Skript zur Vorlesung

Geometrie linearer Matrixungleichungen

Universität Konstanz

Sommersemester 2024

Fachbereich Mathematik und Statistik

geTEXt von

Tobias Retzlaff

Stand: 1. November 2024

Inhaltsverzeichnis

§1 Allgemeine Fakten zu hyperbolischen Formen

1.1 Wiederholung (siehe Konvexität, II.6)**.** Sei $e \in \mathbb{R}^n$. Eine Form $f \in \mathbb{R}[\mathbf{x}] = \mathbb{R}[x_1, \dots, x_n]$ heißt *hyperbolisch bzgl. e*, falls $f(e) \neq 0$ gilt und für jedes $u \in \mathbb{R}^n$ das univariate Polynom $f(te-u) \in \mathbb{R}[t]$ nur reelle Nullstellen besitzt. Sind für alle $u \in \mathbb{R}^n \setminus \mathbb{R}e$ die Nullstellen von $f(te-u)$ sogar reell und einfach, so heißt f *strikt hyperbolisch bzgl. e*. (Beachte, dass man $u \notin \mathbb{R}e$ fordern muss, denn für $u = ce$ mit $c \in \mathbb{R}$ gilt $f(te-u) = f((t-c)e) = (t-c)^{\deg(f)} f(e)$.) Schreibe $\mathcal{H}_d(e)$ (bzw. $\mathcal{H}s_d(e)$) für die Menge aller bzgl. e hyperbolischen Formen (bzw. strikt hyperbolischen Formen) $f \in \mathbb{R}[\mathbf{x}]_d$. Es heißen

$$U_e(f) := \{u \in \mathbb{R}^n \colon \forall x \in [e,u] \colon f(x) \neq 0\}$$

der *offene Hyperbolizitätskegel* von $f \in \mathcal{H}_d(e)$ bzgl. e und

$$C_e(f) := \overline{U_e(f)} = \{u \in \mathbb{R}^n \colon \forall x \in [e,u) \colon f(x) \neq 0\}$$

der *abgeschlossene Hyperbolizitätskegel* (oder schlicht *Hyperbolizitätskegel*) von f bzgl. e. Es gelten

$$\begin{aligned} U_e(f) &= \{u \in \mathbb{R}^n \colon \text{Alle Nullstellen von } f(te-u) \in \mathbb{R}[t] \text{ sind } > 0\} \quad \text{und} \\ C_e(f) &= \{u \in \mathbb{R}^n \colon \text{Alle Nullstellen von } f(te-u) \in \mathbb{R}[t] \text{ sind } \geq 0\}. \end{aligned}$$

Weiter ist $U_e(f)$ auch die Zusammenhangskomponente von e in $\{u \in \mathbb{R}^n \colon f(u) \neq 0\}$.

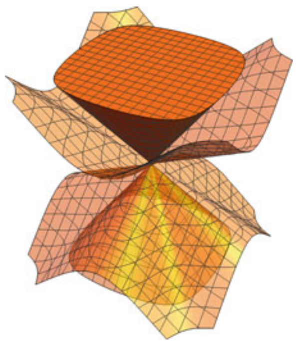


Abbildung 1: Der Hyperbolizitätskegel eines quartischen Polynoms (Quelle: [1], Fig. 2.6)

Verwende die Notationen $(\partial_e f)(x) = \frac{d}{dt} f(x+te)\big|_{t=0}$ und $\partial_e = \sum_{i=1}^n e_i \frac{\partial}{\partial x_i}$, dann gelten

$$\begin{aligned} U_e(f) &= \left\{u \in \mathbb{R}^n \colon \forall k \in [d] \colon \left(\partial_e^k f\right)(u) > 0\right\} \quad \text{und} \\ C_e(f) &= \left\{u \in \mathbb{R}^n \colon \forall k \in [d] \colon \left(\partial_e^k f\right)(u) \geq 0\right\} \end{aligned}$$

mit $d := \deg(f)$. Ist f hyperbolisch bzgl. e, so auch $\partial_e f$ und es gelten $U_e(f) \subseteq U_e(\partial_e f)$, sowie $C_e(f) \subseteq C_e(\partial_e f)$. Weiter ist f in diesem Fall hyperbolisch bzgl. jedem $u \in U_e(f)$.

1.2 Standardbeispiel. Seien $A_1, \dots, A_n \in \mathsf{S}^d = \mathrm{Sym}_d(\mathbb{R})$ und

$$f(x) := \det A(x) \quad \text{mit} \quad A(x) := x_1 A_1 + \dots + x_n A_n.$$

Die Form f hat den Grad d und ist hyperbolisch bzgl. jedem $e \in \mathbb{R}^n$ mit $A(e) = \sum_{i=1}^n e_i A_i \succ 0$, denn nach einem Basiswechsel kann man $A(e) = I_d$ annehmen, woraus

$$f(te-u) = \det A(te-u) = \det(tI_d - A(u)) = \mathrm{charpol}_{A(u)}(t)$$

für alle $u \in \mathbb{R}^n$ folgt und dieses Polynom besitzt (aufgrund der Symmetrie der A_i) nur reelle

Nullstellen. Der Hyperbolizitätskegel von f ist $C_e(f) = \{u \in \mathbb{R}^n : A(u) \succeq 0\}$.
Wir sagen, dass eine Form $f \in \mathbb{R}[\mathbf{x}]_d$ eine *definite Determinantendarstellung bzgl.* $e \in \mathbb{R}^n$ besitzt, falls ein lineares (symmetrisches) Matrixpolynom $A(x) = \sum_{i=1}^n x_i A_i$ mit $A_i \in \mathsf{S}^d$, $f(x) = \det A(x)$ und $A(e) \succ 0$ existiert. In diesem Fall ist f hyperbolisch bzgl. e und für den Hyperbolizitätskegel gilt $C_e(f) = \{u \in \mathbb{R}^n : A(u) \succeq 0\}$.

Die Lax[1]-Vermutung (1958) fragt nach der Umkehrung für $n = 3$:

1.3 Theorem (Helton[2]-Vinnikov[3] 2006)**.** Sei $f \in \mathbb{R}[x_1, x_2, x_3]$ hyperbolisch bzgl. $e \in \mathbb{R}^3$ mit $f(e) > 0$ und sei $d := \deg(f)$. Dann gibt es $A_1, A_2, A_3 \in \mathsf{S}^d$ mit $f(x) = \det(x_1 A_1 + x_2 A_2 + x_3 A_3)$ und $e_1 A_1 + e_2 A_2 + e_3 A_3 \succ 0$.

1.4 Für $n \geq 4$ kann diese Aussage im Allgemeinen nicht gelten. Beispielsweise ist die quadratische Form $f(x) := x_1^2 - x_2^2 - \cdots - x_n^2$ für jedes $n \in \mathbb{N}$ hyperbolisch bzgl. $e_1 = (1, 0, \ldots, 0)^\top \in \mathbb{R}^n$ und hat Rang $n \geq 4$. Aber für beliebige Linearformen ℓ_1, ℓ_2, ℓ_3 hat das Polynom

$$\det \begin{pmatrix} \ell_1(x) & \ell_2(x) \\ \ell_2(x) & \ell_3(x) \end{pmatrix} = \ell_1(x)\ell_3(x) - \ell_2(x)^2$$

in jedem Fall höchstens Rang 3.

1.5 Verallgemeinerte Lax-Vermutung. Ist $f \in \mathbb{R}[\mathbf{x}]_d$ hyperbolisch bzgl. $e \in \mathbb{R}^n$, so ist $C_e(f)$ ein Spektraederkegel (vgl. Aufgabe 4).

1.6 Lokale Vielfachheiten von Hyperflächen. Sei k ein Körper mit $\mathrm{char}(k) = 0$, sei $0 \neq f \in k[\mathbf{x}]$ und sei $X := \mathcal{V}(f) \subseteq \mathbb{A}^n$ (eine Hyperfläche). Für $u \in k^n$ heißt

$$\mu(f, u) := \mu(X, u) := \max\left\{i \in \mathbb{N}_0 : f \in \mathfrak{m}_u^i\right\} \geq 0$$

die *Multiplizität* (oder *Vielfachheit*) von f (oder X) in u. Dabei sei $\mathfrak{m}_u := \langle x_1 - u_1, \ldots, x_n - u_n \rangle$ (ein maximales Ideal in $k[\mathbf{x}]$). Ist

$$f(u + x) = f_m(x) + f_{m+1}(x) + \cdots \tag{T}$$

mit homogenen $f_i(x)$ vom Grad i und $f_m(x) \not\equiv 0$, so ist $m = \mu(f, u)$. Wir sehen also

$$\begin{aligned} &\mu(f, u) = 0 \iff f(u) \neq 0 \iff u \notin X, \\ &\mu(f, u) = 1 \iff f(u) = 0 \wedge \nabla f(u) \neq 0 \iff u \text{ ist ein glatter Punkt von } X. \end{aligned}$$

Jede Gerade durch u schneidet X in u mindestens mit Vielfachheit $m = \mu(f, u)$, d.h. für $0 \neq v \in \mathbb{R}^n$ ist $\mathrm{ord}_{t=0} f(tv + u) \geq m$, denn aus (T) folgt

$$f(tv + u) = f_m(tv) + f_{m+1}(tv) + \cdots = t^m f_m(v) + t^{m+1} f_{m+1}(v) + \cdots .$$

Für alle v in einer Zariski-offenen (dichten) Menge ist $\mathrm{ord}_{t=0} f(tv + u) = m = \mu(f, u)$.

1.7 Beispiel. Für $f := x_1^2 - x_2^2 - x_1^3$ und $u := 0 \in \mathbb{R}^2$ ist $\mu(f, u) = 2$ und $f(tv) = t^2(v_1^2 - v_2^2) - t^3 v_1^3$.

1.8 Satz. Sei $f \in \mathbb{R}[\mathbf{x}]$ eine bzgl. $e \in \mathbb{R}^n$ hyperbolische Form. Dann ist $\mu(f, u) = \mathrm{ord}_{t=0} f(te - u)$ für jeden Punkt $u \in \mathbb{R}^n$.

---

[1] Peter Lax (*1926)
[2] John William "Bill" Helton (*1945)
[3] Victor Vinnikov (*1967)

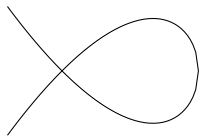


Abbildung 2: Die Hyperfläche $\mathcal{V}(f) \cap \mathbb{R}^2$ aus Beispiel 1.7 (Quelle: [2], 6.3.16)

Beweis. Wir verwenden Theorem 1.3. Man kann $f(e) = 1$ annehmen. Sei $d \coloneqq \deg(f)$, fixiere $0 \neq u \in \mathbb{R}^n$ und wähle ein $v \in U_e(f)$ mit $\mu(f,u) = \operatorname{ord}_{t=0} f(tv-u)$ (ein solches existiert, da $U_e(f) \neq \varnothing$ offen ist). Das Polynom $F(r,s,t) \coloneqq f(re+su+tv)$ erfüllt $\deg(F) = d$ und ist hyperbolisch bzgl. $e_1 \in \mathbb{R}^3$: Für $w \in \mathbb{R}^3$ ist $F(te_1 + w) = F(t+w_1, w_2, w_3) = f(te + \widetilde{w})$ mit $\widetilde{w} \coloneqq w_1 e + w_2 u + w_3 v \in \mathbb{R}^3$. Nach Theorem 1.3 gibt es Matrizen $U, V \in \mathsf{S}^d$ mit

$$F(r,s,t) = \det\big(\underbrace{rI_d + sU + tV}_{=:\,A(r,s,t)}\big).$$

Wegen $v \in U_e(f)$ ist $f \neq 0$ auf $[e,v]$, also $F \neq 0$ auf $[e_1,e_3] \subseteq \mathbb{R}^3$. Da $A(r,s,t)$ auf $[e_1,e_3]$ nirgends singulär ist, folgt mit einem Stetigkeitsargument $V \succ 0$. Nach Aufgabe 3 ist

$$\operatorname{ord}_{t=0} f(tv-u) = \operatorname{ord}_{t=0} \det(tV-U) \overset{\text{A3}}{=} \dim\ker(U).$$

Andererseits ist ebenfalls $\operatorname{ord}_{t=0} f(te-u) = \operatorname{ord}_{t=0}(tI_d - U) = \dim\ker(U)$. Also gilt nach Wahl von v bereits $\mu(f,u) = \operatorname{ord}_{t=0} f(tv-u) = \operatorname{ord}_{t=0} f(te-u)$. $\square$

1.9 Korollar. Sei $f \in \mathbb{R}[\mathbf{x}]$ hyperbolisch bzgl. $e \in \mathbb{R}^n$ und sei $u \in \mathbb{R}^n$ mit $f(u) = 0$. Dann ist u genau dann ein glatter Punkt von $\mathcal{V}(f)$, wenn $\operatorname{ord}_{t=0} f(te-u) = 1$ gilt.

Beweis. Es gilt u ist glatt $\overset{1.6}{\iff} \mu(f,u) = 1 \overset{1.8}{\iff} \operatorname{ord}_{t=0} f(te-u) = 1$. $\square$

1.10 Korollar. Sei $f \in \mathbb{R}[\mathbf{x}]$ hyperbolisch bzgl. $e \in \mathbb{R}^n$. Dann sind äquivalent:

(i) f ist strikt hyperbolisch bzgl. e.

(ii) f ist strikt hyperbolisch bzgl. jedem $u \in U_e(f)$.

(iii) Die Hyperfläche $\mathcal{V}(f) \subseteq \mathbb{A}^n$ hat keine reellen singulären Punkte $\neq 0$.

Beweis. (i) besagt, dass die Nullstellen von $f(te-v) \in \mathbb{R}[t]$ für alle $v \in \mathbb{R}^n \setminus \mathbb{R}e$ reell und einfach sind. Daraus folgt (iii), denn ist $f(v) = 0$ mit $v \neq 0$, so ist $\mu(f,v) = \operatorname{ord}_{t=0} f(te-v) = 1$. Der Beweis von (iii) $\Rightarrow$ (ii) erfolgt analog und (ii) $\Rightarrow$ (i) ist trivial. $\square$

1.11 Theorem (Nuij[4]). Seien $d \geq 1$ und $0 \neq e \in \mathbb{R}^n$ und setze $\mathcal{H}_d^1(e) \coloneqq \{f \in \mathcal{H}_d(e) \colon f(e) = 1\}$.

(a) $\mathcal{H}s_d(e)$ ist offen in $\mathbb{R}[\mathbf{x}]_d$ und dicht in $\mathcal{H}_d(e)$ und $\mathcal{H}_d^1(e)$ ist abgeschlossen in $\mathbb{R}[\mathbf{x}]_d$.

(b) Die Räume $\mathcal{H}_d^+(e) \coloneqq \{f \in \mathcal{H}_d(e) \colon f(e) > 0\}$ und $\mathcal{H}s_d^+(e) \coloneqq \mathcal{H}s_d(e) \cap \mathcal{H}_d^+(e)$ sind kontrahierbar, also insbesondere zusammenhängend und einfach zusammenhängend.

Dabei heißt ein topologischer Raum X *kontrahierbar*, falls es eine *Deformationsretraktion* von X auf einen Punkt $x_0 \in X$ gibt, d.h. eine stetige Abbildung $h \colon X \times [0,1] \to X$ mit $h(x,0) = x$ und $h(x,1) = x_0$ für alle $x \in X$. Alle Homotopiegruppen (insbesondere die Fundamentalgruppe) eines solchen Raums sind trivial.

1.12 Lemma. $p \in \mathbb{R}[t]$ habe nur reelle Nullstellen. Sei $s \in \mathbb{R}$ und $g(t) \coloneqq p(t) + sp'(t)$.

[4] Wim Nuij (*19??)

(a) Alle Nullstellen von g sind reell.

(b) Ist $s \neq 0$ und ist $k \geq 1$ die maximale Vielfachheit einer Nullstelle von $p(t)$, so ist $\max\{k-1, 1\}$ die maximale Vielfachheit einer Nullstelle von $g(t)$.

Beweis. Seien $\alpha_1 < \cdots < \alpha_n$ die verschiedenen Nullstellen von p und sei $m_i \geq 1$ die Vielfachheit von α_i, dann ist $\deg(p) = \sum_{i=1}^n m_i =: m$, also $\deg(g) = m$. Jedes α_i ist eine Nullstelle von g der Vielfachheit $m_i - 1$. Für $i \in [n]$ springt $\frac{p'(t)}{p(t)}$ an der Stelle $t = \alpha_i$ von $-\infty$ nach $+\infty$ (RAG, 1.3.7). Es ist

$$\frac{g(t)}{p(t)} = 1 + s \cdot \frac{p'(t)}{p(t)},$$

also hat $g(t)$ für alle $i \in [n-1]$ im Intervall (α_i, α_{i+1}) (mindestens) eine reelle Nullstelle β_i. Die α_i und β_i sind schon $m-1$ reelle Nullstellen von $g(t)$, es fehlt nur noch eine Nullstelle. Zeige, dass eine solche stets existiert und nicht im Intervall $[\alpha_1, \alpha_n]$ liegt. Für $t \to \pm\infty$ gilt $\frac{g(t)}{p(t)} \to 1$. Ist $s > 0$, so gibt es also eine Nullstelle $\beta < \alpha_1$. Ist $s < 0$, so gibt es eine Nullstelle $\beta > \alpha_n$. $\square$

1.13 Zeige zunächst die letzte Aussage in (a). Sei $f \in \mathbb{R}[\mathbf{x}]_d \setminus \mathcal{H}_d^1(e)$. Ist $f(e) \neq 1$, so gilt das auch in einer Umgebung von f. Ist $f(e) = 1$, so ist also $f \notin \mathcal{H}_d(e)$ (und $f(e) \neq 0$), d.h. es gibt ein $u \in \mathbb{R}^n$, sodass $f(te - u)$ eine nichtreelle Nullstelle besitzt. Nach RAG I, Aufgabe 52 gilt dasselbe in einer Umgebung von f. Somit ist $\mathbb{R}[\mathbf{x}]_d \setminus \mathcal{H}_d^1(e)$ offen in $\mathbb{R}[\mathbf{x}]_d$, d.h. $\mathcal{H}_d^1(e)$ ist abgeschlossen in $\mathbb{R}[\mathbf{x}]_d$. Zeige nun, dass $\mathcal{H}s_d(e)$ offen ist. Wähle einen Unterraum $L \subseteq \mathbb{R}^n$ mit $\dim(L) = n-1$ und $e \notin L$ (d.h. $\mathbb{R}^n = L \oplus \mathbb{R}e$). Sei $S_L \coloneqq \{u \in L\colon |u| = 1\}$ und betrachte die stetige Abbildung

$$\phi\colon \mathbb{R}[\mathbf{x}]_d \times \mathbb{R}^n \to \mathbb{R}[t]_{\leq d}, \quad (g, v) \mapsto g(te - v).$$

Sei $H_d \subseteq \mathbb{R}[t]_{\leq d}$ die Menge aller $p \in \mathbb{R}[t]$ mit $\deg(p) = d$, die nur reelle und einfache Nullstellen haben. Erneut nach RAG I, Aufgabe 52 ist H_d offen in $\mathbb{R}[t]_{\leq d}$. Nach Definition ist ein $g \in \mathbb{R}[\mathbf{x}]_d$ genau dann strikt hyperbolisch bzgl. e, wenn $\phi(\{g\} \times (\mathbb{R}^n \setminus \mathbb{R}e)) \subseteq H_d$ gilt. Dafür ist die Bedingung $\phi(\{g\} \times S_L) \subseteq H_d$ hinreichend (und notwendig), denn jedes $u \in \mathbb{R}^n \setminus \mathbb{R}e$ hat die Form $u = ae + bv$ mit $v \in S_L$, $a, b \in \mathbb{R}$ und $b \neq 0$ und es folgt

$$g(te - u) = g(te - (ae + bv)) = b^d \cdot g\left(\frac{t-a}{b} e - v\right).$$

Hat also $g(te - v)$ die Nullstellen $\lambda_1, \ldots, \lambda_d$, so hat $g(te - u)$ die Nullstellen $\mu_i = a + b\lambda_i$. Sind die λ_i reell und paarweise verschieden, so gilt dies auch für die μ_i. Sind X, K topologische Räume, ist K kompakt und $W \subseteq X \times K$ offen, so ist die Menge $\{x \in X\colon \{x\} \times K \subseteq W\}$ offen in X. Daher ist die Menge

$$\mathcal{H}s_d(e) = \{g \in \mathbb{R}[\mathbf{x}]_d\colon \phi(\{g\} \times S_L) \subseteq H_d\} = \big\{g \in \mathbb{R}[\mathbf{x}]_d\colon \{g\} \times S_L \subseteq \underbrace{\phi^{-1}(H_d)}_{\text{offen}}\big\}$$

offen in $\mathbb{R}[\mathbf{x}]_d$.

1.14 Für $s \in \mathbb{R}$ und $u \in \mathbb{R}^n$ mit $\langle u, e\rangle = 0$ sei $T_{u,s}\colon \mathbb{R}[\mathbf{x}]_d \to \mathbb{R}[\mathbf{x}]_d$ der durch

$$T_{u,s} f(x) \coloneqq f(x) + s \cdot \langle x, u\rangle \cdot \partial_e f(x)$$

definierte lineare Operator. Für alle $f \in \mathbb{R}[\mathbf{x}]_d$ und $s \in \mathbb{R}$ ist $T_{u,0} f = f$ und es gilt $T_{u,s} f(e) = f(e)$. Weiter gilt:

(1) $T_{u,s}$ bildet $\mathcal{H}_d(e)$ und $\mathcal{H}s_d(e)$ in sich ab. Denn sei $f \in \mathcal{H}_d(e)$, sei $g \coloneqq T_{u,s} f$ und sei $v \in \mathbb{R}^n$.

Wir zeigen, dass alle Nullstellen von $g(te-v)$ reell sind. Es gilt

$$g(te-v) = \underbrace{f(te-v)}_{=:p(t)} - s \cdot \langle u, v\rangle \cdot \underbrace{(\partial_e f)(te-v)}_{=p'(t)} = p(t) + cp'(t)$$

mit $c := -s \cdot \langle u, v\rangle \in \mathbb{R}$. Nach Lemma 1.12(a) hat $g(te-v)$ also reelle Nullstellen, d.h. es folgt $g \in \mathcal{H}_d(e)$. Analog zeigt man $g \in \mathcal{H}s_d(e)$, falls $f \in \mathcal{H}s_d(e)$.

(2) Ist α eine Nullstelle von $p(t) = f(te-v)$ der Vielfachheit $m \geq 1$, so ist die Vielfachheit von α als Nullstelle von $g(te-v)$ nach Lemma 1.12(b) genau $m-1$, außer im Fall $s \cdot \langle u, v\rangle = 0$.

(3) Sei $u_1, \dots, u_{n-1}$ eine Basis von $e^\perp := (\mathbb{R}e)^\perp \subsetneq \mathbb{R}^n$ (beachte $e \neq 0$). Für $s \in \mathbb{R}$ betrachte den Operator

$$F_s := (T_{u_1,s})^d \circ \dots \circ (T_{u_{n-1},s})^d$$

auf $\mathbb{R}[\mathbf{x}]_d$. Für $s \to 0$ konvergiert F_s gegen $F_0 = \mathrm{id}_{\mathbb{R}[\mathbf{x}]_d}$. Für $s \neq 0$ gilt $F_s(\mathcal{H}_d(e)) \subseteq \mathcal{H}s_d(e)$, denn ist $v \in \mathbb{R}^n \setminus \mathbb{R}e$, so gibt es ein $i \in [n-1]$ mit $\langle v, u_i\rangle \neq 0$. Für $g \in \mathcal{H}_d(e)$ ist

$$(T_{u_i,s}g)(te-v) = g(te-v) + s \cdot \langle te-v, u_i\rangle \cdot (\partial_e g)(te-v) = g(te-v) + s' \cdot (\partial_e g)(te-v)$$

mit $s' \neq 0$. Nach Lemma 1.12 verringert $T_{u_i,s}$ die Vielfachheiten der Nullstellen um 1, während alle neuen Nullstellen Vielfachheit 1 haben und reell sind. Nach d-maligem Anwenden sind also alle Nullstellen (reell und) einfach. Für alle $j \neq i$ ist weiter $T_{u_j,s}f = f$, denn es gilt $\langle v, u_j\rangle = 0$. Für $f \in \mathcal{H}_d(e)$ und alle $s \neq 0$ gilt damit $F_s f \in \mathcal{H}s_d(e)$ und $F_s f \to f$ für $s \to 0$. Insbesondere ist $\mathcal{H}s_d(e)$ dicht in $\mathcal{H}_d(e)$.

1.15 Noch zu zeigen ist 1.11(b). Ohne Einschränkung sei $e = e_1$. Betrachte die Homotopie

$$\mathbb{R}[\mathbf{x}]_d \times [0,1] \to \mathbb{R}[\mathbf{x}]_d,\ (f,s) \mapsto H_s f \quad \text{mit} \quad H_s f(x_1, \dots, x_n) := f(x_1, sx_2, \dots, sx_n).$$

Es gilt $H_1 f = f$ und $H_0 f = x_1^d \cdot f(e)$. Ist f hyperbolisch bzw. strikt hyperbolisch bzgl. e, so gilt dasselbe für $H_s f$, denn es gilt

$$(H_s f)(te-u) = f(t-u_1, -su_2, \dots, -su_n) = f(te-u')$$

mit $u' := (u_1, su_2, \dots, su_n)$. Ist f hyperbolisch bzgl. e mit $f(e) > 0$, so ist

$$h_s f := \left((1-s) + \frac{s}{f(e)}\right) \cdot F_s H_{1-s} f \quad (s \in [0,1])$$

ein Pfad von $h_0 f = f$ nach

$$h_1 f = \frac{1}{f(e)} F_1 H_0 f = F_1(x_1^d) =: g.$$

Dieser verläuft ganz innerhalb $\mathcal{H}_d(e)$, denn für alle $s \in [0,1]$ ist $(1-s) + \frac{s}{f(e)} > 0$ (beachte auch 1.14(1)). Außerdem ist $h_s f$ strikt hyperbolisch für alle $s > 0$ nach 1.14(3). Damit ist

$$h\colon \mathcal{H}_d^+(e) \times [0,1] \to \mathcal{H}_d^+(e), \quad (f,s) \mapsto h_s f$$

eine Deformationsretraktion von $\mathcal{H}_d^+(e)$ (bzw. $\mathcal{H}s_d^+(e)$) auf $g = h_1 f \in \mathcal{H}s_d^+(e)$. □

Damit ist Theorem 1.11 vollständig bewiesen.

Bemerkung. Ursprünglich wurde in der Vorlesung $\mathcal{H}_d(e) \cup \{0\} = \overline{\mathcal{H}s_d(e)}$ behauptet. Das ist falsch, betrachte beispielsweise $f_t := tx_1^2 - x_2^2 - x_3^2$. Für $t > 0$ ist diese Form hyperbolisch bzgl. e_1, dagegen ist f_0 bzgl. keinem Punkt hyperbolisch (d.h. $\mathcal{H}_2(e_1) \cup \{0\}$ ist nicht abgeschlossen).

1.16 Bemerkung. Es ist $\mathcal{H}_d(e) = \mathcal{H}_d^+(e) \cup \mathcal{H}_d^-(e)$ und die beiden Mengen $\mathcal{H}_d^\pm(e)$ sind kontrahierbar. Typische Beispiele von kontrahierbaren Mengen sind sternförmige Mengen. $\mathcal{H}_d(e)$ ist für $d \geq 2$ jedoch nicht sternförmig (Aufgabe 5). Die Formen

$$f_\pm := x^2 \pm 2axy + y^2 = (x \pm ay)^2 - (a^2-1)y^2 \in \mathbb{R}[x,y]$$

sind für $a > 1$ strikt hyperbolisch bzgl. e_1, jedoch ist $\frac{1}{2}f_+ + \frac{1}{2}f_- = x^2 + y^2$ nicht hyperbolisch bzgl. e_1. (Das zeigt (die schwächere Aussage), dass $\mathcal{H}_2(e_1)$ nicht konvex ist.)

§2 Beweis der (Ex-)Lax-Vermutung

2.1 Theorem (Helton-Vinnikov 2006)**.** Sei $f \in \mathbb{R}[x,y,z]$ eine bzgl. $e \in \mathbb{R}^3$ hyperbolische Form mit $\deg(f) = d$ und $f(e) > 0$. Dann gibt es Matrizen $A, B, C \in \mathsf{S}^d$ mit

$$f(x,y,z) = \det(xA + yB + zC)$$

und $e_1A + e_2B + e_3C \succ 0$ (siehe Theorem 1.3).

Wir diskutieren den algebraischen Beweis von Hanselka[5] (2015).

2.2 Ein Polynom $f \in \mathbb{R}[\mathbf{x}]$ heißt ein *RZ-Polynom* („real zero“) bzgl. $u \in \mathbb{R}^n$, falls $f(u) \neq 0$ gilt und das Polynom $f(u+tv) \in \mathbb{R}[t]$ für alle $v \in \mathbb{R}^n$ nur reelle Nullstellen hat. Ist $f \in \mathbb{R}[\mathbf{x}]$ RZ bzgl. $u \in \mathbb{R}^n$, so ist die Form

$$f^h := x_0^{\deg(f)} f\left(\frac{x_1}{x_0}, \ldots, \frac{x_n}{x_0}\right) \in \mathbb{R}[x_0, \mathbf{x}]$$

hyperbolisch bzgl. $(1,u) \in \mathbb{R}^{n+1}$. Ist umgekehrt $f \in \mathbb{R}[x_0,\mathbf{x}]$ eine bzgl. $(1,u)$ hyperbolische Form, so ist $f(1,x_1,\ldots,x_n) \in \mathbb{R}[\mathbf{x}]$ RZ bzgl. $u \in \mathbb{R}^n$. Das Analogon des Hyperbolizitätskegels für ein RZ-Polynom f (bzgl. $u \in \mathbb{R}^n$) ist die starr konvexe Menge $S_u(f) := \{v \in \mathbb{R}^n \colon \forall x \in [u,v) \colon f(x) \neq 0\}$.

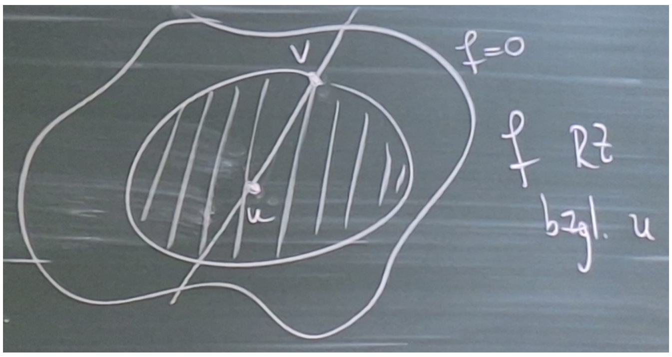


Abbildung 3: Der „Hyperbolizitätskegel“ eines RZ-Polynoms

Für $v \in \mathbb{R}^n$ (und f RZ bzgl. u) ist $\mu(f,v) = \operatorname{ord}_{t=0} f(tu + (1-t)v)$ (Satz 1.8).
Sind beispielsweise $A_0, \ldots, A_n \in \mathsf{S}^d$, $A(x) := A_0 + \sum_{i=1}^n x_iA_i$ und ist $u \in \mathbb{R}^n$ mit $A(u) \succ 0$, so ist $f(x) := \det A(x)$ RZ bzgl. u (das folgt aus 1.2, denn es gilt $f(x) = g(1,x_1,\ldots,x_n)$ mit $g(x_0,\ldots,x_n) := \det(x_0A_0 + \cdots + x_nA_n) \in \mathbb{R}[x_0,\mathbf{x}]$). Die zugehörige starr konvexe Menge ist (ebenfalls nach 1.2) $S_u(f) = \{v \in \mathbb{R}^n \colon A(v) \succeq 0\}$.

Wie betrachten die folgende Umformulierung von Theorem 2.1:

2.3 Theorem. Sei $f \in \mathbb{R}[x,y]$ RZ bzgl. 0 mit $f(0) = 1$ und sei $d = \deg(f)$. Dann gibt es Matrizen $A, B \in \mathsf{S}^d$ mit $f(x,y) = \det(I_d + xA + yB)$.

[5] Christoph Hanselka (*19??)

Lemma. Die Aussagen der beiden Theoreme 2.1 und 2.3 sind äquivalent.

Beweis. Gelte zunächst 2.1 und sei f gegeben wie in 2.3. Die Form $f^h(x,y,z)$ ist hyperbolisch bzgl. e_3, nach 2.1 gibt es also $A,B,C \in \mathsf{S}^d$ mit $C \succ 0$ und $f^h(x,y,z) = \det(xA+yB+zC)$. Wähle ein $S \in \mathrm{GL}_d(\mathbb{R})$ mit $C = SS^\top$, dann ist $xA+yB+zC = S(xA'+yB'+zI)S^\top$ für $A' := S^{-1}AS^{-\top}$ und $B' := S^{-1}BS^{-\top}$. Wegen $\det(SS^\top) = \det(C) = f^h(e_3) = f(0) = 1$ folgt $f(x,y) = f^h(x,y,1) = \det(xA'+yB'+I_d)$.
Gelte nun 2.3 und sei f gegeben wie in 2.1. Wähle entsprechende Koordinaten mit $e = e_3$, dann ist $g(x,y) := \frac{1}{f(e)} f(x,y,1)$ RZ bzgl. 0 mit $g(0) = 1$. Nach 2.3 ist also $g(x,y) = \det(I_d + xA + yB)$ für geeignete $A,B \in \mathsf{S}^d$, d.h. $f(x,y,1) = \det(\alpha I_d + x\alpha A + y\alpha B)$ für ein $\alpha \in \mathbb{R}$ mit $\alpha^d = f(e)$. $\square$

2.4 Sei $h(x,y,z) \in \mathbb{R}[x,y,z]$ eine bzgl. e_3 hyperbolische Form mit $\deg(h) = d$ und $h(e_3) = 1$. Nach Aufgabe 4(a) kann man annehmen, dass h irreduzibel ist. Dann ist $f(x,t) := h(x,1,t) \in \mathbb{R}[x,t]$ irreduzibel und normiert vom Grad d in der Variable t. Wir suchen Matrizen $A,B \in \mathsf{S}^d$ mit $f(x,t) = \det(tI_d + xA + B)$. Das Polynom $f(x,t)$ hat die folgende Eigenschaft: Für jedes $\xi \in \mathbb{R}$ hat das Polynom $f_\xi(t) := f(\xi,t) \in \mathbb{R}[t]$ nur reelle Nullstellen und ist normiert vom Grad d. Denn es gilt $f_\xi(t) = h(te_3 - u)$ für $u := (-\xi,-1,0)$.

Für den Beweis nehmen wir an, dass die affine Kurve $C := \mathcal{V}(f) \subseteq \mathbb{A}^2$ nichtsingulär ist (d.h. sie besitzt keine singulären $\mathbb{C}$-Punkte (Punkte $(a,b) \in \mathbb{C}^2$ mit $f(a,b) = \frac{\partial f}{\partial x}(a,b) = \frac{\partial f}{\partial t}(a,b) = 0$)). Aus 1.8 folgt daher, dass für jedes $\xi \in \mathbb{R}$ alle (reellen) Nullstellen von $f_\xi(t) = f(\xi,t)$ einfach sind.

2.5 Setze $A := \mathbb{R}[x]$. Die A-Algebra $B := \mathbb{R}[x,t]/\langle f\rangle = \mathbb{R}[C]$ ist als A-Modul frei vom Rang d mit A-Basis $1,t,\dots,t^{d-1}$. Sei $f(x,t) = t^d + a_1(x)t^{d-1} + \cdots + a_d(x)$ mit $a_i = a_i(x) \in A$. Dabei ist $\deg(a_i(x)) \le i$ für alle $i \in [d]$. Die Multiplikation $\mu_t \colon B \to B$, $p \mapsto tp$ wird (bzgl. obiger A-Basis) durch die Matrix

$$\begin{pmatrix} 0 & \cdots & 0 & -a_d(x) \\ 1 & & & \vdots \\ & \ddots & & \vdots \\ & & 1 & -a_1(x) \end{pmatrix} \in A^{d\times d}$$

beschrieben. Das charakteristische Polynom ist $\det(t\,\mathrm{id} - \mu_t) = f(x,t)$, wie man per Induktion sieht. Auf dem A-Modul B betrachten wir die symmetrische Bilinearform

$$b\colon B\times B \to A, \quad b(p,q) := \mathrm{tr}_{B/A}(pq) = \mathrm{tr}\begin{pmatrix} B \to B \\ b \mapsto pqb \end{pmatrix}.$$

Weiter betrachten wir für $\xi \in \mathbb{R}$ das folgende Tensorprodukt:

$$\begin{array}{ccc} B & \longrightarrow & B_\xi := B \otimes_{A,\xi} \mathbb{R} = \mathbb{R}[t]/\langle f_\xi\rangle \\ \uparrow & & \uparrow \\ \mathbb{R}[x] = A & \xrightarrow{\ \xi\ } & \mathbb{R} \end{array}$$

Für die zugehörige Spurform $b_\xi \colon B_\xi \times B_\xi \to \mathbb{R}$ gilt explizit (siehe RAG, 1.3.30)

$$b_\xi(t^i,t^j) = \mathrm{tr}_{B_\xi/\mathbb{R}}\left(t^{i+j}\right) = \sum_{k=1}^d \alpha_k^{i+j},$$

wobei $\alpha_1,\dots,\alpha_d \in \mathbb{R}$ die Nullstellen von $f_\xi(t)$ seien (d.h. $b_\xi(t^i,t^j)$ ist gerade die $(i+j)$-te Newtonsumme von $f_\xi(t)$). Die Matrix der symmetrischen Bilinearform b_ξ ist also genau die Hermitematrix $H(f_\xi)$. Da alle Nullstellen von $f_\xi(t)$ reell und einfach sind, ist $b_\xi \succ 0$ (siehe RAG, 1.3.26).

Was bisher geschah. Wir haben $f \in \mathbb{R}[x,t]$, $f = t^d + a_1(x)t^{d-1} + \dots + a_d(x)$ mit $a_i(x) \in \mathbb{R}[x]$ und $\deg(a_i) \leq i$ für alle i. Die affine Kurve $C = \mathcal{V}(f) \subseteq \mathbb{A}^2$ ist glatt. Für $\xi \in \mathbb{R}$ hat $f(\xi, t) = f_\xi(t)$ genau d verschiedene reelle Nullstellen. Das Ziel ist es, Matrizen $M, N \in \mathsf{S}^d$ mit

$$f(x,t) = \det(tI_d + xM + N)$$

zu finden. Es ist $A = \mathbb{R}[x]$ und $B = \mathbb{R}[C] = \mathbb{R}[x,t]/\langle f \rangle$ (ein freier A-Modul mit Basis $1, t, \dots, t^{d-1}$). Wir betrachten die symmetrische Bilinearform $b\colon B \times B \to A$, $(p,q) \mapsto \operatorname{tr}(pq) = \operatorname{tr}_{B/A}(pq)$. Für $\xi \in \mathbb{R}$ und $B_\xi = \mathbb{R}[t]/\langle f_\xi \rangle$ ist $b_\xi\colon B_\xi \times B_\xi \to \mathbb{R}$ die Hermiteform von $f_\xi(t)$. Es gilt $b_\xi \succ 0$.

2.6 Sei R ein Ring und sei M ein endlich erzeugter freier R-Modul. Eine symmetrische Bilinearform $b\colon M \times M \to R$ heißt *unimodular*, falls die induzierte lineare Abbildung

$$\varphi\colon M \to M^\vee \coloneqq \operatorname{Hom}_R(M,R), \quad x \mapsto b(x, \cdot)$$

bijektiv ist. Ist $v_1, \dots, v_n$ eine R-Basis von M mit dualer Basis $v_1^\vee, \dots, v_n^\vee$ von $M^\vee$, so wird φ durch die Matrix $T \coloneqq (b(v_i, v_j))_{i,j}$ beschrieben. b ist also genau dann unimodular, wenn $\det(T) \in R^*$ ist. In unserer Situation 2.5 wird $b\colon B \times B \to A$ im Allgemeinen **nicht** unimodular sein (denn es kann ein $\xi \in \mathbb{C}$ geben, sodass $f(\xi,t)$ eine mehrfache $\mathbb{C}$-Nullstelle hat und in diesem Fall ist $\operatorname{rk}(b_\xi) < d$, also ist b nicht unimodular). Zunächst nehmen wir aber an, dass b unimodular ist.

2.7 Theorem (Harder[6]). Sei k ein Körper mit $\operatorname{char}(k) \neq 2$ und sei $T \in \operatorname{Sym}_n(k[x])$ mit $\det(T) \in k^*$. Dann existiert ein $S \in \operatorname{GL}_n(k[x])$, sodass $S^\top T S$ eine Diagonalmatrix (mit Einträgen in k^*) ist.

Beweis. Siehe [3], Chapter 6, Theorem 3.3. □

2.8 In 2.5 sei b also unimodular. Wegen $b_\xi \succ 0$ für $\xi \in \mathbb{R}$ hat der A-Modul B nach Theorem 2.7 eine Orthonormalbasis $p_1, \dots, p_d$ (d.h. es gilt $b(p_i, p_j) = \delta_{ij}$ für alle i, j). Die A-lineare Abbildung $\mu_t\colon B \to B$, $p \mapsto tp$ ist selbstadjungiert bzgl. b, denn es gilt $b(tp, q) = \operatorname{tr}((tp)q) = b(p, tq)$. Bzgl. der obigen Orthonormalbasis wird μ_t also durch eine symmetrische Matrix $M(x) \in \operatorname{Sym}_d(A)$ beschrieben. Es ist $f(x,t) = \det(t \operatorname{id} - \mu_t) = \det(tI_d - M(x))$. Diese symmetrische Determinantendarstellung von $f(x,t)$ ist automatisch linear, d.h. es gilt $M(x) = xM_1 + M_2$ für geeignete $M_1, M_2 \in \mathsf{S}^d$: Es ist $f = t^d + \sum_{i=1}^d a_i(x) t^{d-i}$ mit $a_i \in A$ und $\deg(a_i) \leq i$. Wende Aufgabe 7 auf $K \coloneqq \operatorname{Quot}(A) = \mathbb{R}(x)$ und die diskrete Bewertung

$$v\left(\frac{p(x)}{q(x)}\right) = \deg(q) - \deg(p)$$

auf K mit Restklassenkörper $\mathbb{R}$ (betrachte $\frac{p}{q} \mapsto \frac{p_m}{\operatorname{LC}(q)}$ für $p = \sum_{i=0}^{\deg(p)} p_i x^i$ und $m = \deg(q)$) an. Es folgt

$$v(M(x)) = \min_{i \in [d]} \frac{v(a_i(x))}{i} \geq -1,$$

denn es ist $\deg(a_i) \leq i$, also $v(a_i) \geq -i$ bzw. $\frac{v(a_i)}{i} \geq -1$.

2.9 Satz (Eulers[7] Lemma). Sei L/K eine endliche separable Körpererweiterung mit Körpergrad $n \coloneqq [L:K] < \infty$, sei $u \in L$ mit $L = K(u)$ und sei $f(t) \coloneqq \operatorname{MinPol}(u/K) \in K[t]$. Dann gelten

$$\operatorname{tr}_{L/K}\left(\frac{u^{n-1}}{f'(u)}\right) = 1 \qquad \text{und} \qquad \operatorname{tr}_{L/K}\left(\frac{u^i}{f'(u)}\right) = 0$$

für alle $i \in \{0, \dots, n-2\}$.

[6] Günter Harder (*1938)
[7] Leonhard Euler (1707–1783)

Beweis. Sei $f(t) = \prod_{j=1}^n (t - u_j)$ mit $u_j \in \overline{K}$.Da f separabel ist, ist $f'(u_j) \neq 0$ für alle j. Es gilt

$$\sum_{j=1}^n \frac{1}{f'(u_j)} \prod_{k \neq j} (t - u_k) = 1, \tag{$*$}$$

denn die linke Seite in $(*)$ hat Grad $\leq n-1$ und setzt man $t = u_\ell$ ein, erhält man für jedes ℓ den Wert 1 (beachte $f'(t) = \sum_{j=1}^n \prod_{k \neq j}(t - u_k)$). Dividiere $(*)$ durch $f(t)$ und erhalte

$$\sum_{j=1}^n \frac{1}{f'(u_j)(t - u_j)} = \frac{1}{f(t)}.$$

Schreibe beide Seiten als Polynom in $\frac{1}{t}$ (Taylorentwicklung um $t = \infty$). Die rechte Seite ist

$$\frac{1}{f(t)} = \left(\frac{1}{t}\right)^n \prod_{j=1}^n \left(\frac{t}{t - u_j}\right) = \left(\frac{1}{t}\right)^n + \left(\text{höhere Potenzen von } \frac{1}{t}\right).$$

Betrachte andererseits die linke Seite: Mit $s := \frac{1}{t}$ ist

$$\frac{1}{t-u} = \frac{1}{s^{-1} - u} = \frac{s}{1 - su} = s + s^2 u + s^3 u^2 + \cdots,$$

also ist die linke Seite

$$\sum_{j=1}^n \frac{1}{f'(u_j)} \left(\frac{1}{t} + \frac{u_j}{t^2} + \frac{u_j^2}{t^3} + \cdots \right).$$

Der Koeffizient von $\left(\frac{1}{t}\right)^i$ in dieser Reihe ist

$$\sum_{j=1}^n \frac{u_j^{i-1}}{f'(u_j)} = \operatorname{tr}_{L/K}\left(\frac{u^{i-1}}{f'(u)}\right)$$

(siehe erneut RAG, 1.3.30), das liefert per Koeffizientenvergleich die Behauptung. $\square$

Sei jetzt wieder $f(x,t) \in \mathbb{R}[x,t]$, $A = \mathbb{R}[x]$, $B = A[t]/\langle f \rangle$, $K := \operatorname{Quot}(A)$ und $L := \operatorname{Quot}(B)$.

2.10 Korollar. Sei $\delta := \frac{\partial f}{\partial t}$ (als Element in B). Die skalierte Spurform

$$\tau \colon B \times B \to A, \quad \tau(p,q) := \operatorname{tr}(pq\delta^{-1}) = \operatorname{tr}_{L/K}(pq\delta^{-1})$$

ist (wohldefiniert und) unimodular.

Beweis. Es ist $\delta \neq 0$ (als Element in B). Nach Eulers Lemma 2.9 ist $\operatorname{tr}(t^i \delta^{-1}) \in A$ für alle $i \in \{0, \dots, d-1\}$ (nämlich 0 für $i = 0, \dots, d-2$ und 1 für $i = d-1$), also folgt $\operatorname{tr}(\delta^{-1}B) \subseteq A$, denn es ist $B = \bigoplus_{i=0}^{d-1} At^i$. Daher ist τ wohldefiniert als Bilinearform $B \times B \to A$. Die Matrix von τ bzgl. $1, t, \dots, t^{d-1}$ ist

$$\begin{pmatrix} 0 & & 1 \\ & \iddots & \\ 1 & & * \end{pmatrix}, \quad \text{denn es gilt} \quad \tau(t^i, t^j) = \operatorname{tr}(t^{i+j}\delta^{-1}) = \begin{cases} 0, & \text{falls } i+j < d-1, \\ 1, & \text{falls } i+j = d-1. \end{cases}$$

Somit ist $\det(\tau) = \pm 1 \in A^*$, d.h. τ ist unimodular. $\square$

Durch den Übergang von b zu τ haben wir leider die positive Definitheit verloren (vgl. Aufgabe 8).

2.11 Satz. Sei C eine nichtsinguläre irreduzible affine Kurve über einem vollkommenen Körper k. Dann ist $k[C]$ ein Dedekindring. Ist $k = \mathbb{C}$, so ist die Divisorenklassengruppe $\mathrm{Cl}(C)$ dividierbar.

Erklärung. Ein *Dedekindring* ist ein eindimensionaler integrer noetherscher Ring R, der ganz abgeschlossen ist (äquivalent: für alle $\mathfrak{m} \in \mathrm{Max}(R)$ ist R_m ein diskreter Bewertungsring). Sei $k = \mathbb{C}$ und setze $\mathrm{Div}(C) := \bigoplus_{\xi \in C} \mathbb{Z}[\xi]$. Zu $f \in k(C)^*$ ist $\mathrm{div}(f) := \sum_{\xi \in C} v_\xi(f) \cdot [\xi] \in \mathrm{Div}(C)$. Die Sequenz

$$k(C)^* \xrightarrow[f \mapsto \mathrm{div}(f)]{\mathrm{div}} \mathrm{Div}(C) \longrightarrow \mathrm{Cl}(C) \longrightarrow 0$$

ist exakt. Die *gebrochenen Ideale* von $k[C]$ bilden eine zu $\mathrm{Div}(C)$ isomorphe abelsche Gruppe. In einem Dedekindring R hat jedes gebrochene Ideal I eine eindeutige Darstellung

$$I = \prod_{\mathfrak{p} \in \mathrm{Max}(R)} \mathfrak{p}^{n(\mathfrak{p})}$$

mit $n(\mathfrak{p}) \in \mathbb{Z}$ und $n(\mathfrak{p}) = 0$ für fast alle $\mathfrak{p} \in \mathrm{Max}(R)$. Dabei ist $\mathfrak{p}^{-1} := \mathrm{Hom}(\mathfrak{p}, R)$. Die Gruppe

$$\mathrm{Cl}(C) := \mathrm{Cl}(k[C]) := \frac{\{\text{gebrochene Ideale von } k[C]\}}{\{\text{gebrochene Hauptideale von } k[C]\}}$$

heißt die *Idealklassengruppe* (oder *Divisorenklassengruppe*) von $k[C]$. Eine abelsche Gruppe $(G, +)$ heißt *dividierbar*, falls $\forall n \in \mathbb{N} \colon \forall x \in G \colon \exists y \in G \colon x = ny$ gilt. Die Dividierbarkeit von $\mathrm{Cl}(C)$ (in 2.11) bedeutet, dass es für jedes gebrochene Ideal I von $k[C]$ und jedes $n \in \mathbb{N}$ ein gebrochenes Ideal J und ein $f \in k(C)^*$ mit $I = fJ^n$ gibt. Wir werden dies nur für $n = 2$ verwenden.

2.12 Lemma. Es gibt ein gebrochenes Ideal I von B und eine Quadratsumme $c \neq 0$ in L mit $I^2 = \left\langle \frac{c}{\delta} \right\rangle$ (ein gebrochenes Hauptideal).

Beweis. Sei $B_\mathbb{C} := B \otimes_\mathbb{R} \mathbb{C} = \mathbb{C}[C]$ und sei $\mathcal{I}(B_\mathbb{C})$ die Gruppe der gebrochenen Ideale von $B_\mathbb{C}$. Es ist $\mathrm{Quot}(B_\mathbb{C}) = \mathbb{C}(C) = L(\sqrt{-1})$. Das Hauptideal $\langle \delta \rangle = B\delta$ in $B = \mathbb{R}[C]$ hat keine $\mathbb{R}$-Punkte, d.h. für $(\xi, s) \in \mathbb{R}^2$ mit $f(\xi, s) = 0$ ist $\delta(\xi, s) \neq 0$ (da $f_\xi(t)$ nur einfache Nullstellen hat). Das Hauptideal $\langle \delta \rangle_\mathbb{C} := B_\mathbb{C}\delta$ ist deshalb ein Produkt von Primidealen und ihren Konjugierten. Also gilt $\langle \delta \rangle_\mathbb{C} = J_0\overline{J_0}$ für ein Ideal $J_0 \subseteq B_\mathbb{C}$. Da $\mathrm{Cl}(B_\mathbb{C})$ 2-dividierbar ist, ist $J_0 = qJ_1^2$ mit $J_1 \in \mathcal{I}(B_\mathbb{C})$ und $q \in \mathbb{C}(C)^*$. Es folgt $\langle \delta \rangle_\mathbb{C} = c(J_1\overline{J_1})^2$. Dabei ist $c := q\overline{q} \in L^*$ eine Summe von zwei Quadraten in L (wegen $(a + ib)(a - ib) = a^2 + b^2$). Es folgt $\langle \delta \rangle = cJ^2$ (in $\mathcal{I}(B)$) mit dem gebrochenen Ideal $J := J_1\overline{J_1} \cap L$ von B. Für das gebrochene Ideal $I := J^{-1}$ von B ist also $I^2 = \langle \frac{c}{\delta} \rangle$ (in $\mathcal{I}(B)$). $\square$

Seien nun ein gebrochenes Ideal I und eine Quadratsumme c wie in Lemma 2.12 fixiert.

2.13 Lemma. Die symmetrische Bilinearform

$$\beta \colon I \times I \to A, \quad (p, q) \mapsto \mathrm{tr}\left(\frac{pq}{c}\right)$$

über A ist unimodular und der A-Modul I besitzt eine Orthonormalbasis bzgl. β.

Beweis. Der A-Modul I ist frei vom Rang d. Es gilt

$$\beta(p, q) = \mathrm{tr}\left(\frac{pq}{c}\right) \in \mathrm{tr}\left(\frac{1}{c}I^2\right) = \mathrm{tr}\left(\frac{1}{\delta}B\right) \subseteq A$$

nach 2.9. Lokalisiere in einem Primideal von B: Lokal ist $I = B'g$ (ein gebrochenes Hauptideal der Lokalisierung B' von B). Es ist $I^2 = \langle \frac{c}{\delta} \rangle$, also $g^2 = u \cdot \frac{c}{\delta}$ mit $u \in (B')^*$. Lokal ist β die Abbildung

$$B' \times B' \to A', \quad (p, q) \mapsto \beta(pg, qg) = \mathrm{tr}\left(\frac{pqg^2}{c}\right) = \mathrm{tr}\left(\frac{upq}{\delta}\right)$$

und diese ist nach 2.10 unimodular.
Sei $0 \neq p \in L$. Es gibt ein $\xi \in \mathbb{R}$ so, dass p und c weder Null- noch Polstellen in einem Punkt (ξ, s) für $s \in \mathbb{C}$ haben. Da c eine Quadratsumme ist, gilt für jedes solche ξ

$$\beta(p,p)(\xi) = \operatorname{tr}\left(\frac{p^2}{c}\right)(\xi) = \sum_{j=1}^{d} \frac{p(\xi, s_j)^2}{c(\xi, s_j)} > 0,$$

wobei $s_1, \dots, s_d$ die Nullstellen von $f_\xi(t) = f(\xi, t)$ seien. Nach dem Satz von Harder hat I also eine A-Basis $p_1, \dots, p_d$ mit $\beta(p_i, p_j) = \delta_{ij} a_i$ und $a_i > 0$, d.h. I besitzt eine Orthonormalbasis. □

2.14 Beweisende. Verfahre wie in 2.8 (mit β statt b): Die A-lineare Abbildung $\mu_t\colon I \to I$ ist selbstadjungiert bzgl. β und hat das charakteristische Polynom $\det(t\,\mathrm{id} - \mu_t) = f(x,t)$. Die Matrix $M(x)$ von μ_t bzgl. $p_1, \dots, p_d$ ist symmetrisch. Nach Aufgabe 7 haben alle Einträge von $M(x)$ Grad ≤ 1, also erhalten wir eine Darstellung $f(x,t) = \det(tI_d + xA + B)$. Im Fall, dass f strikt hyperbolisch ist, ist damit die Lax-Vermutung bewiesen.

2.15 Allgemeiner Fall. Sei $f \in \mathbb{R}[x,y,z]_d$ hyperbolisch bzgl. $e = e_1$ mit $f(e) = 1$. Nach dem Theorem von Nuij existiert eine Folge $(f_j)_{j \geq 1} \subseteq \mathcal{H}s_d(e)$ mit $f_j(e) = 1$ für alle $j \in \mathbb{N}$, die gegen f konvergiert. Es existiert sogar eine solche Folge, dass alle $\mathcal{V}(f_j)$ glatt sind. Der Grund hierfür ist, dass sich jedes f_j durch glatte Formen approximieren lässt und diese sind automatisch strikt hyperbolisch, falls sie nahe bei f_j liegen. Für alle j ist also $f_j(x,y,z) = \det(xA_j + yB_j + zI_d)$ und die Abbildung $\phi\colon \mathsf{S}^d \times \mathsf{S}^d \to \mathbb{R}[x,y,z]_d$, $(A,B) \mapsto \det(xA + yB + zI_d)$ ist eigentlich (Aufgabe 6).

$$\begin{tikzcd} \mathsf{S}^d \times \mathsf{S}^d \arrow[d, "\phi"', "\text{eigentlich}"] \\ \mathbb{R}[x,y,z]_d \end{tikzcd} \qquad \begin{tikzcd} \{(A,B)\} \arrow[d] & \{(A_j,B_j)\} \arrow[l, "\text{konvergente}", "\text{Teilfolge}"'] \arrow[d] \\ f & \{f_j\} \arrow[l] \end{tikzcd}$$

§3 Spektraederschatten: Die Sätze von Helton-Nie

Erinnerung. Ein *Spektraeder* ist eine Menge $\{\xi \in \mathbb{R}^n \colon A(\xi) \succeq 0\}$ mit $A_i \in \mathsf{S}^d$ und $A(x) = A_0 + \sum_{i=1}^n x_i A_i$. Spektraeder sind basisch abgeschlossen, semialgebraisch, konvex und sogar starr konvex (für $n = 2$ ist (nach §2) auch jede Menge mit diesen Eigenschaften ein Spektraeder, für $n > 2$ ist die Frage noch offen). Ein *Spektraederschatten* ist eine Menge $\{\xi \in \mathbb{R}^n \colon \exists \eta \in \mathbb{R}^m \colon A(\xi, \eta) \succeq 0\}$ mit $A_i, B_j \in \mathsf{S}^d$ und

$$A(x,y) = A_0 + \sum_{i=1}^{n} x_i A_i + \sum_{j=1}^{m} y_j B_j.$$

Spektraederschatten sind konvex und semialgebraisch. Die Sätze von Helton-Nie[8] besagen, dass eine Menge $K \subseteq \mathbb{R}^n$ ein Spektraederschatten ist, falls K kompakt, konvex und semialgebraisch ist und ∂K „gutartig" ist (Krümmung, Singularitäten, …).

3.1 Sei $\boldsymbol{g} = (g_1, \dots, g_r)$ ein Tupel in $\mathbb{R}[\mathbf{x}]^r$, sei $S \coloneqq \mathcal{S}(\boldsymbol{g}) = \{\xi \in \mathbb{R}^n \colon \forall i \in [r] \colon g_i(\xi) \geq 0\}$ und sei $M \coloneqq QM(\boldsymbol{g}) = \{s_0 + s_1 g_1 + \dots + s_r g_r \mid s_0, \dots, s_r \in \Sigma\mathbb{R}[\mathbf{x}]^2\}$. Setze $g_0 \coloneqq 1$ und betrachte

$$M_k \coloneqq \left\{ \sum_{i=0}^{r} s_i g_i \;\middle|\; \forall i \in \{0, \dots, r\} \colon s_i \in \Sigma\mathbb{R}[\mathbf{x}]^2 \wedge \deg(s_i g_i) \leq k \right\} \subseteq M \cap \mathbb{R}[\mathbf{x}]_{\leq k}$$

(die trunkierten quadratischen Moduln) für $k \in \mathbb{N}_0$. M_k ist eine semialgebraische Menge in $\mathbb{R}[\mathbf{x}]_{\leq k}$. Für $f \in \mathbb{R}[\mathbf{x}]$, etwa $f = \sum_\alpha c_\alpha \mathbf{x}^\alpha$ mit $c_\alpha \in \mathbb{R}$ sei $\|f\| \coloneqq \max_\alpha |c_\alpha|$ (die L^∞-Norm).

[8] Jiawang Nie (*19??)

3.2 Theorem. Seien $\boldsymbol{g}$, S und M gegeben wie in 3.1 und sei M archimedisch. Für jedes $d \geq 1$ und jedes $c \in (0, \infty)$ gibt es ein $k = k(\boldsymbol{g}, d, c) \in \mathbb{N}$ so, dass M_k jedes $f \in \mathbb{R}[\mathbf{x}]_{\leq d}$ mit $f \geq \frac{1}{c}$ auf S und $\|f\| \leq c$ enthält.

3.3 Formuliere Theorem 3.2 zunächst koordinatenfrei: Sei A eine endlich erzeugte $\mathbb{R}$-Algebra, sei $M := QM(\boldsymbol{g}) \subseteq A$ ein archimedischer quadratischer Modul mit $\boldsymbol{g} = (g_1, \ldots, g_r) \in A^r$ und sei $S := X_M = \{\alpha \in \operatorname{Hom}(A, \mathbb{R})\colon \forall i \in [r]\colon \alpha(g_i) \geq 0\}$.
Wir zeigen: Für jeden linearen Teilraum $U \subseteq A$ mit $\dim(U) < \infty$ und jedes $c > 0$ existiert ein linearer Teilraum $V \subseteq A$ mit $\dim(V) < \infty$, sodass gilt: Zu jedem $f \in U$ mit $f \geq \frac{1}{c}$ auf S und $\|f\| \leq c$ (bzgl. einer fixierten Basis von U) gibt es $s_0, \ldots, s_r \in \Sigma V^2$ mit $f = s_0 + s_1 g_1 + \cdots + s_r g_r$.

3.4 Ein *Matrixpolynom* vom Format $r \times s$ ist eine Matrix $T = (t_{ij}) \in \mathbb{R}[\mathbf{x}]^{r \times s}$. Für ein solches T setzen wir $\deg(T) := \max_{i,j} \deg(t_{ij})$ und $\|T\| := \max_{i,j} \|t_{ij}\|$. Ist $T \in \operatorname{Sym}_m(\mathbb{R}[\mathbf{x}])$, so heißt T *(matrix-)sos*, falls es $m \times m$-Matrixpolynome T_ν mit $T = \sum_\nu T_\nu T_\nu^\top$ gibt oder äquivalent, falls $T = UU^\top$ für ein Matrixpolynom U mit m Zeilen gilt.

3.5 Theorem. Sei $\boldsymbol{g} = (g_1, \ldots, g_r) \in \mathbb{R}[\mathbf{x}]^r$, sei der Modul $M = QM(\boldsymbol{g})$ archimedisch und sei $S = \mathcal{S}(\boldsymbol{g}) \subseteq \mathbb{R}^n$. Für alle $d, m \geq 1$ und alle $0 < c \in \mathbb{R}$ gibt es ein $k = k(\boldsymbol{g}, d, m, c) \in \mathbb{N}$, sodass gilt: Ist $T \in \operatorname{Sym}_m(\mathbb{R}[\mathbf{x}])$ mit $\deg(T) \leq d$, $\|T\| \leq c$ und $T \succeq \frac{1}{c} I_m$ auf S (d.h. $T(\xi) - \frac{1}{c} I_m \succeq 0$ für alle $\xi \in S$), so gibt es sos Matrixpolynome $T_0, \ldots, T_r \in \operatorname{Sym}_m(\mathbb{R}[\mathbf{x}])$ mit $\deg(T_i) \leq k$ und $T = T_0 + g_1 T_1 + \cdots + g_r T_r$.

3.6 Für den Beweis von 3.5 genügt es, Satz 3.2 zu zeigen: Betrachte den kommutativen Teilring $B := \mathbb{R}[\mathbf{x}, T]$ von $\mathbb{R}[\mathbf{x}]^{m \times m}$ (aus symmetrischen Matrizen). Die Ringerweiterung $\mathbb{R}[\mathbf{x}] \subseteq B$ ist endlich, der von $M \subseteq \mathbb{R}[\mathbf{x}]$ in B erzeugte quadratische Modul M^B ist also archimedisch. Aus der Voraussetzung $T \succ \frac{1}{c} I$ auf S folgt $T > \frac{1}{c}$ auf

$$X_{M^B} = \{\beta\colon B \to \mathbb{R} \mid \forall i \in [r]\colon \beta(g_i) \geq 0\}$$

(siehe RAG, 5.5.21, der Grund war, dass $\beta(T)$ für ein solches β ein Eigenwert von $T(u)$ ist ($\beta|_{\mathbb{R}[\mathbf{x}]}$ korrespondiert zu einer Auswertung in $u \in S$)). Angenommen Theorem 3.2 ist bewiesen (und damit auch 3.3), dann kann man 3.3 auf den quadratischen Modul $M^B \subseteq B$ und auf T anwenden, das liefert Theorem 3.5. (Die Matrizen T_i kann man also sogar als sos in B finden.) □

3.7 *Beweis von Theorem 3.2.* Für $\boldsymbol{g}$, S und M wie in der Voraussetzung, $d \geq 1$ und $c > 0$ sei

$$P_{d,c} := \left\{ f \in \mathbb{R}[\mathbf{x}]_{\leq d} \colon f|_S \geq \frac{1}{c},\ \|f\| \leq c \right\} \subseteq M \cap \mathbb{R}[\mathbf{x}]_{\leq d},$$

also

$$P_{d,c} \subseteq \bigcup_{k \geq 1} (M_k \cap \mathbb{R}[\mathbf{x}]_{\leq d}) \tag{$*$}$$

(eine Vereinigung von semialgebraischen Mengen). Dabei sei M_k wie in 3.1. Wir wollen $P_{d,c} \subseteq M_k$ für ein $k \geq 1$ zeigen. Nach RAG I, Aufgabe 43 ist das äquivalent dazu, dass $(*)$ nach beliebiger reell abgeschlossener Grundkörpererweiterung gilt. Zu zeigen ist also: Ist $R \supseteq \mathbb{R}$ reell abgeschlossen, so gilt

$$(P_{d,c})_R \subseteq \bigcup_{k \geq 1} \left(\left(M^{R[\mathbf{x}]}\right)_k \cap R[\mathbf{x}]_{\leq d} \right).$$

Daher ist zu zeigen: Für $f \in R[\mathbf{x}]_{\leq d}$ mit $f|_{S_R} \geq \frac{1}{c}$ und $\|f\| \leq c$ ist $f \in M^{R[\mathbf{x}]} := QM_{R[\mathbf{x}]}(\boldsymbol{g})$. Sei $\mathcal{O}$ die konvexe Hülle von $\mathbb{R}$ in R, sei $M^{\mathcal{O}} := QM_{\mathcal{O}[\mathbf{x}]}(\boldsymbol{g}) \subseteq \mathcal{O}[\mathbf{x}]$ und behaupte, dass $M^{\mathcal{O}}$ ein archimedischer quadratischer Modul in $\mathcal{O}[\mathbf{x}]$ ist. $O(M^{\mathcal{O}}) := \{f \in \mathcal{O}[\mathbf{x}]\colon \exists n \in \mathbb{N}\colon n \pm f \in M^{\mathcal{O}}\}$ ist ein Teilring von $\mathcal{O}[\mathbf{x}]$ und es gilt $\mathbb{R}[\mathbf{x}] \subseteq O(M^{\mathcal{O}})$, sowie $\mathcal{O} \subseteq O(M^{\mathcal{O}})$, also folgt $O(M^{\mathcal{O}}) = \mathcal{O}[\mathbf{x}]$, d.h. $M^{\mathcal{O}}$ ist archimedisch. Jedes $f \in R[\mathbf{x}]$ mit $f|_{S_R} \geq \frac{1}{c}$ und $\|f\| \leq c$ liegt in $\mathcal{O}[\mathbf{x}]$. Aus $f|_{S_R} \geq \frac{1}{c}$ und $S_R = \{\xi \in R^n : \forall i \in [r] : g_i(\xi) \geq 0\}$ folgt $f \geq \frac{1}{c}$ auf

$$X_{M^{\mathcal{O}}} = \{\alpha : \mathcal{O}[\mathbf{x}] \to \mathbb{R} \mid \forall i \in [r] : \alpha(g_i) \geq 0\},$$

denn jedes $\alpha \in X_{M^{\mathcal{O}}}$ faktorisiert als $\mathcal{O}[\mathbf{x}] \to \mathbb{R}[\mathbf{x}] \xrightarrow{u} \mathbb{R}$ mit $u \in X_M = S$, d.h. $\alpha(f) = \overline{f(u)} \geq \frac{1}{c} > 0$. Somit ist $f > 0$ auf $X_{M^{\mathcal{O}}}$, d.h. $f \in M^{\mathcal{O}}$ nach dem archimedischen Positivstellensatz (für $M^{\mathcal{O}}$). Insbesondere ist $f \in M^{R[\mathbf{x}]}$. $\square$

3.8 Sei $K \subseteq \mathbb{R}^n$ eine konvexe Menge. Eine Funktion $f : K \to \mathbb{R}$ heißt *konvex*, falls

$$f((1-t)x + ty) \leq (1-t)f(x) + tf(y)$$

für alle $x, y \in K$ und alle $t \in [0,1]$ gilt. Gilt für alle $t \in (0,1)$ und $x \neq y$ sogar stets '$<$', so heißt f *strikt konvex*. f heißt *(strikt) konkav*, falls $-f$ (strikt) konvex ist.

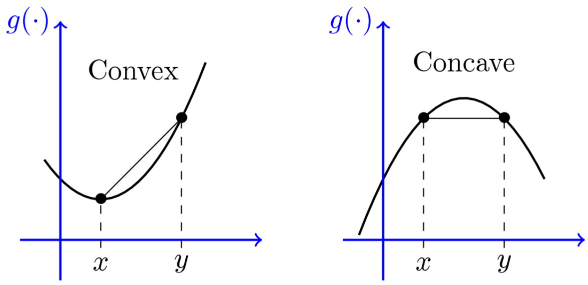


Abbildung 4: (Quelle: https://www.probabilitycourse.com/chapter6/6_2_5_jensen%27s_inequality.php)

Sei $K \subseteq \mathbb{R}^n$ konvex und offen. Ist $f : K \to \mathbb{R}$ eine C^1-Funktion, so ist f genau dann konvex, wenn $\langle \nabla f(u), v - u \rangle \leq f(v) - f(u)$ für alle $u, v \in K$ gilt. Ist f sogar eine C^2-Funktion, so ist f genau dann konvex, wenn $D^2 f(u) \succeq 0$ für alle $u \in K$ gilt. Ist $D^2 f(u) \succ 0$ für alle $u \in K$, so ist f strikt konvex, die Umkehrung ist jedoch im Allgemeinen falsch (z.B. $f(x) = x^4$). Ist f konvex, so ist $\{u \in K : f(u) \leq c\}$ eine konvexe Menge.

3.9 Definition. Ein Polynom $f \in \mathbb{R}[\mathbf{x}]$ heißt *sos-konvex*, falls das Matrixpolynom $D^2 f$ matrix-sos ist. Klar ist, dass jedes sos-konvexe Polynom konvex (auf $\mathbb{R}^n$) ist.

3.10 Lemma. Ist $K \subseteq \mathbb{R}^n$ kompakt und konvex und ist $f : K \to \mathbb{R}$ konkav, so nimmt f sein Minimum auf K in einem Extremalpunkt von K an.

Beweis. Sei $u \in K$ mit $f(u) = \min f(K)$, etwa $u = \sum_{i=0}^r a_i u_i$ mit $u_i \in \mathrm{Ex}(K)$, $a_i \geq 0$ und $\sum_{i=0}^r a_i = 1$ (nach dem Satz von Krein-Milman gilt $K = \mathrm{conv}(\mathrm{Ex}(K))$). Da f konkav ist, folgt $f(u) \geq \sum_{i=0}^r a_i f(u_i) \geq f(u_j)$ für alle $j \in [r]$ mit $a_j \neq 0$, d.h. $f(u_j) = f(u)$ für diese j. $\square$

3.11 Lemma (Lagrange[9]-Multiplikatoren)**.** Sei $K = \mathcal{S}(g_1, \ldots, g_r) \subseteq \mathbb{R}^n$ kompakt und konvex mit $\mathrm{int}(K) \neq \varnothing$. Die Polynome $g_1, \ldots, g_r$ seien konkav auf K und es seien $f \in \mathbb{R}[\mathbf{x}]$ und $u \in K$ mit $f(u) = \min f(K)$. Dann ist $\nabla f(u) = \sum_{i=1}^r b_i \nabla g_i(u)$ mit $b_i \geq 0$ und $b_i g_i(u) = 0$ für alle $i \in [r]$.

Beweis. Wähle ein $v \in K$ mit $g_i(v) > 0$ für alle $i \in [r]$ und setze $w := v - u$. Für jedes i mit $g_i(u) = 0$ gilt

$$0 < g_i(v) \leq g_i(u) + \langle \nabla g_i(u), w \rangle = \langle \nabla g_i(u), w \rangle,$$

denn $g_i|_K$ ist konkav. Ohne Einschränkung sei $g_i(u) = 0$ für $i \in [p]$ und $g_j(u) > 0$ für $j \in [r] \setminus [p]$. Zu zeigen ist $\nabla f(u) \in \mathrm{cone}(\nabla g_1(u), \ldots, \nabla g_p(u)) =: C$ (ein endlich erzeugter und damit

[9] Joseph-Louis Lagrange (1736–1813)

abgeschlossener Kegel (Konvexität, I.6.7)). Angenommen $\nabla f(u) \notin C$, dann existiert ein $z \in C^* \subseteq \mathbb{R}^n$ mit $\langle z, \nabla f(u)\rangle < 0$. Wähle ein $s > 0$ so klein, dass $\langle z + sw, \nabla f(u)\rangle < 0$ ist. Für hinreichend kleine $t > 0$ ist dann $u + t(z + sw) \in K$, denn für $i \in [p]$ und $j \in [r] \setminus [p]$ ist $\langle \nabla g_i(u), z + sw\rangle > 0$ und $g_j(u + t(z + sw)) > 0$. Andererseits ist

$$f(u + t(z + sw)) = f(u) + t \cdot \underbrace{\langle \nabla f(u), z + sw\rangle}_{<0} + (\text{höhere Potenzen von } t),$$

im Widerspruch zu $f(u) = \min f(K)$. $\square$

3.12 Lemma. Sei $F \in \mathrm{Sym}_m(\mathbb{R}[\mathbf{x}])$ matrix-sos und sei $u \in \mathbb{R}^n$. Dann ist auch das Matrixpolynom

$$G_u(x) \coloneqq \int_0^1 \int_0^t F(u + s(x - u))\, ds\, dt$$

matrix-sos und es gilt $\deg G_u(x) = \deg F(x) =: d$.

Beweis. Zeige zunächst die Grad-Aussage: Für jedes Monom $\mathbf{x}^\alpha$ ist

$$\begin{aligned}\int_0^1 \int_0^t \underbrace{(u + s(x - u))^\alpha}_{=\prod_{i=1}^n (u_i + s(x_i - u_i))^{\alpha_i}}\, ds\, dt &= \int_0^1 \int_0^t \left(s^{|\alpha|}\mathbf{x}^\alpha + h_1\right) ds\, dt \\ &= \int_0^1 \left(\frac{t^{|\alpha|+1}}{|\alpha| + 1}\mathbf{x}^\alpha + h_2\right) dt = \frac{\mathbf{x}^\alpha}{(|\alpha| + 1)(|\alpha| + 2)} + (\cdots)\end{aligned}$$

mit zwei Polynomen h_1, h_2 (in (u, x)) und $\deg_x(h_i) < |\alpha|$. Ohne Einschränkung sei $F(x) = v(x)v(x)^\top$ mit $v(x) = (v_1(x), \ldots, v_m(x))^\top$ und $v_i(x) \in \mathbb{R}[x]$ (für $x = (x_1, \ldots, x_n)$ und $y = (y_1, \ldots, y_m)$). Nach Aufgabe 13 ist zu zeigen, dass das Polynom $g(x, y) \coloneqq y^\top G_u(x)y$ sos in $\mathbb{R}[x, y]$ ist. Es gilt $y^\top F(x)y = \left(\sum_{i=1}^m y_i v_i(x)\right)^2$ und damit

$$g(x, y) = \int_0^1 \int_0^t \left(\sum_{i=1}^m y_i v_i(u + s(x - u))\right)^2 ds\, dt.$$

Sei $v_i(u + s(x - u)) = \sum_{k=0}^d p_{ik}(x, u)s^k$ mit $p_{ik} \in \mathbb{R}[x, u]$, dann gilt

$$g(x, y) = \sum_{i=1}^m \sum_{j=1}^m \sum_{k=0}^d \sum_{\ell=0}^d y_i y_j p_{ik}(x, u) p_{j\ell}(x, u) \underbrace{\int_0^1 \int_0^t s^{k+\ell}\, ds\, dt}_{=:\, a_{k\ell}}.$$

Die symmetrische reelle $(d + 1) \times (d + 1)$-Matrix $A_d \coloneqq (a_{k\ell})_{k,\ell=0,\ldots,d}$ ist positiv definit, denn für $z = (z_0, \ldots, z_d)^\top \in \mathbb{R}^{d+1}$ mit $z \neq 0$ ist

$$z^\top A_d z = \int_0^1 \int_0^t \left(z_0 + z_1 s + \cdots + z_d s^d\right)^2 ds\, dt > 0.$$

Damit gilt $A_d = BB^\top$ für eine Matrix $B \in \mathbb{R}^{(d+1)\times(d+1)}$, d.h.

$$\int_0^1 \int_0^t s^{k+\ell}\, ds\, dt = \sum_r b_{kr} b_{\ell r}.$$

Es ist

$$\int_0^1 \int_0^t s^{k+\ell}\, ds\, dt = \int_0^1 \left[\frac{s^{k+\ell+1}}{k + \ell + 1}\right]_0^t dt = \int_0^1 \frac{t^{k+\ell+1}}{k + \ell + 1}\, dt = \frac{1}{(k + \ell + 1)(k + \ell + 2)}$$

und daher

$$g(x,y) = \sum_{i,j}\sum_{k,\ell} y_i y_j p_{ik} p_{j\ell} \sum_r b_{kr} b_{\ell r} = \sum_r \left(\sum_i \sum_k y_i p_{ik} b_{kr} \right)^2. \qquad \square$$

3.13 Theorem. Sei $K = \mathcal{S}(\boldsymbol{g}) \subseteq \mathbb{R}^n$ eine kompakte und konvexe Menge (mit $\boldsymbol{g} = (g_1, \ldots, g_r)$) und sei $M := QM(\boldsymbol{g})$ archimedisch. Für jedes $i \in [r]$ gelte eine der beiden folgenden Bedingungen:

(1) g_i ist sos-konkav (d.h. $-D^2 g_i$ ist matrix-sos).

(2) g_i ist konkav auf K und es gilt $D^2 g_i(u) \prec 0$ für alle $u \in \overline{\mathcal{Z}(g_i) \cap \mathrm{Ex}(K)}$.

Dann hat K eine exakte Momentenrelaxierung bzgl. $\boldsymbol{g}$. Insbesondere ist K ein Spektraederschatten.

3.14 Lemma. Seien die Voraussetzungen von Theorem 3.13 erfüllt. Für jedes $i \in [r]$ gibt es ein $N_i \geq 1$ so, dass für alle $u \in \mathrm{Ex}(K)$ gilt: Das Matrixpolynom

$$G_{i,u}(x) := -\int_0^1 \int_0^t D^2 g_i(u + s(x-u))\, ds\, dt$$

hat die Form

$$G_{i,u}(x) = \sum_{j=0}^r g_j(x) S_{i,u,j}(x)$$

mit $g_0 := 1$ und Matrixpolynomen $S_{i,u,j}$ (für $j \in \{0, \ldots, r\}$), die matrix-sos vom Grad $\leq N_i$ sind.

3.15 *Beweis von Theorem 3.13 unter Verwendung von Lemma 3.14.* Man muss ein $d \geq 0$ so finden, dass der trunkierte quadratische Modul M_d jedes lineare $f \in \mathbb{R}[\mathbf{x}]$ mit $f|_K \geq 0$ enthält (RAG, 8.5.14). Ein solches f nimmt sein Minimum auf K in einem Punkt $u \in \mathrm{Ex}(K)$ an (Lemma 3.10). Da $g_i|_K$ konkav ist, ist $\nabla f(u) = \sum_{i=1}^r b_i \nabla g_i(u)$ mit $b_i \geq 0$ und $b_i g_i(u) = 0$ (Lemma 3.11). Das Polynom $h_u(x) := f(x) - f(u) - \sum_{i=1}^r b_i g_i(x)$ erfüllt $h_u(u) = 0$, $\nabla h_u(u) = 0$ und $D^2 h_u(u) = -\sum_{i=1}^r b_i D^2 g_i(u)$. Für das Matrixpolynom

$$H_u(x) := \int_0^1 \int_0^t D^2 h_u(u + s(x-u))\, ds\, dt$$

gilt $h_u(x) = (x-u)^\top H_u(x)(x-u)$ (Aufgabe 9). Aus 3.14 folgt also

$$H_u(x) = \sum_{i=1}^r b_i G_{i,u}(x) = \sum_{i=1}^r \sum_{j=0}^r b_i g_j(x) S_{i,u,j}(x)$$

mit Matrixpolynomen $S_{i,u,j}$, die matrix-sos vom Grad $\leq \max\{N_1, \ldots, N_r\} =: N$ sind. Damit ist

$$f(x) = f(u) + \sum_{i=1}^r b_i g_i(x) + (x-u)^\top H_u(x)(x-u),$$

mit $f(u) > 0$ und $b_i \geq 0$, d.h. f liegt in $M_{\delta+N+2}$ mit $\delta := \max_i \deg(g_i)$ (Lemma 3.12). $\square$

3.16 *Beweis von Lemma 3.14.* Sei $i \in [r]$ und gelte zunächst (1) für g_i. Dann ist $-D^2 g_i$ matrix-sos mit $\deg(D^2 g_i) = \deg(g_i) - 2$. Nach Lemma 3.12 gilt dasselbe für $G_{i,u}(x)$ (für alle $u \in \mathbb{R}^n$). Gelte jetzt (2) für g_i und sei $Z_i := \overline{\mathcal{Z}(g_i) \cap \mathrm{Ex}(K)}$. Das Matrixpolynom $G_{i,u}(x)$ hat Grad $\deg(g_i) - 2$ und ist polynomial in (x, u). Für $u \in K$ gelten $D^2 g_i(u) \preceq 0$ und $D^2 g_i(u) \prec 0$ für $u \in Z_i$. Für $u \in Z_i$ und $v \in K$ folgt $G_{i,u}(v) \succ 0$, denn für $0 \neq w \in \mathbb{R}^n$ ist

$$w^\top G_{i,u}(v) w = \int_0^1 \int_0^t \underbrace{w^\top(-D^2 g_i)(u + s(x-u))w}_{\substack{>0 \text{ für } s=0 \\ \geq 0 \text{ für alle } 0 \leq s \leq t}}\, ds\, dt > 0.$$

Da K und Z_i kompakt sind, gibt es ein $\varepsilon > 0$ mit $G_{i,u}(v) \succeq \varepsilon I$ für alle $(u,v) \in Z_i \times K$. Nach Theorem 3.5 gibt es für jedes $u \in Z_i$ eine gewichtete sos-Matrixdarstellung

$$G_{i,u}(x) = \sum_{j=0}^{r} g_j(x) S_{i,u,j}(x)$$

mit Matrixpolynomen $S_{i,u,j}$, die matrix-sos vom Grad

$$\deg(S_{i,u,j}) \leq k\left(\boldsymbol{g}, \deg(g_i) - 2, n, \max\left\{\varepsilon^{-1}, \|G_{i,u}\|\right\}\right)$$

sind. Da Z_i kompakt ist, ist $\|G_i\| \coloneqq \max_{u \in Z_i} \|G_{i,u}\| < \infty$, d.h. man hat die uniforme Schranke

$$\deg(S_{i,u,j}) \leq k\left(\boldsymbol{g}, \deg(g_i) - 2, n, \max\left\{\varepsilon^{-1}, \|G_i\|\right\}\right). \qquad \square$$

3.17 Definition. Sei $U \subseteq \mathbb{R}^n$ offen. Eine C^2-Funktion $f\colon U \to \mathbb{R}$ heißt *strikt quasikonkav in* $u \in U$, falls $v^\top D^2 f(u) v < 0$ für alle $0 \neq v \in \mathbb{R}^n$ mit $\langle v, \nabla f(u)\rangle = 0$ gilt. f heißt *strikt quasikonkav*, falls f in jedem $u \in U$ strikt quasikonkav ist. Ist U konvex, so heißt f *quasikonkav*, falls die „superlevel sets“ $\{u \in U\colon f(u) \geq c\}$ für alle $c \in \mathbb{R}$ konvex sind. Ist f eine quasikonkave C^2-Funktion, so gilt $v^\top D^2 f(u) v \leq 0$ für alle $u \in U$ und alle $v \in \mathbb{R}^n$ mit $\langle v, \nabla f(u)\rangle = 0$.

3.18 Lemma. Sei $\mathcal{S}(g_1, \ldots, g_r)$ eine *archimedische Beschreibung* der konvexen Menge $K \subseteq \mathbb{R}^n$ (d.h. es gilt $K = \mathcal{S}(g_1, \ldots, g_r)$ und der quadratische Modul $M \coloneqq QM(g_1, \ldots, g_r)$ ist archimedisch). Jedes g_i sei strikt quasikonkav in jedem Punkt von K. Dann gibt es eine andere archimedische Beschreibung $K = \mathcal{S}(h_0, \ldots, h_r)$ mit $h_i \in M$ und $D^2 h_i \prec 0$ auf K für alle $i \in \{0, \ldots, r\}$.

Beweis. Da M archimedisch ist, gibt es ein $b > 0$ in $\mathbb{R}$ mit $h_0 \coloneqq b^2 - \sum_{i=1}^n x_i^2 \in M$. Skaliere die g_i so, dass $|g_i(u)| \leq 1$ für $|u| \leq b$ und alle $i \in [r]$ gilt und sei $c > 0$ (groß). Nach Aufgabe 12 gibt es ein sos-Polynom $h \in \mathbb{R}[t]$ mit

$$(1)\ h(t) > 0, \qquad (2)\ h(t) + th'(t) > 0, \qquad (3)\ \frac{2h'(t) + th''(t)}{h(t) + th'(t)} \leq -c$$

für $|t| \leq 1$. Setze $h_i(t) \coloneqq g_i(t) \cdot h(g_i(t))$ für $i \in [r]$ und schreibe $\boldsymbol{h} \coloneqq (h_0, h_1, \ldots, h_r)$. Es sind $h_0, \ldots, h_r \in M$, d.h. $K \subseteq \mathcal{S}(\boldsymbol{h})$. Umgekehrt sei $u \in \mathbb{R}^n \setminus K$. Ist $|u| > b$, so ist $h_0(u) < 0$. Ist $|u| \leq b$, so ist $-1 \leq g_i(u) < 0$ für ein $i \in [r]$, d.h. $h_i(t) < 0$. Somit ist $K = \mathcal{S}(\boldsymbol{h})$. Es ist $D^2 h_0 = -2I_n \prec 0$. Für $i \in [r]$ zeigen wir, dass $D^2 h_i \prec 0$ auf K ist, falls $c > 0$ hinreichend groß gewählt war. Es gilt

$$D^2 h_i = (h \circ g_i) \cdot D^2(g_i) + \left(\nabla(g_i) \cdot \nabla(h \circ g_i)^\top\right)^{\mathrm{sym}} + g_i \cdot D^2(h \circ g_i)$$

mit $M^{\mathrm{sym}} \coloneqq M + M^\top$ für $M \in \mathbb{R}^{n \times n}$. Wegen $\nabla(h \circ g_i) = (h' \circ g_i) \cdot \nabla(g_i)$ und

$$D^2(h \circ g_i) = (h'' \circ g_i) \cdot (\nabla g_i)(\nabla g_i)^\top + (h' \circ g_i) \cdot D^2(g_i)$$

ist also

$$D^2(h_i) = p_i \cdot D^2(g_i) + q_i \cdot (\nabla g_i)(\nabla g_i)^\top$$

mit $p_i \coloneqq (h \circ g_i) + g_i \cdot (h' \circ g_i)$ und $q_i \coloneqq 2(h' \circ g_i) + g_i \cdot (h'' \circ g_i)$. Dabei ist $p_i|_K > 0$ nach (2). Da g_i strikt quasikonkav auf K ist, gibt es nach Aufgabe 11 eine Konstante $\kappa_i > 0$ mit $D^2(g_i) \prec \kappa_i \cdot (\nabla g_i)(\nabla(g_i))^\top$ in allen Punkten aus K. Sei $c > 0$ in $\mathbb{R}$ mit $c > \max\{\kappa_1, \ldots, \kappa_r\}$. Nach (3) gilt $\frac{q_i}{p_i} \leq -c < -\kappa_i$ auf K für alle $i \in [r]$, also ist $\kappa_i p_i + q_i < 0$ auf K. Daher gilt

$$D^2(h_i) = p_i \cdot D^2(g_i) + q_i \cdot (\nabla g_i)(\nabla g_i)^\top \prec (\kappa_i p_i + q_i) \cdot (\nabla g_i)(\nabla g_i)^\top \preceq 0$$

auf K, also ist $D^2(h_i) \prec 0$ auf K für alle $i \in [r]$. □

3.19 Theorem. Die Aussage von Theorem 3.13 bleibt richtig, wenn man neben (1) und (2) für die g_i auch die Bedingung (3) zulässt:

(3) g_i ist strikt quasikonkav auf K.

Beweis. Sei (3) erfüllt für $g_1, \ldots, g_s$ und sei (1) oder (2) erfüllt für $g_{s+1}, \ldots, g_r$. Wähle ein $b \in \mathbb{R}$ mit $b^2 - \sum_{i=1}^n x_i^2 \in QM(\boldsymbol{g})$. Nach Lemma 3.18 kann man die Folge $b^2 - \sum_i x_i^2, g_1, \ldots, g_s$ durch eine andere Folge in $QM(\boldsymbol{g})$ so ersetzen, dass jedes der neuen Polynom auf K eine negativ definite Hesse-Matrix hat. Jedes Element der neuen Gesamtfolge für K erfüllt (1) oder (2), also folgt die Behauptung aus Theorem 3.13. □

Zusammenfassung. Nach den Theoremen 3.13 und 3.19 gilt also: Sei $K = \mathcal{S}(g_1, \ldots, g_r)$ eine archimedische Beschreibung und sei K konvex. Für jedes $i \in [r]$ gelte eine dieser Bedingungen:

(1) $-D^2(g_i)$ ist matrix-sos.

(2) g_i ist konkav auf K und für alle $u \in \overline{\mathcal{Z}(g_i) \cap \operatorname{Ex}(K)}$ gilt $D^2 g_i(u) \prec 0$.

(3) g_i ist strikt quasikonkav auf K.

Dann besitzt K eine exakte Momentenrelaxierung (bzgl. $\boldsymbol{g}$), d.h. K ist ein Spektraederschatten.

3.20 Beispiel.

1. Sei $K := \{(x, y) \in \mathbb{R}^2 \colon g(x, y) \geq 0\}$ mit $g := 1 - x^m - y^n$ und geraden Zahlen $m, n \geq 4$ (der „TV-hyperscreen“). Nach (1) ist K ein Spektraederschatten, denn es gilt
$$-D^2(g) = \begin{pmatrix} m(m-1)x^{m-2} & 0 \\ 0 & n(n-1)y^{n-2} \end{pmatrix}.$$
Jedoch sind (2) und (3) verletzt in den Punkten $u \in \{(\pm 1, 0), (0, \pm 1)\}$: Dass (2) verletzt ist, ist klar. (3) verlangt $w^\top D^2 g(u) w < 0$ für alle $0 \neq w \in \mathbb{R}^2$ mit $w \perp \nabla g(u)$, aber für $u = (1,0)^\top$ ist $\nabla g(u) = (-m, 0)^\top$, d.h. die Bedingung ist für $w = (0,1)^\top$ nicht erfüllt.

2. Sei $g := x^a y^b - 1 \in \mathbb{R}[x, y]$ mit $a, b \geq 1$. Auf $(0, \infty)^2$ ist g strikt quasikonkav, aber nicht konkav, denn für $(x, y) \neq 0$ gilt $u = (-bx, ay)^\top \perp \nabla g(x, y)$, aber die Determinante der Hessematrix ist negativ, sie ist also indefinit, d.h. g ist nicht konkav. Für $h := 1 - (x-1)^2 - (y-1)^2$ ist $K := \mathcal{S}(g, h)$ ein Spektraederschatten nach (3), aber (1) und (2) sind verletzt.

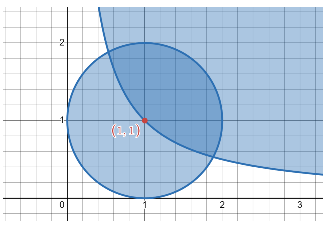


Abbildung 5: Der Spektraederschatten (die dunkelblaue Menge) aus Beispiel 3.20.2

3. Zu einer konvexen archimedischen Beschreibung $K = \mathcal{S}(\boldsymbol{g})$ kann es eine exakte Relaxierung geben, obwohl (1), (2) und (3) nicht anwendbar sind. Für ein Beispiel siehe Aufgabe 15.

4. Im Allgemeinen kann $K = \mathcal{S}(\boldsymbol{g})$ eine archimedische Beschreibung einer konvexen Menge sein, für die keine Relaxierung exakt ist, die aber dennoch ein Spektraederschatten ist. Betrachte beispielsweise $M := QM(\boldsymbol{g}) \subseteq \mathbb{R}[x, y]$ mit $\boldsymbol{g} := (y - x^3, y, 1 - y, x + 1)$. Nach Aufgabe 16 ist M archimedisch und $K := \mathcal{S}(\boldsymbol{g})$ ein Spektraederschatten.

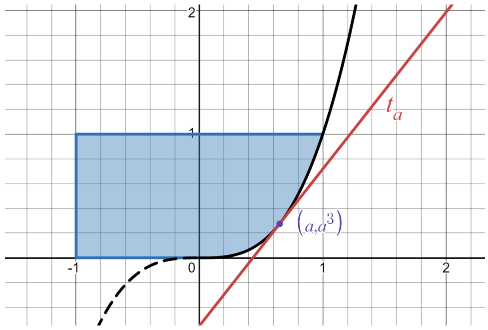


Abbildung 6: Der Spektraederschatten aus Beispiel 3.20.4

Angenommen die Relaxierung ist exakt in Grad d, d.h. nach RAG, 8.5.14 enthält M_d jedes lineare $f \in \mathbb{R}[x,y]$ mit $f|_K \geq 0$. Zu $a \in (0,1]$ seien $g := y - x^3$ und

$$t_a := \frac{\partial g}{\partial x}(a,a^3)\cdot(x-a) + \frac{\partial g}{\partial y}(a,a^3)\cdot(y-a^3) = -3a^2(x-a) + (y-a^3) = 2a^3 + 3a^2x + y,$$

dann ist $t_a \in M_d$. Sei $\varphi\colon \mathbb{R}[x,y] \to \mathbb{R}[x]$, $f \mapsto f(x,0)$ und sei $M' := \varphi(M) \subseteq \mathbb{R}[x]$. Es gilt $\varphi(M_d) \subseteq (M')_d$ und es ist $\varphi(t_a) = 2a^3 - 3a^2x = a^2(2a-3x)$, also folgt $c - x \in (M')_d$ für alle $c > 0$. Nach RAG, 8.5.11 ist $(M')_d$ abgeschlossen (in $\mathbb{R}[x]_{\leq d}$), d.h. es ist auch $-x \in (M')_d$. Daher gilt $-x = s_0 - s_1 \cdot x^3 + s_2 \cdot (x+1)$ mit $s_0, s_1, s_2 \in \mathbb{R}[x]$, aber es ist $s_0(0) = s_2(0) = 0$, d.h. die rechte Seite hat keinen linearen Term, Widerspruch.

§4 Glatte Hyperbolizitätskegel

Erinnerung: Die (offene) verallgemeinerte Lax-Vermutung besagt, dass für jede bzgl. $e \in \mathbb{R}^n$ hyperbolische Form $f \in \mathbb{R}[\mathbf{x}]$ der Hyperbolizitätskegel $C_e(f)$ ein Spektraederkegel ist.

4.1 Theorem (Netzer[10]-Sanyal[11])**.** Die Form $f \in \mathbb{R}[\mathbf{x}]$ sei strikt hyperbolisch bzgl. $e \in \mathbb{R}^n$. Dann ist $C_e(f)$ ein Spektraederschatten.

Nach Korollar 1.10 ist die Voraussetzung äquivalent dazu, dass f hyperbolisch bzgl. e ist und die Menge $\mathcal{V}(f)(\mathbb{R}) = \{u \in \mathbb{R}^n \colon f(u) = 0\}$ keine singulären Punkte $\neq 0$ besitzt.

4.2 Satz. Sei $f \in \mathbb{R}[\mathbf{x}]$ ein RZ-Polynom bzgl. $e \in \mathbb{R}^n$ (siehe 2.2).

(a) Enthält $S_e(f)$ keine Gerade, so ist f strikt quasikonkav in jedem inneren Punkt von $S_e(f)$.

(b) Ist $u \in \partial S_e(f)$ ein glatter Punkt von $\mathcal{V}(f)$ und verschwindet f auf keiner Gerade durch u, so ist f strikt quasikonkav in u.

Beweis. (a) Sei $u \in \operatorname{int} S_e(f)$. Ohne Einschränkung sei $f(u) = 1$. Die Taylorentwicklung von f um u sei $f(u+x) = 1 + f_1(x) + f_2(x) + \cdots$ mit homogenen f_i von Grad i, explizit ist dann $f_1(x) = \langle x, \nabla f(u)\rangle$ und $f_2(x) = \frac{1}{2}x^\top D^2 f(u)x$. Für $u \neq v \in \mathbb{R}^n$ hat $f(u+tv) \in \mathbb{R}[t]$ nur reelle Nullstellen und ist nicht konstant (sonst wäre $u + \mathbb{R}v \subseteq S_e(f)$). Es gilt also

$$f(u+tv) = 1 + \sum_{i\geq 1} f_i(v)t^i = \prod_{i=1}^{k}(1+\lambda_i t)$$

[10]Tim Netzer (*1980)
[11]Raman Sanyal (*1978)

mit $k \geq 1$ und $0 \neq \lambda_1, \ldots, \lambda_k \in \mathbb{R}$ (sind $0 \neq \alpha_1, \ldots, \alpha_k \in \mathbb{R}$ die Nullstellen von $f(u+tv)$, so ist $\lambda_i = -1/\alpha_i$). Es folgt $f_1(v) = \sum_{i=1}^k \lambda_i$ und

$$f_2(v) = \sum_{1 \leq i < j \leq k} \lambda_i \lambda_j = \frac{1}{2}\left(f_1(v)^2 - \sum_{i=1}^k \lambda_i^2\right).$$

Ist also $f_1(v) = 0$, so ist $f_2(v) < 0$.

(b) Man kann $e = 0$ und $f(e) = 1$ annehmen. Sei u wie in in der Voraussetzung gegeben und sei $0 \neq v \in \mathbb{R}^n$ mit $\langle v, \nabla f(u)\rangle = 0$. Nach der Lax-Vermutung (Theorem 2.3) gibt es $U, V \in \mathsf{S}^k$ mit $f(su+tv) = \det(I_k + sU + tV)$ (denn $f|_{\mathbb{R}u+\mathbb{R}v}$ ist ein RZ-Polynom bzgl. $e = 0$). Es gilt $(\mathbb{R}u + \mathbb{R}v) \cap S_e(f) = \{su + tv \mid I_k + sU + tV \succeq 0\}$. Wegen $u \in \partial S_e(f)$ ist $I_k + U \succeq 0$ und $\det(I_k + U) = 0$. Da u ein glatter Randpunkt ist, hat das Polynom

$$f(e+su) = \det(I_k + sU) = (-s)^k \cdot \det\left(-\frac{1}{s}I_k - U\right) = (-s)^k \cdot p_U\left(-\frac{1}{s}\right)$$

für $s = 1$ eine einfache Nullstelle, d.h. $\dim\ker(I_k + U) = 1$. Wegen $p_U(x) = \det(xI_k - U)$ folgt $I_k + U = \operatorname{diag}(0, 1, \ldots, 1)$ bzgl. einer geeigneten Orthonormalbasis von $\mathbb{R}^n$. Ist $V = (v_{ij})$, so folgt

$$f(u+tv) = \det(I_k + U + tV) = \det\begin{pmatrix} tv_{11} & tv_{12} & \cdots & tv_{1k} \\ tv_{12} & 1 + tv_{22} & \cdots & tv_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ tv_{1k} & tv_{2k} & \cdots & 1 + tv_{kk} \end{pmatrix}.$$

Die Leibnizformel liefert $f(u+tv) = f_1(v)t + f_2(v)t^2 + \cdots$ mit $f_1(v) = v_{11} = \langle v, \nabla f(u)\rangle$ und

$$f_2(v) = v_{11}\sum_{i=2}^n v_{ii} - \sum_{i=2}^n v_{1i}^2.$$

$\langle v, \nabla f(u)\rangle = 0$ bedeutet $v_{11} = 0$, d.h. $f_2(v) = -\sum_{i=2}^n v_{1i}^2$. Wäre $v_{11} = \cdots = v_{1n} = 0$, so wäre $f(u+tv) = \det(U + tV) = 0$ für alle $t \in \mathbb{R}$, d.h. f verschwindet auf $u + \mathbb{R}v$. Das ist nach Voraussetzung nicht der Fall, also folgt $f_2(v) < 0$. □

4.3 Beispiel. 1. Zu 4.2(a): Das Polynom

$$f := \det\begin{pmatrix} x & y \\ y & 1 \end{pmatrix} = x - y^2 \in \mathbb{R}[x,y,z]$$

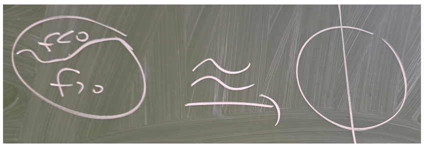
ist ein RZ-Polynom bzgl. $e = e_1$. Jedoch ist f nicht strikt quasikonkav in $u = e_1$, denn $f(e_1 + te_3) = f(e_1)$ ist konstant.

2. Zu 4.2(b): Verschwindet f auf einer Gerade $u + \mathbb{R}v$ durch $u \in \partial S_e(f)$, so ist $f(u+tv) \equiv 0$, also ist f nicht strikt quasikonkav.

4.4 Theorem. Sei $f \in \mathbb{R}[\mathbf{x}]$ ein RZ-Polynom bzgl. $e \in \mathbb{R}^n$ und sei $S_e(f)$ kompakt. Ist jeder Randpunkt von $S_e(f)$ ein glatter Punkt von $\mathcal{V}(f)$, so ist $S_e(f)$ ein Spektraederschatten.

4.5 Lemma. Sei $f \in \mathbb{R}[\mathbf{x}]$ ein RZ-Polynom bzgl. $e \in \mathbb{R}^n$ mit $f(e) > 0$ und sei $u \in \partial S_e(f)$ ein glatter Punkt von $\mathcal{V}(f)$. Dann existiert eine Umgebung U von u in $\mathbb{R}^n$ mit $U \cap S_e(f) = \{v \in U \colon f(v) \geq 0\}$.

Beweis. Es ist $\nabla f(u) \neq 0$, nach dem Satz über implizite Funktionen existieren also eine offene Umgebung U von u und ein Diffeomorphismus $\phi\colon U \to B := B_\varepsilon(0) \subseteq \mathbb{R}^n$ für ein $\varepsilon > 0$, sodass $(U, U \cap \{f = 0\})$ unter ϕ diffeomorph auf $(B, B \cap \{x_1 = 0\})$ abgebildet wird:



Weiter ist $S_e(f)$ der Abschluss einer Zusammenhangskomponente von $\{f > 0\}$ in $\mathbb{R}^n$. $\square$

Abbildung 7: Für glatte Punkte (links) ist Lemma 4.5 richtig, für singuläre Punkte (rechts) ist die Aussage jedoch im Allgemeinen falsch, wie das Bild illustriert.

4.6 Lemma. Sei $f \in \mathbb{R}[\mathbf{x}]$ ein RZ-Polynom bzgl. $e \in \mathbb{R}^n$ und sei $S_e(f)$ kompakt. Besteht $\partial S_e(f)$ aus glatten Punkten von $\mathcal{V}(f)$, so verschwindet f auf keiner Gerade durch einen Punkt aus $\partial S_e(f)$.

Beweis. Angenommen L ist eine Gerade durch $u \in \partial S_e(f)$ mit $f|_L \equiv 0$. Nach Lemma 4.5 existiert dann eine Umgebung U von u mit $U \cap \partial S_e(f) = U \cap \{f = 0\}$. Somit enthält $\partial S_e(f)$ eine Umgebung von u in L. Daher ist $\varnothing \neq L \cap \partial S_e(f)$ relativ offen in L und ist auch abgeschlossen, d.h. $L \cap \partial S_e(f) = L$. Daraus folgt $L \subseteq \partial S_e(f)$, im Widerspruch zur Kompaktheit von $S_e(f)$. $\square$

4.7 Theorem. Sei $f \in \mathbb{R}[\mathbf{x}]$ ein RZ-Polynom bzgl. $e \in \mathbb{R}^n$ und sei $S_e(f)$ kompakt. Für jeden Randpunkt $u \in \partial S_e(f)$ und jeden irreduziblen Faktor g von f mit $g(u) = 0$ sei u ein glatter Punkt von $\mathcal{V}(g)$ und g verschwinde auf keiner Gerade durch u. Dann ist $S_e(f)$ ein Spektraederschatten.

Beweis von Theorem 4.4 unter Verwendung von Theorem 4.7. Sei $u \in \partial S_e(f)$ (ein glatter Punkt von f). Ist g ein irreduzibler Faktor von f mit $g(u) = 0$, so gilt $g|_L \not\equiv 0$ für jede Gerade L durch u (Lemma 4.6). Damit folgt die Behauptung aus Theorem 4.7. $\square$

Beweis von Theorem 4.7. Sei $f = f_1 \cdots f_m$ mit irreduziblen Polynomen f_i. Jedes f_i ist RZ bzgl. e und es gilt

$$S_e(f) = \bigcap_{i=1}^{m} S_e(f_i).$$

Seien $f_1, \ldots, f_k$ diejenigen f_i mit $f_i(u) = 0$. Es genügt, für jedes $u \in \partial S_e(f)$ und jedes $i \in [k]$ zu zeigen, dass $\overline{B}_{r_i}(u) \cap S_e(f_i)$ für ein $r_i > 0$ ein Spektraederschatten ist, denn dann gibt es für jedes $u \in \partial S_e(f)$ ein $r > 0$, sodass $\overline{B}_r(u) \cap S_e(f)$ ein Spektraederschatten ist (wähle $r \leq \min\{r_1, \ldots, r_k\}$ so klein, dass $f_i > 0$ auf $\overline{B}_r(u)$ für alle $i \in \{k+1, \ldots, m\}$ gilt). Überdeckt man $\partial S_e(f)$ nun durch endlich viele dieser $\overline{B}_r(u)$ ($\partial S_e(f)$ ist kompakt), so folgt mit Satz II.4.5 (Konvexität), dass auch $S_e(f) = \operatorname{conv}(\partial S_e(f))$ ein Spektraederschatten ist.

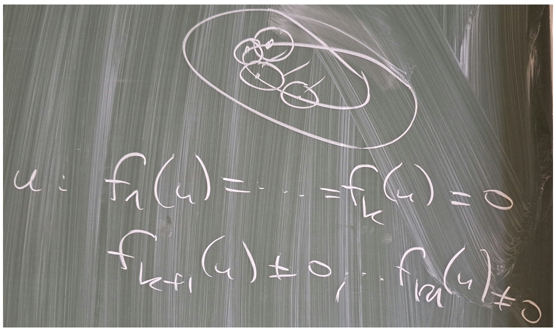
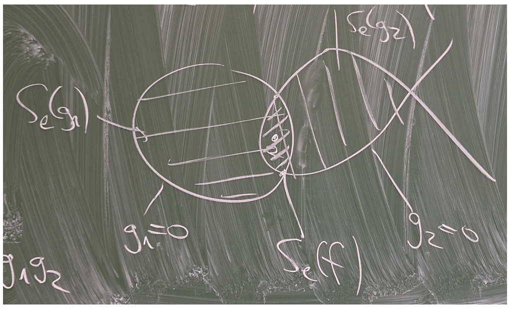


Abbildung 8: Zerlegung von $S_e(f)$ in kleinere Spektraederschatten

Sei f also nun ohne Einschränkung ein irreduzibles RZ-Polynom mit $f(e) > 0$ und sei $u \in \partial S_e(f)$ ein glatter Punkt von f mit $f \not\equiv 0$ auf jeder Gerade durch u. Es reicht zu zeigen, dass ein $r > 0$ existiert, sodass $\overline{B}_r(u) \cap S_e(f)$ ein Spektraederschatten ist. Nach Satz 4.2(b) ist f strikt quasikonkav in u, nach Aufgabe 17 ist f also auch strikt quasikonkav auf $\overline{B}_r(u)$ für ein kleines $r > 0$. Das Polynom $r^2 - |x-u|^2$ ist überall strikt quasikonkav. Daher folgt mit Theorem 3.13/3.19, dass die konvexe Menge $\overline{B}_r(u) \cap S_e(f) = \mathcal{S}(f, r^2 - |x-u|^2)$ ein Spektraederschatten ist. □

4.8 Bemerkung. $\partial S_e(f)$ in Theorem 4.7 muss im Allgemeinen **nicht** aus glatten Punkten von f bestehen, z.B. für $g_1 \coloneqq (3/2)^2 - (x+2)^2 - y^2$, $g_2 \coloneqq (x-1)^2(x+1) - y^2$ und $S_e(f) = \mathcal{S}(g_1, g_2)$:



4.9 –

Mithilfe von Theorem 4.4 können wir nun folgende Verallgemeinerung von Theorem 4.1 beweisen:

4.10 Theorem. Sei $f \in \mathbb{R}[\mathbf{x}]$ eine bzgl. $e \in \mathbb{R}^n$ hyperbolische Form. Jeder Punkt $\neq 0$ von $\partial C_e(f)$ sei ein glatter Punkt von $\mathcal{V}(f)$. Dann ist $C_e(f)$ ein Spektraederschatten.

Beweis. Sei $C \coloneqq C_e(f)$, sei $L \coloneqq C \cap (-C)$ und sei $\mathbb{R}^n = L \oplus W$ mit einem linearen Unterraum $W \subseteq \mathbb{R}^n$ mit $e \in W$. Dann gilt $C = L \oplus (C \cap W)$, es genügt also zu zeigen, dass $C \cap W$ ein Spektraederschatten ist. Da $f|_W$ hyperbolisch bzgl. e mit Hyperbolizitätskegel $C \cap W$ ist, kann man C durch $C \cap W$ ersetzen. Alle $u \in \partial(C \cap W)$ sind glatte Punkte von f. Ohne Einschränkung sei C daher spitz. Nach Satz I.5.12 (Konvexität) gibt es also eine affine Hyperebene $H \subseteq \mathbb{R}^n$ mit $e \in H$, sodass $K \coloneqq C \cap H$ kompakt ist. $f|_H$ ist ein RZ-Polynom bzgl. e mit $S_e(f|_H) = C \cap H = K$ und glatten Randpunkten. Nach Theorem 4.4 ist K daher ein Spektraederschatten. Wegen $C = K^h$ ist damit auch C ein Spektraederschatten (Konvexität, II.4.8). □

§5 Interlacer

5.1 Definition. (a) Ein Polynom $0 \neq f \in \mathbb{R}[t]$ heißt *real-rooted*, falls alle Nullstellen von f (in $\mathbb{C}$) reell sind. Sind zusätzlich alle Nullstellen von f einfach, so heißt f *strictly real-rooted*.

(b) Sei $f \in \mathbb{R}[t]$ real-rooted mit $d = \deg(f) \geq 1$. Ein Polynom $g \in \mathbb{R}[t]$ heißt ein *interlacer* von f, falls $g = 0$ gilt, oder g real-rooted ist mit $\deg(g) = d - 1$, sodass gilt: Sind $\alpha_1 \leq \cdots \leq \alpha_d$ die Nullstellen von f und sind $\beta_1 \leq \cdots \leq \beta_{d-1}$ die Nullstellen von g, so gilt

$$\alpha_1 \leq \beta_1 \leq \alpha_2 \leq \beta_2 \leq \cdots \leq \alpha_{d-1} \leq \beta_{d-1} \leq \alpha_d. \qquad (*)$$

Sind alle Ungleichungen in $(*)$ strikt, so heißt g ein *strikter interlacer* von f. Schreibe $\operatorname{Int}(f)$ für die Menge aller interlacer von f.

5.2 Bemerkung. Ist f real-rooted, so ist $f' = \frac{\partial f}{\partial t}$ nach dem Satz von Rolle ein interlacer von f. Dieser ist genau dann strikt, wenn f stricly real-rooted ist. Ist $g := \operatorname{ggT}(f, f')$ und ist $f = f_1 g$, so hat f_1 nur einfache Nullstellen und es gilt $\operatorname{Int}(f) = g \cdot \operatorname{Int}(f_1)$. Die Menge $\operatorname{Int}(f)$ ist abgeschlossen in $\mathbb{R}[t]_{\leq d-1}$ und ihr Inneres besteht aus den strikten interlacern von f.

5.3 Satz. Seien $f, g \in \mathbb{R}[t]$ mit $\deg(f) = d$ und $\deg(g) = d - 1$ und sei f strictly real-rooted. Dann sind die folgenden Bedingungen äquivalent:

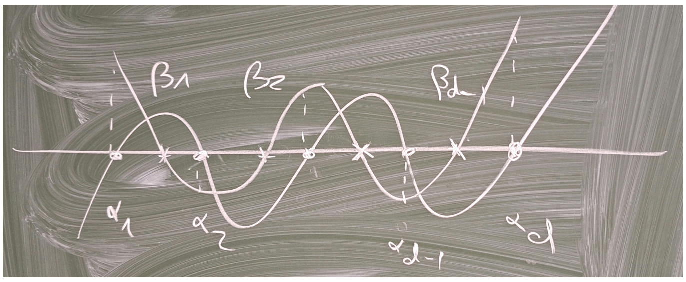
(i) g ist ein strikter interlacer von f.

(ii) $f'g$ ist strikt definit auf $\mathcal{V}(f) := \{\alpha \in \mathbb{R} \colon f(\alpha) = 0\}$.

(iii) Das *Wronksi*[12]*-Polynom* $W(f, g) := f'g - fg'$ besitzt keine reellen Nullstellen (d.h. ist definit).

Die nicht-strikten Versionen von (i)–(iii) sind ebenfalls äquivalent.

Beweis. Zeige die Äquivalenz der strikten Versionen. Ohne Einschränkung haben f und g positive Leitkoeffizienten (sonst ersetze f durch $-f$ bzw. g durch $-g$). Sei $\mathcal{V}(f) = \{\alpha_1 < \cdots < \alpha_d\}$. Die Vorzeichen $\operatorname{sign} f'(\alpha_i) = (-1)^{d-i} \neq 0$ alternieren:



Gilt (i), so gilt $f'(\alpha_i)g(\alpha_i) > 0$ für alle i, also (ii). (ii) $\Rightarrow$ (i) ist klar nach dem Zwischenwertsatz. (iii) $\Rightarrow$ (ii) ist ebenfalls klar. (ii) $\Rightarrow$ (iii): Für $j \in [d]$ gilt

$$\frac{g}{f} = \sum_{j=1}^{d} \frac{c_j}{t - \alpha_j} \quad \text{mit} \quad c_j := \frac{g(\alpha_j)}{f'(\alpha_j)}, \qquad (*)$$

denn nach Multiplikation mit f sind beide Seiten Polynome vom Grad $\leq d - 1$, die in $\alpha_1, \ldots, \alpha_d$ denselben Wert haben (aus $f = (t - \alpha_j)h$ mit $h \in \mathbb{R}[t]$ folgt mit der Produktregel $f'(\alpha_j) = h(\alpha_j)$, d.h. es gilt $(f/(t - \alpha_j))(\alpha_j) = f'(\alpha_j)$). Aus (ii) folgt $c_j \neq 0$ für alle $j \in [d]$ und dass alle c_j dasselbe Vorzeichen haben. Ableiten von $(*)$ liefert

$$\frac{g'f - gf'}{f^2} = -\sum_{j=1}^{d} \frac{c_j}{(t - \alpha_j)^2},$$

[12] Jósef Maria Hoëné-Wroński (1776–1853)

also hat

$$W(f,g) = f'g - fg' = f^2 \sum_{j=1}^{d} \frac{c_j}{(t-\alpha_j)^2}$$

keine Nullstelle in $\mathbb{R}$, d.h. es gilt (iii). Die Äquivalenz der nicht-strikten Versionen folgt durch Grenzübergang. $\square$

Bemerkung. Ist f real-rooted mit einer mehrfachen Nulstelle, so gilt noch (i) $\Rightarrow$ (ii) $\wedge$ (iii) für die nicht-strikten Versionen, aber im Allgemeinen nicht umgekehrt. Beispielsweise für $f = t^3$ und $g = t^2 + t + 1$ gilt

$$W(f,g) = f'g - fg' = 3t^2(t^2+t+1) - t^3(2t+1) = t^2(\underbrace{t^2+2t+3}_{>0 \text{ auf } \mathbb{R}}).$$

Hier sind (ii) und (iii) erfüllt, aber (i) nicht.

5.4 Definition. Seien $f, g \in \mathbb{R}[\mathbf{x}]$ homogen mit $\deg(f) = d$ und $\deg(g) = d-1$ und sei f hyperbolisch bzgl. $e \in \mathbb{R}^n$. g heißt ein *interlacer* von f bzgl. e, falls das univariate Polynom $g(te+u) \in \mathbb{R}[t]$ für jedes $u \in \mathbb{R}^n$ ein interlacer von $f(te+u) \in \mathbb{R}[t]$ ist. Ein interlacer g von f heißt ein *strikter interlacer* von f bzgl. e, falls $g(te+u)$ für ein (sic!) $u \in \mathbb{R}^n$ ein strikter interlacer von $f(te+u)$ ist. Sei

$$\mathrm{Int}_e(f) \coloneqq \{0\} \cup \{g \in \mathbb{R}[\mathbf{x}]_{d-1} \colon f(e)g(e) > 0,\ g \text{ ist ein interlacer von } f \text{ bzgl. } e\}.$$

5.5 Bemerkung. Sei $f \in \mathcal{H}_d(e)$.

1. Ist $0 \neq g \in \mathrm{Int}_e(f)$, so ist $g \in \mathcal{H}_{d-1}(e)$. Es ist $\partial_e f \in \mathrm{Int}_e(f)$ und es gilt $f(e) \cdot (\partial_e f)(e) > 0$. $\partial_e f$ ist genau dann ein strikter interlacer von f, wenn f quadratfrei ist: Ist f quadratfrei, so hat $V \coloneqq \mathcal{V}(f)$ glatte $\mathbb{R}$-Punkte (V_{reg} ist offen-dicht in V und $V(\mathbb{R})$ ist Zariski-dicht in V). Es folgt die Existenz eines $u \in V(\mathbb{R})$, sodass $f(te+u)$ nur einfache Nullstellen hat. Die Abbildung
$$V(\mathbb{R}) \to \mathcal{S}^{n-1}, \quad u \mapsto \frac{u-e}{|u-e|}$$
ist surjektiv, also gibt es eine Richtung $u \in \mathbb{R}^n$, in der kein singulärer Punkt von V liegt (es ist $\dim(V(\mathbb{R})) = \dim(\mathcal{S}^{n-1}) = n-1$ und $\dim(V_{\mathrm{sing},\mathbb{R}}) \leq n-2$).
2. Jeder interlacer von f verschwindet in $V_{\mathrm{sing},\mathbb{R}}(f) = \{u \in \mathbb{R}^n \colon f(u) = 0,\ \nabla f(u) = 0\}$.
3. Ist g ein strikter interlacer von f, so ist die Menge
$$\{u \in \mathbb{R}^n \colon g(te+u) \text{ ist ein strikter interlacer von } f(te+u)\}$$
Zarisiki-offen in $\mathbb{R}^n$, denn das ist die Menge $\{u \in \mathbb{R}^n \colon \mathrm{Res}(g(te+u), f(te+u)) \neq 0\}$.
4. Für $g \in \mathbb{R}[\mathbf{x}]_{d-1}$ ist die Menge $\{u \in \mathbb{R}^n \colon g(te+u)$ ist ein interlacer von $f(te+u)\}$ abgeschlossen. Ist diese Menge also dicht in $\mathbb{R}^n$, so ist g ein interlacer von f.
5. Ist $f = f_1 h$ mit $h \mid f_1$, $f \in \mathcal{H}_d(e)$ und $h(e) > 0$, so ist $\mathrm{Int}_e(f) = h \cdot \mathrm{Int}_e(f_1)$.

5.6 Theorem. Sei $f \in \mathcal{H}_d(e)$ quadratfrei. Für $g \in \mathbb{R}[\mathbf{x}]_{d-1}$ sind äquivalent:

(i) Es gilt $g \in \mathrm{Int}_e(f)$.

(ii) Für alle $u \in U_e(f)$ gilt $g \in \mathrm{Int}_u(f)$.

(iii) Es gilt $(\partial_e f) \cdot g \geq 0$ auf $\mathcal{V}_{\mathbb{R}}(f)$.

(iv) Es gilt $(\partial_e f) \cdot g - f \cdot (\partial_e g) \geq 0$ auf $\mathbb{R}^n$.

Beweis. Ohne Einschränkung sei $f(e) > 0$. $g \coloneqq \partial_e f$ ist ein strikter interlacer von f, da f quadratfrei ist (5.5.1). Die Menge $U \coloneqq \{u \in \mathbb{R}^n \colon f(te+u) \text{ hat nur einfache Nullstellen}\}$ ist offen-dicht (Aufgabe 21). Für $u \in U$ kann man also Satz 5.3 auf $f(te+u)$ und $g(te+u)$ anwenden, das liefert (i) $\Leftrightarrow$ (iii) $\Leftrightarrow$ (iv). (ii) $\Rightarrow$ (i) ist klar. (i) $\Rightarrow$ (ii): Sei $g \in \mathrm{Int}_e(f)$ und sei $u \in U_e(f)$. Um $g \in \mathrm{Int}_u(f)$ zu zeigen, zeige $(\partial_u f) \cdot g \geq 0$ auf $\mathcal{V}_{\mathbb{R}}(f)$. Sei also $v \in \mathcal{V}_{\mathbb{R}}(f)$. Ohne Einschränkung sei $g(v) \neq 0$ und sei v ein nichtsingulärer Punkt von f (sonst ist $(\partial_u f)(v) = 0$). Für alle $w \in [e,u]$ ist f hyperbolisch bzgl. w und da v glatt ist, gilt auch $(\partial_w f)(v) \neq 0$. Die Abbildung $w \mapsto (\partial_w f)(v)$ hat daher keine Nullstellen, also ist $(\partial_e f)(v) \cdot (\partial_u f)(v) > 0$ (denn $(\partial_u f)(v)$ hat das gleiche Vorzeichen wie $(\partial_e f)(v)$). Wegen $(\partial_e f)(v) \cdot g(v) > 0$ (nach (iii)) folgt $(\partial_u f)(v) \cdot g(v) > 0$. □

5.7 Korollar. Sei $f \in \mathcal{H}_d(e)$ quadratfrei. Für $g_1, g_2 \in \mathrm{Int}_e(f)$ gilt dann $g_1 g_2 \geq 0$ auf $\mathcal{V}_{\mathbb{R}}(f)$.

Beweis. Sei $v \in \mathcal{V}_{\mathbb{R}}(f)$. Ist v singulär, so ist $g_1(v) = g_2(v) = 0$. Andernfalls ist $(\partial_e f)(v) \neq 0$, d.h. $(g_1 g_2)(v) \geq 0$ folgt aus Theorem 5.6 (i) $\Rightarrow$ (iii) (angewendet auf g_1 bzw. g_2). □

5.8 Korollar. Sei $f \in \mathcal{H}_d(e)$. Dann ist $\mathrm{Int}_e(f)$ ein abgeschlossener konvexer Kegel in $\mathbb{R}[\mathbf{x}]_{d-1}$ und für alle $u \in U_e(f)$ gilt $\mathrm{Int}_e(f) = \mathrm{Int}_u(f)$. Ist f quadratfrei, so ist

$$\mathrm{Int}_e(f) = \{g \in \mathbb{R}[\mathbf{x}]_{d-1} \colon (\partial_e f) \cdot g - f \cdot (\partial_e g) \geq 0 \text{ auf } \mathbb{R}^n\}.$$

Insbesondere ist $\mathrm{Int}_e(f)$ in diesem Fall ein linearer Schnitt von $P_{n,2d-2}$.

Beweis. Nach Bemerkung 5.5.5 kann man annehmen, dass f quadratfrei ist, d.h. alle Aussagen bis auf die letzte folgen aus Theorem 5.6. Die Abbildung $\phi \colon \mathbb{R}[\mathbf{x}]_{d-1} \to \mathbb{R}[\mathbf{x}]_{2d-2}$, $g \mapsto (\partial_e f) \cdot g - f \cdot (\partial_e g)$ ist linear und (für quadratfreies f) ist $\mathrm{Int}_e(f)$ das Urbild von $P_{n,2d-2}$ unter ϕ. ϕ ist injektiv, denn für quadratfreies f gibt es einen irreduziblen Faktor von f der $(\partial_e f) \cdot g$ nicht teilt. □

5.9 Definition. Für $f \in \mathbb{R}[\mathbf{x}]_d$ und $u, v \in \mathbb{R}^n$ sei $\Delta_{u,v} f \coloneqq (\partial_u f) \cdot (\partial_v f) - f \cdot (\partial_u \partial_v f) \in \mathbb{R}[\mathbf{x}]_{2d-2}$. Die Abbildung

$$\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}[\mathbf{x}]_{2d-2}, \quad (u,v) \mapsto \Delta_{u,v} f$$

ist bilinear und symmetrisch. Für $g \in \mathbb{R}[\mathbf{x}]_d$ gilt $\Delta_{u,v}(fg) = f^2 \cdot \Delta_{u,v} g + g^2 \cdot \Delta_{u,v} f$.

5.10 Theorem. Sei $f \in \mathcal{H}_d(e)$ quadratfrei. Dann ist

$$C_e(f) = \{u \in \mathbb{R}^n \colon \partial_u f \in \mathrm{Int}_e(f)\} = \{u \in \mathbb{R}^n \colon \Delta_{e,u} f \geq 0 \text{ auf } \mathbb{R}^n\}$$

(ein linearer Schnitt von $P_{n,2d-2}$).

Beweis. Die Abbildung $\mathbb{R}^n \to \mathbb{R}[\mathbf{x}]_{2d-2}$, $u \mapsto \Delta_{e,u} f$ ist linear und injektiv. Die zweite Gleichheit folgt mit Theorem 5.6 angewandt auf $g = \partial_u f$. Um die erste Gleichheit zu zeigen, kann man ohne Einschränkung $f(e) > 0$ annehmen. '$\subseteq$': Für $u \in U_e(f)$ ist $\partial_u f \in \mathrm{Int}_u(f) = \mathrm{Int}_e(f)$ (5.6). Da $\mathrm{Int}_e(f)$ abgeschlossen ist, gilt das auch für $u \in C_e(f)$. Außerdem ist $\{u \in \mathbb{R}^n \colon \partial_u f \in \mathrm{Int}_e(f)\}$ ein abgeschlossener konvexer Kegel in $\mathbb{R}^n$ (das Urbild von $\mathrm{Int}_e(f)$ unter $u \mapsto \partial_u f$) mit nichtleerem Inneren (enthält $U_e(f)$). '$\supseteq$': Angenommen es gibt ein $u \notin C_e(f)$ mit $\partial_e f \in \mathrm{Int}_e(f)$, dann gibt es auch ein solches u mit $f(u) \neq 0$. Weiter gilt auch $-u \notin C_e(f)$, denn sonst wäre $\partial_{-u} f = -\partial_u f \in \mathrm{Int}_e(f)$ ('$\subseteq$' schon gezeigt), also $(\partial_u f)(e) > 0$ und $(-\partial_u f)(e) > 0$, Widerspruch. Wegen $f(u) \neq 0$ hat das univariate Polynom $f(e+su) \in \mathbb{R}[s]$ Grad d und ist real-rooted, die Ableitung $(\partial_u f)(e+su)$ ist also ein interlacer davon. Seien $b_1 \leq \cdots \leq b_{d-1}$ die Nullstellen von $(\partial_u f)(e+su)$. Diese sind alle von Null verschieden. Die Nullstellen von $(\partial_u f)(te+u)$ sind $b_1^{-1}, \ldots, b_{d-1}^{-1}$. Seien $a_1 \leq \cdots \leq a_d$ die

Nullstellen von $f(te+u)$, diese sind ebenfalls von Null verschieden. Wegen $\pm u \notin C_e(f)$ ist $a_1 < 0$ und $a_d > 0$: f hat eine Nullstelle in (e,u) und eine in $(e,-u)$, denn

$$f(se+(1-s)u) = 0 = f(s'e-(1-s')u) \quad \text{mit} \quad 0 < s, s' < 1$$

ist äquivalent zu

$$f\Big(\underbrace{\frac{s}{1-s}}_{>0}e+u\Big) = 0 = f\Big(\underbrace{\frac{-s'}{1-s'}}_{<0}e+u\Big).$$

Sei $k \in [d]$ mit $a_k < 0 < a_{k+1}$. Die Nullstellen von $f(e+su)$ sind $a_1^{-1}, \ldots, a_d^{-1}$. Wegen interlacing gilt

$$\underbrace{\frac{1}{a_k} \le b_1 \le \frac{1}{a_{k-1}} \le b_2 \le \cdots \le \frac{1}{a_1}}_{<0} \le b_k \le \underbrace{\frac{1}{a_d} \le b_{k+1} \le \cdots \le b_{d-1} \le \frac{1}{a_{k+1}}}_{>0}.$$

Für $b_k \neq 0$ gibt es zwei Möglichkeiten, beide ergeben einen Widerspruch:

- Ist $b_k < 0$, so ist b_k^{-1} die kleinste Nullstelle von $(\partial_u f)(te+u)$ und b_1^{-1} ist die k-te Nullstelle:

  $$a_1 \ge \frac{1}{b_k} \ge a_2 \ge \frac{1}{b_{k-1}} \ge \cdots \ge a_k \ge \frac{1}{b_1} < 0$$

  Es folgt $a_1 = b_k^{-1}$, also (a_1 ist eine mehrfache Nullstelle von $f(te+u)$) $a_1 = a_2 = b_{k-1}^{-1}$. Induktiv folgt $a_1 = \cdots = a_k = b_1^{-1} = \cdots = b_k^{-1} < 0$ und diese Zahl ist eine genau k-fache Nullstelle sowohl von $f(te+u)$, als auch von seiner Ableitung, Widerspruch.

- Ist $b_k > 0$, so erhält man analog einen Widerspruch. □

5.11 Korollar. Ist $f \in \mathcal{H}_d(e)$ und sind $u, v \in C_e(f)$, so ist $\Delta_{u,v} f \in \mathbb{R}[\mathbf{x}]_{2d-2}$ psd auf $\mathbb{R}^n$.

Beweis. Sei $X = (x_{ij})$ die allgemeine symmetrische $d \times d$-Matrix (d.h. die $x_{ij} = x_{ji}$ sind Variablen) über $\mathbb{R}[X]$. Die Form $\det(X) \in \mathbb{R}[X]_d$ ist hyperbolisch bzgl. I_d mit $C_{I_d}(\det(X)) = \mathsf{S}_+^d$. Für jedes $A \in \mathsf{S}_+^d$ ist $\partial_A \det(X)$ ein interlacer von $\det(X)$. Nach Aufgabe 24 gilt $\partial_A \det(X) = \operatorname{tr}(A \cdot X^{\mathrm{adj}})$. □

5.12 Korollar. Ist $e \in \mathbb{R}^n$ und $A(x) = \sum_{i=1}^n x_i A_i$ mit $A_i \in \mathsf{S}^d$ und $A(e) \succ 0$, so ist $\operatorname{tr}(B \cdot A(x)^{\mathrm{adj}})$ für jedes $B \in \mathsf{S}_+^d$ ein interlacer von $\det A(x)$ bzgl. e. □

5.13 Satz. Sei X die allgemeine symmetrische $d \times d$-Matrix wie im Beweis von Korollar 5.11. Für $U, V \in \mathsf{S}_+^d$ ist das Polynom $\Delta_{U,V} \det(X) \in \mathbb{R}[X]_{2d-2}$ sos.

Beweis. Es genügt, $U = uu^\top$ und $V = vv^\top$ mit $u, v \in \mathbb{R}^n$ zu betrachten. Es gilt

$$\partial_U \det(X) = -\begin{vmatrix} X & u \\ u^\top & 0 \end{vmatrix},$$

anwenden von Aufgabe 24 auf

$$\begin{pmatrix} X & u \\ u^\top & 0 \end{pmatrix} \quad \text{und} \quad \begin{pmatrix} v \\ 0 \end{pmatrix} \quad \text{liefert also} \quad \partial_U \partial_V \det(X) = \begin{vmatrix} X & u & v \\ u^\top & 0 & 0 \\ v^\top & 0 & 0 \end{vmatrix}.$$

Nun gilt

$$\Delta_{u,v} \det(X) = (\partial_U \det(X)) \cdot (\partial_V \det(X)) - \det(X) \cdot \partial_U \partial_V \det(X)$$

$$= \begin{vmatrix} X & u \\ u^\top & 0 \end{vmatrix} \cdot \begin{vmatrix} X & v \\ v^\top & 0 \end{vmatrix} - |X| \cdot \begin{vmatrix} X & u & v \\ u^\top & 0 & 0 \\ v^\top & 0 & 0 \end{vmatrix} \overset{5.14}{=} \begin{vmatrix} X & v \\ u^\top & 0 \end{vmatrix} \cdot \begin{vmatrix} X & u \\ v^\top & 0 \end{vmatrix} = \begin{vmatrix} X & v \\ u^\top & 0 \end{vmatrix}^2 . \qquad \square$$

5.14 Lemma. Seien $X \in \mathbb{R}^{n\times n}$ und $a, b, c, d \in \mathbb{R}^n$. Dann gilt

$$\begin{vmatrix} X & b \\ a^\top & 0 \end{vmatrix} \cdot \begin{vmatrix} X & d \\ c^\top & 0 \end{vmatrix} - \begin{vmatrix} X & d \\ a^\top & 0 \end{vmatrix} \cdot \begin{vmatrix} X & b \\ c^\top & 0 \end{vmatrix} = |X| \cdot \begin{vmatrix} X & b & d \\ a^\top & 0 & 0 \\ c^\top & 0 & 0 \end{vmatrix} .$$

Beweis. Durch Anwendung des Schur-Komplements (B1, Aufgabe 40) auf die rechte Seite folgt

$$D := \begin{vmatrix} X & b & d \\ a^\top & 0 & 0 \\ c^\top & 0 & 0 \end{vmatrix} = |X| \cdot \det\left(-\begin{pmatrix} a^\top \\ c^\top \end{pmatrix} X^{-1}(b, d)\right)$$
$$= |X|^{-1} \cdot \det\left(-\begin{pmatrix} a^\top \\ c^\top \end{pmatrix} X^{\mathrm{adj}}(b, d)\right) = |X|^{-1} \cdot \begin{vmatrix} a^\top X^{\mathrm{adj}} b & a^\top X^{\mathrm{adj}} d \\ c^\top X^{\mathrm{adj}} b & c^\top X^{\mathrm{adj}} d \end{vmatrix} .$$

Nach Aufgabe 24 ist $a^\top X^{\mathrm{adj}} b = -\left|\begin{smallmatrix} X & b \\ a^\top & 0 \end{smallmatrix}\right|$, also folgt

$$D = |X|^{-1} \cdot \left(\begin{vmatrix} X & b \\ a^\top & 0 \end{vmatrix} \cdot \begin{vmatrix} X & d \\ c^\top & 0 \end{vmatrix} - \begin{vmatrix} X & d \\ a^\top & 0 \end{vmatrix} \cdot \begin{vmatrix} X & b \\ c^\top & 0 \end{vmatrix} \right) . \qquad \square$$

5.15 Theorem. Sei $f \in \mathcal{H}_d(e)$. Hat f^r eine definite Determinantendarstellung für ein $r \geq 1$, so ist $\Delta_{u,v} f$ sos für beliebige $u, v \in C_e(f)$.

Beweis. Nach Voraussetzung ist $f = \det A(x)$ mit $A(e) \succ 0$. Sei zunächst $r = 1$ und argumentiere wie im Beweis von Korollar 5.12. Die Abbildung $\phi\colon \mathbb{R}[X] \to \mathbb{R}[x_1, \dots, x_n]$, $x_{ij} \mapsto A(x)_{ij}$ erfüllt

$$\phi(\det(X)) = \det A(x) = f, \quad \phi(\partial_{A(w)} \det(X)) = \partial_w f \quad \text{und} \quad \phi(\Delta_{A(u),A(v)} \det(X)) = \Delta_{u,v} f.$$

Sind $U, V \succeq 0$, so ist $\Delta_{U,V} \det(X)$ nach Satz 5.13 sos. Seien $u, v \in C_e(f)$, dann gilt $A(u) \succeq 0$ und $A(v) \succeq 0$, denn ϕ bildet $\Delta_{U,V} \det(X)$ auf $\Delta_{u,v} f$ ab. Sei nun $r \geq 1$. Nach 5.9 ist

$$\Delta_{u,v}(fg) = f^2 \Delta_{u,v}(g) + g^2 \Delta_{u,v}(f), \quad \text{d.h.} \quad \Delta_{u,v}(f^2) = 2f^2 \Delta_{u,v}(f).$$

Zeige induktiv $\Delta_{u,v}(f^r) = r f^{2r-2} \Delta_{u,v}(f)$ für alle r. Sei $f^r = \det A(x)$ mit $A(e) \succ 0$. Wegen $u, v \in C_e(f) = C_e(f^r)$ folgt aus den Fall $r = 1$, dass $\Delta_{u,v}(f^r) = r f^{2r-2} \Delta_{u,v}(f)$ sos ist, also ist $\Delta_{u,v} f$ sos, denn jeder irreduzible Faktor von f ist reell. $\square$

§6 Gegenbeispiele zur Helton-Nie-Vermutung

6.1 Notation. Für $S \subseteq \mathbb{R}^n$ schreibe

$$P_S := \{f \in \mathbb{R}[\mathbf{x}]_{\leq 1} \colon f|_S \geq 0\} = \mathbb{R}[\mathbf{x}]_{\leq 1} \cap \mathcal{P}(S) \quad \text{und}$$
$$P_{S,0} := \{f \in P_S \colon f(0) = 0\},$$

dann gilt $\overline{\mathrm{conv}(S)} = \{u \in \mathbb{R}^n \colon \forall f \in P_S\colon f(u) \geq 0\}$ und $\overline{\mathrm{cone}(S)} = \{u \in \mathbb{R}^n \colon \forall f \in P_{S,0}\colon f(u) \geq 0\}$.

6.2 Lemma. Seien $S \subseteq \mathbb{R}^n$, $K := \overline{\mathrm{conv}(S)}$ und $C := \overline{\mathrm{cone}(S)}$.

(a) K ist ein affin-linearer Schnitt von $(P_S)^*$ und K ist genau dann ein Spektraederschatten, wenn P_S einer ist.

(b) Es gilt $C = (P_{S,0})^*$ und C ist genau dann ein Spektraederschatten, wenn $P_{S,0}$ einer ist.

Beweis. Es ist $(P_S)^* = \{(v_0, \dots, v_n) \in \mathbb{R}^{n+1} : \forall f \in P_S : \widetilde{f}(v) \geq 0\}$ (mit $\widetilde{f} := \sum_{i=0}^n a_i x_i \in \mathbb{R}[x_0, \mathbf{x}]$ für $f = a_0 + \sum_{i=1}^n a_i x_i$) und damit $K = \{u \in \mathbb{R}^n : (1, u_1, \dots, u_n) \in P_S^*\}$. Wegen $P_S = P_K = (K_h)^*$ folgt die Äquivalenz in (a) aus Konvexität. (b) folgt analog. □

6.3 Definition. Sei V eine affine $\mathbb{R}$-Varietät, sei $S \subseteq V(\mathbb{R})$ und sei $L \subseteq \mathbb{R}[V]$ ein Untervektorraum mit $\dim(L) < \infty$. S hat *uniforme sos-Darstellungen* für L, falls eine affine $\mathbb{R}$-Varietät X, ein Morphismus $\phi : X \to V$ von affinen $\mathbb{R}$-Varietäten und ein Untervektorraum $U \subseteq \mathbb{R}[X]$ mit

$$(1)\ \dim(U) < \infty, \qquad (2)\ S \subseteq \phi(X(\mathbb{R})), \qquad (3)\ \phi^*(L \cap \mathcal{P}(S)) \subseteq \Sigma U^2$$

existieren. Dabei ist $\phi^* : \mathbb{R}[V] \to \mathbb{R}[X]$ der zu ϕ gehörige duale Ringhomomorphismus.

6.4 Bemerkung. Für eine semialgebraische Menge $S \subseteq V(\mathbb{R})$ sei $\mathcal{A}_0(S)$ der Ring der definierbaren Funktionen $S \to \mathbb{R}$. Dann sind die Bedingungen aus Definition 6.3 für S und L genau dann erfüllt, wenn es $h_1, \dots, h_N \in \mathcal{A}_0(S)$ gibt, sodass $f|_S$ für jedes $f \in L \cap \mathcal{P}(S)$ eine Summe von Quadraten von Linearkombinationen von $h_1, \dots, h_N$ ist.

Beweis. '$\Rightarrow$': Gelte 6.3 mit $\phi : X \to V$ und $U \subseteq \mathbb{R}[X]$ ($\dim(U) < \infty$). Wegen $S \subseteq \phi(X(\mathbb{R}))$ gibt es nach RAG I einen definierbaren Schnitt $\sigma : S \to X(\mathbb{R})$ von π über S (d.h. $\pi \circ \sigma = \mathrm{id}_S$). Ist $g_1, \dots, g_N$ eine Basis von U, so gilt 6.4 für $h_i := g_i \circ \sigma : S \to \mathbb{R}$.
'$\Leftarrow$': Gelte 6.4 mit $h_1, \dots, h_N : S \to \mathbb{R}$ und wähle $X := \overline{\mathrm{graph}(h_1, \dots, h_N)} \subseteq V \times \mathbb{A}^N$ (Zariski-Abschluss). Es gilt

$$\mathrm{graph}(h_1, \dots, h_N) = \{(s, h_1(s), \dots, h_N(s)) \mid s \in S\} \subseteq S \times \mathbb{R}^N \subseteq (V \times \mathbb{A}^N)(\mathbb{R}).$$

Ist $\phi : X \to V$ kanonisch, so ist 6.3 erfüllt für $U := \mathrm{span}(g_1, \dots, g_N)$, wobei $g_i : X \to \mathbb{A}^i$ die Projektion auf die i-te Komponente ist. (Ist $f|_S \geq 0$ und $f = \sum_{i,j} (a_{ij} h_i)^2$, so ist $\phi^* = \sum_{i,j} (a_{ij} g_i)^2$ mit $a_{ij} g_i \in U$.) □

6.5 Theorem. Sei $S \subseteq \mathbb{R}^n$ eine semialgebraische Menge und sei $K := \overline{\mathrm{conv}(S)}$. Dann sind äquivalent:

(i) K ist ein Spektraederschatten.

(ii) S hat uniforme sos-Darstellungen für $L := \mathbb{R}[\mathbf{x}]_{\leq 1}$.

Die analoge Äquivalenz gilt für $C := \overline{\mathrm{cone}(S)}$ und $L_1 := \mathbb{R}[\mathbf{x}]_1$ (Linearformen).

6.6 Satz. In Theorem 6.5 gilt (ii) $\Rightarrow$ (i).

Beweis. Sei $\phi : X \to \mathbb{A}^n$ mit $S \subseteq \phi(X(\mathbb{R}))$ und sei $U \subseteq \mathbb{R}[X]$ mit $\dim(U) < \infty$ und $\phi^*(P_S) \subseteq \Sigma U^2$. Setze $W := UU + \phi^*(L_1) \subseteq \mathbb{R}[X]$ und weiter $\varphi := \phi^*|_{L_1} : L_1 \to W$. Wegen $\varphi(P_S) \subseteq \Sigma U^2$ ist $P_S \subseteq \varphi^{-1}(\Sigma U^2)$. Umgekehrt gilt '$\supseteq$' wegen $S \subseteq \phi(X(\mathbb{R}))$. Da ΣU^2 ein Spektraederschatten ist (Konvexität), ist auch $P_S = \varphi^{-1}(\Sigma U^2)$ einer, d.h. nach 6.2 ist K ein Spektraederschatten. □

Für den Beweis von (i) $\Rightarrow$ (ii) in Theorem 6.5 starten wir mit der homogenen Version:

6.7 Satz. Sei $S \subseteq \mathbb{R}^n$ semialgebraisch und sei $C := \overline{\mathrm{cone}(S)}$ ein Spektraederschatten. Dann hat S (und C) uniforme sos-Darstellungen für $L := \mathrm{span}(x_1, \dots, x_n)$.

Beweis. Ohne Einschränkung sei $\operatorname{aff}(S) = \mathbb{R}^n$. Nach Voraussetzung gibt es eine lineare Abbildung $\pi\colon \mathbb{R}^p \to \mathbb{R}^n$ und einen Spektraederkegel $T \subseteq \mathbb{R}^p$ mit $C = \pi(T)$. Man kann $\operatorname{int}(T) \neq \varnothing$ annehmen und dass T eine Beschreibung durch eine **strikt** lösbare LMI hat, etwa

$$T = \{(\xi, \eta) \in \mathbb{R}^n \times \mathbb{R}^m \colon M(\xi) + N(\eta) \succeq 0\}$$

mit $M(x) = \sum_i x_i M_i$, $N(y) = \sum_j y_j N_j$ und $M(\xi) + N(\eta) \succ 0$ für ein $(\xi, \eta) \in T$. Betrachte die folgende Untervarietät X von $\mathbb{A}^n \times \mathbb{A}^m \times \operatorname{Sym}_d$: Die $\mathbb{C}$-Punkte von X seien die Tupel $(\xi, \eta, A) \in \mathbb{C}^n \times \mathbb{C}^m \times \operatorname{Sym}_d(\mathbb{C})$ mit $A^2 = M(\xi) + N(\eta)$ und die Koordinatenfunktionen auf X seien $(x_1, \dots, x_n, y_1, \dots, y_m; z_{\mu\nu} = z_{\nu\mu}\ (\mu, \nu \in [d]))$. Definiere $\phi\colon X \to \mathbb{A}^n$, $(\xi, \eta, A) \mapsto \xi$, dann gilt $\phi(X(\mathbb{R})) = \pi(T) = C$. Setze $U := \operatorname{span}(z_{\mu\nu} \mid \mu, \nu \in [d]) \subseteq \mathbb{R}[X]$ und behaupte, (3) aus Definition 6.3 ist erfüllt. Sei $f = \sum_{i=1}^n a_i x_i$ mit $f|_S \geq 0$ (d.h. $f|_C \geq 0$), dann liegt der Punkt $(a, 0) = (a_1, \dots, a_n, 0, \dots, 0) \in \mathbb{R}^n \times \mathbb{R}^m$ in T^*: Aus $(\xi, \eta) \in T$ folgt $\xi \in C$, d.h. $0 \leq f(\xi) = \langle \xi, a \rangle = \langle (\xi, \eta), (a, 0) \rangle$. Nach Konvexität existiert ein $B \in \mathsf{S}^d_+$ mit $a_i = \langle B, M_i \rangle$ und $0 = \langle B, N_j \rangle$ für alle $j \in [m]$. Sei $W = (w_{k\ell}) \in \mathsf{S}^d$ mit $W^2 = B$. Dann ist $\phi^*(f)$ gleich

$$\sum_{i=1}^n \langle B, M_i \rangle x_i + \sum_{j=1}^m \langle B, N_j \rangle y_j = \langle B, M(x) + N(y) \rangle = \langle W^2, Z^2 \rangle = \langle ZW, ZW \rangle$$

(wegen $\langle W^2, Z^2 \rangle = \operatorname{tr}(W^2 Z^2) = \operatorname{tr}((ZW)(ZW)^\top) = \langle ZW, ZW \rangle$). Als Element in $\mathbb{R}[X]$ ist das

$$\phi^*(f) = \sum_{\mu,\nu=1}^d ((ZW)_{\mu,\nu})^2 = \sum_{\mu,\nu=1}^d \left(\sum_k z_{\mu k} w_{k\nu} \right)^2. \qquad \square$$

Jetzt können wir auch die inhomogene Version beweisen:

6.8 Satz. Sei $S \subseteq \mathbb{R}^n$ semialgebraisch und sei $K := \operatorname{conv}(S)$ ein Spektraederschatten. Dann hat S (und K) uniforme sos-Darstellungen für $L := \operatorname{span}(1, x_1, \dots, x_n)$.

Beweis. Das lässt sich leicht aus Satz 6.7 herleiten. $\square$

6.9 Satz. Sei $S \subseteq \mathbb{R}^n$ eine semialgebraische Menge und sei $L \subseteq \mathbb{R}[\mathbf{x}]$ ein Untervektorraum mit $m := \dim(L) < \infty$. Sei $p_1, \dots, p_m$ eine Basis von L und setze

$$\varphi_L \colon \mathbb{R}^n \to \mathbb{R}^m,\ u \mapsto (p_1(u), \dots, p_m(u)).$$

Dann ist $\overline{\operatorname{conv}(\varphi_L(S))}$ genau dann ein Spektraederschatten, wenn S uniforme sos-Darstellungen für $L_1 := \operatorname{span}(1, p_1, \dots, p_m)$ hat.

Beweis. Übung. $\square$

Definition. Für eine offene und semialgebraische Menge $U \subseteq \mathbb{R}^n$ sei $\mathcal{N}(U)$ der Ring der *Nashfunktionen*[13] auf U. Für $u \in U$ sei $\mathcal{O}_u := \mathbb{R}[\mathbf{x}]_{\mathfrak{m}_u}$, d.h. $\widehat{\mathcal{O}}_u = \mathbb{R}[\![x_1 - u_1, \dots, x_n - u_n]\!] = \mathbb{R}[\![\mathbf{x} - u]\!]$. Taylorentwicklung von $f \in \mathcal{N}$ in u liefert den Ringhomomorphismus

$$\mathcal{N} \to \widehat{\mathcal{O}}_u, \quad f \mapsto \tau_u(f).$$

Betrachte weiter die Veroneseabbildung[14] $\varphi_{n,d} \colon \mathbb{R}^n \to \mathbb{R}^N$, $u \mapsto (u^\alpha)_{1 \leq |\alpha| \leq d}$ mit $N := \binom{n+d}{d} - 1$.

6.10 Theorem. Sei $S \subseteq \mathbb{R}^n$ eine semialgebraische Menge mit $\operatorname{int}(S) \neq \varnothing$. Ist $n \geq 3$ und $d \geq 6$ oder $n \geq 4$ und $d \geq 4$, so ist $\overline{\operatorname{conv}(\varphi_{n,d}(S))}$ kein Spektraederschatten.

[13] John Forbes Nash (1928–2015)
[14] Giuseppe Veronese (1854–1917)

Beweis. Angenommen doch, dann existiert ein Unterraum $U \subseteq \mathcal{A}_0(S)$ mit $\dim(U) < \infty$ so, dass für jedes $f \in \mathbb{R}[\mathbf{x}]_{\leq d}$ mit $f|_S \geq 0$ die Funktion $f|_S$ in ΣU^2 liegt. Jedes $h \in U$ ist Nash auf einer offen-dichten Teilmenge von $\operatorname{int}(S)$ (RAG, 4.5.7). Daher existiert eine offene Menge $\varnothing \neq W \subseteq S$, sodass $h|_W$ für alle $h \in U$ Nash ist. Fixiere ein $u \in U$ und sei $p \in \mathbb{R}[\mathbf{x}]$ eine psd Form mit $\deg(p) \leq d$, welche nicht sos ist. Dann ist $f := p(y_1, \ldots, y_n)$ mit $y_i := x_i - u_i$ eine Summe von Quadraten in $\widehat{\mathcal{O}}_u = \mathbb{R}[\![\mathbf{x} - u]\!] = \mathbb{R}[\![\mathbf{y}]\!]$, d.h. f ist sos von Formen in $(y_1, \ldots, y_n)$ (ist $f(y) = h_1^2 + \cdots + h_N^2$ mit $\deg(f) = 2d$ und $h_i = h_{id}(y) + h_{i,d+1}(y) + \cdots$, so ist $f = \sum_i h_{id}(y)^2$), Widerspruch. $\square$

Bemerkung. Für $(n, d) = (3, 6)$ bzw. $(n, d) = (4, 4)$ liefert das Theorem also konvexe semialgebraische Mengen, die keine Spektraederschatten sind (nämlich abgeschlossene konvexe Hüllen von semialgebraischen Mengen von Dimension ≥ 3) in $\mathbb{R}^{\binom{9}{3}-1} = \mathbb{R}^{83}$ bzw. $\mathbb{R}^{\binom{8}{4}-1} = \mathbb{R}^{69}$.

Fixiere einen reell abgeschlossenen Körper $R \supsetneq \mathbb{R}$ und sei $B \subseteq R$ die konvexe Hülle von $\mathbb{R}$ in R.

6.11 Lemma. Sei A eine reell reduzierte $\mathbb{R}$-Algebra. Sind $f_1, \ldots, f_m \in A \otimes R = A \otimes_{\mathbb{R}} R =: A_R$ mit $\sum_{i=1}^r f_i^2 \in A_B := A \otimes B$, so gilt $f_i \in A_B$ für alle $i \in [r]$.

Beweis. Wähle linear unabhängige $g_1, \ldots, g_S \in A$ und $c_{ij} \in R$ mit $f_i = \sum_j c_{ij} g_j$. Angenommen $c_{ij} \notin B$ für ein Paar (i, j), dann ist $c := \max_{ij} |c_{ij}| \notin B$, d.h. $h_i := \frac{1}{c} f_i \in A_B$ für alle $i \in [r]$ und $\frac{1}{c^2} f = \sum_i h_i^2$ hat Koeffizienten in $\mathfrak{m}_B$. Reduktion modulo $\mathfrak{m}_B$ liefert $0 = \sum_i \overline{h_i}^2$ in $A_B/\mathfrak{m}_B A_B = A$ und $\overline{h_i} \neq 0$ für mindestens ein $i \in [r]$, im Widerspruch dazu, dass A reell reduziert ist. $\square$

6.12 Satz. Sei $f \in \mathbb{R}[t, \mathbf{x}]$ homogen vom Grad d in $(t, \mathbf{x})$ und sei f nicht sos in $\mathbb{R}[t, \mathbf{x}]$. Für alle $0 < \varepsilon \in \mathfrak{m}_B$ ist $f(\varepsilon, x) \in B[\mathbf{x}]$ nicht sos modulo $\langle \mathbf{x} \rangle^{d+1} := \langle x_1, \ldots, x_n \rangle^{d+1}$.

Beweis. Angenommen $f(\varepsilon, x) + g(x) = \sum_j p_j(x)^2$ mit $p_j(x) \in B[\mathbf{x}]$ und $g(x) \in \langle \mathbf{x} \rangle^{d+1} B[\mathbf{x}]$. Sei

$$g_1(x) := \frac{1}{\varepsilon^{d+1}} g(\varepsilon x) \in B[\mathbf{x}].$$

Es folgt $\varepsilon^d f(1, x) + g(\varepsilon x) = f(\varepsilon, \varepsilon x) + g(\varepsilon x) = \sum_j p_j(\varepsilon x)^2$, Division durch ε^d liefert also

$$f(1, x) + \underbrace{\varepsilon g_1(x)}_{\in\, \varepsilon B[\mathbf{x}]} = f(1, x) + \frac{g(\varepsilon x)}{\varepsilon^d} = \sum_j \varepsilon^{-d} p_j(\varepsilon x)^2 = \sum_j \left(\left(\sqrt{\varepsilon}\right)^{-d} p_j(\varepsilon x) \right)^2. \qquad (*)$$

Nach Lemma 6.11 liegen alle Summanden der rechten Seite von $(*)$ in $B[\mathbf{x}]$. Reduziert man beide Seiten modulo $\mathfrak{m}_B$, so folgt, dass $f(1, x)$ sos in $\mathbb{R}[\mathbf{x}]$ ist, im Widerspruch dazu, dass $f(t, x)$ nicht sos ist. $\square$

6.13 Theorem. Sei $p(t, x) \in \mathbb{R}[t, \mathbf{x}]$ eine psd Form, die nicht sos ist und sei $L \subseteq \mathbb{R}[\mathbf{x}]$ ein Unterraum mit $m := \dim(L) < \infty$ und $p(c, x - u) \in L$ für alle $c \in \mathbb{R}$ und $u \in \mathbb{R}^n$. Für jede semialgebraische Menge $S \subseteq \mathbb{R}^n$ mit $\operatorname{int}(S) \neq \varnothing$ ist $\overline{\operatorname{conv}(\varphi_L(S))} \subseteq \mathbb{R}^m$ kein Spektraederschatten.

Beweis. Angenommen doch, dann gibt es wie im Beweis von Theorem 6.10 eine offene semialgebraische Menge $\varnothing \neq W \subseteq S$ und einen Unterraum $U \subseteq \mathcal{N}(W)$ mit $\dim(U) < \infty$ und $f|_W \in \Sigma U^2$ für jedes $f \in L + \mathbb{R} \cdot 1$ mit $f|_S \geq 0$. Seien $R \supsetneq \mathbb{R}$ und $B \subseteq R$ wie zuvor und sei $0 < \varepsilon \in \mathfrak{m}_B$. Für jedes $f \in L_R \subseteq R[\mathbf{x}]$ mit $f|_{S_R} \geq 0$ ist $f|_{W_R}$ sos von Elementen aus U_R. Sei $d := \deg(p)$ und sei $u \in W$. Das Polynom $f := p(\varepsilon, x_1 - u_1, \ldots, x_n - u_n) \in B[\mathbf{x}]$ liegt in $L_R + R \cdot 1$ und ist ≥ 0 auf R^n. Damit ist $f|_W \in \Sigma(U_R)^2$, d.h. f ist sos in $\widehat{\mathcal{O}}_u \otimes R$. Nach Lemma 6.11 ist f daher sos in $\widehat{\mathcal{O}}_u \otimes B$, also auch modulo $\langle \mathbf{x} - u \rangle^{d+1}$. Damit ist f sos in $B[\mathbf{x}]/\langle \mathbf{x} - u \rangle^{d+1}$, im Widerspruch zu Satz 6.12. $\square$

6.14 Bemerkung. Sei $L := \operatorname{span}(\mathbf{x}^\alpha \mid 1 \leq |\alpha| \leq d) \subseteq \mathbb{R}[\mathbf{x}]$ mit $n = 2$ und $d \geq 6$ oder $n \geq 3$ und $d \geq 4$. Dann erhalten wir erneut Beispiele von konvexen semialgebraischen Mengen, die keine Spektraederschatten sind, jetzt jedoch in Dimension $\binom{8}{2} - 1 = 27$ oder $\binom{7}{3} - 1 = 34$.

6.15 Theorem. Der Kegel $P_{n,2d} \subseteq \mathbb{R}[\mathbf{x}]$ ist genau dann ein Spektraederschatten, wenn $P_{n,2d} = \Sigma_{n,2d}$ gilt und das ist genau dann der Fall, wenn $n \leq 2$ oder $2d = 2$ oder $(n, 2d) = (3, 4)$.

Beweis. Das folgt direkt aus Bemerkung 6.14, zusammen mit Hilbert 1888. □

Bemerkung. Es sind noch kleinere Dimensionen von Gegenbeispielen möglich: In Dimension 12 betrachte $\varphi_L \colon \mathbb{R}^2 \to \mathbb{R}^{12}$ mit

$$L = \operatorname{span}\left(x, x^2, x^3, x^4, x^5, x^6, y, y^2, xy, xy^2, 3x^2y - 2y^3, 3x^2y^2 - y^4\right)$$

und nimm $f(t, x, y) = x^6 + t^2x^4 + t^4y^2 - 3t^2x^2y^2$. Auch in Dimension 11 gibt es noch Beispiele, siehe z.B. [2], Exercise 8.7.4.

Eine offene Frage ist, ob es in $\mathbb{R}^d$ für $3 \leq d \leq 10$ auch konvexe semialgebraische Mengen gibt, die keine Spektraederschatten sind.

Literatur

[1] Tim NETZER, Daniel PLAUMANN: *Geometry of Linear Matrix Inequalities.* Birkhäuser, 2023.

[2] Claus SCHEIDERER: *A Course in Real Algebraic Geometry.* Springer, 2024.

[3] Winfried SCHARLAU: *Quadratic and Hermitian Forms.* Springer, 1985.