts $\langle x, y\rangle = x^T y$*, sondern bezüglich des Skalarprodukts $\langle\cdot,\cdot\rangle_A$.*

2. *Da orthogonale Vektoren immer* ***linear unabhängig*** *sind*[176] *und da Konjugiertheit wie eben gesehen nichts anderes als Orthogonalität ist, sind also auch konjugierte Vektoren immer linear unabhängig.*

Diese Beobachtungen legen es nahe, unser Optimierungsproblem zur Abwechslung mal als ein **Approximationsproblem** aufzufassen – eigentlich sind ja Approximationsproblem auch umgekehrt nur spezielle Optimierungsprobleme. Was uns hier interessiert ist das folgende Problem:

> Sei $X \subset \mathbb{R}^n$ ein Unterraum[177] mit $\dim X = m \le n$ und sei $\langle\cdot,\cdot\rangle$ ein **Skalarprodukt**[178], sowie $\|\cdot\|$ die dadurch induzierte Norm. Zu $y \in \mathbb{R}^n$ bestimme man $x^* \in X$, so daß
>
> $$\|y - x^*\| = \min_{x\in X} \|y - x\|\,. \tag{6.8}$$

[175] Das heißt, es ist $x^T Ax > 0$ für alle $x \in \mathbb{R}^n \setminus \{0\}$. Manchmal wird das als **strik positiv definit** bezeichnet.

[176] Und wer's nicht glaubt, darf es gerne selbst beweisen. Wobei man in der Mathematik ja nie etwas unbewiesen glauben sollte.

[177] Selbstverständlich kann man das Problem auch auf unendlichdimensionalen (Funktionen)–Räumen betrachten.

[178] Also eine symmetrische, definite Bilinearform, oder, was hier dasselbe ist, eine nichtentartete Sesquilinearform, siehe z.B. (Brieskorn, 1985, S. 314).

Die Lösung dieses Problems ist recht einfach, denn Normen, die von Skalarprodukten herrühren sind sehr stark mit Orthogonalität verknüpft, mit deren Hilfe man die Lösungen dann auch explizit angeben kann[179].

Lemma 6.13 *Sei* $X \subset \mathbb{R}^n$ *mit* $\dim X = m$. *Dann ist für jedes* $y \in \mathbb{R}^n$

$$\|y - x^*\| = \min_{x \in X} \|y - x\| \qquad \Leftrightarrow \qquad \langle y - x^*, X \rangle = 0. \tag{6.9}$$

Ist darüberhinaus $\{x_1, \ldots, x_m\}$ *eine* ***Orthonormalbasis***[180] *von* X, *dann ist*

$$x^* = \sum_{j=1}^{m} \langle y, x_j \rangle \, x_j \tag{6.10}$$

die Minmallösung von (6.8).

Bemerkung 6.14 *Bevor wir an den Beweis von Lemma 6.13 herangehen, sollten wir uns erst einmal klarmachen, warum es so hilfreich für unser Problem ist und warum die konjugierten Vektoren auftauchen: Die konjugierten Vektoren sind ja gerade* ***orthogonal*** *bezüglich des Skalarprodukts* $\langle \cdot, \cdot \rangle_A$, *so daß wir mit ihrer Hilfe ohne Probleme die Minimallösungen von (6.8) bestimmen können.*

Beweis von Lemma 6.13: Sei $\{x_1, \ldots, x_n\}$ eine Orthonormalbasis von X und sei $y \in \mathbb{R}^n$. Dann ist, für jede Wahl von Zahlen $a_1, \ldots, a_m \in \mathbb{R}$,

$$\begin{aligned}
\left\| y - \sum_{j=1}^{m} \langle y, x_j \rangle \, x_j \right\|^2 &= \left\langle y - \sum_{j=1}^{m} \langle y, x_j \rangle \, x, y - \sum_{j=1}^{m} \langle y, x_j \rangle \, x_j \right\rangle \\
&= \langle y, y \rangle - 2 \sum_{j=1}^{m} \langle y, x_j \rangle \langle y, x_j \rangle + \sum_{j,k=1}^{m} \langle y, x_j \rangle \langle y, x_k \rangle \underbrace{\langle x_j, x_k \rangle}_{\delta_{jk}} \\
&= \|y\|^2 - \sum_{j=1}^{n} \langle y, x_j \rangle^2 \leq \|y\|^2 - \sum_{j=1}^{m} \langle y, x_j \rangle^2 + \sum_{j=1}^{m} (a_j - \langle y, x_j \rangle)^2 \\
&= \|y\|_2^2 - 2 \sum_{j=1}^{m} a_j \langle y, x_j \rangle + \sum_{j=1}^{m} a_j^2 \langle x_j, x_j \rangle = \left\| y - \sum_{j=1}^{m} a_j \, x_j \right\|^2,
\end{aligned}$$

weswegen die Bestapproximation gerade der Fall $a_j = \langle y, x_j \rangle$, $j = 1, \ldots, m$ ist. Und das ist für $k = 1, \ldots, m$ äquivalent zu

$$\left\langle y - \sum_{j=1}^{m} \langle y, x_j \rangle \, x_j, x_k \right\rangle = \langle y, x_k \rangle - \sum_{j=1}^{m} \langle y, x_j \rangle \langle x_j, x_k \rangle = \langle y, x_k \rangle - \langle y, x_k \rangle = 0.$$

□

[179] Das macht dann auch die *Approximation in Hilberträumen* zu einem vergleichsweise leichten Problem, beispielsweise hat man keine Probleme mit Existenz und *Eindeutigkeit* der **Bestapproximation**.

[180] Und eine solche kann man ja über das **Gram–Schmidt–Verfahren** immer konstruieren.

Nun erinnern wir uns kurz daran, daß sich die eindeutige Minimalstelle von $f(x) = \frac{1}{2}x^T Ax - b^T x$ dadurch auszeichnet, daß[181]

$$0 = \nabla f(x) = Ax - b \qquad \Leftrightarrow \qquad Ax = b \quad \text{oder} \quad x = A^{-1}b.$$

Wären jetzt also $p_1, \ldots, p_n \in \mathbb{R}^n$ eine **Orthogonalbasis**[182] bezüglich $\langle \cdot, \cdot \rangle_A$, dann besteht die Idee darin, ausgehend von $x^{(0)}$ die Werte $x^{(k)} = x^{(0)} + a_1 p_1 + \cdots + a_k p_k$ so zu wählen, daß

$$\left\| x^{(k)} - A^{-1}b \right\| = \min \left\{ \left\| x - A^{-1}b \right\| \; : \; x \in x^{(0)} + \operatorname{span}\{p_1, \ldots, p_k\} \right\}$$

oder

$$\left\| \xi^{(k)} - \left(x^{(0)} - A^{-1}b \right) \right\| = \min \left\{ \left\| \xi - \left(x^{(0)} - A^{-1}b \right) \right\| \; : \; \xi \in \operatorname{span}\{p_1, \ldots, p_k\} \right\}$$

ist – wir lösen also unser Gleichungssystem $Ax = b$, indem wir sukzessive minimieren, denn das geht nach Lemma 6.13 ja ganz einfach: mit $y = x^{(0)} - A^{-1}b$ ist

$$\begin{aligned} \xi^{(k)} &= \sum_{j=1}^{k} \frac{\left\langle x^{(0)} - A^{-1}b, p_j \right\rangle_A}{\langle p_j, p_j \rangle_A} p_j = \sum_{j=1}^{k-1} \frac{\left\langle x^{(0)} - A^{-1}b, p_j \right\rangle_A}{\langle p_j, p_j \rangle_A} p_j + \frac{\left\langle x^{(0)} - A^{-1}b, p_k \right\rangle_A}{\langle p_k, p_k \rangle_A} p_k \\ &= \xi^{(k-1)} + \frac{\left\langle x^{(0)} - A^{-1}b, p_k \right\rangle_A}{\langle p_k, p_k \rangle_A} p_k, \end{aligned}$$

also

$$x^{(k)} = x^{(0)} - \xi^{(k)} = \underbrace{x^{(0)} - \xi^{(k-1)}}_{= x^{(k-1)}} - \frac{\left\langle x^{(0)} - A^{-1}b, p_k \right\rangle_A}{\langle p_k, p_k \rangle_A} p_k. =: x^{(k-1)} + \alpha_k p_k.$$

In Optimierungsterminologie verwenden wir jetzt also für den Übergang von $x^{(k-1)}$ zu $x^{(k)}$ die **Abstiegsrichtung** p_k und die **Schrittweite**

$$\begin{aligned} \alpha_k &= -\frac{\left\langle x^{(0)} - A^{-1}b, p_k \right\rangle_A}{\langle p_k, p_k \rangle_A} = -\frac{\left\langle x^{(k-1)} - A^{-1}b, p_k \right\rangle_A}{\langle p_k, p_k \rangle_A} = -\frac{p_k^T \left(Ax^{(k-1)} - b \right)}{p_k^T A p_k} \\ &= -\frac{p_k^T r_{k-1}}{p_k^T A p_k}, \qquad r_{k-1} := Ax^{(k-1)} - b \end{aligned}$$

da $x^{(k-1)} - x^{(0)} \in \operatorname{span}\{p_1, \ldots, p_{k-1}\} \perp_A p_k$. Hätten wir also eine Basis aus konjugierten Vektoren, dann wären wir fertig; dann berechnen wir halt so eine Basis. Dazu nehmen wir an, wir hätten schon einen Punkt $x^{(k)}$ und **konjugierte Richtungen** $p_1, \ldots, p_k$ bestimmt und suchen nun eine dazu konjugierte Richtung, in die wir uns weiter verbessern könnten. Ein erster Versuch für eine Abstiegsrichtung wäre natürlich nun wieder der *steilste* Abstieg $-r_k = b - Ax^{(k)}$. Führt diese

[181] Das gute alte „Ableiten und gleich Null setzen".

[182] Wir verzichten aus den diversesten Gründen auf die Normierung $\langle p_j, p_j \rangle = 1$, $j = 1, \ldots, n$.

Richtung zu keiner Verbesserung, dann haben wir unser Minimum gefunden, ergibt sich hingegen eine Verbesserung, dann ist

$$r_k \notin \operatorname{span}\{p_j \,:\, j = 1, \ldots, k\},$$

aber leider auch (noch) nicht konjugiert. Auch kein Problem, dann setzen wir eben[183]

$$p_{k+1} := r_k - \sum_{j=1}^{k} \frac{\langle r_k, p_j\rangle_A}{\langle p_j, p_j\rangle_A}\, p_j, \tag{6.11}$$

und erhalten so, daß $\langle p_{k+1}, P_k\rangle_A = 0$, also unsere konjugierte Richtung. In Wirklichkeit ist das aber sogar noch einfacher.

Lemma 6.15 *Die Vektoren $p_0, \ldots, p_n$, generiert durch die Vorschrift*

$$p_{k+1} := r_k - \frac{\langle r_k, p_k\rangle_A}{\langle p_k, p_k\rangle_A}\, p_k, \qquad p_0 := 0, \tag{6.12}$$

sind konjugiert.

Beweis: Wir definieren die **Krylov–Räume**[184]

$$P_k := \operatorname{span}\{p_1, \ldots, p_k\}$$

und behaupten zuerst einmal, daß

$$P_k = \operatorname{span}\{r_j \,:\, j = 0, \ldots, k-1\} = \operatorname{span}\left\{A^j r_0 \,:\, j = 0, \ldots, k-1\right\}, \tag{6.13}$$

was man durch Induktion über k nachweist: Der Fall $k = 0$ ist die triviale Feststellung, daß $p_1 = r_0 = A^0 r_0$ und für den Schritt $k \to k+1$ verwenden wir die Rekursionsformel

$$\begin{aligned} r_k &= Ax^{(k)} - b = A\left(x^{(k-1)} + \alpha_k p_k\right) - b = Ax^{(k-1)} - b + \alpha_k Ap_k \\ &= \underbrace{r_{k-1}}_{\in P_k} - \frac{p_k^T r_{k-1}}{p_k^T A p_k} \underbrace{Ap_k}_{\in AP_k}, \end{aligned} \tag{6.14}$$

die direkt aus der Definition der $x^{(k)}$ und damit der r_k folgt und die uns zusammen mit der Induktionshypothese liefert, daß $r_k \in P_k + AP_k$, also, nach (6.12), auch $p_{k+1} \in P_k + AP_k$ und daher ist

$$P_{k+1} \subseteq \operatorname{span}\left\{A^j r_0 \,:\, j = 0, \ldots, k\right\},$$

woraus (6.13) aus einfachen Dimensionsgründen folgt – schließlich sind die Vektoren $p_1, \ldots, p_{k+1}$ ja linear unabhängig. Wegen Lemma 6.13, genauer, wegen (6.9), ist

$$r_k^T P_k = 0 \tag{6.15}$$

[183] Wenn das jemandem bekannt vorkommen sollte – stimmt! Das ist die Vorgehensweise wie beim aus der Linearen Algebra bekannten *Gram–Schmidt–Verfahren*.

[184] Warum die so heißen, ist nicht so wichtig (nach einem Herrn Krylov, der sie erfunden hat natürlich), aber den Namen sollte man mal gesehen haben, denn sie sind ein wichtigs Konzept in der numerischen Linearen Algebra.

und da $P_k \supset AP_{k-1}$ erhalten wir auch, daß

$$0 = r_k^T A P_{k-1} = \langle r_k, P_{k-1} \rangle_A. \tag{6.16}$$

□

Das fassen wir jetzt einmal in einem Algorithmus zusammen.

Algorithmus 6.16 *(Lineares CG–Verfahren[185])*
Gegeben: *Symmetrische, positiv definite Matrix* $A \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^n$.

1. *Wähle* beliebigen *Startwert* $x^{(0)} \in \mathbb{R}^n$.
2. *Setze* $p_0 = 0$.
3. *Für* $k = 1, 2, \ldots$
   (a) *Setze*
   $$r := Ax^{(k-1)} - b.$$
   (b) *Setze*
   $$p_k := r - \frac{r^T A p_{k-1}}{p_{k-1}^T A p_{k-1}} p_{k-1}.$$
   (c) *Setze*
   $$x^{(k)} = x^{(k-1)} - \frac{r^T p_k}{p_k^T A p_k} p_k.$$

Ergebnis: *Folge* $x^{(k)}$ *die gegen das Minimum von* f *konvergiert.*

Übung 6.6 Zeigen Sie: Das CG–Verfahren terminiert spätestens nach n Schritten, das heißt, $x^{(n)} = x^{(n+1)} = x^{(n+2)} = \cdots$ ◇

Übung 6.7 Implementieren Sie die konjugierten Gradienten in `Matlab` bzw. `octave`. ◇

Bemerkung 6.17 *Mathematisch–theoretisch terminiert das CG–Verfahren nach* n *Schritten mit einem Minimum von* $f(x) = \frac{1}{2}x^T Ax - b^T x$, *oder, äquivalent, mit einer Lösung von* $Ax = b$. *In der numerischen Praxis ist dem aber leider* nicht *so – allerdings, und hier ist die Überraschung, hat es sich gezeigt, daß die konjugierten Gradienten ein sehr gutes und stabiles* Iterationsverfahren *liefern, wenn man nicht mit dem* n*–ten Schritt aufhört, sondern einfach weiteriteriert.*

Jetzt aber zurück zu unserem allgemeinen Optimierungsproblem. Ist $f \in C^2(\mathbb{R}^n)$, so könnten wir die **CG–Bedingungen** „wörtlich" in

$$\begin{aligned} p_k &= \nabla f\left(x^{(k-1)}\right) - \frac{p_{k-1}^T \nabla^2 f\left(x^{(k-1)}\right) \nabla f\left(x^{(k-1)}\right)}{p_{k-1}^T \nabla^2 f\left(x^{(k-1)}\right) p_{k-1}} p_{k-1} \\ x^{(k)} &= x^{(k-1)} - \frac{p_k^T \nabla f\left(x^{(k-1)}\right)}{p_{k-1}^T \nabla^2 f\left(x^{(k-1)}\right) p_{k-1}} p_k \end{aligned}$$

umschreiben. Allerdings ist das noch nicht so ganz das, was wir wollen, denn

[185] Das englische Schlagwort ist „**Conjugate Gradients**", ein Verfahren, das auf Hestenes & Stiefel (Hestenes & Stiefel, 1952) zurückgeht.

1. Wir brauchen hier überall die zweite Ableitung, und die muß erst einmal existieren und berechenbar sein; unserer generellen Annahmen waren ja "nur" **Differenzierbarkeit** und **Lipschitz–Stetigkeit** der Ableitung.

2. Die Schrittweite, die in der Berechnungsvorschrift für $x^{(k)}$ auftaucht, ist eine *exakte* Schrittweite, die ohnehin nur für den Fall einer quadratischen Zielfunktion sinnvoll. Im nichtlinearen "Normalfall" braucht man hier wahrscheinlich sowieso etwas anderes.

Anders gesagt: die konjugierten Gradienten spielen eigentlich zur Bestimmung der **Abstiegsrichtung** eine Rolle, aber solange zweite Ableitungen darin auftauchen, bleibt ihr Nutzen beschränkt. Doch das kann man glücklicherweise ändern! Wegen (6.12) und (6.15) ist nämlich

$$\begin{aligned} r_{k-1}^T p_k &= r_{k-1}^T \left(r_{k-1} - \frac{\langle r_{k-1}, p_{k-1}\rangle_A}{\langle p_{k-1}, p_{k-1}\rangle_A} p_{k-1} \right) = r_{k-1}^T r_{k-1} - \frac{\langle r_{k-1}, p_{k-1}\rangle_A}{\langle p_{k-1}, p_{k-1}\rangle_A} \underbrace{r_{k-1}^T p_{k-1}}_{=0} \\ &= r_{k-1}^T r_{k-1}, \end{aligned}$$

also

$$\alpha_k = -\frac{r_{k-1}^T r_{k-1}}{\langle p_k, p_k\rangle_A}$$

und da, nach (6.14) $Ap_k = \alpha_k^{-1}(r_k - r_{k-1})$ gilt, ist, wieder mit (6.15),

$$\begin{aligned} \langle r_k, p_k\rangle_A &= r_k^T A p_k = \alpha_k^{-1} r_k^T (r_k - r_{k-1}) = -\frac{\langle p_k, p_k\rangle_A}{r_{k-1}^T r_{k-1}} \left(r_k^T r_k - \underbrace{r_k^T r_{k-1}}_{=0} \right) \\ &= \langle p_k, p_k\rangle_A \frac{r_k^T r_k}{r_{k-1}^T r_{k-1}} \end{aligned}$$

und somit, als direkte Folgerung aus (6.12)

$$p_{k+1} = r_k - \frac{r_k^T r_k}{r_{k-1}^T r_{k-1}} p_k. \tag{6.17}$$

In dieser Darstellung[186] taucht nun *keine* Matrix A und damit auch keine zweiten Ableitungen von f mehr auf, und wir können sie wieder direkt in

$$p_{k+1} = \nabla f\left(x^{(k)}\right) - \frac{\nabla^T f\left(x^{(k)}\right) \nabla f\left(x^{(k)}\right)}{\nabla^T f\left(x^{(k-1)}\right) \nabla f\left(x^{(k-1)}\right)} p_k \tag{6.18}$$

umschreiben, was uns das *nichtlineare* CG–Verfahren liefert. Damit haben wir also eine „vernünftige" Wahl der Abstiegsrichtung[187] gefunden – allerdings ist

[186] Die nebenbei noch effektiver und numerisch stabiler ist.

[187] Selbst wenn das Sprichwort „Runter kommen sie immer" durchaus seine Gültigkeit behält, so haben wir ja wohl doch gesehen, daß das „wie" eine ganz gewaltige Rolle spielen kann und wird.

nicht garantiert, daß diese konjugierten Richtungen auch wirklich Abstiegsrichtungen darstellen! Das ist nur der Fall, wenn auch die Schrittweitensteuerung geeignet ist. Trotzdem hat die Wahl (6.18) laut (Nocedal & Wright, 1999) sogar einen Namen, nämlich **Fletcher–Reeves–Methode**, siehe (Fletcher & Reeves, 1964). Eine andere Methode, die **Pollak–Ribière–Methode** bestimmt die neue **Suchrichtung**[188] als

$$p_{k+1} = \nabla f\left(x^{(k)}\right) - \frac{\nabla^T f\left(x^{(k)}\right)\left(\nabla f\left(x^{(k)}\right) - \nabla f\left(x^{(k-1)}\right)\right)}{\nabla^T f\left(x^{(k-1)}\right)\,\nabla f\left(x^{(k-1)}\right)}\, p_k. \tag{6.19}$$

Sie ist nach der Tabelle in (Nocedal & Wright, 1999, S. 124) wesentlich effektiver und braucht in vielen Fällen nur die Hälfte der Iterationen oder weniger[189] als die Fletcher–Reeves–Methode. Aber um das zu verstehen, brauchen wir etwas mehr Information über die Schrittweitensteuerung.

6.6 Wahl der Schrittweite

Bleibt noch das zweite Problem, nämlich die Bestimmung der **Schrittweite**. Dabei möchte man gerne zwei Fliegen mit einer Klappe schlagen, und zwar:

1. Die Schrittweite soll so *klein* sein, daß eine Verbesserung erzielt wird (siehe Lemma 6.7).
2. Die Schrittweite soll so *groß* sein, daß der neue Punkt $x^{(k+1)}$ sich möglichst signifikant von $x^{(k)}$ unterscheidet, denn sonst würde das Verfahren ja „numerisch stationär“ werden.

Die beste Wahl wäre natürlich die **exakte Schrittweite**, das heißt, wir suchen für einen Punkt $x \in \mathbb{R}^n$ und eine Abstiegsrichtung y nach der ersten Nullstelle von

$$\phi(t) = \frac{d}{dt} f(x + ty) = D_y f(x + ty).$$

Aber die Bestimmung dieser Nullstelle ist erstens schwierig[190], zweitens aufwendig und kann obendrein im allgemeinen nicht mit endlich vielen Schritten durchgeführt werden. Iterationen im Inneren von Iterationen sind immer eine äußerst problematische Angelegenheit[191], denn deren Verhalten beeinflusst ja auch ganz massiv die äußere Iteration. Deswegen verzichtet man auf die Exaktheit zugunsten der Effizienz und verwendet sogenannte *inexakte* Methoden zur Schrittweitensteuerung. Dazu wählt man Konstanten $0 < c_1 < c_2 < 1$ und fordert, daß die Schrittweite α eines der beiden folgenden Kriterien erfüllt:

[188] Nachdem Abstieg nicht sicher gewährleistet werden kann, erscheint mir dieser Begriff angemessener als „Abstiegsrichtung“.

[189] Das Minimum liegt bei $\frac{1}{5}$!

[190] Im allgemeinen gibt es kein Verfahren, die *nächstgelegene* Nullstelle einer Funktion zu ermitteln, weder Newton, noch Bisektion oder Regula Falsi haben hier eine Chance. In Spezialfällen geht das natürlich schon, beispielsweise, wenn ϕ konvex ist, aber wer will schon *dritte* Ableitungen von f berechnen?

[191] Man weiß nie wirklich, wann man aufhören soll, iteriert man zu lange, wird das Verfahren unerträglich lange, bricht man zu früh ab, sind die Ergebnisse, die das „äußere Verfahren“ braucht, nicht genau genug und dann kann alles passieren.

1. Die **Armijo–Bedingung**

$$f(x+\alpha y) - f(x) \le \alpha c_1 y^T \nabla f(x) \tag{6.20}$$

fordert, daß die Verbesserung[192], die man erzielt, proportional zur Länge der Richtungsableitung ist.

2. Die **Wolfe–Bedingungen**[193] oder auch **Powell–Bedingungen**[194]

$$\begin{aligned} f(x+\alpha y) - f(x) &\le \alpha c_1 y^T \nabla f(x), \\ y^T \nabla f(x+\alpha y) &\ge c_2 y^T \nabla f(x), \end{aligned} \tag{6.21}$$

verlangen außerdem, daß wir bis zu einem Punkt marschieren, an dem die Funktion weniger stark abfällt – so weit sollte man schon mindestens gehen.

3. Als **starken Wolfe–Bedingungen** bezeichnet man die Forderungen,

$$\begin{aligned} f(x+\alpha y) - f(x) &\le \alpha c_1 y^T \nabla f(x), \\ |y^T \nabla f(x+\alpha y)| &\le c_2 \, |y^T \nabla f(x)|, \end{aligned} \tag{6.22}$$

die verlangen, daß man nicht bis zu einem Punkt marschiert, an dem es zu steil "nach oben" geht.

Die erste Frage, die man sich natürlich stellt, ist, ob sich diese Forderungen überhaupt erfüllen lassen.

Lemma 6.18 *Ist* $f \in C^1(\mathbb{R}^n)$, $D_y f(x) < 0$ *und ist*

$$\inf\{\phi(t) := f(x+ty) \ : \ t \in \mathbb{R}_+\} > -\infty,$$

dann gibt es für alle $0 < c_1 < c_2 < 1$ *Werte von* $\alpha \in \mathbb{R}_+$, *die (6.21) bzw. (6.22) erfüllen*[195].

Beweis: Es sei $\ell(t) = f(x) + t\, c_1 D_y f(x)$, $t \in \mathbb{R}_+$, die lineare Funktion, die ϕ und ϕ' an der Stelle 0 interpoliert. Da $\ell(t) \to -\infty$ für $t \to \infty$ und da $\phi > M$ für ein $M \in \mathbb{R}$, gibt es einen *kleinsten* Wert $t' > 0$, so daß

$$f(x) + t'\, c_1 D_y f(x) = \ell(t') = \phi(t') = f(x+t'y)$$

und da $c_1 < 1$ ist, muß $\ell(t) \ge \phi(t)$ für alle $t \le t'$ sein, also

$$f(x+\alpha y) - f(x) \le \alpha c_1 y^T \nabla f(x), \qquad \alpha \in (0, t')$$

was nichts anderes als (6.20) ist. Nach dem Zwischenwertsatz existiert außerdem ein $t^* \in (0, t')$, so daß

$$\phi(t') - \phi(0) = (t' - 0)\ \phi'(t^*)$$

[192] Denn die linke Seite ist negativ!
[193] Nach (Wolfe, 1969).
[194] Nach (Powell, 1976).
[195] Und damit natürlich auch (6.20).

ist, also

$$t' y^T \nabla f(x + t^* y) = f(x + t' y) - f(x) = t' \underbrace{c_1}_{<c_2} \underbrace{y^T \nabla f(x)}_{<0} > t' c_2 y^T \nabla f(x).$$

Kürzen wir nun $t' > 0$, dann erhalten wir in einer Umgebung von t^* die zweite Bedingung von (6.21), aber auch von (6.22), denn die linke Seite der letzten Ungleichungskette ist ja auch negativ. □

Der Beweis von Lemma 6.18 sagt uns nicht nur, daß solche Werte von α, also schönen Schrittweiten, immer existieren, er gibt uns sogar ein Rezept, wie man sie berechnet! Und zwar machen wir das in zwei Schritten:

1. Wir setzen $\alpha_0 = 0$, starten mit einem (geratenen) Wert $\alpha_1 > 0$ und vergrößern ihn (z.B. durch Multiplikation mit $\rho > 1$) so lange, bis für den so erhaltenen Wert α_k eine der folgenden Bedingungen erfüllt ist.

   (a) Die Schrittweite passt:

   $$\phi(\alpha_k) - \phi(0) \le \alpha_k c_1 \phi'(0) \qquad \text{und} \qquad |\phi'(\alpha_k)| \le c_2 |\phi'(0)|.$$

   (b) Die Armijo Bedingung ist verletzt:

   $$\phi(\alpha_k) - \phi(0) > \alpha_k c_1 \phi'(0).$$

   (c) Unsere Richtung y ist zur Aufstiegsrichtung mutiert:

   $$\phi'(\alpha_k) \ge 0.$$

   (d) Der letzte Punkt war besser:

   $$\phi(\alpha_k) \ge \phi(\alpha_{k-1}).$$

   Passiert das schon für $k = 1$, dann war α_1 idiotisch gewählt und wir halbieren α_1 so lange, bis $\phi(\alpha_1) < \phi(0)$.

   Dieses Verfahren bricht irgendwann für ein k ab. Ist nicht gerade der erste Fall eingetreten, dann setzen wir $\alpha_- = \alpha_{k-1}$, $\alpha_+ = \alpha_k$.

2. Im Intervall (α_-, α_+) liegen nun zulässige Schrittweiten, die (6.22) erfüllen und die man über ein geeignetes **Bisektionsverfahren** finden kann.

Der Vorteil der Wolfe- oder Powell–Bedingungen liegt nun darin, daß man mit ihnen tatsächlich auch etwas über die Konvergenz des Abstiegsverfahrens sagen kann.

Satz 6.19 *Es sei* $f \in C^1(\mathbb{R}^n)$ *mit Lipschitz–stetigem Gradienten*[196] *nach unten beschränkt:*

$$\inf\{f(x) \ : \ x \in \mathbb{R}^n\} > -\infty.$$

[196] Das sind also unsere "Standardbedingungen" aus dem Anfang dieses Kapitels.

Bildet man zu einem Startwert $x^{(0)}$ die Folge

$$x^{(k+1)} = x^{(k)} + \alpha_k y^{(k)}, \qquad k \in \mathbb{N}_0,$$

so daß $D_{y^{(k)}} f\left(x^{(k)}\right) < 0$ und wählt man die α_k so, daß sie die Bedingung (6.21) erfüllen, dann gilt für jeden Startwert $x^{(0)}$, daß

$$\sum_{k\in\mathbb{N}_0} \cos^2\theta_k \left\|\nabla f\left(x^{(k)}\right)\right\|_2^2 < \infty, \tag{6.23}$$

wobei

$$\cos\theta_k := -\frac{\nabla^T f\left(x^{(k)}\right) y^{(k)}}{\left\|\nabla^T f\left(x^{(k)}\right)\right\| \left\|y^{(k)}\right\|}$$

den Winkel zwischen der Suchrichtung und dem negativen Gradienten[197] *beschreibt.*

Korollar 6.20 *Unter den Voraussetzungen von Satz 6.19 ist*

$$\lim_{k\to\infty} \cos\theta_k \left\|\nabla f\left(x^{(k)}\right)\right\|_2 = 0. \tag{6.24}$$

Bemerkung 6.21 *Leider ist (6.24) noch nicht ganz das, was wir wollen, denn wir erhalten nicht die Konvergenz der Gradienten gegen 0, das heißt, die Konvergenz der $x^{(k)}$ gegen ein lokales Minimum, sondern wir haben das Problem, daß die Winkel θ_k mit ins Spiel kommen. Schaffen wir es allerdings, zu gewährleisten, daß*

$$\inf_{k\in\mathbb{N}_0} |\cos\theta_k| > 0,$$

sind also die Abstiegsrichtungen hinreichend nichtorthogonal *zu den Gradienten, dann sieht die Sache anders aus.*

Beweis von Satz 6.19: Unter Verwendung der Abkürzung $\nabla f_k := \nabla f\left(x^{(k)}\right)$ erhalten wir aus (6.21) und der Lipschitz–Stetigkeit des Gradienten, daß

$$(c_2 - 1)\, \nabla^T f_k\, y^{(k)} \le \left(\nabla f_{k+1} - \nabla f_k\right)^T y^{(k)} \le \gamma \underbrace{\left\|x^{(k+1)} - x^{(k)}\right\|_2}_{=\alpha_k \|y_k\|_2} \left\|y^{(k)}\right\|_2 = \alpha_k \gamma \left\|y^{(k)}\right\|_2^2,$$

also

$$\alpha_k \ge \frac{c_2 - 1}{\gamma} \frac{\nabla^T f_k\, y^{(k)}}{\left\|y^{(k)}\right\|_2^2} > 0,$$

letzteres, da $c_2 < 1$ und $\nabla^T f_k\, y^{(k)} < 0$. Setzen wir das in die erste Ungleichung von (6.21) ein, dann ergibt sich, daß

$$f\left(x^{(k+1)}\right) \;=\; f\left(x^{(k)} + \alpha_k y^{(k)}\right) \le f\left(x^{(k)}\right) - \frac{c_1\,(1 - c_2)}{\gamma} \underbrace{\frac{\left(\nabla^T f_k\, y^{(k)}\right)^2}{\left\|y^{(k)}\right\|_2^2}}_{=\cos^2\theta_k \left\|\nabla f\left(x^{(k)}\right)\right\|_2^2}$$

[197] Also der Richtung des steilsten Abstiegs.

$$
\begin{aligned}
&= f\left(x^{(k)}\right) - \frac{c_1\,(1-c_2)}{\gamma}\left(\cos^2\theta_k \left\|\nabla f\left(x^{(k)}\right)\right\|_2^2\right)\\
&= f\left(x^{(0)}\right) - \frac{c_1\,(1-c_2)}{\gamma}\sum_{j=0}^{k}\cos^2\theta_k \left\|\nabla f\left(x^{(k)}\right)\right\|_2^2
\end{aligned}
$$

und da f nach unten beschränkt ist folgt (6.23). □

6.7 Nochmal konjugierte Gradienten

Man kann nun zeigen[198], daß mit den starken Wolfe-Bedingungen (6.22) und $0 < c_1 < c_2 < \frac{1}{2}$ das Verfahren von Fletcher & Reeves konvergiert, genauer, daß

$$\liminf_{k\to\infty}\left\|\nabla f\left(x^{(k)}\right)\right\|_2 = 0$$

ist. Trotzdem kann es zu Schwierigkeiten kommen, die bei Pollak–Ribière nicht auftreten. Dazu bemerkt man, daß für $c_2 < \frac{1}{2}$ und die Iterationsvorschrift (6.18) die Abschätzungen[199]

$$\underbrace{\frac{1-2c_2}{1-c_2}}_{=:a_1>0} \le -\frac{\nabla^T f\left(x^{(k)}\right) y^{(k)}}{\left\|\nabla^T f\left(x^{(k)}\right)\right\|_2^2} \le \underbrace{\frac{1}{1-c_2}}_{=:a_2>0}$$

gelten, also

$$a_1\frac{\left\|\nabla^T f\left(x^{(k)}\right)\right\|_2}{\left\|y^{(k)}\right\|_2} \le \cos\theta_k \le a_2\frac{\left\|\nabla^T f\left(x^{(k)}\right)\right\|_2}{\left\|y^{(k)}\right\|_2}.$$

Damit bedeutet der „schlechte“ Fall $\cos\theta_k \sim 0$ von Satz 6.19, daß $\left\|\nabla f\left(x^{(k)}\right)\right\|_2 \ll \left\|y^{(k)}\right\|_2$. Dann ist aber auch, wie im Beweis von Satz 6.19, $x^{(k+1)} \sim x^{(k)}$, also $\nabla f\left(x^{(k+1)}\right) \sim \nabla f\left(x^{(k)}\right)$ und damit $\alpha_{k+1} \sim 1$ und somit, nach (6.18),

$$y^{(k+1)} \sim \nabla f\left(x^{(k+1)}\right) + y^{(k)} \sim y^{(k)},$$

der Algorithmus läuft sich also fest! Und hier ist der Vorteil von (6.19): Sind bei zwei aufeinanderfolgenden Iterationsschritten die Gradienten (nahezu) gleich, dann wird die konjugierte Richtung verworfen und das Verfahren mit dem steilsten Abstieg neu gestartet – in den meisten Fällen eine gute Wahl.

Zum Abschluß aber noch ein nettes theoretisches Ergebnis ohne Beweis.

Satz 6.22 *Es gibt eine Funktion* $f \in C^2(\mathbb{R}^3)$ *und einen Startwert* $x^{(0)} \in \mathbb{R}^3$*, so daß für die Pollak–Ribière–Methode mit* exakter *Schrittweitenbestimmung*

$$\inf\left\{\left\|\nabla f\left(x^{(k)}\right)\right\|_2 \;:\; k \in \mathbb{N}_0\right\} > 0$$

ist.

[198] Siehe (Nocedal & Wright, 1999, Theorem 5.8, S. 128).
[199] Siehe (Nocedal & Wright, 1999, S 125).

Noch hat der Name Philosophie bei den Engländern allgemein diese Bestimmung, Newton hat fortdauernd den Ruhm des größten Philosophen; bis in die Preiskurante der Instrumentenmacher herab heißen diejenigen Instrumente, die nicht unter eine besondere Rubrik magnetischen, elektrischen Apparats gebracht werden, die Thermometer, Barometer usf. philosophische Instrumente; freilich sollte nicht eine Zusammensetzung von Holz, Eisen usf., sondern allein das Denken das Instrument der Philosophie genannt werden.

Georg Wilhelm Friderich Hegel,
Enzyklopädie der philosophischen Wissenschaften im Grundrisse

7 Newton–Verfahren und Variationen

Da sich lokale Extrema x^* einer Funktion $f \in C^1(\mathbb{R}^n)$ ja dadurch auszeichnen, daß $\nabla f(x^*) = 0$, können wir also unser Optimierungsproblem auch als die Suche nach einer Nullstelle von $F(x) = \nabla f(x)$, $F : \mathbb{R}^n \to \mathbb{R}^n$ auffassen. Dazu zuerst einmal der nötige Formalismus.

Definition 7.1 *Sei* $F = [F_j : j = 1, \ldots, n] \in C(\mathbb{R}^n)^n$ *ein Vektorfeld. Die* ***Jacobimatrix*** *zu* F *ist definiert als*

$$F' := J[F] := \left[\frac{\partial F_j}{\partial x_k} : j, k = 1, \ldots, n\right], \qquad F \in C^1(\mathbb{R}^n)^n.$$

Besonders einfach ist die Sache natürlich, wenn $F = \nabla f$, $f \in C^2(\mathbb{R}^n)$, denn dann ist

$$F'_{jk} = \frac{\partial}{\partial x_k} F_j = \frac{\partial}{\partial x_k}\frac{\partial f}{\partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_k} = \nabla^2_{jk} f, \qquad j, k = 1, \ldots, n,$$

also

$$F' = J[F] = J[\nabla f] = \nabla^2 f. \tag{7.1}$$

7.1 Das Newton–Verfahren und das Broyden–Verfahren

Zu einer Funktion $F \in C^1(\mathbb{R})$ erzeugt das Newton–Verfahren vermittels der Iterationsvorschrift

$$x^{(k+1)} = x^{(k)} - \frac{F\left(x^{(k)}\right)}{F'\left(x^{(k)}\right)}, \qquad k \in \mathbb{N}_0, \tag{7.2}$$

eine Folge $\left\{x^{(k)} \,:\, k \in \mathbb{N}\right\}$ von Punkten, die für einen günstig gewählten Startwert $x^{(0)}$ auch tatsächlich gegen eine **einfache Nullstelle** x^* mit

$$F(x^*) = 0, \qquad F'(x^*) \neq 0,$$

konvergiert. Für $F \in C^1\left(\mathbb{R}^n\right)^n$ verwendet man hingegen die Iteration

$$x^{(k+1)} = x^{(k)} - \left(F'\right)^{-1}\left(x^{(k)}\right) F\left(x^{(k)}\right), \qquad k \in \mathbb{N}_0, \tag{7.3}$$

allerdings ist jetzt der Begriff der "einfachen" Nullstelle etwas präziser zu fassen, es muß nämlich

$$F(x^*) = 0, \qquad \det F'(x^*) \neq 0$$

sein, die Matrix $F'(x^*)$ muß also **invertierbar** sein. Das führt dann auch dazu, daß für alle $0 \neq y \in \mathbb{R}^n$ die *vektorwertige* Richtungsableitung $D_y F = F'y$ von Null verschieden ist. Eine "typische" allgemeine Konvergenzaussage für das Newton–Verfahren sieht dann wie folgt aus[200].

Satz 7.2 *Ist* $F \in C^2\left(\mathbb{R}^n\right)^n$ *und ist* $x^* \in \mathbb{R}^n$ *eine* ***einfache Nullstelle*** *von* F*, das heißt,*

$$F(x^*) = 0 \qquad \textit{und} \qquad \det\, J[F](x^*) \neq 0, \tag{7.4}$$

dann gibt es eine offene Menge $U \subset \mathbb{R}^n$*,* $x^* \in U$*, so daß*

$$x^{(0)} \in U \qquad \Rightarrow \qquad \lim_{k\to\infty} x^{(k)} = x^*.$$

Nur zur Erinnerung: Iterationsverfahren, die für Startwerte konvergieren, die hinreichend nahe bei der gesuchten Lösung liegen (und dann aber auch gegen diese "naheliegende" Lösung!), bezeichnet man als **lokal konvergent**.

Neben den offensichtlichen Schwierigkeiten der nur lokalen Konvergenz gibt es noch ein weiteres Problem, das die praktische Anwendung des Newton–Verfahrens schwierig macht: die Bestimmung der Jacobimatrix $J[F]$! Schließlich kann man nicht unbedingt davon ausgehen, daß die Ableitungen aller Komponenten von F auch wirklich so einfach verfügbar sind. Möglichkeiten zur praktischen Bestimmung der Jacobimatrix wären

[200] Aus (Sauer, 2000), dort findet sich auch der Beweis, der auf Fixpunktiterationen und dem Banachschen Fixpunktsatz beruht.

Automatische Differentiation: (siehe z.B. (Nocedal & Wright, 1999, Chapter 7.2))
Funktionen werden intern als Kombination elementarer Funktionen dargestellt, deren Ableitungen bei der Auswertung mitberechnet werden:

$$f = gh \qquad \Rightarrow \qquad \nabla f = g\nabla h + h\nabla g$$

oder

$$f = g(h_1, \ldots, h_m) \qquad \Rightarrow \qquad \nabla f = \nabla g\, J[H], \quad H = [h_j : j = 1, \ldots, m].$$

Solche Schemata lassen sich sehr gut in C++ implementieren.

Numerische Differentiation: Man erhält die näherungsweisen Ableitungen durch Differenzenquotienten oder durch Differentiation von Interpolationspolynomen. So kann man für Punkte $\{x_k \in \mathbb{R}^n : k = 1, \ldots, N\}$ ein Polynom p bestimmen, so daß $f(x_j) = p(x_j)$ und dann $\nabla p(x)$ bestimmen. Der Differenzenquotient ist hierbei nur der Spezialfall $N = 2$. Allerdings ist das mit der Interpolation in mehreren Variablen nicht mehr so ganz einfach, siehe (Gasca & Sauer, 2000).

Trotzdem, beide Methoden zur Bestimmung von Ableitungen, insbesondere von höheren Ableitungen sind aufwendig und numerisch nicht immer stabil. Deswegen versucht man, die Bestimmung von Gradienten so weit es geht zu vermeiden. Das führt zu einer wichtigen Variante des Newton–Verfahrens, zum sogenannten **Broyden–Verfahren**, (Broyden, 1965). Dabei wird die Ableitung $F'\left(x^{(k)}\right)$ durch eine Matrix B_k angenähert und anstatt dann im nächsten Schritt die Matrix $F'\left(x^{(k+1)}\right)$ zu bestimmen, bestimmt man einen *näherungsweisen* "Update" B_{k+1}, der nur von der Richtung $y^{(k)}$ und den Werten von F abhängt. Außerdem spendiert man sich wieder eine Schrittweitensteuerung $\alpha_k \in \mathbb{R}_+$. Insgesamt sieht das Ganze dann folgendermaßen aus:

$$\begin{aligned} y &= B_k^{-1} F\left(x^{(k)}\right) \\ x^{(k+1)} &= x^{(k)} - \alpha_k y, \qquad d = F\left(x^{(k+1)}\right) - F\left(x^{(k)}\right) \\ B_{k+1} &= B_k + \frac{1}{y^T y}\left(d - B_k y\right) y^T. \end{aligned} \tag{7.5}$$

Die Transpositionszeichen in der Update–Regel für B_{k+1} sind tatsächlich korrekt: Der Übergang von B_k zu B_{k+1} erfolgt durch Addition einer Matrix der Form zy^T, also einer Matrix *vom Rang* 1. Laut (Stoer, 1983, S. 248) führt man die Iterationen nach der Regel (7.6) aus, wenn die Schrittweite α_k, die durch (näherungsweise, numerische) Lösung des Minimierungsproblems[201]

$$\left\| F\left(x^{(k)} - \alpha_k y\right)\right\|_2^2 = \min_{t \geq 0} \left\| F\left(x^{(k)} - t\, y\right)\right\|_2^2$$

[201] Preisfrage: Warum steht hier das Quadrat? Wer es immer noch nicht verstanden hat: *Gehe zurück zum Anfang des Skripts, gehe nicht über Los, ziehe keinen Schein ein.*

ermittelt wurde, die Bedingung $\frac{1}{2} \leq \alpha_k \leq 1$ erfüllt, ansonsten berechnet man zähneknirschend $B_{k+1} \sim F'\left(x^{(k+1)}\right)$, entweder numerisch oder mit automatischer Differentiation.
Der Grund für die Iterationsvorschrift (7.5) liegt im folgenden Modellproblem, ein Resultat das auf Broyden (Broyden, 1965) zurückgeht.

Proposition 7.3 *Sei* $F(x) = Ax + b$, $A \in \mathbb{R}^{n\times n}$ *und* $B \in \mathbb{R}^{n\times n}$ *eine beliebige Matrix. Für beliebige* $x, x' \in \mathbb{R}^n$ *sei* $y := x - x'$ *und* $d := F(x) - F(x') = Ay$

$$B' := B + \frac{1}{y^T y}\,(d - By)\,y^T.$$

Dann ist

$$\|B' - A\|_2 \leq \|B - A\|_2\,.$$

Beweis: Da $d = Ay$, ist

$$B' = B + (A - B)\,\frac{yy^T}{y^T y} = A\frac{yy^T}{y^T y} + B\left(I - \frac{yy^T}{y^T y}\right),$$

also

$$B' - A = (B - A)\left(I - \frac{yy^T}{y^T y}\right).$$

Schreiben wir nun einen beliebigen Vektor $u \in \mathbb{R}^n$ als $u = y + z$, $y \perp z$, dann ist

$$(B' - A)\,u = (B - A)\Big(z - \frac{1}{y^T y}y\underbrace{y^T z}_{=0}\Big) + (B - A)\overbrace{\Big(y - \underbrace{y\frac{1}{y^T y}y^T y}_{=1}\Big)}^{=0} = (B - A)\,z$$

und da $\|u\|_2^2 = \|y\|_2^2 + \|z\|_2^2$ wegen der Orthogonalität, ergibt sich, daß

$$\left\|(B' - A)\,u\right\|_2 = \left\|(B - A)\,z\right\|_2 \leq \|B - A\|_2\;\|z\|_2 \leq \|B - A\|_2\;\|u\|_2\,,$$

also $\|B' - A\|_2 \leq \|B - A\|_2$. □

7.2 Das Newton–Verfahren zur Minimumsbestimmung

Beschäftigen wir uns aber nun mit unserem "Spezialfall" $F = \nabla f$ und der Iterationsvorschrift

$$x^{(k+1)} = x^{(k)} - \left(\nabla^2 f\left(x^{(k)}\right)\right)^{-1}\,\nabla f\left(x^{(k)}\right) =: x^{(k)} + \alpha_k\,y^{(k)}, \qquad \alpha_k = -1, \qquad (7.6)$$

wobei man die wie in (7.6) gewählte Richtung $y^{(k)}$ als **Newton–Richtung** bezeichnet. Die Newton–Richtung muß übrigens keine Abstiegsrichtung sein. Trotzdem kann man nun Aussagen über (lokale) Konvergenz und Konvergenzgeschwindigkeit machen.

Satz 7.4 *Sei* $f \in C^2(\mathbb{R}^n)$ *und* $\nabla^2 f$ *Lipschitz–stetig in einer Umgebung eines* strikten *Minimums* x^* *von* f. *Dann gibt es eine Umgebung* U *von* x^*, *so daß die Iteration (7.6) für alle* $x^{(0)} \in U$

1. *gegen* x^* *konvergiert:*
$$\lim_{k\to\infty} x^{(k)} = x^*. \tag{7.7}$$

2. quadratisch *konvergiert:*
$$\sup_{k\in\mathbb{N}_0} \frac{\left\|x^{(k+1)} - x^*\right\|}{\left\|x^{(k)} - x^*\right\|^2} < \infty. \tag{7.8}$$

3. *quadratisch gegen Null konvergente Gradienten liefert:*
$$\sup_{k\in\mathbb{N}_0} \frac{\left\|\nabla f\left(x^{(k+1)}\right)\right\|}{\left\|\nabla f\left(x^{(k)}\right)\right\|^2} < \infty. \tag{7.9}$$

Beweis: Wir setzen wieder $\nabla f_k = \nabla f\left(x^{(k)}\right)$ und entsprechend auch $\nabla^2 f_k$ und ∇f_*. Nach (7.6) ist

$$\begin{aligned}
x^{(k+1)} - x^* &= x^{(k)} - x^* - \left(\nabla^2 f_k\right)^{-1} \nabla f_k \\
&= \left(\nabla^2 f_k\right)^{-1} \left(\nabla^2 f_k\left(x^{(k)} - x^*\right) - \nabla f_k + \underbrace{\nabla f_*}_{=0}\right) \\
&= \left(\nabla^2 f_k\right)^{-1} \left(\nabla^2 f_k\left(x^{(k)} - x^*\right) - \int_0^1 \nabla^2 f\left(x^* + t\left(x^{(k)} - x^*\right)\right)\left(x^{(k)} - x^*\right)\, dt\right) \\
&= \left(\nabla^2 f_k\right)^{-1} \left(\nabla^2 f_k - \int_0^1 \nabla^2 f\left(x^* + t\left(x^{(k)} - x^*\right)\right)\, dt\right)\left(x^{(k)} - x^*\right)
\end{aligned}$$

Nun gibt es ein $\delta > 0$ und ein $C > 0$, so daß für alle $x \in \mathbb{R}^n$ mit $\|x - x^*\| \le \delta$ die Abschätzungen

$$\left\|\nabla^2 f(x)\right\| \le C\left\|\nabla^2 f_*\right\| \qquad \text{und} \qquad \left\|\nabla^2 f(x)^{-1}\right\| \le C\left\|\nabla^2 f_*^{-1}\right\|$$

gelten. Angenommen, $\left\|x^{(k)} - x^*\right\| \le \delta$, dann ist

$$\begin{aligned}
&\left\|\nabla^2 f_k - \int_0^1 \nabla^2 f\left(x^* + t\left(x^{(k)} - x^*\right)\right)\, dt\right\| \\
&\quad= \left\|\int_0^1 \nabla^2 f_k - \nabla^2 f\left(x^* + t\left(x^{(k)} - x^*\right)\right)\, dt\right\| \\
&\quad\le \int_0^1 \underbrace{\left\|\nabla^2 f_k - \nabla^2 f\left(x^* + t\left(x^{(k)} - x^*\right)\right)\right\|}_{\le \gamma\left\|x^{(k)} - x^*\right\|}\, dt \le \gamma\left\|x^{(k)} - x^*\right\|,
\end{aligned}$$

also

$$\begin{aligned}\left\|x^{(k+1)}-x^*\right\| &\leq \left\|\nabla^2 f_k^{-1}\right\| \left\|\nabla^2 f_k - \int_0^1 \nabla^2 f_t \,dt\right\| \left\|x^{(k)}-x^*\right\| \\ &\leq C\gamma \left\|\nabla^2 f_*^{-1}\right\| \left\|x^{(k)}-x^*\right\|^2.\end{aligned}$$

Ist nun[202] $\left\|x^{(k)}-x^*\right\| \leq \left(C\gamma \left\|\nabla^2 f_*^{-1}\right\|\right)^{-1}$, dann gilt das auch für $x^{(k+1)}$ und die Iteration bleibt in der "guten" Umgebung, woraus (7.8) und somit auch (7.7) folgen. Unter Verwendung von $y^{(k)} = x^{(k+1)} - x^{(k)} = -\nabla^2 f_k^{-1} \nabla f_k$ ergibt sich dann auch

$$\begin{aligned}\|\nabla f_{k+1}\| &= \left\|\nabla f_{k+1} - \nabla f_k + \nabla^2 f_k \left(\nabla^2 f_k^{-1} \nabla f_k\right)\right\| \\ &= \left\|\int_0^1 \nabla^2 f\left(x^{(k)} + t y^{(k)}\right) y^{(k)} \,dt - \nabla^2 f_k \, y^{(k)}\right\| \\ &\leq \left\|\int_0^1 \nabla^2 f\left(x^{(k)} + t y^{(k)}\right) - \nabla^2 f_k \,dt\right\| \left\|y^{(k)}\right\| \leq \gamma \left\|y^{(k)}\right\|^2 \leq \gamma \left\|\nabla^2 f_k^{-1}\right\|^2 \|\nabla f_k\|^2 \\ &\leq \gamma C^2 \left\|\nabla^2 f_*^{-1}\right\|^2 \|\nabla f_k\|^2,\end{aligned}$$

was (7.9) liefert. □

7.3 Quasi–Newton–Verfahren

Wie wir also in Satz 7.4 gesehen haben, ist das Newton–Verfahren ein "gutes" Verfahren in dem Sinn, daß es lokal, aber vor allem *schnell* gegen eine lokale Minimalstelle konvergiert, was uns natürlich nicht von der Schwierigkeit befreit, so einen Startwert zu finden. Was in solchen Fällen oftmals hilft, ist ein sogenanntes **Hybridverfahren**, bei dem Abstiegsmethode und Newton–Verfahren im Wechsel ausgeführt werden (beispielsweise ein Schritt Abstieg, dann mehrere Schritte Newton), in der Hoffnung, daß das Abstiegsverfahren dafür sorgt, daß man nahe genug an das Minimum kommt, bis dann letztendlich das Newton–Verfahren "greift". Solche Verfahren sind zwar heuristisch naheliegend und auch gut motiviert, aber mathematisch nicht so schön zu untersuchen.

Wir wollen hier noch einmal die Idee des Broyden–Verfahrens ins Spiel bringen, also die Frage, wie man die Berechnung von $\nabla^2 f_k$ nach Möglichkeit vermeiden kann. Dazu betrachten wir im k–ten Schritt an der Stelle $x^{(k)}$ ein **quadratisches Modell**

$$\underbrace{f\left(x^{(k)}+y\right)}_{} \sim \underbrace{f\left(x^{(k)}\right)}_{f_k} + \nabla^T f_k \, y + \frac{1}{2} y^T \nabla^2 f_k \, y \sim f_k + \nabla^T f_k \, y + \frac{1}{2} y^T B_k y =: \widehat{f_k}(y), \quad (7.10)$$

wobei $\nabla^2 f_k \approx B_k \in \mathbb{R}^{n\times n}$ eine *symmetrische, positiv definite*[203] Matrix und Näherung von $\nabla^2 f_k$ sein soll. Dann setzen wir wieder einmal

$$x^{(k+1)} = x^{(k)} + \alpha_k \, y^{(k)}, \qquad y^{(k)} = -B_k^{-1} \nabla f_k, \qquad \alpha_k \in \mathbb{R}_+,$$

[202] Hier nehmen wir an, daß die Konstante $C\gamma \left\|\nabla^2 f_*^{-1}\right\| \geq 1$ ist, ansonsten wäre alles nur noch einfacher.

[203] Wir nehmen also an, wir wären schon nahe genug an einem *strikten* lokalen Minimum.

und stellen uns anhand des „Modells“

$$\widehat{f}_{k+1}(y) = f_{k+1} + \nabla^T f_{k+1}\, y + \frac{1}{2} y^T B_{k+1} y$$

die Frage, wie man nun B_{k+1} wählen sollte. Erinnern wir uns daran, daß wir eine Nullstelle von ∇f berechnen wollen, dann könnten wir beispielsweise fordern, daß die lineare Näherung B_{k+1} der Ableitung $\nabla^2 f$ von ∇f zumindest die **Sekantenbedingung**

$$B_{k+1}\left(x^{(k+1)} - x^{(k)}\right) = \nabla f_{k+1} - \nabla f_k \tag{7.11}$$

erfüllt. Für $n = 1$ ist das das bereits bekannte **Sekantenverfahren**[204] für f', für $n > 1$ reicht das aber natürlich nicht aus, um die Matrix B_{k+1} komplett festzulegen. Multiplikation von links mit $\left(x^{(k+1)} - x^{(k)}\right)^T$ ergibt, zusammen mit der positiven Definitheit von B_{k+1}, daß die Sekantenbedingung nur dann erfüllbar ist, wenn die **Krümmungsbedingung**

$$\left(x^{(k+1)} - x^{(k)}\right)^T \left(\nabla f_{k+1} - \nabla f_k\right) > 0 \tag{7.12}$$

erfüllt ist, was sich mit der richtigen Schrittweitensteuerung erreichen läßt. In der Tat liefert die zweite Wolfe–Bedingung aus (6.21) mit $y = \alpha_k^{-1}\left(x^{(k+1)} - x^{(k)}\right)$, daß

$$\begin{aligned} 0 &\le \alpha_k^{-1}\left(x^{(k+1)} - x^{(k)}\right)^T \left(\nabla f_{k+1} - c_2 \nabla f_k\right) \\ &= \frac{c_2}{\alpha_k}\left(x^{(k+1)} - x^{(k)}\right)^T \left(\nabla f_{k+1} - \nabla f_k\right) + \frac{1 - c_2}{\alpha_k} \underbrace{\left(x^{(k+1)} - x^{(k)}\right)^T \nabla f_{k+1}}_{\le 0} \\ &\le \frac{c_2}{\alpha_k}\left(x^{(k+1)} - x^{(k)}\right)^T \left(\nabla f_{k+1} - \nabla f_k\right), \end{aligned}$$

woraus (7.12) folgt, da $y^{(k)}$ eine Abstiegsrichtung[205] und daher $\alpha_k > 0$ ist. Nun bestimmt die Sekantenbedingung (7.11) aber natürlich die Matrix B_{k+1} nicht vollständig, weswegen man sie wieder einmal als Lösung eines Minimierungsproblems definieren kann, beispielsweise

$$B_{k+1} = \min_B \|B - B_k\|, \qquad B = B^T, \quad B\left(x^{(k+1)} - x^{(k)}\right) = \nabla f_{k+1} - \nabla f_k, \tag{7.13}$$

wobei $\|\cdot\|$ eine *beliebige* Matrixnorm sein kann – und in der Tat liefert verschiedene Normen auch verschiedene *Quasi–Newton–Verfahren*, wie man diese Familie von Iterationsverfahren auch nennt. Eine beliebte Wahl ist eine **gewichtete Frobenius–Norm**[206] der Form

$$\|A\|_W = \left\|W^{1/2} A W^{1/2}\right\|_F, \qquad \|A\|_F^2 = \operatorname{trace}\left(A^T A\right) = \sum_{j,k=1}^{n} a_{jk}^2, \tag{7.14}$$

[204] Zumal alle 1×1–Matrizen immer und automatisch symmetrisch sind.

[205] Schließlich ist ja $B_k y^{(k)} = \nabla f_k$ und im quadratischen Fall ist das sogar der *direkte* Weg zum Minimum.

[206] Ferdinand Georg Frobenius, 1849–1917, promovierte bei Weierstrass und wurde 1874 *ohne Habilitation* in Berlin zum Professor ernannt. Wichtige Beiträge zur Darstellungstheorie von Gruppen (hat ja auch einiges mit Matrizen zu tun), insbesondere Entwicklung der Charakteren–

wobei die **Gewichtsmatrix** $W \in \mathbb{R}^{n\times n}$ symmetrisch und positiv semidefinit sein muß, denn dann ist die (positive) **Wurzel** $W^{1/2}$ wohldefiniert als diejenige symmetrische, positiv definite Matrix B, die $B^2 = W$ erfüllt.

Übung 7.1 Zeigen Sie:

1. Zu jeder symmetrischen positiv semidefiniten Matrix W gibt es eine eindeutige symmetrische positiv semidefinite Matrix $B = W^{1/2}$, so daß $W = B^2$ ist.

2. Für $A \in \mathbb{R}^{n\times n}$ ist
$$\operatorname{trace}\left(A^T A\right) = \sum_{j,k=1}^{n} a_{jk}^2.$$

◇

Bei geeigneter Wahl von W (als eine gemittelte Hessematrix) so daß $W^{-1}\xi_k = \eta_k$, mit $\xi_k := x^{(k+1)} - x^{(k)}$ und $\eta_k := \nabla f_{k+1} - \nabla f_k$, ergibt sich dann die folgende Update–Regel aus (Nocedal & Wright, 1999, S. 196)

$$B_{k+1} = \left(I - \frac{\eta_k\,\xi_k^T}{\eta_k^T\xi_k}\right) B_k \left(I - \frac{\xi_k\,\eta_k^T}{\eta_k^T\xi_k}\right) + \frac{\eta_k\eta_k^T}{\eta_k^T\xi_k}, \tag{7.15}$$

die 1959 von Davidon vorgeschlagen (Davidon, 1959; Davidon, 1991), aber vor allem von Fletcher und Powell (unabhängig) untersucht und popularisiert wurde, weswegen man sie als **DFP–Methode** bezeichnet.

Anstelle mit B_k zu rechnen und in jedem Iterationsschritt das Gleichungssystem $B_k y^{(k)} = -\nabla f_k$ lösen zu müssen, kann man auch *direkt* mit $B_k^{-1} := H_k$ rechnen und nun das Minimierungsproblem

$$H_{k+1} = \min_H \|H - H_k\|, \qquad H = H^T, \quad H\left(\nabla f_{k+1} - \nabla f_k\right) = x^{(k+1)} - x^{(k)}, \tag{7.16}$$

lösen, was zur Update–Regel

$$H_{k+1} = \left(I - \frac{\xi_k\,\eta_k^T}{\eta_k^T\xi_k}\right) H_k \left(I - \frac{\eta_k\,\xi_k^T}{\eta_k^T\xi_k}\right) + \frac{\xi_k\xi_k^T}{\eta_k^T\xi_k} \tag{7.17}$$

führt, in der, wegen des Übergangs zur Inversen, die Rollen von ξ_k und η_k vertauscht sind. Das damit verbundene Verfahren bezeichnet man nach seinen „Vätern" Broyden, Fletcher, Goldfarb und Shanno[207] als **BFGS–Verfahren**.

Theorie. Eine interessante Bemerkung über Frobenius ist:

> *For Frobenius, conceptual argumentation played a somewhat secondary role. Although he argued in a comparatively abstract setting, abstraction was not an end in itself.*

Darüberhinaus konnte er den „neuen mathematischen Stil" aus Göttingen (verkörpert durch Klein und Lie) ganz und gar nicht ausstehen ...

[207] Das Literaturverzeichnis von (Nocedal & Wright, 1999) legt nahe, daß sie das Verfahren nicht gemeinsam sondern aufeinander aufbauend oder in Konkurrenz oder unabhängig oder wie auch immer entwickelt haben.

Man kann nun auch, ausgehend von (7.15) eine inverse Regel für die Updates der entsprechenden H_k im DFP–Verfahren, beziehungsweise, ausgehend von (7.17), eine primäre Regel zur Bestimmung von B_{k+1} aus B_k für das BFGS–Verfahren aufstellen. Das geht ganz einfach unter Verwendung des folgenden Resultats.

Lemma 7.5 („Sherman–Morrison–Woodbury–Formel") [208]
Für eine nichtsinguläre Matrix $A \in \mathbb{R}^{n\times n}$ und $x, y \in \mathbb{R}^n$ ist

$$\left(A + xy^T\right)^{-1} = A^{-1} - \frac{A^{-1}xy^TA^{-1}}{1 + y^TA^{-1}x}. \tag{7.18}$$

Übung 7.2 Beweisen Sie (7.18) (durch Ausmultiplizieren)[209] und charakterisieren Sie, wann $A + xy^T$ invertierbar ist. ◊

Übung 7.3 Zeigen Sie, daß die zu (7.17) äquivalente Update–Regel sich als

$$H_{k+1} = H_k - \frac{H_k\xi_k\,\xi_k^TH_k^T}{\xi_k^TH_k\xi_k} + \frac{\eta_k\eta_k^T}{\xi_k^T\eta_k} \tag{7.19}$$

schreiben läßt. ◊

Wir werden uns nun mit *Konvergenzeigenschaften* des BFGS–Verfahrens befassen, die man unter gewissen (lokalen) Bedingungen auch tatsächlich beweisen kann.

Satz 7.6 *Sei $f \in C^2(\mathbb{R}^n)$ und sei $x^{(0)}$ so gewählt, daß*

$$\Omega := \left\{x \in \mathbb{R}^n \;:\; f(x) \le f\left(x^{(0)}\right)\right\}$$

konvex ist und es Konstanten $0 < m < M$ gibt, so daß

$$m\,\|y\|_2^2 \le y^T\,\nabla^2 f(x)\,y \le M\,\|y\|_2^2, \qquad x \in \Omega. \tag{7.20}$$

Dann konvergiert die Folge

$$x^{(k+1)} := x^{(k)} + \alpha_k\,y^{(k)}, \qquad y^{(k)} = -H_k\,\nabla f_k, \qquad k \in \mathbb{N}_0,$$

unter Beachtung der Wolfe– oder Powell–Bedingungen und unter Verwendung der Update–Regel (7.17) gegen ein Minimum x^* von f.

Bemerkung 7.7 *Die Bedingung (7.20) bedeutet, daß die Funktion f auf der gesamten Niveaumenge* gleichmäßig strikt *konvex (oder auch* stark konvex*) ist. Das ist natürlich ziemlich viel verlangt und eine Bedingung für ein* striktes lokales *Minimum, das, wenn man den Einstiegswert niedrig genug wählt, dann aber auch gefunden wird.*

[208] Normalerweise sollte man sehr skeptisch bei allem sein, was den Namen von mehr als zwei Personen trägt ...

[209] Drei Namen und ein fast trivialer Beweis.

Für den Beweis von Satz 7.6 brauchen wir zuerst eine kleine Hilfsaussage aus der linearen Algebra.

Lemma 7.8 *Seien* $x, y, u, v \in \mathbb{R}^n$. *Dann ist*

$$\det\left(I + uv^T + xy^T\right) = \left(1 + u^T v\right)\left(1 + x^T y\right) - \left(v^T x\right)\left(u^T y\right). \tag{7.21}$$

Beweis: Beginnen wir mit dem Fall $x = 0$ oder $y = 0$. Dazu seien $w_2, \ldots, w_n$ linear unabhängige Vektoren, die senkrecht auf v stehen, dann ist

$$\left(I + uv^T\right) w_j = w_j + u \underbrace{v^T w_j}_{=0} = w_j =: \lambda_j w_j, \qquad j = 2, \ldots, n,$$

sowie

$$\left(I + uv^T\right) u = u + \left(v^T u\right) u = \left(1 + u^T v\right) u =: \lambda_1 u,$$

womit wie alle Eigenwerte und Eigenvektoren identifiziert haben und da die Determinante das Produkt der Eigenvektoren ist, erhalten wir, daß

$$\det\left(I + uv^T\right) = \prod_{j=1}^{n} \lambda_j = 1 + u^T v.$$

Nun nehmen wir an, daß v und y linear unabhängig sind, denn ansonsten könnten wir das auf den einfachen Fall zurückführen, den wir gerade erledigt haben. Nun wählen wir $w_3, \ldots, w_n$ senkrecht zu v und y, was uns sofort

$$\left(I + uv^T + xy^T\right) w_j = w_j, \qquad j = 3, \ldots, n$$

liefert und da

$$\begin{aligned} \left(I + uv^T + xy^T\right) x &= \left(1 + x^T y\right) x + \left(v^T x\right) u \\ \left(I + uv^T + xy^T\right) u &= \left(u^T y\right) x + \left(1 + v^T u\right) u \end{aligned}$$

ist, ergibt sich für die ersten beiden Eigenwerte λ_1 und λ_2, daß

$$\lambda_1 \lambda_2 = \det\begin{bmatrix} 1 + x^T y & v^T x \\ u^T y & 1 + u^T v \end{bmatrix} = \left(1 + x^T y\right)\left(1 + u^T v\right) - \left(v^T x\right)\left(u^T y\right),$$

woraus (7.21) unmittelbar folgt. □

Beweis von Satz 7.6: Wie die Verwendung der **Wolfe–Bedingungen** ja nahelegt, wollen wir Satz 6.19 verwenden – zu diesem Zweck müssen wir aber die Winkel zwischen Gradienten und Abstiegsrichtungen in den Griff bekommen.

Da

$$\eta_k = \nabla f_{k+1} - \nabla f_k = \int_0^1 \nabla\left(\nabla f\right) \underbrace{\left(x^{(k)} + t\xi_k\right)}_{=:x_t} \xi_k \, dt = \int_0^1 \nabla^2 f\left(x_t\right) \xi_k \, dt =: G_k \xi_k,$$

ist wegen der Annahme (7.20)

$$\xi_k^T \eta_k = \int_0^1 \underbrace{\xi_k^T \nabla^2 f(x_t)\ \xi_k}_{\geq m\|\xi_k\|_2^2}\ dt \geq m\,\|\xi_k\|_2^2 \qquad \Rightarrow \qquad m_k := \frac{\xi_k^T \eta_k}{\xi_k^T \xi_k} \geq m,$$

sowie

$$M_k := \frac{\eta_k^T \eta_k}{\eta_k^T \xi_k} = \frac{\xi^T G_k^2 \xi}{\xi^T G_k \xi} = \frac{\left(\sqrt{G_k}\xi\right)^T G_k \left(\sqrt{G_k}\xi\right)}{\left(\sqrt{G_k}\xi\right)^T \left(\sqrt{G_k}\xi\right)} = \frac{1}{\|z\|_2^2} \int_0^1 z^T \nabla^2 f(x_t)\ z\ dt \leq M.$$

Schreiben wir (7.19) in

$$B_{k+1} = B_k - \frac{(B_k\xi_k)\ (B_k\xi_k)^T}{\xi_k^T B_k \xi_k} + \frac{\eta_k \eta_k^T}{\xi_k^T \eta_k} = B_k \left(I - \frac{\xi_k\ (B_k\xi_k)^T}{\xi_k^T B_k \xi_k^T} + \frac{B_k^{-1}\eta_k \eta_k^T}{\xi_k^T \eta_k} \right) \tag{7.22}$$

um und berücksichtigen wir, daß $\operatorname{trace}(xx^T) = \|x\|_2^2$ ist, dann erhalten wir, daß

$$\operatorname{trace} B_{k+1} = \operatorname{trace} B_k - \frac{\|B_k\xi_k\|_2^2}{\xi_k^T B_k \xi_k} + \frac{\|\eta_k\|_2^2}{\xi_k^T \eta_k} = \operatorname{trace} B_k - \frac{\|B_k\xi_k\|_2^2}{\xi_k^T B_k \xi_k^T} + M_k. \tag{7.23}$$

Die zweite Identität in (7.22) und Lemma 7.8 liefern außerdem, daß

$$\begin{aligned} \det B_{k+1} &= \det B \Bigg(\underbrace{\left(1 - \frac{\xi_k^T B_k \xi_k}{\xi_k^T B_k \xi_k}\right)}_{=0} \left(1 + \frac{\eta_k^T B_k^{-1} \eta_k}{\xi_k^T \eta_k}\right) - \frac{-\xi_k^T \eta_k}{\xi_k^T B_k \xi_k} \frac{\xi_k^T B_k B_k^{-1} \eta_k}{\xi_k^T \eta_k} \Bigg) \\ &= \frac{\xi_k^T \eta_k}{\xi_k^T B_k \xi_k} \det B_k \end{aligned} \tag{7.24}$$

Schreiben wir θ_k für den Winkel zwischen ξ_k und $B_k\xi_k$, also

$$\cos\theta_k := \frac{\xi_k^T\, B_k\, \xi_k}{\|\xi_k\|_2\, \|B_k\xi_k\|_2},$$

so erhalten wir für den zweiten Term auf der rechten Seite von (7.23), daß

$$\frac{\|B_k\xi_k\|_2^2}{\xi_k^T B_k \xi_k} = \frac{\|B_k\xi_k\|_2^2\ \|\xi_k\|_2^2}{\left(\xi_k^T\, B_k\, \xi_k\right)^2} \frac{\xi_k^T B_k \xi_k}{\|\xi_k\|_2^2} = \frac{\xi_k^T B_k \xi_k}{\|\xi_k\|_2^2 \cos^2\theta_k} =: \frac{\beta_k}{\cos^2\theta_k}, \tag{7.25}$$

wobei $\beta_k = \|\xi_k\|_2^{-2}\left(\xi_k^T B_k \xi_k\right)$ eine Zahl ist, die nach unten durch den kleinsten[210] Eigenwert von B_k und nach oben durch den größten Eigenwert von B_k beschränkt ist. Damit liefert auch (7.24), daß

$$\det B_{k+1} = \frac{1}{\beta_k} \underbrace{\frac{\xi_k^T \eta_k}{\xi_k^T \xi_k}}_{=m_k} \det B_k = \frac{m_k}{\beta_k} \det B_k. \tag{7.26}$$

[210] Aber immer noch positiven!

Zu einer symmetrischen, positiv (semi-)definiten Matrix $B \in \mathbb{R}^{n\times n}$ mit Eigenwerten $0 \le \lambda_1 \le \cdots \le \lambda_n$ betrachten wir nun die Funktion

$$\psi(B) := \operatorname{trace} B - \log\det B = \sum_{j=1}^{n} \lambda_j - \log\left(\prod_{j=1}^{n} \lambda_j\right) = \sum_{j=1}^{n} (\lambda_j - \log\lambda_j) > 0,$$

da $\log t < t$ ist für $t > 0$. Mit (7.23), (7.24), (7.25) und (7.26) sowie Übung 7.4 erhalten wir somit, daß

$$\begin{aligned}
0 \;<\; & \psi(B_{k+1}) = \operatorname{trace} B_k - \frac{\beta_k}{\cos^2\theta_k} + M_k - \log\det B_k - \log\frac{m_k}{\beta_k} \\
= \; & \psi(B_k) - \frac{\beta_k}{\cos^2\theta_k} + M_k - \log m_k + \log\beta_k \\
= \; & \psi(B_k) + (M_k - \log m_k - 1) + \left(1 - \frac{\beta_k}{\cos^2\theta_k} + \log\frac{\beta_k}{\cos^2\theta_k}\right) + \log\cos^2\theta_k \\
\le \; & \psi(B_k) + (M - \log m - 1) + \underbrace{\left(1 - \frac{\beta_k}{\cos^2\theta_k} + \log\frac{\beta_k}{\cos^2\theta_k}\right)}_{\le 0} + \log\cos^2\theta_k \\
\le \; & \psi(B_k) + (M - \log m - 1) + \log\cos^2\theta_k \\
\le \; & \psi(B_{k-1}) + 2\,(M - \log m - 1) + \log\cos^2\theta_{k-1} + \log\cos^2\theta_k \\
& \vdots \\
\le \; & \psi(B_1) + k\,(M - \log m - 1) + \sum_{j=1}^{k} \log\cos^2\theta_j,
\end{aligned}$$

und indem wir die Schranken m und M hinreichend klein bzw. groß wählen, können wir ohne Einschränkung annehmen, daß $M - \log m - 1 > 0$ ist.

Und damit bekommen wir schließlich unsere Winkel θ_k in den Griff: Wäre nämlich

$$\lim_{k\to\infty} \cos\theta_k = 0 \qquad \Rightarrow \qquad \lim_{k\to\infty} \cos^2\theta_k = 0 \qquad \Rightarrow \qquad \lim_{k\to\infty} \log\cos^2\theta_k = -\infty,$$

also auch

$$\lim_{k\to\infty} \sum_{j=1}^{k} \Big(\log\cos^2\theta_j + (M - \log m - 1)\Big) = -\infty,$$

dann erhielten wir den Widerspruch

$$0 \le \lim_{k\to\infty} \psi(B_{k+1}) = \psi(B_1) + \underbrace{\lim_{k\to\infty} \sum_{j=1}^{k} \Big(\log\cos^2\theta_j + (M - \log m - 1)\Big)}_{=-\infty} = -\infty.$$

Also gibt es zumindest eine Teilfolge $x^{(k_j)}$, $j \in \mathbb{N}_0$, die gegen ein Minimum konvergiert, aber wegen der *starken* Konvexität der Funktion f bleibt dann auch der gesamten Folge nichts anderes übrig, als zu konvergieren. □

Übung 7.4 Zeigen Sie: Für jedes $t > 0$ gilt $\log t \le t - 1$ mit Gleichheit genau dann, wenn $t = 1$ ist. ◊

Bemerkung 7.9 *1. Nach (Nocedal & Wright, 1999) läßt sich dieser Beweis für BFGS mit Powell–Schrittweiten*[211] *auf eine ganze Klasse von Verfahren, die sogenannte* ***Broyden–Klasse*** *ausdehnen, funktioniert aber* nicht *für das DFP–Verfahren.*

2. Man kann auch zeigen[212]*, daß das BFGS–Verfahren* ***superlinear*** *konvergiert, wenn die zweite Ableitung Lipschitz–stetig ist. Genauer: es ist*

$$\lim_{k\to\infty} \frac{\left\|x^{(k+1)} - x^*\right\|}{\left\|x^{(k)} - x^*\right\|} = 0.$$

[211] Oder „Wolfe–Schrittweiten".

[212] Mit noch etwas sorgfältigerer Rechnerei, siehe (Nocedal & Wright, 1999, S. 214–218).

Strafen heißt, absichtlich ein Übel zuzufügen. Wer in diesem Sinne strafen will, muß sich eines höheren Auftrags zuversichtlich bewußt sein.

Gustav Radbruch

8 Strafterme und Barrieren

Wir kehren jetzt nochmal zu der in Kapitel 6 bereits erwähnten Idee zurück, *restringierte* Optimierungsprobleme dadurch zu behandeln, daß man sie in ein oder mehrere *unrestringierte* Approximationsprobleme umwandelt, bei denen die Verletzung der Nebenbedingungen als Bestandteil der Zielfunktion aufgefasst wird. Dabei betrachten wir das restringierte Optimierungsproblem

$$\min f(x), \qquad g(x) = 0, \qquad g : \mathbb{R}^n \to \mathbb{R}^m, \tag{8.1}$$

bzw., wenn wir auch Ungleichungsbedingungen zulassen wollen,

$$\min f(x), \qquad g(x) = 0, \quad h(x) \geq 0, \qquad g, h : \mathbb{R}^n \to \mathbb{R}^m, \tag{8.2}$$

mit den *stetigen* Nebenbedingungsfunktionen[213] g und h für die Gleichheits- und Ungleichungsbedingungen. Um nicht in Existenznöte bezüglich des Minimums zu kommen nehmen wir außerdem an, daß der zulässige Bereich

$$\{x \in \mathbb{R}^n : g(x) = 0\} \qquad \text{bzw.} \qquad \{x \in \mathbb{R}^n : g(x) = 0,\ h(x) \geq 0\}$$

kompakt sein soll.
Die Idee hinter den Straftermen und Barrieren besteht nun darin, anstelle von f eine Funktion $x \mapsto f_\Phi(x) := f(x) + \Phi\left(g(x), h(x)\right)$ zu minimieren, wobei man natürlich $\Phi : \mathbb{R}^{2m} \to \mathbb{R}$ so wählen sollte, daß f_Φ *einfach zu berechnen* und *einfach zu minimieren* ist.

8.1 Quadratische Strafterme

Für $f : \mathbb{R}^n \to \mathbb{R}$ und Nebenbedingungen g und h definiert man die **quadratische Straffunktion**[214] $Q : \mathbb{R}^n \times \mathbb{R}_+ \to \mathbb{R}$ als

$$Q(x, \mu) = f(x) + \frac{1}{2\mu} \left\| g(x) \right\|_2^2 = f(x) + \frac{1}{2\mu} \sum_{j=1}^{m} g_j^2(x) \tag{8.3}$$

[213] Daß beide als Wert m–Vektoren haben, ist keine Einschränkung! Wir brauchen den Restriktionstyp, der aus weniger Nebenbedingungen besteht, durch Einführung von Bedingungen des Typs "$0 = 0$" oder "$0 \geq 0$", also durch $g_j \equiv 0$ oder $h_j \equiv 0$ aufzufüllen.
[214] Englisch: **penalty function**.

beziehungsweise

$$Q(x,\mu) = f(x) + \frac{1}{2\mu}\left(\left\|g(x)\right\|_2^2 + \left\|h_-(x)\right\|_2^2\right) = f(x) + \frac{1}{2\mu}\sum_{j=1}^{m} g_j^2(x) + (h_j)_-^2(x), \quad (8.4)$$

wobei

$$y_- = \frac{1}{2}(y - |y|) = \begin{cases} 0, & y \geq 0, \\ y, & y < 0. \end{cases}$$

Allerdings hat (8.4) im Vergleich zu (8.3) einen wesentlichen Nachteil: Die Funktion $x \mapsto x_-$ ist nicht mehr differenzierbar und so kann die Differenzierbarkeitsordnung von Q *geringer* sein als die von f, g und h.

Zur Bestimmung des restringierten Minimums wählt man nun eine positive Nullfolge $\mu_k \in \mathbb{R}_0$, $k \in \mathbb{N}$, $\mu_k \to 0$, $k \to \infty$ und minimiert[215] $Q(x, \mu_k)$ bezüglich x, bis man eine Näherungslösung $x^{(k)}$ gefunden hat, die

$$\left\|\nabla_x Q\left(\mu_k, x^{(k)}\right)\right\|_2 \leq \tau_k \quad (8.5)$$

für vorgegebene Toleranzen $\tau_k > 0$, $k \in \mathbb{N}_0$, erfüllt. Je kleiner nun μ_k wird, desto weniger darf die Näherungslösung die Nebenbedingungen verletzen und so besteht die Hoffnung, daß $x^{(k)} \to x^*$ für $k \to \infty$, wobei

1. x^* eine *Minimalstelle* von f ist.
2. x^* ein (näherungsweise) **zulässiger Punkt** ist:

$$g(x^*) \sim 0 \qquad \text{und} \qquad h(x^*) \geq -\varepsilon, \quad \varepsilon \sim 0.$$

Und diese Hoffnung besteht zu Recht.

Proposition 8.1 *Sei μ_k, $k \in \mathbb{N}$, eine positive Nullfolge und seien $x^{(k)}$, $k \in \mathbb{N}$, die Minimallösungen von $Q\left(x^{(k)}, \mu_k\right)$ aus (8.4). Dann ist jeder Häufungspunkt x^* der Folge $x^{(k)}$ eine Lösung von (8.1).*

Beweis: Sei $\widehat{x}$ eine globale Lösung von (8.2), das heißt,

$$f(\widehat{x}) \leq f(x), \qquad x \in Z_g \cap Z_{h_-}, \qquad Z_\varphi := \{x' \in \mathbb{R}^n \,:\, \varphi(x') = 0\}.$$

Insbesondere ist also $\widehat{x}$ ein zulässiger Punkt, der $g(\widehat{x}) = h_-(\widehat{x}) = 0$ erfüllt. Nach der Definition der $x^{(k)}$ als Minimallösungen des modifizierten Problems ist außerdem für $k \in \mathbb{N}$

$$\begin{aligned} f\left(x^{(k)}\right) \;+\; & \frac{1}{2\mu_k}\left(\left\|g\left(x^{(k)}\right)\right\|_2^2 + \left\|h_-\left(x^{(k)}\right)\right\|_2^2\right) = Q\left(x^{(k)}, \mu\right) \leq Q(\widehat{x}, \mu) \\ = \; & f(\widehat{x}) + \frac{1}{2\mu_k}\Big(\underbrace{\left\|g(\widehat{x})\right\|_2^2}_{=0} + \underbrace{\left\|h_-(\widehat{x})\right\|_2^2}_{=0}\Big) = f(\widehat{x}), \end{aligned}$$

[215] Mit einem der Verfahren zur unrestringierten Optimierung aus den vorherigen Kapiteln. Oder natürlich mit etwas besserem ...

also

$$\left\|g\left(x^{(k)}\right)\right\|_2^2 + \left\|h_-\left(x^{(k)}\right)\right\|_2^2 \leq 2\mu_k\left(f\left(\widehat{x}\right) - f\left(x^{(k)}\right)\right). \tag{8.6}$$

Sei nun x^* ein Häufungspunkt, d.h., es gibt eine Folge k_j, so daß

$$x^* = \lim_{j\to\infty} x^{(k_j)}.$$

Wegen der Stetigkeit von g und h, der Stetigkeit der Norm und (8.6) ergibt siche

$$\begin{aligned}\left\|g\left(x^*\right)\right\|_2^2 + \left\|h_-\left(x^*\right)\right\|_2^2 &= \lim_{j\to\infty}\left(\left\|g\left(x^{(k_j)}\right)\right\|_2^2 + \left\|h_-\left(x^{(k_j)}\right)\right\|_2^2\right)\\ &\leq \lim_{j\to\infty} 2\underbrace{\mu_{k_j}}_{\to 0}\underbrace{\left(f\left(\widehat{x}\right) - f\left(x^{(k_j)}\right)\right)}_{\to f(\widehat{x})-f(x^*)} = 0,\end{aligned}$$

weswegen $x^* \in Z_g$ ein **zulässiger Punkt** ist. Und da

$$\begin{aligned}f\left(x^*\right) &= \lim_{j\to\infty} f\left(x^{(k_j)}\right) \leq \lim_{j\to\infty} f\left(x^{(k_j)}\right) + \frac{1}{2\mu_{k_j}}\left(\left\|g\left(x^{(k_j)}\right)\right\|_2^2 + \left\|h_-\left(x^{(k_j)}\right)\right\|_2^2\right)\\ &\leq f\left(\widehat{x}\right)\end{aligned}$$

ist, bleibt x^* auch gar nichts anderes übrig, als Minimallösung zu sein. □

Auch für die **Penalty–Methode** kann man wieder die Konvergenz gegen einen **stationären Punkt** beweisen, allerdings müssen wir jetzt wieder die Nebenbedingungen berücksichtigen. Und das erinnert uns deutlich an die verallgemeinerten **Lagrange–Multiplikatoren** aus Satz 5.13, deren Existenz uns eine *notwendige*[216] Bedingung für die Existenz eines lokalen Minimums geliefert hat – ganz genau wie die Forderung $\nabla f = 0$. Trotzdem kann man für die Penalty–Methode einen Konvergenzbeweis führen, was wir allerdings nur für Probleme der Form (8.1), also ohne Verwendung von Ungleichungsnebenbedingungen tun werden.

Satz 8.2 *Sei $g \in C^1(\mathbb{R}^n)$ und seien μ_k und τ_k, $k \in \mathbb{N}$, positive Nullfolgen und sei x^* ein Häufungspunkt der Folge $x^{(k)}$, die (8.5) erfüllt. Sind die Gradienten $\nabla g_j(x^*)$, $j = 1, \ldots, m$, linear unabhängig, dann gibt es einen Vektor $\lambda \in \mathbb{R}^n$, so daß*

$$\nabla f\left(x^*\right) - \nabla g\left(x^*\right)\lambda = 0 \tag{8.7}$$

ist, und es ist

$$\lambda = \lim_{j\to\infty} -\frac{g\left(x^{(k_j)}\right)}{\mu_{k_j}}, \qquad x^* = \lim_{j\to\infty} x^{(k_j)}. \tag{8.8}$$

Bemerkung 8.3 *1. Die Forderung, daß die Gradienten $\nabla g_j : \mathbb{R}^n \to \mathbb{R}^n$ linear unabhängig sind, setzt natürlich voraus, daß $m \leq n$ ist; zu viele Nebenbedingungen dürfen wir also nicht haben.*

[216] Und nicht unbedingt hinreichende!

2. *Daß auch die τ_k eine Nullfolge bilden müssen, ist ziemlich naheliegend, denn ansonsten sind die Lösungen $x^{(k)}$ auch keine hinreichend guten Minimallösungen des modifizierten Problems im k–ten Schritt.*

3. *Die Gleichung (8.7), also der Lagrange–Multiplikator, ist nichts anderes als (5.9) aus Satz 5.13. Da (5.10) trivialerweise erfüllt ist – schließlich gibt es ja $h = 0$ – ist also die einzige Bedingung aus Satz 5.13, um die wir uns herumgemogelt haben, die Bedingung (5.8) an die Kegel, und die interessiert sowieso niemanden*[217].

Beweis: Bildet man von (8.3) den Gradient bezüglich x, dann ergibt sich

$$\nabla_x Q(x,\mu) = \nabla f(x) + \frac{1}{\mu}\sum_{j=1}^{m} g_j(x)\,\nabla g_j(x), \tag{8.9}$$

was zusammen mit (8.5) die Bedingung[218]

$$\begin{aligned}\tau_k &\geq \left\|\nabla_x Q\left(x^{(k)},\mu_k\right)\right\| = \left\|\nabla f\left(x^{(k)}\right) + \frac{1}{\mu_k}\sum_{j=1}^{m} g_j\left(x^{(k)}\right)\nabla g_j\left(x^{(k)}\right)\right\| \\ &\geq \frac{1}{\mu_k}\left\|\sum_{j=1}^{m} g_j\left(x^{(k)}\right)\nabla g_j\left(x^{(k)}\right)\right\| - \left\|\nabla f\left(x^{(k)}\right)\right\|,\end{aligned}$$

also

$$\left\|\sum_{j=1}^{m} g_j\left(x^{(k)}\right)\nabla g_j\left(x^{(k)}\right)\right\| \leq \mu_k\left(\tau_k + \left\|\nabla f\left(x^{(k)}\right)\right\|\right) \tag{8.10}$$

liefert. Wegen der Stetigkeit aller beteiligten Größen ist somit

$$\begin{aligned}\left\|\sum_{j=1}^{m} g_j(x^*)\,\nabla g_j(x^*)\right\| &= \lim_{j\to\infty}\left\|\sum_{j=1}^{m} g_j\left(x^{(k_j)}\right)\nabla g_j\left(x^{(k_j)}\right)\right\| \\ &\leq \lim_{j\to\infty} \underbrace{\mu_{k_j}}_{\to 0}\Big(\underbrace{\tau_k}_{\to 0} + \underbrace{\left\|\nabla f\left(x^{(k_j)}\right)\right\|}_{\to \nabla f(x^*)}\Big) = 0\end{aligned}$$

und die lineare Unabhängigkeit der $\nabla g_j(x^*)$ liefert, daß

$$g_j(x^*) = 0, \qquad j = 1,\ldots,m, \tag{8.11}$$

und somit ist x^* zulässig. Mit $\lambda^{(k)} = -\frac{1}{\mu_k} g\left(x^{(k)}\right) \in \mathbb{R}^m$ ergibt sich aus (8.9), daß

$$\nabla_x Q\left(x^{(k)},\mu_k\right) = \nabla f_k - \nabla g_k\,\lambda^{(k)}, \qquad \nabla g_k := \left[\nabla g_j\left(x^{(k)}\right) : j = 1,\ldots,m\right] \in \mathbb{R}^{n\times m},$$

und da

$$\left\|\nabla f_k - \nabla g_k\,\lambda^{(k)}\right\| = \left\|\nabla_x Q\left(x^{(k)},\mu_k\right)\right\| \leq \tau_k \to 0$$

[217] Oder, um es vornehmer zu formulieren: Diese Bedingung wird in der Praxis meist nicht verifiziert.

[218] Hier ist sie wieder, die ebenso nützliche wie weitgehend unbekannte **Dreiecksungleichung nach unten**.

ist, dann müssen wir aus Stetigkeitsgründen zeigen, daß $\lambda^* = \lim_{k\to\infty} \lambda^{(k)}$ existiert. Da $\nabla g\,(x^*)$ vollen Rang m hat, gilt dies für hinreichend großes k auch für ∇g_k und deswegen ist $\nabla^T g_k \nabla g_k \in \mathbb{R}^{m\times m}$ invertierbar, wenn nur k hinreichend groß gewählt ist. Nun ist dann

$$\nabla g_k\,\lambda^{(k)} = \nabla f_k - \nabla_x Q\left(x^{(k)}, \mu_k\right) \quad \Rightarrow \quad \nabla^T g_k \nabla g_k\,\lambda^{(k)} = \nabla^T g_k \left(\nabla f_k - \nabla_x Q\left(x^{(k)}, \mu_k\right)\right)$$

und somit, für hinreichend großes k,

$$\lambda^{(k)} = \underbrace{\left(\nabla^T g_k \nabla g_k\right)^{-1}}_{\to(\nabla^T g_* \nabla g_*)^{-1}} \underbrace{\nabla^T g_k}_{\to\nabla^T g_*} \Big(\underbrace{\nabla f_k}_{\to\nabla f_*} - \underbrace{\nabla_x Q\left(x^{(k)}, \mu_k\right)}_{\to 0}\Big), \tag{8.12}$$

also

$$\lambda^* = \lim_{j\to\infty} \lambda^{(k_j)} = \left(\nabla^T g_* \nabla g_*\right)^{-1} \nabla^T g_*\, \nabla f_*,$$

was einen wohldefinierten Multiplikator ergibt. □

Allerdings gibt es ein kleines Problem, und zwar ein numerisches Problem bei der Bestimmung der näherungsweisen Minima $x^{(k)}$. Dazu nehmen wir der Einfachheit an, daß $h \equiv 0$ ist, daß also die Nebenbedingungen ausschließlich in Gleichungsform vorliegen, und bilden einmal die Hessematrix

$$\begin{aligned}
\nabla_x^2 Q\,(x,\mu) &= \nabla_x\left(\nabla f(x) + \frac{1}{\mu}\sum_{j=1}^{m} g_j(x)\,\nabla g_j(x)\right) = \nabla^2 f(x) + \frac{1}{\mu}\sum_{j=1}^{m} \nabla\left(g_j(x)\,\nabla g_j(x)\right)\\
&= \nabla^2 f(x) + \frac{1}{\mu}\sum_{j=1}^{m}\left(\nabla g_j(x)\,\nabla^T g_j(x) + g_j(x)\,\nabla^2 g_j(x)\right)\\
&= \nabla^2 f(x) + \frac{1}{\mu}\nabla g(x)\nabla^T g(x) + \sum_{j=1}^{m}\frac{g_j(x)}{\mu}\nabla^2 g_j(x),
\end{aligned} \tag{8.13}$$

die ja bei den meisten Verfahren eine ziemlich entscheidende Rolle gespielt hat. Sei nun x^μ die Optimallösung für ein vorgegebenes $\mu > 0$ und $\lambda = -g\,(x^\mu)\,/\mu$, dann ist für $x \sim x^\mu$

$$\begin{aligned}
A_\mu &:= \nabla_x^2 Q\,(x,\mu) = \nabla^2 f(x) + \frac{1}{\mu}\nabla g(x)\nabla^T g(x) - \sum_{j=1}^{m}\lambda_j \nabla^2 g_j(x)\\
&= \nabla^2\left(f - \lambda^T g\right)(x) + \frac{1}{\mu}\nabla g(x)\nabla^T g(x).
\end{aligned}$$

Da für eine vernünftige Konvergenz $m \le n$ sein muß, siehe Satz 8.2, und im Normalfall sogar $m < n$ sein wird, hat die symmetrische Matrix A_μ gerade $n - m$ Eigenvektoren und Eigenwerte η_j, $j = m+1, \ldots, n$, die nicht von μ abhängen, nämlich diejenigen Vektoren, die zu $(\nabla^T g(x))^\perp \subset \mathbb{R}^n$ gehören, und m Eigenvektoren zu Eigenwerten der Form $\eta_j = \eta_j'/\mu$, $j = 1, \ldots, m$, die für $\mu \to 0$ beliebig groß werden können, für $\mu \to 0$ sind die Hessematrizen also

beliebig schlecht konditioniert! Und das hat natürlich Auswirkungen, wenn man Gleichungssysteme der Form

$$\nabla_x^2 Q\left(x^{(k)}, \mu\right) y^{(k)} = -\nabla_x Q\left(x^{(k)}, \mu\right),$$

beispielsweise beim Newton–Verfahren, lösen will. Glücklicherweise ist das aber beim Newton–Verfahren nun gerade wieder nicht so schlimm: setzen wir nämlich

$$z := \mu^{-1} \nabla^T g(x)\, y$$

und verwenden wir (8.13), dann erhalten wir das äquivalente Gleichungssystem

$$\begin{aligned}\left(\nabla^2 f(x) + \sum_{j=1}^{m} \frac{g_j(x)}{\mu} \nabla^2 g_j(x)\right) y + \nabla g(x)\, z &= -\nabla_x Q\,(x, \mu) \\ \nabla^T g(x)\, y - \mu\, z &= 0,\end{aligned}$$

also

$$\begin{bmatrix} \nabla^2 f(x) + \sum_{j=1}^{m} \frac{g_j(x)}{\mu} \nabla^2 g_j(x) & \nabla g(x) \\ \nabla^T g(x) & -\mu I \end{bmatrix} \begin{bmatrix} y \\ z \end{bmatrix} = \begin{bmatrix} -\nabla_x Q\,(x, \mu) \\ 0 \end{bmatrix}$$

und diese Matrix ist nun wieder gut konditioniert und somit das Gleichungssystem numerisch stabil lösbar.

8.2 Logarithmische Barrieren

Barrieren sind eine gute Methode für restringierte Optimierungsprobleme, die nur durch *Ungleichungen* beschränkt sind[219], also Probleme der Form

$$\min f(x), \qquad h(x) \geq 0, \qquad h : \mathbb{R}^n \to \mathbb{R}^m, \tag{8.14}$$

mit mindestens stetigem h. Dabei setzen wir

$$\Omega := \{x \in \mathbb{R}^n \;:\; h(x) \geq 0\} \qquad \text{sowie} \qquad \Omega^* := \{x \in \mathbb{R}^n \;:\; h(x) > 0\}$$

und nehmen an, daß $\Omega^* \neq \emptyset$, daß also mindestens ein echter **innerer Punkt** von Ω existiert.

Übung 8.1 Zeigen Sie: $\Omega^* \subseteq \Omega^\circ$, aber Gleichheit gilt im allgemeinen nicht[220]. ◊

Definition 8.4 *Sei $\Omega \subseteq \mathbb{R}^n$ wie oben. Eine Funktion $\phi : \Omega \to \mathbb{R}$ heißt **Distanzfunktion** für Ω, wenn*

1. $\phi(x) > 0$, $x \in \Omega^*$.
2. $\phi(x) = 0$, $x \in \partial^*\Omega := \Omega \setminus \Omega^*$.

[219] Im Kontext der Lagrange–Multiplikatoren sind das sogar eher die „bösen“ Probleme.
[220] Die inneren Punkte sind also nicht unbedingt dasselbe wie die Punkte im Inneren.

3. *für* $x, x' \in \Omega$

$$h(x) > h(x') \qquad \Rightarrow \qquad \phi(x) > \phi(x')$$

gilt.

Die letzte Bedingung in der obigen Definition bedeutet, daß ϕ den Abstand vom Rand im Sinne der Nebenbedingungen misst und daß ϕ umso größer ist, je „besser" die Nebenbedingungen erfüllt sind.

Beispiel 8.5 *Die natürliche Distanzfunktion zu der Nebenbedingungsfunktion* h *ist die Funktion*

$$\phi = \phi_h := \prod_{j=1}^{m} h_j.$$

Eine **Barrierefunktion** für Ω ist nun eine möglichst glatte Funktion $\psi : \Omega^* \to \mathbb{R}$ mit der Eigenschaft, daß

$$\psi(x) < \infty, \quad x \in \Omega^* \qquad \text{und} \qquad \lim_{x \to \delta^* \Omega} \psi(x) = \infty.$$

Verwendet man nämlich so eine Barrierefunktion als Strafterm, dann wird bei der Minimumssuche für $f + \psi$, ausgehend von einem Startwert $x^{(0)} \in \Omega^*$ der *strikt zulässige* Bereich Ω^* nie verlassen werden, insbesondere, wenn man ψ außerhalb von Ω glatt mit $\psi = \infty$ fortsetzt.

Beispiel 8.6 *Die* ***natürliche Barrierefunktion*** *zu einer Distanzfunktion* ϕ *ist die* ***logarithmische Barrierefunktion*** $\psi = -\log \phi$*, das heißt*

$$\psi_h = -\log \phi_h = -\log \prod_{j=1}^{m} h_j = -\sum_{j=1}^{m} \log h_j.$$

Ist außerdem

$$\alpha := \sup_{x \in \Omega} \phi(x) < \infty,$$

dann kann man ϕ *durch* $\alpha^{-1}\phi$ *ersetzen und die zugehörige Barrierefunktion* $\psi = \log \alpha - \log \phi$ *wäre sogar nichtnegativ.*

Wie vorher mit den quadratischen Straftermen betrachtet man auch jetzt wieder ein modifiziertes Optimierungsproblem, nämlich

$$\min_x P(x, \gamma) = f(x) - \gamma \log \phi_h(x) = f(x) - \gamma \sum_{j=1}^{m} \log h_j, \qquad \gamma > 0, \tag{8.15}$$

und läßt dann γ schön langsam gegen Null gehen. Dabei erzeugt man *immer* eine Folge von *strikten inneren* Punkten, denn die Funktion ϕ_h nimmt ja nur auf Ω^* endliche Werte an[221]. Die Vorgehensweise ist nun wieder wie vorher bei den Straftermen.

[221] Unter Verwendung der Konvention $\log t = -\infty$ für $t < 0$.

Algorithmus 8.7 Gegeben: *Funktion* $f \in C^1(\mathbb{R}^n)$, *Nebenbedingungen* $h \in C^1(\mathbb{R}^n)^m$.

1. *Wähle* $\gamma_1, \tau_1 \in \mathbb{R}_+$.
2. *Für* $k = 1, 2 \ldots$
   (a) *Bestimme* $x^{(k)} \in \mathbb{R}^n$, *so daß*

$$\left\| \nabla_x P\left(x^{(k)}, \gamma_k\right) \right\| \leq \tau_k. \tag{8.16}$$

   (b) *Wähle*

$$\gamma_{k+1} \in (0, \gamma_k), \qquad \tau_{k+1} \in (0, \tau_k).$$

Ergebnis: *Folge* $x^{(k)}$, *die (hoffentlich) gegen ein Minimum konvergiert.*

Eine Konvergenzanalyse solcher Barrierefunktionen ist haarig und aufwendig, so daß wir sie uns schenken. Allerdings erkennt man ganz gut, *warum* die Sache so problematisch ist. Sehen wir uns nämlich die notwendige Voraussetzung für ein Minimum von $P(x, \gamma)$ an, also

$$0 = \nabla_x P(x, \gamma) = \nabla f(x) - \gamma \sum_{j=1}^{n} \nabla \log h_j(x) = \nabla f(x) - \sum_{j=1}^{n} \frac{\gamma}{h_j(x)} \nabla h_j(x),$$

dann ist, weil $x \in \Omega^*$ ist,

$$\mu := \mu^\gamma = \left[\frac{\gamma}{h_j(x)} \ : \ j = 1, \ldots, m \right] \in \mathbb{R}_+^m$$

ein guter Kandidat für den „Ungleichungsmultiplikator" aus Satz 5.13, denn schließlich ist ja

$$\nabla f(x) - \underbrace{[\nabla h_j \ : \ j = 1, \ldots, m]}_{=\nabla h} \mu = 0;$$

Allerdings folgt aus der Definition von μ^γ, daß

$$\mu^T h(x) = \sum_{j=1}^{m} h_j(x) \frac{\gamma}{h_j(x)} = m\gamma$$

und damit ist die Bedingung (5.10) aus Satz 5.13 leider nicht erfüllt. Na gut, wenn $\gamma \to 0$ geht, dann wird das besser und besser, aber dann muß halt auch

$$\mu_j^* = \lim_{\gamma \to 0} \mu_j^\gamma = \frac{\gamma}{h_j(x^\gamma)}, \qquad j = 1, \ldots, m.$$

existieren. Das ist kein wirkliches Problem, wenn $x^* = \lim x^\gamma$ in Ω^* liegt, aber wenn das Minimum an einem Randpunkt angenommen wird, dann braucht man weitere Bedingungen an die Nebenbedingungen und die Funktion f, um Konvergenz beweisen zu können.

8.3 Erweiterte Lagrange–Multiplikatoren

Als letztes Beispiel betrachten wir eine Methode, die sich in praktischen Anwendungen besonders gut bewährt hat, nämlich die **augmented Lagrangian**, was man als **ergänzte Lagrange–Multiplikatoren** oder **erweiterte Lagrange–Multiplikatoren** übersetzen könnte; wer wirklich verstanden werden möchtes, sollte aber vielleicht den englischen Begriff wählen. Auch wenn man Ungleichungsnebenbedingungen in diesen Rahmen integrieren könnte, wollen wir[222] uns nur auf Gleichungen beschränken, also ein Optimierungsproblem der Form

$$\min f(x), \qquad g(x) = 0, \tag{8.17}$$

zu lösen versuchen. Die **Hilfsfunktion**, die wir jetzt betrachten wollen, hat die Form

$$L(x, \lambda, \mu) = f(x) - \lambda^T g(x) + \frac{1}{2\mu} \left\| g(x) \right\|_2^2, \tag{8.18}$$

wobei man sich unter λ eine Näherung für den Lagrange–Multiplikator vorzustellen hat – man „mischt“ also sozusagen Lagrange–Multiplikatoren mit quadratischen Straftermen. Der Name „erweiterte Lagrange–Funktion“ stammt übrigens daher, daß man die Funktion $L(x, \lambda) = f(x) - \lambda^T g(x)$ auch manchmal als **Lagrange–Funktion** bezeichnet[223]. Mit den schon wohlbekannten Rechnungen ergibt sich dann sofort, daß

$$\nabla_x L(x, \lambda, \mu) = \nabla f(x) - \nabla g(x)\, \lambda + \frac{1}{\mu} \nabla g(x)\, g(x), \tag{8.19}$$

und ist x^* nun eine Minimallösung von (8.18), insbesondere also $x^* \in Z_g$, dann ist

$$0 = \nabla_x L(x^*, \lambda, \mu) = \nabla f(x^*) - \nabla g(x^*)\, \lambda + \frac{1}{\mu} \nabla g(x^*)\, \underbrace{g(x^*)}_{=0} = \nabla f(x^*) - \nabla g(x^*)\, \lambda,$$

λ wäre also ein Lagrange–Multiplikator für f. Haben wir hingegen einen unzulässigen, näherungsweisen Minimalwert $\widehat{x}$ des unrestringierten Hilfsproblems (8.18) gefunden, dann ist also, unter Verwendung von (8.19),

$$0 \sim \nabla_x L(\widehat{x}, \lambda, \mu) = \nabla f(\widehat{x}) - \nabla g(\widehat{x}) \left(\lambda - \frac{1}{\mu}\, g(\widehat{x}) \right),$$

also ist

$$\widehat{\lambda} := \lambda - \frac{g(\widehat{x})}{\mu}$$

eine gute Schätzung für den Lagrange–Multiplikator. Und das können wir auch schon wieder als Basis für ein iteratives Verfahren nehmen.

[222] Schon der Einfachheit halber, ansonsten siehe (Nocedal & Wright, 1999, S. 516–518).

[223] Um die Verwirrung zu komplettieren: Im Zusammenhang mit der **Lagrange–Interpolation**, das ist, im Gegensatz zur **Hermite–Interpolation**, bei der auch Ableitungen interpoliert werden, die Interpolation von Funktionswerten an vorgegebenen Stellen, verwendet man den Begriff “Lagrange–Funktion” gerne für eine Funktion, die an einem der Interpolationspunkte den Wert 1, an allen anderen Interpolationspunkten aber den Wert 0 hat.

Algorithmus 8.8 Gegeben: *Zielfunktion* $f \in C^1(\mathbb{R}^n)$ *und Gleichungsnebenbedingungen* $g \in C^1(\mathbb{R}^n)^m$.

1. *Wähle*
$$\lambda^1 \in \mathbb{R}^m, \qquad \tau_1, \mu_1 \in \mathbb{R}_+$$
2. *Für* $k = 1, 2, \ldots$
   (a) *Bestimme* $x^{(k)} \in \mathbb{R}^n$, *so daß*
$$\left\| \nabla_x L_x \left(x^{(k)}, \lambda^k, \mu_k \right) \right\| \leq \tau_k.$$
   (b) *Setze*
$$\lambda^{k+1} = \lambda^k - \frac{g\left(x^{(k)}\right)}{\mu_k}$$
   (c) *Wähle*
$$\mu_{k+1} \in (0, \mu_k), \qquad \tau_{k+1} \in (0, \tau_k).$$

Im Gegensatz zu den einfacheren quadratischen Straftermen besteht der Reiz dieser Methode darin, daß man μ nicht beliebig verkleinern muß, sondern daß es einen Wert $\bar{\mu}$ gibt, so daß man für alle $\mu < \bar{\mu}$ bei einem lokalen Minimum landet – das läßt auf ein sinnvolles Terminieren des Verfahrens nach endlich vielen Schritten hoffen.

Satz 8.9 *Für* $f \in C^2(\mathbb{R}^n)$ *und* $g \in C^2(\mathbb{R}^n)^m$ *sei* $x^* \in Z_g$ *eine lokale Lösung von (8.17) und* λ^* *der zugehörige Multiplikator. Außerdem seien die Spalten von* $\nabla g(x^*)$ *linear unabhängig*[224] *und es sei*

$$y^T \nabla_x^2 L(x^*, \lambda^*) y := y^T \nabla_x^2 \left(f - g^T \lambda^* \right)(x^*)\, y > 0, \qquad \nabla^T g(x^*)\, y = 0, \quad y \neq 0. \tag{8.20}$$

Dann gibt es einen Wert $\bar{\mu} > 0$, *so daß für alle* $\mu < \bar{\mu}$ *der Punkt* x^* *ein striktes lokales Minimum von* $L(\cdot, \lambda^*, \mu)$ *ist.*

Bemerkung 8.10 *Die Bedingung (8.20) ist eine* hinreichende Bedingung *zweiter Ordnung für das Vorliegen eines Minimums unter Nebenbedingungen. Für Details siehe (Nocedal & Wright, 1999, Theorem 12.6, S. 345).*

Definition 8.11 *Sei* $A \in \mathbb{R}^{n \times n}$ *eine symmetrische Matrix. Wir schreiben* $A \geq 0$ *wenn* A *positiv semidefinit ist und* $A > 0$, *wenn* A *strikt positiv definit ist.*

Beweis: Wir werden zeigen, daß

$$\nabla_x L(x^*, \lambda^*, \mu) = 0 \qquad \text{und} \qquad \nabla_x^2 L(x^*, \lambda^*, \mu) > 0 \tag{8.21}$$

ist; nach Proposition 6.2 ist dann x^* ein striktes lokales Minimum von $L(\cdot, \lambda^*, \mu)$.

[224] Wie in Satz 8.2, das heißt also auch wieder, daß $m \leq n$ ist.

Der erste Teil von (8.21) ist einfach: Da $x^* \in Z_g$ und da λ^* als Lagrange–Multiplikator die Bedingung $\nabla f(x^*) - \nabla g(x^*)\,\lambda^* = 0$ erfüllt, ist nach (8.19)

$$\nabla_x L(x^*, \lambda^*, \mu) = \underbrace{\nabla f(x^*) - \nabla g(x^*)\,\lambda^*}_{=0} + \frac{1}{\mu}\nabla g(x^*)\,\underbrace{g(x^*)}_{=0} = 0$$

und zwar sogar *unabhängig* von μ.

Interessanter wird natürlich die Sache mit der zweiten Ableitung. Da

$$\begin{aligned}
\nabla_x^2 L(x, \lambda, \mu) &= \nabla_x \left(\nabla f(x) - \nabla g(x) \left(\lambda - \frac{1}{\mu}\, g(x) \right) \right) \\
&= \nabla^2 f(x) - \sum_{j=1}^{m} \lambda_j \nabla^2 g_j(x) + \frac{1}{\mu} \sum_{j=1}^{m} \left(g_j(x) \nabla^2 g_j(x) + \nabla g_j(x) \nabla^T g_j(x) \right) \\
&= \nabla_x^2 L(x, \lambda) + \frac{1}{\mu} \sum_{j=1}^{m} \left(g_j(x) \nabla^2 g_j(x) + \nabla g_j(x) \nabla^T g_j(x) \right),
\end{aligned}$$

ist wegen $x^* \in Z_g$ dann

$$\nabla_x^2 L(x^*, \lambda^*, \mu) = \nabla_x^2 L(x^*, \lambda^*) + \frac{1}{\mu} \sum_{j=1}^{m} \nabla g_j(x^*)\, \nabla^T g_j(x^*) = \nabla_x^2 L(x^*, \lambda^*) + \frac{\nabla g_* \nabla^T g_*}{\mu}. \tag{8.22}$$

Da[225]

$$\mathbb{R}^n = \ker \nabla^T g^* \oplus \nabla g_* \mathbb{R}^m$$

ist, können wir also nun ein beliebiges $0 \neq y \in \mathbb{R}^n$ als

$$y = v + \nabla g_*\, w = v + w^*, \qquad \nabla^T g^*\, v = 0, \qquad v \in \mathbb{R}^n,\ w \in \mathbb{R}^m,$$

schreiben und es ist, mit (8.22)

$$\begin{aligned}
y^T \nabla_x^2 L(x^*, \lambda^*, \mu)\, y &= (v + \nabla g_*\, w)^T \left(\nabla_x^2 L(x^*, \lambda^*) + \frac{\nabla g_* \nabla^T g_*}{\mu} \right) (v + \nabla g_*\, w) \\
&= v^T \nabla_x^2 L(x^*, \lambda^*)\, v + 2 v^T \nabla_x^2 L(x^*, \lambda^*)\, \nabla g_*\, w + \nabla^T g_*\, w\, \nabla_x^2 L(x^*, \lambda^*)\, \nabla g_*\, w \\
&\quad + \frac{1}{\mu} \Big(\underbrace{v^T \nabla g^*}_{=0}\, \underbrace{\nabla^T g^* v}_{=0} + 2 \underbrace{v^T \nabla g^*}_{=0} \nabla^T g^* \nabla g^*\, w + w^T \nabla^T g^* \nabla g^* \nabla^T g^* \nabla g^*\, w \Big) \\
&= v^T \nabla_x^2 L(x^*, \lambda^*)\, v + 2 v^T \nabla_x^2 L(x^*, \lambda^*)\, w^* + w^{*T} \nabla_x^2 L(x^*, \lambda^*)\, w^* + \frac{\left\| \nabla^T g_* \nabla g_*\, w \right\|_2^2}{\mu}
\end{aligned}$$

Nach der Voraussetzung (8.20) ist nun

$$v^T \nabla_x^2 L(x^*, \lambda^*)\, v \geq A \|v\|_2^2, \qquad A > 0,$$

sowie

$$\begin{aligned}
v^T \nabla_x^2 L(x^*, \lambda^*)\, w^* &\geq -\left| v^T \nabla_x^2 L(x^*, \lambda^*)\, w^* \right| \geq -\|v\|_2 \underbrace{\left\| \nabla_x^2 L(x^*, \lambda^*)\, \nabla g_* \right\|_2}_{=:B} \|w\|_2 \\
&\geq -B \|v\|_2 \|w\|_2, \qquad B > 0,
\end{aligned}$$

[225] Sollte aus der linearen Algebra bekannt sein!

und

$$w^{*T} \nabla_x^2 L(x^*, \lambda^*)\, w^* \geq -\underbrace{\left\| \nabla^T g_* \nabla_x^2 L(x^*, \lambda^*)\, \nabla g_* \right\|_2}_{=:C} \|w\|_2^2 = -C \|w\|_2^2, \qquad C > 0.$$

Da die Matrix ∇g_* den Maximalrang m hat ist außerdem $\nabla^T g_* \nabla g_*$ strikt positiv definit, weswegen

$$\left\| \nabla^T g_* \nabla g_*\, w \right\|_2^2 \geq D\, \|w\|_2^2, \qquad D > 0,$$

ist. Somit ist

$$\begin{aligned} y^T \nabla_x^2 L(x^*, \lambda^*, \mu)\, y &\geq A\left(\|v\|_2^2 - 2\frac{B}{A} \|v\|_2 \|w\|_2 + \frac{B^2}{A^2} \|w\|_2^2 \right) + \|w\|_2^2 \left(\frac{D}{\mu} - C - \frac{B^2}{A} \right) \\ &= A \underbrace{\left(\|v\|_2^2 - \frac{B}{A} \|v\|_2^2 \right)^2}_{\geq 0} + \|w\|_2^2 \left(\frac{D}{\mu} - C - \frac{B^2}{A} \right), \end{aligned}$$

was ≥ 0 ist, sobald

$$\mu < \bar{\mu} := \frac{D}{C + B^2/A}$$

ist. Außerdem gilt für jedes solche $\mu < \bar{\mu}$, daß

$$y^T \nabla_x^2 L(x^*, \lambda^*, \mu)\, y = 0 \qquad \Leftrightarrow \qquad v = 0, \quad w = 0.$$

□

Bei der wirklichen praktischen Implementierung im Optimierungspaket LANCELOT von Conn, Gould und Toint (Conn *et al.*, 1992) betrachtet man lokalisierte Probleme der Form

$$\min f(x), \qquad g(x) = 0, \quad a \leq x \leq b, \quad a, b \in \mathbb{R}^n,$$

die mit einem geeigneten Iterationsverfahren und Updateregeln für die Multiplikatoren, Toleranzen und Strafparameter (das μ) behandelt werden, wobei eine Unmenge von Details berücksichtigt und aufeinander abgestimmt werden müssen. Für eine erste Idee siehe (Nocedal & Wright, 1999, S. 522–523).

Denn viel größres Vertrauen muß
immer erwecken, was selber
Unabhängig von andrem den Irrtum
schlägt mit der Wahrheit.

Lukrez, *Über die Natur der Dinge*

9 Trust–Region–Verfahren

In diesem Kapitel befassen wir uns mit einer anderen Familie von Methoden zur unrestringierten Optimierung, bei der ein *quadratisches Modell* der Zielfunktion optimiert wird, aber nur Schrittweiten innerhalb eines Bereiches zugelassen werden, auf der das quadratische Modell die Zielfunktion auch halbwegs zuverlässig annähert, der sogenannten **Trust Region**.

9.1 Quadratische Modelle und wem man wo wie vertraut

Wir approximieren wieder einmal die Zielfunktion f lokal um $x \in \mathbb{R}^n$ durch ein **quadratisches Modell** der Form

$$f(x+y) \sim q(y) = f + g^T y + \frac{1}{2} y^T B y, \qquad f \in \mathbb{R},\ g \in \mathbb{R}^n,\ B \in \mathbb{R}^{n\times n}, \quad B^T = B, \tag{9.1}$$

bzw.

$$f\left(x^{(k)} + y\right) \sim q_k(y) = f_k + g_k^T y + \frac{1}{2} y^T B_k y,$$

wenn wir iterative Verfahren herleiten wollen. Dieses quadratische Modell kann man auf die verschiedensten Arten erhalten:

1. Durch *exakte* Kenntnis von $f \in C^2(\mathbb{R}^n)$ und die **Taylorformel**, indem man in (9.1)
$$f = f(x), \qquad g = \nabla f(x), \qquad B = \nabla^2 f(x),$$
setzt.

2. Durch **polynomiale Interpolation** von f. Kennt man f an $\binom{n+2}{2} = \dim \Pi_2$ Stellen $\mathscr{X} \subset \mathbb{R}^n$, dann kann man[226] ein quadratisches Polynom bestimmen $q \in \Pi_2$, das an diesen Stellen interpoliert,
$$q(x) = f(x), \qquad x \in \mathscr{X},$$
und dieses Polynom als Modell verwenden.

[226] Hoffentlich ... Hier spielt nämlich in mehr als zwei Variablen Geometrie eine ganz wesentliche Rolle, siehe (Sauer, 2006).

3. Durch **Least–Squares-Approximation** von f. Kennt man f an *mindestens* $\dim \Pi_2$ Stellen $\mathscr{X} \subset \mathbb{R}^n$, dann sucht man ein Polynom $q \in \Pi_2$, so daß

$$\sum_{x \in \mathscr{X}} (q(x) - f(x))^2 = \min_{q' \in \Pi_2} \sum_{x \in \mathscr{X}} (q'(x) - f(x))^2$$

Die beiden letzten Ansätze haben den Vorteil, daß sie nicht nur für differenzierbare oder zweimal differenzierbare Funktionen verwendet werden können, sondern wir nur die Möglichkeit haben müssen, die Funktion f an gewissen Punkten auszuwerten.

Bemerkung 9.1 *So einfach ist es aber leider doch wieder nicht mit der Erzeugung eines quadratischen Modells durch Interpolation. Es gibt da einiges an Problemen:*

1. *Im Gegensatz zum univariaten Fall spielt die Geometrie der Punkte in $\mathscr{X}$ eine Rolle bereits bei der Frage nach der (eindeutigen)* Lösbarkeit *des Interpolationsproblems. So ist es beispielsweise in zwei Variablen nicht möglich, einen quadratischen Interpolanten an $\dim \Pi_2 = 6$ Punkte zu finden, wenn diese alle auf dem Einheitskreis liegen, denn dann verschwindet ja das quadratische Polynom $x^2 + y^2 - 1$ an all diesen Punkten.*

2. *Auch die Frage, inwieweit so ein Interpolationspolynom überhaupt eine gute Näherung an f darstellt, also Fehlerabschätzungen der Form $\|f - q\| \leq \ldots$ hängen selbst für hinreichende oft differenzierbares f von der Geometrie der Punkte ab, beispielsweise vom Quotienten aus Umkreis- und Inkreisradius, siehe (Ciarlet & Raviart, 1972), und das kann beliebig schlecht werden.*

3. *Auch algorithmisch ist die Polynominterpolation nicht so ganz einfach, für effiziente und stabile Implementierungen muß man sich schon ein bißchen was überlegen, siehe z.B. (Sauer, 1995; Boor, 2000).*

4. *Mehr Information über Trust–Region–Verfahren unter Verwendung polynomialer Interpolation findet sich in (Conn* et al.*, 1997).*

Aber zurück zur Optimierung! Bei einem **Trust–Region–Verfahren** erzeugt man zusätzlich zu einer Folge $x^{(k)} \in \mathbb{R}^n$ von Punkten[227] eine Folge $r_k > 0$ von Radien; die **Trust Region** $T_k = B\left(x^{(k)}, r_k\right)$ ist dann der Kreis vom Radius r_k um $x^{(k)}$ und dieser Radius wird die Schrittweitensteuerung beeinflussen.
Zuerst einmal bestimmt man jetzt $y^{(k)}$ als Optimalstelle des quadratischen Modells innerhalb der Trust Region, also als Lösung von

$$\min_y q_k(y), \qquad \|y\| \leq r_k.$$

Dann überprüft man, inwieweit das quadratische Modell wirklich zutreffend war, indem man den Quotienten

$$\rho_k := \frac{f\left(x^{(k)}\right) - f\left(x^{(k)} + y^{(k)}\right)}{q_k(0) - q_k\left(y^{(k)}\right)}$$

[227] Die natürlich gegen das gewünschte Extremum konvergieren sollte.

bestimmt. Der Nenner von ρ_k ist wegen der Definition von $y^{(k)}$ als *Minimalstelle* übrigens immer positiv; wäre also $\rho_k < 0$, dann muß $f\left(x^{(k)}\right) < f\left(x^{(k)} + y^{(k)}\right)$ gelten und das Modell war offenbar nicht wirklich gut: Die Minimalstelle von q entspricht noch nicht einmal einer Verbesserung von f. Ist hingegen das quadratische Modell exakt, also $f(x + y) = q(y)$, dann ergibt sich natürlich $\rho_k = 1$. Und nun entscheidet man anhand von ρ_k:

1. Ist ρ_k klein und insbesondere negativ, dann verkleinert man die Trust Region und versucht es nochmals[228] mit $x^{(k+1)} = x^{(k)}$.

2. Ist ρ_k groß und hat man für $y^{(k)}$ die zulässige Maximalschrittweite r_k voll ausgenutzt, dann vergrößert man den Radius der Trust Region und setzt $x^{(k+1)} = x^{(k)} + y^{(k)}$.

3. Ist ρ_k weder wirklich groß noch wirklich klein, dann beläßt man den Radius wie er ist[229] und setzt wieder $x^{(k+1)} = x^{(k)} + y^{(k)}$.

Natürlich muß man die Frage was groß und was klein ist und wie man vergrößert und verkleinert spezifizieren. In (Nocedal & Wright, 1999) heißt das $\rho_k < \frac{1}{4}$ für "klein", und dann wird $r_{k+1} = \frac{1}{4} r_k$ gesetzt und $\rho > \frac{3}{4}$ für "groß", wobei der Radius allerdings "nur" verdoppelt wird. Außerdem kann man noch einen Maximalradius r^* vorgeben, der nie überschritten werden darf, was zur Regel

$$r_{k+1} = \min\left(2r_k, r^*\right) \tag{9.2}$$

führt.

9.2 Wahl der Richtung

Als erstes wollen wir uns zwei Methoden zur schnellen näherungsweisen Bestimmung der Minimallösung des Modellproblems und damit der **Suchrichtung** $y^{(k)}$ ansehen.

Definition 9.2 *Der **Cauchy–Punkt** y_c ist definiert als $y_c := \tau^* y^*$, wobei y^*, τ^* Lösungen der (sequentiellen) Optimierungsprobleme*

$$\min_{y \in \mathbb{R}^n} f + g^T y = q(0) + \nabla^T q(0)\, y, \quad \|y\|_2 \le r, \qquad \min_{\tau \in \mathbb{R}_+} q\left(\tau y^*\right), \quad \|\tau y^*\|_2 \le r.$$

Wählt man $x^{(k+1)} = x^{(k)} + y^{(k)}$, wobei $y^{(k)}$ Cauchy–Punkt bezüglich g_k ist, dann ergibt das zwar ein Trust–Region–Verfahren, bei dem ein vernünftiger Abstieg gewählt ist, siehe Lemma 9.7, aber da

$$y^{(k)} = r_k \frac{g}{\|g\|_2} = r_k \frac{\nabla q(0)}{\left\|\nabla q(0)\right\|_2}$$

[228] Das quadratische Modell bleibt dabei unverändert. Man könnte natürlich hier auch einen "Modell–Update" in Betracht ziehen, bei dem z.B. neue, nähere Interpolationspunkte gewählt werden.

[229] Es funktioniert ja „so halbwegs".

ist, erhalten wir, bis auf die Schrittweitensteuerung, eine Variante des steilsten Abstiegs und von dem wissen wir ja, siehe Lemma 6.9 und Beispiel 6.10, daß er nicht so grandios funktioniert.

Heuristisch besser wäre es mit Sicherheit, wie beim Newton–Verfahren die **Newtonrichtung**

$$y = \left(\nabla^2 q\right)^{-1} \nabla q(0) = B^{-1} g$$

zu wählen[230], oder zumindest diese Größe bei der Richtungsbestimmung in Betracht zu ziehen. Zu diesem Zweck sehen wir uns einmal an, wie die Lösung y^r des Minimierungsproblems

$$\min_y q(y) = f + g^T y + \frac{1}{2} y^T B y, \qquad \|y\|_2 \le r, \tag{9.3}$$

eigentlich aussieht. Ist $r = \infty$, betrachten wir also das *unrestringierte* Problem, so kennen wir die Lösung: $y^\infty = -B^{-1}g$, siehe Beispiel 6.8. Das heißt aber, daß $y^r = y^\infty = -B^{-1}g$ gilt, so lange $r \ge \|y^\infty\|_2$ ist. Andererseits liefert uns aber die **Taylorformel**, genauer, die Tatsache, daß der quadratische Anteil nur mit der Größenordnung $\|y\|_2^2$, der lineare Anteil aber von der Größenordnung $\|y\|_2$ beiträgt, daß

$$y^0 := \lim_{r \to 0} y^r = -g = -\nabla q(0).$$

ist. Da nutzt man für die sogenannte **Dogleg–Methode**[231], bei der man in der steilsten Abstiegsrichtung y^0 aus dem Punkt 0 herausfährt, aber dafür sorgt, daß man dennoch in der unrestringierten Optimallösung y^r ankommt. Dazu kombinieren wir die Richtungsvektoren des steilsten Abstiegs und der **Newton–Richtung**

$$y^0 := -\frac{g^T g}{g^T B g}\, g \qquad \text{und} \qquad y^1 := -B^{-1} g$$

in eine stückweise lineare Funktion

$$y(t) := \begin{cases} t\, y^0, & 0 \le t \le 1 \\ (2-t)\, y^0 + (t-1)\, y^1, & 1 \le t \le 2, \end{cases} \qquad t \in [0,2]$$

die die Eigenschaft hat, daß $y(0) = y^0$ und $y(2) = y^1$. Und dann suchen wir das Minimum entlang dieses geknickten Streckenzugs, welches immer eindeutig bestimmt ist.

Proposition 9.3 *Ist* B *positiv definit und ist* $\left\|y^1\right\| \ge r$, *dann gibt es genau einen Wert* $t \in [0,2]$, *so daß* $\left\|y(t)\right\|_2 = r$ *und für genau diesen Wert ist die Funktion* $q\left(y(t)\right)$ *minimal unter der Nebenbedingung* $\left\|y(t)\right\|_2 \le r$.

[230] Denn mit dieser Richtung, die $q(0)$ mit dem Minimum verbindet, wird das quadratische Optimierungsproblem in einem Iterationsschritt *global* gelöst.

[231] Encyclopedia Britannica: „***dog-leg** a thing that bends sharply, in particular a sharp bend in a road or route.*“

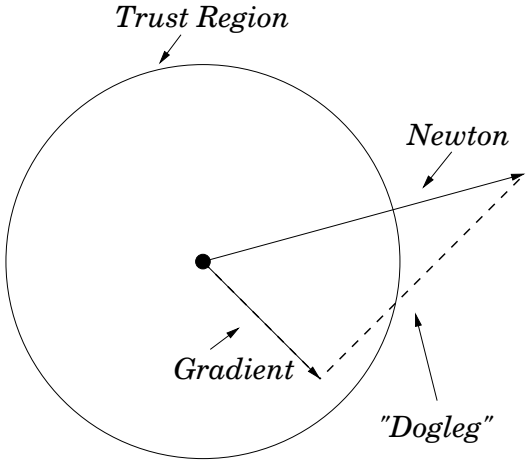

Abbildung 9.1: Der Pfad des "Dogleg"–Verfahrens und die Trust Region.

Beweis: Wir beweisen sogar viel mehr, wir zeigen nämlich, daß

$$t < t' \qquad \Rightarrow \qquad \begin{cases} \left\|y(t)\right\|_2 \le \left\|y(t')\right\|_2 \\ q\left(y(t)\right) \ge q\left(y(t')\right) \end{cases} \tag{9.4}$$

Daß (9.4) für $t \in [0,1]$ gilt, liegt an der Wahl von y^0: Das Minimum von $q\left(t\, y^0\right)$ wird ja gerade für $t = g^T g / g^T B g$ angenommen. Interessant ist also nur der Fall $t \in [1,2]$; schreiben wir $t = 1 + s$, $s \in [0,1]$, dann ist

$$\begin{aligned} \left\|y(t)\right\|_2^2 &= \left\|y(1+s)\right\|_2^2 = \left\|y^0 + s\left(y^1 - y^0\right)\right\|_2^2 \\ &= \left\|y^0\right\|_2^2 + 2s\left(y^1 - y^0\right)^T y^0 + s^2 \left\|y^1 - y^0\right\|_2^2 \end{aligned}$$

und daher

$$\begin{aligned} &\frac{d}{ds}\frac{1}{2}\left\|y(1+s)\right\|_2^2 = \left(y^1 - y^0\right)^T y^0 + s\left\|y^1 - y^0\right\|_2^2 \ge \left(y^1 - y^0\right)^T y^0 \\ &= \frac{g^T g}{g^T B g}\, g^T\left(B^{-1} g - \frac{g^T g}{g^T B g}\, g\right) = \underbrace{\frac{g^T g}{g^T B g}}_{\ge 0}\, \underbrace{g^T B^{-1} g}_{\ge 0}\left(1 - \frac{g^T g}{g^T B g}\,\frac{g^T g}{g^T B^{-1} g}\right). \end{aligned}$$

Seien $0 < \lambda_1 \le \cdots \le \lambda_n$ die Eigenwerte von B, $b_1, \ldots, b_n$ die orthonormalen Eigenvektoren dazu und

$$g = \sum_{j=1}^{n} g_j b_j,$$

dann ist

$$g^T B g = \sum_{j,k=1}^{n} g_j g_k \underbrace{b_j^T B b_k}_{=\lambda_k \delta_{jk}} = \sum_{j=1}^{n} \lambda_j g_j^2 \qquad \text{und} \qquad g^T B^{-1} g = \sum_{j=1}^{n} \lambda_j^{-1} g_j^2$$

und da für $a, b > 0$

$$\frac{a^2+b^2}{ab} = \frac{(a-b)^2+2ab}{ab} \geq \frac{2ab}{ab} = 2,$$

erhalten wir

$$\begin{aligned}\left(g^T B g\right)\left(g^T B^{-1} g\right) &= \sum_{j,k=1}^{n} \frac{\lambda_j}{\lambda_k} g_j^2\, g_k^2 = \sum_{j=1}^{n} g_j^4 + \sum_{1\leq j<k\leq n} \underbrace{\left(\frac{\lambda_j}{\lambda_k}+\frac{\lambda_k}{\lambda_j}\right)}_{=\frac{\lambda_j^2+\lambda_k^2}{\lambda_j\,\lambda_k}\geq 2} g_j^2\, g_k^2 \\ &\geq \sum_{j=1}^{n} g_j^4 + \sum_{1\leq j<k\leq n} 2 g_j^2\, g_k^2 = \sum_{j,k=1}^{n} g_j^2 g_k^2 = \|g\|_2^4,\end{aligned}$$

also ist $\frac{d}{ds}\left\|y(1+s)\right\|_2^2 \geq 0$, was den ersten Teil von (9.4) liefert.
Für den zweiten Teil von (9.4) betrachten wir $y(1+s) = y^0 + s\,(y^1 - y^0)$ und

$$\begin{aligned}&\frac{d}{ds} q\left(y(1+s)\right) \\ &= \frac{d}{ds}\left(f + g^T\left(y^0 + s\left(y^1-y^0\right)\right) + \frac{1}{2}\left(y^0 + s\left(y^1-y^0\right)\right)^T B\left(y^0 + s\left(y^1-y^0\right)\right)\right) \\ &= g^T\left(y^1-y^0\right) + \left(y^1-y^0\right)^T B y^0 + s\underbrace{\left(y^1-y^0\right)^T B\left(y^1-y^0\right)}_{>0} \\ &\leq \left(y^1-y^0\right)^T\left(g + B y^0\right) + \left(y^1-y^0\right)^T B\left(y^1-y^0\right) \\ &= \left(y^1-y^0\right)^T\left(g + B y^1\right) = \left(y^1-y^0\right)^T\left(g - BB^{-1}g\right) = 0,\end{aligned}$$

weswegen wir ständig abfallende Werte erzeugen. □

Der Beweis zeigt: das Minimum auf dem "Dogleg"–Pfad wird genau dort angenommen, wo dieser Pfad die Trust Region verläßt!

9.3 Exakte Lösungen des quadratischen Problems

Cauchy–Punkte und Dogleg–Methode sind nette, aber eher heuristische Ansätze, um eine näherungsweise Optimallösung für das quadratische **Modellproblem** (9.3) zu bestimmen. Besser wäre es abe, mit der *exakten* Lösung zu arbeiten. Und die kann man zumindest beschreiben.

Satz 9.4 *Ein Vektor $y^* \in \mathbb{R}^n$ ist genau dann Lösung von (9.3), wenn $\|y^*\| \leq r$ und es eine Zahl $\lambda \geq 0$ gibt, so daß*

$$\begin{aligned}(B + \lambda I)\, y^* &= -g, &(9.5)\\ \lambda\left(r - \|y^*\|_2\right) &= 0, &(9.6)\\ (B + \lambda I) &\geq 0. &(9.7)\end{aligned}$$

Als Hilfsmittel ein bißchen Analysis quadratischer Funktionen.

Lemma 9.5 *Sei* $B \in \mathbb{R}^{n\times n}$ *symmetrisch und* $q(y) = g^T y + \frac{1}{2} y^T B y$. *Dann gilt:*

1. q *besitzt genau dann ein* globales *Minimum, wenn* B *positiv semidefinit ist und* $g \in B\,\mathbb{R}^n$.
2. q *hat genau dann ein* eindeutiges *globales Minimum, wenn* B *positiv definit ist.*
3. *Ist* B *positiv semidefinit, dann ist jede Lösung* y *von* $By = g$ *ein globales Minimum von* q.

Beweis: 1): Hat q ein Minimum y^*, dann muß

$$0 = \nabla q\,(y^*) = g + By^* \qquad \text{und} \qquad 0 \le \nabla^2 q\,(y^*) = B$$

weswegen $g \in B\,\mathbb{R}^n$ liegen und $B \ge 0$ gelten muß. Für die Umkehrung wählen wir ein $v \in \mathbb{R}^n$, so daß $g = -Bv$ – nach den Voraussetzungen $g \in B\,\mathbb{R}^n$ muß das ja funktionieren. Dann ist, für beliebiges $w \in \mathbb{R}^n$

$$\begin{aligned} q\,(v+w) &= g^T\,(v+w) + \frac{1}{2}\,(v+w)^T\,B\,(v+w) \\ &= g^T v + g^T w + \frac{1}{2} v^T B v + \underbrace{v^T B}_{=g^T} w + \frac{1}{2}\underbrace{v^T B v}_{\ge 0} \ge \underbrace{g^T v \frac{1}{2} v^T B v}_{=v} + g^T w - g^T w \\ &= q(v), \end{aligned}$$

womit v ein Minimum sein muß, was 3) im Übrigen gleich miterledigt.
2): Ist y^* ein striktes Minimum, so muß nach Proposition 6.2 $0 < \nabla^2 q\,(y^*) = B$ sein und umgekehrt ist für eine strikt positiv definite Matrix B ja $B\,\mathbb{R}^n = \mathbb{R}^n$ und in obiger Rechnung gilt die strikte Ungleichung. □

Beweis von Satz 9.4: Ohne Einschränkung nehmen wir an, daß $q(0) = 0$ ist – für die Suche nach dem Minimum ist der konstante Term ja bekanntlich irrelevant. Wir beginnen mit "⇐" und nehmen an, es existiere ein $\lambda \ge 0$, das (9.5)–(9.7) erfüllt. Zusammen mit Lemma 9.5 ergeben (9.5) und (9.7), daß y^* ein globales Minimum der Funktion

$$q_\lambda(y) := g^T y + \frac{1}{2}\, y^T\,(B + \lambda I)\, y = \underbrace{g^T y + \frac{1}{2} y^T B y}_{=q(y)} + \frac{\lambda\,\|y\|_2^2}{2}$$

ist, es gilt also für alle $y \in \mathbb{R}^n$, daß

$$q\,(y^*) + \frac{\lambda}{2}\,\|y^*\|_2^2 \le q\,(y) + \frac{\lambda}{2}\,\|y\|_2^2, \tag{9.8}$$

also für alle y mit $\|y\|_2 \le r$

$$\begin{aligned} q\,(y) &\ge q\,(y^*) + \frac{\lambda}{2}\left(\|y^*\|_2^2 - \|y\|_2^2\right) = q\,(y^*) + \frac{\lambda}{2}\left(\|y^*\|_2^2 - r + r - \|y\|_2^2\right) \\ &= q\,(y^*) + \frac{1}{2}\underbrace{\lambda\left(\|y^*\|_2^2 - r^2\right)}_{=0} + \frac{1}{2}\,\underbrace{\lambda\left(r^2 - \|y\|_2^2\right)}_{\ge 0} \ge q\,(y^*), \end{aligned}$$

weswegen y^* eine Lösung von (9.3) sein muß.
Für die Umkehrung, "⇒", sei y^* eine Lösung von (9.3). Ist $\|y^*\|_2 < r$, dann muß nach (9.6) $\lambda = 0$ sein und dann ergibt sich (9.5) aus $0 = \nabla q(y^*) = g + By^*$ sowie (9.7) aus $0 \le \nabla^2 q(y^*) = B$. Interessant wird es also, wenn $\|y^*\|_2 = r$ ist, dann müssen wir die Existenz des ominösen $\lambda > 0$ nachweisen. Nun ist aber y^* in diesem Fall auch eine Lösung des restringierten Optimierungsproblems

$$\min_y q(y), \qquad \underbrace{\frac{1}{2}\left(\|y\|_2^2 - r^2\right)}_{=:g(y)} = 0,$$

und nach unserem Multiplikatoren–Satz 5.13 muß es ein $\lambda \in \mathbb{R}$ geben, so daß[232]

$$0 = \nabla q(y^*) + \nabla g(y^*)\lambda = g + By^* + \lambda y^* = g + (B + \lambda I)\, y^*,$$

also muß $g = -(B + \lambda I)\, y^*$ gelten. Unter Verwendung von (9.8) gilt wegen der Minimalität von y^*

$$\|y\|_2 = r \qquad \Rightarrow \qquad q(y) \ge q(y^*) + \frac{\lambda}{2}\left(\|y^*\|_2^2 - \|y\|_2^2\right),$$

also

$$\begin{aligned}
0 &\le q(y) - q(y^*) - \frac{\lambda}{2}\left(\|y^*\|_2^2 - \|y\|_2^2\right) \\
&= g^T y + \frac{1}{2} y^T B y - g^T y^* - \frac{1}{2} y^{*T} B y^* - \frac{1}{2} y^{*T} (\lambda I)\, y^* + \frac{1}{2} y^T (\lambda I)\, y \\
&= g^T (y - y^*) + \frac{1}{2} (y - y^*)^T (B + \lambda I)(y - y^*) + y^T (B + \lambda I)\, y^* - y^{*T} (B + \lambda I)\, y^* \\
&= -(y - y^*)^T (B + \lambda I)\, y^* + \frac{1}{2} (y - y^*)^T (B + \lambda I)(y - y^*) + (y - y^*)^T (B + \lambda I)\, y^* \\
&= \frac{1}{2} (y - y^*)^T (B + \lambda I)(y - y^*),
\end{aligned}$$

also ist $(B + \lambda I)$ positiv semidefinit, weil die Vektoren

$$\left\{ \pm \frac{y - y^*}{\|y - y^*\|_2} \;:\; \|y\|_2 = r \right\}$$

eine dichte Teilmenge der Einheitskugel bilden[233]. Bleibt zu zeigen, daß $\lambda \ge 0$ ist. Da (9.5) und (9.7) erfüllt sind, sagt uns Lemma 9.5, 3), daß y^* ein globales Minimum von

$$q_\lambda(y) = g^T y + \frac{1}{2} y^T (B + \lambda I)\, y$$

ist. Könnten wir nun nur $\lambda < 0$ wählen, dann liefert uns wieder einmal (9.8), daß für alle $y \in \mathbb{R}^n$, $\|y\|_2 > r = \|y^*\|_2$, daß

$$q(y) \ge q(y^*) + \underbrace{\overbrace{\frac{\lambda}{2}}^{<0} \overbrace{\left(\|y^*\|_2^2 - \|y\|_2^2\right)}^{<0}}_{>0} > q(y^*),$$

[232] Und hier ersetzen wir das λ in (5.9) durch $-\lambda$.
[233] Nur den Punkt y^* erreichen wir so nicht, der ist der Pol bzw. die Singularität der Parametrisierung.

und da y^* schon das Minimum auf $\{y : \|y\| \le r\}$ war, ist also y^* ein *globales* Minimum von q. Nach Lemma 9.5, 1), wäre dann aber $g = -By^*$ und B wäre positiv semidefinit und wir könnten, im Widerspruch zu unserer Annahme, eben doch $\lambda = 0$ wählen. □

Jetzt können wir also mit Hilfe von Satz 9.4 das lokalisierte Optimierungsproblem (9.3) in Angriff nehmen:

1. Wir bestimmen zuerst y als Lösung von[234] $By = -g$ und testen, ob $\|y\| < r$. Wenn ja, dann können wir nach (9.6) $\lambda = 0$ wählen, und y ist die gesuchte Lösung, außerdem ist B positiv semidefinit.

2. Ansonsten müssen wir einen Wert $\lambda > 0$ bestimmen, so daß $B + \lambda I$ positiv semindefinit, besser (strikt) positiv definit, ist und dann $(B + \lambda I)\, y(\lambda) = -g$ lösen. Allerdings, und das macht die Sache interessant, muß gleichzeitig $\left\|y(\lambda)\right\| = r$ gelten.

Schauen wir uns also mal an, warum es so ein λ immer geben muß. Da B eine **symmetrische Matrix** ist, gibt es eine **orthogonale Matrix** $Q \in \mathbb{R}^{n\times n}$, $Q^TQ = QQ^T = I$, das heißt, eine Matrix mit orthonormalen[235] Spaltenvektoren $q_j \in \mathbb{R}^n$, $j = 1, \ldots, n$, so daß

$$Q^TBQ = \Lambda = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}, \qquad \lambda_1 \le \cdots \le \lambda_n,$$

wobei nicht unbedingt $\lambda_1 \ge 0$ gelten muß – wir haben *nicht* vorausgesetzt, daß $B \ge 0$ sein soll! Ist nun $\lambda > 0$ so gewählt, daß

$$B + \lambda I = Q\Lambda Q^T + \lambda QQ^T = Q\,(\Lambda + \lambda I)\,Q^T$$

invertierbar ist[236], dann ist, mit $Q = \sum q_j\, e_j^T$,

$$\begin{aligned}
y(\lambda) &= -(B + \lambda I)^{-1}\, g = -Q\,(\Lambda + \lambda I)^{-1}\, Q^T g \\
&= -\left(\sum_{j=1}^{n} q_j\, e_j^T\right)(\Lambda + \lambda I)^{-1}\left(\sum_{k=1}^{n} q_k\, e_k^T\right)^T g \\
&= -\sum_{j,k=1}^{n} q_j\, e_j^T \underbrace{\begin{bmatrix} (\lambda_1+\lambda)^{-1} & & \\ & \ddots & \\ & & (\lambda_n+\lambda)^{-1} \end{bmatrix}}_{\delta_{jk}(\lambda_j+\lambda)^{-1}} e_k\, q_k^T g \\
&= -\sum_{j=1}^{n} \frac{q_j\, q_j^T}{\lambda_j + \lambda}\, g = -\sum_{j=1}^{n} \frac{q_j^T\, g}{\lambda_j + \lambda}\, q_j
\end{aligned}$$

[234] Hier können wir unser numerisches „Lieblingsverfahren" verwenden. Symmetrische Matrizen, vor allem dann, wenn sie auch noch positiv semidefinit sind, sind ja dankbare Kandidaten für die **Cholesky–Zerlegung** und für das iterative **Gauß–Seidel–Verfahren**, siehe (Golub & van Loan, 1996).

[235] Ja, da ist die Terminologie (wieder einmal) ein wenig inkonsistent.

[236] Was, unabhängig von Λ für alle $\lambda \in \mathbb{R}$, abgesehen von endlich vielen Ausnahmen, gilt.

was uns, wegen der Orthogonalität der q_j

$$\left\|y(\lambda)\right\|_2^2 = \sum_{j,k=1}^{n} \frac{q_j^T g}{\lambda_j + \lambda} \frac{q_k^T g}{\lambda_k + \lambda} \underbrace{q_j^T q_k}_{=\delta_{jk}} = \sum_{j=1}^{n} \frac{\left(q_j^T g\right)^2}{(\lambda_j + \lambda)^2} \tag{9.9}$$

liefert. Ist nun $q_1^T g \neq 0$, dann hat diese Funktion eine Singularität an $\lambda = -\lambda_1$, ist aber für $\lambda \in (-\lambda_1, \infty)$ wohldefiniert und erfüllt $\left\|y(\lambda)\right\|_2 \to 0$ für $\lambda \to \infty$. Mit anderen Worten:

Ist $q_1^T g \neq 0$, *dann gibt es ein* $\lambda \in (-\lambda_1, \infty)$, *so daß* $\left\|y(\lambda)\right\|_2 = r$.

Was aber passiert, wenn $q_1^T g = 0$ ist? Auch kein Beinbruch, denn dann beginnt eben erst bei $j = 2$, oder, wenn allgemein $q_1^T g = \cdots = q_k^T g = 0$, eben bei $j = k+1$ und wir finden dann halt ein $\lambda \in (-\lambda_{k+1}, \infty)$, das die gewünschte Eigenschaft hat. Allerdings: Die positive Semidefinitheit setzt immer noch voraus, daß $\lambda \geq -\lambda_1$ ist.
Zur Berechnung von λ könnte (und wird) man nun wieder das **Newton–Verfahren** verwenden, um eine Nullstelle von $F(\lambda) = \left\|y(\lambda)\right\|_2 - r$ zu berechnen. Dabei taucht aber ein kleines Problem auf: Für $\lambda \sim -\lambda_1$ ist $F(\lambda) \sim (\lambda + \lambda_1)^{-1}$, was uns noch nicht einmal lokale Konvergenz des Newton–Verfahrens garantiert, denn F' und F'' sind in einer Umgebung von $-\lambda_1$ unbeschränkt. Das ist aber halb so wild, dann betrachten wir eben

$$F(\lambda) = \frac{1}{\left\|y(\lambda)\right\|} - \frac{1}{r}, \tag{9.10}$$

die sich um $-\lambda_1$ wie $\lambda + \lambda_1 + C$ verhält – also wesentlich anständiger. Für die Newton–Iteration

$$\lambda_{k+1} = \lambda_k - \frac{F(\lambda_k)}{F'(\lambda_k)}, \qquad k \in \mathbb{N}_0,$$

brauchen wir also die Ableitung F', die sich mit Hilfe von (9.9) als

$$\begin{aligned} F'(\lambda) &= \frac{d}{d\lambda}\left(\left\|y(\lambda)\right\|_2^2\right)^{-1/2} = -\frac{1}{2}\left(\left\|y(\lambda)\right\|_2^2\right)^{-3/2} \frac{d}{d\lambda}\left\|y(\lambda)\right\|_2^2 \\ &= -\frac{1}{2}\left(\left\|y(\lambda)\right\|_2^2\right)^{-3/2} \frac{d}{d\lambda}\left(\sum_{j=1}^{n} \frac{\left(q_j^T g\right)^2}{(\lambda_j + \lambda)^2}\right) = -\frac{1}{2}\left(\left\|y(\lambda)\right\|_2^2\right)^{-3/2} \left(\sum_{j=1}^{n} -2\frac{\left(q_j^T g\right)^2}{(\lambda_j + \lambda)^3}\right) \\ &= \left\|y(\lambda)\right\|_2^{-3} \left(\sum_{j=1}^{n} \frac{\left(q_j^T g\right)^2}{(\lambda_j + \lambda)^3}\right) \end{aligned}$$

bestimmen läßt. Für den zweiten Term sei $B + \lambda I = G_\lambda^T G_\lambda$ die **Cholesky–Zerlegung** von $B + \lambda I$ und wir setzen $z(\lambda) = G_\lambda^{-T} y(\lambda)$. Dann ist

$$\left\|z(\lambda)\right\|_2^2 = \left(G_\lambda^{-T} y(\lambda)\right)^T \left(G_\lambda^{-T} y(\lambda)\right) = y^T \underbrace{G_\lambda^{-1} G_\lambda^{-T}}_{=(B+\lambda I)^{-1}} y$$

$$
\begin{aligned}
&= \left(-(B+\lambda I)^{-1} g\right)^T (B+\lambda I)^{-1} \left(-(B+\lambda I)^{-1} g\right) = g^T (B+\lambda I)^{-3} g \\
&= g^T Q \underbrace{\left(Q^T (B+\lambda I)^{-1} Q\right)^3}_{=(\Lambda+\lambda I)^{-3}} Q^T g \\
&= \left(\sum_{j=1}^{n} q_j^T g\, e_j\right)^T \begin{bmatrix} (\lambda_1+\lambda)^{-3} & & \\ & \ddots & \\ & & (\lambda_n+\lambda)^{-3} \end{bmatrix} \left(\sum_{j=1}^{n} q_j^T g\, e_j\right) \\
&= \sum_{j=1}^{n} \frac{\left(q_j^T g\right)^2}{(\lambda_j+\lambda)^3}
\end{aligned}
$$

und somit

$$
\frac{F(\lambda)}{F'(\lambda)} = \left(\frac{1}{\|y(\lambda)\|_2} - \frac{1}{r}\right) \frac{\|y(\lambda)\|_2^3}{\|z(\lambda)\|_2^2} = \left(\frac{\|y(\lambda)\|_2}{\|z(\lambda)\|_2}\right)^2 \frac{r - \|y(\lambda)\|_2}{r},
$$

was sich nun ganz gut als Newton–Iteration verwenden läßt:

$$
\lambda_{k+1} = \lambda_k - \left(\frac{\|y(\lambda_k)\|_2}{\|z(\lambda_k)\|_2}\right)^2 \frac{r - \|y(\lambda)\|_2}{r}, \qquad k \in \mathbb{N}_0. \tag{9.11}
$$

Um dafür zu sorgen, daß $B + \lambda_k I$ auch immer positiv definit ist, empfiehlt es sich, mit *großen* Startwerten von λ zu beginnen, indem man z.B.

$$
\lambda_0 \geq \rho(B) := \max\{|\mu| \ : \ \ker(B - \mu I) \neq \{0\},\ \mu \in\}
$$

wählt, wobei $\rho(B)$ den **Spektralradius** der Matrix B bezeichnet; für diesen Wert gibt es Abschätzungen, die man mit verhältnismäßig geringem Aufwand berechnen kann, die *exakte* Berechnung des Spektralradius kann extrem aufwendig werden.

9.4 Konvergenz von Trust–Region–Verfahren

Wir zeigen nun, daß die Trust–Region–Verfahren unter bestimmten Voraussetzungen tatsächlich gegen eine Minimallösung konvergieren. Dazu nehmen wir zuerst an, daß wir den linearen Teil des quadratischen Modells *exakt* wählen, d.h., $f_k = f\left(x^{(k)}\right)$ und $g_k = \nabla f\left(x^{(k)}\right)$, also

$$
q_k(y) = f\left(x^{(k)}\right) + \nabla^T f\left(x^{(k)}\right) y + \frac{1}{2} y^T B_k y. \tag{9.12}
$$

Außerdem seien $0 < \rho_- < \rho_+ < 1$ die Schwellenwerte für ρ_k, nach denen entschieden wird, ob ρ_k als „groß“ ($> \rho_+$) oder als „klein“ ($< \rho_-$) angesehen wird.

Satz 9.6 *Ist* $f \in C^1(\mathbb{R}^n)$ *nach unten beschränkt und gibt es eine Konstante* $\beta > 0$, *so daß* $\|B_k\|_2 \leq \beta$, $k \in \mathbb{N}_0$, *dann gilt für das Trust–Region–Verfahren mit den exakten Lösungen von (9.3), daß*

$$\liminf_{k\to\infty} \left\|\nabla f\left(x^{(k)}\right)\right\| = 0. \tag{9.13}$$

Zuerst halten wir fest, daß Cauchy–Punkte und alles, was besser als diese ist, für eine spürbare Verbesserung des Modells sorgen.

Lemma 9.7 *Für das quadratische Modell (9.12) und den zugehörigen Cauchy–Punkt* $y = y_c(r)$ *gilt*

$$q_k(0) - q_k(y) \geq \frac{1}{2}\left\|\nabla f(x)\right\|_2 \min\left(r, \frac{\left\|\nabla f(x)\right\|_2}{\|B\|_2}\right) \tag{9.14}$$

Da "Dogleg" und exakte Lösung ja nur zu Verbesserung des Ergebnisses führen, das man mit Hilfe des Cauchy–Punkts gewinnt, erhalten wir unmittelbar das folgende Resultat.

Korollar 9.8 *Die Abschätzung (9.14) gilt auch für* $y(r)$, *das mit Hilfe der Dogleg–Methode oder durch exakte Lösung des Modellproblems erhalten wird.*

Beweis von Lemma 9.7: Wir schreiben $g = \nabla f(x)$, was uns die Richtung $y^* = -r\frac{g}{\|g\|_2}$ liefert. Dann ist

$$q(\tau y^*) - q(0) = -\tau\, r\|g\|_2 + \tau^2 \frac{r^2}{\|g\|_2^2} g^T B g.$$

Dieser Ausdruck ist monoton fallend in τ falls $g^T B g \leq 0$ ist — in diesem Fall wählen wir $\tau^* = 1$ — und eine konvexe quadratische Funktion in τ wenn $g^T B g > 0$ ist. In diesem zweiten Fall wird das globale Minimum für

$$\tau^* = \frac{r\|g\|_2}{2}\left(r\frac{g^T B g}{\|g\|_2^2}\right)^{-1} = \frac{1}{r}\frac{\left\|\nabla f(x)\right\|_2^3}{g^T B g} \tag{9.15}$$

oder $\tau^* = 1$ angenommen – je nachdem welcher der beiden Werte eher kommt. Und genau diese drei Fälle müssen wir jetzt (natürlich) auch unterscheiden.

1. Ist $g^T B g \leq 0$, also $\tau^* = 1$, dann ist

$$\begin{aligned} q(y_c(r)) - q(0) &= -r\|g\|_2 + \frac{1}{2}\frac{r^2}{\|g\|_2^2}\underbrace{g^T B g}_{\leq 0} \leq -r\|g\|_2 \leq -\frac{1}{2}r\|g\|_2 \\ &\leq -\frac{1}{2}\left\|\nabla f(x)\right\|_2 \min\left(r, \frac{\left\|\nabla f(x)\right\|_2}{\|B\|_2}\right) \end{aligned}$$

2. Ist $g^T B g > 0$ und erfüllt das τ^* aus (9.15) die Bedingung $\tau^* \leq 1$, dann ist

$$\begin{aligned} q\left(y_c(r)\right) - q(0) &= -\left(r\,\tau^*\right)\|g\|_2 + \frac{1}{2}\frac{\left(r\,\tau^*\right)^2}{\|g\|_2^2} g^T B g \\ &= -\frac{\|g\|_2^3}{g^T B g}\|g\|_2 + \frac{1}{2}\frac{\|g\|_2^6}{\left(g^T B g\right)^2}\frac{g^T B g}{\|g\|^2} = -\frac{1}{2}\frac{\|g\|_2^4}{g^T B g} \leq -\frac{1}{2}\frac{\|g\|_2^4}{\|B\|_2\|g\|_2^2} = -\frac{1}{2}\frac{\|g\|_2^2}{\|B\|_2} \\ &\leq -\frac{1}{2}\left\|\nabla f(x)\right\| \min\left(r, \frac{\left\|\nabla f(x)\right\|_2}{\|B\|_2}\right) \end{aligned}$$

3. Ist schließlich $g^T B g > 0$, aber erfüllt das τ^* aus (9.15) die Bedingung $\tau^* > 1$, das heißt also, daß $g^T B g < \|g\|_2^3/r$, dann müssen wir $\tau^* = 1$ wählen und erhalten, daß

$$\begin{aligned} q\left(y_c(r)\right) - q(0) &= -r\|g\|_2 + \frac{1}{2}\frac{r^2}{\|g\|_2^2} g^T B g < -r\|g\|_2 + \frac{1}{2}\frac{r^2}{\|g\|_2^2}\frac{\|g\|_2^3}{r} \\ &= -r\|g\|_2 + \frac{1}{2}r\|g\|_2 = -\frac{1}{2}r\|g\|_2 \leq -\frac{1}{2}\left\|\nabla f(x)\right\| \min\left(r, \frac{\left\|\nabla f(x)\right\|_2}{\|B\|_2}\right). \end{aligned}$$

In jedem dieser Fälle folgt (9.14) durch Multiplikation mit -1. □

Beweis von Satz 9.6: Wir bemerken zuerst, daß wegen $q_k(0) = f\left(x^{(k)}\right)$

$$\begin{aligned} \left|\frac{q_k\left(y^{(k)}\right) - f\left(x^{(k)} + y^{(k)}\right)}{q_k(0) - q_k\left(y^{(k)}\right)}\right| &= \left|\frac{q_k\left(y^{(k)}\right) - q_k(0) + f\left(x^{(k)}\right) - f\left(x^{(k)} + y^{(k)}\right)}{q_k(0) - q_k\left(y^{(k)}\right)}\right| \\ &= \left|\frac{f\left(x^{(k)}\right) - f\left(x^{(k)} + y^{(k)}\right)}{q_k(0) - q_k\left(y^{(k)}\right)} - \frac{q_k(0) - q_k\left(y^{(k)}\right)}{q_k(0) - q_k\left(y^{(k)}\right)}\right| \\ &= \left|\rho_k - 1\right|. \end{aligned} \tag{9.16}$$

Nach der Taylorformel ist außerdem

$$f\left(x^{(k)} + y^{(k)}\right) = \underbrace{f\left(x^{(k)}\right) + \nabla^T f\left(x^{(k)}\right) y^{(k)}}_{= q_k\left(y^{(k)}\right) - \frac{1}{2} {y^{(k)}}^T B_k y^{(k)}} + \int_0^1 \left(\nabla f\left(x^{(k)} + t y^{(k)}\right) - \nabla f\left(x^{(k)}\right)\right)^T y^{(k)}\,dt,$$

weswegen sich

$$\begin{aligned} &\left|q_k\left(y^{(k)}\right) - f\left(x^{(k)} + y^{(k)}\right)\right| \\ &= \left|\frac{1}{2}{y^{(k)}}^T B_k y^{(k)} - \int_0^1 \left(\nabla f\left(x^{(k)} + t y^{(k)}\right) - \nabla f\left(x^{(k)}\right)\right)^T y^{(k)}\,dt\right| \\ &\leq \left|\frac{1}{2}{y^{(k)}}^T B_k y^{(k)}\right| + \left|\int_0^1 \left(\nabla f\left(x^{(k)} + t y^{(k)}\right) - \nabla f\left(x^{(k)}\right)\right)^T y^{(k)}\,dt\right| \\ &\leq \frac{\beta}{2}\left\|y^{(k)}\right\|_2^2 + \omega\left(\nabla f, \left\|y^{(k)}\right\|_2\right)\left\|y^{(k)}\right\|_2 \qquad (9.17) \\ &= \left\|y^{(k)}\right\|_2\left(\frac{\beta}{2}\left\|y^{(k)}\right\|_2 + \omega\left(\nabla f, \left\|y^{(k)}\right\|_2\right)\right) \leq r_k\left(\frac{\beta}{2}r_k + \omega\left(\nabla f, r_k\right)\right) \qquad (9.18) \end{aligned}$$

ergibt, wobei der **Stetigkeitsmodul** $\omega(F,\delta)$ für $F \in C\left(\mathbb{R}^n\right)^n$ als

$$\omega\left(F,\delta\right) = \sup_x \sup_{\|d\|<\delta} \left\|F\left(x+d\right) - F(x)\right\|_2$$

definiert ist und $\omega(F,\delta) \to 0$ für $\delta \to 0$ erfüllt.
Nach all den Vorbemerkungen nehmen wir nun an, daß (9.13) *nicht* erfüllt wäre, das heißt, es gibt ein $\varepsilon > 0$, so daß $\|\nabla f_k\| \geq \varepsilon$, $k \in \mathbb{N}_0$. Nach Lemma 9.7 ist dann auch der Nenner von (9.16) nach unten beschränkt, denn es ist

$$q_k(0) - q_k\left(y^{(k)}\right) \geq \frac{1}{2}\,\|\nabla f_k\|_2\,\min\left(r_k, \frac{\|\nabla f_k\|_2}{\|B\|}\right) \geq \frac{\varepsilon}{2}\min\left(r_k, \frac{\varepsilon}{\beta}\right). \tag{9.19}$$

Setzen wir nun (9.19) und (9.18) in (9.16) ein, dann erhalten wir, daß

$$|\rho_k - 1| \leq \frac{r_k\left(r_k\,\beta + 2\omega\left(\nabla f, r\right)\right)}{\varepsilon\,\min\left(r_k, \varepsilon/\beta\right)} \tag{9.20}$$

Ist nun $r_k < \epsilon/\beta$, dann wird (9.20) zu

$$|\rho_k - 1| \leq \frac{r_k\beta + 2\omega\left(\nabla f, r\right)}{\varepsilon}$$

und es gibt eine Schranke $\bar{r}$, so daß für jedes $r < \bar{r}$ die Ungleichung $|\rho_k - 1| < |1 - \rho_+|$ erfüllt ist, was dazu führen würde, daß $r_{k+1} > r_k$ ist[237]. Das heißt aber, daß eine Verkleinerung der Trust Region nur dann eintreten kann, wenn $r_k > \bar{r}$ ist, dann aber mit Sicherheit wieder vergrößert werden muß. Sei $0 < \gamma < 1$ dieser Verkleinerungsfaktor[238], dann erhalten wir, daß

$$r_k \geq \min\left(r_0, \gamma\,\bar{r}\right), \qquad k \in \mathbb{N}_0, \tag{9.21}$$

die Radien der Trust Regions sind also nach unten beschränkt. Insbesondere bedeutet (9.21), daß unendlich oft $\rho_k > \rho_-$ gelten muß, denn sonst würde ja für $k \to \infty$ die Folge $r_k \to 0$ konvergieren. Nehmen wir also an, daß, nach eventuellem Übergang zu einer Teilfolge, $\rho_k > \rho_-$, $k \in \mathbb{N}_0$, gilt, dann erhalten wir, daß

$$f\left(x^{(k)}\right) - f\left(x^{(k+1)}\right) \geq \rho_-\left(q_k(0) - q_k\left(y^{(k)}\right)\right) \geq \frac{\rho_-\,\varepsilon}{2}\,\min\left(r_k, \frac{\varepsilon}{\beta}\right)$$

und somit, weil f nach unten beschränkt ist, gibt es ein $C > 0$ so daß

$$\begin{aligned} C \;&>\; f\left(x^{(0)}\right) - f\left(x^{(k+1)}\right) = \sum_{j=0}^{k}\left(f\left(x^{(j)}\right) - f\left(x^{(j+1)}\right)\right) \geq \sum_{j=0}^{k}\frac{\rho_-\,\varepsilon}{2}\,\min\left(r_j, \frac{\varepsilon}{\beta}\right) \\ &=\; \frac{\rho_-\,\varepsilon}{2}\sum_{r_j<\varepsilon/\beta} r_j + \frac{\rho_-\,\varepsilon^2}{2\beta}\,\#\left\{j \;:\; j \leq k,\, r_j \geq \frac{\varepsilon}{\beta}\right\}, \end{aligned}$$

[237] Beispielsweise, indem man dann, wie in (9.2), $r_{k+1} = 2r_k$ wählt.
[238] Im Beispiel (9.2) war dies $\gamma = \frac{1}{4}$.

was für *alle* $k \in \mathbb{N}_0$ gelten muß. Mit $k \to \infty$ erhalten wir somit, daß

$$\# \left\{ j \ : \ r_j \geq \frac{\varepsilon}{\beta} \right\} < \infty$$

und somit

$$\sum_{j=0}^{\infty} r_j < \infty \qquad \Rightarrow \qquad \lim_{j \to \infty} r_j = 0$$

ist, was den langersehnten Widerspruch zu (9.21) liefert, weswegen (9.13) eben doch erfüllt sein muß. □

9 Literatur

Ablay, P. (1989). Optimieren mit Evolutionsstragien. In *Computer–Anwendungen*, Spektrum der Wissenschaft: Verständliche Forschung, pages 162–174. Spektrum–Verlag.

Anderson, E., Bai, Z., Bischof, C., Demmel, J., Dongarra, J., Du Croz, J., Greenbaum, A., Hammarling, S., McKenney, A., Ostrouchov, S., Sorensen, D. (1995). *LAPACK User's Guide*. SIAM, second edition.

Barnes, E. R. (1986). A variation of Karmarkar's algorithm for solving linear programming problems. *Math. Prog.*, **36**:174–182.

Berkelaar, M., Dirks, J., Eikland, K., Notebaert, P. (2000). `lp_solve`. `http://lpsolve.sourceforge.net/5.5`.

Boor, C. (2000). Computational aspects of multivariate polynomial interpolation: Indexing the coefficients. *Advances Comput. Math.*, **12**:289–301.

Brieskorn, E. (1985). *Lineare Algebra und Analytische Geometrie II*. Vieweg.

Broyden, C. G. (1965). A class of methods for solving nonlinear simulaneous equations. *Math. Comp.*, **19**:577–593.

Ciarlet, P. G., Raviart, P. A. (1972). General Lagrange and Hermite interpolation in $\mathbb{R}^n$ with applications to finite element methods. *Arch. Rational Mech. Anal.*, **46**:178–199.

Conn, A. R., Gould, N. I. M., Toint, P. L. (1992). *LANCELOT: a FORTRAN package for large–scale nonlinear optimization*, volume 17 of *Springer Series in Computational Mathematics*. Springer–Verlag.

Conn, A. R., Scheinfeld, K., Toint, P. L. (1997). On the convergence of derivative–free methods for unconstrained optimization. In Buhmann, M. D., Iserles, A., editors, *Approximation Theory and Optimization – Tributes to M. J. D. Powell*, pages 83–108. Cambridge University Press.

Cooley, J. W., Tukey, J. W. (1965). An algorithm for machine calculation of complex Fourier series. *Math. Comp.*, **19**:297–301.

Cox, D., Little, J., O'Shea, D. (1998). *Using Algebraic Geometry*, volume 185 of *Graduate Texts in Mathematics*. Springer Verlag.

Dantzig, G. B. (1963). *Linear programming and extensions*. Pinceton University Press.

Davidon, W. C. (1959). Variable metric method for minimization. Technical Report ANL-5990, Argonne National Laboratory, Argonne, Il.

Davidon, W. C. (1991). Variable metric method for minimization. *SIAM J. Optimization*, **1**:1–17.

Dikin, I. I. (1967). Iterative solution of problems of linear and quadratic programming. *Soviet Math. Doklady*, **8**:674–675.

Duden (1969). *Rechnen und Mathematik. Das Lexikon für Schule und Praxis*. Bibliographisches Institut Mannheim/Wien/Zürich, 3. edition.

Dueck, G., Scheuer, T., Wallmeier, H.-M. (1993). Toleranzschwelle und Sintflut: neue Ideen zur Optimierung. *Spektrum der Wissenschaft*, **1993/3**:42–51.

Fischer, G. (1984). *Lineare Algebra*. Vieweg.

Fletcher, R., Reeves, C. M. (1964). Function minimization by conjugate gradients. *Computer Journal*, **7**:149–154.

Gasca, M., Sauer, T. (2000). Polynomial interpolation in several variables. *Advances Comput. Math.*, **12**:377–410. to appear.

Gass, S. I. (1970). *An Illustrated Guide to Linear Programming*. McGraw–Hill. Republished by Dover, 1990.

Golub, G., van Loan, C. F. (1996). *Matrix Computations*. The Johns Hopkins University Press, 3rd edition.

Guingard, M. (1969). Generalized Kuhn–Tucker conditions for mathematical programmin in a Banach space. *SIAM J. Control*, **7**:232–241.

Hestenes, M. R., Stiefel, E. (1952). Methods of conjugate gradients for solving linear systems. *Journal of Research of the National Bureau of Standards*, **49**:409–436.

Heuser, H. (1983). *Lehrbuch der Analysis. Teil 2*. B. G. Teubner, 2. edition.

Heuser, H. (1984). *Lehrbuch der Analysis. Teil 1*. B. G. Teubner, 3. edition.

Higham, N. J. (1996). *Accuracy and stability of numerical algorithms*. SIAM.

Horn, R. A., Johnson, C. R. (1985). *Matrix Analysis*. Cambridge University Press.

Hoşten, S., Thomas, R. (1998). Gröbner bases and integer programming. In Buchberger, B., Winkler, F., editors, *Gröbner bases and applications*, pages 144–158. Cambridge University Press.

Isaacson, E., Keller, H. B. (1966). *Analysis of Numerical Methods*. John Wiley & Sons.

Karlin, S. (1959). *Mathemtical Methods and Theory in Games, Programming and Economics*. Dover Phoenix Editions. Addison–Wesley. Dover Reprint 2003.

Kolmogoroff, A. N. (1948). A remark on the polynomials fo Chebyshev, deviating at least from a given function. *Ushepi*, **3**:216–221. Probably in Russian.

Kunz, K. S. (1957). *Numerical Analysis*. McGraw-Hill Book Company.

Makhorin, A. (2000). `glpk`. `http://www.gnu.org/software/glpk/`.

Minty, G. J., Klee, V. (1972). How good is the simplex algorithm. In Shisha, O., editor, *Inequalities – III*. Academic Press.

Neumann, J. v. (1928). Zur Theorie der Gesellschaftsspiele. *Math. Annalen*, **100**:295–320.

Neumann, J. v., Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton University Press, sixth paperback printing, 1990 edition.

Nocedal, J., Wright, S. J. (1999). *Numerical Optimization*. Springer Series in Operations Research. Springer.

Powell, M. J. D. (1976). Some global convergence properties of a variable matric algorithm for minimization without exact line searches. *SIAM–AMS Proceedings 9: Nonlinear Programming*, pages 53–72.

Pschyrembel (1994). *Klinisches Wörterbuch*. Walter de Gruyter & Co, 257 edition.

Riese, A. (1574). *Rechenbuch / auff Linien und Ziphren / in allerley Handhierung / Geschäfften unnd Kauffmannschafft*. Franck. Bey. Chr. Egen. Erben. Facsimile: Verlag Th. Schäfer, Hannover, 1987.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton University Press.

Sauer, T. (1995). Computational aspects of multivariate polynomial interpolation. *Advances Comput. Math.*, **3**(3):219–238.

Sauer, T. (2000). Numerische Mathematik II. Vorlesungsskript, Friedrich–Alexander–Universität Erlangen–Nürnberg, Justus–Liebig–Universität Gießen. `http://www.math.uni-giessen.de/tomas.sauer`.

Sauer, T. (2002). Approximationstheorie. Vorlesungsskript, Justus–Liebig–Universität Gießen. `http://www.math.uni-giessen.de/tomas.sauer`.

Sauer, T. (2006). Polynomial interpolation in several variables: Lattices, differences, and ideals. In Buhmann, M., Hausmann, W., Jetter, K., Schaback, W., Stöckler, J., editors, *Multivariate Approximation and Interpolation*, pages 189–228. Elsevier.

Schwarz, H. R. (1988). *Numerische Mathematik*. B. G. Teubner, Stuttgart.

Schwarz, H. R. (1997). *Numerische Mathematik*. B. G. Teubner, 4th edition.

Sigler, L. (2002). *Fibonacci's Liber Abaci. Leoanardo Pisano's Book of Calculation*. Springer.

Spellucci, P. (1993). *Numerische Verfahren der nichtlinearen Optimierung*. Internationale Schriftenreihe zu Numerischen Mathematik. Birkhäuser.

Stoer, J. (1983). *Einführung in die Numerische Mathematik I.* Heidelberger Taschenbücher. Springer Verlag, 4 edition.

Williams, J. (1986). *The Complete Strategyst. Being a Primer on the Theory of Games on Strategy*. Dover Publications. Reprint. Originally Mc–Graw–Hill, 1966.

Wolfe, P. (1969). Convergence conditions for ascent methods. *SIAM Review*, **11**:220–228.

Ye, Y. (1997). *Interior Points Algorithms. Theory and Analysis*. John Wiley & Sons.