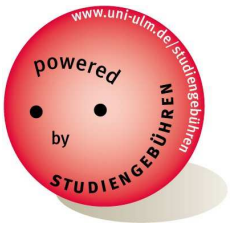
ulm university universität
uulm

Vorlesungsmanuskript zu

Lineare Algebra II

Werner Balser
Institut für Angewandte Analysis

Sommersemester 2008



Inhaltsverzeichnis

Kapitel 1

Mehr zu Normalformen von Matrizen

In diesem Kapitel sei V ein endlich-dimensionaler Vektorraum über $\mathbb{K}$, der nicht nur den Nullvektor enthält. Wir wollen untersuchen, wie die „einfachste Form“ der Darstellungsmatrix eines gegebenen Endomorphismus von V aussieht. Da eine Darstellungsmatrix immer bis auf Ähnlichkeit bestimmt ist, ist dies äquivalent zur Frage nach einer „einfachsten Form“ einer quadratischen Matrix A unter Ähnlichkeit, oder genauer, nach einem möglichst einfachen Repräsentanten innerhalb jeder Ähnlichkeitsklasse von Matrizen. Beachte, dass in vielen der folgenden Resultate vorausgesetzt wird, dass die Nullstellen des charakteristischen Polynoms eines Endomorphismus oder einer Matrix zu dem Skalarenkörper $\mathbb{K}$ gehören; diese Voraussetzung ist natürlich für den Fall $\mathbb{K} = \mathbb{C}$ immer erfüllt. Einige Resultate in diesem Kapitel wurden bereits im ersten Teil dieser Vorlesung behandelt und werden hier nur für „Quereinsteiger“ ohne Beweise wiederholt, auch um einige Sprechweisen und Bezeichnungen klarzustellen.

1.1 Trigonalisierung

Satz 1.1.1 *Sei $n \in \mathbb{N}$, sei V ein n-dimensionaler Vektorraum über $\mathbb{K}$, und sei $T \in L(V)$ so, dass alle Nullstellen des charakteristischen Polynoms von T in $\mathbb{K}$ liegen. Dann gibt es eine Basis $(v_1, \ldots, v_n)$ von V, bezüglich der die Darstellungsmatrix von T obere Dreiecksgestalt hat. Auf der Diagonalen dieser Darstellungsmatrix stehen dann die Eigenwerte von T entsprechend ihrer algebraischen Vielfachheit, und man kann sogar die Reihenfolge dieser Eigenwerte beliebig vorschreiben. Falls V ein Raum mit Skalarprodukt ist, kann zusätzlich $(v_1, \ldots, v_n)$ sogar als Orthonormalbasis gewählt werden.*

Definition 1.1.2 *Wir nennen zwei Matrizen $A, B \in \mathbb{K}^{n\times n}$* unitär ähnlich*, wenn es eine unitäre bzw. orthogonale Matrix $U \in \mathbb{K}^{n\times n}$ gibt, für welche $B = \overline{U}^T A U$ ist. Beachte, dass wegen $U^{-1} = \overline{U}^T$ die Matrizen B und A dann auch ähnlich zueinander sind. Wegen eines Lemmas aus LA I sind die Darstellungsmatrizen eines Endomorphismus bezüglicher zweier Orthonormalbasen immer unitär ähnlich.*

Korollar zu Satz 1.1.1 (Schursches Lemma) *Jede Matrix $A \in \mathbb{K}^{n\times n}$, für welche alle Nullstellen des charakteristischen Polynoms in $\mathbb{K}$ liegen, ist unitär ähnlich zu einer oberen Dreiecksmatrix.*

1.2 Der Satz von Cayley-Hamilton und das Minimalpolynom

Definition 1.2.1 *Sei V ein beliebiger Vektorraum über $\mathbb{K}$, und sei $T \in L(V)$. Für eine natürliche Zahl $k \in \mathbb{N}$ definieren wir T^k als die k-malige Hintereinanderausführung der Abbildung T. Setzt man noch*

T^0 gleich der identischen Abbildung, so können wir dann für jedes Polynom $p(t) = \sum_{k=0}^{m} p_k\, t^k$ mit Koeffizienten $p_k \in \mathbb{K}$ die Abbildung

$$p(T) \;=\; \sum_{k=0}^{m} p_k\, T^k$$

bilden und stellen fest, dass $p(T) \in L(V)$ ist. Wir erhalten somit für jedes feste $T \in L(V)$ eine Abbildung $p \longmapsto p(T)$ von $\mathbb{K}[t]$ in $L(V)$, und es ist nicht schwer zu sehen, dass diese Abbildung linear ist. Sie ist zusätzlich auch multiplikativ.

Satz 1.2.2 (Cayley-Hamilton) *Sei $n \in \mathbb{N}$, sei V ein n-dimensionaler Vektorraum über $\mathbb{K}$, und sei $T \in L(V)$. Wenn p das charakteristische Polynom von T ist, dann folgt $p(T) = 0$, wobei 0 die Nullabbildung in $L(V)$ bezeichnet.*

Definition 1.2.3 (Minimalpolynom) *Nach dem Satz von Cayley-Hamilton gibt es zu jedem Endomorphismus T eines n-dimensionalen Vektorraums V, mit $n \in \mathbb{N}$, ein nicht-triviales Polynom p, für welches $p(T) = 0$ ist. In der Menge aller dieser Polynome gibt es immer eines mit minimalem Grad und höchstem Koeffizienten 1; wie das nächste Lemma zeigt, ist dieses Polynom durch T eindeutig festgelegt, und es heißt* Minimalpolynom zu T. *Als abkürzende Sprechweise wollen wir vereinbaren, dass ein beliebiges Polynom mit dem höchsten Koeffizienten* 1 normalisiert *heißen soll. Das Minimalpolynom ist also immer normalisiert, das charakteristische Polynom dagegen nicht.*

Lemma 1.2.4 *Sei $n \in \mathbb{N}$, sei V ein n-dimensionaler Vektorraum über $\mathbb{K}$, und sei $T \in L(V)$. Dann gibt es ein Polynom $p \in \mathbb{K}[t] \setminus \{0\}$, welches folgende beiden Eigenschaften hat:*

(a) $p(T) = 0$.

(b) *Ist $p_1 \in \mathbb{K}[t]$ so, dass $p_1(T) = 0$ ist, so gibt es ein $q \in \mathbb{K}[t]$ mit $p_1 = q\,p$.*

Das Polynom p ist durch diese beiden Eigenschaften bis auf einen konstanten Faktor eindeutig festgelegt und hat unter allen $p_1 \in \mathbb{K}[t] \setminus \{0\}$ mit der Eigenschaft $p_1(T) = 0$ den kleinsten Grad.

1.3 Hauptachsentransformation

Im folgenden Satz muss nicht vorausgesetzt werden, dass die Eigenwerte von T zu $\mathbb{K}$ gehören, da die Eigenwerte einer selbstadjungierten Abbildung immer reell sind.

Satz 1.3.1 (Satz von der Hauptachsentransformation) *Sei $n \in \mathbb{N}$, sei V ein n-dimensionaler Vektorraum über $\mathbb{K}$ mit Skalarprodukt, und sei $T \in L(V)$ selbstadjungiert. Dann gibt es eine Orthonormalbasis in V, bezüglich der die Darstellungsmatrix von T diagonal ist. Anders ausgedrückt heißt das: Es gibt eine Orthonormalbasis von V, die aus Eigenvektoren von T besteht.*

Definition 1.3.2 *Wir nennen eine Matrix $A \in \mathbb{K}^{n\times n}$* unitär diagonalisierbar, *wenn es eine unitäre bzw. orthogonale Matrix $U \in \mathbb{K}^{n\times n}$ gibt, für welche $B = \overline{U}^T A\,U$ diagonal ist.*

Korollar zu Satz 1.3.1 *Jede hermitesche Matrix A ist unitär diagonalisierbar. Genauer: Ist $n \in \mathbb{N}$, und ist $A \in \mathbb{K}^{n\times n}$ hermitesch, so gibt es eine unitäre Matrix $U \in \mathbb{K}^{n\times n}$ und eine Diagonalmatrix Λ, für welche*

$$A\,U \;=\; U\,\Lambda$$

ist. Die Diagonalelemente von Λ sind gerade die Eigenwerte von A, und die Spalten von U sind die Eigenvektoren von A, in der zu den Diagonalelementen von Λ passenden Reihenfolge.

Beachte, dass im Falle einer reellen symmetrischen Matrix A auch die Transformationsmatrix U im obigen Korollar reell (also orthogonal) gewählt werden kann.

Der Vollständigkeit halber beweisen wir noch folgenden, in der theoretischen Physik sehr wichtigen Satz:

Satz 1.3.3 (Simultane Hauptachsentransformation) *Seien $n, m \in \mathbb{N}$, sei V ein n-dimensionaler Vektorraum über $\mathbb{K}$ mit Skalarprodukt, und seien die Endomorphismen $T_1, \ldots, T_m \in L(V)$ selbstadjungiert und paarweise miteinander vertauschbar; d. h., es gelte*

$$\forall\ 1 \le j < k \le m: \qquad T_j \circ T_k \ = \ T_k \circ T_j \,.$$

Dann gibt es eine Orthonormalbasis in V, bezüglich der die Darstellungsmatrizen aller T_j diagonal sind. Anders ausgedrückt heißt das: Es gibt eine Orthonormalbasis von V, die aus Vektoren besteht, welche Eigenvektoren von allen T_j sind.

Beweis: Induktion über m: Für $m = 1$ ist dieser Satz gleich dem über die Hauptachsentransformation. Sei jetzt $m \ge 2$, und sei der Satz für $m-1$ schon bewiesen. Seien die (reellen) Zahlen $\lambda_1, \ldots, \lambda_s$ die verschiedenen Eigenwerte und $U_1, \ldots, U_s$ die zugehörigen Eigenräume von T_m. Aus dem Satz über die Hauptachsentransformation, angewandt auf T_m, folgt für jedes j, dass die Dimension von U_j, also die geometrische Vielfachheit, gleich der algebraischen Vielfachheit von λ_j ist. Die Räume U_j sind nach Teil I der Vorlesung paarweise zueinander orthogonal, und daher gilt $V = U_1 \oplus \ldots \oplus U_s$. Für $u \in U_j$ folgt aus $T_m \circ T_\nu \, u = T_\nu \circ T_m \, u = \lambda_j \, T_\nu \, u$, dass $T_\nu \, u \in U_j$ ist, für jedes $\nu = 1, \ldots, m-1$ und jedes $j = 1, \ldots, s$. Also bildet jedes T_ν jeden Raum U_j in sich ab. Wendet man die Induktionshypothese auf die Restriktionen der $T_1, \ldots, T_{m-1}$ auf den Raum U_j an, so folgt die Existenz einer Orthonormalbasis von U_j, deren Vektoren Eigenvektoren zu allen T_j sind, und zwar sogar für $j = 1, \ldots, m$. Diese Orthonormalbasen können wir wegen der Orthogonalität der Räume U_j zu einer Orthogonalbasis von V zusammenfassen, und das ist die Behauptung. □

Korollar zu Satz 1.3.3 *Seien $A_1, \ldots, A_m \in \mathbb{K}^{n \times n}$ hermitesche Matrizen, welche miteinander kommutieren, d. h., für welche*

$$\forall\ j, k = 1, \ldots, m: \qquad A_j \, A_k \ = \ A_k \, A_j \,.$$

Dann gibt es eine unitäre Matrix $U \in \mathbb{K}^{n \times n}$ so, dass die Matrizen $\overline{U}^T \, A_k \, U$ alle diagonal sind.

Beweis: Aus der Voraussetzung folgt, dass die Endomorphismen $x \longmapsto A_j \, x$ die Voraussetzung von Satz 1.3.3 erfüllen, und daher gibt es eine Orthonormalbasis von $\mathbb{K}^n$, bezüglich derer die Darstellungsmatrizen dieser Endomorphismen alle diagonal sind. Daraus folgt die Behauptung. □

1.4 Diagonalisierung normaler Endomorphismen

Auch das folgende Ergebnis wurde im ersten Teil dieser Vorlesung gezeigt:

Satz 1.4.1 *Sei $n \in \mathbb{N}$, sei V ein n-dimensionaler Vektorraum über $\mathbb{K}$ mit Skalarprodukt, und sei $T \in L(V)$. Dann sind folgende Aussagen äquivalent:*

(a) *T ist normal, und alle Nullstellen des charakteristischen Polynoms liegen in $\mathbb{K}$.*

(b) *Es gibt eine Orthonormalbasis von V, die aus Eigenvektoren von T besteht.*

Korollar zu Satz 1.4.1 *Eine Matrix $A \in \mathbb{C}^{n \times n}$ ist genau dann normal, wenn sie unitär ähnlich zu einer Diagonalmatrix ist.*

Aus diesem Korollar folgt übrigens, dass bei einer normalen Matrix Eigenvektoren zu verschiedenen Eigenwerten immer orthogonal sind! Dies spielt im Folgenden noch eine Rolle.

1.5 Reelle Normalformen diagonalisierbarer Matrizen

Eine reelle Matrix hat im Allgemeinen auch nicht-reelle Eigenwerte, und deshalb auch nicht-reelle Eigenvektoren. Deshalb sind die bisher betrachteten Normalformen in der Regel Matrizen mit nicht-reellen Einträgen, selbst wenn die Ausgangsmatrix reell war. Es ist deshalb von Interesse, auch eine Normalform einer reellen Matrix unter Ähnlichkeitstransformationen mit reellen invertierbaren Matrizen zu finden. Das soll hier geschehen, falls die Matrix diagonalisierbar ist. Dazu benötigen wir einige Bezeichnungen:

Definition 1.5.1 *Für quadratische Matrizen $A_1, \ldots, A_\nu$ von i. a. unterschiedlicher Größe heißt die Matrix*

$$A \;=\; \left[\begin{array}{c|c|c|c} A_1 & 0 & \ldots & 0 \\ \hline 0 & A_2 & \ldots & 0 \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline 0 & 0 & \ldots & A_\nu \end{array}\right] ,$$

wobei das Symbol 0 jeweils für eine Nullmatrix passender Größe steht, die direkte Summe *der Matrizen $A_1, \ldots, A_\nu$. Eine reelle Zahl λ kann mit der eindimensionalen reellen Matrix $[\lambda]$ identifiziert werden, während eine komplexe Zahl $\lambda = \alpha + i\,\beta$, $\alpha, \beta \in \mathbb{R}$, $\beta \neq 0$, der zweidimensionalen Matrix*

$$\Lambda \;=\; \left[\begin{array}{cc} \alpha & \beta \\ -\beta & \alpha \end{array}\right]$$

zugeordnet sei. Wenn eine Matrix $A \in \mathbb{R}^{n\times n}$ gegeben ist, dann seien ihre reellen Eigenwerte mit $\lambda_1, \ldots, \lambda_s$ bezeichnet, wobei $0 \leq s \leq n$ ist, denn die Menge der reellen, aber auch die der nicht-reellen, Eigenwerte kann leer sein. Da zu jedem nicht-reellen Eigenwert auch die konjugiert-komplexe Zahl ein Eigenwert ist, fassen wir diese Eigenwerte, falls es solche gibt, zu Paaren $(\lambda_j = \alpha_j + i\,\beta_j, \overline{\lambda}_j = \alpha_j - i\,\beta_j)$, $\alpha_j, \beta_j \in \mathbb{R}$, $(j = s+1, \ldots, \mu)$ zusammen und vereinbaren, dass alle $\beta_j > 0$ sein sollen. Wenn wir wie üblich die mehrfachen Eigenwerte so oft aufschreiben, wie es ihrer Vielfachheit entspricht, so muss dann $s + 2\,\mu = n$ gelten. Die direkte Summe aus den eindimensionalen Matrizen $[\lambda_j]$, $1 \leq j \leq s$ und den zweidimensionalen Λ_j, $s < j \leq \mu$, ist also eine reelle Matrix vom Typ $n \times n$, welche in Folgenden reelle Eigenwertmatrix *genannt werden soll. Sie ist bis auf die Numerierung der Eigenwerte eindeutig bestimmt!*

Aufgabe 1.5.2 *Sei $A \in \mathbb{R}^{n\times n}$, und sei λ ein nicht-reeller Eigenwert von A, sowie $(b_1, , \ldots, b_\nu)$ eine Basis des zugehörigen Eigenraums. Zeige, dass dann auch $\overline{\lambda}$ ein Eigenwert von A ist, und dass $(\overline{b}_1, , \ldots, \overline{b}_\nu)$ Basis des Eigenraums zum Eigenwert $\overline{\lambda}$ ist. Zeige weiter: Ist $\lambda = \alpha + i\,\beta$, $\alpha, \beta \in \mathbb{R}$, $\beta > 0$, Eigenwert von A mit zugehörigem Eigenvektor $b = c + i\,d$, $c, d \in \mathbb{R}^n$, so sind c und d linear unabhängig, und es gilt*

$$A \left[\begin{array}{cc} c\,, & d \end{array}\right] \;=\; \left[\begin{array}{cc} c\,, & d \end{array}\right] \left[\begin{array}{cc} \alpha & \beta \\ -\beta & \alpha \end{array}\right] .$$

Zeige schließlich noch: Wenn $c + i\,d$ zu $c - id$ orthogonal ist, was bei einer normalen Matrix A immer der Fall ist, dann sind auch c und d orthogonal, und $\|c\| = \|d\|$.

Satz 1.5.3 *Sei $A \in \mathbb{R}^{n\times n}$ diagonalisierbar. Dann gibt es eine invertierbare Matrix $T \in \mathbb{R}^{n\times n}$ derart, dass $\Lambda = T^{-1}\,A\,T$ reelle Eigenwertmatrix zu A ist. Wenn A normal ist, dann kann T sogar als orthogonale Matrix gewählt werden.*

Beweis: Nach Definition der Diagonalisierbarkeit gibt es eine Basis von $\mathbb{C}^n$, die nur aus Eigenvektoren von A besteht, und bei einer normalen Matrix kann dies sogar eine Orthonormalbasis sein. Wir numerieren die Vektoren der Basis so, dass die ersten b_j, $j = 1, \ldots, s$ zu den reellen Eigenwerten gehören, und somit selbst reell gewählt werden können. Die übrigen dagegen sollen in Paaren $(b_j, \overline{b}_j)$ angeordnet sein, entsprechend einem Paar $(\lambda, \overline{\lambda})$ von nicht-reellen Eigenwerten. Nach Aufgabe 1.5.2 kann man dann (unter Benutzung des Austauschlemmas) zeigen dass $(b_1, \ldots, b_s, \mathrm{Re}\; b_{s+1}, \mathrm{Im}\; b_{s+1}, \ldots, \mathrm{Re}\; b_\mu, \mathrm{Im}\; b_\mu)$ eine Basis von $\mathbb{R}^n$ ist, und sogar eine Orthonormalbasis falls A normal ist. Daraus folgt dann die Behauptung. □

1.6 Orthogonale Matrizen

Wir machen folgende einfache Beobachtung für unitäre bzw. orthogonale Matrizen, die für die Algebra wichtig ist:

Satz 1.6.1 *Die Menge der unitären, der orthogonalen, aber auch die Teilmenge der orthogonalen Matrizen, deren Determinante gleich* 1 *ist, sind Gruppen bezüglich der Matrixmultiplikation.*

Beweis: Da für die Matrixmultiplikation immer ein Assoziativgesetz gilt, ist nur zu zeigen, dass die Einheitsmatrix zu jeder der Mengen gehört, und dass mit einer Matrix auch ihre Inverse in der Menge ist. Dies ist aber klar wegen der Definition der entsprechenden Matrizenmenge sowie dem Determinantenmultiplikationssatz. □

Sei A eine unitäre Matrix. Dann ist bekanntlich $A^{-1} = \overline{A}^T$, und deshalb ist A insbesondere normal, also unitär diagonalisierbar. Außerdem sind alle Eigenwerte von A komplexe Zahlen vom Betrag 1, und daraus folgt $|\det A| = 1$. Für eine orthogonale (also reelle) Matrix A folgt deshalb, dass ihre reelle Eigenwertmatrix außer 1 und -1 auf der Diagonalen aus lauter zweidimensionalen Matrizen der Form

$$D(\phi) \;=\; \begin{bmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{bmatrix}, \qquad \phi \in (0,\pi)$$

aufgebaut ist, die als Abbildung von $\mathbb{R}^2$ in sich jeweils einer Drehung um den Winkel ϕ entsprechen. Beachte dabei, dass der Drehwinkel echt zwischen 0 und π liegt - das liegt daran, dass wir eine Drehung um einen größeren Winkel erhalten, wenn wir zunächst eine Spiegelung, dann eine Drehung und danach nochmals dieselbe Spiegelung ausführen. Die Matrix T im letzten Satz bedeutet geometrisch, dass wir in $\mathbb{R}^n$ statt der üblichen kanonischen Basis eine andere Orthonormalbasis einführen, die gerade den Spalten von T entspricht, und somit setzt sich die durch die Matrix A definierte Abbildung in $\mathbb{R}^n$, in dieser neuen Basis betrachtet, aus einer Anzahl von Spiegelungen und Drehungen zusammen.

Aufgabe 1.6.2 *Zeige dass eine dreidimensionale orthogonale Matrix A mit* $\det A = -1$ *in der Form $A = S_1\, B$ geschrieben werden kann, wobei S_1 eine Spiegelung an der x_3-x_2-Ebene beschreibt, während B eine dreidimensionale orthogonalen Matrix mit* $\det B = 1$ *ist.*

Bemerkung 1.6.3 (Normalform von dreidimensionalen orthogonalen Matrizen)
Eine Matrix $A \in \mathbb{R}^{3\times 3}$ hat mindestens einen reellen Eigenwert, und falls A orthogonal ist, erhält man für die von ihr definierte Abbildung folgende verschiedene Fälle:

(a) *Falls A drei reelle Eigenwerte hat, so hat A entweder den dreifachen Eigenwert* 1*, woraus $A = I$ folgt, oder den dreifachen Eigenwert -1, was $A = -I$ impliziert, oder A hat die Eigenwerte* 1 *und -1, wovon einer die Vielfachheit* 2 *hat. In den beiden letzten Fällen entspricht A einer Drehung um den Winkel π, falls -1 Vielfachheit* 2 *hat, oder einer Spiegelung.*

(b) *Falls A nur einen reellen und zwei konjugiert-komplexe nicht-reelle Eigenwerte hat, sieht man aus der reellen Eigenwertmatrix, dass A eine Drehung um eine geeignete Achse ist, evtl. gefolgt von einer Spiegelung an der zur Drehachse orthogonalen Koordinatenebene, wobei letzteres für* $\det A = -1$ *eintritt.*

Wir sehen also, dass A in jedem der Fälle eine Drehung um eine geeignete Achse darstellt, falls $\det A = 1$ *ist. Für den anderen Fall ist die von A dargestellte Matrix eine* Drehspiegelung*, d. h., eine Drehung gefolgt von einer Spiegelung.*

Die oben gemachten Aussagen beschreiben die von einer orthogonalen dreidimensionalen Matrix A definierte Abbildung in einem entsprechenden Koordinatensystem. Wir wollen jetzt noch die Frage klären, ob man die Abbildung auch im ursprünglichen Koordinatensystem durch Drehungen um die Koordinatenachsen beschreiben kann. Wir beschränken uns dabei auf den Fall $\det A = 1$ und betrachten folgende beiden Typen von orthogonalen Matrizen:

$$D_1(\phi) \;=\; \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\phi & -\sin\phi \\ 0 & \sin\phi & \cos\phi \end{bmatrix}, \qquad D_3(\phi) \;=\; \begin{bmatrix} \cos\phi & -\sin\phi & 0 \\ \sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Man sieht, dass die erste Matrix eine Drehung um die x_1-Achse, die zweite eine um die x_3-Achse darstellt. Es folgt (auch durch einfaches Nachrechnen), dass $D_j(-\phi)$ die inverse Matrix zu $D_j(\phi)$ ist.

Satz 1.6.4 (Eulersche Winkel) *Zu jeder dreidimensionalen orthogonalen Matrix A mit $\det A = 1$ gibt es $\alpha, \beta, \gamma \in [0, 2\pi)$ mit*

$$A \;=\; D_3(\alpha)\, D_1(\beta)\, D_3(\gamma)\ .$$

Beweis: Die Spalten von A sind eine Orthonormalbasis von $\mathbb{R}^3$, und wenn wir A von links her mit einer Drehmatrix multiplizieren, so haben die Spalten der neuen Matrix die gleiche Eigenschaft. Daher können wir ein α so wählen, dass die dritte Spalte von $D_3(-\alpha)\, A$ in der x_3-x_2-Ebene liegt (dass also ihre erste Komponente verschwindet). Anschließend wählen wir β so, dass die dritte Spalte von $D_1(-\beta)\, D_3(-\alpha)\, A$ gleich dem dritten Einheitsvektor wird - die beiden anderen Spalten müssen dann auf e_3 senkrecht stehen, also in der x_1-x_2-Ebene liegen. Danach können wir γ so wählen, dass die zweite Spalte von $D_3(-\gamma)\, D_1(-\beta)\, D_3(-\alpha)\, A$ gleich e_2 wird, wobei die dritte Spalte nach wie vor e_3 ist. Die erste Spalte ist dann ein Einheitsvektor, der auf e_2 und e_3 senkrecht steht, und da nach dem Determinantenmultiplikationssatz klar ist dass die Determinante von $D_3(-\gamma)\, D_1(-\beta)\, D_3(-\alpha)\, A$ gleich der von A, also gleich 1 ist, muss die erste Spalte der erhaltenen Matrix gleich e_1 sein, woraus die Behauptung folgt. □

Die im letzten Satz vorkommenden Zahlen α, β, γ heißen auch *die Eulerschen Winkel* zur Matrix A. Interessanterweise kann also eine allgemeine orthogonale dreidimensionale Matrix mit positiver Determinante als Hintereinanderausführung dreier Drehungen *um nur zwei verschiedene Achsen* aufgefasst werden!

1.7 Jordanmatrizen

Definition 1.7.1 *Eine quadratische Matrix N, deren Elemente unmittelbar oberhalb der Diagonalen alle gleich 1 und alle übrigen gleich 0 sind, heißt* ein nilpotenter Jordanblock. *Für beliebiges $\lambda \in \mathbb{C}$ heißt die Matrix $\lambda I + N$, mit N wie oben, ein* Jordanblock. *Eine Matrix J heißt* Jordanmatrix, *wenn sie als direkte Summe von Jordanblöcken geschrieben werden kann.*

Aufgabe 1.7.2 *Sei N ein nilpotenter Jordanblock. Finde alle Potenzen N^k, d. h., alle Produkte der Form $N \cdot \ldots \cdot N$ mit k Faktoren, für $k \in \mathbb{N}$.*

Aufgabe 1.7.3 *Sei J eine Jordanmatrix mit paarweise verschiedenen Eigenwerten $\lambda_1, \ldots, \lambda_m$. Zu jedem der Eigenwerte λ_j sei der größte Jordanblock in J mit dem Eigenwert λ_j vom Typ $s_j \times s_j$. Zeige: Dann ist das Minimalpolynom von J gleich $(t-\lambda_1)^{s_1} \cdot \ldots \cdot (t-\lambda_m)^{s_m}$.*

Aufgabe 1.7.4 *Zeige mit* Aufgabe 1.7.3, *dass jedes normalisierte Polynom das Minimalpolynom einer geeignet gewählten Matrix A ist.*

Aufgabe 1.7.5 (Kleinstes gemeinsames Vielfaches von Polynomen) *Seien $p_1(t), p_2(t)$ normalisierte Polynome. Wir sagen, dass ein normalisiertes Polynom $p(t)$ das* kleinste gemeinsame Vielfache, *oder kurz* kgV von $p_1(t)$ und $p_2(t)$ *ist, wenn $p(t)$ durch $p_1(t)$ und $p_2(t)$ teilbar ist, und wenn jedes andere Polynom mit dieser Eigenschaft durch $p(t)$ geteilt werden kann. Zeige: Genau dann ist $p(t)$ kgV von $p_1(t)$ und $p_2(t)$, wenn jede Nullstelle von $p_1(t)$ oder $p_2(t)$ auch Nullstelle von $p(t)$ ist und umgekehrt, und wenn die Vielfachheit einer Nullstelle λ von $p(t)$ gleich dem Maximum der Vielfachheiten von λ als Nullstellen von $p_1(t)$ und $p_2(t)$ ist, wobei diese gleich 0 gesetzt seien, falls λ keine Nullstelle von $p_1(t)$ oder $p_2(t)$ ist. Folgere, dass es insbesondere immer genau ein kgV gibt.*

Aufgabe 1.7.6 (Minimalpolynom einer direkten Summe) *Schließe aus der Definition des Minimalpolynoms: Seien $A \in \mathbb{C}^{n\times n}$ und $B \in \mathbb{C}^{m\times m}$, und sei C gleich der direkten Summe von A und B. Dann ist das Minimalpolynom von C gleich dem kgV der Minimalpolynome von A und B.*

1.8 Invariante Unterräume

Definition 1.8.1 *Sei V ein Vektorraum über $\mathbb{K}$, sei U ein Unterraum von V, und sei $T \in L(V)$. Wenn $T(U) \subset U$ ist, d. h., wenn $T(u) \in U$ ist für alle $u \in U$, dann nennen wir U einen* invarianten Unterraum *für T.*

Aufgabe 1.8.2 *Zeige, dass Eigenräume zu beliebigen Eigenwerten eines Endomorphismus T immer invariante Unterräume für T sind.*

Satz 1.8.3 *Sei $n \in \mathbb{N}$, sei V ein n-dimensionaler Vektorraum über $\mathbb{K}$, und sei $T \in L(V)$. Genau dann existiert für ein $m \in \{1, \ldots, n-1\}$ ein m-dimensionaler invarianter Unterraum für T, wenn eine Basis von V existiert, bezüglich der die Darstellungsmatrix A von T die Form*

$$A = \left[\begin{array}{ccc|ccc} a_{11} & \ldots & a_{1m} & a_{1,m+1} & \ldots & a_{1n} \\ \vdots & & \vdots & \vdots & & \vdots \\ a_{m1} & \ldots & a_{mm} & a_{m,m+1} & \ldots & a_{mn} \\ \hline 0 & \ldots & 0 & a_{m+1,m+1} & \ldots & a_{m+1,n} \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \ldots & 0 & a_{n,m+1} & \ldots & a_{nn} \end{array}\right]$$

hat.

Beweis: Falls A bezüglich einer Basis $(v_1, \ldots, v_n)$ von V die angegebene Form hat, dann folgt dass die Bilder der ersten m Basisvektoren auf Linearkombinationen derselben abgebildet werden. Daraus wiederum folgt, dass $\mathcal{L}(v_1, \ldots, v_m)$ ein invarianter Unterraum für T der Dimension m ist. Umgekehrt, wenn U ein invarianter Unterraum für T dieser Dimension ist, kann eine Basis $(v_1, \ldots, v_m)$ von U zu einer Basis von V ergänzt werden, und bezüglich dieser Basis hat A die gewünschte Form. □

Aufgabe 1.8.4 *Gib ein Beispiel eines Endomorphismus in $\mathbb{R}^2$, der keinen nicht-trivialen invarianten Unterraum hat. Warum ist dies anders bei Endomorphismen von Räumen über $\mathbb{C}$, aber auch bei Räumen über $\mathbb{R}$, wenn die Dimension ungerade ist?*

Aufgabe 1.8.5 *Sei $n \in \mathbb{N}$, sei V ein n-dimensionaler Vektorraum über $\mathbb{K}$, und sei $T \in L(V)$. Zeige: Wenn es für T invariante Unterräume $U_1, \ldots, U_m$ von V gibt, für die $V = U_1 \oplus \ldots \oplus U_m$ ist, dann existiert eine Basis von V, bezüglich der die Darstellungsmatrix die direkte Summe von Matrizen $A_1, \ldots, A_m$ ist; dabei ist die Größe der Matrix A_k gerade gleich der Dimension von U_k. Untersuche, ob auch die Umkehrung gilt.*

1.9 Ketten von Hauptvektoren

Definition 1.9.1 *Sei V ein Vektorraum über $\mathbb{K}$. Wir nennen $v_1, \ldots, v_\mu \in V$ eine* Kette von Hauptvektoren *eines Endomorphismus $T \in L(V)$ zu einem $\lambda \in \mathbb{K}$, wenn für alle $j = 1, \ldots, \mu$ gilt*

$$v_j \neq 0\,, \qquad (T - \lambda\, id)\, v_j = v_{j-1}\,,$$

wobei wir $v_0 = 0$ setzen. Beachte, dass dann v_1 immer ein Eigenvektor von T, also λ ein Eigenwert von T sein muss. Die Zahl μ heißt auch Länge der Kette.

Bemerkung 1.9.2 *Wenn $v_1, \ldots, v_\mu \in V$ eine Kette von Hauptvektoren eines Endomorphismus $T \in L(V)$ zu einem $\lambda \in \mathbb{K}$ bilden, so tun dies auch die Vektoren $v_1, \ldots, v_j$ für $1 \leq j < \mu$. Wir können also Ketten verkürzen.*

Aufgabe 1.9.3 *Sei V ein Vektorraum über $\mathbb{K}$, und seien $v_1, \ldots, v_\mu \in V$ eine Kette von Hauptvektoren eines Endomorphismus $T \in L(V)$ zu einem Eigenwert $\lambda \in \mathbb{K}$. Zeige, dass $U = \mathcal{L}(v_1, \ldots, v_\mu)$ ein invarianter Unterraum von V ist, und dass das System $(v_1, \ldots, v_\mu)$ linear unabhängig, also sogar eine Basis von U ist. Finde die Darstellungsmatrix der Restriktion von T auf U.*

Der wesentliche Inhalt des noch zu beweisenden Satzes von der Jordanschen Normalform einer Matrix kann auch so ausgedrückt werden, dass es zu einem Endomorphismus $T \in L(V)$, für den die Nullstellen des charakteristischen Polynoms alle in $\mathbb{K}$ liegen, immer eine Basis von V gibt, welche aus Ketten von Hauptvektoren besteht. Wir werden dies im nächsten Abschnitt zeigen und dabei folgendes Resultat zur linearen Unabhängigkeit solcher Ketten benutzen:

Lemma 1.9.4 *Sei V ein Vektorraum über $\mathbb{K}$, und sei $T \in L(V)$. Für natürliche Zahlen $\mu_1 \geq \ldots \geq \mu_\nu \geq 1$ und ein $\lambda \in \mathbb{K}$ seien $v_1^{(j)}, \ldots, v_{\mu_j}^{(j)}$, für $j = 1, \ldots, \nu$, lauter Ketten von Hauptvektoren zum Eigenwert λ. Wenn das System der Eigenvektoren $(v_1^{(1)}, \ldots, v_1^{(\nu)})$ linear unabhängig ist, dann ist auch das gesamte System $(v_1^{(1)}, \ldots, v_{\mu_1}^{(1)}, \ldots, v_1^{(\nu)}, \ldots, v_{\mu_\nu}^{(\nu)})$ aller Hauptvektoren in den Ketten linear unabhängig.*

Beweis: O. B. d. A. sei $\lambda = 0$ angenommen; sonst kann man T durch $T - \lambda\, id$ ersetzen. Wir führen den Beweis mit Induktion über μ_1, also über die maximale Länge der Ketten: Für $\mu_1 = 1$ ist nichts zu zeigen, da dann alle $\mu_j = 1$ sind und das Gesamtsystem nur aus Eigenvektoren besteht. Sei jetzt $\mu_1 \geq 2$ vorausgesetzt, und seien $\alpha_{jk} \in \mathbb{K}$ so, dass

$$0 = \sum_{j=1}^{\nu} \sum_{k=1}^{\mu_j} \alpha_{jk} v_k^{(j)}\,. \tag{1.9.1}$$

Für $\mu \in \mathbb{N}$ ist $T^\mu\, v_k^{(j)} = v_{k-\mu}^{(j)}$, wobei Vektoren mit unteren Indizes $k - \mu \leq 0$ als der Nullvektor zu lesen sind. Also folgt aus (1.9.1)

$$0 = T^\mu\, 0 = \sum_{j=1}^{\nu} \sum_{k=\mu+1}^{\mu_j} \alpha_{jk} v_{k-\mu}^{(j)}\,.$$

Für $\mu = \mu_1 - 1$ ist die innere Summe leer, falls $\mu_j < \mu_1$ ist. Für alle anderen j besteht aber die innere Summe nur aus dem einen Term mit $k = \mu_1$, und dann ist $k - \mu = 1$. Also reduziert sich die rechte Seite auf eine Linearkombination von Eigenvektoren, die nach Voraussetzung linear unabhängig sind, und deshalb müssen die entsprechenden $\alpha_{j\mu_1} = 0$ sein. Damit treten in (1.9.1) nur noch Hauptvektoren aus Ketten mit Längen $\leq \mu_1 - 1$ auf, und diese sind per Induktionshypothese linear unabhängig. □

Vereinfacht ausgedrückt besagt obiges Lemma, dass Ketten von Hauptvektoren zu einem festen Eigenwert linear unabhängig sind, wenn sie zu linear unabhängigen Eigenvektoren gehören. Man kann weiter zeigen, dass Ketten zu verschiedenen Eigenwerten stets linear unabhängig sind (der Beweis hierfür ist ähnlich wie der, dass Eigenvektoren zu verschiedenen Eigenwerten immer linear unabhängig sind). Dies wird hier aber nicht benötigt.

1.10 Die Jordansche Normalform

Folgender Satz hat zahlreiche Anwendungen in verschiedenen Bereichen der Mathematik:

Satz 1.10.1 (Satz von der Jordanschen Normalform)
Sei $n \in \mathbb{N}$, und sei $A \in \mathbb{K}^{n\times n}$ so, dass alle Nullstellen des charakteristischen Polynoms von A zu $\mathbb{K}$ gehören. Dann gibt es eine invertierbare Matrix $B \in \mathbb{K}^{n\times n}$ so, dass $B^{-1}AB$ eine Jordanmatrix ist.

Um diesen Satz zu beweisen, machen wir folgende allgemeine Voraussetzungen und benutzen weitere Bezeichungen:

(a) V sei ein n-dimensionaler Vektorraum über $\mathbb{K}$, und $T \in L(V)$ sei ein fest gewählter Endomorphismus für V.

(b) Die Zahlen $\lambda_1, \ldots, \lambda_\mu$ seien die verschiedenen Nullstellen des charakteristischen Polynoms p_T von T, und $m_1, \ldots, m_\mu$ seien die entsprechenden algebraischen Vielfachheiten. Dann gilt also

$$\sum_{j=1}^{\mu} m_j \;=\; n\,, \qquad p_T(t) \;=\; \prod_{j=1}^{\mu} \big(\lambda_j - t\big)^{m_j}\,. \tag{1.10.1}$$

(c) Als weitere entscheidende Voraussetzung nehmen wir an, dass alle Nullstellen von p_T in $\mathbb{K}$ liegen. Diese Voraussetzung ist natürlich immer erfüllt, wenn $\mathbb{K} = \mathbb{C}$ ist.

(d) Mit Hilfe von T und unter Beachtung von (c) bilden wir folgende weitere Endomorphismen für V:

$$T_j \;=\; \big(\lambda_j\, id \;-\; T\big)^{m_j}\,, \qquad \tilde{T}_j \;=\; \prod_{k\neq j} T_k\,, \tag{1.10.2}$$

wobei das rechts stehende Produkt von Endomorphismen wie üblich als die Hintereinanderausführung zu verstehen ist. Der Index k in diesem Produkt läuft dabei von 1 bis n, wobei wie angegeben der Wert j ausgelassen wird. Beachte auch, dass die Endomorphismen T_k alle miteinander kommutieren, sodass es nicht nötig ist, eine bestimmte Reihenfolge der Faktoren in diesem Produkt festzulegen. Wichtig ist auch, dass im Fall $\mu = 1$ der Endomorphismus $\tilde{T}_1$ durch ein leeres Produkt definiert ist, welches wie üblich die Identität ergibt.

(e) Mit $U_j = \tilde{T}_j(V)$ bezeichnen wir das Bild des Endomorphismus $\tilde{T}_j$. Nach Teil I der Vorlesung sind die U_j Unterräume von V, für alle $j = 1, \ldots, \mu$. Falls $\mu = 1$ ist, ist offenbar $U_1 = V$.

Unter den oben gemachten Voraussetzungen gilt nun folgendes Resultat:

Lemma 1.10.2 *Für alle $j = 1, \ldots, \mu$ gilt:*

(a) *Die U_j sind invariante Unterräume für T.*

(b) *Die Restriktion von T auf den Unterraum U_j ist ein Endomorphismus von U_j mit dem einzigen Eigenwert λ_j.*

(c) *Die Restriktion von $\tilde{T}_j$ auf U_j ist injektiv, also sogar ein Automorphismus für U_j.*

(d) *Es gilt $\dim U_j = m_j$ und $V = U_1 \oplus \ldots \oplus U_\mu$.*

Beweis: Für den Fall $\mu = 1$ sind die gemachten Aussagen alle trivial erfüllt, und deshalb sei jetzt $\mu \geq 2$ angenommen. Für $u \in U_j$ gibt es ein $v \in V$ mit $u = \tilde{T}_j\, v$. Da T mit jedem $\tilde{T}_j$ kommutiert, folgt $T\,u = \tilde{T}_j\,(T\,v)$, und daher ist $T\,u \in U_j$. Also gilt (a). Da $T_j \circ \tilde{T}_j = p_T(T)$ ist, folgt mit dem Satz

von Cayley-Hamilton, dass $T_j \circ \tilde{T}_j = 0$ ist. Also ist $T_j(U_j) = \{0\}$, woraus folgt dass die Restriktion von $\lambda_j\, id - T$ auf U_j ein nilpotenter Endomorphismus von U_j ist und daher den einzigen Eigenwert $\lambda = 0$ hat, was gleichbedeutend mit (b) ist. Aus (b) folgt aber sofort (c), da ja $T - \lambda_k\, id$ auf U_j bijektiv ist, sofern nur $k \neq j$ ist, denn die Zahlen $\lambda_1, \ldots, \lambda_\mu$ sind alle voneinander verschieden. Für den Beweis von (d) benutzen wir ein Resultat, welches auch im Beweis der Partialbruchzerlegung rationaler Funktionen eine wichtige Rolle spielt und sich einfach durch Koeffizientenvergleich bzw. vollständige Induktion über n ergibt (vergleiche auch Bemerkung 3.2.2 und Aufgabe 3.2.4):

- Es gibt eindeutig bestimmte Zahlen $a_{jk} \in \mathbb{K}$, so dass

$$1 \;=\; \sum_{j=1}^{\mu} \Big(\prod_{\ell \neq j} (\lambda_\ell - t)^{m_\ell} \Big) \sum_{k=1}^{m_j} a_{jk}\, (\lambda_j - t)^{m_j - k} \qquad \forall\, t \in \mathbb{K}\,.$$

Aus dieser Identität folgt, wenn wir $p_j(t) = \sum_{k=1}^{m_j} a_{jk}\, (\lambda_j - t)^{m_j - k}$ setzen, die Beziehung

$$id \;=\; \sum_{j=1}^{\mu} \tilde{T}_j \circ p_j(T)\,.$$

Da $\tilde{T}_j$ mit $p_j(T)$ kommutiert, folgt hieraus unter Benutzung von (a):

$$\forall\, v \in V: \qquad v \;=\; \sum_{j=1}^{\mu} u_j\,, \qquad u_j \;=\; p_j(T)\,\big(\tilde{T}_j\, v\big) \in U_j\,.$$

Dies bedeutet, dass $V = U_1 + \ldots + U_\mu$ ist. Um zu zeigen, dass diese Summe direkt ist, seien $u_j \in U_j$ so, dass $0 = u_1 + \ldots + u_\mu$ ist. Sei jetzt ein $j \in \{1, \ldots, n\}$ ausgewählt. Da $(\lambda_k\, id - T)^{m_k}$ auf dem Unterraum U_k gleich der Nullabbildung ist, folgt $\tilde{T}_j\, u_k = 0$ für alle $k \neq j$, $1 \leq k \leq n$, und daher ist

$$0 \;=\; \tilde{T}_j\,(u_1 + \ldots + u_\mu) \;=\; \tilde{T}_j\, u_j\,.$$

Auf dem Unterraum U_j ist $\tilde{T}_j$ nach (c) bijektiv, so dass $u_j = 0$ sein muss. Da aber j beliebig war, folgt dass die Gleichung $0 = u_1 + \ldots + u_\mu$ nur bestehen kann, wenn alle $u_j = 0$ sind. Um jetzt noch $\dim U_j = m_j$ zu zeigen, wählen wir eine Basis von V so, dass die Darstellungsmatrix A von T eine direkte Summe von Matrizen $A_1, \ldots, A_\mu$ ist, deren Größe mit den Dimensionen von $U_1, \ldots, U_\mu$ übereinstimmen; dass dies möglich ist, wurde in Aufgabe 1.8.5 gezeigt. Dann ist aber jedes A_j die Darstellungsmatrix der Restriktion von T auf U_j und hat deshalb nach (b) das charakteristische Polynom $(\lambda_j - t)^{d_j}$, mit $d_j = \dim U_j$. Daraus folgt aber dann dass das charakteristische Polynom von A, also auch von T, von der Form $\prod_{j=1}^{\mu} (\lambda_j - t)^{d_j}$ ist, und das impliziert $d_j = m_j$ für $j = 1, \ldots, \mu$. Damit ist alles gezeigt. □

Aufgabe 1.10.3 *Zeige: Wenn $A, B \in \mathbb{K}^{n \times n}$ ähnlich sind, dann sind für jedes $\lambda \in \mathbb{K}$ auch $A - \lambda I$ und $B - \lambda I$ ähnlich.*

Aufgabe 1.10.4 *Zeige: Ist $A \in \mathbb{K}^{n \times n}$ direkte Summe von Matrizen $A_1, \ldots, A_\mu$, und ist jedes A_j ähnlich zu einer Matrix B_j, dann ist die direkte Summe der $B_1, \ldots, B_\mu$ ähnlich zu A.*

Aus diesen beiden Aufgaben und dem vorausgegangenen Lemma folgt, dass es zum Beweis des Satzes von der Jordanschen Normalform ausreicht, jetzt noch einen nilpotenten Endomorphismus von V zu betrachten. Mit den Bezeichnungen vom Anfang dieses Abschnittes bedeutet das, dass jetzt $\mu = 1$ und $\lambda_1 = 0$ sein sollen.

Aufgabe 1.10.5 *Sei $T \in L(V)$ nilpotent, und sei $U \subset V$ ein invarianter Unterraum, für den die Restriktion von T auf U injektiv ist. Zeige $U = \{0\}$.*

Lemma 1.10.6 *Sei T nilpotent, und seien $U_m = T^m(V)$ für alle $m \in \mathbb{N}_0$. Dann sind alle U_m Unterräume von V, und für ein ℓ gilt $U_\ell = \{0\}$. Ist dabei ℓ minimal gewählt, so folgt*

$$n \;=\; \dim U_0 \;>\; \dim U_1 \;>\; \ldots \;>\; \dim U_{\ell-1} \;>\; \dim U_\ell \;=\; 0\,.$$

Außerdem gilt $U_m = T(U_{m-1})$ und $U_m \subset U_{m-1}$ für $m \geq 1$.

Beweis: Nach Definition ist $T^0 = id$, also $U_0 = V$. Da ein nilpotenter Endomorphismus sicher nicht injektiv ist, folgt dass $\dim U_1 < n = \dim V$ ist. Aus der Definition der U_m folgt $T(U_{m-1}) = U_m$, und daher folgt allgemein $\dim U_m < \dim U_{m-1}$, solange die Restriktion von T auf U_{m-1} nicht injektiv ist, d. h., solange $U_{m-1} \neq \{0\}$ ist. Außerdem ist $U_m = T^{m-1}(T(V))$, und da $T(V) \subset V$ ist, folgt $U_m \subset U_{m-1}$ für $m \geq 1$. □

Satz 1.10.7 *Wenn T ein nilpotenter Endomorphismus von V ist, dann existiert für jeden Unterraum $U_m = T^m(V)$, mit $0 \leq m < \ell$ und ℓ wie oben, also auch für V selber, eine Basis, welche aus Ketten von Hauptvektoren für T zum einzigen Eigenwert $\lambda = 0$ besteht.*

Beweis: Zur Wahl dieser Basen von U_m beginnen wir mit $m = \ell - 1$: In diesem Fall wählen wir eine beliebige Basis $(u_1, \ldots, u_s)$ von $U_{\ell-1}$. Da $T(U_{\ell-1}) = \{0\}$ ist, sind alle diese Basisvektoren Eigenvektoren von T zum (einzigen) Eigenwert $\lambda = 0$, und ein Eigenvektor bildet immer auch eine Kette von Hauptvektoren der Länge 1. Also haben wir unser Ziel für $m = \ell - 1$ erreicht. Sei jetzt für ein $m \in \{0, \ldots, \ell - 1\}$ eine Basis aus Ketten von Hauptvektoren gewählt; die Anzahl der Ketten sei dabei ν genannt, und für $j = 1, \ldots, \nu$ bestehe die entsprechende Kette aus den Vektoren $u_1^{(j)}, \ldots, u_{\mu_j}^{(j)}$. Dann hat die Basis von U_m gerade $d = \mu_1 + \ldots + \mu_\nu$ Elemente, und somit ist $d = \dim U_m$. Wenn $m = 0$ ist, ist nichts mehr zu zeigen, und daher sei $m \geq 1$ angenommen. Wegen $U_{m-1} \supset U_m$ liegen die Vektoren aller Ketten auch in U_{m-1}, bilden dort aber noch keine Basis. Da $T(U_{m-1}) = U_m$ ist, gibt es zu jedem Vektor $u_{\mu_j}^{(j)}$ (mindestens) ein $u_{\mu_j+1}^{(j)} \in U_m$ mit $T\,u_{\mu_j+1}^{(j)} = u_{\mu_j}^{(j)}$. Daher können wir jede der Ketten um den Vektor $u_{\mu_j+1}^{(j)}$ verlängern, und die Gesamtheit dieser Vektoren ist nach Lemma 1.9.4 weiterhin linear unabhängig, da die Eigenvektoren $(u_1^{(1)}, \ldots, u_1^{(\nu)})$ linear unabhängig sind. Diese Eigenvektoren gehören auch zum Kern K der Restriktion von T auf den Raum U_{m-1}; durch eventuelle Wahl weiterer Vektoren $u_{\nu+1\,(1)}, \ldots, u_{\sigma\,(1)}$ erhalten wir eine Basis $(u_1^{(1)}, \ldots, u_1^{(\sigma)})$ von K. Die hinzugekommenen Vektoren sind aber wieder Eigenvektoren von T, also auch Ketten von Hauptvektoren der Länge 1, und durch Hinzunahme dieser neuen Ketten erhalten wir ein System von Ketten von Hauptvektoren, welche wegen Lemma 1.9.4 linear unabhängig sind. Die Anzahl dieser Vektoren ist jetzt gleich $d + \sigma$. Nach der Dimensionsformel für lineare Abbildungen ist die Summe aus den Dimensionen von Bild und Kern von T, eingeschränkt auf U_{m-1}, gleich der Dimension von U_{m-1}. Da das Bild aber gleich U_m ist und deshalb die Dimension d hat, und da die Dimension des Kerns mit σ bezeichnet war, folgt dass die Hauptvektoren aus allen gewählten Ketten eine Basis von U_{m-1} bilden. Das war zu zeigen. □

Aufgabe 1.10.8 (Beweis des Satzes von der Jordanschen Normalform) *Zeige, dass aus den Resultaten dieses Abschnitts der* Satz 1.10.1 *folgt.*

Aufgabe 1.10.9 *Schließe aus* Aufgabe 1.7.3 *und dem Satz von der Jordanschen Normalform: Eine Matrix A ist genau dann diagonalisierbar, wenn ihr Minimalpolynom lauter einfache Nullstellen hat.*

Aufgabe 1.10.10 (Jordansche Normalform mit MAPLE) *Finde den Befehl, mit dem* MAPLE *die Jordansche Normalform einer Matrix berechnen kann.*

1.11 Praktische Berechung der Jordan-Normalform

Im Folgenden sei A eine n-reihige quadratische Matrix, und J bezeichne ihre (im allgemeinen unbekannte) Jordansche Normalform. Wir nehmen an, dass man die Eigenwerte von A berechnen kann, was bei großem n durchaus nicht selbstverständlich ist. Wir wollen jetzt sehen, wie man zu einem Eigenwert λ feststellen kann, wie die Anzahl und Größe der zugehörigen Jordanblöcke in J ist, ohne die Transformationsmatrix T mit $T^{-1}\,A\,T = J$ zu berechnen. Falls man dann doch T, oder jedenfalls die zu λ gehörigen Spalten von T berechnen will, so ist das nach den bereits gezeigten Resultaten dazu äquivalent, ein maximales System von zu λ gehörigen Ketten von linear unabhängigen Hauptvektoren zu berechnen, wobei die Gesamtzahl der zu findenden Hauptvektoren gleich der algebraischen Vielfachheit des Eigenwertes ist. Diese Berechnung kann einerseits so erfolgen, dass man die im letzten Abschnitt eingeführten invarianten Unterräume $U_1, \ldots, U_\mu$ findet und anschließend die einzelnen Schritte des Beweises von Satz 1.10.7 für jeden der verschiedenen Eigenwerte von A durchführt. Wir wollen aber ein etwas einfacheres Verfahren kennen lernen, das im Wesentlichen "nur" die Lösung einer Anzahl von homogenen linearen Gleichungssystemen erfordert.

Wir wollen die folgenden Überlegungen auf den Fall beschränken, dass der Eigenwert λ gleich 0 ist; der allgemeine Fall kann auf diesen zurückgeführt werden, indem man A durch $A - \lambda\,I$ ersetzt.

Definition 1.11.1 *Für $k \geq 0$ sei $r_k = \operatorname{rang} A^k$, und V_k sei die Lösungsmenge von $A^k\,x = 0$. Dann ist also V_k ein Unterraum von $\mathbb{C}^n$ der Dimension $n - r_k$.*

Da ähnliche Matrizen gleichen Rang haben, gilt $r_k = \operatorname{rang} J^k$, und daraus liest man ab, dass $r_1 = n - a_1$ ist, wobei a_1 gleich der Anzahl der Jordanblöcke in J zum Eigenwert $\lambda = 0$ ist. Da die Ränge der Potenzen eines nilpotenten Jordanblockes immer um 1 geringer werden, solange der Exponent nicht die Blockgröße übersteigt, während die Ränge der Potenzen eines anderen, also eines invertierbaren Jordanblockes immer voll sind, ergeben sich folgende Tatsachen:

1. Wenn a_k die Anzahl der nilpotenten Jordanblöcke der Größen $s_j \geq k$ bezeichnet, so ist

   $$r_k \;=\; r_{k-1} \;-\; a_k \qquad \forall\; k \geq 0\,.$$

   Also kann alleine aus den Zahlen r_k die Anzahl und Größe aller zu $\lambda = 0$ gehörigen Jordanblöcke bestimmt werden.

2. Aus Punkt 1 folgt weiter, dass

   $$n \;=\; r_0 \;>\; r_1 \;> \ldots > \; r_\ell \;=\; r_{\ell+1} \;=\; \ldots$$

   wobei ℓ das Maximum der Größen aller Jordanblöcke (zum Eigenwert $\lambda = 0$) ist. Es reicht also, so viele Zahlen r_k zu berechnen, bis zum ersten Mal zwei aufeinander folgende Zahlen gleich sind!

3. Da $A^k\,x = A\,(A^{k-1}\,x)$ ist, folgt

   $$\{0\} \;=\; V_0 \subset\; V_1 \;\subset \ldots \subset\; V_\ell \;=\; V_{\ell+1} \;=\; \ldots$$

   wobei die ersten Inklusionen strikt sind, da ja die Dimensionen der Räume streng wachsen. Also muss man auch nur endlich viele Räume V_k berechnen, das heißt die allgemeine Lösung von endlich vielen Gleichungssystemen ausrechnen.

4. Für $k \leq \ell$ ist V_{k-1} ein echter Teilraum von V_k, und deshalb gibt es ein W_k der Dimension $n - r_k - (n - r_{k-1}) = a_k$, so dass

   $$V_k \;=\; V_{k-1} \;\oplus\; W_k\,.$$

   Eine Basis $(w_1, \ldots, w_{a_k})$ von W_k erhält man, indem man eine Basis von V_{k-1} berechnet und zu einer Basis von V_k ergänzt. Wir nehmen an, dass eine solche Basis gegeben ist. Setzt man

   $$v_\nu^{(j)} \;=\; A^{k-\nu}\,w_j \qquad \forall\; j = 1, \ldots, a_k\,, \quad 1 \leq \nu \leq k\,,$$

so bilden die Vektoren $v_1^{(j)}, \ldots, v_k^{(j)}$ eine Kette aus Hauptvektoren. Um zu zeigen, dass alle diese berechneten Hauptvektoren linear unabhängig sind, genügt es, die lineare Unabhängigkeit der Eigenvektoren $v_1^{(j)}$ zu beweisen. Dazu seien $\alpha_j \in \mathbb{C}$ so, dass $0 = \sum \alpha_j \, v_1^{(j)} (= A^{k-1} \sum \alpha_j \, w_j)$ gilt. Dann folgt aber $\sum \alpha_j \, w_j \in V_{k-1}$, was nach Definition einer direkten Summe nur sein kann, wenn $\sum \alpha_j \, w_j = 0$ ist. Das impliziert aber dass alle $\alpha_j = 0$ sind. Also kann man zu jeder Basis von W_k Ketten von Hauptvektoren der Länge k berechnen, wobei die so gefundenen Vektoren alle linear unabhängig sind.

5. Wenn man im letzten Punkt $k = \ell$ wählt, so erhält man alle Ketten von Hauptvektoren maximaler Länge. Wenn man anschließend für irgend ein $k \geq \ell - 1$ annimmt, dass man schon Ketten der Längen $\geq k+1$ berechnet hat, so kann man durch Verkürzung der Ketten eine entsprechende Zahl von Ketten mit Länge k erhalten. Die letzten Vektoren dieser Ketten liegen dann in W_k. Falls sie sogar eine Basis bilden, gibt es keine weiteren linear unabhängigen Ketten dieser Länge. Im anderen Fall kann man diese Vektoren zu einer Basis von W_k ergänzen und dann zu jedem neuen Basisvektor eine weitere Kette der Länge k finden. Auf diese Weise erhält man ein maximales System linear unabhängiger Ketten zum Eigenwert $\lambda = 0$.

Kapitel 2

Singulärwertzerlegung und Pseudo-Inverse

Wir wollen in diesem Kapitel zeigen, dass es auch für nicht-invertierbare, insbesondere sogar für nicht-quadratische Matrizen A möglich ist, eine andere Matrix B zu definieren, die ähnliche Eigenschaften wie eine inverse Matrix hat. Insbesondere wird sich ergeben, wie man bei linearen Gleichungssystemen, welche nicht eindeutig lösbar sind, eine in gewissem Sinne optimale Lösung bestimmen kann. Tatsächlich gilt etwas analoges sogar für unlösbare Systeme – wie dies zu verstehen ist, wird später noch klar. Jedenfalls spielen solche Resultate z. B. in der numerischen Mathematik eine wichtige Rolle!

2.1 Die Singulärwertzerlegung beliebiger Matrizen

Im Folgenden seien $n, m \in \mathbb{N}$ immer fest gewählt, wobei $n \neq m$ sein kann.

Lemma 2.1.1 *Für jede Matrix $A \in \mathbb{K}^{n\times m}$ ist $\overline{A}^T A$ hermitesch und positiv semidefinit, und sogar positiv definit, falls* $\operatorname{rang} A = m$ *ist – dies kann allerdings nur gelten, falls $n \geq m$ ist. Jedenfalls sind alle Eigenwerte von $\overline{A}^T A$ nicht-negative reelle Zahlen.*

Beweis: Aus den Rechenregeln für transponierte Matrizen folgt dass $\overline{A}^T A$ hermitesch ist. Für $y = A\,x$ ist $\overline{x}^T \overline{A}^T A\; x = \overline{y}^T y = \|y\|^2 \geq 0$, und daher ist $\overline{A}^T A$ positiv semidefinit. Falls $\operatorname{rang} A = m$ ist, folgt aus der Dimensionsformel für lineare Gleichungssysteme dass $x = 0$ aus $y = 0$ folgt, und dann ist die Matrix sogar positiv definit. □

Definition 2.1.2 *Wir nennen eine Matrix $\Sigma = [\sigma_{jk}] \in \mathbb{K}^{n\times m}$ eine* allgemeine Diagonalmatrix, *falls für alle $j \neq k$ immer $\sigma_{jk} = 0$ ist. Die Einträge σ_{jj} werden* Diagonalelemente *genannt. Eine allgemeine Diagonalmatrix ist also genau dann eine Diagonalmatrix im üblichen Sinn, wenn $n = m$ ist, d. h., wenn sie quadratisch ist.*

Aufgabe 2.1.3 *Zeige: Anders als eine "richtige" Diagonalmatrix ist eine nicht quadratische allgemeine Diagonalmatrix Σ nicht symmetrisch. Es gilt aber dass $\Sigma^T \Sigma$ eine Diagonalmatrix ist. Finde heraus, wie die Diagonalelemente von $\Sigma^T \Sigma$ mit den Quadraten der σ_{jj} zusammenhängen.*

Satz 2.1.4 (Singulärwertzerlegung) *Zu jeder Matrix $A \in \mathbb{K}^{n\times m}$ gibt es unitäre Matrizen $U_1 \in \mathbb{K}^{n\times n}$, $U_2 \in \mathbb{K}^{m\times m}$, sowie eine allgemeine Diagonalmatrix $\Sigma \in \mathbb{R}^{n\times m}$ mit nicht-negativen Diagonalelementen, so dass gilt*

$$A \;=\; U_1\,\Sigma\,\overline{U}_2^T\,. \tag{2.1.1}$$

Dabei sind die Spalten von U_1, bzw. von U_2, Eigenvektoren zu $A\,\overline{A}^T$, bzw. zu $\overline{A}^T A$, und auf der Diagonalen von Σ stehen die Quadratwurzeln aus den positiven Eigenwerten von $A\,\overline{A}^T$, bzw. $\overline{A}^T A$, die evtl. noch mit Nullen aufgefüllt sind. Insbesondere stimmen die positiven Eigenwerte, einschließlich ihrer Vielfachheit, von $A\,\overline{A}^T$ und $\overline{A}^T A$ überein.

Beweis: Wenn A wie behauptet zerlegt werden kann, dann folgt dass $\overline{A}^T = U_2\,\Sigma^T\,\overline{U}_1^T$, und dann ist

$$\overline{A}^T\,A\,U_2 = \;U_2\,\Sigma^T\Sigma\,, \qquad A\,\overline{A}^T\,U_1 = \;U_1\,\Sigma\,\Sigma^T\,.$$

Daher sind die Spalten von U_1, U_2 so, wie im Satz behauptet wird. Um jetzt die Existenz der Zerlegung zu zeigen, sei $U_2 = [u_1^{(2)}, \ldots, u_m^{(2)}]$ so, dass die Spalten eine Orthonormalbasis von Eigenvektoren zur Matrix $\overline{A}^T A$ bilden. Es gilt also $\overline{A}^T A\,u_j^{(2)} = \lambda_j\,u_j^{(2)}$, $1 \le j \le m$, wobei die Anordnung der Spalten noch so sei, dass genau die ersten r der λ_j positiv sind (beachte, dass alle Eigenwerte nach Lemma 2.1.1 nicht-negativ sind). Wir definieren jetzt $u_j^{(1)} = \lambda_j^{-1/2}\,A\,u_j^{(2)}$, $1 \le j \le r$, und rechnen nach dass diese Vektoren ein Orthonormalsystem in $\mathbb{K}^n$ sind. Daher können wir weitere Vektoren so wählen, dass $(u_1^{(1)}, \ldots, u_n^{(1)})$ eine Orthonormalbasis in $\mathbb{K}^n$ ist, und wir setzen $U_1 = [u_1^{(1)}, \ldots, u_n^{(1)}]$. Die Matrix $\Sigma := \overline{U_1}^T A\,U_2$ ist dann in der Tat eine allgemeine Diagonalmatrix, auf deren Diagonale die Quadratwurzeln der λ_j stehen. □

Definition 2.1.5 (Singulärwerte) *Man nennt die positiven Elemente auf der Diagonalen von Σ auch die* Singulärwerte *von A; ihre Anzahl ist gleich dem Rang von A. Man kann o. B. d. A. annehmen, dass sie so angeordnet sind, dass*

$$\sigma_{11} \;\ge\; \sigma_{22} \;\ge\; \ldots \;\ge\; \sigma_{rr}\,.$$

Der in der Numerik wichtigste Fall der Singulärwertzerlegung ist der einer reellen Matrix A, und dann zeigt der Beweis, dass man auch die Matrizen U_1 und U_1 reell wählen kann. In diesem Fall kann man aus der Zerlegung folgendes ablesen:

- Die ersten r Spalten von U_1 sind eine Orthonormalbasis des *Spaltenraums* d. i. die Menge aller Linearkombinationen der Spalten, von A. Die übrigen Spalten sind eine Orthonormalbasis des *Linksnullraums* von A, d. i. die Menge der Vektoren y mit $y^T A = 0$.
- Da (bei reellem A) gilt $A^T = U_2\,\Sigma^T\,U_1^T$, gilt eine entsprechende Aussage für die Spalten von U_2.

Die Singulärwerte einer Matrix sind bis auf ihre Reihenfolge eindeutig bestimmt. Aus dem Beweis des Satzes sieht man, dass man bei der Wahl der unitären Matrizen noch gewisse Freiheiten hat; z. B. kann man im reellen Fall für jede Spalte von U_2 noch ein Vorzeichen wählen, d. h. einige (oder alle) der Spalten mit -1 multiplizieren. Falls A vollen Rang hat, und falls alle Singulärwerte verschieden sind, ist dies die einzige Freiheit, die man hat. Dies spielt aber im Folgenden keine Rolle.

2.2 Die Polarzerlegung

Eine komplexe Zahl $z = x + i\,y = r\,\mathrm{e}^{i\alpha}$, $x, y, \alpha \in \mathbb{R}$, $r \in \mathbb{R}_+$, entspricht in gewissen Sinn einer Matrix

$$Z \;=\; \begin{bmatrix} x & y \\ -y & x \end{bmatrix} \;=\; \begin{bmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{bmatrix}\;(r\,I)\;,$$

und die Drehmatrix ist unitär (sogar orthogonal), während $r\,I$ eine (sehr einfache) positiv definite Matrix ist. In Analogie spricht man beim folgenden Resultat von der *Polarzerlegung* einer quadratischen Matrix.

Satz 2.2.1 *Eine Matrix $A \in \mathbb{K}^{n\times n}$ kann geschrieben werden als $A = U\,S$, wobei U unitär und S hermitesch und positiv semidefinit ist.*

Beweis: Aus der Singulärwertzerlegung von A folgt

$$A \;=\; (U_1\,\overline{U}_2^T)\,(U_2\,\Sigma\,\overline{U}_2^T) \;=\; U\,S\,,$$

wobei $U = U_1\,\overline{U}_2^T$ unitär und $S = U_2\,\Sigma\,\overline{U}_2^T$ hermitesch und positiv semidefinit ist, was zu zeigen war. □

2.3 Die Pseudo-Inverse beliebiger Matrizen

Definition 2.3.1 (Optimale Lösung unlösbarer Gleichungssysteme) *Für $A \in \mathbb{K}^{n\times m}$ und $b \in \mathbb{K}^n$ kann das inhomogene lineare Gleichungssystem $A\,x = b$ unlösbar sein, und dies ist genau dann der Fall, wenn b nicht im Spaltenraum von A liegt. In diesem Fall sucht man oft einen oder alle Vektoren x, für welche der "Fehler" $\|b - A\,x\|$ minimal wird. Nach dem Satz über die beste Aproximation ist dies genau dann der Fall, wenn x eine Lösung von $A\,x = p$ ist, wobei p die orthogonale Projektion von b auf den Spaltenraum, also die lineare Hülle der Spalten von A, bezeichnet. In diesem Fall ist $A\,x = p$ zwar lösbar, aber im Allgemeinen nicht eindeutig lösbar, und wir nennen eine Lösung x^+* optimal, *wenn die Norm von x^+, verglichen mit der Norm jeder anderen Lösung von $A\,x = p$, minimal ist.*

Definition 2.3.2 (Pseudo-Inverse) *Wenn A so wie in* (2.1.1) *zerlegt ist, und wenn $\sigma_1, \ldots, \sigma_r$ die Singulärwerte von A bezeichnen, dann setzen wir*

$$A^+ \;=\; U_2\,\Sigma^+\,\overline{U}_1^T\,, \tag{2.3.1}$$

wobei $\Sigma^+ \in \mathbb{K}^{m\times n}$ die allgemeine Diagonalmatrix bezeichnet, die zu den Singulärwerten $\sigma_1^{-1}, \ldots, \sigma_r^{-1}$ gehört. Beachte, dass Σ^+ also aus Σ durch Transposition und Kehrwertbildung der singulären Werte entsteht! Wir nennen A^+ die Pseudo-Inverse *von A. In der Literatur findet man oft auch die Bezeichnung* Moore-Penrose-Inverse. *Da die Matrizen U_1, U_2 in der Singulärwertzerlegung nicht eindeutig bestimmt sind, ist noch nicht klar, ob A^+ eindeutig festliegt, dies wird aber im nächsten Satz mitbewiesen.*

Wir zeigen jetzt, dass die optimale Lösung eines linearen Gleichungssystems eindeutig festgelegt ist und sogar berechnet werden kann, wenn man A^+ kennt.

Satz 2.3.3 *Sei eine Matrix $A \in \mathbb{K}^{n\times m}$ gegeben. Zu jedem $b \in \mathbb{K}^n$ gibt es ein eindeutig bestimmtes $x^+ \in \mathbb{K}^m$, welches die im Sinne der obigen Definition optimale Lösung des inhomogenen Gleichungssystems $A\,x = p$ ist, wobei p die orthogonale Projektion des Vektors b auf den Spaltenraum von A bezeichnet. Die Abbildung $b \longmapsto x^+$ ist linear und gegeben durch*

$$x^+ \;=\; A^+\,b\,.$$

Insbesondere ist die Pseudo-Inverse A^+ als die Darstellungsmatrix dieser linearen Abbildung bezüglich der kanonischen Basis in $\mathbb{R}^n$ eindeutig festgelegt.

Beweis: Sei A so wie in (2.1.1) zerlegt. Mit $y = \overline{U}_2^T x$, $q = \overline{U}_1^T p$ ist die Gleichung $A\,x = p$ äquivalent zu $\Sigma\,y = q$. Da eine Lösung existieren muss, folgt hieraus $q_{r+1} = \ldots = q_n = 0$ (was aber nicht gebraucht

wird), und wir erhalten $y_j = \sigma_j^{-1} q_j$ für $1 \leq j \leq r$, während die übrigen y_j beliebig gewählt werden können. Da U_2 unitär, also längentreu ist, folgt $\|y\| = \|x\|$, und diese Norm ist genau dann minimal, wenn wir $y_{r+1} = \ldots = y_n = 0$ setzen. Dies zeigt sowohl Existenz als auch Eindeutigkeit der optimalen Lösung x^+, und es folgt $x^+ = A^+ p$. Nach Wahl von p ist $d = b - p$ orthogonal zu allen Spalten von A, und das bedeutet das gleiche wie $\overline{A}^T d = 0$, was wiederum zu $\Sigma^T \overline{U}_1^T d = 0$ äquivalent ist. Dies gilt aber genau dann, wenn d zu den ersten r Spalten von U_1 orthogonal ist, woraus sofort $A^+ d = 0$ folgt. Also ist $x^+ = A^+ b$, und somit ist die Abbildung $b \longmapsto x^+$ linear und hat die Darstellungsmatrix A^+. Dies wiederum legt A^+ eindeutig fest, unabhängig von der Wahl der Matrizen U_1 und U_2. □

Aufgabe 2.3.4 *Sei $U \subset \mathbb{R}^m$, und sei $A \in \mathbb{R}^{m \times n}$ so, dass die Spalten von A ein Erzeugendensystem von U sind. Zeige, dass für ein beliebiges $b \in \mathbb{R}^m$ der Vektor $A\,A^+\,b$ gleich der orthogonalen Prokektion von b auf U ist.*

Aufgabe 2.3.5 *Sei $A \in \mathbb{K}^{n \times m}$, und sei x so, dass $x^T A \overline{A}^T = 0$ ist. Zeige dass dann bereits $x^T A = 0$ sein muss.*

Satz 2.3.6 (Eigenschaften der Pseudo-Inversen) *Für jede Matrix $A \in \mathbb{K}^{n \times m}$ gilt:*

(a) $A\,A^+\,A \;=\; A$.

(b) $A^+\,A\,A^+ \;=\; A^+$.

(c) $A^+\,A$ *und* $A\,A^+$ *sind hermitesch.*

Diese Eigenschaften legen A^+ fest, d. h., ist $B \in \mathbb{K}^{m \times n}$ so, dass (a)-(c) *für B anstelle von A^+ gelten, so folgt $B = A^+$.*

Beweis: Man rechnet einfach nach, dass (a)-(c) gelten, weil nämlich die entsprechenden Aussagen für Σ und Σ^+ richtig sind. Sei jetzt B so, dass (a)-(c) ebenfalls erfüllt sind. Dann folgt mit (a) und (c), mit B anstelle von A^+:

$$\overline{A}^T \;=\; \overline{A}^T \overline{B}^T \overline{A}^T \;=\; \left(\overline{B\,A}\right)^T \overline{A}^T \;=\; B\,A\,\overline{A}^T\,,$$

und da gleiches auch gilt für A^+ anstelle von B, folgt $(B - A^+)\,A\,\overline{A}^T = 0$. Mit Aufgabe 2.3.5 gilt dies aber nur, falls $(B - A^+)\,A = 0$ ist, und dies bedeutet dass die Zeilen von $B - A^+$ auf den Spalten von $\overline{A}$ senkrecht stehen. Aus (b) und (c) folgt analog dass $\overline{B}^T = A\,B_1$ mit einer geeigneten Matrix B_1, und gleiches gilt wieder für A^+ anstelle von B und einer anderen Matrix B_2. Daraus folgt wiederum dass $B - A^+ = C\,\overline{A}^T$ gilt für eine geeignete Matrix C. Dies bedeutet, dass die Zeilen von $B - A^+$ Linearkombinationen der Zeilen von $\overline{A}^T$ (also den Transponierten der Spalten von $\overline{A}$) sind, und daher folgt $B - A^+ = 0$. □

Aufgabe 2.3.7 *Benutze den letzten Satz, um zu zeigen: Wenn A quadratisch und invertierbar ist, dann ist $A^+ = A^{-1}$. Zeige, dass dies auch bereits aus dem vorangegangenen Satz folgt. Zeige weiter für allgemeines A dass die Pseudoinverse zu A^+ gleich A ist.*

Aufgabe 2.3.8 *Finde die Singulärwertzerlegung und die Pseudoinverse für folgende Matrizen:*

$$A \;=\; \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix}\,, \qquad B \;=\; \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}\,, \qquad C \;=\; \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}\,.$$

Aufgabe 2.3.9 *Finde die optimale Lösung von $A\,x = b$ für*

$$A \;=\; \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix}\,, \qquad b \;=\; \begin{bmatrix} 0 \\ 2 \\ 2 \end{bmatrix}\,.$$

Aufgabe 2.3.10 *Zeige: Wenn die Spalten von A linear unabhängig sind, dann ist $\overline{A}^T A$ invertierbar, und $A^+ = (\overline{A}^T A)^{-1} \overline{A}^T$. Finde eine analoge Aussage, falls die Zeilen von A linear unabhängig sind.*

Kapitel 3

Matrixfunktionen

In diesem Kapitel soll definiert werden, was man unter $f(A)$ verstehen kann, wenn f eine geeignete Funktion und A eine quadratische Matrix ist. Im Zusammenhang mit dem Satz von Cayley-Hamilton haben wir bereits definiert, was $f(A)$ bedeutet, falls f ein Polynom ist. Dies soll jetzt wiederholt und zur Grundlage einer allgemeinen Definition gemacht werden.

3.1 Matrixpolynome

Definition 3.1.1 (Matrixpolynome) *Für ein Polynom $p(t) = \sum_{j=0}^{m} p_j\, t^j \in \mathbb{C}[t]$ und eine Matrix $A \in \mathbb{C}^{n\times n}$ sei*

$$p(A) \;=\; \sum_{j=0}^{m} p_j\, A^j\,, \tag{3.1.1}$$

wobei die Potenzen von A wie üblich zu verstehen sind, also insbesondere $A^0 = I$ ist. Offenbar ist dann $p(A) \in \mathbb{C}^{n\times n}$.

Aufgabe 3.1.2 (Direkte Summe von Matrixpolynomen) *Zeige: Wenn C direkte Summe zweier quadratischer Matrizen A und B ist, dann ist für jedes Polynom $p(t)$ die Matrix $p(C)$ direkte Summe von $p(A)$ und $p(B)$.*

Proposition 3.1.3 (Eigenschaften von Matrixpolynomen) *Sei $A \in \mathbb{C}^{n\times n}$, und sei $\psi_A(t)$ das Minimalpolynom von A. Dann gilt:*

(a) *Die Abbildung $p(t) \longmapsto p(A)$ ist linear von $\mathbb{C}[t]$ nach $\mathbb{C}^{n\times n}$.*

(b) *Sind $p_1(t), p_2(t) \in \mathbb{C}[t]$, und ist $p(t) = p_1(t)\, p_2(t)$, so ist $p(A) = p_1(A)\, p_2(A)$.*

(c) *Sind $p_1(t), p_2(t) \in \mathbb{C}[t]$, und ist $p(t) = p_1(p_2(t))$, so ist $p(A) = p_1(p_2(A))$.*

(d) *Sind $p_1(t), p_2(t) \in \mathbb{C}[t]$, so ist $p_1(A) = p_2(A)$ genau dann, wenn die Differenz $d(t) := p_1(t) - p_2(t)$ durch das Minimalpolynom $\psi_A(t)$ teilbar ist.*

Beweis: Seien $p_1(t) = \sum_j p_j\, t^j, p_2(t) = \sum_j q_j\, t^j \in \mathbb{C}[t]$, wobei wir uns merken, dass die Summation über j formal über alle nicht-negativen ganzen Zahlen erstreckt werden soll, und dass von den Koeffizienten p_j, q_j höchstens endlich viele von 0 verschieden sind. Dann ist

$$p_1(t) \;+ p_2(t) \;=\; \sum_j (p_j + q_j)\, t^j\,, \qquad p_1(t)\, p_2(t) \;=\; \sum_j \Big(\sum_{k=0}^{j} p_{j-k}\, q_k \Big)\; t^j\,.$$

Mit diesen Darstellungen zeigt man leicht (a) und (b). Für den Nachweis von (c) muss man nur beachten, dass $p_1(p_2(t)) = \sum_j p_j[p_2(t)]^j$ ist, und dass das gleiche auch für A anstelle von t gilt. Da aus $p_1(A) = p_2(A)$ folgt, dass die Differenz $d(t)$ ein anullierendes Polynom für A ist, ergibt sich aus den Ergebnissen des ersten Teils der Vorlesung die letzte der Behauptungen. □

Aufgabe 3.1.4 (Eigenwerte von Matrixpolynomen) *Sei $A \in \mathbb{C}^{n\times n}$, und sei $p(t) \in \mathbb{C}[t]$. Zeige: Ist λ ein Eigenwert von A, so ist $p(\lambda)$ ein Eigenwert von $p(A)$, und der Eigenraum von $p(A)$ zum Eigenwert $p(\lambda)$ enthält den Eigenraum von A zum Eigenwert λ. Gib ein Beispiel, für welches beide Räume nicht gleich sind.*

Aufgabe 3.1.5 (Matrixpolynome und Ähnlichkeit) *Zeige: Sind $A, B, T \in \mathbb{C}^{n\times n}$ so, dass $\det T \neq 0$ und $B = T^{-1}\,A\,T$ ist, so folgt für jedes $p(t) \in \mathbb{C}[t]$, dass $p(B) = T^{-1}\,p(A)\,T$ gilt.*

3.2 Das Lagrange-Sylvestersche Interpolationspolynom

Für die Definition von $f(A)$ für Funktionen, welche keine Polynome sind, benötigen wir ein Interpolationspolynom, welches allgemeiner ist als das im Teil I der Vorlesung behandelte sog. Lagrangesche Interpolationspolynom:

Im Folgenden seien verschiedene *Stützstellen* $\lambda_1, \ldots, \lambda_s \in \mathbb{C}$ und *Vielfachheiten* $m_1, \ldots, m_s \in \mathbb{N}$ gegeben. Das *Lagrange-Sylvestersche Interpolationsproblem* lautet dann wie folgt:

- Für beliebige *Werte* $f_{jk} \in \mathbb{C}$, für $0 \leq k \leq m_j - 1$ und $1 \leq j \leq s$, finde ein Polynom $p(t)$ kleinsten Grades, welches die *Interpolationsbedingungen*

$$p^{(k)}(\lambda_j) \;=\; f_{jk} \qquad \forall\; k = 0, \ldots, m_j - 1\,, \quad j = 1, \ldots, s$$

  erfüllt.

Wenn alle Vielfachheiten $m_j = 1$ sind, ist aus dem ersten Teil der Vorlesung bekannt, dass es genau ein Polynom vom Grad $\leq s - 1$ gibt, welches das Interpolationsproblem löst. Wir zeigen jetzt das analoge Resultat für den allgemeinen Fall:

Satz 3.2.1 *Es gibt genau ein Polynom vom Grade echt kleiner als $m := m_1 + \ldots + m_s$, welches den oben formulierten Interpolationsbedingungen genügt.*

Beweis: Sei $\psi(t) = (t - \lambda_1)^{m_1} \cdot \ldots \cdot (t - \lambda_s)^{m_s}$ gesetzt, und sei für den Moment angenommen, dass wir das gesuchte Polynom $p(t)$ bereits gefunden hätten. Dann ist $p(t)/\psi(t)$ eine rationale Funktion und kann in ihre Partialbrüche zerlegt werden, d. h., es gibt eindeutig bestimmte Zahlen $a_{jk} \in \mathbb{C}$, so dass

$$\frac{p(t)}{\psi(t)} \;=\; \sum_{j=1}^{s} \sum_{k=1}^{m_j} \frac{a_{jk}}{(t - \lambda_j)^k}\;. \tag{3.2.1}$$

Die Bestimmung des Polynoms $p(t)$, falls es denn existiert, ist somit äquivalent zur Bestimmung der Zahlen $a_{jk} \in \mathbb{C}$, und um diese zu finden, bilden wir die Hilfspolynome $\psi_\nu(t) = \psi(t)/(t - \lambda_\nu)^{m_\nu}$ für $\nu = 1, \ldots, s$ und schreiben (3.2.1) in der Form

$$\frac{p(t)}{\psi_\nu(t)} \;=\; \sum_{k=1}^{m_\nu} a_{\nu k}\,(t - \lambda_\nu)^{m_\nu - k} \;+\; (t - \lambda_\nu)^{m_\nu}\, r_\nu(t)\,,$$

wobei die $r_\nu(t)$, genau wie die linke Seite, rationale Funktionen sind, die bei λ_ν keine Pole haben. Somit können wir $t = \lambda_\nu$ in diese Beziehung einsetzen und erhalten (da ja $p(\lambda_\nu) = f_{\nu 0}$ sein soll), dass $a_{\nu m_\nu} = f_{\nu 0}/\psi_\nu(\lambda_\nu)$. Durch Differenzieren der Gleichung und anschließendes Einsetzen findet man dann die Zahl $a_{\nu,m_\nu-1}$, und so fort. Damit ist der Satz bewiesen. □

Bemerkung 3.2.2 *Mit dem obigen Beweis wird eigentlich auch der Satz über die Partialbruchzerlegung bewiesen: Wenn man $p(t)$ kennt, kennt man ja auch die Werte f_{jk} und kann daraus die Zahlen a_{jk} ausrechnen. Der Identitätssatz für Polynome zeigt dann die Richtigkeit der Gleichung* (3.2.1).

Definition 3.2.3 *Wenn wir zwei ganze Zahlen ν, μ mit $1 \le \nu \le s$ und $0 \le \mu \le m_\nu - 1$ wählen, dann gibt es nach dem obigen Satz genau ein Polynom $\psi_{\nu\mu}(t)$ vom Grad echt kleiner als m, welches die Bedingungen*

$$\psi_{\nu\mu}^{(k)}(\lambda_j) \;=\; \delta_{j\nu}\,\delta_{k\mu} \qquad \forall\; k = 0, \ldots, m_j - 1\,, \quad j = 1, \ldots, s$$

erfüllt. Diese Polynome heißen auch die Basispolynome *des Interpolationsproblems, denn mit ihrer Hilfe erhält man die Lösung $p(t)$ der allgemeinen Interpolationsaufgabe in der Form*

$$p(t) \;=\; \sum_{\nu=1}^{s} \sum_{\mu=0}^{m_\nu} f_{\nu\mu}\; \psi_{\nu\mu}(t)\;. \tag{3.2.2}$$

Diese Darstellung ist grundlegend für Abschnitt 3.5.

Aufgabe 3.2.4 (Zerlegung der Eins) *Zeige für die oben eingeführten Basispolynome $\psi_{\nu\mu}(t)$ die Identität*

$$\sum_{\nu=1}^{s} \psi_{\nu 0}(t) \;=\; 1 \qquad \forall\; t \in \mathbb{C}\,.$$

Anleitung: Benutze, dass die Darstellung (3.2.2) auch für $p(t) \equiv 1$ gelten muss, und finde heraus, wie dann die Werte $f_{\nu\mu}$ aussehen müssen.

3.3 Definition der Matrixfunktion

Wir schreiben das Minimalpolynom $\psi_A(t)$ der Matrix $A \in \mathbb{C}^{n\times n}$ in der Form

$$\psi_A(t) \;=\; (t-\lambda_1)^{m_1} \cdot \ldots \cdot (t-\lambda_s)^{m_s}\,, \tag{3.3.1}$$

mit verschiedenen Zahlen $\lambda_j \in \mathbb{C}$ und natürlichen Zahlen m_j. Die Werte λ_j sind dann gerade die verschiedenen Eigenwerte von A, und die m_j sind die entsprechenden Vielfachheiten der Nullstellen λ_j von $\psi_A(t)$. Man erkennt die Vielfachheit der Nullstelle λ_j gerade daran, dass λ_j Nullstelle aller Ableitungen $\psi_A^{(k)}(t)$, $k = 0, \ldots, m_j - 1$, des Minimalpolynoms ist. Man sieht außerdem, dass $\psi_A(t)$ genau dann ein Polynom $p(t)$ teilt, wenn $p(t)$ (mindestens) die Nullstellen λ_j mit einer Vielfachheit $\ge m_j$ hat, für $j = 1, \ldots, s$. Da das Minimalpolynom das charakteristische Polynom von A teilt, ist m_j höchstens gleich der algebraischen Vielfachheit des Eigenwertes λ_j.

Die Aussage (d) von Proposition 3.1.3 lässt sich auch so ausdrücken:

- Genau dann gilt $p_1(A) = p_2(A)$, wenn $p_1^{(k)}(\lambda_j) = p_2^{(k)}(\lambda_j)$ ist für $k = 0, \ldots, m_j - 1$ und $j = 1, \ldots, s$.

Diese Tatsache motiviert die folgende Definition der Matrixfunktion:

Definition 3.3.1 *Sei $A \in \mathbb{C}^{n\times n}$, und sei das Minimalpoynom $\psi_A(t)$ wie in* (3.3.1) *geschrieben. Wir sagen dann, dass eine auf einer Menge $D \subset \mathbb{C}$ definierte, reell- oder komplexwertige Funktion f* auf dem Spektrum von A definiert ist*, wenn f an jeder Stelle λ_j definiert und dort mindestens $(m_j - 1)$-mal differenzierbar ist, und wir nennen die Zahlen*

$$f^{(k)}(\lambda_j)\,, \qquad k = 0, \ldots, m_j - 1\,, \quad j = 1, \ldots, s\,,$$

auch gelegentlich die Werte von f auf dem Spektrum von A. *Wenn $p(t)$ ein Polynom ist, welches auf dem Spektrum von A dieselben Werte wie f hat, dann definieren wir*

$$f(A) \;:=\; p(A)\,. \tag{3.3.2}$$

Aus der oben stehenden Beobachtung folgt dann, dass die Definition von $f(A)$ nicht von der Wahl des Polynoms $p(t)$ abhängt, und aus Satz 3.2.1 *folgt dass es immer ein solches Polynom gibt. Außerdem stimmt die Definition von f, im Fall dass f ein Polynom ist, mit der früher gegebenen Definition von Matrixpolynomen überein. Eine Möglichkeit zur Berechnung von $f(A)$ besteht darin, das zugehörige Interpolationspolynom zu berechnen und in dieses A einzusetzen; dies ist aber nicht die beste Methode. Auf andere Berechnungsmethoden wird etwas später eingegangen.*

Bemerkung 3.3.2 *Für die Definition von $f(A)$ ist es nicht entscheidend, ob man die Ableitung als Grenzwert des Differenzenquotienten in $\mathbb{C}$ oder, falls $D \subset \mathbb{R}$ ist, in $\mathbb{R}$ auffasst. Man kann sogar Funktionen erlauben, die z. B. an einem oder mehreren Punkten nur rechtsseitig stetig bzw. differenzierbar sind, in jedem Fall müssen die in der Definition vorkommenden Größen definiert sein. Darüber hinaus müssen aber für die betrachteten Ableitungen die üblichen Rechenregeln gelten, da man sonst die im nächsten Abschnitt betrachteten Eigenschaften nicht beweisen kann – diese sind für die reelle und die komplexe Ableitung erfüllt, aber die Kettenregel ist für einseitige Ableitungen i. A. nicht richtig! Als ein wichtiges Beispiel betrachten wir die Funktion $f(t) = \sqrt{t}$ für $t \geq 0$. Sie ist für $t > 0$ beliebig oft differenzierbar, und deshalb kann $f(A)$ für alle Matrizen, deren Eigenwerte positiv sind, gebildet werden, dies sind z. B. alle positiv definiten hermiteschen Matrizen. Falls 0 ebenfalls ein Eigenwert ist, gibt es ein Problem, wenn das Minimalpolynom im Nullpunkt eine mehrfache Nullstelle hat, da dann f entsprechend oft differenzierbar sein müsste, was aber nicht der Fall ist. Falls die Nullstelle dagegen einfach ist, gibt es kein Problem mit der Definition von $f(A)$. Dies ist z. B. so, wenn A hermitesch und positiv semidefinit ist. Wenn A negative oder komplexe Eigenwerte hat, muss man die komplexe Funktion $f(t) = t^{1/2}$ benutzen, diese wird z. B. in der Vorlesung* Elemente der Funktionentheorie *behandelt. In jedem Fall wird die entsprechende Matrix in der Regel mit $A^{1/2}$ bezeichnet. Ob dann $A^{1/2}\,A^{1/2} = A$ gilt, wird noch diskutiert.*

Bemerkung 3.3.3 *Da man nicht immer das Minimalpolynom von A kennt, kann man auch nicht immer entscheiden, ob eine gegebene Funktion $f(t)$ auf dem Spektrum von A definiert ist. Da aber die algebraische Vielfachheit eines Eigenwertes jedenfalls nicht größer als n sein kann, ist klar* dass alle mindestens $(n-1)$-mal differenzierbaren Funktionen auf dem Spektrum von A definiert sind, wenn nur alle Eigenwerte von A im Definitionsbereich D liegen.

Aufgabe 3.3.4 *Zeige, dass* e^A, $\sin A$ *und* $\cos A$ *für alle A definiert sind. Wie ist es mit* $\log A$*?*

Aufgabe 3.3.5 (Matrixfunktionen eines Jordanblockes) *Zeige, dass das Minimalpolynom des Jordanblockes $\lambda I + N$ von der Größe $\mu \times \mu$ gleich $(t-\lambda)^\mu$ ist. Also ist f genau dann auf dem Spektrum von A definiert, wenn f an der Stelle λ mindestens $(\mu - 1)$-mal differenzierbar ist. Zeige weiter, dass dann*

$$f(\lambda I + N) \;=\; \begin{bmatrix} f(\lambda) & f'(\lambda) & \ldots & \frac{f^{(\mu-1)}(\lambda)}{(\mu-1)!} \\ 0 & f(\lambda) & \ldots & \frac{f^{(\mu-2)}(\lambda)}{(\mu-2)!} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & f(\lambda) \end{bmatrix} \;=\; \sum_{j=0}^{\mu-1} \frac{f^{(j)}(\lambda)}{j!}\, N^j\,.$$

Untersuche, wann $f(\lambda I + N)$ nilpotent ist, und bestimme dann den Nilpotenzgrad.

3.4 Eigenschaften von Matrixfunktionen

Wir wollen nun die Proposition 3.1.3 von Polynomen auf allgemeinere Funktionen übertragen; dies ist einfach, soweit es die Regeln (a) und (b) betrifft, die wir hier aber etwas anders formulieren. Außerdem beweisen wir noch, wie Matrixfunktionen ähnlicher Matrizen zusammenhängen, was bei Polynomen der Aufgabe 3.1.5 entspricht, sowie eine Regel für direkte Summen von Matrizen:

Satz 3.4.1

(a) *Seien $A, B, T \in \mathbb{C}^{n\times n}$ so, dass $\det T \neq 0$ und $B = T^{-1}AT$ ist. Wenn f auf dem Spektrum von A definiert ist, dann gilt dasselbe auch für B, und $f(B) = T^{-1} f(A) T$.*

(b) *Seien f_1 und f_2 auf dem Spektrum einer Matrix A definiert. Dann sind auch $f_+ := f_1 + f_2$ und $f_\cdot := f_1 f_2$ auf dem Spektrum von A definiert, und es gilt $f_+(A) = f_1(A) + f_2(A)$, $f_\cdot(A) = f_1(A) f_2(A)$.*

(c) *Seien $A \in \mathbb{C}^{n\times n}$ und $B \in \mathbb{C}^{m\times m}$, und sei C gleich der direkten Summe von A und B. Wenn f auf den Spektren von A und B definiert ist, dann gilt dies auch für das Spektrum von C, und $f(C)$ ist die direkte Summe von $f(A)$ und $f(B)$.*

Beweis: Der Beweis von (a) folgt direkt mit der Definition von Matrixfunktionen und der Aufgabe 3.1.5. Auch die Regel für die Summe von Funktionen ergibt sich ohne Schwierigkeit. Für die Produktregel benutzen wir die sogenannte *Leibnizregel für höhere Ableitungen* eines Produktes, sie folgt per Induktion aus der normalen Produktregel: *Falls sowohl f_1 als auch f_2 mindestens k-mal differenzierbar sind, dann gilt dies auch für $f_\cdot$, und*

$$f_\cdot^{(k)}(t) \;=\; \sum_{j=0}^{k} \binom{k}{j} f_1^{(k-j)}(t)\, f_2^{(j)}(t) \qquad \forall\, k \geq 0\,.$$

Aus dieser Tatsache ergibt sich, dass $f_\cdot$ auf dem Spektrum von A definiert ist, und dass das Produkt der beiden Interpolationspolynome zu f_1, f_2 gleich dem Interpolationspolynom für $f_\cdot$ ist. Mit Proposition 3.1.3 folgt dann die Behauptung (b). Für (c) seien $\psi_A(t)$, $\psi_B(t)$ und $\psi_C(t)$ die Minimalpolynome von A, B und C. Nach Aufgabe 1.7.6 ist $\psi_C(t)$ das kgV von $\psi_A(t)$ und $\psi_B(t)$, und daraus folgt, dass f auf dem Spektrum von C definiert ist. Wenn $p(t)$ auf dem Spektrum von C mit f übereinstimmt, dann tut es dies erst recht auf den Spektren von A und B, und somit folgt $f(C) = p(C)$, $f(A) = p(A)$, $f(B) = p(B)$. Mit Aufgabe 3.1.2 folgt also (c). □

Bemerkung 3.4.2 *Die obigen Regeln, zusammen mit dem Satz von der Jordanschen Normalform und* Aufgabe 3.3.5*, erlauben grundsätzlich die Berechnung von $f(A)$ für allgemeines A. Dazu muss man aber zuerst eine Matrix T bestimmen, welche A auf Jordannormalform bringt, und dies ist im Allgemeinen schwierig, so dass sich dieses Berechnungsverfahren nur in einfachen Fällen anbietet. Es ist aber theoretisch wichtig zu sehen, wie $f(A)$ aussieht, falls man T als gegeben annimmt. Z. B. kann man ablesen, dass die Eigenwerte von $f(A)$ genau die Zahlen $f(\lambda_1), \ldots, f(\lambda_s)$ sind. Allerdings können unter diesen Werten natürlich auch einige, oder sogar alle, gleich sein. Das Minimalpoynom von $f(A)$ erhält man wie folgt: Sind $\mu_1, \ldots, \mu_\sigma$ die verschiedenen unter den Werten $f(\lambda_1), \ldots, f(\lambda_s)$, und wird $\tilde{m}_j$ gleich dem Maximum der m_k für alle solchen k mit $f(\lambda_k) = \mu_j$ gesetzt, so ist $p(t) = (t-\mu_1)^{\tilde{m}_1} \cdot \ldots \cdot (t-\mu_\sigma)^{\tilde{m}_\sigma}$ jedenfalls annullierendes Polynom für $f(A)$, und sogar gleich dem Minimalpolynom, falls $f'(\lambda_j) \neq 0$ ist für alle $j = 1, \ldots, s$. Hieraus folgt, dass jede Funktion g, welche an allen Stellen $f(\lambda_j)$ mindestens m_j-mal differenzierbar ist, auf dem Spektrum von $f(A)$ definiert ist, so dass $g(f(A))$ gebildet werden kann. Vergleiche hierzu auch die nächste Aufgabe.*

Aufgabe 3.4.3 *Benutze den letzten Satz um zu zeigen, dass e^A immer invertierbar ist, und dass e^{-A} die inverse Matrix ist.* Hinweis: Wende den Satz auf die Funktionen $f(t) = \mathrm{e}^t$ und $g(t) = \mathrm{e}^{-t}$ an.

Aufgabe 3.4.4 *Benutze den letzten Satz um zu zeigen: Für jede Matrix A ist* $(\sin A)^2 + (\cos A)^2 = I$.

Bemerkung 3.4.5 *Die Produktregel gilt genauso für rechtsseitige Ableitungen, und daher kann man sehen, dass die Aussagen des letzten Satzes richtig bleiben, falls f und/oder g an einigen Stellen des Spektrums von A nur rechtsseitig stetig bzw. differenzierbar ist - das gleiche gilt natürlich auch für die linke Seite. Daher kann man aus* (b) *folgern, dass die in* Bemerkung 3.3.2 *definierte Matrix $A^{1/2}$ die Eigenschaft $A^{1/2}\,A^{1/2} = A$ besitzt. Es gibt im Allgemeinen aber Matrizen B mit $B\,B = A$, die sich nicht in der Form $f(A)$ mit $f(t) = t^{1/2}$ schreiben lassen; dies ist z. B. so, falls $A = I$ und $n \geq 2$ ist.*

Aufgabe 3.4.6 (Kettenregel für höhere Ableitungen) *Seien f an der Stelle λ und g an der Stelle $f(\lambda)$ beide k-mal differenzierbar. Zeige: Dann ist $g \circ f$ an der Stelle λ ebenfalls k-mal differenzierbar, und es gilt*

$$\frac{d^k}{dt^k}\, g(f(\lambda)) \;=\; \sum_{j=1}^{k} \; g^{(j)}(f(\lambda))\; p_{kj}(f'(\lambda), \ldots, f^{(k)}(\lambda))\,,$$

mit Polynomen p_{kj} in k Variablen. Speziell ist

$$p_{k1}(f'(\lambda), \ldots, f^{(k)}(\lambda)) \;=\; f^{(k)}(\lambda)\,, \qquad p_{kk}(f'(\lambda), \ldots, f^{(k)}(\lambda)) \;=\; (f'(\lambda))^k\,.$$

Finde eine Rekursionsformel zur Berechnung der übrigen p_{kj}.

Der Beweis der noch ausstehenden Regel für die Hintereinanderausführung von Funktionen erfordert nicht nur etwas mehr Aufwand, wir brauchen auch etwas stärkere Voraussetzungen zur Differenzierbarkeit der beiden Funktionen, da aus der k-maligen Differenzierbarkeit der Hintereinanderausführung $f \circ g$ nicht die der beiden Funktionen f und g folgt.

Satz 3.4.7 *Sei f auf dem Spektrum von A definiert, und sei g an allen Stellen $\mu_j = f(\lambda_j)$ mindestens m_j-mal differenzierbar, wobei m_j wie in* (3.3.1) *ist. Dann sind g auf dem Spektrum von $f(A)$, und $h := g \circ f$ auf dem Spektrum von A definiert, und es gilt $h(A) = g(f(A))$.*

Beweis: Aus Aufgabe 3.4.6 folgt zunächst, dass $g \circ f$ auf dem Spektrum von A definiert ist, und in Bemerkung 3.4.2 wurde gezeigt, dass g auf dem Spektrum von $f(A)$ ebenfalls definiert ist. Seien jetzt $p(t)$ und $q(t)$ Polynome, die auf dem Spektrum von A mit f bzw. auf dem Spektrum von $f(A)$ mit g übereinstimmen, so dass also $f(A) = p(A)$ und $q(f(A)) = g(f(A))$ $[\,= q(p(A))]$ ist. Aus Aufgabe 3.4.6 ergibt sich dann, dass $q \circ p$ auf dem Spektrum von A mit $g \circ f$ übereinstimmt, und deshalb folgt die Behauptung mit Hilfe von Proposition 3.1.3, Teil (c). □

Aufgabe 3.4.8 (Inverse einer Matrixfunktion) *Sei $f(t)$ auf dem Spektrum von A definiert, und seien $f(\lambda_j) \neq 0$ für $1 \leq j \leq s$. Zeige: Dann ist der Kehrwert $g(t) = 1/f(t)$ auf dem Spektrum von A definiert, und $g(A)$ ist die inverse Matrix zu $f(A)$.*

Aufgabe 3.4.9 *Zeige: Wenn der Logarithmus auf dem Spektrum von A definiert ist, dann gilt immer*

$$\mathrm{e}^{\log A} \;=\; A\,.$$

3.5 Potenzreihen von Matrizen

Im Vektorraum der $n \times m$-Matizen kann man viele unterschiedliche Normen, sog. *Matrixnormen*, betrachten. Wir definieren hier

$$\|A\| \;:=\; \Big(\; \sum_{j,k} \; |a_{jk}|^2 \;\Big)^2\,,$$

was gleich der euklidischen Norm desjenigen Vektors aus $\mathbb{C}^{nm}$ ist, der entsteht, wenn man alle Spalten von A untereinander schreibt.

Aufgabe 3.5.1 (Submultiplikativität der Matrixnorm) *Zeige mit Hilfe der Cauchy-Schwarzschen Ungleichung: Sind die Matrizen A und B so, dass $C = A\,B$ definiert ist, dann gilt $\|C\| \leq \|A\|\,\|B\|$. Speziell gilt dies, wenn B ein Spaltenvektor passender Länge ist. Benutze dies, um weiter zu zeigen: Ist x ein Eigenvektor einer quadratischen Matrix A, und ist λ der zugehörige Eigenwert, so folgt $|\lambda|\,\|x\| \leq \|A\|\,\|x\|$, also $|\lambda| \leq \|A\|$.*

Die folgende Definition entspricht der aus *Analysis II* für die Konvergenz einer Folge von Vektoren:

Definition 3.5.2 *Wir sagen, dass eine Folge (A_N) von Matrizen aus $\mathbb{C}^{n\times m}$ gegen $A \in \mathbb{C}^{n\times m}$ konvergiert, wenn $\|A_N - A\|$ für $N \longrightarrow \infty$ gegen 0 geht. Wie in Analysis sagt man dann, dass eine Reihe von Matrizen konvergiert, wenn die Folge ihrer Partialsummen konvergent ist.*

Wenn wir die Elemente von A_N bzw. A mit $a_{jk}^{(N)}$ bzw. a_{jk} bezeichnen, so wissen wir aus Analysis II, dass die Folge (A_N) genau dann gegen A konvergiert, wenn für alle festen $j,k \in \{1,\ldots,n\}$ gilt $a_{jk}^{(N)} \longrightarrow a_{jk}$ für $N \longrightarrow \infty$.

Definition 3.5.3 (Konvergenz auf dem Spektrum) *Gegeben sei eine Matrix $A \in \mathbb{C}^{n\times n}$, deren Minimalpolynom in der Form* (3.3.1) *geschrieben sei. Wir sagen, dass eine Funktionenfolge $(f_N(t))$* auf dem Spektrum von A konvergiert*, wenn folgendes gilt:*

(a) *Alle $f_N(t)$ sind auf dem Spektrum von A definiert.*

(b) $\forall\ j = 1,\ldots,s \quad \forall\ k = 0,\ldots,m_j - 1\ :\quad \lim\limits_{N\longrightarrow\infty} \dfrac{d^k}{dt^k}\, f_N(\lambda_j)$ *existiert.*

Wenn man die Eigenwerte einer Matrix nicht kennt, kann man jedenfalls mit Hilfe von Aufgabe 3.5.1 folgende hinreichende Bedingung für die Konvergenz der Folge auf dem Spektrum von A geben:

- Wenn alle $f_N(t)$ für $|t| < r$ mindestens $(n-1)$-mal differenzierbar sind, und wenn die Folgen $(f_N^{(j)}(t))$ für $0 \leq j \leq n-1$ dort punktweise konvergieren, dann konvergiert $(f_N(t))$ auf dem Spektrum jeder Matrix $A \in \mathbb{C}^{n\times n}$ mit $\|A\| < r$.

Wir beweisen jetzt den entscheidenden Satz dieses Abschnitts:

Satz 3.5.4 *Gegeben sei eine Matrix $A \in \mathbb{C}^{n\times n}$ sowie eine Folge $(f_N(t))$ von Funktionen, die alle auf dem Spektrum von A definiert sind. Genau dann ist die Matrixfolge $(f_N(A))$ konvergent, wenn die Funktionenfolge $(f_N(t))$ auf dem Spektrum von A konvergiert.*

Beweis: Aus (3.2.2) folgt die Darstellung

$$f_N(A) \;=\; \sum_{\nu=1}^{s} \sum_{\mu=0}^{m_\nu} f_N^{(\mu)}(\lambda_\nu)\ \psi_{\nu\mu}(A)\ . \tag{3.5.1}$$

Dabei hängen die Matrizen $\psi_{\nu\mu}(A)$ nicht von N ab, so dass sich offenbar aus der Konvergenz der Funktionenfolge auf dem Spektrum die Konvergenz der Matrixfolge ergibt. Für die Umkehrung kann man sich wegen Satz 3.4.1 (a) auf den Fall beschränken, dass A in Jordannormalform, ja sogar ein einziger Jordanblock ist, und dann kann man aus der expliziten Form von $f_N(A)$ die Konvergenz der Funktionenfolge auf dem Spektrum ablesen. □

Korollar zu Satz 3.5.4 *Wenn eine Potenzreihe $\sum_0^\infty f_j\,(t-t_0)^j$ den Konvergenzradius $r>0$ hat, und wenn alle Eigenwerte von A in der offenen Kreisscheibe $|t-t_0|<r$ liegen, dann konvergiert auch die Matrixreihe*

$$\sum_0^\infty f_j\,(A-t_0\,I)^j\,.$$

Bemerkung 3.5.5 *Wenn eine Funktion $f(t)$ in eine Potenzreihe entwickelbar ist, d. h., wenn gilt*

$$f(t) \;=\; \sum_{j=0}^\infty f_j\,(t-t_0)^j \qquad \forall\; t \quad \text{mit} \quad |t-t_0|<r\,,$$

so ist $f(t)$ auf der Kreisscheibe $D=\{|t-t_0|<r\}$ beliebig oft differenzierbar, und alle Ableitungen von f ergeben sich durch gliedweises Differenzieren der Potenzreihe. Das bedeutet für eine Matrix, deren Eigenwerte alle zu D gehören, dass die Partialsummen der Potenzreihe auf dem Spektrum gegen $f(t)$ konvergieren, und deshalb folgt

$$f(A) \;=\; \sum_{j=0}^\infty f_j\,(A-t_0\,I)^j\,.$$

Daher erhält man speziell dass

$$\mathrm{e}^A \;=\; \sum_{j=0}^\infty \frac{A^j}{j!} \qquad \forall\; A\in\mathbb{C}^{n\times n} \tag{3.5.2}$$

Man kann natürlich auch $\sin A$ oder $\cos A$ über die entsprechenden Potenzreihen darstellen, aber das soll im Moment nicht geschehen.

Mit Hilfe von (3.5.2) können wir weitere Eigenschaften der Matrixexponentialfunktion herleiten:

Satz 3.5.6 (Eigenschaften der Matrixexponentialfunktion) *Für zwei Matrizen $A,B\in\mathbb{C}^{n\times n}$, welche miteinander kommutieren, folgt*

$$\mathrm{e}^{A+B} \;=\; \mathrm{e}^A\,\mathrm{e}^B\,. \tag{3.5.3}$$

Wenn $A(t)$ eine quadratische Matrix ist, deren Elemente differenzierbare Funktionen auf einem Intervall $D\subset\mathbb{R}$ sind, wenn $A'(t)$ auf D beschränkt ist, und wenn $A(t)\,A'(t)=A'(t)\,A(t)$ für alle $t\in D$ gilt, folgt

$$\frac{d}{dt}\;\mathrm{e}^{A(t)} \;=\; A'(t)\,\mathrm{e}^{A(t)} \;=\; \mathrm{e}^{A(t)}\,A'(t) \qquad \forall\; t\in D\,. \tag{3.5.4}$$

Beweis: Mit Induktion zeigt man dass für kommutierende Matrizen A und B die binomische Formel gilt, d. h. also

$$(A\;+\;B)^n \;=\; \sum_{k=0}^j \binom{k}{j}\,A^{k-j}\,B^j \qquad \forall\; j\in\mathbb{N}_0\,.$$

Daraus folgt, genau wie in *Analysis I*, die Gleichung (3.5.3). Weiter zeigt man, ebenfalls mit Induktion, dass unter den im Satz gemachten Voraussetzungen gilt

$$\frac{d}{dt}\;(A(t))^j \;=\; j\,A'(t)\,(A(t))^{j-1} \;=\; j\,(A(t))^{j-1}\,A'(t) \qquad \forall\; j\in\mathbb{N}_0\,,$$

woraus folgt dass

$$\frac{d}{dt}\;\sum_{j=0}^N \frac{(A(t))^j}{j!} \;=\; A'(t)\;\sum_{j=0}^{N-1} \frac{(A(t))^j}{j!} \;=\; \Big(\;\sum_{j=0}^{N-1} \frac{(A(t))^j}{j!}\Big)\;A'(t) \qquad \forall\; N\in\mathbb{N}_0\,.$$

Da $A'(t)$ auf D beschränkt ist, gilt das gleiche wegen des Mittelwertsatzes auch für $A(t)$. Aus der Abschätzung

$$\Big\| A'(t) \sum_{j=M}^{N-1} \frac{(A(t))^j}{j!} \Big\| \;=\; \Big\| \Big(\sum_{j=M}^{N-1} \frac{(A(t))^j}{j!} \Big)\, A'(t) \Big\| \;\leq\; \|A'(t)\| \sum_{j=M}^{N-1} \frac{\|A(t)\|^j}{j!}$$

folgt die gleichmäßige Konvergenz der gliedweise differenzierten Reihe auf jedem beschränkten Teil von D, und mit dem entsprechenden Satz aus *Analysis I* folgt für $N \longrightarrow \infty$ die Beziehung (3.5.4). □

Bemerkung 3.5.7 (Lineare Differentialgleichungssysteme) *Als Anwendung betrachten wir ein sogenanntes lineares Differentialgleichungssystem. Dazu sei eine auf einem abgeschlossenen Intervall $D = [a, b]$ stetige Matrix $A(t)$ gegeben. Die Gleichung*

$$x' \;=\; A(t)\, x$$

heißt dann ein lineares Differentialgleichungssystem *zur Bestimmung von Vektoren $x(t)$, welche auf D differenzierbar sind und dieser Gleichung genügen. Eine Matrix $X(t)$, die auf D differenzierbar ist, und für die $X'(t) = A(t)\, X(t)$ für alle $t \in D$ gilt, heißt* Fundamentalsystem*, wenn ihre Determinante auf D keine Nullstelle hat. Man sieht, dass jeder Vektor $x(t)$, der eine Linearkombination der Spalten von $X(t)$ ist, eine Lösung des Differentialgleichungssystems ist, und man kann beweisen, dass es keine weiteren Lösungen gibt. Es folgt jetzt mit* (3.5.4)*:*

- *Ist $B(t)$ eine Stammfunktion zu $A(t)$, also $B'(t) = A(t)$ für $t \in D$, und kommutieren $A(t)$ und $B(t)$, was z. B. für eine konstante Matrix A und $B(t) = t\, A$ immer erfüllt ist, so ist $X(t) = \mathrm{e}^{B(t)}$ ein Fundamentalsystem. Speziell für den Fall einer konstanten Matrix A ist also $X(t) = \mathrm{e}^{tA}$ immer ein Fundamentalsystem von $x' = A\, x$.*

3.6 Frobeniussche Kovarianten und Berechnung von Matrixfunktionen

Im Folgenden sei $A \in \mathbb{C}^{n\times n}$ gegeben. Wie bisher sei $\psi_A(t)$ das Minimalpolynom von A, geschrieben in der Form (3.3.1). Weiter sei $\psi(t)$ ein beliebiges Polynom der Form

$$\psi(t) \;=\; (t-\lambda_1)^{\mu_1} \cdot \ldots \cdot (t-\lambda_s)^{\mu_s}\,, \qquad \mu_\nu \geq m_\nu\,, \quad 1 \leq \nu \leq s\,. \tag{3.6.1}$$

In anderen Worten ist $\psi(t)$ ein beliebiges Polynom, welches durch das Minimalpolynom teilbar ist, so dass also $\psi(A) = 0$ ist. Zum Beispiel kann $\psi(t)$, bis auf den Faktor $(-1)^n$, das charakteristische Polynom von A sein, was den Vorteil hat, dass die Vielfachheiten μ_ν einfacher zu berechnen sind. Allerdings *kann* $\psi(t)$ auch das Minimalpolynom sein, und es ist wichtig zu beachten, dass die noch zu definierenden Froebeniusschen Kovarianten nicht von der Wahl des Polynoms $\psi(t)$ abhängen! Siehe dazu auch die nächste Aufgabe.

Definition 3.6.1 (Frobeniussche Kovarianten) *Wir definieren Polynome*

$$q_\nu(t) \;=\; \frac{\psi(t)}{(t-\lambda_\nu)^{\mu_\nu}} \;=\; \prod_{1\leq j\leq s,\; j\neq\nu} (t-\lambda_j)^{\mu_j}\,, \quad 1 \leq \nu \leq s\,, \qquad q(t) = \sum_{\nu=1}^{s} q_\nu(t)\,, \tag{3.6.2}$$

und mit diesen berechnen wir die Matrizen

$$B_\nu \;=\; q_\nu(A)\,, \quad 1 \leq \nu \leq s\,, \qquad B \;=\; q(A) \;=\; \sum_{\nu=1}^{s} B_\nu\,. \tag{3.6.3}$$

Da die Eigenwerte von $B = q(A)$ genau die Zahlen $q(\lambda_j) = q_j(\lambda_j)$ und deshalb alle $\neq 0$ sind, ist B invertierbar, und wir können deshalb die sogenannten Frobeniusschen Kovarianten *definieren durch*

$$C_\nu \;:=\; B_\nu\, B^{-1} \qquad 1 \le \nu \le s\,. \tag{3.6.4}$$

Aufgabe 3.6.2 *Sei $A = T^{-1}\, J\, T$, wobei J die direkte Summe von Matrizen $J_1, \ldots, J_s$ der Form $J_k = \lambda_k\, I + N_k$ ist, mit einer nilpotenten Matrix N_k, so dass $N_k^{\mu_k} = 0$ ist. Beachte, dass die Existenz von T und J aus dem Satz über die Jordannormalform folgt. Zeige:*

(a) *$B_\nu = T^{-1}\, q_\nu(J)\, T$ und $B = T^{-1}\, q(J)\, T$.*

(b) *$q_\nu(J)$ bzw. $q_\nu(J)$ ist die direkte Summe von $q_\nu(J_1), \ldots, q_\nu(J_s)$ bzw. $q(J_1), \ldots, (J_s)$.*

(c) *$q_\nu(J_\mu) = 0$ falls $\nu \neq \mu$ ist, und $q_\nu(J_\nu)$ ist invertierbar.*

(d) *$C_\nu = T^{-1}\, D_\nu\, T$, mit einer Diagonalmatrix D_ν, welche die direkte Summe aus Nullmatrizen und einer Einheitsmatrix ist, wobei die Einheitsmatrix gerade an der ν-ten Stelle auf der Blockdiagonalen steht, und die Blockgrößen denen in der Matrix J entsprechen.*

Schließe hieraus, dass die Matrizen C_ν nicht davon abhängen, ob $\psi(t)$ das charakteristische oder das Minimalpolynom von A, oder irgendein anderes annullierendes Polynom für A ist.

Proposition 3.6.3 (Eigenschaften der Frobeniusschen Kovarianten) *Für die oben definierten C_ν gilt immer:*

(a) $\displaystyle\sum_{\nu=1}^{s} C_\nu = I,$

(b) $C_\nu\, C_\mu = 0$ *falls* $\nu \neq \mu$ *ist,*

(c) $C_\nu^2 = C_\nu$ *für alle* $\nu = 1, \ldots, s,$

(d) $(A - \lambda_\nu\, I)^\mu\, C_\nu = 0$ *für alle* $\mu \ge m_\nu$.

Beweis: Die Gleichung (a) folgt direkt aus der Definition von B, und (b) gilt, da $q_\nu(t)\, q_\mu(t)$ durch $\psi_A(t)$ teilbar ist. Die Gleichung (c) folgt aus (a) und (b), da $C_\nu = C_\nu\, I = C_\nu\, (C_1 + \ldots + C_s) = C_\nu^2$ ist. Es ist $C_\nu = q_\nu(A)\, B^{-1}$, und wenn wir für $\psi(t)$ das Minimalpolynom von A wählen, dann ist $(t - \lambda_\nu)^\mu\, q_\nu(t) = (t - \lambda_\nu)^{\mu - m_\nu}\, \psi_A(t)$, woraus (d) folgt. □

Wir zeigen jetzt, dass zur Berechnung einer Matrixfunktion, außer den Werten der Funktion und den entsprechenden Ableitungen auf dem Spektrum, nur die Frobeniusschen Kovarianten sowie einige Potenzen von $A - \lambda_\nu\, I$ benötigt werden:

Satz 3.6.4 *Sei $A \in \mathbb{C}^{n \times n}$, und sei $f(t)$ auf dem Spektrum von A definiert. Seien weiter die m_ν wie in* (3.3.1)*, also die Vielfachheiten der Nullstellen des Minimalpolynoms. Dann gilt*

$$f(A) \;=\; \sum_{\nu=1}^{s} \sum_{\mu=0}^{m_\nu - 1} \frac{f^{(\mu)}(\lambda_\nu)}{\mu!}\; (A - \lambda_\nu\, I)^\mu\, C_\nu\,. \tag{3.6.5}$$

Beweis: Wenn man mit $p(t)$ das zu $f(t)$ und A gehörige Interpolationspolynom bezeichnet, so sieht man dass es ausreicht, die Gleichung (3.6.5) für $p(t)$ anstelle von $f(t)$ zu beweisen. Auf Grund von Teil

(d) der letzten Proposition können wir dann die innere Summe auf der rechten Seite von (3.6.5) bis ∞ erstrecken, da die hinzukommenden Terme alle verschwinden. Aus der Taylorschen Formel folgt aber

$$p(t) \;=\; \sum_{\mu=0}^{m-1} \frac{p^{(\mu)}(\lambda_\nu)}{\mu!}\,(t-\lambda_\nu)^\mu\,,$$

und deshalb ergibt sich unter Benutzung von Gleichung (a) aus der letzten Proposition

$$\sum_{\nu=1}^{s}\sum_{\mu=0}^{m_\nu-1} \frac{p^{(\mu)}(\lambda_\nu)}{\mu!}\,(A-\lambda_\nu\,I)^\mu\,C_\nu \;=\; \sum_{\nu=1}^{s} p(A)\,C_\nu \;=\; p(A)\,.$$

Damit ist der Satz bewiesen. □

Aus (3.6.5) ergibt sich, dass die Berechnung einer Matrixfunktion gelingt, wenn man die Eigenwerte der Matrix und ihre Frobeniusschen Kovarianten berechnet hat. Diese Darstellung hat den Vorteil dass die Berechnung der Kovarianten einfach ist, sofern man die Eigenwerte und ihre Vielfachheiten kennt.

Kapitel 4

Konvexe Polyeder

Im nächsten Kapitel werden einige Hilfmittel gebraucht, die wir hier bereitstellen wollen. Dazu untersuchen wir Begriffe und Resultate aus der sogenannten *konvexen Analysis* in $\mathbb{R}^n$. Da aber viele der Definitionen auch allgemeiner sinnvoll sind, betrachten wir zunächst einen beliebigen Vektorraum V über dem Körper $\mathbb{R}$.

4.1 Einige Bezeichnungen

Definition 4.1.1 (Konvexkombinationen etc.) *Für $v_1, \ldots, v_\nu \in V$ und $\alpha_1, \ldots, \alpha_\nu \in \mathbb{R}$ heißt die Linearkombination $\sum_{k=1}^{\nu} \alpha_\nu v_\nu$ auch*

- affine Kombination *von $v_1, \ldots, v_\nu$, falls $\sum_{k=1}^{\nu} \alpha_k = 1$ ist.*
- konische Kombination *von $v_1, \ldots, v_\nu$, falls alle $\alpha_k \geq 0$ sind.*
- Konvexkombination *von $v_1, \ldots, v_\nu$, falls alle $\alpha_k \geq 0$ sind, und zusätzlich $\sum_{k=1}^{\nu} \alpha_k = 1$ ist.*

Für ein $M \subset V$ bezeichne $\mathrm{lin}(M)$ *die Menge aller Linearkombinationen,* $\mathrm{aff}(M)$ *die Menge aller affinen Kombinationen,* $\mathrm{cone}(M)$ *die Menge aller konischen Kombinationen, bzw.* $\mathrm{conv}(M)$ *die Menge aller Konvexkombinationen, von Elementen aus M. Wir nennen diese Mengen auch die* lineare, *bzw.* affine, *bzw.* konische, *bzw.* konvexe Hülle von M. *Entsprechend der Konvention über leere Summen bestehen die lineare und die konische Hülle genau aus dem Nullvektor, falls die Menge M leer ist. Insbesondere sind diese Hüllen niemals gleich der leeren Menge. Anders ist es bei der affinen und der konvexen Hülle, denn bei der leeren Summe ist die Bedingung, dass sich die Koeffizienten zu Eins addieren, immer verletzt. Deshalb setzen wir* $\mathrm{aff}(\emptyset) = \mathrm{conv}(\emptyset) = \emptyset$. *Die konvexe Hülle einer* endlichen *Menge M heißt auch* Polytop. *Falls $M = \mathrm{lin}(M)$, bzw. $M = \mathrm{aff}(M)$, bzw. $M = \mathrm{cone}(M)$, bzw. $M = \mathrm{conv}(M)$ ist, dann heißt M* Unterraum, *bzw.* affiner Raum, *bzw.* Kegel, *bzw.* konvex. *Wir sagen auch, dass eine konvexe Menge M* endlich erzeugt *ist, falls M ein Polytop ist.*

Klar ist, dass die hier gegebene Unterraumdefinition gleich der aus dem ersten Teil der Vorlesung ist. Für die weiteren Begriffe vergleiche auch die folgende Aufgabe.

Aufgabe 4.1.2 *Zeige für ein beliebiges $M \subset V$:*

(a) *Die Menge aller affinen Kombinationen von $v_0, \ldots, v_\nu \in V$ ist genau dann ein Unterraum von V, wenn das System $(v_0, \ldots, v_\nu)$ linear abhängig ist, und sonst ist sie ein um einen Vektor, z. B. um*

v_0, *verschobener Unterraum.* Anleitung: *Zeige zunächst, dass jede affine Kombination in der Form* $v_0 + \sum_1^\nu \alpha_j\,(v_j - v_0)$ *geschrieben werden kann.*

(b) *Die Menge* $\mathrm{aff}(M)$ *ist immer ein affiner Raum. Genauer: Ist* $v_0 \in \mathrm{aff}(M)$*, so ist* $U := \{u \in V : \exists\, v \in \mathrm{aff}(M) \text{ mit } u = v - v_0\}$ *ein Unterraum von* V*. Hieraus folgt dann* $\mathrm{aff}(M) = v_0 + U$*.*

(c) *Die Menge* M *ist genau dann konvex, wenn für* $v, w \in M$ *und* $0 \le \alpha \le 1$ *auch* $\alpha\, v + (1-\alpha)\, w \in M$ *ist. Das bedeutet anschaulich, dass die* Verbindungsstrecke *von* v *nach* w *ganz zu* M *gehört.*

(d) *Die Mengen* $\mathrm{cone}(M)$ *und* $\mathrm{aff}(M)$ *sind konvex.*

(e) *Es gilt* $M \subset \mathrm{lin}(M) = \mathrm{lin}(\mathrm{lin}(M))$*,* $M \subset \mathrm{aff}(M) = \mathrm{aff}(\mathrm{aff}(M))$*,* $M \subset \mathrm{cone}(M) = \mathrm{cone}(\mathrm{cone}(M))$*,* $M \subset \mathrm{conv}(M) = \mathrm{conv}(\mathrm{conv}(M))$*.*

(f) *Die Menge* M *ist genau dann ein Kegel, wenn für alle* $x, y \in M$ *und alle* $\alpha, \beta \in \mathbb{R}_+ := \{x \in \mathbb{R} : x \ge 0\}$ *der Vektor* $\alpha\, x + \beta\, y$ *wieder in* M *liegt. Insbesondere ist jeder Kegel konvex und enthält den Nullvektor.*

Veranschauliche diese Begriffe anhand von Beispielen in $\mathbb{R}^2$ *bzw.* $\mathbb{R}^3$*. Zeige weiter, dass die Menge* $\mathbb{R}^n_+ := \{x = (x_1, \ldots, x_n)^T :\ x_j \ge 0 \quad \forall\ j = 1, \ldots, n\}$ *gleich der konischen Hülle der kanonischen Basivektoren und deshalb ein Kegel ist. Gib ein Beispiel dafür, dass ein Kegel nicht abgeschlossen sein muss.*

Definition 4.1.3 *Da nach der letzten Übungsaufgabe ein affiner Raum* A *durch eine Verschiebung eines Unterraumes* U *entsteht, setzen wir* $\dim A := \dim U$*. Ein eindimensionaler affiner Raum heißt* Gerade*, ein zweidimensionaler heißt* Ebene*, einer der Dimension* $n-1$ *heißt* Hyperebene*.*

Aufgabe 4.1.4 *Wiederhole die Ergebnisse über lineare Gleichungssysteme um zu sehen, dass die Lösungsmenge eines* lösbaren *inhomogenen System immer ein affiner Raum ist. Zeige umgekehrt, dass jeder affine Raum in* $\mathbb{R}^n$ *auch Lösungsmenge eines geeigneten linearen Gleichungssystems ist, dass dieses aber durch den Raum nicht eindeutig festgelegt ist.*

Satz 4.1.5 *Für jede nicht-leere Menge* $M \subset \mathbb{R}^n$ *gelten folgende Aussagen:*

(a) *Zu jedem* $x \in \mathrm{conv}(M)$ *gibt es* $n+1$ *Elemente* $x_0, \ldots, x_n \in M$ *so, dass* x *eine Konvexkombination dieser Punkte ist.*

(b) *Wenn* M *kompakt ist, dann ist auch* $\mathrm{conv}(M)$ *kompakt.*

(c) *Wenn* M *abgeschlossen und konvex ist, und wenn* $x_0 \notin M$ *ist, dann gibt es ein* $c \in \mathbb{R}^n \setminus \{0\}$ *so, dass* $c^T\,(x - x_0) > 0$ *ist für alle* $x \in M$*.*

Beweis: Zunächst gibt es jedenfalls $x_0, \ldots, x_m \in M$ und $\alpha_0, \ldots, \alpha_m \in \mathbb{R}_+$ mit $\sum_0^m \alpha_j = 1$ so, dass $x = \sum_0^m \alpha_j\, x_j$ ist. Falls $m < n$ ist, wählen wir weitere Punkte $x_{m+1}, \ldots, x_n \in M$ und setzen $\alpha_{m+1} = \ldots = \alpha_n = 0$. Falls $m > n$ ist, können wir zunächst o. B. d. A. annehmen dass alle $\alpha_j > 0$ sind, denn sonst gilt eine entsprechende Darstellung mit weniger Summanden. In jeden Fall sind die Vektoren $x_j - x_0$, $1 \le j \le m$, linear abhängig, und daher gibt es $\beta_j \in \mathbb{R}$, die nicht alle verschwinden, so dass $\sum_1^m \beta_j\,(x_j - x_0) = 0$ ist. Wenn wir β_0 entsprechend wählen, gilt $\sum_0^m \beta_j = 0$, und dann folgt $\sum_0^m \beta_j\, x_j = 0$, also $x = \sum_0^m (\alpha_j - \lambda\,\beta_j)\, x_j$ für beliebiges $\lambda \in \mathbb{R}$. Es folgt dass $\sum_0^m (\alpha_j - \lambda\,\beta_j) = 1$ ist, und für hinreichend kleines λ sind alle $\alpha_j - \lambda\,\beta_j$ nicht negativ. Falls man evtl. λ vergrößert, kann man erreichen dass mindestens eine der Zahlen verschwindet, und dann ist x Konvexkombination von weniger als m Elementen. Durch Iteration dieses Schrittes folgt dann (a). Um (b) zu beweisen, definieren wir

$$K \;=\; \{(\alpha_0, \ldots, \alpha_n)^T \in \mathbb{R}^{n+1}_+ :\ \sum_j \alpha_j = 1\}\,.$$

Diese Menge ist abgeschlossen und beschränkt und deshalb kompakt, und daher ist auch $K \times M^{n+1}$ kompakt. Da die Funktion

$$f(\alpha_0, \ldots, \alpha_n, x_0, \ldots, x_n) \;=\; \sum_{j=0}^{n} \alpha_j \, x_j$$

auf $K \times M^{n+1}$ stetig ist, ist ihre Bildmenge nach Analysis ebenfalls kompakt, und wegen (a) ist diese Bildmenge genau die konvexe Hülle von M. Für den Beweis von (c) sei $r > 0$ so groß dass $K := M \cap \{\|x - x_0\| \leq r\}$ nicht leer ist. Dann ist K abgeschlossen und beschränkt und somit kompakt, und $x_0 \notin K$. Die stetige Funktion $f(x) = \|x - x_0\|$ nimmt auf K ein Minimum an, und dieses kann nicht $= 0$ sein. Also gibt es ein $x_1 \in K$ mit $\|x_1 - x_0\| \leq \|x - x_0\|$ für alle $x \in K$, und dasselbe gilt erst recht für alle $x \in M \setminus K$, da ja $\|x_1 - x_0\| \leq r$ ist. Da M konvex ist, gilt für alle $\alpha \in (0,1)$ und alle $x \in M$ (nach Wahl von x_1)

$$\|x_1 - x_0\|^2 \;\leq \|(1-\alpha)\, x_1 + \alpha\, x - x_0\|^2 \;=\; \|x_1 - x_0\|^2 \;+\; \alpha^2 \, \|x - x_1\|^2 \;+\; 2\,\alpha\, (x_1 - x_0)^T \, (x - x_1)\,.$$

Wenn man $c = x_1 - x_0$ setzt, folgt hieraus dass

$$0 \;\leq\; \alpha \, \|x - x_1\|^2 \;+\; 2\, c^T \, (x - x_1)\,,$$

was für $\alpha \longrightarrow 0$ impliziert dass $c^T \, (x - x_0) \geq c^T \, (x_1 - x_0) = \|x_1 - x_0\|^2 > 0$ ist, und da $x \in M$ beliebig war, folgt die Behauptung. □

Bemerkung 4.1.6 *Aussage* (c) *des letzten Satzes impliziert dass für genügend kleines $\delta > 0$ die Hyperebene $c^T \, (x - x_0) = \delta$ den Punkt x_0 von der konvexen Menge M* trennt, *in dem Sinne dass beide auf verschiedenen Seiten der Hyperebene liegen – siehe dazu auch die unten stehende Definition von Halbräumen. Man nennt deshalb diese Aussage auch den* Trennungssatz.

4.2 Lineare Systeme von Ungleichungen

Seien im Folgenden n und m zwei natürliche Zahlen. Für Matrizen $A \in \mathbb{R}^{m \times n}$ bzw. Vektoren $b \in \mathbb{R}^m$ schreiben wir immer a_{jk} bzw. b_k für die Elemente von A bzw. Koordinaten von b und verfahren sinngemäß, wenn eine Matrix oder ein Vektor mit einem anderen Buchstaben bezeichnet ist. Außerdem bezeichnen wir die Spalten von A^T mit $a_1, \ldots, a_m$, sodass also immer gilt

$$A\,x \;=\; b \qquad \Longleftrightarrow \qquad b_j \;=\; a_j^T \, x \qquad \forall\; j = 1, \ldots, m\,. \tag{4.2.1}$$

Definition 4.2.1 *Für $A, B \in \mathbb{R}^{m \times n}$ definieren wir*

$$A \;\leq\; B \qquad \Longleftrightarrow \qquad \forall\; j = 1, \ldots, m\,, \quad k = 1, \ldots, n \;:\qquad a_{jk} \;\leq\; b_{jk}\,.$$

Wir schreiben dann auch $A \geq B$ falls $B \leq A$ ist.[1] *Damit ist natürlich auch definiert, was eine Ungleichung zwischen Zeilen- oder Spaltenvektoren bedeutet. Wir sagen deshalb auch kurz, dass wir* Ungleichungen zwischen Vektoren und Matrizen koordinatenweise verstehen. *Dies bedeutet abstrakt formuliert, dass wir in $\mathbb{R}^{m \times n}$ eine* partielle Ordnung *einführen, d. h., eine Relation "$\leq$" mit den Eigenschaften*

- $\forall\; A \in \mathbb{R}^{m \times n} : \quad A \leq A$,
- $\forall\; A, B \in \mathbb{R}^{m \times n} : \quad A \leq B$ und $B \leq A \Longrightarrow A = B$,
- $\forall\; A, B, C \in \mathbb{R}^{m \times n} : \quad A \leq B$ und $B \leq C \Longrightarrow A \leq C$,

[1] Die Schreibweisen $A < B$ bzw. $A > B$ könnten einmal bedeuten dass alle Elemente von A echt kleiner bzw. größer als die von B sind, oder es könnte heißen dass $A \leq B$ bzw. $A \geq B$ gilt und außerdem $A \neq B$ ist. Da dies nur für $n = m = 1$, also für Zahlen, dasselbe ist, wollen wir diese Schreibweisen bei Matrizen vermeiden.

während für $A, B \in \mathbb{R}^{m\times n}$ weder $A \leq B$ noch $B \leq A$ zu gelten braucht. Für $A \in \mathbb{R}^{m\times n}$ und $b \in \mathbb{R}^m$ heißt die Menge $P(A,b)$ aller Lösungen $x \in \mathbb{R}^n$ der Ungleichung, besser: des linearen Ungleichungssystems

$$A\,x \;\leq\; b \tag{4.2.2}$$

ein (konvexes) Polyeder. *Ausgeschrieben bedeutet dies $a_j^T\, x \leq b_j$ für $j = 1, \ldots, m$. Falls $m = 1$ ist, also falls wir nur eine Ungleichung $a^T\, x \leq b$ mit $a \in \mathbb{R}^n$ und $b \in \mathbb{R}$ haben, treten folgende Fälle auf:*

- *Wenn $a = 0$ ist, ist die Lösungsmenge gleich $\mathbb{R}^n$ bzw. $\emptyset$ falls $b \geq 0$ bzw. $b < 0$ ist.*
- *Wenn $a \neq 0$ ist, ist die Lösungsmenge weder leer noch der gesamte Raum $\mathbb{R}^n$ und wird auch* Halbraum *genannt.*

Woher der Name Halbraum kommt, ist klar: Für ein beliebiges $x \in \mathbb{R}^n$ ist $a^T\, x - b$ entweder positiv oder negativ oder $= 0$, und der letzte Fall tritt nur "für wenige x" ein, wenn $a \neq 0$ ist. Statt "der Halbraum aller x mit $a^T\, x \leq b$" schreiben wir auch kürzer "der Halbraum $a^T\, x \leq b$". Die Lösungsmenge der Gleichung $a^T\, x = b$ mit $a \neq 0$ ist eine Hyperebene, und wir schreiben auch analog zu oben "die Hyperebene $a^T\, x = b$" bzw. "das Polyeder $A\,x \leq b$". Offenbar ist die Hyperebene $a^T\, x = b$ gleich dem Durchschnitt der Halbräume $a^T\, x \leq b$ und $-a^T\, x \leq -b$.

Aufgabe 4.2.2 *Seien $A \in \mathbb{R}^{m\times n}$ und $b \in \mathbb{R}^m$ gegeben. Seien, wie bereits oben vereinbart, $a_1, \ldots, a_m \in \mathbb{R}^n$ so, dass a_j^T die j-te Zeile von A ist. Zeige:*

(a) *Die Lösungsmenge eines Systems linearer Ungleichungen* (4.2.2) *bleibt unverändert, wenn wir bei der erweiterten Matrix $[A, b]$*

- *zwei Zeilen vertauschen,*
- *eine Zeile mit einer positiven reellen Zahl multiplizieren.*

Beachte aber, dass Multiplikation einer Zeile mit einer negativen Zahl die Lösungsmenge im Allgemeinen ändert, und dasselbe gilt, wenn man eine Zeile der erweiterten Matrix zu einer anderen addiert, oder von dieser subtrahiert.

(b) *Wenn $x \in P(A,b)$ und $a = \sum_j \alpha_j\, a_j$ ist, mit $\alpha_j \in \mathbb{R}_+$, dann folgt $a^T\, x \leq \sum_j \alpha_j\, b_j$.*

(c) *Wenn $-a_j = \sum_{k\neq j} \alpha_k\, a_k$ ist, mit $\alpha_k \in \mathbb{R}_+$, so folgt dass $-\sum_{k\neq j} \alpha_k\, b_k \leq a_j^T\, x \leq b_j$ für alle $x \in P(A,B)$ gilt. Insbesondere ist $P(A,b)$ leer falls $-\sum_{k\neq j} \alpha_k\, b_k > b_j$ ist.*

(d) *Wenn $a_j = \sum_{k\neq j} \alpha_k\, a_k$ ist, mit $\alpha_k \in \mathbb{R}_+$, und wenn $\sum_{k\neq j} \alpha_k\, b_k \leq b_j$ ist, dann kann man die j-te Zeile aus der erweiterten Matrix entfernen, ohne die Lösungsmenge von* (4.2.2) *zu ändern.*

Um die Begriffe Halbraum und Polyeder zu veranschaulichen, wählen wir ein $a \in \mathbb{R}^n \setminus \{0\}$ sowie eine Zahl $b \in \mathbb{R}$ und beachten, dass die Hyperebene $a^T\, x = b$ eine $(n-1)$-dimensionale lineare Mannigfaltigkeit in $\mathbb{R}^n$ ist. Für $b = 0$ ist diese Hyperebene ein Unterraum und besteht aus allen Vektoren, die auf a senkrecht stehen. Für $b > 0$ enthält die Hyperebene den Vektor $x_0 = b\,\|a\|^{-2}a$, entsteht also aus der vorigen Hyperebene durch Verschiebung in Richtung von a um die Strecke der Länge $\|x_0\| = b\,\|a\|^{-1}$. Für $b < 0$ wird die Hyperebene in die entgegengesetzte Richtung um die gleiche Strecke verschoben. Daraus ergibt sich jetzt folgende Information über die Lösungsmenge einer linearen Ungleichung:

- Der Halbraum $a^T\, x \leq b$ ist zusammenhängend, abgeschlossen und konvex, und sein Rand ist die Hyperebene $a^T\, x = b$. Wir stellen fest, dass der Nullpunkt genau dann in diesem Halbraum liegt, wenn $b \geq 0$ ist.
- Für $A \in \mathbb{R}^{m\times n}$ und $b \in \mathbb{R}^m$ ist das Polyeder $A\,x \leq b$ ein Durchschnitt von m Halbräumen. Dieser Durchschnitt ist zusammenhängend, abgeschlossen und konvex. Er kann leer, beschränkt oder unbeschränkt sein.

Aufgabe 4.2.3 *Zeige dass ein Polyeder leer sein kann, und dass jeder affine Raum in $\mathbb{R}^n$ auch ein Polyeder ist. Zeige weiter dass auch die Lösungsmenge von $A\,x \geq b$ ein Polyeder ist.*

Aufgabe 4.2.4 *Nach der vorausgegangenen Aufgabe ist die Lösungsmenge eines linearen Gleichungssystems immer ein Polyeder. Zeige, dass umgekehrt nicht jedes Polyeder als Lösungsmenge eines Gleichungssystems auftreten kann. Zeige weiter dass der Durchschnitt endlich vieler Polyeder wieder ein Polyeder ist.*

4.3 Alternativsätze

Im Folgenden wollen wir einige wichtige Resultate aus der Theorie der linearen Ungleichungen bzw. der *konvexen Analysis* kennen lernen. Beachte dabei die Analogie, aber auch die Unterschiede zu den entsprechenden Ergebnissen für lineare Gleichungssysteme!

Aufgabe 4.3.1 (Lösbarkeit linearer Gleichungssysteme) *Zeige dass das lineare Gleichungssystem $A\,x \;=\; b$ genau dann eine Lösung besitzt, wenn jedes $y \in \mathbb{R}^m$ mit $y^T\,A = 0$ zum Vektor b orthogonal ist. Zeige weiter, dass dieses Ergebnis auch wie folgt ausgedrückt werden kann:*
Es gilt immer genau eine der folgenden beiden Aussagen:

(a) Das lineare Gleichungssystem $A\,x \;=\; b$ ist lösbar.

(b) Es gibt ein $y \in \mathbb{R}^m$ mit $y^T\,A = 0$ und $y^T\,b = 1$.

Ein Ergebnis dieser Form heißt aus naheliegendem Grund auch ein Alternativsatz. *Vergleiche mit dem unten stehenden Lemma von Farkas.*

Aufgabe 4.3.2 *Zeige: Wenn gilt $y^T\,b \geq 0$ für alle $y \in \mathbb{R}^m_+$, dann folgt $b \geq 0$.*

Wir zeigen jetzt folgendes Resultat über Kegel, welches im Beweis des sogenannten *Lemma von Farkas* eine zentrale Rolle spielt:

Satz 4.3.3 (Fundamentalsatz für lineare Ungleichungen) *Seien $v, v_1, \ldots, v_\nu \in \mathbb{R}^m$. Wenn v nicht im von $v_1, \ldots, v_\nu$ aufgespannten Kegel liegt, d. h., wenn $v \notin \mathrm{cone}\{v_1, \ldots, v_n\}$ ist, dann gibt es ein $y \in \mathbb{R}^m$ mit*

$$y^T\,v_k \;\geq\; 0 \quad \forall\; k = 1, \ldots, \nu\,, \qquad y^T\,v \;<\; 0\,.$$

Beweis: Wenn v nicht in der linearen Hülle der $v_1, \ldots, v_\nu$ liegt, folgt die Behauptung aus Aufgabe 4.3.1. Den anderen Fall beweisen wir durch Induktion über ν: Für $\nu = 1$ können wir auf Grund der Vorbemerkung annehmen, dass $v = \alpha\,v_1$ ist, wobei aber nach Voraussetzung $\alpha < 0$ sein muss. In diesem Fall ergibt sich die Behauptung mit $y = -v$. Sei jetzt $\nu \geq 2$ und der Satz für $\nu - 1$ bereits bewiesen. Da v nicht in dem von $v_1, \ldots, v_{\nu-1}$ aufgespannten Kegel liegen kann, folgt also die Existenz von $\hat{y} \in \mathbb{R}^n$ mit $\hat{y}^T\,v_k \geq 0$ für $k = 1, \ldots, \nu - 1$ und $\hat{y}^T\,v < 0$. Wenn auch $\hat{y}^T\,v_\nu \geq 0$ ist, ist nichts mehr zu zeigen, und daher sei $\hat{y}^T\,v_\nu < 0$ angenommen. Wir setzen

$$\tilde{v}_k := (\hat{y}^T\,v_k)\,v_\nu \;-\; (\hat{y}^T\,v_\nu)\,v_k\,, \quad 1 \leq k \leq \nu - 1\,, \qquad \tilde{v} \;=\; (\hat{y}^T\,v)\,v_\nu \;-\; (\hat{y}^T\,v_\nu)\,v\,.$$

Dann prüft man nach, dass $\tilde{v}$ nicht im von $\tilde{v}_1, \ldots, \tilde{v}_{\nu-1}$ aufgespannten Kegel liegen kann, und somit existiert ein $\tilde{y} \in \mathbb{R}^m$ mit $\tilde{y}^T\,\tilde{v}_k \geq 0$ für $k = 1, \ldots, \nu - 1$ und $\tilde{y}^T\,\tilde{v} < 0$. Wenn man jetzt $y := (\tilde{y}^T\,v_\nu)\,\hat{y} - (\hat{y}^T\,v_\nu)\,\tilde{y}$ setzt, folgt

$$\begin{aligned}
y^T\,v_k &= (\tilde{y}^T\,v_\nu)\,(\hat{y}^T\,v_k) \;-(\hat{y}^T\,v_\nu)\,(\tilde{y}^T\,v_k) \;=\; \tilde{y}^T\,\tilde{v}_k \;\geq\; 0\,, \qquad 1 \leq k \leq \nu - 1\,,\\
y^T\,v_\nu &= (\tilde{y}^T\,v_\nu)\,(\hat{y}^T\,v_\nu) \;-\; (\hat{y}^T\,v_\nu)\,(\tilde{y}^T\,v_\nu) \;=\; 0\,,\\
y^T\,v &= (\tilde{y}^T\,v_\nu)\,(\hat{y}^T\,v) \;-(\hat{y}^T\,v_\nu)\,(\tilde{y}^T\,v) \;=\; \tilde{y}^T\,\tilde{v} \;<\; 0\,.
\end{aligned}$$

Das war zu zeigen! □

Jetzt können wir den folgenden wichtigen Satz beweisen:

Satz 4.3.4 (Lemma von Farkas) *Es gilt immer genau eine der folgenden Aussagen:*

(a) *Das lineare Gleichungssystem* $A\,x = b$ *hat eine Lösung* $x \in \mathbb{R}^n_+$.

(b) *Es gibt ein* $y \in \mathbb{R}^m$ *mit* $y^T\,A \geq 0, \quad y^T\,b < 0$.

Beweis: Wenn (a) richtig ist, und wenn $A^T\,y \geq 0$ ist, dann folgt $y^T\,b = y^T\,A\,x \geq 0$, und deshalb kann (b) nicht gelten. Sei jetzt vorausgesetzt dass (a) nicht gilt. Das bedeutet genau, dass b nicht in dem von den Spalten $a_1, \ldots, a_n$ von A aufgespannten Kegel liegt. Nach Satz 4.3.3 gibt es also ein $y \in \mathbb{R}^m$ mit $y^T\,a_k \geq 0$, $1 \leq k \leq n$, sowie $y^T\,b < 0$. Dies ist aber zu (b) äquivalent. □

Aus dem Lemma von Farkas erhalten wir weitere Alternativsätze für lineare Ungleichungen:

Satz 4.3.5 *Es gilt immer genau eine der folgenden Aussagen:*

(a) *Das lineare Ungleichungssystem* $A\,x \leq b$ *hat eine Lösung* $x \in \mathbb{R}^n_+$.

(b) *Es gibt ein* $y \in \mathbb{R}^m_+$ *mit* $y^T\,A \geq 0, \quad y^T\,b < 0$.

Beweis: Wenn (a) richtig ist, dann gilt für alle $y \in \mathbb{R}^m_+$: Wenn $y^T\,A \geq 0$ ist, dann folgt $y^T\,b \geq y^T\,A\,x \geq 0$, und daher kann (b) nicht gelten. Wenn (a) falsch ist, dann hat das Gleichungssystem $A\,x + I\,\tilde{x} = b$ keine Lösung in $\mathbb{R}^{2n}_+$. Mit dem Farkaslemma folgt deshalb die Existenz von $y \in \mathbb{R}^m$ mit $y^T\,[A, I] \geq 0$, $y^T\,b < 0$, d. h. (b) gilt. □

Satz 4.3.6 *Es gilt immer genau eine der folgenden Aussagen:*

(a) *Das lineare Ungleichungssystem* $A\,x \leq b$ *hat eine Lösung* $x \in \mathbb{R}^n$.

(b) *Es gibt ein* $y \in \mathbb{R}^m_+$ *mit* $y^T\,A = 0, \quad y^T\,b < 0$.

Beweis: Wenn (a) richtig ist, dann gilt für alle $y \in \mathbb{R}^m_+$: Wenn $y^T\,A = 0$ ist, dann folgt $y^T\,b \geq y^T\,A\,x = 0$, und daher kann (b) nicht gelten. Wenn (a) falsch ist, dann hat das Gleichungssystem $A\,x^+ - A\,x^- + I\,\tilde{x} = b$ keine Lösung in $\mathbb{R}^{3n}_+$, denn sonst wäre $x = x^+ - x^-$ eine Lösung für $A\,x \leq b$. Mit dem Farkaslemma folgt deshalb die Existenz von $y \in \mathbb{R}^m$ mit $y^T\,[A, -A, I] \geq 0$, $y^T\,b < 0$. Das bedeutet aber dass $\pm y^T\,A \geq 0$, woraus $y^T\,A = 0$ folgt, und darum gilt (b). □

4.4 Die Darstellungssätze für polyedrische Kegel und Polyeder

Definition 4.4.1 *Wenn* $K \subset \mathbb{R}^n$ *ein Kegel ist, dann heißt*

$$K^* \;:=\; \{x \in \mathbb{R}^n \;:\quad x^T\,y \;\geq\; 0 \quad \forall\, y \in K\}$$

der Dualkegel *zu* K. *Den Dualkegel zu* K^* *nennen wir* Bidualkegel *und schreiben dafür* K^{**}; *beachte, dass im nächsten Resultat gezeigt wird, dass* K^* *ebenfalls ein Kegel ist.*

Lemma 4.4.2 *Für jeden Kegel* $K \subset \mathbb{R}^n$ *gilt:*

(a) *Der Dualkegel K^* ist abgeschlossen und ein Kegel.*

(b) *$K \subset K^{**}$, und genau dann ist $K = K^{**}$, wenn K abgeschlossen ist.*

Beweis: Seien $x_1, x_2 \in K^*$ und $\alpha, \beta \in \mathbb{R}_+$. Dann folgt für alle $y \in K$ dass $y^T(\alpha x_1 + \beta x_2) = \alpha y^T x_1 + \beta y^T x_2 \geq 0$ ist, und deshalb ist K^* ein Kegel. Wenn $x \notin K^*$ ist, dann gibt es ein $y \in K$ so, dass $y^T x \neq 0$ ist. Da die Abbildung $x \longmapsto y^T x$ stetig ist, gibt es ein $\delta > 0$ so, dass für alle $\tilde{x}$ mit $\|\tilde{x} - x\| < \delta$ folgt $y^T \tilde{x} \neq 0$, also $\tilde{x} \notin K^*$. Daher ist das Komplement von K^* offen, also K^* selber abgeschlossen. Damit ist (a) bewiesen. Aus der Definition des Dualkegels folgt dass alle x, welche auf allen $y \in K^*$ senkrecht stehen, zu K^{**} gehören, und dies trifft auf alle $x \in K$ zu. Also ist $K \subset K^{**}$. Da K^{**} nach (a) abgeschlossen ist, folgt aus $K = K^{**}$ die Abgeschlossenheit von K. Sei jetzt $x_0 \notin K$. Wenn K abgeschlossen ist, dann zeigt man mit Hilfe von Satz 4.1.5 (c) die Existenz von $c \in \mathbb{R}^n$ mit $c^T(x - x_0) > 0$ für alle $x \in K$. Da $0 \in K$ ist, folgt $c^T x_0 < 0$. Für $x \in K$ und $\alpha \in \mathbb{R}_+$ folgt aus der Kegeleigenschaft $\alpha x \in K$. Da $c^T(\alpha x) = \alpha c^T x$ ist, folgt dass $c^T x \geq 0$ gelten muss, und deshalb ist $c \in K^*$. Dann kann aber x_0 nicht zu K^{**} gehören, und deshalb ist (b) bewiesen. □

Aufgabe 4.4.3 *Sei $K = \{x = B y : y \geq 0\}$ mit einer n-reihigen Matrix B. Zeige dass K ein Kegel ist, und dass $K^* := \{x^* : B^T x^* \geq 0\}$ der zu K duale Kegel ist.*

Definition 4.4.4 *Ein Kegel $K \subset \mathbb{R}^n$, der gleichzeitig ein Polyeder ist, heißt* polyedrisch. *Er heißt* endlich erzeugt*, falls er die konische Hülle endlich vieler Vektoren ist.*

Aufgabe 4.4.5 *Zeige, dass sowohl jeder Unterraum von $\mathbb{R}^n$ als auch jeder Halbraum der Form $a^T x \leq 0$ mit $a \neq 0$ ein endlich erzeugter Kegel ist.Zeige weiter, dass ein Kegel $K \in \mathbb{R}^n$ genau dann endlich erzeugt ist, wenn es für ein $k \in \mathbb{N}$ eine Matrix $B \in \mathbb{R}^{n \times k}$ gibt, sodass $x \in K$ genau dann gilt, wenn $x = B y$ ist für ein geeignetes $y \geq 0$.*

Definition 4.4.6 *Eine Lösung $x \in \mathbb{R}^n_+$ von $A x = b$ heißt* zulässige Basislösung*, wenn eine Indexmenge $J \subset \{1, \ldots, n\}$ existiert, so dass die Spalten a_j von A mit $j \in J$ linear unabhängig sind, und $x_j = 0$ für $j \notin J$.*

Satz 4.4.7 *Wenn das Gleichungssystem $A x = b$ eine Lösung $x \in \mathbb{R}^n_+$ besitzt, dann existiert auch eine zulässige Basislösung.*

Beweis: Sei eine Lösung $x \in \mathbb{R}^n_+$ gegeben, und sei J die Menge der Indizes k mit $x_k > 0$. Falls x keine zulässige Basislösung ist, dann gilt $\sum_{k \in J} \alpha_k a_k = 0$, wobei mindestens ein $\alpha_k \neq 0$ ist, und o. B. d. A sei $\alpha_k > 0$ angenommen. Seien $\alpha_k = 0$ für $k \notin J$, und sei $\tilde{x}$ der Vektor mit den Koordinaten α_k. Dann folgt $A \tilde{x} = 0$, und daher ist $\hat{x} = x - \lambda \tilde{x}$ ebenfalls eine Lösung des inhomogenen Systems, für alle $\lambda \in \mathbb{R}$. Wir können jetzt $\lambda > 0$ so klein wählen, dass $\hat{x} \geq 0$ ist, und durch evtl. Vergrößern von λ können wir erreichen, dass eine weitere Koordinate von $\hat{x}$ verschwindet. Auf diese Weise erhalten wir eine neue Lösung, die wir wieder x nennen, für welche die entsprechende Indexmenge J ein Element weniger enthält. Setzt man dies fort, so erhält man nach endlich vielen Schritten eine zulässige Basislösung. □

Bemerkung 4.4.8 *Wegen der linearen Unabhängigkeit der entsprechenden Spalten von A folgt, dass eine zulässige Basislösung durch die entsprechende Indexmenge J eindeutig festgelegt ist. Aus diesem Grund ist die Anzahl der verschiedenen zulässige Basislösungen höchstens gleich der Anzahl der Teilmengen von $\{1, \ldots, n\}$, also endlich. Jede zulässige Basislösung kann durch Invertieren der zur Indexmenge J gehörigen Untermatrix von A berechnet werden, und da die Anzahl dieser Matrizen endlich ist, folgt dass die Normen ihrer inversen Matrizen durch eine nur von A abhängige Zahl beschränkt sind. Daraus ergibt sich:* Es gibt eine Konstante $c > 0$, welche von A aber nicht von b abhängt, derart dass für jede zulässige Basislösung x gilt $\|x\| \leq c \|b\|$.

Aufgabe 4.4.9 (Abgeschlossenheit endlich erzeugter Kegel) *Benutze die obige Bemerkung um zu zeigen, dass ein endlich erzeugter Kegel immer abgeschlossen ist.*

Für den Beweis des nächsten Satzes zeigen wir folgendes Hilfsresultat:

Lemma 4.4.10 *Für $a \in \mathbb{R}^n$ ist die Menge $K := \{x \in \mathbb{R}^n_+ : \quad a^T x \geq 0\}$ ein endlich erzeugter Kegel.*

Beweis: Seien J_0, J_+, J_- die Mengen der Indizes $j \in \{1, \ldots, n\}$ mit $a_j = 0$, bzw. $a_j > 0$ bzw. $a_j < 0$. Falls die Menge J_- leer ist, wird K von den kanonischen Basisvektoren erzeut. Falls J_+ leer ist, kann man prüfen, dass die Vektoren e_j für $j \in J_0$ bereits den Kegel erzeugen. Seien jetzt $J_\pm$ beide nicht leer. Die Vektoren

$$e_j \quad (j \in J_0 \cup J_+)\,, \qquad e_{jk} \;:=\; \frac{1}{a_j}\, e_j \;-\; \frac{1}{a_k}\, e_k \quad (j \in J_+\,,\; k \in J_-)$$

liegen alle in K, und ihre Anzahl ist endlich. Ein $x \in K$ ist genau dann eine Linearkombination dieser Vektoren mit Koeffizienten α_j bzw. α_{jk}, wenn

$$x_j \;=\; \begin{cases} \alpha_j & (j \in J_0) \\ \alpha_j + a_j^{-1} \displaystyle\sum_{k \in J_-} \alpha_{jk} & (j \in J_+) \\ -a_j^{-1} \displaystyle\sum_{k \in J_+} \alpha_{kj} & (j \in J_-) \end{cases}$$

und da alle $x_j \geq 0$ sind, kann man $\alpha_{jk} \geq 0$ und $\alpha_j \geq 0$ (in dieser Reihenfolge) immer so wählen, dass diese Gleichungen erfüllt sind (allerdings sind diese Koeffizienten im Allgemeinen nicht eindeutig bestimmt, was aber hier nicht wichtig ist). Ist dies geschehen, so folgt

$$a^T x \;=\; \sum_{j \in J_+} a_j\, \alpha_j \;\geq\; 0\,.$$

Also ist K die konische Hülle dieser Vektoren e_j und e_{jk}, und somit insbesondere ein Kegel und endlich erzeugt. □

Satz 4.4.11 (Minkowski-Weyl) *Für $K \subset \mathbb{R}^n$ sind folgende Aussagen äquivalent:*

(a) *K ist ein polyedrischer Kegel.*

(b) *Es gibt $A \in \mathbb{R}^{m \times n}$ mit geeignetem $m \in \mathbb{N}$ so, dass K die Lösungsmenge von $A\,x \leq 0$ ist.*

(c) *K ist ein endlich erzeugter Kegel.*

Beweis: Sei (a) erfüllt. Dann gibt es ein $A \in \mathbb{R}^{m \times n}$ und ein $b \in \mathbb{R}^m$ so, dass $K = P(A, b)$ ist. Aus der Kegeldefinition folgt dass $\alpha\, x \in K$ ist für alle $x \in K$ und alle $\alpha > 0$, und wegen $A\,(\alpha\, x) = \alpha\,(A\, x) \leq b$ ergibt sich $A\, x \leq 0$ für alle $x \in K$. Da $0 \in K$ ist, folgt $b \geq 0$, und deshalb folgt aus $A\, x \leq 0$ auch umgekehrt $x \in K$. Also ist $K = P(A, 0)$, und deshalb gilt (b). Sei jetzt $K = P(A, 0)$. Wir zeigen (c) durch Induktion über m. Für $m = 1$ ist K entweder ganz $\mathbb{R}^n$ oder ein Halbraum, und in beiden Fällen ist (c) richtig wegen Aufgabe 4.4.5. Wenn die Richtigkeit für ein m bewiesen ist, gibt es nach der gleichen Aufgabe ein $B \in \mathbb{R}^{n \times k}$, für welche

$$K \;=\; \{\; x = B\, y \;:\quad y \geq 0 \;\}\,.$$

Wenn wir jetzt zur Matrix A eine weitere Zeile a^T hinzufügen, dann ist der neue Kegel $\tilde{K}$ der Durchschnitt von K mit dem Halbraum $a^T\, x \leq 0$. Wenn wir $c^T = -a^T\, B$ setzen, dann folgt

$$\tilde{K} \;=\; \{\; x = B\, y \;:\quad y \geq 0\,,\quad c^T\, y \geq 0 \;\}\,.$$

Nach Lemma 4.4.10 ist die Menge dieser y ein endlich erzeugter Kegel, und daher gibt es ein C so, dass $\{ y \geq 0\,, \quad c^T y \geq 0 \} = \{C z \,:\, z \geq 0\}$. Daraus folgt aber $\tilde{K} = \{x \,:\, x = B C z \,:\, z \geq 0\}$, woraus (c) für $m+1$ anstelle von m folgt. Wenn schließlich (c) erfüllt ist, d. h., wenn $K = \{x = B y \,:\, y \geq 0\}$, dann ist nach Aufgabe 4.4.3 $K^* = \{\tilde{x} \in \mathbb{R}^n \,:\, \tilde{x}^T B \geq 0\}$ der zu K duale Kegel. Offenbar ist K^* polyedrisch und deshalb nach dem bereits gezeigten Teil des Satzes endlich erzeugt. Es gibt also eine Matrix $A \in \mathbb{R}^{m \times n}$, mit geeignetem $m \in \mathbb{N}$, so dass

$$K^* \;=\; \{\; \tilde{x} = -A^T z \;:\quad z \geq 0 \;\}\,.$$

Sei jetzt $x \in K$, also $x = B y$ für ein geeignetes $y \geq 0$. Dann gilt für alle $z \geq 0$ dass $z^T A B y = -\tilde{x}^T B y \leq 0$ ist, und daraus folgt $A x \leq 0$. Ist umgekehrt $A x \leq 0$, dann ist $-z^T A x \geq 0$ für alle $z \geq 0$. Das bedeutet dass x zum Bidualkegel K^{**} gehört, und da K nach Aufgabe 4.4.9 abgeschlossen ist, gilt $K^{**} = K$. Also ist K gleich dem Polyeder $A x \leq 0$, und deshalb gilt (a). □

Definition 4.4.12 (Projektion) *Für $x \in \mathbb{R}^n$ und $1 \leq k \leq n$ sei $\pi_k(x)$ der Vektor aus $\mathbb{R}^k$, dessen Koordinaten mit den ersten k Koordinaten von x identisch sind. Dies ergibt eine lineare Abbildung von $\mathbb{R}^n$ nach $\mathbb{R}^k$, welche wir* Projektion *nennen, obwohl sie nicht die im ersten Teil der Vorlesung verlangten Eigenschaften hat. Weiter nennen wir für $C \in \mathbb{R}^{k \times m}$ und $d \in \mathbb{R}^k$, mit einer beliebigen natürlichen Zahl k, die Abbildung $x \longmapsto C x + d$ eine* affine Abbildung. *In anderen Worten ist eine affine Abbildung die Hintereinanderausführung einer linearen Abbildung und einer Translation.*

Lemma 4.4.13 *Das Bild eines Polyeders unter einer affinen Abbildung ist ein Polyeder. Die Summe zweier Polyeder ist ebenfalls ein Polyeder.*

Beweis: Seien $A \in \mathbb{R}^{m \times n}$ und $b \in \mathbb{R}^m$, sowie C und d wie oben. Die Menge der Paare (x, y) mit $A x \leq b$ und $y = C x + d$ ist ein Polyeder in $\mathbb{R}^{n+k}$, und die Bildmenge von $P = P(A, b)$ unter der affinen Abbildung $x \longmapsto C x + d$ ist eine Projektion dieses Polyeders. Also reicht es zu zeigen, dass die Projektion eines Polyeders wieder ein Polyeder ist. Dazu sei $A = [A_1, A_2]$ gesetzt, wobei A_1 die ersten k Spalten von A enthält. Dann ist

$$\pi_k(P) \;=\; \{\tilde{x} \in \mathbb{R}^k \;:\quad \exists\, \hat{x} \in \mathbb{R}^{n-k} \quad \text{mit} \quad A_1 \tilde{x} + A_2 \hat{x} \leq b \;\}\,.$$

Aus Satz 4.3.6 folgt jetzt

$$\pi_k(P) \;=\; \{\tilde{x} \in \mathbb{R}^k \;:\quad \forall\, y \in \mathbb{R}^m_+ \quad \text{mit} \quad y^T A_2 \;=\; 0 \quad \text{gilt} \quad y^T (b - A_1 \tilde{x}) \;\geq\; 0 \;\}\,.$$

Die Menge $K := \{y \geq 0\,,\; y^T A_2 = 0\}$ ist ein polyedrischer Kegel, und deshalb nach Satz 4.4.11 endlich erzeugt. Daher gibt es $y_1, \ldots, y_r \in \mathbb{R}^m$ mit $K = \text{cone}\{y_1, \ldots, y_r\}$, und man erkennt dass die Bedingung $y^T (b - A_1 \tilde{x}) \;\geq\; 0$ für alle $y \in K$ genau dann gilt, wenn $y_j^T (b - A_1 \tilde{x}) \;\geq\; 0$ für alle $j = 1, \ldots, r$ erfüllt ist. Dies zeigt, dass die Menge $\pi_k(P)$ durch die Ungleichungen

$$(y_j^T A_1)\, \tilde{x} \;\leq\; (y_j^T b) \qquad \forall\, j = 1, \ldots, r$$

charakterisiert und deshalb ein Polyeder ist. Auch die Summe $P_1 + P_2$ zweier Polyeder $P_1 = P(A_1, b_1)$ und $P_2 = P(A_2, b_2)$ ist die Projektion des Polyeders aller Tripel (x, y, z) mit

$$\begin{bmatrix} A_1 & 0 & 0 \\ 0 & A_2 & 0 \\ I & I & -I \\ -I & -I & I \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} \quad \leq \quad \begin{bmatrix} b_1 \\ b_2 \\ 0 \\ 0 \end{bmatrix}$$

und ist deshalb ein Polyeder. □

Satz 4.4.14 (Darstellungssatz für Polyeder) *Jedes Polyeder $P = P(A, b)$ ist die Summe aus einem Polytop und dem sogenannten* charakteristischen Kegel *$P(A, 0)$. Umgekehrt ist die Summe eines Polytops und eines endlich erzeugten Kegels immer ein Polyeder.*

Beweis: Ein Polytop Q kann geschrieben werden als $Q = \{x = Bz : z \in \mathbb{R}^k_+, \sum_j z_j = 1\}$, ist also das Bild des Polyeders $P = \{z : z \in \mathbb{R}^k_+, \sum_j z_j = 1\}$ unter der linearen Abbildung $z \longmapsto Bz$. Nach Lemma 4.4.13 ist Q also ein Polyeder. Nach Satz 4.4.11 ist ein endlich erzeugter Kegel ebenfalls ein Polyeder, und dann gilt dies auch für die Summe eines Polytops und eines endlich erzeugten Kegels. Sei jetzt $P = P(A, b)$ gegeben, und sei $K = P(A, 0)$. Die Menge $\tilde{P}$ aller Paare (x, α) mit $\alpha \in \mathbb{R}$ und $Ax - \alpha b \leq 0$, $-\alpha \leq 0$ ist ein polyedrischer Kegel und somit endlich erzeugt. Also gibt es endlich viele Paare (x_j, α_j), deren konische Hülle gleich $\tilde{P}$ ist, und wir können o. B. d. A. annehmen dass die α_j entweder gleich 0 oder gleich 1 sind. Der erste Fall bedeutet genau dass $x_j \in P(A, 0)$ ist, während der andere Fall zu $x_j \in P(A, b)$ äquivalent ist. Weil außerdem $x \in P(A, b)$ zu $(x, 1) \in \tilde{P}$ gleichwertig ist, folgt die Behauptung. □

Aufgabe 4.4.15 *Zeige: Ein nicht-leeres Polyeder ist genau dann beschränkt, wenn es ein Polytop ist.*

4.5 Ecken von Polyedern

Definition 4.5.1 (Ecken einer konvexen Menge) *Sei $M \subset V$ eine konvexe Menge. Eine Konvexkombination, also eine endliche Summe der Form $v = \sum_j \alpha_j v_j$, mit $v_j \in M$, $\alpha_j \in \mathbb{R}_+$ und $\sum_j \alpha_j = 1$ heißt* echt*, falls alle $\alpha_j > 0$ sind. Diejenigen Punkte $v \in M$, welche keine echten Konvexkombinationen von zwei verschiedenen $v_1, v_2 \in M$ sind, heißen* Ecken *von M.*

Aufgabe 4.5.2 *Gib Beispiele einer konvexen Menge in $\mathbb{R}^n$ mit einer unendlichen Anzahl von Ecken, und zeige dass affine Räume keine Ecken besitzen, außer wenn sie nur ein Element enthalten. Zeige genauer: Ist A ein affiner Raum mit positiver Dimension, dann gibt es zu jedem $x_0 \in A$ und jedem $\varepsilon > 0$ zwei Punkte $x_1, x_2 \in A$, mit $\|x_j - x_0\| < \varepsilon$, so dass x_0 echte Konvexkombination von x_1 und x_2 ist. Zeige weiter, dass die Ecken einer konvexen Menge $K \subset \mathbb{R}^n$ niemals innere Punkte von K sind.*

Definition 4.5.3 (Aktive Nebenbedingungen) *Seien $A \in \mathbb{R}^{m \times n}$ und $b \in \mathbb{R}^m$, und sei $P = P(A, b)$ das Polyeder $Ax \leq b$. Wir nennen die m Ungleichungen $a_j^T x \leq b_j$ auch* Nebenbedingungen *für P und bezeichnen diejenigen, für die bei gegebenem $x \in P$ das Gleichheitszeichen gilt, als die für x* aktiven *Nebenbedingungen. Klar ist, dass die Menge der für ein x aktiven Nebenbedingungen leer sein kann. Weiter nennen wir aktive Nebenbedingungen* unabhängig*, wenn das zugehörige System von Vektoren a_j linear unabhängig ist.*

Aufgabe 4.5.4 *Seien $a, x \in \mathbb{R}^n$ und $b \in \mathbb{R}$ so, dass $a \neq 0$ und $a^T x = b$ ist, und sei x echte Konvexkombination zweier Vektoren $x_1, x_2 \in \mathbb{R}^n$. Zeige: Wenn $a^T x_1 < b$ ist, dann muss $a^T x_2 > b$ sein.*

Lemma 4.5.5 *Seien $A \in \mathbb{R}^{m \times n}$ und $b \in \mathbb{R}^m$, und sei $P = P(A, b)$ das Polyeder $Ax \leq b$. Genau dann ist $x \in P$ eine Ecke, wenn es n Zahlen $j_k \in \{1, \ldots, m\}$ gibt, für welche die zugehörigen Nebenbedingungen aktiv und unabhängig sind. Insbesondere hat P keine Ecken, falls* $\operatorname{rang} A < n$ *ist.*

Beweis: Sei $x_0 \in P$, und sei $J \subset \{1, \ldots, m\}$ die (eventuell leere) Menge der für x aktiven Nebenbedingungen. Für alle $j \notin J$ ist dann x_0 ein innerer Punkt der Halbräume $a_j^T x \leq 0$, und daher gibt es ein $\varepsilon > 0$ derart, dass alle x mit $\|x - x_0\| < \varepsilon$ ebenfalls innere Punkte dieser Halbräume sind. Also sind für solche x, soweit sie in P liegen, alle $j \notin J$ ebenfalls inaktiv. Somit ist x_0 keine Ecke von P, falls J leer ist. Für nicht-leeres J legen die aktiven Nebenbedingungen einen affinen Raum A fest, der x_0 enthält und dessen Dimension gleich n minus der maximalen Zahl von linear unabhängigen Vektoren im System $(a_j, j \in J)$ ist. Falls $\dim A > 0$ ist, dann folgt aus Aufgabe 4.5.2 dass x_0 keine Ecke sein kann. Falls dagegen $\dim A = 0$ ist, dann folgt mit Aufgabe 4.5.4 dass x_0 keine echte Konvexkombination von verschiedenen Punkten aus P sein kann und somit eine Ecke ist. □

Satz 4.5.6 *Seien $A \in \mathbb{R}^{m\times n}$ und $b \in \mathbb{R}^m$, und sei $P = P(A,b)$ das Polyeder $Ax \leq b$. Falls $P \neq \emptyset$ ist, sind folgende Aussagen äquivalent:*

(a) $\operatorname{rang} A = n$.

(b) *P besitzt eine Ecke.*

(c) *P enthält keine Gerade.*

(d) *Der Nullpunkt ist Ecke von $P(A,0)$.*

In jedem Fall ist die Anzahl der Ecken von P endlich.

Beweis: Sei $\operatorname{rang} A = n$. Dann gibt es genau ein $x \in P$ mit $Ax = b$, und dieses x ist eine Ecke von P. Also folgt (b) aus (a). Wenn (c) nicht gilt, d. h., wenn P eine Gerade enthält, dann gibt es Vektoren x_0 und $x \neq 0$ derart, dass $x_0 + \alpha x \in P$ ist für alle $\alpha \in \mathbb{R}$. Wegen $A(x_0 + \alpha x) = Ax_0 + \alpha Ax \leq b$ kann dies nur gelten, wenn $Ax = 0$ ist. Also folgt $\operatorname{rang} A < n$, und dann hat P nach dem letzten Lemma keine Ecke, d. h., (b) ist falsch. Somit folgt (c) aus (b). Der Unterraum $Ax = 0$ ist in $P(A,0)$ enthalten und enthält genau dann keine Gerade, wenn seine Dimension gleich 0, d. h., der Rang von A gleich n ist. Dies ist genau dann der Fall, wenn $x = 0$ eine Ecke von $P(A)$ ist, da ja für $x = 0$ alle Nebenbedingungen aktiv sind. Eine Gerade aus $P(A,0)$ kann aber durch Translation zu einer Geraden in P verschoben werden, und deshalb ist (c) falsch, falls der Nullpunkt keine Ecke von $P(A,0)$ ist. Im anderen Fall muss es n unabhängige Nebenbedingungen geben, und deshalb folgt (a) aus (d). Die Endlichkeit der Eckenzahl ist klar nach dem letzten Lemma, weil n aktive und unabhängige Nebenbedingungen eine Ecke eindeutig festlegen. □

Kapitel 5

Lineare Optimierung

In diesem Kapitel werden wir uns ausschließlich mit $\mathbb{R}^n$, also dem sogenannten *n-dimensionalen Euklidischen Raum,* befassen, wobei $n \geq 2$ in der Regel eine fest gewählte aber beliebige natürliche Zahl ist. Wir wollen das sogenannte *Simplex-Verfahren* zur (numerischen) Bestimmung eines Minimums eines linearen Funktionals, meist als *Kostenfunktional* bezeichnet, unter linearen Nebenbedingungen kennen lernen. Die Nebenbedingungen sind dabei stets durch lineare (Un-)gleichungen beschrieben. Die Behandlung eines solchen, aber auch weit allgemeinerer Probleme ist eine Frage, die zum Gebiet der *linearen Optimierung* gehört. Da typischerweise die Dimension n sehr groß ist, ist die Bestimmung dieses Minimums ein Problem, das theoretisch einfach zu verstehen, dessen Lösung aber ohne Rechner normalerweise unmöglich ist. Da das Kostenfunktional ein lineares Funktional auf $\mathbb{R}^n$ ist, gibt es nach dem sogenannten *Darstellungssatz* einen Vektor $c \in \mathbb{R}^n$, so dass diese Funktion als $x \longmapsto c^T x$, $x \in \mathbb{R}^n$, dargestellt werden kann. Die auftretenden linearen Nebenbedingungen kann man immer in der Form

$$A\,x \;\geq b\,, \qquad x \geq 0$$

schreiben, wobei $A \in \mathbb{R}^{m\times n}$ und $b \in \mathbb{R}^m$ gegeben sind.

Wenn nichts anderes gesagt ist, schreiben wir immer x_j bzw. b_j bzw. c_j für die Koordinaten der Vektoren x bzw. b bzw. c, und a_{jk} bezeichnet das Element von A in der entsprechenden Position. Wir definieren als das *Grundproblem der linearen Programmierung* die folgende Aufgabe:

- Für gegebene $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ und $c \in \mathbb{R}^n \setminus \{0\}$, betrachte die Menge $\mathcal{Z}$ aller $x \in \mathbb{R}^n$ mit

  $$x \;\geq\; 0\,, \qquad A\,x \;\geq\; b\,. \tag{5.0.1}$$

  Die *Nebenbedingungen* in (5.0.1) entsprechen genau $n+m$ Ungleichungen der Form $a^T x \geq \beta$, wobei a entweder das Transponierte einer Zeile von A oder ein Vektor der kanonischen Basis von $\mathbb{R}^n$ ist, und β ist im ersten Fall die entsprechende Koordinate von b, bzw. $= 0$ im zweiten Fall. Es gilt nun, wenn möglich, ein Element $x^* \in \mathcal{Z}$ so zu bestimmen, dass das Funktional

  $$f(x) = c^T\,x = \langle c, x\rangle \tag{5.0.2}$$

  für $x = x^*$ minimal wird. Dabei heißt $\mathcal{Z}$ auch *zulässiger Bereich* und ist ein Polyeder, und f heißt *Kostenfunktional* oder *Zielfunktion.* Ein $x^* \in \mathcal{Z}$ mit $c^T\,x \geq c^T\,x^*$ für alle $x \in \mathcal{Z}$ heißt auch *ein Optimum für das Problem* (P), welches wir auch gelegentlich *primales Problem* nennen wollen. Als eine suggestive Formulierung dieses Problems ist es üblich zu schreiben:

  $$\text{(P)} \qquad \left\{\begin{array}{rcl} c^T\,x & \to & \min \\ A\,x & \geq & b \\ x & \geq & 0 \end{array}\right. \tag{5.0.3}$$

Aufgabe 5.0.7 *Gib ein einfaches Beispiel dafür, dass ein Optimum, wenn es existiert, nicht eindeutig zu sein braucht.*

Für den Moment wollen wir der Einfachheit halber annehmen, dass der zulässige Bereich $\mathcal{Z} \neq \emptyset$ ist, da sonst auch keine Lösung des Optimierungsproblems existiert. Unter dieser Annahme muss es Werte f_0 geben, für welche die Hyperebene $c^T x = f_0$ den Bereich $\mathcal{Z}$ schneidet, und dann ergeben sich folgende zwei Fälle:

- Falls ein f_0 existiert, für welches $\mathcal{Z}$ ganz im Halbraum $c^T x \geq f_0$ enthalten ist, dann gilt dies erst recht für alle kleineren Werte von f_0, und somit folgt für $f_1 < f_0$, dass die Hyperebene $c^T x = f_1$ den Bereich $\mathcal{Z}$ nicht schneidet (denn ein möglicher Schnittpunkt wäre ja auch im Halbraum $c^T x \geq f_0$ enthalten, was nicht sein kann). Also ist das Kostenfunktional nach unten beschränkt, und eine Minimalstelle x_0 (wenn sie existiert, was noch nicht bewiesen ist) liegt auf der Hyperebene $c^T x = f_0^*$, wobei f_0^* das Infimum aller f_0 ist, für welche sich $\mathcal{Z}$ und die Hyperebene $c^T x = f_0$ schneiden.
- Falls *kein* f_0 existiert, für welches $\mathcal{Z}$ ganz im Halbraum $c^T x \geq f_0$ enthalten ist, dann ist das Zielfunktional nach unten unbeschränkt, und somit existiert kein Optimum.

Aufgabe 5.0.8 *Gib einfache Beispiele dafür, dass die beiden obigen Fälle in der Tat eintreten können.*

Bemerkung 5.0.9 *Falls der zulässige Bereich $\mathcal{Z}$ nicht leer und beschränkt ist, folgt aus Ergebnissen der Analysis die Existens eines Optimums $x^* \in \mathcal{Z}$, aber bei unbeschränktem $\mathcal{Z}$ ist dies noch zu zeigen, selbst wenn der erste der beiden obigen Fälle eintritt.*

5.1 Äquivalente Formulierungen des Optimierungsproblems

In den Anwendungen kann manchmal ein Problem auftreten, in dem zum Beispiel nur Gleichungen der Form $Ax = b$ als Nebenbedingungen auftreten, oder für welches die Bedingung $x \geq 0$ fehlt. In solchen Fällen kann man das Optimierungsproblem immer auf die *Normalform* (P) bringen. Umgekehrt kann man aber auch ein Problem in Normalform in ein äquivalentes Problem einer anderen Form umschreiben. Bei allen diesen Transformationen tritt allerdings eine Erhöhung der Anzahl der Variablen und/oder Nebenbedingungen auf, so dass diese Umschreibung nicht unbedingt für die praktische Behandlung zu empfehlen ist. Es ist aber wichtig zu wissen, dass die hier noch zu zeigenden allgemeinen Resultate für Probleme in Normalform, bzw. in der unten definierten Standardform auch auf andere Fälle übertragbar sind.

1. Vorgelegt sei ein Optimierungsproblem der Form

$$\left\{ \begin{array}{rcl} c^T x & \to & \min \\ Ax & \geq & b \end{array} \right.$$

Hier ist also nicht verlangt, dass die Vektoren des zulässigen Bereichs in $\mathbb{R}^n_+$ liegen. Ein solches Problem ist äquivalent zu einem anderen in der Normalform (P), aber in $\mathbb{R}^{2n}$: Wir setzen (mit nicht eindeutig bestimmten Vektoren $x^{\pm}$)

$$x = x^+ - x^-, \quad \tilde{x} = \begin{bmatrix} x^+ \\ x^- \end{bmatrix}, \quad \tilde{c} = \begin{bmatrix} c \\ -c \end{bmatrix}, \quad \tilde{b} = b, \quad \tilde{A} = [A, -A].$$

Dann ist obiges Problem äquivalent zu

$$\left\{ \begin{array}{rcl} \tilde{c}^T \tilde{x} & \to & \min \\ \tilde{A}\tilde{x} & \geq & \tilde{b} \\ \tilde{x} & \geq & 0. \end{array} \right.$$

2. Wenn ein Problem in Normalform (P) gegeben ist, dann kann man setzen

$$\tilde{A} \;=\; \begin{bmatrix} A \\ I \end{bmatrix}, \quad \tilde{b} \;=\; \begin{bmatrix} b \\ 0 \end{bmatrix},$$

und findet die Äquivalenz von (P) zum Problem

$$\left\{ \begin{array}{rcl} c^T x & \to & \min \\ \tilde{A}\, x & \geq & \tilde{b} \end{array} \right.$$

3. Wenn ein Problem in der sogenannten *Standardform*

$$\text{(S)} \qquad \left\{ \begin{array}{rcl} c^T x & \to & \min \\ A\, x & = & b \\ x & \geq & 0 \end{array} \right.$$

gegeben ist, dann kann man setzen

$$\tilde{A} \;=\; \begin{bmatrix} A \\ -A \end{bmatrix}, \quad \tilde{b} \;=\; \begin{bmatrix} b \\ -b \end{bmatrix},$$

und findet die Äquivalenz zum Problem

$$\left\{ \begin{array}{rcl} c^T x & \to & \min \\ \tilde{A}\, x & \geq & \tilde{b} \\ x & \geq & 0 \end{array} \right.$$

was einProblem in Normalform ist.

4. Für ein Problem in Normalform (P) kann man sogenannte *Schlupfvariable* einführen, um ein äquivalentes Problem zu erhalten, in dem Nebenbedingungen in Gleichungsform vorliegen: Für $y \in \mathbb{R}^n$ seien $\tilde{A} = [A, -I]$ und

$$\tilde{x} \;=\; \begin{bmatrix} x \\ y \end{bmatrix}, \qquad \tilde{c} \;=\; \begin{bmatrix} c \\ 0 \end{bmatrix}.$$

Dann ist (P) offenbar äquivalent mit

$$\left\{ \begin{array}{rcl} \tilde{c}^T \tilde{x} & \to & \min \\ \tilde{A}\, \tilde{x} & = & b \\ \tilde{x} & \geq & 0 \end{array} \right.$$

was offenbar ein Problem in Standardform ist.

Als Fazit dieser Diskussion halten wir fest, dass alle Probleme der oben betrachteten Formen wahlweise in eines in Normal- oder in Standardform umgewandelt werden können! Außerdem sind alle Probleme so, dass der zulässige Bereich ein Polyeder in $\mathbb{R}^n$ ist. Daher können wir allgemein auch ein Problem der folgenden abstrakten Form betrachten:

- Gegeben sei ein Polyeder $\mathcal{Z} \subset \mathbb{R}^n$ und ein Vektor $c \in \mathbb{R}^n$. Dann lautet das zugehörige Optimierungsproblem

$$\left\{ \begin{array}{rcl} \tilde{c}^T \tilde{x} & \to & \min \\ x & \in & \mathcal{Z} \end{array} \right.$$

5.2 Das duale Problem und der schwache Dualitätssatz

Um die Existenz eines Optimums untersuchen zu können, benötigen wir einige Vorbereitungen. Insbesondere ist es hilfreich, das sogenannte duale Problem zu betrachen.

Definition 5.2.1 (Duales Problem) *Das Optimierungsproblem*

$$(\mathrm{D}) \qquad \left\{ \begin{array}{rcl} b^T y & \to & \max \\ A^T y & \leq & c \\ y & \geq & 0 \end{array} \right. \tag{5.2.1}$$

heißt das zu (P) duale Optimierungsproblem. *Dabei ist natürlich* $y \in \mathbb{R}^m$, *und* (D) *kann auf Normalform gebracht werden, indem man die ersten beiden Bedingungen mit* -1 *multipliziert. Man sieht dann, dass das zu dem so umgeschriebenen* (D) *duale Problem wiederum äquivalent zu* (P) *ist. Um Missverständnisse auszuschließen, bezeichnen wir die zulässigen Bereiche des primalen bzw. dualen Problems auch mit* $\mathcal{Z}_P$ *bzw.* $\mathcal{Z}_D$.

Satz 5.2.2 (Schwacher Dualitätssatz) *Für alle* $x \in \mathcal{Z}_P$ *und* $y \in \mathcal{Z}_D$ *gilt immer* $b^T y \leq c^T x$. *Wenn für ein Paar* $x \in \mathcal{Z}_P$ *und* $y \in \mathcal{Z}_D$ *sogar Gleichheit gilt, dann ist* x *optimal für* (P) *und* y *optimal für* (D).

Beweis: Da $x \geq 0$ und $y \geq 0$ ist, folgt $y^T b \leq y^T A x \leq c^T x$. Wenn x nicht das Optimum von (P) ist, gibt es ein $\tilde{x}$ mit $c^T \tilde{x} < c^T x$, und dann gilt $b^T y \leq c^T \tilde{x} < c^T x$. Genauso schließt man, wenn y nicht optimal für (D) ist. □

Korollar zu Satz 5.2.2 *Wenn für* $x \in \mathcal{Z}_P$ *und* $y \in \mathcal{Z}_D$ *gilt*

$$y^T (A x - b) = 0, \qquad x^T (c - A^T y) = 0, \tag{5.2.2}$$

dann ist x *optimal für* (P) *und* y *optimal für* (D).

Beweis: Offenbar gilt

$$c^T x - b^T y = x^T (c - A^T y) + y^T (A x - b).$$

Falls (5.2.2) gilt, folgt $b^T y = c^T x$, also die Behauptung. □

Definition 5.2.3 (Komplementaritätsbedingungen) *Die Beziehungen* (5.2.2) *heißen auch* Komplementaritätsbedingungen. *Wir werden noch zeigen, dass sie auch notwendig dafür sind, dass* x *und* y *optimal sind.*

Bemerkung 5.2.4 *Die Komplementaritätsbedingungen sind ausgeschrieben gleich den folgenden Aussagen*

$$\forall\, j = 1, \ldots, m: \qquad \sum_{k=1}^{n} a_{jk} x_k > b_j \implies y_j = 0,$$

$$\forall\, k = 1, \ldots, n: \qquad \sum_{j=1}^{m} a_{jk} y_j < c_j \implies x_k = 0.$$

5.3 Existenz- und starker Dualitätssatz

Definition 5.3.1 (Optimalwerte) *Mit den Bezeichnungen vom Anfang dieses Kapitels seien, unter Beachtung der üblichen Konvention für Supremum und Infimum der leeren Menge,*

$$v(P) \;=\; \inf_{x\in\mathcal{Z}_P} c^T x\,, \qquad v(D) \;=\; \sup_{x\in\mathcal{Z}_D} b^T x\,.$$

Wir nennen diese Zahlen dann auch Optimalwerte *des entsprechenden Problems; beachte aber, dass diese Werte auch $\pm\infty$ sein können.Die Lösbarkeit des primalen bzw. dualen Problems ist also äquivalent damit, dass die jeweilige Zielfunktion den Optimalwert annimmt.*

Satz 5.3.2 (Starker Dualitätssatz) *Folgende Aussagen sind äquivalent:*

(a) *Das primale Problem hat eine optimale Lösung.*

(b) *Das duale Problem hat eine optimale Lösung.*

(c) *Die zulässigen Bereiche $\mathcal{Z}_P$ und $\mathcal{Z}_D$ sind beide nicht leer.*

(d) *Das primale und das duale Problem haben beide optimale Lösungen, und die Optimalwerte stimmen überein.*

Beweis: Sei zunächst angenommen, dass (c) erfüllt ist. Dann folgt aus dem schwachen Dualitätssatz

$$\forall\, x\in\mathcal{Z}_P\,,\; y\in\mathcal{Z}_D\,: \qquad -\infty \;<\; b^T y \;\le\; v(D) \;\le\; v(P) \;\le\; c^T x < \;+\infty\,.$$

Das bedeutet insbesondere, dass $v(P)$ und $v(D)$ endlich sind. Wenn das Gleichungssystem

$$\begin{bmatrix} c^T & 0 \\ A & -I \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \;=\; \begin{bmatrix} v(D) \\ b \end{bmatrix}$$

keine nicht-negative Lösung hätte, müsste es nach dem Lemma von Farkas eine Zahl δ und ein $y\in\mathbb{R}^m$ geben, für welche die Ungleichungen

$$\delta\, c^T \;+\; y^T A \;\ge\; 0\,, \quad -y \;\ge 0\,, \quad v(D)\,\delta \;+\; y^T b \;<\; 0$$

erfüllt sind. Wenn $\delta\le 0$ wäre, dann müsste für $x\in\mathcal{Z}_P$ wegen $v(D)\le v(P)\le c^T x$ sowie $y\le 0$ gelten

$$v(D)\,\delta \;+\; y^T b \;\ge \delta\, c^T x \;+\; y^T b \;\ge\; -y^T A\, x \;+\; y^T b \;\ge\; -y^T b \;+\; y^T b \;=\; 0\,,$$

was nicht sein kann. Also müsste $\delta > 0$ sein, woraus mit $\tilde y := -\delta^{-1} y$ aber folgt dass $\tilde y\in\mathcal{Z}_D$ und $b^T\tilde y > v(D)$ sein müsste, was ebenfalls ein Widerspruch ist. Also muss das obige Gleichungssystem eine nicht-negative Lösung haben, und daraus folgt dass $x\in\mathcal{Z}_P$ und $v(D) = c^T x$ ist. Also ist x ein Optimum für (P), und $v(D) = v(P)$. Durch Vertauschen der Rollen von (P) und (D) folgt genauso die Existenz eines Optimums von (D), und deshalb gilt Aussage (d). Klar ist, dass aus (d) sowohl (a) als auch (b) folgen, und daher reicht es aus, noch folgendes zu zeigen: Wenn $\mathcal{Z}_P = \emptyset$ ist, dann hat das duale Problem keine optimale Lösung (durch Vertauschen von primalem und dualem Problem folgt dann auch, dass das primale Problem kein Optimum hat, falls $\mathcal{Z}_D$ leer ist, und deshalb können weder (a) noch (b) richtig sein, falls (c) nicht erfüllt ist). Sei also angenommen, dass $\mathcal{Z}_P = \emptyset$, dass also das System $Ax\ge b$ keine Lösung in $\mathbb{R}^n_+$ besitzt. Dann folgt aus Satz 4.3.5 (durch Übergang zu $-A$ und $-b$) die Existenz von $y\in\mathbb{R}^m_+$ mit $y^T A\le 0$ und $y^T b > 0$. Falls auch $\mathcal{Z}_D$ leer ist, ist nichts zu zeigen, also sei das Gegenteil angenommen. Für $\tilde y\in\mathcal{Z}_D$ und $\alpha > 0$ setzen wir $z = \tilde y + \alpha y$. Dann ist $z\ge 0$, sowie $z^T A\le \tilde y A\le c$, und deshalb ist $\tilde y\in\mathcal{Z}_D$. Weiter ist $b^T z\longrightarrow +\infty$ für $\alpha\longrightarrow+\infty$, woraus $v(D) = +\infty$ folgt. □

Aus dem Beweis des starken Dualitätssatzes erhalten wir jetzt sofort folgenden Existenzsatz:

Satz 5.3.3 (Existenzsatz) *Sei $\mathcal{Z}_P \neq \emptyset$. Falls die Zielfunktion $c^T x$ auf $\mathcal{Z}_P$ nach unten beschränkt ist, dann existiert ein Optimum für das primale Problem. Die Beschränktheit der Zielfunktion ist insbesondere dann gewährleistet, wenn der zulässige Bereich $\mathcal{Z}_P$ beschränkt ist.*

Beweis: Wenn $\mathcal{Z}_P \neq \emptyset$ und $v(P) > -\infty$ gilt, dann folgt aus dem Beweis des letzten Satzes dass auch $\mathcal{Z}_D$ nicht leer ist. Aus dem starken Dualitätssatz selber folgt dann die Behauptung. □

5.4 Das Simplexverfahren

Die Basis des Simplexverfahrens beruht auf der folgenden einfachen aber zentralen Feststellung:

Satz 5.4.1 (Optimale Eckenlösung) *Sei $\mathcal{Z} \subset \mathbb{R}^n$ ein Polyeder mit einer Ecke, und besitze das Problem*

$$\left\{ \begin{array}{lcl} c^T x & \rightarrow & \min \\ x & \in & \mathcal{Z} \end{array} \right.$$

eine optimale Lösung. Dann existiert auch mindestens eine Ecke von $\mathcal{Z}$, welche optimal ist.

Beweis: Sei α der Optimalwert der Zielfunktion, und sei $\tilde{\mathcal{Z}}$ der Durchschnitt von $\mathcal{Z}$ und der Hyperebene $c^T x = \alpha$. Dann ist $\tilde{\mathcal{Z}}$ ein nicht-leeres Polyeder, welches keine Gerade enthält, da ja $\mathcal{Z}$ selber nach Satz 4.5.6 keine Gerade enthalten kann. Also folgt mit dem gleichen Satz die Existenz einer Ecke x^* von $\tilde{\mathcal{Z}}$, welche gleichzeitig Optimum unseres Problems ist. Für $x_1, x_2 \in \mathcal{Z}$ und $0 < \lambda < 1$ sei $x^* = \lambda x_1 + (1-\lambda) x_2$. Dann folgt $c^T x^* = \lambda (c^T x_1) + (1-\lambda)(c^T x_2)$. Weil $c^T x_j \geq c^T x^*$ gelten muss, folgt daraus aber $c^T x_1 = c^T x_2 = \alpha$. Also sind $x_1, x_2 \in \tilde{\mathcal{Z}}$, und da x^* eine Ecke von $\tilde{\mathcal{Z}}$ ist, folgt $x_1 = x_2$. Deshalb ist x^* auch eine Ecke von $\mathcal{Z}$. □

Aufgabe 5.4.2 *Gib Beispiele für c und $\mathcal{Z}$, bei denen es mehrere optimale Ecken gibt.*

Im Hinblick auf diesen Satz ist klar, dass man grundsätzlich ein Optimum für ein Optimierungsproblem berechnen kann, indem man die Werte der Zielfunktion an allen (endlich vielen) Ecken des zulässigen Bereiches vergleicht. Allerdings kann es sehr viele solcher Ecken geben, so dass eine unsystematische Suche nach einer optimalen Ecke zu aufwändig wäre. Außerdem hat die Zielfunktion an den Ecken immer irgendwo einen kleinsten Wert, auch wenn sie nach unten unbeschränkt ist. Unser Ziel ist es also, ein Kriterium zu entwickeln, das entscheidet, ob eine gegebene Ecke optimal ist, und das im anderen Fall erlaubt eine neue Ecke zu finden, an der die Zielfunktion einen echt kleineren Wert annimmt. Dabei gehen wir von folgender Lage aus:

- Gegeben ist ein Problem in der Standardform (S), und wir können o B. d. A. annehmen, dass das Gleichungssystem $A x = b$ mehr als nur eine Lösung hat, und dass $\operatorname{rang} A = m < n$ ist.

Da die Ecken des zulässigen Bereiches eine zentrale Rolle spielen zeigen wir:

Lemma 5.4.3 (Ecken eines Problems in Standardform) *Für ein Problem in Standardform sind die Ecken des zulässigen Bereichs $\mathcal{Z}$ genau die zulässigen Basislösungen des Gleichungssystems $A x = b$.*

Beweis: Der Beweis dieses Lemmas ist ähnlich zu dem von Lemma 4.5.5: Sei $x \in \mathcal{Z}$, und sei J die (eventuell leere) Menge der Indizes j mit $x_j > 0$. Falls $(a_j, j \in J)$ linear abhängig ist, dann hat der affine

Raum $\mathcal{A}$ aller $\tilde{x}$ mit $\tilde{x}_j = 0$ für $j \notin J$ und $A\,\tilde{x} = 0$ eine positive Dimension, und es gibt ein $\varepsilon > 0$ derart, dass $x + \tilde{x} \in \mathcal{Z}$ ist für alle $\tilde{x} \in \mathcal{A}$ mit $\|\tilde{x}\| < \varepsilon$. Daraus folgt, dass x keine Ecke sein kann. Sei jetzt $(a_j, j \in J)$ linear unabhängig, also $\dim \mathcal{A} = 0$, und seien $\tilde{x}, \hat{x} \in \mathcal{Z}$ so, dass $x = \alpha\,\tilde{x} + (1-\alpha)\,\hat{x}$ für ein α mit $0 < \alpha < 1$ gilt. Dann folgt für $j \notin J$ dass $\tilde{x}_j = \hat{x}_j = 0$ sein muss (da sonst einer der Werte negativ sein müsste, was wegen $\tilde{x} \geq 0$ und $\hat{x} \geq 0$ nicht sein kann). Dann sind aber $\tilde{x}, \hat{x} \in \mathcal{A}$, und wegen $\dim \mathcal{A} = 0$ bedeutet das $\tilde{x} = \hat{x} = x$: Also ist x Ecke. □

Wir benutzen im Weiteren immer folgende Bezeichungen:

- Mit $J = \{j_1 < \ldots < j_m\}$ bezeichnen wir eine beliebige m-elementige Teilmenge von $\{1, \ldots, n\}$ und schreiben $N = \{n_1 < \ldots < n_{n-m}\}$ für die komplementäre Menge.
- Seien $s_1, \ldots, s_n$ die Spalten von A, und für J und N wie oben seien $A_J = [s_{j_1}, \ldots, s_{j_m}]$ bzw. $A_N = [s_{n_1}, \ldots, s_{n_{n-m}}]$. Beachte, dass A_J immer eine quadratische, aber vielleicht nicht invertierbare Matrix ist.
- Für einen beliebigen Vektor $x \in \mathbb{R}^n$, und J und N wie oben, seien analog $x_J = (x_{j_1}, \ldots, x_{j_m})^T$ bzw. $x_N = (x_{n_1}, \ldots, x_{n_{n-m}})^T$.
- Solche J, für welche A_J invertierbar ist, heißen *Basismengen*, und der Vektor x mit $x_N = 0$ und $x_J = A_J^{-1}\,b$ heißt die *Basislösung zu J*. Somit ist x genau dann zulässige Basislösung im früher definierten Sinn, falls $x_J \geq 0$ ist, und diese sind genau die Ecken des zuässigen Bereiches. Beachte aber, dass hier, im Gegensatz zur Definition 4.4.6, die Menge J immer m Elemente hat, und dass es sein kann dass $x_j = 0$ ist, auch wenn $j \in J$ liegt; vergleiche hierzu den nächsten Begriff.
- Eine zur Basismenge J gehörige Ecke x von $\mathcal{Z}$ heißt *entartet*, falls mindestens ein $j \in J$ mit $x_j = 0$ existiert. Das bedeutet, dass es unter Umständen ein anderes $\tilde{J}$ geben kann, welches die gleiche Ecke x definiert, während im nicht entarteten Fall das x die Basismenge J festlegt.

Definition 5.4.4 (Reduzierte Kosten) *Sei $\hat{x}$ eine zu einer Basismenge J gehörige Ecke von $\mathcal{Z}$. Mit $t_N = (t_{n_1}, \ldots, t_{n_{n-m}})^T = (A_J^{-1}\,A_N)^T\,c_J$ heißt der Vektor $c_N - t_N$ auch* (der Vektor der) reduzierte(n) Kosten.

Mit Hilfe der reduzierten Kosten können wir feststellen, wann eine gegebene Ecke ein Optimum ist, aber auch wann die Zielfunktion nicht nach unten beschränkt ist:

Satz 5.4.5 (Optimalitäts- und Unlösbarkeitskriterium) *Sei $\hat{x}$ eine zu einer Basismenge J gehörige Ecke von $\mathcal{Z}$.*

(a) *Falls $c_N - t_N \geq 0$ ist, ist $\hat{x}$ Optimum von* (S).

(b) *Falls für ein $n_k \in N$ die Zahl $c_{n_k} - t_{n_k}$ negativ ist, während $A_J^{-1}\,s_{n_k} \leq 0$, dann ist die Zielfunktion nicht nach unten beschränkt, und daher existiert kein Optimum von* (S).

Beweis: Sei $x \in \mathcal{Z}$. Dann ist $A_J\,x_J + A_N\,x_N = b$, und wegen $\hat{x}_N = 0$ ist $A_J\,\hat{x}_J = b$, und daher folgt $x_J + A_J^{-1}\,A_N\,x_N = \hat{x}_J$. Also ergibt sich

$$c^T\,x \;=\; c^T\,\hat{x} \;+\; (c_N - t_N)^T\,x_N\,.$$

Falls $c_N - t_N \geq 0$ ist, folgt $c^T\,x \geq c^T\,\hat{x}$, was die Optimalität von $\hat{x}$ zeigt. Im Fall (b) ist für alle $\alpha \geq 0$ der Vektor x, gegeben durch $x_J = \hat{x}_J - \alpha\,A_J^{-1}\,s_{n_k}$ bzw. $x_N = \alpha\,e_k$, d. h., $A_N\,x_N = s_{n_k}$, im zulässigen Bereich $\mathcal{Z}$, und $c^T\,x = c^T\,\hat{x} + \alpha\,(c_{n_k} - t_{n_k})$, woraus für $\alpha \longrightarrow \infty$ die Behauptung folgt. □

Wir betrachten jetzt den Hauptfall, in dem die Voraussetzungen von (a) und (b) nicht erfüllt sind. Dazu seien die Elemente der Matrix $A_J^{-1}\,A_N$ mit $\alpha_{j_\nu n_k}$ bezeichnet, für $\nu = 1, \ldots, m$ und $k = 1, \ldots, n-m$.

Dann ist $A_J^{-1}\, s_{n_k} = (\alpha_{j_1 n_k}, \ldots, \alpha_{j_m n_k})^T$, und der Hauptfall liegt genau dann vor, wenn für ein $n_k \in N$ gilt $c_{n_k} - t_{n_k} < 0$, während wenigstens ein $\alpha_{j_\nu n_k} > 0$ ist. In diesem Fall definieren wir x wie im Beweis von Satz 5.4.5, mit

$$\alpha \;:=\; \min\left\{ \hat{x}_{j_\nu} / \alpha_{j_\nu n_k} \;:\quad 1 \le \nu \le m\,,\quad \alpha_{j_\nu n_k} > 0 \right\}. \tag{5.4.1}$$

Anders ausgedrückt wählen wir α maximal, so dass das zugehörige x gerade noch in $\mathcal{Z}$ liegt. Wie oben folgt dann $c^T x = c^T \hat{x} + \alpha\,(c_{n_k} - t_{n_k}) \le c^T \hat{x}$. Falls die Ecke $\hat{x}$ nicht degeneriert ist, ist der gewählte Wert von α positiv, und dann gilt sogar $c^T x < c^T \hat{x}$. In jedem Fall ist aber der Vektor x wieder eine Ecke von $\mathcal{Z}$, und zwar zu einer Basismenge, welche aus J entsteht, wenn man einen der Werte $j_\nu \in J$, für die $\alpha = \hat{x}_{j_\nu} / h_{j_\nu}$ gilt, durch n_k ersetzt, denn nach Wahl von j_ν folgt mit dem Basisaustauschlemma, dass die Ersetzung von s_{j_ν} durch s_{n_k} wieder zu einer Basis von $\mathbb{R}^m$ führt.

Im obigen Hauptfall haben wir gesehen, dass es möglich ist, von einer nicht optimalen Ecke aus eine andere anzusteuern, ohne dass der Wert der Zielfunktion wächst. Dabei wird nur ein Wert der Basismenge J durch eine Zahl aus N ersetzt, und deshalb heißt die neue Ecke *zur alten benachbart*. Im Fall von nicht entarteten Ecken wird der Wert der Zielfunktion beim Übergang zur Nachbarecke sogar echt kleiner, und daher ist klar dass man nach endlich vielen Schritten eine Ecke erreicht, für die einer der Fälle (a) und (b) aus Satz 5.4.5 eintritt. Unklar ist allerdings, ob man beim Auftreten von entarteten Ecken nicht auch im Kreis laufen und daher nicht zu einem Ende kommen kann. Beispiele zeigen, dass dies tatsächlich so sein kann, aber durch Anwendung einer zusätzlichen *Pivotregel von Charnes* verhindert werden kann - dies soll erst später besprochen werden.

5.5 Bestimmung einer Ausgangsecke

Bei einem Problem in Standardform haben wir gesehen, wie man ein Optimum bestimmen kann, indem man, ausgehend von einer Ecke des zulässigen Bereiches, von dieser zu einer Nachbarecke geht, für welche die Zielfunktion kleiner, oder jedenfalls nicht größer wird. *Allerdings ist im Allgemeinen unklar, ob der zulässige Bereich nicht leer ist und wie man zunächst eine Ausgangsecke finden kann.* Dies soll jetzt gekärt werden. Dazu setzen wir o. B. d. A. voraus dass $b \ge 0$ ist - sonst kann man die entsprechenden Gleichungen des Systems $A\,x = b$ mit dem Faktor -1 multiplizieren.

- Für ein Problem in Standardform betrachten wir das Hilfsproblem

$$\text{(H)} \qquad \left\{ \begin{array}{rcl} \sum y_j & \to & \min \\ A\,x + y & = & b \\ x & \ge & 0 \\ y & \ge & 0 \end{array} \right.$$

  Dies ist selber ein Problem in Standardform, aber in Dimension $2\,n$, und hier ist das Paar ($x = 0$, $y = b$) eine Ecke des zulässigen Bereichs. Die Zielfunktion ist hier durch 0 nach unten beschränkt, und wenn es ein zulässiges x für (S) gibt, dann ist das Paar $(x, 0)$ zulässig für (H), was zur Folge hat dass der Optimalwert $v(H) = 0$ ist. Ausgehend von der Ecke ($x = 0$, $y = b$) kann man mit dem Simplexalgorithmus eine Ecke ($x^* \ge 0$, $y^* \ge 0$) berechnen, welche ein Optimum für (H) ist. Falls $y^* \ne 0$ ist, kann das Ausgangsproblem (S) keinen zulässigen Punkt haben. Im anderen Fall ist x^* eine Ecke für (S), denn die zu den positiven Koordinaten von x^* gehörigen Spalten von A müssen dann linear unabhängig sein (weil sonst das Paar $(x^*, 0)$ keine Ecke von (H) gewesen wäre), und durch eventuelle Hinzunahme weiterer Spalten findet man eine Basismenge zu x^* – allerdings ist dann die Ecke entartet, was aber nicht stört.

Wir sehen also, dass ein Optimierungsproblem in Standardform folgendermaßen gelöst werden kann: In einer ersten Phase löst man das Hilfsproblem (H) und stellt fest, ob (S) überhaupt zulässige Punkte hat. Falls ja, erhält man gleichzeitig eine Ausgangsecke für (S) und kann dann von dieser aus mit dem Simplexalgorithmus entscheiden, ob (S) ein Optimum besitzt. Falls dies so ist, berechnet der Algorithmus auch eine Ecke, welche optimal ist.

5.6 Das Simplextableau und die Pivotregel

Zur schematischen Durchführung des Simplexalgorithmus benutzt man Tableaus der folgenden Gestalt:

	n_1	$\dots$	n_{n-m}	
j_1	$\alpha_{j_1 n_1}$	$\dots$	$\alpha_{j_1 n_{n-m}}$	$\hat{x}_{j_1}$
$\vdots$	$\vdots$		$\vdots$	$\vdots$
j_m	$\alpha_{j_m n_1}$	$\dots$	$\alpha_{j_m n_{n-m}}$	$\hat{x}_{j_m}$
	$t_{n_1} - c_{n_1}$	$\dots$	$t_{n_{n-m}} - c_{n_{n-m}}$	$c^T \hat{x}$

Dabei sind die Elemente α_{jk} die Einträge der Matrix $A_J^{-1} A_N$. Diesem Tableau kann man entnehmen, ob man die gesuchte optimale Ecke schon gefunden hat, ob das Problem kein Optimum besitzt, oder ob als nächstes ein Austauschschritt zu erfolgen hat. Im letzten Fall bestimmt man die sogenannten *Pivotelemente* n_k und j_ν wie oben beschrieben, und kann dann das ganze Tableau auf die neue Ecke umrechnen. Einzelheiten hierzu findet man in der Literatur.

Um bei entarteten Ecken zu verhindern, dass das Verfahren in einen Zykel hineinläuft, wählt man zunächst n_k wie üblich und wählt anschließend j_ν nach folgender Regel:

1. Falls es genau ein $j_\nu \in J$ gibt, für welches $\hat{x}_{j_\nu}/\alpha_{j_\nu n_k}$ minimal ist, nimmt man dieses.
2. Falls es mehrere $j_\nu \in J$ gibt, für welches $\hat{x}_{j_\nu}/\alpha_{j_\nu n_k}$ minimal ist, sucht man eines unter diesen, für welches $\alpha_{j_\nu n_1}/\alpha_{j_\nu n_k}$ minimal ist. Falls es auch dann noch mehrere Möglichkeiten gibt, sucht man eines unter diesen, für welches $\alpha_{j_\nu n_2}/\alpha_{j_\nu n_k}$ minimal ist, und so weiter.

Diese sogenannte lexikographische Auswahlregel verhindert, dass das Verfahren zyklisch verläuft, da man von einer Basismenge niemals zur gleichen Menge zurückkommt.

Kapitel 6

Ergänzungen

6.1 Multilinearformen

Definition 6.1.1 *Sei V ein Vektorraum über $\mathbb{K}$, und sei $n \in \mathbb{N}$. Für das n-fache kartesische Produkt von V mit sich selber schreiben wir kurz V^n. Also besteht V^n per Definition aus allen n-Tupeln $(v_1, \ldots, v_n)$ mit $v_k \in V$. Eine Abbildung $f : V^n \longrightarrow \mathbb{K}$, $(v_1, \ldots, v_n) \longmapsto f(v_1, \ldots, v_n)$ heißt dann eine* n-stellige Form *auf V. Für ein $k \in \{1, \ldots, n\}$ heißt eine solche Form* linear in der k-ten Stelle, *falls folgendes gilt:*

(M1) Für jede Wahl von $v_1, \ldots, v_n, \tilde{v_k} \in V$ gilt

$$f(v_1, \ldots, v_{k-1}, v_k + \tilde{v}_k, v_{k+1}, \ldots, v_n) \;=\; f(v_1, \ldots, v_{k-1}, v_k, v_{k+1}, \ldots, v_n) + f(v_1, \ldots, v_{k-1}, \tilde{v}_k, v_{k+1}, \ldots, v_n).$$

(M2) Für jede Wahl von $v_1, \ldots, v_n \in V$ und $\lambda \in \mathbb{K}$ gilt

$$f(v_1, \ldots, v_{k-1}, \lambda\, v_k, v_{k+1}, \ldots, v_n) \;=\; \lambda\, f(v_1, \ldots, v_{k-1}, v_k, v_{k+1}, \ldots, v_n).$$

Falls dies sogar für jedes $k = 1, \ldots, n$ gilt, falls also die Form in jeder Stelle linear ist, nennt man sie auch kurz multilinear *oder eine* Multilinearform *auf V. Folgende Spezialfälle sind besonders wichtig:*

(a) *Falls $n = 1$ ist, ist eine solche Multilinearform das gleiche wie eine lineare Abbildung von V in $\mathbb{K}$.*

(b) *Für $n = 2$ heißt eine Multilinearform auch* Bilinearform *auf V.*

Eine Form auf V heißt alternierend, *falls für alle $1 \leq j < k \leq n$ und alle $v_1, \ldots, v_n \in V$ gilt*

$$f(v_1, \ldots, v_{k-1}, v_k, v_{k+1}, \ldots, v_{j-1}, v_j, v_{j+1}, \ldots, v_n) \;=\; -f(v_1, \ldots, v_{k-1}, v_j, v_{k+1}, \ldots, v_{j-1}, v_k, v_{j+1}, \ldots, v_n).$$

In Worten heißt das, dass die Form beim Vertauschen von zwei Stellen das Vorzeichen wechselt. Da eine beliebige Permutation ein Produkt von endlich vielen Transpositionen ist, folgt sofort für jede alternierende Form, dass für jede Permutation $\sigma \in S_n$ und alle $v_1, \ldots, v_n \in V$ gilt

$$f(v_{\sigma(1)}, \ldots, v_{\sigma(n)}) \;=\; \operatorname{sgn}(\sigma)\, f(v_1, \ldots, v_n)\,.$$

Wenn eine Form beim Vertauschen ihrer Stellen unverändert bleibt, spricht man auch von einer symmetrischen Form.

Aufgabe 6.1.2 *Zeige: Für jede Multilinearform ist $f(v_1, \ldots, v_n) = 0$, falls mindestens ein v_j der Nullvektor ist.*

Aufgabe 6.1.3 *Zeige: Für jede alternierende Form ist $f(v_1,\dots,v_n)=0$, falls mindestens ein Paar (j,k) mit $j\neq k$ aber $v_j=v_k$ existiert.*

Satz 6.1.4 *Sei $n\in\mathbb{N}$, und sei V ein n-dimensionaler Vektorraum über $\mathbb{K}$, sowie $(v_1^{(0)},\dots,v_n^{(0)})$ eine Basis von V. Dann gibt es auf V genau eine n-stellige alternierende Multilinearform f mit der Normierungseigenschaft*

$$f(v_1^{(0)},\dots,v_n^{(0)})=1\,.$$

Für beliebige $v_1,\dots,v_n\in V$ sei $A=[a_{jk}]$ durch $v_k=\sum_{j=1}^n a_{jk}\,v_j^{(0)}$ definiert. Dann gilt

$$f(v_1,\dots,v_n)\;=\;\det A\,.$$

Beweis: Definiert man A wie oben, so folgt für jede Multilinearform f:

$$f(v_1,\dots,v_n)\;=\;\sum_{k_1,\dots,k_n} a_{k_1 1}\cdot\ldots\cdot a_{k_n n}\;f(v_{k_1}^{(0)},\dots,v_{k_n}^{(0)})\;,$$

wobei alle Indizes $k_1,\dots,k_n$ voneinander unabhängig von 1 bis n laufen. Wenn die Form alternierend ist, folgt aus Aufgabe 6.1.3 dass $f(v_{k_1}^{(0)},\dots,v_{k_n}^{(0)})=0$ ist, außer wenn das Tupel $(k_1,\dots,k_n)$ eine Permutation der Zahlen $1,\dots,n$ ist. Für $\sigma\in S_n$ folgt $f(v_{\sigma(1)}^{(0)},\dots,v_{\sigma(n)}^{(0)})=\operatorname{sgn}(\sigma)\;f(v_1^{(0)},\dots,v_n^{(0)})$, und wenn die Form die Normierungsbedingung erfüllt, folgt hieraus mit der Definition der Determinante $f(v_1,\dots,v_n)=\det A^T=\det A$. Wenn man umgekehrt $f(v_1,\dots,v_n)=\det A$ definiert, dann folgt aus den Rechenregeln für Determinanten, dass f in der Tat multilinear ist. □

6.2 Hermitesche Formen

Definition 6.2.1 *Eine zweistellige Form h auf einem Vektorraum V über $\mathbb{K}$ heißt eine* hermitesche Form, *falls sie in der zweiten Stelle linear ist, und falls zusätzlich gilt*

$$\forall\; v_1,v_2\in V:\qquad h(v_1,v_2)=\overline{h(v_2,v_1)}\,. \tag{6.2.1}$$

Aus diesen beiden Bedingungen folgt dann, dass die Form in der ersten Stelle semilinear ist. Falls V ein Vektorraum über $\mathbb{K}=\mathbb{R}$ ist, ist $h(v_1,v_2)$ immer eine reelle Zahl. Deshalb ist (6.2.1) *dann gleichwertig mit*

$$\forall\; v_1,v_2\in V:\qquad h(v_1,v_2)=h(v_2,v_1)\,.$$

Daher ist die Form im oben definierten Sinn symmetrisch, und man schreibt dann meist auch s an Stelle von h. Also sind für reelle Vektorräume hermitesche Formen das gleiche wie symmetrische Bilinearformen. Für jede hermitesche Form auf V heißt die Abbildung $q:\;V\longrightarrow\mathbb{K}$ mit $q(v)=h(v,v)$ für alle $v\in V$ auch die zugehörige quadratische Form. *Die quadratische Form, oder wahlweise die hermitesche Form, heißt*

(a) *positiv definit*, falls $q(v)>0$ gilt für alle $v\in V\setminus\{0\}$,

(b) *negativ definit*, falls $q(v)<0$ gilt für alle $v\in V\setminus\{0\}$,

(c) *positiv semidefinit*, falls $q(v)\geq 0$ gilt für alle $v\in V$,

(d) *negativ semidefinit*, falls $q(v)\leq 0$ gilt für alle $v\in V$,

(e) *indefinit*, falls $q(v_1)>0$ gilt für mindestens ein $v_1\in V$ und $q(v_2)<0$ für mindestens ein (anderes) $v_2\in V$.

Beachte, dass für eine beliebige quadratische Form immer mindestens einer dieser Fälle eintritt.

Beispiel 6.2.2 *Ein inneres Produkt ist eine positiv definite hermitesche Form auf einem Vektorraum. Ist V ein Vektorraum mit innerem Produkt, und ist $T \in L(V)$ selbstadjungiert, so ist die Abbildung $(v_1, v_2) \longmapsto \langle v_1, T\, v_2\rangle$ eine hermitesche Form auf V.*

Satz 6.2.3 *Sei $n \in \mathbb{N}$, und sei V ein n-dimensionaler Vektorraum über $\mathbb{K}$, sowie $(v_1, \ldots, v_n)$ eine Basis von V. Zu einer hermiteschen Form h auf V sei $A = [a_{jk}] \in \mathbb{K}^{n\times n}$ definiert durch*

$$h(v_j, v_k) \;=\; a_{jk} \qquad \forall\; j,k = 1, \ldots, n\,. \tag{6.2.2}$$

Dann ist A eine hermitesche Matrix, und für $v, \tilde{v} \in V$ seien $x = (x_1, \ldots, x_n)^T$ und $y = (y_1, \ldots, y_n)^T$ so, dass $v = \sum_{j=1}^n x_j\, v_j$ und $\tilde{v} = \sum_{k=1}^n y_k\, v_k$. Dann gilt

$$h(v, \tilde{v}) \;=\; \overline{x}^T\, A\, y\,.$$

Beweis: Wenn (6.2.2) gilt, folgt sofort mit der Definition einer hermiteschen Form $\overline{A}^T = A$, und

$$h(v, \tilde{v}) \;=\; \sum_{k=1}^n \sum_{j=1}^n \overline{x_j}\, y_k\, a_{jk} \;=\; \overline{x}^T\, A\, y\,,$$

und das war zu zeigen. □

Definition 6.2.4 *Sei V ein n-dimensionaler Vektorraum V über $\mathbb{K}$ mit Basis $(v_1, \ldots, v_n)$, und sei h eine hermitesche Form auf V. Dann heißt die durch* (6.2.2) *definierte Matrix A* Strukturmatrix *von h bezüglich dieser Basis.*

Aufgabe 6.2.5 *Gegeben sei ein beliebiges Skalarprodukt $\langle\,\cdot\,,\,\cdot\,\rangle$ auf $\mathbb{C}^n$. Zeige: Es gibt eine positiv definite Matrix A, für welche*

$$\langle x, y\rangle \;=\; \overline{x}^T\, A\, y \qquad \forall\; x, y \in \mathbb{C}^n\,.$$

Wie sieht A für das kanonische Skalarprodukt aus?

Satz 6.2.6 *Sei $n \in \mathbb{N}$, sei V ein n-dimensionaler Vektorraum über $\mathbb{K}$, sowie $(v_1, \ldots, v_n)$ und $(\tilde{v}_1, \ldots, \tilde{v}_n)$ zwei Basen von V. Weiter sei $B = [b_{jk}] \in \mathbb{K}^{n\times n}$ die zugehörige Umrechnungsmatrix. Schließlich sei h eine hermitesche Form auf V, und A bzw. $\tilde{A}$ seien ihre Strukturmatrizen bezüglich der Basen $(v_1, \ldots, v_n)$ bzw. $(\tilde{v}_1, \ldots, \tilde{v}_n)$. Dann gilt*

$$\tilde{A} \;=\; \overline{B}^T\, A\, B\,.$$

Beweis: Für $1 \le j, k \le n$ folgt aus den Eigenschaften von h

$$\tilde{a}_{jk} \;=\; h(\tilde{v}_j, \tilde{v}_k) \;=\; \sum_{\nu=1}^n \sum_{\mu=1}^n \overline{b_{\nu j}}\, a_{\nu\mu}\, b_{\mu k}\,,$$

und das ist die Behauptung. □

Definition 6.2.7 *Zwei Matrizen $A, B \in \mathbb{K}^{n\times n}$ heißen* kongruent, *falls es eine invertierbare Matrix $C \in \mathbb{K}^{n\times n}$ gibt, für die*

$$B \;=\; \overline{C}^T\, A\, C\,.$$

Falls C sogar unitär ist, dann sind A und B unitär ähnlich zueinander, aber allgemein müssen kongruente Matrizen keinesfalls ähnlich sein. In jedem Fall sind kongruente Matrizen Strukturmatrizen derselben hermiteschen Form zu unterschiedlichen Basen eines Raumes.

Aufgabe 6.2.8 *Zeige, dass kongruente Matrizen denselben Rang haben.*

Aufgabe 6.2.9 *Zeige, dass die Matrix*

$$A \;=\; \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$$

zur Einheitsmatrix kongruent, aber nicht zu ihr ähnlich ist. Zeige allgemeiner: Ist A eine Diagonalmatrix mit reellen Zahlen $\lambda_1, \ldots, \lambda_n$ entlang der Diagonalen, so ist A kongruent zu einer Matrix B, auf deren Diagonalen nur die Zahlen 0, 1 und -1 auftreten.

Satz 6.2.10 (Sylvesterscher Trägheitssatz) *Sei $n \in \mathbb{N}$, sei V ein n-dimensionaler Vektorraum über $\mathbb{K}$, sowie h eine hermitesche Form auf V. Dann gibt es eine Basis $(v_1, \ldots, v_n)$ von V, bezüglich der die Strukturmatrix A von h eine Diagonalmatrix ist, auf deren Diagonalen, in dieser Reihenfolge, q-mal eine 0, r-mal eine 1 und s-mal eine -1 steht, wobei eine oder mehrere dieser Anzahlen auch $= 0$ sein dürfen. Die Zahlen q, r, s sind dabei nur von h, nicht aber von der Wahl der Basis abhängig.*

Beweis: Sei zunächst irgendeine Basis von V gewählt, und sei A die entsprechende Strukturmatrix von h. Nach dem Satz von der Hauptachsentransformation ist A unitär ähnlich zu einer Diagonalmatrix, und diese ist wiederum Darstellungsmatrix von h zu einer entsprechenden Basis von V. Sei deshalb jetzt A bereits als diagonal vorausgesetzt. Die Diagonalelemente, also die Nullstellen des charakteristischen Polynoms von A, sind in $\mathbb{R}$, und q bzw. r bzw. s sei die Zahl derjenigen Eigenwerte, welche $= 0$ bzw. positiv bzw. negativ sind. Durch evtl. Umnummerierung der Basisvektoren können wir erreichen, dass zuerst die Nullen, dann die positiven und schließlich die negativen Eigenwerte entlang der Diagonalen stehen. Nach Aufgabe 6.2.9 wiederum ist eine solche Matrix kongruent zu einem A wie im Satz. Sei jetzt auch B eine Diagonalmatrix der im Satz vorkommenden Art, wobei die Anzahl der 0-en, 1-en und (-1)-en auf der Diagonalen von B mit $\tilde{q}$, $\tilde{r}$ und $\tilde{s}$ bezeichnet seien. Sind A und B kongruent, so haben sie nach Aufgabe 6.2.8 insbesondere denselben Rang, und deshalb folgt $\tilde{q} = q$. Aus der Form von A folgt, dass das System $(v_1, \ldots, v_q)$ aus der zugehörigen Basis von V einen Unterraum U aufspannen, auf dem die Form h nur den Wert 0 annimmt. Analog spannt $(v_{q+1}, \ldots, v_{q+r})$ einen Raum U^+ auf, auf dem h positiv definit ist, und auf $U^- = \mathcal{L}(v_{q+r+1}, \ldots, v_n)$ ist h negativ definit. Offenbar ist $V = U \oplus U^+ \oplus U^-$. Analog gibt es zu B einen Raum $\tilde{U}^+$ der Dimension $\tilde{r}$, der von den entsprechenden Vektoren einer anderen Basis von V aufgespannt ist, und auf diesem ist h ebenfalls positiv definit. Wäre $\tilde{r} = \dim \tilde{U}^+ > r$, so würde aus den Resultaten für direkte Summen folgen, dass $U = \tilde{U}^+ \cap \left(U \oplus U^-\right)$ ein Unterraum positiver Dimension wäre, was ein Widerspruch zur positiven Definitheit von h auf $\tilde{U}^+$ wäre, denn auf $U \oplus U^-$ ist h negativ semidefinit. Also muss $\tilde{r} \leq r$ gelten, und durch Vertauschung der Bezeichungen ergibt sich dann sogar die Gleichheit. Damit folgt aber auch $\tilde{s} = s$, da die Summe der drei Anzahlen in jedem Fall n ergeben muss. □

6.3 Der Rayleigh-Quotient

Definition 6.3.1 *Sei eine Matrix $A \in \mathbb{K}^{n \times n}$ gegeben. Die durch*

$$R_A(x) \;=\; \frac{\overline{x}^T A\, x}{\|x\|^2} \qquad \forall\; x \in \mathbb{K}^n \setminus \{0\}$$

definierte Abbildung heißt der Rayleigh-Quotient *für A.*

Aufgabe 6.3.2 *Zeige für $A \in \mathbb{K}^{n \times n}$, $x \in \mathbb{K}^n$ und $\alpha > 0$, dass immer $R_A(\alpha\, x) = R_A(x)$ gilt. Schließe hieraus, dass es bei Untersuchung der Wertemenge des Rayleigh-Quotienten genügt, Vektoren x auf der sogenannten* Einheitssphäre $S_n = \{x \in \mathbb{K}^n :\; \|x\| = 1\}$ *zu untersuchen.*

Wir wollen nun für eine hermiteschen Matrix $A \in \mathbb{K}^{n\times n}$ einen wichtigen Zusammenhang zwischen ihren Eigenwerten und den Werten des Rayleigh-Quotienten herleiten. Nach dem Satz über die Hauptachsentransformation gibt es zu A eine Orthonormalbasis $x_1, \ldots, x_n$ von $\mathbb{K}^n$ aus Eigenvektoren von A, und wir numerieren die Basisvektoren so, dass die (reellen) Eigenwerte die Bedingung

$$\lambda_1 \ \leq \ \lambda_2 \ \leq , \ldots , \ \leq \ \lambda_n \tag{6.3.1}$$

erfüllen. Mit diesen Bezeichnungen gilt dann:

Satz 6.3.3 (Rayleighsches Prinzip) *Sei $A \in \mathbb{K}^{n\times n}$ eine hermitesche Matrix. Wir wählen eine Orthonormalbasis $x_1, \ldots, x_n$ wie oben beschrieben, und setzen*

$$U_j^- \ = \ \mathcal{L}(x_1, \ldots, x_j)\,, \qquad U_j^+ = \mathcal{L}(x_j, \ldots, x_n) \qquad \forall\ j = 1, \ldots, n$$

wobei offenbar $U_1^+ = U_n^- = \mathbb{K}^n$ ist. Dann gilt für $j = 1, \ldots, n$

$$\lambda_j \ = \ R_A(x_j) \ = \ \min\{R_A(x) : \ x \in U_j^+ \setminus \{0\}\} \ = \ \max\{R_A(x) : \ x \in U_j^- \setminus \{0\}\}\,.$$

Beweis: Sei $j \in \{1, \ldots, n\}$. Für $x \in U_j^+$ gilt dann $u = \sum_{k=j}^n \alpha_k\, x_k$, mit $\alpha_k \in \mathbb{K}$, und es gilt $\|u\|^2 = \sum_{k=j}^n |\alpha_k|^2$. Wir können wegen Aufgabe 6.3.2 annehmen, dass $\sum_{k=j}^n |\alpha_k|^2 = 1$ ist, und finden mit den Regeln der Matrixmultiplikation $R_A(x) = \sum_{k=j}^n \lambda_k\, |\alpha_k|^2$. Wegen der Anordnung der Eigenvektoren bzw. Eigenwerte folgt daraus sofort

$$R_A(x) \ = \ \sum_{k=j}^{n} \lambda_k\, |\alpha_k|^2 \ \geq \ \lambda_j \ \sum_{k=j}^{n} |\alpha_k|^2 \ = \ \lambda_j\,,$$

wobei Gleichheit offenbar für $u = x_j$ eintritt. Ganz analog zeigt man die andere Ungleichung. □

Bemerkung 6.3.4 (Praktische Anwendung des Rayleigh-Prinzips)
Meist kennt man weder die Eigenvektoren noch, vor allem, die Eigenwerte einer hermiteschen Matrix $A \in \mathbb{K}^{n\times n}$. Man kann aber den obigen Satz, wenigstens theoretisch, zur Berechnung der Eigenwerte und Eigenvektoren einsetzen. Dabei ist es wichtig zu beachten, dass $U_{j+1}^+ = (U_j^-)^\perp$ ist.

(a) Berechne $\lambda_1 = \min\{R_A(x) : \ x \in \mathbb{K}^n \setminus 0\} = \min\{R_A(x) : \ x \in S^n\}$. Dieser Wert ist nach obigem Satz gleich dem kleinsten Eigenwert von A.

(b) Finde eine Orthonormalbasis $(x_1, \ldots, x_s)$ des Eigenraums von A zum Eigenwert λ_1. Da A diagonalisierbar ist, ist die geometrische Vielfachheit von λ_1 gleich der algebraischen Vielfachheit, weshalb wir schließen, dass in der (unbekannten) Liste der Eigenwerte $\lambda_1, \ldots, \lambda_n$ gelten muss: $\lambda_1 = \ldots = \lambda_s$.

(c) Wir können jetzt annehmen, dass wir für ein $m \in \{1, \ldots, n\}$ bereits ein Orthonormalsystem $(x_1, \ldots, x_m)$ von Eigenvektoren für A berechnet haben, wobei die zugehörigen Eigenwerte $\lambda_1, \ldots, \lambda_m$ monoton wachsend sind. Falls $m = n$ ist, sind wir fertig. Sonst berechne $U_{m+1}^+ = (U_m^-)^\perp$ sowie $\lambda_{m+1} = \min\{R_A(x) : x \in U_{m+1}^+ \setminus \{0\}\} = \min\{R_A(x) : x \in U_{m+1}^+ \cap S_n\}$. Damit kennen wir nach dem Rayleigh-Prinzip auch den nächst-größeren Eigenwert von A.

(d) Berechne jetzt eine Orthonormalbasis $x_{m+1}, \ldots, x_{m+s}$ des Eigenraumes von A zum Eigenwert λ_{m+1} und schließe wie vorher, dass dann $\lambda_{m+1} = \ldots = \lambda_{m+s}$ ist. Wiederhole nun Schritt (c) mit $m + s$ an Stelle von m.

Auf diese Weise finden wir in endlich vielen Schritten alle Eigenvektoren und Eigenwerte von A. Man kann natürlich auch zuerst den größten der Eigenwerte als das Maximum des Rayleigh-Quotienten berechnen und danach den zweitgrößten, u. s. w. Die Berechnung der Eigenräume geschieht durch Lösen des entsprechenden linearen Gleichungssystems, aber die Hauptschwierigkeit besteht natürlich im Berechnen der entsprechenden Minima bzw. Maxima. Darauf soll hier nicht weiter eingegangen werden.

Literaturverzeichnis

[1] **L. Collatz und W. Wetterling**, *Optimierungsaufgaben*, Springer-Verlag, Berlin, 1971. Zweite Auflage, Heidelberger Taschenbücher, Band 15.

[2] **F. R. Gantmacher**, *Matrizentheorie*, Springer-Verlag, Berlin, 1986. With an appendix by V. B. Lidskij, With a preface by D. P. Želobenko, Translated from the second Russian edition by Helmut Boseck, Dietmar Soyka and Klaus Stengert.

[3] **P. Kall**, *Mathematische Methoden des Operations Research*, B. G. Teubner, Stuttgart, 1976. Eine Einführung, Leitfäden der angewandten Mathematik und Mechanik, Band 27.

[4] **F. Lorenz**, *Lineare Algebra. II*, Bibliographisches Institut, Mannheim, 2. Aufl., 1989.

[5] **G. Strang**, *Linear algebra and its applications*, Academic Press [Harcourt Brace Jovanovich Publishers], New York, zweite Aufl., 1980.

Index