Exponentielle Approximation: Theorie und Numerische Verfahren

Franz J. Polster

15.02.2023

Prolog

Dieses Dokument ist eine "Überarbeitung" meiner Diplomarbeit vom Januar 1973 [1]*, angefertigt am Institut für Angewandte Mathematik I der Universität Erlangen-Nürnberg. Es wurde erstellt mit LaTeX in den Pandemiejahren 2020-2022.*

Die Aufgabe bestand in der Implementierung und Erprobung von Verfahren zur Exponentiellen Approximation:

- *Es war die Theorie der Exponentialapproximation darzustellen: "Teil I"*
- *Auf dieser Grundlage waren Verfahren zur Exponentiellen Approximation zu beschreiben: "Teil II"*
- *Diese Verfahren waren mittels numerischer Experimente zu erproben: "Teil III".*

Änderungen, Abweichungen gegenüber dem Original:

Teil I und Teil II dieses Dokuments sind inhaltlich identisch mit den entsprechenden Teilen der Diplomarbeit von 1973.

- *Layout und Formatierung ergeben sich aus den Fähigkeiten und Möglichkeiten von LaTeX . Die Fehlerfunktionen wurden mit LaTeX-package "pgfplots" erstellt.*[2]
- *Vereinzelt wurden textuelle Änderungen zur Verbesserung der Lesbarkeit vorgenommen.*
- *Die Verwendung von LaTeX-Zählern hat Nummerierungsfehler im Original (bei Sätzen, Korollaren, ...) aufgezeigt, diese wurden korrigiert.*
- *Einige (zum Teil offensichtliche) Unstimmigkeiten im Original wurden korrigiert, auf diese weisen entsprechende Fußnoten hin.*
- *Beispiele 5.4, 5.5 von Kapitel 5 sind Erweiterungen des Originals*

Für Teil III, Kapitel 7 und 8 gilt zusätzlich:
Für die Berechnungen wird auf das Original verwiesen, Anhang B enthält den Textteil von §7 des Originals.

Anhang A ist eine Erweiterung des Originals, ebenso wie die Quellenangaben des Literaturverzeichnisses ab [26].

---

[1] *nachstehend auch "Original" genannt*

[2] *das Original wurde auf einer Schreibmaschine geschrieben, die Zeichnungen des Originals wurden mit dem Plotter des Rechenzentrums der Universität gezeichnet*

Inhaltsverzeichnis

Bezeichnungen und Abkürzungen

$\mathbb{N}$	Menge der natürlichen Zahlen
$\mathbb{R}$	Menge der reellen Zahlen
$\mathbb{C}$	Menge der komplexen Zahlen
C(X)	Menge der auf X stetigen Funktionen
xEy	$x \times 10^y$
$xE \pm y$	$x \times 10^{\pm y}$
$\setminus$	Mengendifferenz
$\Re(z)$	Realteil von $z \in \mathbb{C}$
$\Im(z)$	Imaginärteil von $z \in \mathbb{C}$
$\|x\|$	Betrag von x
$\| f \|_X$	Tschebyscheff-Norm (p. 5)
$\frac{d}{dx}$, D^n	Differentialoperator (p. 8)
$grd(P)$	Grad des Polynom P (p. 9)
$grd(E)$	Grad von $E \in V_N$ (p. 9)
$\equiv$	Gleichheit von Funktionen
$sign(x)$	Vorzeichen von x (p. 20)
$sign(E)$	Vorzeichenvektor der Exponentialsumme E (p. 20)
o.B.d.A.	ohne Beschränkung der Allgemeinheit

Teil I

Theorie der Exponentialapproximation

1 Einleitung, Hilfsmittel

1.1 Die Mengen V_N, V_N^0, V_N^+

Es sei X stets eine kompakte Teilmenge der reellen Zahlen $\mathbb{R}$ und C(X) die Menge der auf X stetigen, reellen Funktionen; C(X) sei mit der von der Tschebyscheff-Norm $\| \, . \, \|$ induzierten Metrik versehen:

$$\| \, f \, \|_X = \sup_{x \in X} \mid f(x) \mid \;\; , \;\; f \in C(I)$$

In $\mathbb{R}^n$, $n \in \mathbb{N}$, wird die Euklidische Metrik bzw. Norm verwendet.
I bezeichnet stets ein abgeschlossenes Intervall [a,b] mit $a < b$, $a, b \in \mathbb{R}$.

In dieser Arbeit wird die Tschebyscheff-Approximation stetiger, reeller Funktionen durch Exponentialsummen - ein nichtlineares Approximationsproblem - behandelt; wie üblich sei also definiert:

Definition 1.1
Es sei $V \subseteq C(X)$.

a. $v_0 \in V$ heißt *beste Approximation* an $f \in C(X)$ bezüglich V auf X oder auch *Minimallösung* für f bezüglich V auf X, wenn gilt:

$$\| \, f - v_0 \, \|_X = \inf_{v \in V} \| \, f - v \, \|_X = R(f)$$

b. R(f) ist die *Minimalabweichung* von f bezüglich V auf X.

c. Die Folge $(v_n)_{n \in \mathbb{N}} \subseteq V$ ist eine *Minimalfolge* in V für f auf X, wenn gilt:

$$\lim_{n \to \infty} \| \, f - v_n \, \|_X = \; R(f)$$

Wenn man von Exponentialsummen spricht, denkt man zunächst an Summen

$$\sum_{i=1}^{n} a_i e^{t_i x} \text{ mit } a_i, t_i \in \mathbb{R} \text{ für } 1 \leqq i \leqq n \ , \ n \in \mathbb{N}$$

Dies soll im Folgenden präzisiert werden.

Definition 1.2
Es sei $N \in \mathbb{N}$ und $a = (a_1, \ldots, a_N, t_1, \ldots, t_N) \in \mathbb{R}^{2N}$.

a.

$$E(a,x) \quad := \quad \sum_{i=1}^{N} a_i e^{t_i x} \qquad x \in \mathbb{R}$$

Die hiermit auf $\mathbb{R}$ definiert Funktion wird mit $E(a)$ bezeichnet; $a \in \mathbb{R}^{2N}$ ist der *Parametervektor* von $E(a)$[3].

b.

$$V_N^0 := \{E(a) \mid \ a = (a_1, \ldots, a_N, t_1, \ldots, t_N) \in \mathbb{R}^{2N}, \\ t_i < t_{i+1} \text{ für } 1 \leqq i \leqq N-1, \ a_i = 0 \Rightarrow a_{i+1} = 0\}$$

V_0^0 enthält nur die Nullfunktion.

c. L ist die *Länge* von $E(a,x) = \sum_{i=1}^{L} a_i e^{t_i x} \in V_N^0$, wenn $a_i \neq 0$ für $1 \leqq i \leqq L$ erfüllt ist.

Bemerkung:
Jede Funktion $\sum_{i=1}^{N} a_i e^{t_i x}$ mit $a_i, t_i \in \mathbb{R}$ für $1 \leqq i \leqq N$ ist in V_N^0 enthalten; $E(a) \in V_N^0$ bedeutet nur, daß die Indizierung der Summanden gemäß Definition 1.2b durchgeführt wird. Dies ist offensichtlich stets möglich und man sieht unmittelbar ein, daß $E(a) \in V_N^0$ mit der Länge $L < N$ keinen eindeutig bestimmten Parametervektor $a \in \mathbb{R}^{2N}$ besitzt, da zum Beispiel $t_{L+1} \in \mathbb{R}$ mit $t_{L+1} > t_L$ beliebig gewählt werden kann.

Das folgende Bespiel zeigt, daß die Existenz einer besten Approximation bezüglich V_N^0 allgemein nicht gesichert ist:

[3] der Parametervektor a umfasst also auch die t_i: $a = (a_1, \ldots, a_N, t_1, \ldots, t_N)$

Beispiel 1.1 (Meinardus, [12])
Es sei $X = [0,1]$, $f(x) = xe^x$ und für $m \in \mathbb{N}$ sei $E(a_m) \in V_2^0$ gegeben durch

$$E(a_m, x) = me^{(1+\frac{1}{m})x} - me^x$$

Für $x \in X$ gilt:

$$\begin{aligned} \mid f(x) - E(a_m, x) \mid &= e^x \mid x - me^{\frac{x}{m}} + m \mid = e^x \mid m \sum_{i=2}^{\infty} \frac{x^i}{i! m^i} \mid \\ &\leqq e^x m^{-1} e^x \leqq m^{-1} e^2 \end{aligned}$$

Es folgt also $\lim\limits_{n\to\infty} \parallel f - E(a_m) \parallel_X = 0$.
Die Minimalabweichung von f bezüglich V_2^0 ist demnach Null, f ist jedoch nicht in V_2^0 enthalten. Damit gibt es keine beste Approximation für f bezüglich V_2^0 auf X.

Anders formuliert besagt Beispiel 1.1, daß jede Umgebung von $f(x) = xe^x$ in $C([0,1])$ eine Funktion aus V_2^0 enthält.
Dies ergibt sich auch unmittelbar aus der Tatsache, daß die betrachteten Funktionen Lösungen homogener, linearer Differentialgleichungen zweiter Ordnung sind.

Allgemein für $n \in \mathbb{N}$ entnimmt man der Theorie der linearen Differentialgleichungen N-ter Ordnung mit konstanten Koeffizienten (man vergleiche hierzu [5], [7]):

1. Die Lösungen h der Differentialgleichungen

$$\sum_{i=0}^{N} d_i D^i h = \prod_{i=1}^{L} (D - t_i)^{m_i+1} h = 0 \tag{1.1}$$

$$\text{mit } t_i \in \mathbb{R} \text{ für } \quad 1 \leqq i \leqq L \text{ und } \sum_{i=1}^{L} (m_i + 1) = N$$

sind von der Gestalt

$$h(x) = \sum_{i=1}^{L} P_i(x) e^{t_i x}$$

mit reellen Polynomen $P_i(x) = \sum\limits_{n=0}^{m_i} a_{i,n} x^n$ für $1 \leqq i \leqq L$.

Liegt ein Anfangswertproblem vor, so sind die Lösungen von (1.1) eindeutig bestimmt.

(Wie üblich wird der Differentialoperator $\frac{d}{dx}$ mit D bezeichnet. Es ist $D^0 f = f$ und $D^n = D^{n-1}D$.)

2. Die Lösungen der Differentialgleichungen (1.1) mit den Anfangswerten $D^i h(x_0) = h_i \in \mathbb{R}$, $0 \leqq i \leqq N-1$, hängen stetig von den Koeffizienten d_i und den Anfangswerten ab. Dies ergibt sich aus dem entsprechenden Satz für Differentialgleichungssysteme.

Hat man also eine Funktion $h(x) = \sum_{i=1}^{L}(\sum_{n=0}^{m_i} a_{i,n}x^n)e^{t_i x}$ mit $m_i \geqq 1$ für ein i und $\sum_{i=1}^{L}(m_i+1) = N$ gegeben, so ist

$$\sum_{i=0}^{N} d_i D^i h := \prod_{i=1}^{L}(D-t_i)^{m_i+1} h = 0 \tag{1.2}$$

erfüllt.

Die Differentialgleichung

$$\sum_{i=0}^{N} c_i D^i g := \prod_{i=1}^{N}(D-s_i)g$$

mit $\mid t_i - s_i \mid < \epsilon$ für $1 \leqq i \leqq N$, $s_i \neq s_j$ für $i \neq j$, besitzt Lösungen der Form

$$g(x) = \sum_{i=0}^{N} a_i e^{s_i x}$$

Legt man ein abgeschlossenes Intervall $I \subset \mathbb{R}$ und die Anfangswerte

$$D^i g(x_0) = D^i h(x_0) \qquad 0 \leqq i \leqq N-1, \quad x_0 \in I$$

zugrunde, dann kann $\parallel h - g \parallel_I$ bei geeigneter Wahl von ϵ beliebig klein gemacht werden, da $\lim_{\epsilon \to 0} c_i = d_i$, $0 \leqq i \leqq N-1$, gilt.

Damit ist gezeigt:

(1.3) In jeder Umgebung von $\sum_{i=1}^{L} P_i(x)e^{t_i x}$ in C(I) mit $\sum_{i=1}^{L}(grd(P_i)+1) = N$ liegt ein Element von V_N^0

Hierbei wird der Grad des Polynoms $P(x) = \sum_{i=0}^{m} a_i x^i$ mit *grd(P)* bezeichnet; mit $a_m \neq 0$ gilt also $grd(P) = m$.
Das Nullpolynom $P(x) \equiv 0$ habe den Grad -1.

Will man bei der Approximation stetiger Funktionen durch Exponentialsummen die Existenz einer besten Approximation sichern, so muß man also V_N^0 mindestens um die oben in (1.3) betrachteten, verallgemeinerten Exponentialsummen erweitern.

Definition 1.3
a.

$$V_N := \{E(a) \mid \quad E(a,x) = \sum_{i=1}^{L} P_i(x) e^{t_i x}, P_i \text{ ist ein reelles Polynom}$$
$$\text{mit } grd(P_i) = m_i, \ \sum_{i=1}^{L}(m_i + 1) \leqq N,$$
$$t_i \in \mathbb{R} \text{ für } 1 \leqq i \leqq L \text{ mit } t_i < t_{i+1}\}, \ N \in \mathbb{N} \cup \{0\}$$

b.
Für $E(a,x) = \sum_{i=1}^{L} P_i(x) e^{t_i x} \in V_N$ werden folgende Bezeichnungen eingeführt:
Es sei $grd(E(a)) \quad := \quad \sum_{i=1}^{L}(grd(P_i) + 1)$ der *Grad* von E(a);
ist $grd(P_i) \geqq 0$ für $1 \leqq i \leqq L$ erfüllt, dann ist L die *Länge* von E(a).
Das *Spektrum* von E(a) ist gegeben durch $\{t_i \mid 1 \leqq i \leqq L\}$, die Elemente dieser Menge seien die *Frequenzen* von E(a).

Bemerkung 1.1

1. Es ist nicht ohne weiteres möglich, den Elementen von V_N einen Parametervektor zuzuordnen, wie dies in Definition 1.1 für V_N^0 geschah. Dies wird deutlich wenn man bedenkt, daß neben $(\sum_{i=0}^{N-1} a_{i+1} x^i) e^{t_1 x}$ mit $a_N \neq 0$ auch $\sum_{i=1}^{N} a_i e^{t_i x} \in V_N$ enthalten ist; im ersten Fall bestimmen $N+1$, im zweiten Fall $2N$ Parameter die Funktion.
   Auch die Beschränkung auf Elemente von V_N mit einer festen Anzahl von Parametern bringt keine Änderung:

Es seien $a, b, c, u, v \in \mathbb{R}$ mit $u < v$ gegeben. Mit (a, b, c, u, v) als Parametervektor lassen sich die Exponentialsummen

$$ae^{ux} + (b + cx)e^{vx} \in V_3 \qquad (a + bx)e^{ux} + ce^{vx} \in V_3$$

bilden, die im Spektrum übereinstimmen, im allgemeinen jedoch nicht gleich sind.
Für $E(a) \in V_N \setminus V_N^0$ hat also der Parameter a nicht die gleiche Bedeutung wie für die Elemente von V_N^0; es dient hier nur der Bezeichnung einer speziellen Funktion im Sinne einer Indizierung.

2. Für $E(a) \in V_N^0$ fällt die Länge mit dem Grad zusammen. Hat $E(a) \in V_N \setminus V_N^0$ die Länge L, so folgt $L < grd(E(a))$ für $N \geqq 2$.

3. Aus der Definition der Polynomgrade ergibt sich, daß V_0 nur die Nullfunktion enthält.

In Kapitel 2 wird gezeigt, daß auch bei der Approximation bezüglich der in Definition 1.4 beschriebenen Teilmengen von V_N^0 eine Minimallösung für $f \in C(I)$ existiert:

Definition 1.4

$$\begin{aligned} V_N^+ &:= \{\ E(a) \quad | \ E(a,x) = \sum_{i=1}^{N} a_i e^{t_i x}, a_i \geqq 0 \text{ für } 1 \leqq i \leqq N\ \} \\ V_N^- &:= \{\ E(a) \quad | \ -E(a) \in V_N^+\ \} \end{aligned}$$

Bemerkung 1.2 (Werner, [21])
Für $n \in \mathbb{N}$ ist $E \in V_N$ n-fach differenzierbar in $\mathbb{R}$ und es gilt:

$$D^n E \in V_N$$

Beweis: Vollständige Induktion nach n:
$E(x) = \sum_{i=1}^{L} P_i(x)e^{t_i x} \in V_N$ ist in $\mathbb{R}$ differenzierbar:

$$DE(x) = \sum_{i=1}^{L} (DP_i(x) + P_i(x)t_i)e^{t_i x}$$

Wegen $grd(DP_i + P_i t_i) \leqq grd(P_i)$ ist $DE \in V_N$ erfüllt.
Gilt die Behauptung für n, also $z = D^n h \in V_N$, dann ist z in $\mathbb{R}$ differenzierbar und es ist $Dz = D^{n+1}h \in V_N$. Damit ist die Behauptung vollständig gezeigt.

Von entscheidender Bedeutung ist der nun folgende Satz über die Nullstellen von Exponentialsummen:

Satz 1.1 (Meinardus, [12])
Jede nicht identisch verschwindende Funktion in V_N besitzt höchstens $N-1$ reelle Nullstellen.

Beweis: Durch vollständig Induktion nach L wird gezeigt:

$$\begin{aligned} &\sum_{i=1}^{L} P_i(x)e^{t_i x} \in V_N \text{ mit} E \not\equiv 0 \text{ hat höchstens} \\ &(\sum_{i=1}^{L}(grd(P_i)+1)) - 1 \text{ reelle Nullstellen.} \end{aligned} \tag{1.4}$$

Für $L = 1$ ist nichts zu zeigen: $P(x)e^{tx} \in V_N$ besitzt höchstens $grd(P)$ reelle Nullstellen oder verschwindet identisch.
Es gelte (1.4) für $L-1$.
$E(x) = \sum_{i=1}^{L} P_i(x)e^{t_i x} \in V_N$ habe M reelle Nullstellen und es sei $m_L = grd(P_L) \geqq 0$ (sonst gilt die Behauptung nach Induktionsannahme).

$$\begin{aligned} E_1(x) &:= D^{m_L+1}(e^{-t_L x}E(x)) = D^{m_L+1}(\sum_{i=1}^{L-1} P_i(x)e^{(t_i-t_L)x}) = \\ &= \sum_{i=1}^{L-1} Q_i(x)e^{(t_i-t_L)x} \end{aligned}$$

Nach Bemerkung 1.2 gilt

$$\sum_{i=1}^{L-1}(grd(Q_i)+1) \leqq \sum_{i=1}^{L-1}(grd(P_i)+1)$$

Die Funktion $e^{-t_L}E(x)$ besitzt M reelle Nullstellen; nach dem Satz von Rolle hat dann E_1 mindestens $M-(m_L+1)$ reelle Nullstellen. Die Induktionsannahme, auf E_1 angewandt, ergibt

$$M - m_L - 1 \leqq (\sum_{i=1}^{L-1}(grd(P_i)+1)) - 1 = (\sum_{i=1}^{L}(grd(P_i)+1)) - 1 - m_L - 1$$

woraus $M \leqq (\sum_{i=1}^{L}(grd(P_i)+1)) - 1$ folgt.
Damit ist (1.4) vollständig bewiesen und die Behauptung des Satzes gilt.

Die Theorie der nichtlinearen Approximation, wie sie in [11] dargestellt ist, wird in Kapitel 3 zur Herleitung von Charakterisierungs- und Eindeutigkeitssätzen für die Approximation bezüglich V_N herangezogen; die Theorie von Rice, [15], ist hier nicht anwendbar, da V_N im allgemeinen die Varisolvency-Eigenschaft nicht besitzt. Braess gibt dazu in [1] folgendes Beispiel an:

Beispiel 1.2
Betrachtet wird V_2 im Intervall $[-3,3]$. In V_2 besitzt $E_0(x) = x \in V_2$ die Varisolvency-Eigenschaft nicht:

a. Es sei
$x_1 = -2, \quad x_2 = -1, x_3 = 1, x_4 = 2 \quad$ und
$y_1 = -2+\delta, y_2 = -1, y_3 = 1, y_4 = 2-\delta \quad$ mit $\delta > 0$ gegeben.
Es kann o.B.d.A. $\delta < 1$ angenommen werden
Gesucht sind die Funktionen $E \in V_2$, die

$$E(x_i) = y_i \text{ für } 1 \leqq i \leqq 4 \tag{1.5}$$

erfüllen.
Gilt $E_1(x_i) = E_2(x_i) = y_i$, $1 \leqq i \leqq 4$, für $E_1, E_2 \in V_2$, so folgt $E_1 = E_2$ nach Satz 1.1. Für jede Lösung E des Problems (1.5) gilt $E(x_i) = -E(-x_i)$, $1 \leqq i \leqq 4$, und man erhält

$$E(x) = -E(-x) \qquad x \in I \tag{1.6}$$

Die Funktionen in V_2^0 und $V_2 \setminus V_2^0$, die (1.6) erfüllen, sind gegeben durch $E(x) = a(e^{sx} - e^{-sx})$ und $E(x) = ax$ mit $a, s \in \mathbb{R}$. Gilt weiter $E(x) \geqq 0$ für $x \geqq 0$, dann sind diese Funktionen in $[0,3]$ konvex und lösen daher für kein $\delta > 0$ das Problem (1.5).

b. Mit $E(x) = \frac{1}{4}(e^x - e^{-x}) \in V_2$ besitzt $E_0 - E$ in I drei Nullstellen:

x	-3	-1	0	+1	+3
$E_0(x) - E(x)$	+2.009	-0.412	0.0	+0.412	-2.009

Nach a. ist also der Solvency-Grad von E_0 in V_2 kleiner als 4, es gibt aber Elemente $E \in V_2$, so daß $E_0 - E$ in I 4–1 Nullstellen hat; damit ist für E_0 die Varisolvency-Eigenschaft, wie sie in [15] definiert ist, nicht erfüllt.

Bemerkung:
In "On Extended Varisolvent Families" zeigen R. B. Barrar und H. L. Loeb, daß die reellen Funktionen

$$\frac{1}{2\pi i}\int_K \frac{C(z)}{A(z)} e^{zx} dz \ , \qquad x \in \mathbb{R}$$

ein Funktionensystem bilden, das die Varisolvency-Eigenschaft besitzt; hierbei gilt

1. C(z) und A(z) sind Polynome mit reellen Koeffizienten, die keine Nullstellen gemeinsam haben; es sei $grd(c) < grd(A) = n \leqq N \in \mathbb{N}$ und der Koeffizient bei z^n in A ist Eins.
2. Für den Imaginärteil $\Im(z_0)$ jeder Nullstelle z_0 von A gilt:
$$| \Im(z_0) | < r$$
3. Der geschlossene Weg K sei der Rand eines Gebietes in $\mathbb{C}$, das alle Nullstellen von A im Innern enthält.

Jedes Element E von V_N^0 ist in dem beschriebenen Funktionensystem enthalten: Man hat hierzu das Polynom A so zu wählen, daß die n Frequenzen von E die (reellen) Nullstellen von A sind; dann kann das Polynom C so bestimmt werden, daß man (mit dem Cauchyschen Integralsatz) die restlichen Parameter von E erhält.

1.2 Darstellung der Exponentialsummen durch Differenzenquotienten

Die Darstellung von Funktionen aus V_N gemäß Definition 1.3 ist für topologische Untersuchungen oft nicht sehr geeignet, was auch Beispiel 1.1 zeigt. Deshalb wird nun durch Verwendung von Differenzenquotienten eine andere Darstellung für die Elemente von V_N entwickelt, die für Konvergenzbetrachtungen günstiger ist.

Definition 1.5
Die Funktion f sei auf $T := \{\ t_i \mid t_i \in \mathbb{R}, i \in \mathbb{N}, t_i \neq t_j \text{ für } i \neq j\}$ definiert.

$$\begin{aligned} \Delta^0(t_1)f &:= f(t_1) \\ \Delta^n(t_1,\ldots,t_{n+1})f &:= \frac{\Delta^{n-1}(t_1,\ldots,t_n)f - \Delta^{n-1}(t_2,\ldots,t_{n+1})f}{t_1 - t_{n+1}} \qquad n \in \mathbb{N} \end{aligned}$$

Es werden die folgenden **Eigenschaften der Differenzenquotienten** benötigt (man vergleiche hierzu [19], [22], [23]):

1. Mit

$$P_j(t) = \prod_{i=1, i\neq j}^{n+1} (t - t_i)\ ,\ 1 \leqq j \leqq n+1 \qquad \text{gilt}$$

$$\Delta^n(t_1,\ldots,t_{n+1})f = \sum_{j=1}^{n+1} \frac{f(t_j)}{P_j(t_j)} \tag{1.7}$$

Es sei $I \subseteq \mathbb{R}$ ein Intervall und f eine Funktion von t in I.

2. Für $f \in C^n(I)$, d.h. f ist in I n-mal stetig differenzierbar, gilt mit $\{t_i \mid 1 \leqq i \leqq n+1\} \subseteq I$:

$$\Delta^n(t_1,\ldots,t_{n+1})f =$$

$$= \int_0^1 \int_0^{s_1} \ldots \int_0^{s_{n-1}} \frac{d^n}{dt^n} f(t_1 + \sum_{i=1}^{n} (t_{i+1} - t_i)s_i) ds_n \ldots ds_1 \tag{1.8}$$

$$\Delta^n(t_1,\ldots,t_{n+1})f = \frac{1}{n!}\frac{d^n}{dt^n} f(\tau) \qquad \text{für ein } \tau \in I \tag{1.9}$$

$$\Delta^n(t_1,\ldots,t_1)f = \frac{1}{n!}\frac{d^n}{dt^n} f(t_1) \tag{1.10}$$

Die Gleichung (1.8) ermöglicht für $f \in C^n(I)$ folgende Verallgemeinerung:

Definition 1.6
Es sei $n \in \mathbb{N}$, $f \in C^n(I)$ und $\{\ t_i \mid 1 \leqq i \leqq n+1\} \subseteq I$.

$$\Delta^m(t_1, \ldots, t_{m+1})f := \int_0^1 \int_0^{s_1} \ldots \int_0^{s_{m-1}} \frac{d^m}{dt^m} f(t_1 + \sum_{i=1}^{m}(t_{i+1} - t_i)s_i) ds_m \ldots ds_1$$

mit $m \leqq n$.

3. Nach (1.8) sind die Definitionen 1.5 und 1.6 äquivalent, falls $t_i \neq t_j$ für $i \neq j$ gilt.
Für $t_1 \neq t_{n+1}$ folgt auch für die Differenzenquotienten nach Definition 1.6:

$$\Delta^n(t_1, \ldots, t_{n+1})f = \frac{\Delta^{n-1}(t_1, \ldots, t_n)f - \Delta^{n-1}(t_2, \ldots, t_{n+1})f}{t_1 - t_{n+1}}\ , \qquad f \in C^n(I)$$

Bei der Anwendung der Differenzenquotienten auf e^{tx}, $x \in \mathbb{R}$, wird t als Variable betrachtet; die Differenzierbarkeitsbedingungen sind hier erfüllt und die Gleichungen (1.8), (1.9) und (1.10) sind unmittelbar anzuwenden; insbesondere hat man also:

$$\begin{aligned} &\Delta^n(t_1, \ \ldots\ , t_{n+1})e^{tx} = \\ &\quad = x^n \int_0^1 \int_0^{s_1} \ldots \int_0^{s_{n-1}} e^{(t_1 + \sum_{i=1}^{n}(t_{i+1} - t_i)s_i)x} ds_n \ldots ds_1 \end{aligned} \tag{1.11}$$

Vereinbarung:
Da die Differenzenquotienten unabhängig von der Anordnung der Argumente t_i sind, seien diese o.B.d.A. stets der Größe nach geordnet: $t_1 \leqq t_2 \leqq t_3 \leqq \ldots$.

Hilfsatz 1.1 (Braess, [2])
Es sei $N \in \mathbb{N}$ und $t_i \in \mathbb{R}$ für $1 \leqq i \leqq N$ mit $t_i \leqq t_{i+1}$ gegeben.
Behauptung:

$$\Delta^{N-1}(t_1, \ldots, t_N)e^{tx} = \sum_{i=1}^{L} P_i(x)e^{T_i x} \in V_N\ ;$$

dabei gilt: $T_j \in S := \{t_i \mid 1 \leqq i \leqq N\}$ für $1 \leqq j \leqq L$ und für jedes Element $t \in S$ gibt es ein j mit $1 \leqq j \leqq L$, so daß $t = T_j$ erfüllt ist; L ist also die Zahl der verschiedenen Elemente von S.
Ferner gilt:
der Grad von P_i ist kleiner als die Zahl der Elemente von S, die gleich T_i

sind.
Beweis: Vollständige Induktion nach N:

$N = 1$:

$$\Delta^0(t_1)e^{tx} = e^{t_1 x}$$

Die Behauptung gilt hier offensichtlich; sie sei auch für $N-1$ gezeigt.
Induktionsschluß:
1. Fall: $t_1 = t_N$
Aus (1.10 folgt

$$\Delta^{N-1}(t_1, \ldots, t_N)e^{tx} = \frac{1}{(N-1)!}x^{N-1}e^{t_1 x}$$

2. Fall: $t_1 \neq t_N$
Es wird hier

$$\Delta^{N-1}(t_1, \ldots, t_N)e^{tx} = \frac{\Delta^{N-2}(t_1, \ldots, t_{N-1})e^{tx} - \Delta^{N-2}(t_2, \ldots, t_N)e^{tx}}{t_1 - t_N}$$

benutzt.
Aus $t_1 \neq t_N$ folgt für $1 \leqq i \leqq N$: Gibt es m_i Komponenten in $(t_1, \ldots, t_N)$ mit dem Wert t_i, dann gilt dies mindestens für eines der (N-1)-Tupel $(t_1, \ldots, t_{N-1})$ und $(t_2, \ldots, t_N)$. Mit der Induktionsannahme ergibt sich damit die Behauptung, da für Polynome Q_1, Q_2 der Differenz auf der rechten Seite die Beziehung $grd(Q_1+Q_2) \leqq max(grd(Q_1), grd(Q_2))$ gilt.
Die Behauptung ist somit vollständig bewiesen.

Hilfsatz 1.2 (Braess, [2])
Für $N \in \mathbb{N}$ und mit $t_i \in \mathbb{R}$ für $1 \leqq i \leqq N$ ist die Menge

$$B = \{\Delta^i(t_1, \ldots, t_{i+1})e^{tx} \mid 0 \leqq i \leqq N-1\}$$

linear unabhängig.
Beweis:
Nach Hilfssatz 1.1 ist die Behauptung gezeigt, wenn die Wronski-Determinante der Elemente von B für $x = 0$ von Null verschieden ist (vgl. [5], [7]).
Für $0 \leqq m < n$ gilt wegen (1.11):

$$D^m\Delta^n(t_1, \ldots, t_{n+1})e^{tx} = x^{n-m}R_{mn}(x)\ ,\ n \leqq N-1\ ,$$

wobei $R_{mn}(x)$ bis auf den Faktor x^{n-m} die Summe darstellt, die durch m-fache Anwendung der Produktregel entsteht.
Für $0 \leqq m = n \leqq N-1$ erhält man

$$D^m\Delta^n(t_1, \ldots, t_{n+1})e^{tx} = n! \int_0^1 \int_0^{s_1} \ldots \int_0^{s_{n-1}} e^{(t_1 + \sum_{i=1}^n (t_{i+1}-t_i)s_i)x} ds_n \ldots ds_1 + xR_{mm}(x)$$

$R_{mm}(x)$ stellt bis auf den Faktor x den Rest der bei der Differentiation entstehenden Summe dar.

Für $x=0$ folgt: $D^m\Delta^n(t_1,\ldots,t_{n+1})e^{tx} = \begin{cases} 0 & : 0 \leqq m < n \\ & \\ 1 & : 0 \leqq m = n \end{cases} \qquad n \leqq N-1$

Setzt man für $x = 0$

$$a_{ij} := D^i\Delta^j(t_1,\ldots,t_{j+1})e^{tx} \ , \ 0 \leqq i,j \leqq N-1 \ ,$$

so erhält man eine nichtsinguläre Dreiecksmatrix $(a_{i,j})_{0\leqq i,j\leqq N-1}$, womit die lineare Unabhängigkeit von B bewiesen ist.

Satz 1.2 (Braess, [2])
$E(x) = \sum_{i=1}^{L} P_i(x)e^{T_i x} \in V_N \setminus V_{N-1}$ besitzt eine eindeutige Darstellung durch Differenzenquotienten: $E(x) = \sum_{i=0}^{N-1} b_i\Delta^i(t_1,\ldots,t_{i+1})e^x$.
Die Koeffizienten $b_i \in \mathbb{R}$ sind für $0 \leqq i < N-1$ eindeutig bestimmt; es gilt $t_j = T_i$ für $\sum_{k=1}^{i-1}(grd(P_k)+1) \leqq j \leqq \sum_{k=1}^{i}(grd(P_k)+1)$ und $1 \leqq i < L$; dabei sei $L \in \mathbb{N}$ die Länge von E.
Beweis:
Die Menge

$$U_1 = \{\sum_{i=1}^{L} Q_i(x)e^{T_i x} \mid Q_i\text{ist ein reelles Polynom mit } grd(Q_i) \leqq grd(P_i) \text{ für } 1 \leqq i \leqq L\} \subseteq V_N$$

stellt nach [5], [7] und der Definition von V_N einen N-dimensionalen Vektorraum dar. Nach Hilfsatz 1.2 ist $\{\Delta^i(t_1,\ldots,t_{i+1})e^{tx} \mid 0 \leqq i \leqq N-1\}$ Basis eines N-dimensionalen reellen Vektorraums U_2 mit $U_2 \subseteq U_1$ nach Hilfssatz 1.1, falls für die t_i die Beziehung der Behauptung gilt.
Daraus folgt $U_1 = U_2$ und damit der erste Teil der Behauptung.
Nach Hilfssatz 1.1 und Satz 1.1 sind auch die t_i für $0 \leqq 1 \leqq N$ eindeutig bestimmt. Damit ist der Satz gezeigt.

Bemerkung:
Es ist also möglich, jedem Element von $V_N \setminus V_{N-1}$ die 2N Parameter

$$b_0,\ldots,b_{N-1},t_1,\ldots,t_N$$

gemäß Satz 1.2 zuzuordnen; verwendet man im Beispiel 1.1 die Darstellung durch Differenzenquotienten, so erhält man:

$$\begin{aligned} E(a_m,x) &= me^{(1+\frac{1}{m})x} - me^x = 0 \times \Delta^0(1)e^{tx} + 1 \times \Delta^1(1,1+\frac{1}{m})e^{tx} \\ f(x) &= xe^x \qquad\qquad\quad = 0 \times \Delta^0(1)e^{tx} + 1 \times \Delta^1(1,1\qquad)e^{tx} \end{aligned}$$

Damit ist $E(a_m)$ der Parametervektor$(0,1,1,1+\frac{1}{m}) \in \mathbb{R}^4$, $m \in \mathbb{N}$, und f der Parametervektor $(0,1,1,1) \in \mathbb{R}^4$ zugeordnet; die Konvergenz der Folge ist hier an den Parametern zu erkennen.

Es wird hierzu später benötigt:

Hilfsatz 1.3
Mit $n \in \mathbb{N}$ seien in $\mathbb{R}$ die Folgen $(t_{j,m})_{m\in\mathbb{N}}$, $1 \leqq j \leqq n+1$, mit $t_{j,m} < t_{j+1,m}$ und $\lim\limits_{m\to\infty} t_{j,m} = t_j \in \mathbb{R}$, $1 \leqq j \leqq n+1$, gegeben; weiter sei entweder $t_j < t_{j+1}$, $1 \leqq j \leqq n$, oder $t_j = t$, $1 \leqq j \leqq n+1$, erfüllt.
Dann konvergiert die Folge $(\Delta^n(t_{1,m},\ldots,t_{n+1,m})e^{tx})_{m\in\mathbb{N}}$ auf $I=[a,b]$ gleichmäßig gegen $\Delta^n(t_1,\ldots,t_{n+1})e^{tx}$.

Beweis: (nach [13])
Die Funktion f sei holomorph in $\mathbb{C}$.
Für $s_j \in \mathbb{R}$, $1 \leqq j \leqq n$, mit $s_j < s_{j+1}$, $1 \leqq j \leqq n-1$, oder $s_j = s$, $1 \leqq j \leqq n$, wird durch vollständige Induktion nach n gezeigt:

$$\Delta^{n-1}(s_1,\ldots,s_n)f = \frac{1}{2\pi \mathrm{i}} \int_K f(z) \prod_{j=1}^{n} (z-s_j)^{-1} dz \tag{1.12}$$

wobei $K = \{z \in \mathbb{C} \mid R = \mid z \mid\}$ mit $\max\limits_{1\leqq j\leqq n} \mid s_j \mid < R < \infty$ sei.

$n=1$:

$$\Delta^0(s_1)f = f(s_1) = \frac{1}{2\pi \mathrm{i}} \int_K f(z)(z-s_1)^{-1} dz \ .$$

Gilt (1.12) für n, dann folgt:
Für $s_1 \neq s_{n+1}$ erhält man

$$\Delta^n(s_1,\ldots,s_{n+1})f = \frac{\Delta^{n-1}(s_1,\ldots,s_n)f - \Delta^{n-1}(s_2,\ldots,s_{n+1})f}{s_1 - s_{n+1}} =$$

$$= \frac{1}{s_1-s_{n+1}} \frac{1}{2\pi \mathrm{i}} \int_K [f(z)(z-s_{n+1}) - f(z)(z-s_1)] \prod_{j=1}^{n+1} (z-s_j)^{-1} dz =$$

$$= \frac{1}{2\pi \mathrm{i}} \int_K f(z) \prod_{j=1}^{n+1} (z-s_j)^{-1} dz$$

Aus (1.10) und dem Cauchyschen Integralsatz für Ableitungen folgt für $s_j = s$, $1 \leqq j \leqq n+1$, unmittelbar

$$\Delta^n(s,\ldots,s)f = \frac{1}{2\pi \mathrm{i}} \int_K f(z) \prod_{j=1}^{n+1} (z - s_j)^{-1} dz$$

Damit ist (1.12) vollständig gezeigt.

Es sei nun $\max\{\max\limits_{1\leqq i\leqq n+1} \mid t_i \mid , 1\} < R < \infty$.

Es kann o.B.d.A. $\mid t_{j,m} \mid < \dfrac{3}{2}R$ für $1 \leqq j \leqq n+1$ angenommen werden.
Für $x \in I$ und $z \in \mathbb{C}$ mit $\mid z \mid = 2R$ erhält man:

$$\begin{aligned}
&\mid e^{zx} \prod_{j=1}^{n+1} (z - t_j)^{-1} - e^{zx} \prod_{j=1}^{n+1} (z - t_{j,m})^{-1} \mid = \\
&= \mid e^{zx} \mid \mid \prod_{j=1}^{n+1} (z - t_{j,m}) - \prod_{j=1}^{n+1} (z - t_j) \mid \prod_{j=1}^{n+1} (\mid z - t_j \mid \mid z - t_{j,m} \mid)^{-1} \leqq \\
&\leqq e^{2R(|a|+|b|)} (R\frac{R}{2})^{-n-1} \mid \sum_{j=0}^{n} z^j (S_{j,m} - S_j) \mid \leqq \\
&\leqq e^{2R(|a|+|b|)} (R\frac{R}{2})^{-n-1} (2R)^{n+1} \sum_{j=0}^{n} \mid S_{j,m} - S_j \mid := M \sum_{j=0}^{n} \mid S_{j,m} - S_j \mid \ ;
\end{aligned}$$

dabei sind S_j bzw. $S_{j,m}$ die elementarsymmetrischen Funktionen von t_k, $1 \leqq k \leqq n+1$. Mit den Ergebnissen von oben erhält man für alle $x \in I$

$$\begin{aligned}
\mid \Delta^n(t_1,\ldots,t_{n+1})e^{tx} \ - \ \Delta^n(t_{1,m},\ldots,t_{n+1,m})e^{tx} \mid \leqq \\
\leqq \ \frac{1}{2\pi} 4R\pi M \sum_{j=0}^{n} \mid S_{j,m} - S_j \mid
\end{aligned}$$

Wegen $\lim\limits_{m\to\infty} \sum\limits_{j=0}^{n} \mid S_{j,m} - S_j \mid = 0$ folgt damit die Behauptung.

1.3 Vorzeichenklassen in V_N

Ein weiteres Hilfsmittel zur Beschreibung der Struktur von V_N, soweit dies hier erforderlich ist, liefert die Einführung von Vorzeichenklassen in V_N nach Braess, [2].
Da, wie in Abschnitt 1.1 angekündigt, die Existenz einer besten Approximation einmal durch Erweiterung von V_N^0 und zum anderen durch Beschränkung auf geeignete Teilmengen, zum Beispiel V_N^+ oder V_N^- gesichert werden kann, liegt es nahe, die Vorzeichen der Summanden in Exponentialsummen zu untersuchen.

Jeder Exponentialsumme wird rekursiv ein *Vorzeichenvektor* zugeordnet:

Definition 1.7

a. $E(x) = (\sum_{i=0}^{n} a_i x^i)e^{tx}$ mit $a_n \neq 0$ und $s = sign(a_n)$ hat den Vorzeichenvektor $sign(E)$ mit $n+1$ Komponenten:

$$sign(E) := ((-1)^n s, \ldots, (-1)^1 s, s)$$

Dabei ist für $a \in \mathbb{R}$ $sign(a) := \begin{cases} 1 & \text{für } a > 0 \\ 0 & \text{für } a = 0 \\ -1 & \text{für } a < 0 \end{cases}$

b. Für $E_1, E_2 \in V_N$ seien die Frequenzen von E_1 kleiner als die von E_2; $sign(E_j)$ habe m_j Komponenten, $j = 1, 2$.
Dann hat E_1+E_2 den Vorzeichenvektor $sign(E_1+E_2)$ mit m_1+m_2 Komponenten; dabei sind die ersten m_1 Komponenten die von $sign(E_1)$, die restlichen m_2 Komponenten die von $sign(E_2)$:

$$sign(E_1 + E_2) := (sign(E_1), sign(E_2))$$

Beispiele hierzu:

- $sign((x+5)e^{-3x}) = (-1, 1)$
- $sign(4e^x) = (1)$
- $sign(-x^2e^{2x}) = (-1, 1, -1)$
- $sign((x+5)e^{-3x} + 4e^x - x^2e^{2x}) = (-1, 1, 1, -1, 1, -1)$

Bemerkung:

a. Die Zahl der Komponenten des Vorzeichenvektors $sign$(E) ist gleich dem Grad von E.

b. Für $E(x) = \sum_{i=1}^{N} a_i e^{t_i x}$ mit $grd(E) = N$ gibt die i-te Komponente von $sign$(E) das Vorzeichen von a_i an.

Definition 1.8
Es sei S ein Vorzeichenvektor mit N Komponenten.

a. $V_N(S) := \{E \mid E \in V_N, (sign(E) = S) \vee$ (man erhält sign(E) aus S durch Streichen von Komponenten in S) $\}$
$V_N(S)$ ist die *Vorzeichenklasse* zu S in V_N.

b. $V_N^0(S) := V_N(S) \cap V_N^0$

Bemerkung 1.3

a. $V_N^+ = V_N^0(1, 1, \ldots, 1) = V_N(1, 1, \ldots, 1)$
Entsprechendes gilt für V_N^-.

b. $E \in V_N$ mit $grd(E) < N$ gehört zu mehreren Vorzeichenklassen in V_N, wie die Beispiele zu Definition 1.6 zeigen.
Nur die Elemente von $V_N \setminus V_{N-1}$ gehören zu genau einer Vorzeichenklasse, so daß damit in $V_N \setminus V_{N-1}$ eine Einteilung in disjunkte Teilmengen erklärt ist.

Der folgende Satz stellt eine Beziehung zwischen der Zahl der Nullstellen einer Exponentialsumme und ihrem Vorzeichenvektor her.

Satz 1.3 (Braess, [1])
Voraussetzung:
Es sei $E \in V_N$ mit $grd(E) = N \geqq 1$ gegeben und E besitze m reelle Nullstellen.
Behauptung:
In $S = sign(E)$ gibt es mindestens m Vorzeichenwechsel, d.h. es gibt mindestens m Komponenten S_i von S mit $S_i \neq S_{i+1}$.

Beweis:
Vollständige Induktion nach N:

N=1: $E(x) = ae^{tx}$ besitzt nach Voraussetzung keine reelle Nullstelle.
Der Satz sei gezeigt für N-1.

Induktionsschluß:
$E(x) = \sum_{i=1}^{L} P_i(x)e^{t_i x} \in V_N$ mit der Länge L habe $m > 0$ reelle Nullstellen; für $m = 0$ ist nichts zu zeigen.
In S liegt damit mindestens ein Vorzeichenwechsel vor, denn sonst gilt nach Bemerkung 1.3a $\mid E(x) \mid > 0$ für $x \in \mathbb{R}$ im Widerspruch zu $m > 0$.
Es sei also $w > 0$ die Zahl der Vorzeichenwechsel in S; k sei der kleinste Index mit $S_k = -S_{k+1}$.

Fall 1:

$$E(x) = \sum_{i=1}^{k} a_i e^{t_i x} + \sum_{i=k+1}^{L} P_i(x)e^{t_i x}$$

Für $E_1(x) = e^{-t_k x}E(x)$ gilt

$$DE_1(x) = \sum_{i=1}^{k-1}(t_i - t_k)a_i e^{(t_i - t_k)x} + \sum_{i=k+1}^{L} DP_i(x) + P_i(x)(t_i - t_k))e^{(t_i - t_k)x}$$

E_1 hat die gleichen Nullstellen wie E und $sign(E_1) = sign(E)$. Beachtet man $t_i < t_k$ für $i < k$, so liegen in $sign(DE_1)$ $w - 1$ Vorzeichenwechsel vor. Nach dem Satz von Rolle besitzt DE_1 mindestens $m - 1$ reelle Nullstellen. Wegen $DE_1 \in V_{N-1}$ erhält man nach Induktionsannahme $w - 1 \geqq m - 1$, also $w \geqq m$.

Fall 2:

$$E(x) = \sum_{i=1}^{L} P_i(x)e^{t_i x} \text{ mit } grdP_1 \geqq 1$$

Es ist also $k = 1$.

$$\begin{aligned} E_1(x) &= e^{-t_1 x}E(x) = P_1(x) + \sum_{i=2}^{L} P_i(x)e^{(t_i - t_1)x} \\ DE_1(x) &= DP_1(x) + \sum_{i=2}^{L}(DP_i(x) + (t_i - t_1)P_i(x))e^{(t_i - t_1)x} \end{aligned}$$

In $sign(DE_1)$ liegen $w - 1$ Vorzeichenwechsel vor, wie man durch Vergleich mit E_1 erkennt; damit folgt wie in Fall 1: $w \geqq m$.

Fall 3:

$$E(x) = \sum_{i=1}^{k-1} a_i e^{t_i x} + \sum_{i=k}^{L} P_i(x) e^{t_i x} \text{ mit } grd(P_k) \geqq 1$$

$$\begin{aligned} E_1(x) &= e^{-t_k x} E(x) = \sum_{i=1}^{k-1} a_i e^{(t_i - t_k)x} + P_k(x) + \sum_{i=k+1}^{L} P_i(x) e^{(t_i - t_k)x} \\ DE_1(x) &= \sum_{i=1}^{k-1} a_i (t_i - t_k) e^{(t_i - t_k)x} + DP_k(x) + \\ &\quad + \sum_{i=k+1}^{L} (DP_i(x) + P_i(x)(t_i - t_k)) e^{(t_i - t_k)x} \end{aligned}$$

Mit den gleichen Überlegungen wie oben erhält man auch hier $w \geqq m$.

Damit ist der Satz bewiesen.

2 Existenzsätze

2.1 Gleichmäßige Konvergenz im Innern

In diesem Kapitel sei I stets das Intervall $[a, b]$ mit $a < b$.
Während bei linearer Approximation die Existenz einer Minimallösung stets gesichert ist, können Existenzsätze für nichtlineare Approximation nur unter einschneidenden Voraussetzungen bewiesen werden.

Für $V \subseteq C(I)$ hat Rice in [15] gezeigt:
Es existiert eine beste Approximation an $f \in C(I)$ bezüglich V auf I, wenn V den beiden folgenden Forderungen genügt:

1. Für $u, v \in V$ gilt: $u - v$ besitzt in I höchstens n Nullstellen oder verschwindet auf I identisch.
2. Jede punktweise konvergente Folge in V konvergiert gegen ein Element von V.

Aus Forderung 1 folgt für jede auf I gleichmäßig beschränkte Teilmenge von V die Existenz einer auf I punktweise konvergenten Folge und mit Forderung 2 folgt daraus die Existenz einer besten Approximation bezüglich V ([15], Theorem 7-2).
Dabei heißt $U \subseteq C(I)$ *gleichmäßig beschränkt* auf I, wenn es ein $K < \infty$ gibt mit

$$\| u \|_I \leqq K \text{ für } u \in U$$

V_N genügt der ersten Forderung mit $n = 2N - 1$ nach Satz 1.1, nicht jedoch der zweiten:

Beispiel 2.1
Man betrachte in $[0, 1]$ die Funktionen $E(a_m, x) = e^{-mx} + e^{m(x-1)} \in V_2^0$ für $m \in \mathbb{N}$:

$$\lim_{m\to\infty} E(a_m, x) = \begin{cases} 1 & : x = 0 \\ 0 & : x \in (0, 1) \\ 1 & : x = 1 \end{cases}$$

Will man beim Nachweis der Existenz einer Minimallösung bezüglich V_N analog vorgehen, so muß man also einen anderen Konvergenzbegriff verwenden; es wird definiert:

Definition 2.1
Eine Folge von Funktionen aus C(I) *konvergiert gleichmäßig im Innern* von I gegen $g \in C(I)$, wenn sie auf jedem abgeschlossenen Teilintervall von (a, b) gleichmäßig gegen g konvergiert.

Die in Beispiel 2.1 gegebene Folge $(E(a_m))_{m\in\mathbb{N}}$ konvergiert im Innern von $[0, 1]$ gleichmäßig gegen ein Element von V_2, nämlich die Nullfunktion;
es zeigt sich, daß dies allgemein für jede gleichmäßig beschränkte Folge in V_N gilt und daß diese Eigenschaft die Existenz einer besten Approximation gewährleistet. Der Beweis wird für eine allgemeinere Menge von Funktionen durchgeführt, die eine der Forderung 2 entsprechende Eigenschaft besitzt:

Definition 2.2

$$DZ^n(I) := \{h \mid h \in C^{n+1}(I), \quad D^{n+1}h \text{ hat in I höchstens } n-1 \text{ Nullstellen oder verschwindet auf I identisch } \}, \ n \in \mathbb{N}$$

Bemerkung 2.1
Nach Satz 1.1 und Bemerkung 1.2 gilt $V_N \subseteq DZ^N(I)$ für $N \in \mathbb{N}$.

Definition 2.3
Für $d > 0$, $h \in C^m(I)$ und $n \in \mathbb{N}$ wird gesetzt:

$$Z^m(h) := \begin{cases} \emptyset & : D^m h \equiv 0 \text{ auf I} \\ \{x \in I \mid D^m h(x) = 0\} & : D^m h \not\equiv 0 \text{ auf I} \end{cases} \qquad 0 \leqq m \leqq n$$

$$I^m(h,d) := \{x \in I \mid (m \times d) \leqq | \, x - z \, | \text{ für } z \in \{a, b\} \cup \bigcup_{i=0}^{m+1} Z^i(h)\}, \qquad 0 \leqq m \leqq n-1$$

Hilfsatz 2.1 (Werner, [20])
Es sei $n \in \mathbb{N}$ und $h \in C^{n+1}(I)$ gegeben. Mit $d > 0$ sei $I^i(h, d)$ nicht leer, $0 \leqq i \leqq n$.
Behauptung:

$$\| D^i h \|_{I^i(h,d)} \leqq d^{-i} \| h \|_I \qquad 0 \leqq i \leqq n$$

Beweis: Vollständige Induktion nach i:

Man setze $K_{-1} := d^{-(i-1)} \| h \|_I$ für $i \in \mathbb{N}$.

Für $i = 0$ ist nichts zu zeigen.

Die Behauptung sei für $i - 1 < n$ gezeigt, d.h. mit $g = D^{i-1}h$ gilt:
$\| g \|_{I^{i-1}(h,d)} \leqq K_{i-1}$

Induktionsschluß:
In I sei $Dg \not\equiv 0$, sonst ist nichts zu zeigen.
Mit den Punkten von $Z^{i+1}(h)$ erhält man in $I^{i-1}(h,d)$ abgeschlossene Teilintervalle J_k; falls $Z^{i+1}(h)$ leer ist oder für $x \in I^{i-1}(h,d)$ stets $D^{i+1}h(x) \neq 0$ gilt, sind die J_k die Teilintervalle, die $I^{i-1}(h,d)$ bilden.
Die Teilintervalle, die mit $I^i(h,d)$ einen nichtlinearen Durchschnitt besitzen, seien o.B.d.A. die Intevalle J_k, $1 \leqq k \leqq K < \infty$. Nach Definition 2.3 gilt für $1 \leqq k \leqq K$:

a. Dg ist in J_k monoton und wechselt sein Vorzeichen nicht.

b. In J_k nimmt $\mid Dg \mid$ sein Maximum in einem Randpunkt $u_k \in J_k$ an, wobei gilt: $u_k \in Z^{i+1}(h)$ oder u_k gehört zum Rand von $I^{i-1}(h,d)$.

Es sei nun $x \in J_k$. Nach Induktionsannahme gilt $\mid g(x) - g(u_k \mid \leqq K_{i-1}$ und mit dem Mittelwertsatz der Intergralrechnung erhält man

$$\mid g(x) - g(u_k) \mid = \mid \int_{u_k}^{x} Dg(t)dt \mid = \mid (x - u_k)Dg(t_0 \mid$$

für ein $t_0 \in (u_k, x)$ bzw. $t_0 \in (x, u_k)$.
Da Dg in J_k sein Vorzeichen nicht wechselt, ergibt sich mit Punkt b:
$\mid Dg(t_0) \mid \geqq \mid Dg(x) \mid$ und damit $K_{i-1} \geqq \mid x - u_k \mid \ \mid Dg(x) \mid$.

Gilt insbesondere $x \in I^i(h,d) \cap J_k$, dann folgt $\mid x - u_k \mid \geqq d$ und damit $K_{i-1} \geqq d \mid Dg(x) \mid$.
Für $x \in I^i(h,d)$ hat man also erhalten: $\mid D^i h(x) \mid \leqq d^{-1}K_{i-1} = d^{-i} \parallel h \parallel_I$; dies war zu zeigen.

Eine weitere Abschätzung für Dh liefert Hilfssatz 2.2:

Hilfsatz 2.2 (Schmidt, [17])
Es sei $n \in \mathbb{N}$ und $h \in DZ^n(I)$ gegeben. Mit $d = b - a$ und

$$K := \max_{1 \leqq i \leqq n} \max\{\mid D^i h(a) \mid, \mid D^i h(b) \mid\}$$

gilt:

$$\parallel Dh \parallel_I \leqq \ K \sum_{j=0}^{n-1} d^j$$

Beweis:
Nach Definition von K gilt

$$\mid Dh(a) \mid \leqq K \sum_{j=0}^{n-1} d^j \geqq \mid Dh(b) \mid$$

Es ist also noch zu zeigen:

$$\mid Dh(x) \mid \leqq K \sum_{j=0}^{n-1} d^j \qquad x \in (a,b)$$

Beweisschritt 1:
Rekursiv wird definiert:

$$\begin{aligned} K_1 &:= K \sum_{j=0}^{n-1} d^j \\ K_i &:= (K_{i-1} - K)d^{-1} \qquad 2 \leqq i \leqq n \end{aligned}$$

Vollständige Induktion nach i ergibt:

$$K_i = K \sum_{j=0}^{n-i} d^j \qquad 1 \leqq i \leqq n \tag{2.1}$$

Für $i = 1$ ist nichts zu zeigen; es gelte (2.1) für $i < n$.
Induktionsschluß:

$$K_{i+1} = (K_i - K)d^{-1} = Kd^{-1}(\sum_{j=0}^{n-i} d^j - 1) = Kd^{-1} \sum_{j=1}^{n-i} d^j = K \sum_{j=0}^{n-i-1} d^j$$

Damit ist (2.1) vollständig gezeigt.
Unmittelbar folgt hieraus:

$$K_i \geqq K \qquad 1 \leqq i \leqq n \tag{2.2}$$

Beweisschritt 2:
Annahme: die Behauptung des Satzes ist falsch, es gibt ein $x_0 \in (a,b)$ mit

$$\mid Dh(x_0) \mid > K \sum_{j=0}^{n-1} d^j$$

Durch vollständige Induktion nach i wird gezeigt:

$$\begin{array}{l} D^i h \text{ besitzt in } (a,b)\ i-1 \text{ Nullstellen} x_1^i < x_2^i < \ldots < x_{i-1}^i \\ \text{und es ist } \mid D^{i-1}h(x_r^i) \mid > K_{i-1},\ r \in \{1, i-1\}, \text{ für } 2 \leqq i \leqq n+1 \end{array} \tag{2.3}$$

$i = 2$: $\mid Dh(a) \mid \leqq K_1 \geqq \mid Dh(b) \mid$ gilt nach Voraussetzung.
Wegen $\mid Dh(x_0) \mid > K_1 \geqq K$ und $h \in C^{n+1}(I)$ gibt es ein $x_1^2 \in (a,b)$ mit $\mid Dh(x_1^2) \mid > K_1$ und $D^2h(x_1^2) = 0$.
Damit ist (2.3) für $i = 2$ gezeigt.
Induktionsannahme: (2.3) gilt für $i < n+1$.
Induktionsschluß: Für $i+1$ erhält man mit dem Satz von Rolle:

$$\begin{array}{c} D^{i+1}h \text{ hat in } (x_1^i, x_{i-1}^i) \text{ mindestens } i-2 \\ \text{Nullstellen } x_2^{i+1} < x_3^{i+1} < \ldots < x_{i-1}^{i+1} \end{array} \tag{2.4}$$

Der Mittelwertsatz der Differentialrechnung liefert mit der Induktionsannahme:
Es gibt ein $x_1 \in (a, x_1^i)$ und ein $x_2 \in (x_{i-1}^i, b)$ mit

$$D^i h(x_1) = \frac{D^{i-1}h(x_1^i) - D^{i-1}h(a)}{x_i^i - a} \text{ und } D^i h(x_2) = \frac{D^{i-1}h(b) - D^{i-1}h(x_{i-1}^i)}{b - x_{i-1}^i}$$

Es folgt:

$$\begin{array}{rcl} \mid D^i h(x_1) \mid & \geqq & \dfrac{\mid D^{i-1}h(x_1^i) \mid - \mid D^{i-1}h(a) \mid}{\mid x_i^i - a \mid} \text{ und} \\ \mid D^i h(x_2) \mid & \geqq & \dfrac{\mid D^{i-1}h(x_{i-1}^i) \mid - \mid D^{i-1}h(b) \mid}{\mid x_{i-1}^i - b \mid} \end{array}$$

Mit $D^i h(x_1^i) = D^i h(x_{i-1}^i) = 0$, $\mid D^{i-1}h(a) \mid \leqq K \geqq \mid D^{i-1}h(b) \mid$, $\mid x_1^i - a \mid < d > \mid b - x_{i-1}^i \mid$ und der Induktionsannahme erhält man:
$\mid D^i h(x_1) \mid > (K_{i-1} - K)d^{-1} = K_i < \mid D^i h(x_2) \mid$
Wie oben folgt mit (2.2): Es gibt ein $x_1^{i+1} \in (a, x_1^i)$ und ein $x_i^{i+1} \in (x_{i-1}^i, b)$ mit

$$\begin{array}{ll} \mid D^i h(x_1^{i+1}) \mid > K_i\ ,\ D^{i+1}h(x_1^{i+1}) = 0 & \text{und} \\ \mid D^i h(x_i^{i+1}) \mid > K_i\ ,\ D^{i+1}h(x_i^{i+1}) = 0 & \end{array}$$

Mit (2.4) erhält man insgesamt: $D^{i+1}h$ besitzt in (a,b) $(i-2)+2=i$ Nullstellen $x_1^{i+1} < x_2^{i+1} < \ldots < x_i^{i+1}$ mit

$$\mid D^i h(x_r^{i+1}) \mid > K_i,\ r \in \{1,i\}$$

Damit ist (2.3) gezeigt.

Auf $i = n+1$ angewandt erhält man:
$D^{n+1}h$ besitzt in (a,b) n Nullstellen $x_1^{n+1} < x_2^{n+1} < \ldots < x_n^{n+1}$ und $D^{n+1}h$ verschwindet auf I nicht identisch. Dies steht im Widerspruch zu $h \in DZ^n(I)$. Damit ist die Annahme widerlegt und die Behauptung des Satzes gilt.

Nach Bemerkung 1.2 folgt aus Hilfssatz 2.2 wegen $V_N \subseteq DZ^N(I)$ unmittelbar
Korollar 2.1
Für $h \in V_N$ mit $N \in \mathbb{N}$ gilt

$$\parallel D^m h \parallel_I \leqq\ K_m \sum_{j=0}^{N-1} d^j \qquad m \in \mathbb{N}$$

wobei $d = b-a$ und $K_m = \max\limits_{m \leqq i \leqq m+N-1} \max\{\mid D^i h(a) \mid, \mid D^i h(b) \mid\}$ ist.

Satz 2.1 (Schmidt, [17])
Es sei $n \in \mathbb{N}$ und $I_1 = [a_1, b_1] \subseteq (a,b)$ mit $d_0 = \min\{a_1 - a, b - b_1\} > 0$ gegeben, $d = b-a$.
Behauptung: Es gibt ein $K(I, I_1) < \infty$, so daß jedes $h \in DZ^n(I)$ erfüllt:

$$\parallel Dh \parallel_{I_1} \leqq K(I, I_1) \parallel Dh \parallel_I$$

Beweis:
Beweisschritt 1:
Für $h \in DZ^n(I)$ gilt $\mid Z^{n+1}(h) \mid \leqq n-1$ und aus dem Satz von Rolle folgt $\mid Z^i(h) \mid \leqq \mid Z^{i+1}(h) \mid +1$ für $0 \leqq i \leqq n$. Damit erhält man:

(2.5) $\qquad \mid Z^{n+1-k}(h) \mid \leqq \mid Z^{n+1}(h) \mid +k \leqq n-1+k$ für $0 \leqq k \leqq n+1$

Für $k = 0$ gilt (2.5) unmittelbar.
Für $k < n+1$ sei (2.5) gezeigt; es ergibt sich damit $\mid Z^{n+1-(k+1)}(h) \mid \leqq \mid Z^{n+1-k}(h) \mid +1 \leqq n-1+(k+1)$ nach Induktionsannahme und (2.5) ist gezeigt.
Für jedes $h \in DZ^n(I)$ erhält man damit

$$s = \sum_{k=0}^{n+1}(n-1+k) \;=\; (n+2)(n-1) + \sum_{k=0}^{n+1} k = (n+2)\frac{3n-1}{2} \geqq$$

$$\geqq\ \sum_{i=0}^{n+1} Z^i(h)$$

Beweisschritt 2:
Wegen $3 \leqq s < \infty$ für $n \in \mathbb{N}$ existieren für jedes $h \in DZ^n(I)$ abgeschlossene Intervalle $I_1(h)$, $I_2(h)$ mit der Länge $d_1 = \dfrac{d_0}{2(s+1)} > 0$, die $I_1(h) \subseteq (a, a_1]$, $I_2(h) \subseteq [b_1, b)$ und $I_j(h) \cap \bigcup_{i=0}^{n+1} Z^i(h) = \emptyset$ für $j = 1, 2$ erfüllen.
($[a, b]$ hat die Länge $b - a$, $\dfrac{a+b}{2}$ ist die Intervallmitte)
Es seien $a(h)$ bzw. $b(h)$ die Intervallmitten von $I_1(h)$ bzw. $I_2(h)$ und es gilt daher

$$| a(h) - z | > \frac{d_1}{2} < | b(h) - z | \qquad \text{für } z \in \bigcup_{i=0}^{n+1} Z^i(h)$$

Hilfssatz 2.1 ergibt für $a(h), b(h)$:

$$| D^i h(a(h)) | \leqq \| h \|_I \, (\frac{d_1}{2n})^{-i} \geqq | D^i h(b(h)) | \,, \quad 0 \leqq i \leqq n$$

Mit $K := \| h \|_I \max_{1 \leqq i \leqq n} (\frac{d_1}{2n})^{-i}$ gilt also für $1 \leqq i \leqq n$

$$| D^i h(a(h)) | \leqq K \geqq | D^i h(b(h)) |$$

Hilfssatz 2.2, angewandt auf $I(h) = [a(h), b(h)]$, ergibt

$$\| Dh \|_{I(h)} \leqq K \sum_{j=0}^{n-1} (b(h) - a(h))^j \leqq K \sum_{j=0}^{n-1} d^j$$

Aus $I_1 \subseteq I(h)$ folgt

$$\| Dh \|_{I_1} \leqq \| h \|_I \max_{1 \leqq i \leqq n} (\frac{d_1}{2n})^{-i} \sum_{j=0}^{n-1} d^j$$

Mit $K(I, I_1) := \max_{1 \leqq i \leqq n} (\frac{d_1}{2n})^{-i} \sum_{j=0}^{n-1} d^j$ erhält man die Behauptung für jedes $h \in DZ^n(I)$, da $K(I, I_1)$ allein durch die Größen d und d_0 und damit durch I und I_1 bestimmt ist.

Speziell für $h \in V_N$ läßt sich mehr aussagen:

Satz 2.2 (Werner, [21])
Es sei $n \in \mathbb{N}$ und I_1, d und d_0 wie in Satz 2.1. Für jedes $j \in \mathbb{N}$ gibt es ein $K(j, I, I_1) < \infty$, so daß für $h \in V_N$ gilt:

$$\| D^j h \|_{I_1} \leqq K(j, I, I_1) \| h \|_I$$

Beweis: (analog zu Satz 2.1)
Beweisschritt 1:
Wegen Satz 1.1 gilt hier für alle $h \in V_N$:

$$s_n = (N-1)(n+2) \geqq | \bigcup_{i=0}^{n+1} Z^i(h) | \qquad n \in \mathbb{N}$$

Beweisschritt 2:
Es sei $n \in \mathbb{N}$.
Mit $0 \leqq s_n \leqq \infty$ existieren für jedes $h \in V_N$ Intervalle $I_1(h)$ und $I_2(h)$ mit der Länge $d_n = \dfrac{d_0}{2(s_n+1)}$, so daß gilt:

$$I_1(h) \subseteq (a, a_1],\ I_2(h) \subseteq [b_1, b) \text{ und } I_j(h) \cap \bigcup_{i=0}^{n+1} Z^i(h) = \emptyset, \quad j = 1, 2$$

Man erhält wieder (Bezeichnungen wie oben):

$$| D^i h(a(h)) | \leqq \| h \|_I \, (\frac{d_n}{2n})^{-i} \geqq | D^i h(b(h)) | \qquad i \in \mathbb{N}\ ,\ i \leqq n$$

Mit $K_j := \max\limits_{j \leqq i \leqq N+j-1} (\frac{d_n}{2n})^{-i} \| h \|_I$ für $j \in \mathbb{N}$ gilt für $j \leqq i \leqq N+j-1$ mit $N+j-1 \leqq n$: $| D^i h(a(h)) | \leqq K_j \geqq | D^i h(b(h)) |$.
Nach Korollar 2.1 folgt für $N+j-1 \leqq n$:

$$\begin{aligned} \| D^j h \|_{I(h)} &\leqq [\sum_{j=0}^{N-1} (b(h) - a(h))^j] K_j \text{ , also} \\ \| D^j h \|_{I_1} &\leqq \| h \|_I \, K(j, I, I_1) \end{aligned}$$

wobei für $j \in \mathbb{N}$ mit $N+j-1 \leqq n$ gesetzt wird: $K(j, I, I_1) := \max\limits_{j \leqq i \leqq N+j-1} (\frac{d_n}{2n})^{-i} \sum\limits_{j=0}^{N-1} d^j$

Da $n \in N$ beliebig angenommen war, gilt die Behauptung.

Wir sind nun in der Lage, einen Satz zu beweisen, der Theorem 7-2 in [15] entspricht:

Satz 2.3 (Schmidt, [17])
Mit $n \in \mathbb{N}$ sei $M \subseteq DZ^n(I)$ auf I gleichmäßig beschränkt.
Behauptung:
Es gibt eine Folge von Elementen aus M, die im Innern von I gleichmäßig konvergiert.

Beweis:
Es sei $(I_n)_{n\in\mathbb{N}}$ eine Intervallfolge in I mit $(I_n) = [a_n, b_n]$,

$$a_n = a + \frac{b-a}{n+2} \text{ , } b_n = b + \frac{b-a}{n+2}$$

Nach Voraussetzung gibt es ein $K < \infty$ mit $\| h \|_I < K$ für alle $h \in M$. Damit ist M auch gleichmäßig stetig in I_n für $n \in \mathbb{N}$:
Denn nach Satz 2.1 gibt es ein $K(I, I_n) < \infty$, so daß für $h \in M$ gilt:

$$\| Dh \|_{I_n} \leqq K(I, I_n) \| h \|_I \leqq K(I, I_n)K := K_n < \infty$$

Für $n \in \mathbb{N}$ und $x_1, x_2 \in I$ mit $x_1 < x_2$ existiert nach dem Mittelwertsatz der Differentialrechnung ein $x \in (x_1, x_2)$ mit $\mid h(x_1 - h(x_2 \mid = \mid Dh(x) \mid\mid x_1 - x_2 \mid$. Hieraus folgt für alle $h \in M \mid h(x_1 - h(x_2 \mid \leqq K_n \mid x_1 - x_2 \mid$, womit die gleichmäßige Stetigkeit von M auf I_n, $n \in \mathbb{N}$, nachgewiesen ist.

Nach dem Satz von Ascoli ([6], S. 127) erhält man:
In M existiert eine auf I_1 gleichmäßig konvergente Folge $(h_{1,m})_{m\in\mathbb{N}}$.
$(h_{1,m})_{m\in\mathbb{N}} \subseteq M$ ist gleichmäßig beschränkt auf I und gleichmäßig stetig auf I_2. Damit enthält $(h_{1,m})_{m\in\mathbb{N}}$ eine auf I_2 gleichmäßig konvergente Folge $(h_{2,m})_{m\in\mathbb{N}}$. Wegen $(h_{2,m})_{m\in\mathbb{N}} \subseteq M$ enthält wiederum $(h_{2,m})_{m\in\mathbb{N}}$ eine Folge $(h_{3,m})_{m\in\mathbb{N}}$, die auf I_3 gleichmäßig konvergiert, usw.; man erhält so für $n \in \mathbb{N}$ Teilfolgen $(h_{n,m})_{m\in\mathbb{N}} \subseteq M$, die auf I_n, und damit auch auf I_j, $1 \leqq j \leqq n$, gleichmäßig konvergieren.

Wegen $h_{m,m} \in (h_{n,j})_{j\in\mathbb{N}}$ für $m \geqq n$ folgt (Auswahl nach dem Diagonalverfahren): $(h_{m,m})_{m\in\mathbb{N}} \subseteq M$ konvergiert gleichmäßig auf I_n für alle $n \in \mathbb{N}$.

Nach Konstruktion von $(I_n)_{n\in\mathbb{N}}$ gibt es für jedes abgeschlossene Intervall $J \subseteq (a, b)$ ein $n_0 \in \mathbb{N}$, so daß $J \subseteq I_n$ für $n \geqq n_0$ erfüllt ist; damit gilt die Behauptung.

Wegen $V_N \in DZ^N(I)$ ist der Satz unmittelbar auf gleichmäßig beschränkte Teilmengen von V_N anwendbar; es gilt sogar Korollar 2.2:

Korollar 2.2 (Werner, [21])
Es sei $N, J \in \mathbb{N}$.
Ist $V \subseteq V_N$ auf I gleichmäßig beschränkt, dann enthält V eine Folge $(h_n)_{n\in\mathbb{N}}$, so daß $(D^j h_n)_{n\in\mathbb{N}}$, $0 \leqq j \leqq J-1$, gleichmäßig im Innern von I konvergiert.

Beweis: Vollständige Induktion nach J:

Für $J = 1$ gilt die Behauptung nach Satz 2.3.
Die Behauptung gelte für $J \in \mathbb{N}$:
Es gebe also eine Folge $(h_m)_{m\in\mathbb{N}}$ in V, so daß $(D^j h_m)_{m\in\mathbb{N}}$ für $0 \leqq j \leqq J-1$ gleichmäßig im Innern von I konvergiert. $(I_n)_{n\in\mathbb{N}}$ sei die Intervallfolge von Satz 2.3; nach Satz 2.2 gibt es für alle $n \in \mathbb{N}$ ein $K(J, I, I_n) < \infty$ mit

$$\| D^J h_m \|_{I_n} \leqq K(J, I, I_n) \| h_m \|_I \qquad m \in \mathbb{N}$$

Da V auf I gleichmäßig beschränkt ist, ist auch $(D^J h_m)_{m\in\mathbb{N}}$ auf I_n für $n \in \mathbb{N}$ gleichmäßig beschränkt. Damit gibt es in $(h_m)_{m\in\mathbb{N}}$ eine Teilfolge $(h_{1,m})_{m\in\mathbb{N}}$, so daß $(D^J h_{1,m})_{m\in\mathbb{N}}$ im Innern von I_1 gleichmäßig konvergiert; $(h_{1,m})_{m\in\mathbb{N}}$ wiederum enthält eine Teilfolge $(h_{2,m})_{m\in\mathbb{N}}$, so daß $(D^J h_{2,m})_{m\in\mathbb{N}}$ im Innern von I_2 gleichmäßig konvergiert, usw.:
Für $n > 1$ gibt es also eine Folge $(h_{n,m})_{m\in\mathbb{N}} \subseteq (h_{n-1,m})_{m\in\mathbb{N}}$, so daß $D^J(h_{n,m})_{m\in\mathbb{N}}$ im Innern von I_n gleichmäßig konvergiert.
Die Folge $(D^J h_{m,m})_{m\in\mathbb{N}}$ konvergiert also gleichmäßig im Innern von I_n für alle $n \in \mathbb{N}$ und damit folgt wie in Satz 2.3 für diese Folge die gleichmäßige Konvergenz im Innern von I.
Nach Konstruktion gilt:
$(D^j h_{m,m})_{m\in\mathbb{N}}$ konvergiert im Innern von I für $0 \leqq j \leqq J$ gleichmäßig.

Damit ist der Satz bewiesen.

2.2 Abgeschlossenheit, Existenzsätze

Mit Satz 2.3 erhält man die Existenz einer besten Approximation für $f \in C(I)$, wenn M abgeschlossen ist in Sinne der Definition 2.4 (man vergleiche hierzu die Forderung 2):

Definition 2.4
Eine Menge $V \subseteq C(I)$ heißt *abgeschlossen*, wenn gilt:
Konvergiert eine Folge von Elementen aus V auf einem abgeschlossenen Intervall $I_1 \subseteq (a,b)$ gleichmäßig gegen h, dann stimmt h auf I_1 mit einem Element von V überein.

Satz 2.4 (Schmidt, [18])
Es sei $M \subseteq DZ^n(I)$ mit $n \in \mathbb{N}$ abgeschlossen.
Behauptung: Jede Funktion $f \in C(I)$ besitzt eine Minimallösung bezüglich M auf I.
Beweis:
$(h_m)_{m\in\mathbb{N}} \subseteq M$ sei eine Minimalfolge für f:

$$\lim_{m\to\infty} \| f - h_m \|_I = \inf_{h\in M} \| f - h \|_I = R$$

Damit gibt es für $\delta > 0$ ein $m(\delta) \in \mathbb{N}$ mit

$$\| h_m \|_I \leqq \| f \|_I + \| f - h_m \|_I \leqq \| f \|_I + R + \delta \text{ für } m \geqq m(\delta)$$

$(h_m)_{m\in\mathbb{N}}$ ist also auf I gleichmäßig beschränkt:

$$\| h_m \|_I \leqq K, \text{ wobei}$$

$K := max\{\ max\{\| h_m \|_I \mid 1 \leqq m \leqq m(\delta)\ \},\ \| f \|_I + R + \delta\ \}$ ist.
Nach Satz 2.3 existiert in $(h_m)_{m\in\mathbb{N}}$ eine Teilfolge $(h_m)_{m\in J}$, $J \subseteq \mathbb{N}$, die gleichmäßig im Innern von I konvergiert. Da M abgeschlossen ist, konvergiert $(h_m)_{m\in J}$ auf jedem abgeschlossenen Intervall in (a,b) gleichmäßig gegen ein Element $h \in M$. Für $x \in (a,b)$ gilt also:

$$| f(x) - h(x) | = \lim_{m\to\infty} | f(x) - h_m(x) | \leqq \lim_{m\to\infty} \| f - h_m \|_I = R\ , \quad m \in J$$

Da $| f - h |$ auf I stetig ist, folgt hieraus

$$\| f - h \|_I = max\{ \begin{array}{ll} |f(a) - h(a)| & \leqq R \geqq | f(b) - h(b) | \quad \text{und damit} \\ |f(a) - h(a)|, & \\ | f(b) - h(b) |, & \sup\limits_{x\in(a,b)} | f(x) - h(x) | \} \leqq R \end{array}$$

Damit ist h eine Minimallösung für f bezüglich M.

Bemerkung:
Die Stetigkeit von f geht entscheidend bei der Ermittlung von $\| f-h \|_I$ ein.

Nach dem bisher Gezeigten ist die Aufgabe, die Existenz einer Minimallösung bezüglich V_N nachzuweisen, zurückgeführt auf das Problem, die Abgeschlossenheit von V_N zu zeigen. Es wird entscheidend benutzt, daß die Exponentialsummen Lösungen linearer Differentialgleichungen sind.

Satz 2.5 (Schmidt, [17]; Werner, [21])
Es sei $N \in \mathbb{N}$, $(h_n)_{n\in\mathbb{N}} \subseteq V_N$ konvergiere auf I gleichmäßig gegen die Funktion h.
Behauptung: $h \in V_N$
Beweis:
Die Länge und der Grad von h_n sind für alle $n \in \mathbb{N}$ höchstens gleich N; es kann daher o.B.d.A. angenommen werden:

$$h_n(x) = \sum_{i=1}^{L} P_{i,n}(x) e^{t_{i,n}x}$$

mit $L \neq 0$ die Länge von h_n und $grd(h_n) = k < N$ für $n \in \mathbb{N}$.

Für $L = 0$ ist nichts zu zeigen. Aus der gleichmäßigen Konvergenz der Folge erhält man $h \in C(I)$ und wie oben (mit $R = 0$) die gleichmäßige Beschränktheit von $(h_n)_{n\in\mathbb{N}}$:

$$\| h_n \|_I \leqq K < \infty \qquad n \in \mathbb{N}$$

Nach Korollar 2.2 gibt es eine Teilfolge $(h_n)_{n\in J_1}$, $J_1 \subseteq \mathbb{N}$, so daß die Folgen $(D^r h_n)_{n\in J_1}$, $0 \leqq r \leqq k$, gleichmäßig im Innern von I konvergieren. Für jedes abgeschlossene Intervall I_0 in (a, b) erhält man durch k-fache Anwendung des Satzes von der gliedweisen Differentiation:

(2.6)
In I_0 ist h k-fach differenzierbar
und
$(D^r h_n)_{n\in J_1}$ konvergiert auf I_0 gleichmäßig gegen $D^r h$ für $0 \leqq r \leqq k$

Es sei für $1 \leqq i \leqq L$, $n \in \mathbb{N}$ (wie üblich wird $\sum_{j=1}^{n} a_j := 0$ für $n < i$ gesetzt):

$$s_{m,n} = t_{i,n} \ , \ \sum_{j=1}^{i-1}(grd(P_{j,n}) + 1) < m \leqq \sum_{j=1}^{i}(grd(P_{j,n}) + 1)$$

Mit $C_n(t) = \prod_{m=1}^{k}(t - s_{m,n})$, $n \in \mathbb{N}$, gilt also $C_n(D)h_n = 0$, $n \in \mathbb{N}$; C_n ist das zu h_n gehörige charakteristische Polynom.
Die Folgen $(s_{m_i,n})_{n\in\mathbb{N}}$ seien für $1 \leqq i \leqq d$ beschränkt, für $d < i \leqq k$ unbeschränkt mit $d \in \mathbb{N} \cup \{0\}$; es gibt also eine (unendliche) Teilmenge $J_2 \subseteq J_1$ mit

$$\begin{aligned} \lim_{n\to\infty} s_{m_i,n} = &\quad s_i \in \mathbb{R} \quad , 1 \leqq i \leqq d, \text{ und} \\ \lim_{n\to\infty} \mid s_{m_i,n} \mid = &\quad \infty \qquad , d < i \leqq k \text{ , für } n \in J_2 \end{aligned} \tag{2.7}$$

Somit existiert ein $n_0 \in \mathbb{N}$ mit $s_{m_i,n} \neq 0$ für $d < i \leqq k$ und $n \geqq n_0$, $n \in J_2$; für $n \in J_2$, $n \geqq n_0$ gilt auf I:

$$(\prod_{i=d+1}^{k} s_{m_i,n}^{-1})C_n(D)h_n = (\prod_{i=1}^{d}(D - s_{m_i,n}))(\prod_{i=d+1}^{k}(\frac{D}{s_{m_i,n}} - 1))h_n = 0$$

(Für $n < 1$ wird $\prod_{j=1}^{n} a_j := 1$ gesetzt)

Aus (2.6) und (2.7) folgt auf I_0:

$$\lim_{n\to\infty}(\prod_{i=d+1}^{k} s_{m_i,n}^{-1})C_n(D)h_n = (-1)^{k-d}(\prod_{i=1}^{d}(D - s_i)h = 0 \text{ , } n \in J_2$$

Da $I_0 \subseteq (a, b)$ beliebig gewählt war, ist h in (a, b) und wegen $h \in C(I)$ in ganz I identisch mit einem Element aus $V_d \subseteq V_N$; dies war zu zeigen.

Dem Beweis zu Satz 2.5 entnimmt man unmittelbar:
Korollar 2.3 (Schmidt, [17])
Die Folge $(h_n)_{n\in\mathbb{N}} \subseteq V_N$ konvergiere auf I gleichmäßig gegen h; für $n \in \mathbb{N}$ gelte $h_n(x) = \sum_{i=1}^{L} P_{i,n}(x)e^{t_{i,n}x}$, $grd(h_n) = N$ und $grd(P_{i,n}) = k_i \in \mathbb{N} \cup \{0\}$.
Weiterhin seien m Folgen $(t_{i_j,n})_{n\in\mathbb{N}}$, $1 \leqq j \leqq m$, unbeschränkt.
Behauptung:
Es gilt $h \in V_{N-d}$ mit $d = \sum_{j=1}^{m}(k_{i_j} + 1)$

Speziell für V_N^+ folgt:

Korollar 2.4 (Schmidt, [17]; Werner, [21])
Konvergiert $(h_n)_{n\in\mathbb{N}} \subseteq V_N^+$ ($(h_n)_{n\in\mathbb{N}} \subseteq V_N^-$) auf I gleichmäßig gegen h, dann gilt $h \in V_N^+$ ($h \in V_N^-$).

Beweis:
Der Beweis wird für $(h_n)_{n\in\mathbb{N}} \subseteq V_N^+$ durchgeführt; im anderen Fall hat man nur $a_{i_j} \geqq 0$ bzw. $a_j \geqq 0$ durch $a_{i_j} \leqq 0$ bzw. $a_j \leqq 0$ zu ersetzen.

Aus der gleichmäßigen Konvergenz folgt wieder die gleichmäßige Beschränktheit von $(h_n)_{n\in\mathbb{N}}$:

$$\| h_n \|_I \leqq K < \infty \ , \ n \in \mathbb{N}$$

Mit $h_n(x) = \sum_{i=1}^{N} a_{i,n} e^{t_{i,n}x}$ erhält man nach Voraussetzung

$$\| a_{i,n} e^{t_{i,n}x} \|_I \leqq K \text{ und } | a_{i,n} | \leqq K e^{-t_{i,n}a} \text{ für } n \in \mathbb{N}$$

Nach Korollar 2.2 gibt es damit (nach N-maliger Auswahl von Teilfolgen) eine (unendliche) Teilmenge $J \subseteq \mathbb{N}$, so daß die Folgen $(a_{i,n}e^{t_{i,n}x})_{n\in J}$, $1 \leqq i \leqq N$, im Innern von I gleichmäßig konvergieren. Die Folgen $(t_{i_j,n})_{n\in\mathbb{N}}$ seien für $1 \leqq j \leqq N-d$ beschränkt, für $N-d < j \leqq N$ unbeschränkt; es gibt daher ein $T < \infty$ mit

$$| t_{i_j,n} | \leqq T \ , \ 1 \leqq k \leqq N-d \ , \ n \in \mathbb{N} \ ;$$

damit gilt $| a_{i_j,n} | \leqq K e^{T|a|}$ für $1 \leqq j \leqq N-d$, $n \in \mathbb{N}$. Es gibt daher eine Teilfolge mit

$$\begin{aligned} \lim_{n\to\infty} a_{i_j,n} &= a_j \in \mathbb{R} \qquad \text{und} \\ \lim_{n\to\infty} t_{i_j,n} &= t_j \in \mathbb{R} \text{ für } 1 \leqq j \leqq N-d \ , \ n \in J_1 \subseteq J \ ; \end{aligned}$$

wegen $a_{i_j,n} > 0$ für $n \in \mathbb{N}$ konvergieren also die Folgen $(a_{i_j,n}e^{t_{i_j,n}x})_{n\in J_1}$ im Innern von I gleichmäßig gegen $a_j e^{t_j x}$ mit $a_j \geqq 0$ für $1 \leqq j \leqq N-d$, für $N-d < j \leqq N$ gegen die Nullfunktion (als einzigem Element von V_0) nach Korollar 2.3. Damit gilt also

$$h(x) = \sum_{i=1}^{N-d} a_i e^{t_i x} \quad \text{für } x \in (a,b) \ ;$$

wegen $h \in C(I)$ ist dies für ganz I richtig und die Behauptung ist gezeigt.

Es ist damit gezeigt:

Korollar 2.5 (Schmidt, [17])
Die Mengen V_N, V_N^+, V_N^- sind abgeschlossen für $N \in \mathbb{N}$.

Beweis:
Es sei V eine Menge der Behauptung.
Konvergiert $(h_n)_{n\in\mathbb{N}} \subseteq V$ im Innern von I gleichmäßig, dann konvergiert diese Folge auf jedem abgeschlossenen Intervall in (a, b) gleichmäßig gegen ein $h \in V$, wie oben gezeigt; damit ist V abgeschlossen.

Aus Satz 2.4 folgt somit:
Satz 2.6 (Schmidt, [17]; Werner, [21])
Es sei $N \in \mathbb{N}$ und $f \subseteq C(I)$.
Behauptung:
Auf I besitzt f bezüglich jeder der drei Mengen V_N, V_N^+, V_N^- eine Minimallösung.

Bemerkung:
Es ist wesentlich, daß I ein reelles Intervall $[a, b]$ ist, wie Beispiel 6.1 zeigt.

Es folgen nun Sätze, die für spätere Untersuchungen benötigt werden, bei deren Beweis jedoch Satz 2.5 eingeht.
Wichtig sind Aussagen über den Vorzeichenvektor der Grenzfunktion einer gleichmäßig konvergenten Folge in V_N, die Verallgemeinerungen von Korollar 2.4 darstellen.

Hilfsatz 2.3 (Braess, [2]; Werner, [21])
Es sei $N \in \mathbb{N}$ und $h(x) = \sum_{i=0}^{N-1} a_i x^i e^{t_0 x} \subseteq V_N \setminus V_{N-1}$ mit $S = sign(h)$ gegeben;
die Folge $(h_n)_{n\in\mathbb{N}} \subseteq V_N^0 \setminus V_{N-1}$ mit $h_n(x) = \sum_{i=1}^{N} a_{i,n} e^{t_{i,n} x}$ konvergiere auf I gleichmäßig gegen h.
Behauptung:
Es gibt ein $n_0 \in \mathbb{N}$, so daß für $n \geqq n_0$ gilt: $h_n \in V_N^0(S)$.

Beweis:
Zum Beweis wird für h und h_n, $n \in \mathbb{N}$, die Darstellung durch Differenzenquotienten (nach Satz 1.2) benutzt:

$$h(x) = \sum_{i=0}^{N-1} b_i \Delta^i(t_0, \ldots, t_0) e^{tx}$$

$$h_n(x) = \sum_{i=0}^{N-1} b_{i,n} \Delta^i(t_{1,n}, \ldots, t_{i+1,n}) e^{tx}$$

Aus $\Delta^{N-1}(t_0, \ldots, t_0)e^{tx} = \dfrac{1}{(N-1)!} x^{N-1} e^{t_0 x}$ folgt

$$sign(b_{N-1}) = sign(a_{N-1})$$

Es wird nun gezeigt: Es gibt eine Teilfolge $(h_n)_{n \in J}$, $J \subseteq \mathbb{N}$ mit

$$\lim_{n \to \infty} b_{i,n} = b_i \ , \ 0 \leqq i \leqq N-1 \ , \ n \in J \tag{2.8}$$

Hierzu wähle man N Punkte x_j, $1 \leqq j \leqq N$, mit $x_1, x_2, \ldots, x_N$ in I; aus der gleichmäßigen Konvergenz folgt dann

$$\lim_{n \to \infty} h_n(x_j) = h(x_j) \ , \ 1 \leqq j \leqq N \tag{2.9}$$

Nach Korollar 2.3 sind die Folgen $(t_{i,n})_{n \in \mathbb{N}}$, $1 \leqq i \leqq N$, beschränkt und nach dem Beweis von Satz 2.5 gibt es eine Teilfolge $(h_n)_{n \in J}$, $J \subseteq \mathbb{N}$, mit

$$\lim_{n \to \infty} t_{i,n} = t_0 \ , \ n \in J \ , \ 1 \leqq i \leqq N$$

Aus Hilfssatz 1.3 folgt für $1 \leqq j \leqq N$ und $n \in J$:

$$\lim_{n \to \infty} \Delta^i(t_{1,n}, \ldots, t_{i+1,n} e^{tx_j} = \Delta^i(t_0, \ldots, t_0 e^{tx_j} \ , \ i < N \tag{2.10}$$

Für $0 \leqq i \leqq N-1$ und $1 \leqq j \leqq N$ wird definiert:

$$d^n_{j,i+1} := \Delta^i(t_{1,n}, \ldots, t_{i+1,n}) e^{tx_j} \ , \ d_{j,i+1} := \Delta^i(t_0, \ldots, t_0) e^{tx_j}$$

$$D_n := \begin{pmatrix} d^n_{1,1} & \ldots & d^n_{1,N} \\ \vdots & \vdots & \vdots \\ d^n_{N,1} & \ldots & d^n_{N,N} \end{pmatrix} \quad D := \begin{pmatrix} d_{1,1} & \ldots & d_{1,N} \\ \vdots & \vdots & \vdots \\ d_{N,1} & \ldots & d_{N,N} \end{pmatrix}$$

$$B_n := \begin{pmatrix} b_{0,n} \\ \vdots \\ b_{N-1,n} \end{pmatrix} \ B := \begin{pmatrix} b_0 \\ \vdots \\ b_{N-1} \end{pmatrix} \ H_n := \begin{pmatrix} h_n(x_1) \\ \vdots \\ h_n(x_N) \end{pmatrix} \ H := \begin{pmatrix} h(x_1) \\ \vdots \\ h(x_N) \end{pmatrix}$$

Es gilt also:

$$D_n B_n = H_n \ , \ DB = H \tag{2.11}$$

Die (N-N)-Matrix $C_{i,n}$ bzw. C_i erhält man aus D_n bzw. D, indem in D_n bzw. D die i-te Spalte durch H_n bzw. H ersetzt wird, $1 \leqq i \leqq N$.

Die Gleichungssysteme (2.11) sind eindeutig lösbar, da der von

$$\begin{aligned} \{\Delta^i(t_0, \ldots, t_0 e^{tx} \quad &| \quad 0 \leqq i \leqq N-1\} \text{ bzw.} \\ \{\Delta^i(t_{1,n}, \ldots, t_{i+1,n} e^{tx} \quad &| \quad 0 \leqq i \leqq N-1\} \end{aligned}$$

erzeugte N-dimensionale reelle Vektorraum die Haarsche Bedingung erfüllt (Satz 1.1 und Hilfssatz 1.2). Damit folgt:

$$\det(D_n) \neq 0 \neq \det(D) \ , \ n \in \mathbb{N}$$

Nach der Cramerschen Regel gilt

$$b_{i,n} = \frac{\det(C_{i+1,n})}{\det(D_n)} \ , \ n \in \mathbb{N}, \text{ und } b_i = \frac{\det(C_{i+1})}{\det(D)} \text{ für } n \in \mathbb{N}$$

Für (n-n)-Matrizen mit reellen Koeffizienten ist die Abbildung $A \to \det(A)$ eine stetige Funktion der n^2 Koeffizienten von A; aus (2.9) und (2.10) erhält man daher für $0 \leqq i \leqq N$

$$\lim_{n \to \infty} \frac{\det(C_{i+1,n})}{\det(D_n)} = \frac{\det(C_{i+1})}{\det(D)} \ , \ n \in J$$

Somit ist (2.8) gezeigt.

Durch vollständige Induktion nach N ergibt sich unter Verwendung von (1.7) für $p_i, s_i \in \mathbb{R}$, $i \leqq i \leqq N$, mit $s_i \neq s_j$ für $i \neq j$:

$$\sum_{i=1}^{N} p_{i-1} \Delta^{i-1}(s_1, \ldots, s_i) e^{tx} =$$

(2.12)

$$= \sum_{i=1}^{N} e^{s_i x} \prod_{m=1, m \neq i}^{N} (s_i - s_m)^{-1} [p_{N-1} + \sum_{k=i}^{N-1} p_{k-1} \prod_{m=k+1}^{N} (s_i - s_m)]$$

$N = 1$:

$$\sum_{i=1}^{1} p_{i-1} \Delta^{i-1}(s_1, \ldots, s_i) e^{tx} = p_0 s^{s_1 x} =$$

$$= \sum_{i=1}^{1} e^{s_i x} \prod_{m=1, m \neq i}^{1} (s_i - s_m)^{-1} [p_{N-1} + \sum_{k=i}^{0} p_{k-1} \prod_{m=k+1}^{1} (s_i - s_m)]$$

Es gelte (2.12) für $N \in \mathbb{N}$.

Induktionsschluß:

$$\begin{aligned}
&\sum_{i=1}^{N} p_{i-1}\Delta^{i-1}(s_1,\ldots,s_i)e^{tx} + p_N\Delta^N(s_1,\ldots,s_{N+1})e^{tx} = \\
&= \sum_{i=1}^{N} e^{s_i x} \prod_{m=1,m\neq i}^{N} (s_i - s_m)^{-1}[p_{N-1} + \sum_{k=i}^{N-1} p_{k-1} \prod_{m=k+1}^{N} (s_i - s_m)] + \\
&\quad + p_N \sum_{i=1}^{N+1} e^{s_i x} \prod_{m=1,m\neq i}^{N+1} (s_i - s_m)^{-1} = \\
&= \sum_{i=1}^{N} e^{s_i x} \prod_{m=1,m\neq i}^{N+1} (s_i - s_m)^{-1} \sum_{k=i}^{N} p_{k-1} \prod_{m=k+1}^{N+1} (s_i - s_m) + \\
&\quad + \sum_{i=1}^{N+1} e^{s_i x} p_N \prod_{m=1,m\neq i}^{N+1} (s_i - s_m)^{-1} = \\
&= \sum_{i=1}^{N+1} e^{s_i x} \prod_{m=1,m\neq i}^{N+1} (s_i - s_m)^{-1}[p_N + \sum_{k=i}^{N} p_{k-1} \prod_{m=k+1}^{N+1} (s_i - s_m)]
\end{aligned}$$

Damit ist (2.12) vollständig bewiesen.

Es gilt also

$$h_n(x) = \sum_{i=1}^{N} e^{t_{i,n}x} \prod_{m=1,m\neq i}^{N} (t_{i,n} - t_{m,n})^{-1}[b_{N-1,n} + \sum_{k=i}^{N-1} b_{k-1,n} \prod_{m=k+1}^{N} (t_{i,n} - t_{m,n})]$$

Wegen $\lim\limits_{n\to\infty} t_{i,n} = t_0$ für $n \in J$, $1 \leqq i \leqq N$, und (2.8) existiert ein $n_0 \in \mathbb{N}$, so daß für $n \geqq n_0$, $n \in J$, erfüllt ist:

$$\begin{aligned}
sign(b_{N-1,n} + \sum_{k=i}^{N-1} b_{k-1,n} \prod_{m=k+1}^{N} (t_{i,n} - t_{m,n})\ ) &= sign(b_{N-1}) \\
&= sign(a_{N-1}) \text{ für } 1 \leqq i \leqq N;
\end{aligned}$$

weiter gilt für $1 \leqq i \leqq N$:

$$sign(\prod_{m=1,m\neq i}^{N} (t_{i,n} - t_{m,n})^{-1}) = (-1)^{N-i}\ ,\ n \in \mathbb{N}$$

Wegen

$$a_{i,n} = \sum_{m=1,m\neq i}^{N} (t_{i,n} - t_{m,n})^{-1}[b_{N-1,n} + \sum_{k=i}^{N-1} b_{k-1,n} \prod_{m=k+1}^{N} (t_{i,n} - t_{m,n})]$$

erhält man $sign(a_{i,n}) = (-1)^{N-i} sign(a_{N-1})$ für $n \geqq n_0$, $n \in J$;
die Behauptung gilt also für die Teilfolge $(h_n)_{n\in J}$. Nimmt man an, daß die Behauptung nicht für die Folge $(h_n)_{n\in\mathbb{N}}$ gilt, dann gibt es eine Teilfolge $(h_n)_{n\in J_1} \subseteq (h_n)_{n\in\mathbb{N}}$ mit $sign(h_n) \neq S$ für $n \in J_1$. Da aber $(h_n)_{n\in J_1}$ ebenfalls gegen h gleichmäßig konvergiert, liefern die soeben durchgeführten Überlegungen die Existenz einer Teilfolge in $(h_n)_{n\in J_1}$, deren Elemente den Vorzeichenvektor S besitzen, was im Widerspruch zu $sign(h_n) \neq S$ steht.
Damit ist die Behauptung gezeigt.

Satz 2.7 (Braess, [2]; Werner, [21])
Es sei $N \in \mathbb{N}$ und S ein Vorzeichenvektor mit N Komponenten; die Folge $(h_n)_{n\in\mathbb{N}} \subseteq V_N^0(S)$ konvergiere auf I gleichmäßig gegen $h \in V_N \setminus V_{N-1}$.

Behauptung: $h \in V_N(S)$

Beweis:
Es sei $h_n(x) = \sum_{i=1}^{N} a_{i,n} e^{t_{i,n}x}$, $n \in \mathbb{N}$. Nach Korollar 2.3 sind die Folgen $(t_{i,n})_{n\in N}$, $1 \leqq i \leqq N$, beschränkt und o.B.d.A. gilt $grd(h_n) = N$ für $n \in \mathbb{N}$. Es gibt also eine Teilfolge $(h_n)_{n\in J}$ mit $J \subseteq \mathbb{N}$, so daß erfüllt ist:

$$grd(h_n) = N \text{ und } (t_{i,n})_{n\in J} \text{ ist konvergent für } 1 \leqq i \leqq N$$

Es wird nun die Folge $(h_n)_{n\in J}$ betrachtet, es ist also stets $n \in J$ erfüllt.
Es gebe L verschiedene Grenzwerte der Folgen $(t_{i,n})_{n\in J}$:
$\lim_{n\to\infty} t_{j,n} = t_i$ für $j_i \leqq j < j_{i+1}$ mit $1 \leqq i \leqq L$, wobei $j_1 = 1$, $j_i < j_{i+1}$ und $j_{L+1} - 1 = N$ sei und weiter $t_i < t_{i+1}$ gilt. (Im allgemeinen ist hierzu für endlich viele Elemente der Folge die Indizierung der Summanden, wie sie durch Definition 1.2 gegeben ist, abzuändern.)
Mit $h_{i,n}(x) = \sum_{j=j_i}^{j_{i+1}-1} a_{j,n} e^{t_{j,n}x}$, $1 \leqq i \leqq L$, erhält man $h_n = \sum_{i=1}^{L} h_{i,n}(x)$ und es gibt ein $n_0 \in \mathbb{N}$ mit $S = (sign(h_{1,n}), \ldots, sign(h_{L,n}))$ für $n \geqq n_0$. Der Beweis von Satz 2.5 zeigt, daß $\prod_{i=1}^{L} (D - t_i)^{j_{i+1}-j_i} h = 0$ und damit $h(x) = \sum_{i=1}^{L} P_i(x) e^{t_i x}$ mit $grd(P_i) = j_{i+1} - j_i - 1 := k_i$ erfüllt ist.

Mit $h_i(x) = P_i(x)e^{t_i x}$ gilt $h(x) = \sum_{i=1}^{L} h_i(x)$.

Für $L = 1$, d.h. $\lim_{n\to\infty} t_{i,n} = t_1$ für $1 \leqq i \leqq N$, liefert Hilfssatz 2.3 die Behauptung. Es gelte also $2 \leqq L \leqq N$.

Es wird nun ähnlich wie in Hilfssatz 2.3 die Darstellung durch Differenzenquotienten verwendet:

$$h_{i,n}(x) = \sum_{j=0}^{k_i} b_{i,j}^n \Delta^j(t_{j_i,n}, \ldots, t_{j_i+j,n})e^{tx}$$

(2.13)

$$h_i(x) = \sum_{j=0}^{k_i} b_{i,j} \Delta^j(t_i, \ldots, t_i)e^{tx}$$

Es seien N Punkte x_i, $1 \leqq i \leqq N$, in I gegeben mit $x_1 < x_2 < \ldots, < x_N$; nach Voraussetzung gilt damit für $1 \leqq i \leqq N$

$$\lim_{n\to\infty} h_n(x_i) = h(x_i) \tag{2.14}$$

Hiermit wird definiert:

$$d_{j,m,i}^n := \Delta^{m-1}(t_{j_i,n}, \ldots, t_{j_i+m-1,n})e^{tx_j} \text{ und}$$
$$d_{j,m,i} := \Delta^{m-1}(t_i, \ldots, t_i)e^{tx_j}$$

für $1 \leqq m \leqq k_i + 1$, $1 \leqq i \leqq L$, $1 \leqq j \leqq N$, $n \in J$.

Weiter sei für $1 \leqq i \leqq L$

$$D_{i,n} := \begin{pmatrix} d_{1,1,i}^n & \ldots & d_{1,k_i+1,i}^n \\ \vdots & \vdots & \vdots \\ d_{N,1,i}^n & \ldots & d_{N,k_i+1,i}^n \end{pmatrix}, n \in J \text{ und } D_i := \begin{pmatrix} d_{1,1,i} & \ldots & d_{1,k_i+1,i} \\ \vdots & \vdots & \vdots \\ d_{N,1,i} & \ldots & d_{N,k_i+1,i} \end{pmatrix}$$

Mit den (N-N)-Matrizen $D_n = (D_{1,n}, \ldots, D_{L,n})$, $n \in J$, und $D = (D_1, \ldots, D_L)$ gilt nach (2.13)

$$D_n \begin{pmatrix} B_{1,n} \\ \vdots \\ B_{L,n} \end{pmatrix} = \begin{pmatrix} h_n(x_{1)} \\ \vdots \\ h_n(x_N) \end{pmatrix} := H_n \text{ und } D \begin{pmatrix} B_{1,} \\ \vdots \\ B_{L,} \end{pmatrix} = \begin{pmatrix} h(x_{1)} \\ \vdots \\ h(x_N) \end{pmatrix} := H \ ,$$

wobei für $1 \leqq i \leqq L$

$$B_{i,n} = \begin{pmatrix} b_{i,0}^n \\ \vdots \\ b_{i,k_i}^n \end{pmatrix} \text{ und } B_i = \begin{pmatrix} b_{i,0} \\ \vdots \\ b_{i,k_i} \end{pmatrix} \text{ ist.}$$

Die (N-N)-Matrizen $C_{i,n}$ und C_i, $1 \leqq j \leqq N$, sind definiert wie in Hilfsatz 2.3. Mit (2.14) und Hilfsatz 1.3 folgt wie oben aus der Cramerschen Regel

$$\lim_{n\to\infty} b_{i,j}^n = \lim_{n\to\infty} \frac{\det(C_{j_i+j,n})}{\det(D_n)} = \frac{\det(C_{j_i+j})}{\det(D)} = b_{i,j} \ , \ 0 \leqq j \leqq k_i \ , \ 1 \leqq i \leqq L \ , \ n \in J \ ,$$

da die entsprechenden Gleichungssysteme wieder eindeutig lösbar sind.

Nach Hilfsatz 1.3 konvergieren daher die Folgen $(h_{i,n})_{n\in J}$ auf I gleichmäßig gegen h_i für $1 \leqq i \leqq L$, und Hilfsatz 2.3 ergibt $sign(h_{i,n}) = sign(h_i)$, $1 \leqq i \leqq L$, für $n \geqq n_0$, $n \in J$, woraus $sign(h) = S$ folgt, was zu zeigen war.

Es gilt noch allgemeiner:

Korollar 2.6 (Braess, [2]; Werner, [21])
Es seien die Voraussetzungen von Satz 2.7 mit $(h_n)_{n\in\mathbb{N}} \subseteq V_N(S)$ erfüllt.

Behauptung: $h \in V_N(S)$

Beweis:
Es sei $n \in \mathbb{N}$; für jedes $m \in \mathbb{N}$ existiert ein $g_{m,n} \in V_N^0$ mit $\| h_n - g_{m,n} \|_I \leqq \frac{1}{m}$ und nach Satz 2.7 gibt es ein $m(n) \in \mathbb{N}$ mit $sign(g_{m,n}) = sign(h_n)$ für $m \geqq m(n)$.
Mit $m_n := max\{n, m(n)\}$, $n \in \mathbb{N}$, erhält man so eine Folge $(g_{m_n,n})_{n\in\mathbb{N}}$ in $V_N^0(S)$, die gleichmäßig auf I gegen h konvergiert; Satz 2.7 ergibt damit die Behauptung.

Es wird noch benötigt:

Satz 2.8

$$(h_n)_{n\in\mathbb{N}} \subseteq V_N \text{ mit } h_n(x) = \sum_{i=1}^{L} \sum_{j=0}^{k_i} a_{i,j}^n x^j e^{t_{i,n}x} \text{ und } L \in \mathbb{N}$$

konvergiere auf I gleichmäßig gegen

$$h(x) = \sum_{i=1}^{L} \sum_{j=0}^{k_i} a_{i,j} x^j e^{t_i x}$$

und es gelte

$$\lim_{n\to\infty} t_{i,n} = t_i \in \mathbb{R} \text{ für } 1 \leqq i \leqq L$$

Behauptung:

$$\lim_{n\to\infty} a_{i,n}^n = a_{i,j} \text{ für } 0 \leqq j \leqq k_i, 1 \leqq i \leqq L$$

Beweis:
Mit $k = \sum_{i=1}^{L}(k_i + 1)$ wähle man k Punkte x_i, $1 \leqq i \leqq k$, mit $x_i < x_{i+1}$ in I; es gilt

$$\sum_{i=1}^{L}\sum_{j=0}^{k_i} a_{i,j}^n x_m^j e^{t_{i,n}x_m} = h_n(x_m)\ , \quad 1 \leqq m \leqq k\ , n \in \mathbb{N}$$

(2.15)

$$\sum_{i=1}^{L}\sum_{j=0}^{k_i} a_{i,j} x_m^j e^{t_i x_m} = h(x_m)\ , \qquad 1 \leqq m \leqq k$$

Betrachtet man wieder die Koeffizienten $a_{i,j}^n$ bzw. $a_{i,j}$ als die eindeutig bestimmten Lösungen der Gleichungssysteme (2.15), dann erhält man wegen $\lim_{n\to\infty} h_n(x_j) = h(x_j)$, $1 \leqq j \leqq k$, und $\lim_{n\to\infty} t_{i,n} = t_i$, $1 \leqq i \leqq L$, aus der Cramerschen Regel die Behauptung des Satzes.

Einen Zusammenhang zwischen der gleichmäßigen Konvergenz auf I und der gleichmäßigen Konvergenz im Innern von I gibt der folgende Satz für Exponentialsummen an:

Satz 2.9 (Werner, [21])
Es sei $N \in \mathbb{N}$; konvergiert die Folge $(h_n)_{n\in\mathbb{N}} \subseteq V_N$ gleichmäßig im Innern von I gegen $h \in V_N \setminus V_{N-1}$, dann konvergiert $(h_n)_{n\in\mathbb{N}}$ auf I gleichmäßig gegen h.

Beweis:
Für jedes abgeschlossene Intervall $I_1 = [a_1, b_1] \subseteq (a, b)$ sei $I_1' = [\frac{a_1 + a}{2}, \frac{b_1 + b}{2}]$; für $j \in \mathbb{N}$ gibt es nach Satz 2.2 ein $K(j, I_1', I_1) < \infty$ mit $\| D^j(h_n - h) \|_{I_1} \leqq K(j, I_1', I_1) \| h - h_n \|_{I_1'}$. Aus der gleichmäßigen Konvergenz im Innern von I folgt daher für $j \in \mathbb{N}$:

$$\lim_{n\to\infty} \| D^j h_n - D^j h \|_{I_1} = 0$$

Für $x_0 \in (a, b)$ gilt daher

$$\lim_{n\to\infty} D^j h_n(x_0) = D^j h(x_0) \qquad j \in \mathbb{N} \tag{2.16}$$

Wegen $h \in V_N \setminus V_{N-1}$ erhält man mit dem Beweis von Satz 2.5, daß die Frequenzen der Elemente von $(h_n)_{n\in\mathbb{N}}$ beschränkt sind: Es gilt also

$$\prod_{i=1}^{N}(D - t_{i,n})h_n = 0 \text{ mit } \mid t_{i,n} \mid < T < \infty \text{ für} 1 \leqq i \leqq N \text{ und } n \in \mathbb{N}$$

und es gibt daher eine Teilfolge mit $\lim_{n\to\infty} t_{i,n} = t_i \in \mathbb{R}$, $n \in J \subseteq \mathbb{N}$.

Mit (2.16) ergibt der Satz von der stetigen Abhängigkeit der Lösungen eines Anfangswertproblems gewöhnlicher Differentialgleichungen die gleichmäßige Konvergenz von $(h_n)_{n\in J}$ auf I; da somit in jeder Teilfolge von $(h_n)_{n\in\mathbb{N}}$ eine auf I gleichmäßig gegen h konvergente Teilfolge existiert, gilt die Behauptung.

Bemerkung 2.2
Entscheidend für den Beweis von Satz 2.9 ist die Beschränktheit der Folgen $(t_{i,n})_{n\in\mathbb{N}}$; man beachte hierzu Beispiel 2.1. Setzt man dies voraus, dann wird die Voraussetzung $h \in V_N \setminus V_{N-1}$ nicht benötigt.

3 Eindeutigkeitssätze und Charakterisierung der Minimallösungen

Nach dem Existenzproblem sollen nun folgende Fragen behandelt werden:

1. Sind die Minimallösungen bezüglich V_N, V_N^0, V_N^+ eindeutig bestimmt?
2. Wodurch sind die Minimallösungen gekennzeichnet?
3. Unter welchen Voraussetzungen ist eine Minimallösung bezüglich V_N^+ oder V_N^0 auch Minimallösung bezüglich V_N?

Da die Existenz einer Minimallösung vorausgestzt wird, können die entsprechenden Sätze, allgemeiner als in [1] für die Approximation über kompakten, reellenTeilmengen angegeben werden.

Um, wie in Kapitel 1 angekündigt, die Ergebnisse von [11] anwenden zu können, werden folgende Begriffe benötigt:

Definition 3.1
Mit $n \in \mathbb{N}$ sei A eine offene Teilmenge des n-dimensionalen Raumes $\mathbb{R}^n$ und V sei eine Menge von Funktionen $v(a) = v(a, x)$ aus C(I) auf dem reellen Intervall I, die von $a \in A$ abhängen; ferner existiere für jedes $a = (a_1, \ldots, a_n) \in A$ und $x \in I$ die partielle Ableitung

$$D_i v(a, x) = \frac{\partial v(a, x)}{\partial a_i}, 1 \leqq i \leqq n$$

Diese Ableitungen werden als Funktionen auf I mit $D_i v(a), a \in A$, bezeichnet und es sei $D_i v(a) \in C(I)$ für $a \in A$ und $1 \leqq i \leqq n$ erfüllt.

a. Der *Gradient* von $v(a) \in V$ ist gegeben durch

$$grad(v(a)) := (D_1 v(a), \ldots D_n v(a)).$$

Mit $r = (r_1, \ldots, r_n) \in \mathbb{R}^n$ sei $(r, grad(v(a))) \in C(I)$ die Funktion

$$(r, grad(v(a))) := \sum_{i=1}^{n} r_i D_i v(a)$$

Für $x \in I$ gilt also $grad(v(a, x)) = (D_1 v(a, x), \ldots D_n v(a, x))$ und

$$(r, grad(v(a, x))) = \sum_{i=1}^{n} r_i D_i v(a, x)$$

b. Der *Gradientenraum* $W(a)$ von $v(a) \in V$ ist der von den Funktionen $D_i v(a), 1 \leqq i \leqq n$, erzeugte lineare Raum:

$$W(a) = \{(r, grad(v(a))) \mid r \in \mathbb{R}^n\}$$

c. V erfüllt die *lokale Haarsche Bedingung*, wenn für alle $a \in A$ der lineare Raum $W(a)$ der Haarschen Bedingung genügt, also:
Für $a \in A$ besitzt jedes Element von $W(a)$ in I höchstens $m-1$ Nullstellen, wenn m die Dimension von $W(a)$ ist, oder verschwindet auf I identisch.

Bemerkung 3.1

a. Für die Bildung des Gradienten nach Definition 3.1a ist es also notwendig, daß für die Elemente des betrachteten Funktionensystems eine Parametrisierung wie oben angegeben ist; wie in Satz 3.1 erfordert dies für $V = V_N$ eine Beschränkung auf geeignete Teilmengen.

b. Es sei V das Funktionensystem von Definition 3.1; ist weiterhin die Abbildung $a \longrightarrow D_i v(a) \in C(I)$, $1 \leqq i \leqq n$, stetig für alle $a \in A$, gibt es also für jedes $\epsilon > 0$ und $a \in A$ ein $\delta = \delta(a, \epsilon) > 0$, so daß $\| D_i v(a) - D_i v(b) \|_I < \epsilon$ für $b \in A$ mit $\| a - b \| < \delta$ erfüllt ist, dann besitzt $v(a)$ für jedes $a \in A$ eine Fréchet-Ableitung (nach dem Parameter a); dabei ist $\| b \|$ die Euklidische Norm von $b \in \mathbb{R}^n$:
Wie in der Analysis gezeigt wird (z.B. H. Bauer, Differential- und Integralrechnung II, S. 97), folgt aus diesen Voraussetzungen für $a, b \in \mathbb{R}^n$, $a = (a_1, \ldots, a_n)$, $b = (b_1, \ldots, b_n)$:

$$\frac{\| v(a+b) - v(a) - (b, grad(v(a))) \|_I}{\| b \|} \leqq \sum_{i=1}^{n} \| D_i v(c_i(b)) - D_i v(a)) \|_I$$

für $\| b \|$ hinreichend klein; dabei ist

$$c_i(b) = (a_1 + b_1, \ldots, a_{i-1} + b_{i-1}, c_i, a_{i+1}, \ldots, a_n) \quad \text{mit } c_i \in (a_i, a_i + b_i), \; 1 \leqq i \leqq n$$

Aus der Stetigkeit in den Parametervektoren folgt

$$\lim_{\|b\| \to 0} \frac{\| v(a+b) - v(a) - (b, grad(v(a))) \|_I}{\| b \|} = 0$$

Die Abbildung $b \longrightarrow (b, grad(v(a))) \in C(I)$ für $b \in \mathbb{R}^n$ ist linear und beschränkt und stellt somit die (eindeutig bestimmte) Fréchet-Ableitung von v(a) nach dem Parametervektor a dar.

Definition 3.2
Gegeben sei $F \in C(X), X \subseteq \mathbb{R}$.

a. $x \in X$ ist ein *Extremalpunkt* von F in X, falls gilt:

$$| F(x) | = \| F \|_X$$

b. F besitzt in X eine *Alternante* der Länge n mit $n \in \mathbb{N}$, falls F n Extremalpunkte $x_i, 1 \leqq i \leqq n$, in X besitzt und mit $n > 1$ gilt:

$$x_i < x_{i+1} \text{ und } sign(F(x_i)) = -sign(F(x_{i+1})) \qquad 1 \leqq i \leqq n-1$$

Die Punkte $x_i, 1 \leqq i \leqq n$, werden als *Alternantenpunkte* von F bezeichnet.

c. F besitzt in X eine *positive (negative) Alternante der Länge n*, wenn F in X eine Alternante der Länge n besitzt, so daß im Alternantenpunkt x_n gilt:

$$F(x_n) > 0 \qquad (F(x_n) \leqq 0)$$

Vereinbarung:
Ab jetzt sei $N \in \mathbb{N}$ und X stets eine kompakte, reelle Teilmenge mit der Mächtigkeit $|X| > 2N$; I sei das Intervall $[p, q]$ mit $p = \min X$ und $q = \max X$.
Weiter sei $f \in C(I)$.

Zunächst werden die Sätze 11 und 12 aus [11] für kompakte Mengen formuliert:

Hilfsatz 3.1 ([11], Satz 11)
Es sei $V \subseteq C(I)$; für $v \in V$ gebe es ein $N(v) \in \mathbb{N}$, so daß $u - v$ für jedes $u \in V$ höchstens $N(v) - 1$ Nullstellen in I besitzt oder auf I identisch verschwindet. In X gebe es $N(v) + 1$ Punkte x_i mit $1 \leqq i \leqq N(v) + 1$ und $x_1 < x_1 \ldots < x_{N(v)+1}$, so daß

$$sign(f(x_i) - v(x_i)) = -sign(f(x_{i+1}) - v(x_{i+1})) \neq 0$$

für $1 \leqq i \leqq N(v)$ erfüllt ist; $D := \{x_i \mid 1 \leqq i \leqq N(v) + 1\}$.
Behauptung:

$$\inf_{u \in V} \| f - u \|_X \geqq \min_{x \in D} | f(x) - v(x) |$$

Beweis:
Mit

$$S(u_1, u_2, x) := (f(x) - u_1(x))(u_2(x) - u_1(x)) \text{ für } u_1, u_1 \in V$$

gibt es für alle $u \in V$ mindestens ein $x_u \in D$, so daß $S(v, u, x_u) \leqq 0$ erfüllt ist; denn gilt für ein $u \in V$ und alle $x \in D$

$$sign(f(x) - v(x)) = sign(u(x) - v(x)) \neq 0$$

dann folgt im Widerspruch zu Voraussetzung, daß $u - v$ in $J := (x_1, x_{N(v)+1})$ mindestens $N(v)$ Nullstellen besitzt, aber auf I nicht identisch verschwindet. Es ist also $\min_{x \in D} S(v, u, x) \leqq 0$ für all $u \in V$ erfüllt und Satz 1 in [11] ergibt die Behauptung.

Bemerkung:
Für den Beweis von Hilfssatz 3.1 wird eine Parametrisierung von V, wie sie in Definition 3.1 angenommen war, nicht benötigt; deshalb ist dieser Hilfssatz auf V_N mit $N(v) \leqq 2N$ für $v \in V_N$ anzuwenden und ergibt eine **untere Schranke für die Minimalabweichung** von f bezüglich V_N.

Bemerkung 3.2
Zum Beweis von Hilfsatz 3.2 wird benötigt:
Der lineare Raum $U \subseteq C(I)$ mit der Dimension $n + 1$ erfülle die Haarsche Bedingung. Es gilt:

1. Jedes Element $u \in U$, das auf I nicht identisch verschwindet, besitzt in I höchstens n Nullstellen, wobei die Nullstellen im Innern von I, an denen das Vorzeichen nicht wechselt, doppelt gezählt werden. ([8], Theorem 4.2)

2. Es sei $T = \{t_i \mid 1 \leqq i \leqq k\} \subseteq I$, $k \in \mathbb{N}$, eine Menge paarweise verschiedener Punkte von $I = [p, q]$ mit $\sum_{i=1}^{k} w(t_i) \leqq n$, wobei gelte:

   $$w(t) = \begin{cases} 1 & : t \in \{p, q\} \\ 2 & : t \in (p, q) \end{cases}$$

   Dann gibt es ein nichtnegatives Element $u^+ \in U$, das auf I nicht identisch verschwindet, so daß u^+ genau die k Nullstellen t_i, $1 \leqq i \leqq k$, besitzt; die einzige Aussnahme besteht darin, daß, falls genau ein Element von {p,q } in T enthalten und n gerade ist, u^+ auch im anderen Randpunkt von I verschwinden kann. ([8], Theorem 4.1)

3. Mit diesen beiden Sätzen folgt die Existenz eines Elements $v \in U$ mit $v(x) > 0$ für $x \in I$.

Hilfsatz 3.2 ([11], Satz 12)
Es seien V und A gegeben wie in Definition 3.1 und Bemerkung 3.1b. V erfülle die lokale Haarsche Bedingung und $v(a) \in V$ sei eine Minimallösung für f bezüglich V; es sei hier $\mid X \mid > n$.
Dann gibt es in X für $f - v(a)$ eine Alternante der Länge $d+1$, wobei der Gradientenraum W von $v(a)$ die Dimension d besitzt.

Beweis:
Nach Voraussetzung ist $v(b) \in V$ Fréchet-differenzierbar in $b \in A$. Es sei o.B.d.A. $\| f - v(a) \|_X \neq 0$.
Für $d = 0$ ist nichts zu zeigen; es sei also $d \geqq 1$.

Nach Satz 9 in [11] besitzt $F := f - v(a)$ in X mindestens $d+1$ Extremalpunkte; D sei die Menge der Extremalpunkte von F in X.

Nimmt man an, daß $F(x) = \| F \|_X$ ($F(x) = - \| F \|_X$) für alle $x \in D$ erfüllt ist , dann gilt mit der besten Approximation w an F bezüglich W auf D (W erfüllt die Haarsche Bedingung nach Voraussetzung):

$$\| F \|_X = \| F \|_D > \| F - w \|_D$$

Hieraus folgt $w(x) > 0$ ($w(x) < 0$) für $x \in D$ und damit

$$sign(w(x)) = sign(F(x)) \neq 0 \qquad x \in D$$

Dies steht im Widerspruch zur Minimalität von v(a) nach Satz 8 in [11]. Dieser Widerspruch ergibt sich auch aus der Existenz eines Elementes $u \in W$ mit $u(x) > 0$ ($u(x) < 0$) für $x \in I$ nach Bemerkung 3.2. Es folgt also, daß es in D mindestens zwei Punkte x_1, x_2 gibt mit $x_1 < x_2$ und $sign(F(x_1) = -sign(F(x_2) \neq 0$.
Die Behauptung ist damit für $d = 1$ gezeigt; es sei nun $d \geqq 2$:

D ist abgeschlossen und es sei $J = [d_1, d_2]$ mit $d_1 = \min D$ und $d_2 = \max D$.

Annahme: Es gibt keine Alternante der Länge $d+1$ für F in X.
F ist stetig in I und man kann daher J in $m+1 \geqq 2$ Teilintervalle $[y_i, y_{i+1}]$, $0 \leqq i \leqq m$, mit $m < d$ so aufteilen, daß gilt:

1. Auf $D \cap [y_i, y_{i+1}]$ besitzt F gleiches Vorzeichen für $0 \leqq i \leqq m$
2. Für $x \in D \cap [y_{i-1}, y_i]$ und $y \in D \cap [y_i, y_{i+1}]$ gilt für $1 \leqq i \leqq m$:
$$sign(F(y)) = -sign(F(x))$$
3. $F(y_i) \neq 0$ und $y_i \notin D$ für $1 \leqq i \leqq m$

In $[y_{m-1}, d_2]$ besitzt F mindestens eine Nullstelle z: $F(z) = 0$.

Fall 1: $m = d - 2r - 1$, $r \in \mathbb{N} \cup \{0\}$.
In J wähle man 2r Punkte p_i, $1 \leqq i \leqq 2r$, mit $z < p_1 < p_2 \ldots < p_{2r}$ und

$$p_{2r} - z < min\{\mid z - x \mid \mid x \in D \cup \{y_m\}\}$$

Da W mit der Dimension d die Haarsche Bedingung erfüllt, folgt zunächst, daß es ein $b \in \mathbb{R}^d$ gibt mit

$$(3.1) \qquad (b, grad(v(a, d_1))) = sign(F(d_1))$$

$$(3.2) \qquad \begin{aligned} (b, grad(v(a, y_i))) &= 0 \qquad & 1 \leqq i \leqq m \\ (b, grad(v(a, p_i))) &= 0 \qquad & 1 \leqq i \leqq 2r \end{aligned}$$

Nach Bemerkung 3.2 besitzt $(b, grad(v(a)))$ in I genau die durch (3.2) gegebenen $d - 1$ Nullstellen und wechselt in diesen das Vorzeichen, weshalb mit (3.1) folgt:

$$(3.3) \qquad sign(b, grad(v(a, x))) = sign(F(x)) \quad x \in D$$

Fall 2: $m = d - 2r - 2$, $r \in \mathbb{N} \cup \{0\}$.
Wie oben gibt es ein $b \in \mathbb{R}^d$, so daß mit den Punkten p_i von oben gilt:

$$(3.4) \qquad (b, grad(v(a, d_1))) = sign(F(d_1))$$

$$(3.5) \qquad \begin{aligned} (b, grad(v(a, y_i))) &= 0 \qquad & 1 \leqq i \leqq m \\ (b, grad(v(a, p_i))) &= 0 \qquad & 1 \leqq i \leqq 2r \end{aligned}$$

$$(3.6) \qquad (b, grad(v(a, d_2))) = sign(F(d_2))$$

Nach Bemerkung 3.2 besitzt $(b, grad(v(a)))$ wegen (3.4) und (3.6) in J genau die durch (3.5) gegebenen $d-2$ Nullstellen und wechselt dort sein Vorzeichen. Damit gilt auch hier (3.3).

In beiden Fällen liegt also ein Widerspruch zur Minimalität wie oben vor und die Behauptung ist bewiesen.

Wir sind nun in der Lage, eine Charakterisierung für Minimallösungen bezüglich V_N anzugeben.

Satz 3.1 (Braess, [1]. **Minimallösung bezüglich $\mathbf{V_N}$**)
Es sei

$$E(x) = \sum_{i=1}^{L} P_i(x)e^{t_i x} \in V_N \text{ mit grd(E)=K und der Länge L}$$

gegeben.

Behauptung:

a. Besitzt $f-E$ in X eine Alternante der Länge $N+K+1$, dann ist E die eindeutig bestimmte Minimallösung für f bezüglich V_N auf X.

b. Ist E Minimallösung für f bezüglich V_N auf X, dann gibt es für $f-E$ eine Alternante der Länge $N+L+1$ in X.

Beweis:
Für jedes $h \in V_N$ gilt $grd(h-E) \leqq N+K$ und $h-E \not\equiv 0$ besitzt nach Satz 1.1 höchstens $N+K-1$ reelle Nullstellen. Hat also $f-E$ in X eine Alternante der Länge $N+K+1$, dann folgt mit Hilfssatz 3.1

$$\inf_{h \in V_N} \| f-h \|_X \geqq \| f-E \|_X$$

E ist also eine Minimallösung für f auf X bezüglich V_N.
Annahme: $E_1 \in V_N$ mit $E_1 \not\equiv E$ ist eine weitere Minimallösung:

$$\| f-E_1 \|_X = \| f-E \|_X \tag{3.7}$$

Der Vorzeichenvektor S von $E_1 - E$ hat maximal $N+K$ Komponenten und damit höchstens $N+K-1$ Vorzeichenwechsel. Andererseits hat $(f-E)-(f-E_1) = E_1 - E \not\equiv 0$ in $[x_1, x_{N+K+1}]$ wegen (3.7) mindestens $N+K$ Nullstellen, wobei x_1 bzw. x_{N+K+1} der kleinste bzw. größte Alternantenpunkt einer Alternante von $f-E$ in X ist. Nach Satz 1.3 stellt dies einen Widerspruch dar; damit ist die erste Behauptung gezeigt.

Es gilt

$$E(x) = \sum_{i=1}^{L} P_i(x)e^{t_i x} \; + \; \sum_{i=K+1}^{N} a_i e^{t_i x}$$

mit $a_i = 0$, $K+1 \leqq i \leqq N$, und $t_i < t_{i+1}$, $K+1 \leqq i \leqq N$, $t_i \in \mathbb{R}$ für $K+1 \leqq i \leqq N$.

Ist E Minimallösung für f bezüglich V_N auf X, dann ist E auch Minimallösung für bezüglich f der Menge

$$\begin{aligned} V = \quad \{ \quad & E(b) \in V_N \mid E(b,x) = \sum_{i=1}^{L} Q_i(x) e^{s_i x} \; + \sum_{i=K+1}^{N} b_i e^{s_i x} \; , \\ & Q_i \text{ ist reelles Polynom mit } grd(Q_i \leqq m_i \; , \\ & s_i \in \mathbb{R} \; , \; 1 \leqq i \leqq L \; , \\ & s_i, b_i \in \mathbb{R} \text{ für } K+1 \leqq i \leqq N \text{ und } s_i < s_j \text{ für } i < j \\ \} \; , & \end{aligned}$$

wobei $m_i = grd(P_i)$ für $1 \leqq i \leqq L$ ist.

Jedem Element $E(b) \in V$ ist ein Parametervektor $b \in \mathbb{R}^{2N-K+L}$ zugeordnet:

$$\begin{aligned} b = (\quad & b_{1,0}, \quad \ldots, b_{1,m_1}, b_{2,0}, \ldots \; \ldots, b_{L,0}, \ldots, b_{L,m_L}, b_{k+1}, \ldots, b_N, \\ & s_1, \quad \ldots, s_L, s_{K+1}, \ldots, s_N) \qquad \longleftrightarrow \\ \longleftrightarrow \quad & E(b,x) = \sum_{i=1}^{L} (\sum_{j=0}^{m_i} b_{i,j} x^j) e^{s_i x} \; + \sum_{i=K+1}^{N} b_i e^{s_i x} \in V \end{aligned}$$

Damit genügt V den Voraussetzungen von Hilfssatz 3.2. Der Gradientenraum von E als Element von V hat die Dimension $N+L$ und es gibt daher für $f-E$ in X eine Alternante der Länge $N+L+1$; dies ist die zweite Behauptung des Satzes.

Bemerkung: Minimallösung bezüglich $\mathbf{V_N^0}$
In V_N stimmen also im allgemeinen die notwendigende und hinreichende Bedingung für eine Minimallösung nicht überein; da für $E \in V_N^0$ die Länge mit dem Grad zusammenfällt, ist in diesem Fall die Existenz einer Alternante der Länge $N+K+1 = N+L+1$ von $f–E$ in X notwendig *und* hinreichend dafür, daß E Minimallösung für f auf X bezüglich V_N ist.

Zu folgendem Korollar vergleiche man Satz 16 in [11].

Korollar 3.1 (**Charakterisierung & Eindeutigkeit bezüglich $\mathbf{V_N^0}$**)
Es sei

$$E(a,x) = \sum_{i=1}^{L} a_i e^{t_i x} \in V_N^0$$

mit der Länge L gegeben.
Behauptung:

- **a.** Ist E(a) auf X Minimallösung für f bezüglich V_N^0, dann ist E(a) die eindeutig bestimmte Minimallösung für f bezüglich V_N und V_N^0
- **b.** E(a) ist genau dann beste Approximation auf X an f bezüglich V_N^0, wenn f–$E(a)$ in X eine Alternante der Länge N+L+1 besitzt.

Beweis:
V_N^0 erfüllt die Voraussetzungen von Hilfssatz 3.2 und der Gradientenraum von E(a) besitzt die Dimension $N + L$.
Ist daher $E(a)$ eine Minimallösung für f auf X bezüglich V_N^0, dann besitzt $f - E(a)$ in X eine Alternante der Länge $N + L + 1$ und Satz 3.1a ergibt den ersten Teil der Behauptung.
Liegt umgekehrt eine solche Alternante vor, so ist $E(a)$ Minimallösung bezüglich V_N, also insbesondere auch bezüglich V_N^0. Damit ist auch der zweite Teil der Behauptung gezeigt.

Satz 3.2 (Braess, [1])
$E(a) \in V_N^+$ ($E(a) \in V_N^-$) mit der Länge L sei eine beste Approximation an f auf X bezüglich V_N^+ (V_N^-).

Behauptung:

- **a.** E(a) ist auch die beste Approximation an f bezüglich V_L^0 und V_L auf X.
- **b.** E(a) ist die eindeutig bestimmte Minimallösung für f bezüglich V_N^+ (V_N^-) auf X.

Beweis:
Zu a.: Für $L = 0$ ist nichts zu zeigen; es sei also $L \geqq 1$.
Nach Voraussetzung ist $E(a)$ eine beste Approximation an f bezüglich $V := V_L^+ \backslash V_{L-1}$, $(V_L^- \backslash V_{L-1})$. V erfüllt die Voraussetzungen von Hilfssatz 3.2 und der Gradientenraum von $E(a)$ als Element von V hat die Dimension $2L$. Daher besitzt $f - E(a)$ eine Alternante der Länge $2L + 1$ in X. Aus Korollar 3.1 folgt hiermit der erste Teil der Behauptung.

Zu b.: $E(a)$ ist also insbesondere die eindeutig bestimmte Minimallösung bezüglich V_L^+ (V_L^-); $E(b) \in V_N^+$ ($E(b) \in V_N^-$) mit der Länge L_1 sei eine

weitere Minimallösung:

$$\| f - E(b) \|_X = \| f - E(a) \|_X \tag{3.8}$$

Damit ist $E(b)$ auch Minimallösung für f bezüglich $V_{L_1}^0$. Aus der Eindeutigkeitsaussage von Korollar 3.1 folgt wegen (3.8)

$$L = L_1 \ , \ E(a) = E(b)$$

Damit ist auch der zweite Teil der Behauptung gezeigt.

Bemerkung:

1. Satz 3.2 besagt für die Approximation durch Exponentialsummen, daß ein Algorithmus, der eine beste Approximation an f bezüglich V_N ermittelt, zugleich für die Approximation bezüglich V_N^+ und V_N^0 zu verwenden ist. Hat man umgekehrt ein Verfahren zur Berechnung der besten Approximation bezüglich V_N^+, so erhält man damit auch die beste Approximation bezüglich V_L, wobei L nach Satz 3.2 bestimmt ist.

2. Im Beweis von Satz 3.2 wird $V_L^+ \setminus V_{L-1}$ betrachtet, da die für V_L^+ definierte Parametermenge in $\mathbb{R}^{2L}$ nicht offen ist; Hilfssatz 3.2 ist damit nicht auf V_L^+ anzuwenden. Diese Tatsache wirkt sich auch bei der Charakterisierung der Minimallösung bezüglich V_N^+ in Satz 3.4 aus.

Nach Korollar 3.1 und Satz 3.2 sind also die Minimallösungen für f bezüglich V_N^0 und V_N^+, V_N^- eindeutig bestimmt; es soll nun gezeigt werden, daß dies in V_N allgemein nicht gilt[4].

Hilfsatz 3.3 Es sei J=[-y,y] und $V \subseteq C(J)$; ferner sei $v \in V$ die eindeutig bestimmte beste Approximation an $f \in C(J)$ bezüglich V auf J und es sei die Funktion w mit $w(x) = v(-x)$ für $x \in J$ in V enthalten. Für f gelte $f(x) = f(-x)$, $x \in J$.

Behauptung: Für $x \in J$ gilt $v(x) = v(-x)$
Beweis:
Mit $x \in J \Rightarrow -x \in J$ erhält man

$$\begin{aligned} \| f - v \|_J &= \max_{x \in J} | f(x) - v(x) | = \max_{x \in J} | f(-x) - v(-x) | = \\ &= \max_{x \in J} | f(x) - w(x) | = \| f - w \|_J \end{aligned}$$

Aus der Eindeutigkeit der Minimallösung folgt

$$v(x) = w(x) = v(-v) \ , \ x \in J$$

und damit die Behauptung.

[4] s. Satz 3.3

Satz 3.3 (**Nichteindeutigkeit**; Braess, [1])
Es sei J=[-1,1]; ist $f \in C(J)$ streng monoton fallend auf [0,1] und gilt $f(x) = f(-x) > 0$ für $x \in J$, dann gibt es für f auf J mindestens zwei Minimallösungen bzgl V_2.

Beweis:
Die Existenz einer besten Approximation $E \in V_2$ für f auf J folgt aus Satz 2.6.
Annahme:
E ist die eindeutig bestimmte Minimallösung für f bezüglich V_2 auf J.

Nach Hilfssatz 3.3 folgt

$$E(x) = E(-x),\ x \in J \tag{3.9}$$

Es gilt $E \not\equiv 0$ nach Satz 3.1 und Voraussetzung.
Fall 1: $E(x) = (a + bx)e^{tx}$
Nach (3.9) folgt also $(a + bx)e^{tx} - (a - bx)e^{-tx} = 0$ für $x \in J$. Mit Satz 1.1 erhält man daraus $E(x) = a \in \mathbb{R}$.

Fall 2: $E(x) = a_1 e^{t_1 x} + a_2 e^{t_2 x}$
Aus (3.9) folgt für alle $x \in J$

$$a_1(e^{t_1 x} - e^{-t_1 x}) + a_2(e^{t_2 x} - e^{-t_2 x}) = 0$$

Beachtet man weiter $E \not\equiv 0$ und $t_2 \neq t_1$, so ergibt sich $a_1 = a_2 = a$ und $-t_2 = t = t_1$ und damit

$$E(x) = a(e^{tx} + e^{-tx}) \text{ oder } E(x) = a \in \mathbb{R}$$

Aus (3.9) folgt also $E(x) = a(e^{tx} + e^{-tx})$, $a, t \in \mathbb{R}$ (mit $t = 0$ erhält man die konstante Funktion).
Wegen der Monotonie von f besitzt $f - E$ in jedem Fall höchstens zwei Nullstellen in J. Dies steht im Widerspruch zur Aussage von Satz 3.1, wonach $f - E$ in J eine Alternante hat, deren Länge mindestens 4 ist. Damit ist die Annahme widerlegt.

Zur Charakterisierung der Minimallösung bezüglich V_N^+ beachte man die Bemerkung von oben.

Satz 3.4 (**Charakterisierung bezüglich $\mathbf{V_N^+}$**; Braess, [1])
Gegeben sei $E(a,x) = \sum_{i=1}^{L} a_i e^{t_i x} \in V_N^+$ mit der Länge L.

Behauptung:
Für $L = N$ ist E(a) genau dann Minimallösung für f bezüglich V_N^+ auf X, wenn $f - E(a)$ in X eine *Alternante der Länge 2N + 1* besitzt.
Gilt $L < N$, so ist E(a) genau dann Minimallösung für f bezüglich V_N^+ auf X, wenn $f - E(a)$ in X eine *negative Alternante der Länge 2L + 1* besitzt.

Beweis:
Es sei $F := f - E(a)$.

Aus Korollar 3.1 und Satz 3.2 folgt die Behauptung für $L = N$.

Für $L = 0$ ist nichts zu zeigen. Es sei $1 \leqq L < N$: $E(a) \in V_{N-1}^+$.

a. ($\Rightarrow$ ist Minimallösung)

$$\begin{gathered} \text{F besitze in X eine negative Alternante der Länge } 2L+1 \\ \text{mit den Alternantenpunkten } x_i \in X,\ 1 \leqq i \leqq 2L+1 \end{gathered} \tag{3.10}$$

Annahme:

$$\begin{gathered} E(b,x) = \sum_{i=1}^{L_1} b_i e^{s_i x} \in V_N^+ \text{ mit der Länge } L_1 \\ \text{ist eine bessere Approximation an f:} \\ \| F \|_X > \| f - E(b) \|_X \end{gathered} \tag{3.11}$$

Damit gilt nach Satz 3.2 $L_1 > L$ und

$$\begin{gathered} E := E(b) - E(a) = (f - E(a)) - (f - E(b)) \\ \text{besitzt in } (x_1, x_{2L+1}) \text{ mindestens 2L Nullstellen} \end{gathered} \tag{3.12}$$

Im Vorzeichenvektor $S := sign(E)$ liegen also mindestens 2L Vorzeichenwechsel vor. Da aber S genau L negative Komponenten enthält, muß gelten:

1. $s_1 < t_1$ und $t_L < s_{L_1}$
2. Für jedes i mit $1 \leqq i \leqq L-1$ gibt es im Spektrum von E(b) ein s(i) mit $t_i < s(i) < t_{i+1}$.

Damit erhält man

$$\lim_{x\to-\infty} E(x)e^{-s_1 x} = \lim_{x\to-\infty}(b_1 + \sum_{i=2}^{L_1} b_i e^{(s_i-s_1)x} - \sum_{i=1}^{L} a_i e^{(t_i-s_1)x}) = b_1 > 0$$

und entsprechend

$$\lim_{x\to\infty} E(x)e^{-s_{L_1}x} = b_{L_1} > 0$$

Da in S höchstens 2L Vorzeichenwechsel auftreten können, besitzt $E \not\equiv 0$ genau 2L reelle Nullstellen, die nach (3.12) in (x_1, x_{2L+1}) liegen.
Damit gilt für $i \in \{1, 2L+1\}$ $E(x_i) > 0$, also $E(b, x_i) > E(a, x_i)$, und wegen Annahme (3.11) $f(x_i) > E(a, x_i)$ [5]. Dies steht jedoch im Widerspruch zu (3.10)[6]; die Annahme (3.11) ist damit falsch und E(a) ist Minimallösung für f bezüglich V_N^+.

b. ($\Rightarrow$ negative Alternante der Länge 2L+1)

(3.13) Es sei $E(a) \in V_{N-1}^+$ Minimallösung bezüglich V_N^+ für f auf X

Nach Satz 3.2a ist E(a) dann auch Minimallösung bezüglich V_L^0 und F besitzt eine Alternante der Länge $2L+1$ in X.
Annahme:

(3.14) F besitzt in X keine negative Alternante der Länge $2L+1$

Daraus folgt, daß jede Alternante von F in X höchstens die Länge $2L+1$ hat, denn in jeder Alternante mit größerer Länge ist eine negative Alternante der Länge $2L+1$ enthalten. E(a) ist daher keine Minimallösung bezüglich V_{L+1}; es gilt also

$$\inf_{v\in V_{L+1}} \| f - v \|_X < \| F \|_X = \inf_{v\in V_L^+} \| f - v \|_X$$

und V_{L+1} enthält eine bessere Approximation E(b) an f auf X:

$$\| f - E(b) \|_X < \| F \|_X$$

Es sei x_1 bzw. x_{2L+1} der kleinste bzw. größte Alternantenpunkt einer Alternante der Länge $2L+1$ von F an X.

[5] $E(a) \in V_N^+$, $E(b) \in V_N^+ \Rightarrow E(a,x) > 0 < E(b,x)$
[6] die Alternante von F ist eine *negative* Alternante

Nach Satz 3.2 gilt $grd(E(b)) = L + 1$ und mit $E = E(b) - E(a)$ erhält man für $i \in \{1, 2L + 1\}$ wegen (3.14)

$$E(x_i) > 0 \tag{3.15}$$

Wie oben besitzt E mindestens 2L Nullstellen in (x_1, x_{2L+1}) und wegen $E \not\equiv 0$ treten in $S = sign(E)$ mindesten 2L Vorzeichenwechsel auf. Daraus folgt, daß die Spektren von E(a) und E(b) disjunkt sind; deshalb besteht S aus $2L + 1$ Komponenten.

(3.16) In S treten also genau 2L Vorzeichenwechsel auf und

(3.17) E besitzt daher genau 2L reelle Nullstellen, die in (x_1, x_{2L+1}) liegen

Es wird gezeigt: Aus (3.16) und (3.17) folgt $E(b) \in V^+_{L+1} \setminus V^+_L$.

Damit für $E(b) \in V_{L+1} \setminus V^0_{L+1}$ (3.16) gilt, muß erfüllt sein:

$$E(b,x) = (b_1 + b_2 x)e^{s_1 x} + \sum_{i=2}^{L} b_{i+1} e^{s_i x} \quad \text{mit } b_2 > 0 \text{ und} \tag{3.18}$$
$$s_1 < t_1 < s_2 < t_2 < \ldots < s_L < t_L$$

oder

$$E(b,x) = \sum_{i=1}^{L-1} b_i e^{s_1 x} + (b_L + b_{L+1} x)e^{s_L x} \quad \text{mit } b_{L+1} < 0 \text{ und} \tag{3.19}$$
$$t_1 < s_1 < t_2 < s_2 < \ldots < t_L < s_L$$

oder o.B.d.A.

$$E(b,x) = \sum_{i=0}^{2} b_{i+1} x^i e^{s_1 x} + \sum_{i=2}^{L-1} b_{i+2} e^{s_i x} \quad \text{mit } b_3 > 0 \text{ und} \tag{3.20}$$
$$t_1 < s_1 < t_2 < s_2 < \ldots < s_{L-1} < t_L$$
(wesentlich für das Folgende ist nur $t_1 < s_1 < t_L$)

Die Vorzeichen der übrigen Parameter sind damit, ebenso wie weiter unten, positiv, was jedoch nicht benötigt wird.

E(b) nach (3.18) ergibt

$$\begin{aligned}\lim_{x\to\infty} E(x)e^{-t_L x} &= \lim_{x\to\infty}((b_1+b_2x)e^{(s_1-t_L)x} + \sum_{i=2}^{L} b_{i+1}e^{(s_i-t_L)x} - \\ &- \sum_{i=1}^{L-1} a_i e^{(t_i-t_L)x} - a_L) = -a_L < 0 \\ \lim_{x\to\infty} E(x)e^{-s_1 x} &= \lim_{x\to\infty}((b_1+b_2x) + \sum_{i=2}^{L} b_{i+1}e^{(s_i-s_1)x} - \\ &- \sum_{i=1}^{L} a_i e^{(t_i-s_1)x}) = -\infty\end{aligned}$$

Mit E(b) gemäß (3.19) erhält man entsprechend

$$\lim_{x\to\infty} E(x)e^{-s_L x} = -\infty \text{ und } \lim_{x\to\infty} E(x)e^{-t_1 x} = a_1 < 0$$

und mit E(b) nach (3.20)

$$\lim_{x\to\infty} E(x)e^{-t_L x} = -a_L < 0 \text{ und } \lim_{x\to\infty} E(x)e^{-t_1 x} = a_1 < 0$$

Dies steht im Widerspruch zu (3.15) und (3.17) und damit ist gezeigt:

$$\mathbf{E(b) \notin V_{L+1} \setminus V^0_{L+1}}.$$

Für $E(b,x) = \sum_{i=1}^{L+1} b_i e^{s_i x} \in V^0_{L+1}$ gibt es genau drei Möglichkeiten, so daß (3.16) erfüllt ist:

1. Es gilt $s_1 < s_2 < t_1 < s_3 < \ldots < s_{L+1} < t_L$ und $b_1 < 0, b_2 > 0$.
   Damit folgt aber

   (3.21) $$\lim_{x\to\infty} E(x)e^{-t_L x} = a_L < 0 \text{ und } \lim_{x\to\infty} E(x)e^{-s_1 x} = b_1 < 0$$

2. Es gilt $t_1 < s_1 < t_2 < \ldots < s_{L-1} < t_L < s_{L+1}$ mit $b_L > 0$, $b_{L+1} < 0$.
   Man erhält hier analog:

   (3.22) $$\lim_{x\to\infty} E(x)e^{-s_{L+1} x} = b_{L+1} < 0 \ , \ \lim_{x\to\infty} E(x)e^{-t_1 x} = -a_1 < 0$$

   Wie oben liefern (3.21) und (3.22) den gewünschten **Widerspruch**.

3. Es ist $s_1 < t_1 < s_2 < \ldots < s_L < t_L < s_{L+1}$. Hier enthält S genau dann 2L Vorzeichenwechsel, wenn $E(b) \in V^+_{L+1}$ erfüllt ist.

Damit ist gezeigt:

(3.23) Unter der Annahme, daß (3.14) gilt, gibt es in V_{L+1} eine bessere Approximation an f auf X und jede bessere Approximation in V_{L+1} ist Element von $V_{L+1}^+ \subseteq V_N^+$

Dies steht aber im Widerspruch zur Voraussetzung (3.13) und die Annahme (3.14) ist daher falsch.
Damit ist der Satz ganz gezeigt.

Aus Satz 3.4 erhält man unmittelbar:
Korollar 3.2 (Braess, [1])
$E(a) \in V_N^+$ mit dem Grad L sei Minimallösung für f auf X bezüglich V_N^+.
Behauptung:

a. Gilt $L < N$, dann ist E(a) beste Approximation an f auf X bezüglich V_n^+ für alle $n > L$, $n \in \mathbb{N}$.

b. Gilt $L = N$, dann hat für $n < L$ die beste Approximation an f auf X bezüglich V_n^+ die Länge n.

Beweis:
Für $L < N$ besitzt $f - E(a)$ in X nach Satz 3.4 eine negative Alternante der Länge $2L + 1$, woraus unmittelbar die erste Teilbehauptung folgt.

Es sei $L = N$; ist für ein $i \in \mathbb{N}$ mit $i < N$ $E(b) \in V_i^+$ mit der Länge $i - 1$ Minimallösung für f bezüglich V_i^+, dann ist E(b) also auch beste Approximation an f bezüglich V_N^+. Wegen $E(b) \not\equiv E(a)$ steht dies im Widerspruch zu Satz 3.2.

Korollar 3.3 (Braess[7])
$E(a) \in V_{N-1}^+$ sei Minimallösung für f auf X bezüglich V_{N-1}^+, nicht aber bezüglich V_N^+; $E(b) \in V_N$ sei eine bessere Approximation an f.

Behauptung:

a.
$$E(b) \in V_N^+ \setminus V_{N-1}^+$$

b.
$$\text{Mit } E(a,x) = \sum_{i=1}^{N-1} a_i e^{t_i x} \text{ und } E(b,x) = \sum_{i=1}^{N} b_i e^{s_i x} \text{ gilt}$$
$$s_1 < t_1 < s_2 < t_2 < \ldots < t_{N-1} < s_N$$

[7] wahrscheinlich [1]

Beweis:
Nach der Voraussetzung folgt: E(a) besitzt nach Korollar 3.2 die Länge $2N-1$ und $f - E(a)$ hat in X eine Alternante der Länge $2N-1$; jede Alternante der Länge $2N-1$ von $f - E(a)$ in X ist nicht negativ und es gibt keine Alternante größerer Länge in X für $f - E(a)$.
Daraus ergibt sich wie im Beweis zu Satz 3.4 die Behauptung[8].

Es wird nun eine Verallgemeinerung von Korollar 3.3 angegeben, die Aussagen über die Vorzeichenvektoren von besten Approximationen ermöglicht[9]. Es wird hierzu benötigt:

Bemerkung 3.3 (Braess, [1])
Besitzt ein Vorzeichenvektor mit k Komponenten w Vorzeichenwechsel, so gilt $|k^+ - k^-| \leqq k - w$, wobei k^+ bzw. k^- die Zahl der positiven bzw. negativen Komponenten ist.

Beweis:
1. Fall: $k^- \leqq k^+$
Es gilt $w \leqq 2k$; unter Beachtung von $k^+ + k^- = k$ folgt damit

$$|k^+ - k^-| = k^+ - k^- = k^+ + k^- - 2k^- \leqq k - w$$

2. Fall: $k^- > k^+$
Es gilt entsprechend $w \leqq 2k^+$ und daher wie oben

$$|k^+ - k^-| = k^- - k^+ = k^- + k^+ - 2k^+ \leqq k - w$$

Satz 3.5 (Braess, [1])
$E(a) \in V_N^0$ mit $S_1 = sign(E(a))$ und $grd(E(a)) = N$ sei Minimallösung für f bezüglich V_N^0 auf X. $E(b) \in V_M$ mit $grd(E(b)) = M$ und $S_2 = sign(E(b))$ sei eine bessere Approximation an f.

Behauptung:
S_2 enthält mindestens ebensoviele positive und negative Komponenten wie S_1.
Beweis:
Da E(b) besser als E(a) approximiert, gilt $M > N$ und $E := E(b) - E(a) \not\equiv 0$ besitzt mindestens 2N reelle Nullstellen (man vergleiche hierzu (3.12)). Daher besitzt $S = sign(E)$ mindestens 2N Vorzeichenwechsel.

[8] s. Feststellung (3.23) im Beweis zu Satz 3.4
[9] s. Satz 3.5

Die Zahl der positiven (negativen) Komponenten in S_1 bzw. S_2 bzw. S sei gegeben durch p_1 (n_1) bzw. p_2 (n_2) p (n). Es gilt also

$$p = n_1 + p_2 \quad \text{und} \quad n = n_2 + p_1 \tag{3.24}$$

Mit $grd(E) \leqq M + N$ erhält man nach Bemerkung 3.3

$$| p - n | \leqq M + N - 2N = M - N \tag{3.25}$$

Fall 1: $p - n =| p - n |$
Mit (3.25) gilt also

$$N - M \leqq n_1 + p_2 - n_2 - p_1 \leqq M - N \tag{3.26}$$

Aus der ersten Ungleichung ergibt sich

$$\begin{aligned} p_2 - n_2 + p_2 - p_2 &\geqq N - M - n_1 + p_1 = N - M - n_1 - p_1 + 2p_1 \\ 2p_2 - M &\geqq -M + 2p_1 \\ p_2 &\geqq p_1 \end{aligned}$$

Entsprechend ergibt die zweite Ungleichung in (3.26)

$$\begin{aligned} 2n_1 - N &\leqq M - N - p_2 - n_2 + 2n_2 = -N + 2n_2 \\ n_1 &\leqq n_2 \end{aligned}$$

Fall 2: $p - n = - | p - n |$
Auch hier gilt (3.26) wegen $p - n \leqq| p - n |$; es folgt also

$$p_2 \geqq p_1 \quad \text{und} \quad n_1 \leqq n_2$$

Damit ist die Behauptung gezeigt.

Bemerkungen:

1. Es sei darauf hingewiesen, daß für V_N^- die Satz 3.4, Korollar 3.2 und Korollar 3.3 entsprechenden Aussagen gelten; die Beweise verlaufen analog, man hat nur die Vorzeichen zu vertauschen.
2. Für Fragestellungen, wie sie zum Beispiel in der Physik oder Chemie auftreten (Behandlung von Abklingvorgängen, radioaktiver Zerfall), sind besonders die Ergebnisse wichtig, die die Approximation bezüglich V_N^+ betreffen, da hier die linearen Koeffizienten a_i meist Massen darstellen und damit im allgemeinen nicht negativ sein sollen.

4 Lokale Minima

Es sei hier wieder $I \subseteq \mathbb{R}$ ein kompaktes Intervall und $f \in C(I)$; weiter gelte $N \in \mathbb{N}$.

Definition 4.1
Eine Funktion $E_0 \in V_N$ ($E_0 \in V_N^0$) heißt ein *lokales Minimum* für f in V_N (V_N^0), wenn es eine Umgebung U von E_0 gibt, so daß für alle $E \in U$ erfüllt ist:

$$\| f - E_0 \|_I \leqq \| f - E \|_I$$

Es wird gezeigt, daß im allgemeinen bei der Approximation von f durch Exponentialsummen mit der Existenz lokaler Minima nach Definition 4.1 zu rechnen ist; dies ist von großer Bedeutung für die Konstruktion einer Minimallösung durch Iterationsverfahren.

Einfache Verhältnisse liegen in V_N^0 vor:
Satz 4.1 (Werner, [21])
Jedes lokale Minimum für f in V_N^0 ist die beste Approximation an f bezüglich V_N^0 auf I.

Beweis:
Es sei $E(a,x) = \sum_{i=1}^{N} a_i e^{t_i x} \in V_N^0$ mit $grd(E(a)) = K$ ein lokales Minimum für f. Wählt man $\epsilon > 0$ hinreichend klein, dann ist E(a) nach Voraussetzung beste Approximation an f auf I bezüglich

$$V := \{\sum_{i=1}^{N} b_i e^{s_i x} \in V_N^0 \mid \quad \mid a_i - b_i \mid < \epsilon > \mid s_i - t_i \mid \text{für } 1 \leqq K,$$
$$\mid b_i \mid < \epsilon > \mid s_i \mid \text{für } K < i \leqq N\}$$

V erfüllt die Voraussetzungen von Hilfssatz 3.2 und der Gradientenraum von E(a) als Element von V hat die Dimension $N + K$. Damit besitzt $f - E(a)$ in I eine Alternante der Länge $N + K + 1$ und nach Korollar 3.1 ist E(a) die (eindeutig bestimmte) beste Approximation an f bezüglich V_N^0.

Aus Satz 4.1 folgt mit Korollar 3.1 unmittelbar

Korollar 4.1
Ist $E(a) \in V_N^0$ ein lokales Minimum für f in V_N, dann ist $E(a)$ die beste Approximation für f bezüglich V_N

Damit ist gezeigt, daß lokale Minima für f in V_N, die nicht zugleich Minimallösungen sind, nur in $V_N \setminus V_N^0$ enthalten sein können.

Hilfsatz 4.1 (Braess, [2]
Es sei $E(a)$ mit $grd(E(a)) = K$ auf I Minimallösung für f bezüglich V_{N-1}^0, nicht aber bezüglich V_N^0. Die Menge

$$\{t_i \mid t_i \in \mathbb{R} \ , \ K+1 \leqq i \leqq N \ , \ t_i < t_{i+1}\}$$

sei so gegeben, daß sie mit dem Spektrum von $E(a)$ kein Element gemeinsam hat.

Behauptung:
In jeder Umgebung von $E(a)$ in V_N gibt es ein $E(b) \in V_K^0$ und es gibt ein $a_i \in \mathbb{R}$, $K+1 \leqq i \leqq N$, so daß für $E(\overline{a}, x) := E(b, x) + \sum_{i=K+1}^{N} a_i e^{t_i x}$ gilt:

$$\| f - E(a) \|_I > \| f - E(\overline{a}) \|_I$$

Beweis: (Man vergleiche hierzu den Beweis von Satz 8 und 9 in [11] und Satz 2.10 in [24].)
Man betrachte die Menge

$$V := \{E(d) \in V_N \mid E(d,x) = \sum_{i=1}^{N} d_i e^{s_i x} \quad , \quad (d_1, \ldots, d_N, s_1, \ldots, s_K) \in \mathbb{R}^{N+K}, \\ s_i = t_i \text{ für } K < i \leqq N\}$$

Die Elemente von V werden mit v(p) bezeichnet, wobei der Parametervektor $p \in \mathbb{R}^{N+K}$ dem Element $v(p) \in V$ zugeordnet ist durch

$$p = (d_1, \ldots, d_N, s_1, \ldots, s_K) \in \mathbb{R}^{N+K} \leftrightarrow v(p,x) = \sum_{i=1}^{K} d_i e^{s_i x} + \sum_{i=K+1}^{N} d_i e^{t_i x} \in V$$

E(a) ist ein Element von V und der dazugehörige Parametervektor sei wieder mit a bezeichnet: $E(a) = v(a) \in V$.

Der Gradientenraum W von v(a) hat die Dimension $N+K$; nach Hilfssatz 3.2 ist daher v(a) nicht Minimallösung für f bezüglich V, da $F = f - v(a)$ in diesem Fall, im Widerspruch zur Voraussetzung über E(a), eine Alternante der Länge $N+K+1$ hätte.

Die Punkte $x_i \in I,\ 1 \leqq i \leqq N+K$, seien Alternantenpunkte einer Alternante der Länge $N+K$ von F in I, die nach Voraussetzung existiert; D sei die Menge der Extremalpunkte von F in I:

$$D = \{x \in I \mid \quad |F(x)| = \| F \|_I \}$$

Aus $F \in C(I)$ folgt die Existenz von $N+K-1$ Nullstellen $z_1 < z_2 < \ldots < z_{N+K-1}$ von F in I, so daß F auf $D \cap [z_i, z_{i+1}]$ für $0 \leqq i \leqq N+K-1$ das Vorzeichen nicht wechselt; dabei sei $z_0 = q_1, z_{N+K} = q_2$ mit $I = [q_1, q_2]$.

Da W die Haarsche Bedingung erfüllt, gibt es mit $d_1 = min\ D$ (D ist kompakt) ein $b \in \mathbb{R}^{N+K}$, so daß gilt:

$$\begin{aligned} (b, grad(v(a, d_1))) &= sign(F(d_1)) \\ (b, grad(v(a, z_i))) &= 0 \qquad\qquad 1 \leqq i \leqq N+K-1 \end{aligned}$$

Nach Bemerkung 3.2 hat (b,grad(v(a))) in I genau die $N+K-1$ Nullstellen $z_i, 1 \leqq i \leqq N+K-1$, und wechselt dort sein Vorzeichen. Daraus folgt wie im Hilfsatz 3.2:

$$\min_{x \in D}(\ F(x) \times (b, grad(v(a,x)))\ ) > 0 \tag{4.1}$$

Es wird nun benutzt, daß v(a) Fréchet-differenzierbar ist:

$$\| v(a+r) - v(a) - (r, grad(v(a))) \|_I = \ o(\| r \|)$$

wobei $r \in \mathbb{R}^{N+K}$ und $\| r \|$ die Euklidische Norm von r ist.

Da D kompakt ist, gibt es eine offene Umgebung U von D inI, so daß wegen (4.1) für ein $c > 0$, $c \in \mathbb{R}$, gilt:

$$F(x)(b, grad(v(a,x))) \geqq 2c > 0 \qquad x \in U$$

Für $x \in U$ folgt mit $t \in \mathbb{R}$, $t > 0$

$$\begin{aligned} F(x)(v(a+tb,x) \ &- \ v(a,x) = \\ &= \ F(x)t(b, grad(v(a,x))) + \\ &\quad +F(x)v(a+tb,x) - v(a,x) - (tb, grad(v(a,x)))) \geqq \\ &\geqq \ 2ct - \| F \|_I \| \ v(a+tb) - v(a) - (tb, grad(v(a)))\ \|_I \end{aligned}$$

Daher gibt es ein $T_1 > 0$, so daß für $t \in (0, T_1]$ für alle $x \in U$ gilt:

$$F(x)(v(a+tb,x) - v(a,x)) \geqq ct$$

Die Menge $I \setminus U$ ist abgeschlossen in I und für $x \in I \setminus U$ gilt

$$\mid F(x) \mid < \parallel F \parallel_I \ ;$$

daraus folgt $h := \parallel F \parallel_I - \parallel F \parallel_{I \setminus U} > 0$.

Die Dreiecksungleichung ergibt für $t \in \mathbb{R}$

$$\begin{aligned} \parallel v(a+tb) - v(a) \parallel_I \ \leqq \ & \parallel (tb, grad(v)a))) \parallel_I + \\ & + \parallel v(a+tb) - v(a) - (tb, grad(v(a))) \parallel_I \ ; \end{aligned}$$

daher gibt es ein $T_2 > 0$, so daß für $t \in (0, T_2]$ gilt:

$$\parallel v(a+tb) - v(a) \parallel_I \leqq \ 2t \parallel (b, grad(v(a))) \parallel_I$$

Für $T \in \mathbb{R}$ gelte

$$0 < T < min\{ \ T_1, T_2, \frac{c}{4 \parallel (b, grad(v(a))) \parallel_I^2}, \frac{h}{4 \parallel (b, grad(v(a))) \parallel_I} \ \}$$

Für $x \in U$ erhält man also

$$\begin{aligned} \mid f(x) \ - \ & v(a+Tb, x) \mid^2 = \\ = \ & \mid f(x) - v(a,x) \mid^2 - 2F(x)(v(a+Tb,x) - v(a,x)) + \\ & + \mid v(a,x) - v(a+Tb,x) \mid^2 \leqq \\ \leqq \ & \parallel F \parallel_I^2 -2cT + 4T^2 \parallel \ (b, grad(v(a))) \parallel_I^2 \ \leqq \\ \leqq \ & \parallel F \parallel_I^2 -2cT + cT < \\ < \ & \parallel f - v(a) \parallel_I^2 \end{aligned}$$

Für $x \in I \setminus U$ gilt die Abschätzung

$$\begin{aligned} \mid f(x) \ - \ & v(a+Tb, x) \mid \leqq \\ \leqq \ & \mid F(x) \mid + \mid v(a,x) - v(a+Tb,x) \mid \leqq \\ \leqq \ & \parallel F \parallel_I - \parallel F \parallel_I + \parallel F \parallel_{I \setminus U} + \parallel v(a) - v(a+Tb) \parallel_I \leqq \\ \leqq \ & \parallel F \parallel_I -h + 2T \parallel (b, grad(v(a))) \parallel_I \leqq \parallel F \parallel_I -h + \frac{h}{2} < \\ < \ & \parallel f - v(a) \parallel_I \end{aligned}$$

Da $T > 0$ beliebig klein gewählt werden kann, ist die Behauptung nach Definition von V gezeigt.

Satz 4.2 (Braess, [2])
$E(a,x) = \sum_{i=1}^{K} a_i e^{t_i x}$ mit dem Grad K und $S(a) = sign(E(a))$ sei Minimallösung für f bezüglich V_{N-1}^0, nicht aber bezüglich V_N. In I habe $f - E(a)$ eine positive (negative) Alternante der Länge $N + K$, deren Alternantenpunkte mit x_i, $1 \leqq i \leqq N + K$, bezeichnet werden. $S = (S_1, \ldots, S_N)$ sei ein Vorzeichenvektor mit N Komponenten, so daß erfüllt ist:

1. Durch Streichen von $N - K$ Komponenten S_{i_k}, $1 \leqq k \leqq N - K$, in S erhält man $S(a)$ und es gilt $i_{k-1} < i_k$.
2. $S_{i_{N-K}} = +1$ $(S_{i_{N-K}} = -1)$ und $sign(S_{i_{k-1}}) = -sign(S_{i_k}), 1 < k \leqq N–K$.

Behauptung:
$V_N(S)$ enthält eine bessere Approximation als $E(a)$.
Beweis:

(4.2) Die Menge $T = \{s_i \mid K + 1 \leqq i \leqq N, s_i < s_j \text{ für } i < j\} \subseteq \mathbb{R}$ sei so gegeben, daß T und das Spektrum von E(a) disjunkt sind.

Ordnet man die Frequenzen von E(a) und die Elemente von T der Größe nach, so erhält man eine aufsteigende Kette $s_1' < s_2' < s_3' < \ldots < s_N'$;
für die Elemente von T gelte hierbei

(4.3) $$s_{i_k}' = s_{K+k} \ , \ 1 \leqq k \leqq N-K$$

Nach Hilfssatz 4.1 gibt es in V_N eine bessere Approximation
$E(c,x) = E(b,x) + \sum_{i=K+1}^{N} c_i e^{s_i x}$ mit $E(b,x) = \sum_{i=1}^{K} b_i e^{s_i x}$, so daß gilt:

1. $sign(E(a)) = sign(E(b))$
2. (4.4) $| s_i - t_i | < | s_i - t |$ für alle $t \in T$ und für alle i mit $1 \leqq i \leqq K$

Dies ergibt sich aus der Konstruktion der besseren Approximation E(c) nach Hilfssatz 4.1. Nach einem mehrfach benutzten Schluß besitzt $E := E(c) - E(a) \not\equiv 0$ in (x_1, x_{N+K}) mindestens $N+K-1$ Nullstellen und $S(E) = sign(E)$ hat nach Satz 1.3 daher mindestens $N + K - 1$ Vorzeichenwechsel;

(4.5) S(E) besteht also aus genau $N + K$ Komponenten, die abwechselnd positiv und negativ sind.

Nach Satz 1.1 hat E genau $N+K-1$ reelle Nullstellen. Ist die Alternante der Voraussetzung positiv (negativ), so gilt

$$E(x_{N+K}) > 0 \qquad (\ E(x_{N+K}) < 0\ )$$

Man beachte hierbei, daß nach Voraussetzung alle Alternanten von $f - E(a)$ in I mit der Länge $N+K$ positiv (negativ) sind, falls eine Alternante dieser Länge positiv (negativ) ist.
Aus (4.5) folgt weiter $c_i \neq 0$ für $K+1 \leqq i \leqq N$.

Wie im ersten Teil des Beweises zu Satz 3.4 erhält man damit

$$\text{für } s_N > t_K\text{:} \quad sign(c_N) = +1 \quad (sign(c_N) = -1)$$
$$\text{für } s_N < t_K : \begin{cases} sign(b_K) = +1 & (sign(b_K) = -1) \quad \text{falls } s_K > t_K \text{ gilt} \\ sign(a_K) = -1 & (sign(a_K) = +1) \quad \text{falls } s_K < t_K \text{ gilt.} \end{cases}$$

Die (N+K)-te Komponente von S(E) ist damit positiv (negativ).

In (s_i, s_{i+1}) liegt für $K+1 \leqq i \leqq N-1$ stets eine gerade Anzahl von Elementen aus dem Spektrum von E(a)+E(b); dies folgt aus (4.4). Mit (4.5) erhält man also für $K+1 \leqq i \leqq N$:

$$sign(c_i) = -sign(c_{i+1}) \text{ und } \quad sign(c_N) = +1 \ \ (\ sign(c_N) = -1\ ).$$

Damit ist die Behauptung gezeigt, denn $E(c) \in V_N(\overline{S})$ ist eine bessere Approximation an f und wegen (4.3) erfüllt $\overline{S} = sign(E(c))$ die Voraussetzung des Satzes.

Bemerkung 4.1 (Braess, [2])
Für $K = N-1$ läßt sich die Zahl der verschiedenen Vorzeichenvektoren S, die man nach der in Satz 4.2 verwandten Konstruktion erhalten kann, genau bestimmen:
Es liege eine positive Alternante von $f - E(a)$ vor und $S(a)$ habe genau k^- negative Komponenten; die entsprechenden Frequenzen von $E(a)$ teilen $\mathbb{R}$ in $k^- + 1$ Intervalle. Die Menge T besteht hier nur aus dem Element s_N und die Wahl von s_N bestimmt, wie oben gezeigt, den Vorzeichenvektor S der besseren Approximation $E(c)$. Daher gibt es also $k^- + 1$ verschiedene Vorzeichenklassen, die bessere Approximationen enthalten.
Ist die betrachtete Alternante negativ, so ergeben sich analog $k^+ + 1$ Vorzeichenklassen, wenn $S(a)$ genau k^+ positive Komponeten enthält.

Satz 4.3 (Braess, [2])
$E(a) \in V_{N-1}^0$ mit $grd(E(a)) = K \neq 0$ und dem Vorzeichenvektor $S(a) = (S_1, \ldots, S_K)$ sei die Minimallösung bezüglich V_{N-1} für f auf I.
Erfüllt jede Minimallösung E für f bezüglich V_N die Beziehung $E \notin V_{N-1} \cup V_N^+ \cup V_N^-$, dann gibt es in V_N mindestens zwei lokale Minima für f.

Beweis:
Es sei wieder $F = f - E(a)$. Nach Voraussetzung besitzt F in I eine Alternante der Länge $N + K$ und falls diese positiv (negativ) ist, sind alle Alternanten von F in I mit dieser Länge positiv (negativ).

In V_N werden zwei verschiedene Vorzeichenklassen angegeben, die bessere Approximationen als E(a) enthalten:

Fall 1: $K < N - 1$
Es sei $s = +1$, falls F in I eine positive Alternante der Länge $N + K$ besitzt, und $s = -1$, falls diese negativ ist. Die $N - K$ Komponenten von

$$((-1)^{N-K-1}s, \ldots, (-1)^1 s, (-1)^0 s)$$

lassen sich auf mindestens zwei Arten in S(a) einfügen, so daß man die verschiedenen Vorzeichenvektoren

$$(S_1, (-1)^{N-K-1}s, \ldots)$$
$$((-1)^{N-K-1}s, (-1)^{N-K-2}s, S_1, \ldots)$$

erhält

Fall 2: $K = N - 1$
F habe in I eine positive Alternante der Länge $N + K$. Hieraus folgt $E(a) \notin V_{N-1}^+$: Denn mit $E(a) \in V_{N-1}^+$ erhält man durch Einfügen von $s = +1$ in S(a) stets einen Vorzeichenvektor mit N positiven Komponenten, so daß nach Satz 4.2 in $V_N^+ \setminus V_{N-1}$ eine bessere Approximation als E(a) enthalten ist; da f bezüglich V_N^+ eine beste Approximation auf I besitzt und diese nach Satz 3.2 Element von $V_N^+ \setminus V_{N-1}$ und auch beste Approximation bezüglich V_N ist, erhält man einen Widerspruch zur Voraussetzung.
Es gibt also ein $i \in \{1, \ldots, N-1\}$ mit $S_i = -1$.

Durch Einfügen von $s = +1$ in S(a) erhält man die Vorzeichenvektoren

$$(S_1, \ldots, S_{i-1}, -1, S_{i+1}, \ldots, S_{N-1}, +1)$$
$$(S_1, \ldots, S_{i-1}, +1, -1, \ldots, S_{N-1})$$

Hat F eine negative Alternante der Länge $N + K$ in I, so erhält man analog $E(a) \notin V_{N-1}^-$; es gibt also ein $i \in \{1, \ldots, N-1\}$ mit $S_i = +1$ und man erhält

durch Einfügen von $s = -1$ entsprechend die Vorzeichenvektoren

$$(S_1, \ldots, S_{i-1}, +1, S_{i+1}, \ldots, S_{N-1}, -1)$$
$$(S_1, \ldots, S_{i-1}, -1, +1, \ldots, S_{N-1})$$

Es gibt somit für $K = N - 1$ und $K < N - 1$ nach Satz 4.2 mindestens zwei verschiedene Vorzeichenklassen in V_N, die bessere Approximationen an f als E(a) enthalten.

Nun wird gezeigt, daß es bezüglich jeder der soeben konstruierten Vorzeichenklassen eine beste Approximation an f gibt:
Es sei also $V := V_N(S)$ so gegeben, daß V eine bessere Approximation an f enthält als E(a); nach Voraussetzung hat diese den Grad N. Die Folge $(h_m)_{m\in\mathbb{N}}$ sei eine Minimalfolge in V; es gilt also

$$\lim_{m\to\infty} \| f - h_m \|_I = \inf_{v\in V} \| f - v \|_I$$

und es kann o.B.d.A. $grd(h_m) = N$ für $m \in \mathbb{N}$ angenommen werden.

Nach Satz 2.3 gibt es eine Teilfolge $(h_m)_{m\in\mathbb{N}}$, die auf jedem abgeschlossenen Teilintervall von I gleichmäßig konvergiert und die Grenzfunktion ist ein Element von V nach Korollar 2.5; daraus folgt mit $f \in C(I)$ wie im Beweis zu Satz 2.4 die Existenz einer besten Approximation h bezüglich V. Nach Voraussetzung gilt $\| f - E(a) \|_I > \| f - h \|_I$ und mit

$$0 < \epsilon = \| f - E(a) \|_I - \| f - h \|_I$$

gilt für $h_1 \in V_N$ mit $\| h - h_1 \|_I < \epsilon$

$$\| f - h_1 \|_I \leqq \| f - h \|_I + \| h - h_1 \|_I < \| f - E(a) \|_I$$

Es gibt es also eine Umgebung U von h in V_N mit $U \subseteq V_N(S) \setminus V_{N-1}$, so daß gilt[10]:

$$\| f - h_1 \|_I \geqq \| f - h \|_I \qquad h_1 \in U$$

Dies war zu zeigen.

[10] Korrektur von [26]: dort wird an dieser Stelle auf ein (in [26] nicht existierendes) "Korrollar 2.7" Bezug genommen.

Teil II

Numerische Verfahren

5 Konstruktion von Näherungen

Es sei I das Intervall $[a, b] \in \mathbb{R}$ und $f \in C(I)$; $N \in \mathbb{N}$.

Es werden nun Verfahren zur Konstruktion von Näherungen beschrieben; das entscheidende Problem hierbei ist die Behandlung eines nichtlinearen Gleichungssystems.

5.1 Näherungen nach Meinardus

Gegeben sei eine Punktmenge $X = \{x_i \mid 0 \leqq i \leqq 2N-1\} \subseteq I$; gesucht ist eine Funktion

$$\begin{aligned} E(x) &= \sum_{i=1}^{N} a_i e^{t_i x} \in V_N^0 \qquad \text{mit} \\ E(x_i) &= f(x_i) \qquad 0 \leqq i \leqq 2N-1 \end{aligned} \tag{5.1}$$

Nach [13] wird eine zumindest formale Lösung dieses Interpolationsproblems angegeben; hierzu wird X als äquidistante Punktmenge angenommen:

$$x_i = a + ih \qquad 0 \leqq i \leqq 2N-1,\ h > 0\ .$$

Setzt man für $1 \leqq j \leqq N$

$$\begin{aligned} A_j &:= a_j e^{t_j a}\ ,\ E_j := e^{t_j h} \\ f_j &:= f(x_j) \quad \text{für } 0 \leqq j \leqq 2N-1\ , \end{aligned}$$

so erhält man aus (5.1)

$$\sum_{j=1}^{N} A_j E_j^i = f_i \qquad 0 \leqq i \leqq 2N-1 \tag{5.2}$$

Zur Lösung dieses nichtlinearen Gleichungssystems mit den 2N Unbekannten A_j und E_j wird ein Verfahren angewandt, das auf Srinvasa Ramanujan, [14], zurückgeht:

Die rationale Funktion

$$R(x) := \sum_{j=1}^{N} \frac{A_j}{1 - E_j x}$$

besitzt für $x \in U$, $U = \{x \in \mathbb{R} \mid \max_{1 \leqq j \leqq N} \mid E_j x \mid < 1\}$, die Darstellung [11]

$$R(x) = \sum_{j=1}^{N} A_j \sum_{i=0}^{\infty} E_j^i x^i = \sum_{i=0}^{\infty} x^i \sum_{j=1}^{N} A_j E_j^i$$

Mit $f_i := \sum_{j=1}^{N} A_j E_j^i$ für $i \geqq 2N$ erhält man unter Beachtung von (5.2) für $x \in U$:

$$R(x) = \sum_{i=0}^{\infty} x^i f_i$$

R läßt sich weiter schreiben in der Form $R(x) = \frac{P(x)}{Q(x)}$ mit

$$P(x) := \sum_{i=0}^{N-1} p_i x^i := \sum_{j=1}^{N} A_j \prod_{i=1, i \neq j}^{N} (1 - E_i x) \qquad \text{und} \tag{5.3}$$

$$Q(x) := \sum_{i=0}^{N} q_i x^i := \prod_{i=1}^{N} (1 - E_i x) \tag{5.4}$$

Es gilt somit für $x \in U$:

$$\sum_{i=0}^{N-1} p_i x^i = (\sum_{i=0}^{N} q_i x^i)(\sum_{i=0}^{\infty} x^i f_i) = \sum_{i=0}^{\infty} x^i \sum_{\substack{k+j=i \\ k \geqq 0 \\ 0 \leqq j \leqq N}} q_i f_k$$

und Koeffizientenvergleich ergibt

$$\sum_{j=0}^{i} q_j f_{i-j} = p_i \qquad 0 \leqq i \leqq N-1$$

$$\sum_{j=0}^{N} q_j f_{i-j} = 0 \qquad N \leqq i \leqq 2N-1$$

[11] wegen: $\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n$

Aus (5.4) folgt $q_0 = 1$ und man erhält so ein lineares Gleichungssystem mit den 2N Unbekannten q_i, $1 \leqq i \leqq N$, und p_i, $0 \leqq i \leqq N-1$:

$$\begin{aligned} p_0 &= f_0 \\ \sum_{j=1}^{i} q_j f_{i-j} - p_i &= -f_i \qquad 1 \leqq i \leqq N-1 \qquad (5.5) \\ \sum_{j=1}^{N} q_j f_{i-j} &= -f_i \qquad N \leqq i \leqq 2N-1 \qquad (5.6) \end{aligned}$$

Die in (5.5) auftretende Dreiecksmatrix mit N-1 Zeilen und Spalten sei mit L_{N-1} bezeichnet; die (N-N)-Matrix F_N des Systems (5.6) ist eine Hankelsche Matrix:

$$L_{N-1} = \begin{pmatrix} f_0 & 0 & 0 & 0 & \dots & 0 \\ f_1 & f_0 & 0 & 0 & \dots & 0 \\ f_2 & f_1 & f_0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & 0 \\ f_{N-3} & f_{N-4} & \dots & \dots & f_0 & 0 \\ f_{N-2} & f_{N-3} & f_{N-4} & f_{N-5} & \dots & f_0 \end{pmatrix}$$

$$F_N = \begin{pmatrix} f_{N-1} & f_{N-2} & \dots & f_0 \\ f_N & f_{N-1} & \dots & f_1 \\ \vdots & \vdots & \vdots & \vdots \\ f_{2N-2} & f_{2N-3} & \dots & f_{N-1} \end{pmatrix}$$

5.1.1 Der Algorithmus

Zur Lösung von (5.2) und damit des Ausgangsproblems geht man vor wie folgt:

1. Ermittlung der Frequenzen

Zuerst bestimmt man eine Lösung q des linearen Gleichungssystems (5.6):

$$F_N \times q = -\begin{pmatrix} f_N \\ f_{N+1} \\ \vdots \\ f_{2N-1} \end{pmatrix} \text{ mit } q = \begin{pmatrix} q_1 \\ q_2 \\ \vdots \\ q_N \end{pmatrix}$$

Existiert eine Lösung q, so werden die Nullstellen des Polynoms $Q(x) = \sum_{i=0}^{N} q_i x^i$ mit $q_0 = 1$ ermittelt.

Besitzt Q die N Nullstellen z_i, $1 \leqq i \leqq N$, so sind diese wegen $q_0 = 1$ von Null verschieden und nach (5.4) sei o.B.d.A.

$$E_i = z_i^{-1} \qquad 1 \leqq i \leqq N \tag{5.7}$$

Besitzt Q nur einfache, reelle Nullstellen und ist weiter $z_i > 0$ für $1 \leqq i \leqq N$ erfüllt, so erhält man die reellen Frequenzen

$$t_i = h^{-1} \ln E_i = -h^{-1} \ln z_i \ , \ 1 \leqq i \leqq N$$

2. Die Koeffizienten A_i und a_i

Die Koeffizienten p_i, $1 \leqq i \leqq N-1$, erhält man aus Gleichungssystem (5.5):

$$L_{N-1} \times \begin{pmatrix} q_1 \\ q_2 \\ \vdots \\ q_{N-1} \end{pmatrix} + \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_{N-1} \end{pmatrix} = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_{N-1} \end{pmatrix}$$

Mit $p_0 = f_0$ ergeben sich nunmehr aus (5.3) und (5.7) durch Koeffizientenvergleich die Größen A_i, $1 \leqq i \leqq N$; es genügt dazu, daß Q genau N reelle Nullstellen besitzt und so E_i, $1 \leqq i \leqq N$, nach (5.7) bestimmt ist.
Für den Fall, daß Q nur einfache, reelle Nullstellen besitzt, kann man nach Satz 5.1 A_i explizit angeben.

Sind N Frequenzen $t_i \in \mathbb{R}$, $1 \leqq i \leqq N$, ermittelt, so gilt

$$a_i = A_i e^{-t_i a} \in \mathbb{R} \ , \ 1 \leqq i \leqq N$$

Bemerkungen:

1. Sind die Frequenzen t_i, $1 \leqq i \leqq N$ bekannt, dann können die Koeffizienten a_i, $1 \leqq i \leqq N$ auch durch Lösung des (nunmehr linearen) Gleichungssystems (5.1) bestimmt werden.

2. Ein nichtlineares Gleichungssystem wie (5.2) erhält man, wenn man an Stelle des Interpolationsproblems (5.1) ein Approximationsproblem auf einer Punktmenge

   $$\{x_i \mid 0 \leqq i \leqq 2N,\ x_i = x_0 + ih\}$$

   zu lösen versucht; man vergleiche hierzu Abschnitt 5.2.

3. Kelly beschreibt in [9] zur Lösung von (5.1) ein Verfahren von Prony ("Prony's method"):
   Es wird benutzt, daß die f_i, $0 \leqq i \leqq 2N-1$, falls $f_i = E(x_i)$ mit $E \in V_N^0$ erfüllt ist, einer Differenzengleichung N-ter Ordnung genügen. Die Bestimmung der Koeffizienten dieser Differenzengleichung führt auf das Gleichungssystem (5.6) und die Frequenzen ergeben sich wie oben beschrieben.

5.1.2 Zur Theorie des Verfahrens

Die Bezeichnungen von oben werden beibehalten.
Formeln zur direkten Bestimmung der A_i gibt Satz 5.1 an:

Satz 5.1
Q besitze die N einfachen, reellen Nullstellen E_j^{-1}, $1 \leqq j \leqq N$; E_j ist für $1 \leqq j \leqq N$ wegen $q_0 = 1$ definiert.
Behauptung:

$$A_k = \frac{\sum_{j=0}^{N-1} p_j E_k^{N-1-j}}{\prod_{\substack{j=1 \\ j \neq k}}^{N} (E_k - E_j)} \qquad 1 \leqq k \leqq N$$

Beweis:
Wegen $E_j^{-1} \neq 0$ für $1 \leqq j \leqq N$ erhält man aus (5.3) für $1 \leqq k \leqq N$:

$$\sum_{j=0}^{N-1} p_j E_k^{-j} = \sum_{i=1}^{N} A_i \prod_{\substack{j=1 \\ j \neq i}}^{N} (1 - \frac{E_j}{E_k}) = A_k \prod_{\substack{j=1 \\ j \neq k}}^{N} (1 - \frac{E_j}{E_k})$$

$$A_k = \sum_{j=0}^{N-1} p_j E_k^{-j} \prod_{\substack{j=1 \\ j \neq k}}^{N} E_k (E_k - E_j)^{-1} = \sum_{j=0}^{N-1} p_j E_k^{N-1-j} \prod_{\substack{j=1 \\ j \neq k}}^{N} (E_k - E_j)^{-1}$$

Damit ist die Behauptung gezeigt.

Es folgen nun Aussagen zur Existenz von Lösungen für das Gleichungssystem (5.6) und deren Eindeutigkeit.

Hilfsatz 5.1

Es sei $n \in \mathbb{N}$; besitzt das Polynom $\sum_{i=0}^{n} c_i x^i$ eine Nullstelle $z \neq 0$, dann ist z^{-1} eine Nullstelle des Polynoms $\sum_{i=0}^{n} c_{n-i} x^i$.

Beweis:

Nach Voraussetzung gilt

$$z^n \sum_{i=0}^{n} c_{n-i}(z^{-1})^i = \sum_{i=0}^{n} c_{n-i}(z^{n-i}) = 0$$

Wegen $z^n \neq 0$ folgt daraus die Behauptung.

Satz 5.2

Es gelte $f_i = E(x_i) = \sum_{j=1}^{n} a_j e^{t_i x_i}$ für $0 \leqq i \leqq 2N-1$ mit $grd(E) = n \leqq N$; A_i und E_i sind für $1 \leqq i \leqq n$ definiert wie oben. Es sei $z_i \in \mathbb{R}$ mit $z_i \neq 0$ für $n+1 \leqq i \leqq N$ beliebig gewählt. Die Koeffizienten q_i, $0 \leqq i \leqq N$, seien gegeben durch

$$\sum_{i=0}^{N} q_i x^i := \prod_{i=1}^{n}(1 - E_i x) \prod_{i=n+1}^{N} (1 - z_i x)$$

Behauptung:

(5.8) Es gilt $$F_N \begin{pmatrix} q_1 \\ q_2 \\ \vdots \\ q_N \end{pmatrix} = - \begin{pmatrix} f_N \\ f_{N+1} \\ \vdots \\ f_{2N-1} \end{pmatrix}$$ und

für $n = N$ bilden die Koeffizienten q_i, $1 \leqq i \leqq N$ die eindeutig bestimmte Lösung von (5.8).

Beweis:

Das Polynom $\sum_{i=0}^{N} q_i x^i$ besitzt nach Voraussetzung die n Nullstellen E_i^{-1}, $1 \leqq i \leqq n$ (E_i^{-1} ist wegen $E_i = e^{t_i h}$ stets definiert). Nach Hilfssatz 5.1 gilt für $1 \leqq k \leqq n$:

$$0 = \sum_{i=0}^{N} q_{N-i} E_k^i = \sum_{i=1}^{N} q_i E_k^{N-i} + q_0 E_k^N \ ;$$

wegen $q_0 = 1$ folgt für $1 \leqq k \leqq n$:

$$\sum_{i=1}^{N} q_i E_k^{N-i} = -E_k^N$$

Für $1 \leqq j \leqq N$ gilt hiermit

$$\begin{aligned}\sum_{i=1}^{N} f_{N+j-1-i} q_i &= \sum_{i=1}^{N}(\sum_{k=1}^{n} A_k E_k^{N+j-1-i}) q_i = \sum_{k=1}^{n}(\sum_{i=1}^{N} q_i E_k^{N-i}) A_k E_k^{j-1} = \\ &= \sum_{k=1}^{n} A_k E_k^{j-1}(-E_k^N) = f_{N+j-1} \ ;\end{aligned}$$

damit ist (5.8) gezeigt.

Es sei nun $n = N$; es gilt also $\sum_{i=0}^{N} q_i x^i = \prod_{i=1}^{N}(1 - E_i x)$ und es sei $\begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_N \end{pmatrix}$ eine zweite Lösung von (5.8); für $1 \leqq j \leqq N$ folgt also

$$\sum_{i=1}^{N} f_{N+j-1-i} v_i = \sum_{k=1}^{N} A_k E_k^{j-1}(\sum_{i=1}^{N} v_i E_k^{N-i}) = -f_{N+j-1} \tag{5.9}$$

Die (N-N)-Matrix dieses Gleichungssystems besteht aus den Elementen $A_k E_k^{j-1}$, $1 \leqq k, j \leqq N$. Wegen $E \in V_N^0 \setminus V_{N-1}$ gilt

$$A_k \neq 0 \ , \ 1 \leqq k \leqq N, \text{ und } 0 < E_k < E_{k+1}, \ 1 \leqq k \leqq N-1 \ . \tag{5.10}$$

Somit erhält man

$$\det \begin{pmatrix} A_1 E_1^0 & A_2 E_2^0 & \dots & A_N E_N^0 \\ A_1 E_1^1 & A_2 E_2^1 & \dots & A_N E_N^1 \\ \vdots & \vdots & \vdots & \vdots \\ A_1 E_1^{N-1} & A_2 E_2^{N-1} & \dots & A_N E_N^{N-1} \end{pmatrix} =$$

$$= (\prod_{i=1}^{N} A_i) \ \det \begin{pmatrix} 1 & 1 & \dots & 1 \\ E_1 & E_2 & \dots & E_N \\ \vdots & \vdots & \vdots & \vdots \\ E_1^{N-1} & E_2^{N-1} & \dots & E_N^{N-1} \end{pmatrix} \neq 0 \ ,$$

da eine Vandermondesche Determinante vorliegt und die Ungleichungen (5.10) bestehen. Das Gleichungssystem (5.9) ist also eindeutig lösbar und es gilt daher für $1 \leqq k \leqq N$

$$\sum_{i=1}^{N} v_i E_k^{N-i} = -E_k^N \ ,$$

da $-\begin{pmatrix} E_1^N \\ \vdots \\ E_N^N \end{pmatrix}$ eine Lösung von (5.9) ist.

Das Polynom $\sum_{i=0}^{N} v_i x^{N-i} = \sum_{i=0}^{N} v_{N-i} x^i$ mit $v_0 = 1$ besitzt damit die N Nullstellen E_k, $1 \leqq k \leqq N$, und nach Hilfssatz 5.1 gilt für $x = E_k^{-1}$:

$$\sum_{i=0}^{N} v_i x^i = 0 \qquad 1 \leqq k \leqq N$$

Nach Voraussetzung stimmen die Nullstellen von $\sum_{i=0}^{N} v_i x^i$ und $\sum_{i=0}^{N} q_i x^i$ überein, womit die Eindeutigkeit gezeigt ist:

$$q_i = v_i \ , \qquad 1 \leqq i \leqq N$$

Bemerkung:
Nimmt man an, daß $f_i = E(x_i)$ für $0 \leqq i \leqq 2N-1$ mit $E \in V_N^0 \setminus V_{N-1}$ erfüllt ist - (5.1) besitze also die eindeutig bestimmte Lösung E - , so folgt mit Satz 5.2, daß das beschriebene Verfahren die Lösung des Interpolationsproblems (5.1) ergibt.

Es soll nun noch gezeigt werden, daß die Aussage von Satz 5.2 zur Existenz einer Lösung nicht nur für $E \in V_N^0$ richtig ist.

Hilfsatz 5.2
Für $n \in \mathbb{N}$ und $t \in \mathbb{R}$ gilt mit $0 \leqq m < n$:

$$\sum_{k=0}^{n} (-1)^k \binom{n}{k} (t-k)^m = 0$$

Beweis (Meinardus, [13]):
Es sei $K_r = \{z \in \mathbb{C} \mid \ |z| = r\}$, $r \in \mathbb{R}$. Unter Verwendung des Residuenkalküls erhält man für $r > n$:

$$\begin{aligned}\frac{1}{2\pi \mathrm{i}} \int_{K_r} (t-z)^m \prod_{j=0}^{n} (z-j)^{-1} dz &= \sum_{k=0}^{n} Res_{z=k} (t-z)^m \prod_{j=0}^{n} (z-j)^{-1} = \\ = \sum_{k=0}^{n} \lim_{\substack{z \to k \\ z \neq k}} (z-k)(t-z)^m \prod_{j=0}^{n} (z-j)^{-1} &= \sum_{k=0}^{n} (t-k)^m \prod_{\substack{j=0 \\ j \neq k}}^{n} (k-j)^{-1} = \\ = \sum_{k=0}^{n} \frac{(t-k)^m}{k!(n-k)!} (-1)^{n-k} &= (-1)^n \frac{1}{n!} \sum_{k=0}^{n} (-1)^k (t-k)^m \binom{n}{k}\end{aligned}$$

Es sei $A(z) := \sum_{j=0}^{m} a_j z^j := (t-z)^m$ und $B(z) := \sum_{j=0}^{n+1} b_j z^j := \prod_{j=0}^{n} (z-j)$.
Es gilt $a_m = (-1)^m$, $b_{n+1} = 1$ und man erhält für $z \in \mathbb{C}$, $z \neq 0$:

$$\begin{aligned} \mid A(z) \mid &\leqq \mid z \mid^m (1 + \sum_{j=0}^{m-1} \mid a_j \mid\mid z \mid^{i-m}) \\ \mid B(z) \mid &\geqq \mid z \mid^{n+1} (1 - \sum_{i=0}^{n} \mid b_i \mid\mid z \mid^{i-n-1}) \end{aligned}$$

Wählt man daher $r_o \in \mathbb{R}$, $r_0 > n$, hinreichend groß, so gilt $\mid A(z) \mid \leqq 2 \mid z \mid^m$ und $\mid B(z) \mid \geqq \frac{1}{2} \mid z \mid^{n+1}$ für $z \in \mathbb{C}$ mit $\mid z \mid \geqq r_0$. Daraus folgt mit $r \geqq r_0$:

$$\mid \int_{K_r} \frac{A(z)}{B(z)} dz \mid \leqq 8\pi r^{m-n}$$

Wegen $m < n$ erhält man somit für alle $r \in \mathbb{R}$ mit $r \geqq r_0$

$$\frac{1}{n!} \mid \sum_{k=0}^{n} (-1)^k (t-k)^m \binom{n}{k} \mid \leqq 4r^{-1} \ ,$$

woraus die Behauptung folgt.

Bemerkung:
Diesen Hilfssatz erhält man auch aus der Beziehung (s. Willers, [23], S. 73)

$$\Delta^n (0, 1, \ldots, n) q = \frac{1}{n!} \sum_{k=0}^{n} (-1)^{n-k} \binom{n}{k} q(k)$$

mit $q(x) = (t-x)^m$; denn aus $m < n$ folgt nach (1.9) $\Delta^n(0,1,\ldots,n)q = 0$ und damit $\sum_{k=0}^{n}(-1)^k\binom{n}{k}(t-k)^m = 0$.

Satz 5.3

Es sei $N \geqq 2$ und $f_i = E(x_i)$ mit $E(x) = (\sum_{i=0}^{N-1} a_i x^i)e^{tx} \in V_N \setminus V_{N-1}$ für $0 \leqq i \leqq 2N-1$; ferner sei mit $E_1 = e^{th}$ für $0 \leqq i \leqq N$ q_i gegeben durch $\sum_{i=0}^{N} q_i x^i = (1 - E_1 x)^N$.

Behauptung:

$$F_N \begin{pmatrix} q_1 \\ q_2 \\ \vdots \\ q_N \end{pmatrix} = - \begin{pmatrix} f_N \\ f_{N+1} \\ \vdots \\ f_{2N-1} \end{pmatrix}$$

Beweis:

Mit dem binomischen Lehrsatz erhält man

$$\sum_{i=0}^{N} q_i x^i = \sum_{i=0}^{N}(-1)^i\binom{N}{i}E_1^i x^i$$

und daraus durch Koeffizientenvergleich

$$q_i = (-1)^i\binom{N}{i}E_1^i\ , \qquad 0 \leqq i \leqq N$$

Für $1 \leqq j \leqq N$ gilt daher mit $A = e^{ta}$:

$$\begin{aligned}
\sum_{i=1}^{N} f_{N+j-1-i}q_i &= A\sum_{i=1}^{N}(\sum_{k=0}^{N-1} a_k x_{N+j-1-i}^k)E_1^{N+j-1-i}q_i = \\
&= AE_1^{N+j-1}\sum_{k=0}^{N-1} a_k(\sum_{i=1}^{N} E_1^{-i}q_i x_{N+j-1-i}^k) = \\
&= AE_1^{N+j-1}\sum_{k=0}^{N-1} a_k(\sum_{i=1}^{N}(-1)^i\binom{N}{i}x_{N+j-1-i}^k)
\end{aligned}$$

Mit $0 \leqq k < N$ erhält man durch Anwendung von Hilfssatz 5.2:

$$\begin{aligned}
\sum_{i=1}^{N}(-1)^i\binom{N}{i}x_{N+j-1-i}^k &= \sum_{i=1}^{N}(-1)^i\binom{N}{i}\sum_{m=0}^{k}\binom{k}{m}a^{k-m}(N+j-1-i)^m h^m = \\
&= \sum_{m=0}^{k} h^m\binom{k}{m}a^{k-m}\sum_{i=1}^{N}(-1)^i\binom{N}{i}((N+j-1)-i)^m = \\
&= -\sum_{m=0}^{k}\binom{k}{m}a^{k-m}(N+j-1)^m h^m = \\
&= -x_{N+j-1}^k
\end{aligned}$$

Damit gilt

$$\sum_{i=1}^{N} f_{N+j-1-i}\ q_i = AE_1^{N+j-1}\sum_{k=0}^{N-1} a_k(-x_{N+j-1}^k) = -f_{N+j-1}\ ,$$

was zu zeigen war.

Bemerkung:
Versucht man (5.2) mit $A_i > 0$ für $1 \leqq i \leqq N$ zu lösen ($0 < E_i < E_{i+1}$ gilt nach Definition), dann liegt ein endliches Momentenproblem vor (Gantmacher, [25], S. 211); nach [25] besitzt (5.2) in diesem Fall genau dann eine Lösung, wenn die quadratischen Formen $\sum_{j,k=0}^{N-1} f_{j+k}y_jy_k$, $\sum_{j,k=0}^{N-1} f_{j+k+1}y_jy_k$ positiv definit sind. Läßt sich f als Derichletsche Reihe mit positiven Koeffizienten darstellen, gilt also

$$f(x) = \sum_{n=1}^{\infty} c_n e^{s_n x} \text{ mit } c_n, s_n \in \mathbb{R}, c_i > 0 \text{ für } 1 \leqq i \leqq N \text{ und } c_n \geqq 0,\ n \in \mathbb{N},$$

dann ist (5.2) lösbar, wie in [11] gezeigt ist:

Es gilt $f_i > 0$ für $0 \leqq i \leqq 2N-1$; mit $f_{j+k} = \sum_{n=1}^{\infty} c_n e^{s_n a}e^{(j+k)hs_n}$ gilt

$$\begin{aligned}
\sum_{j,k=0}^{N-1} f_{j+k}y_jy_k &= \sum_{n=1}^{\infty} c_n e^{s_n a}\sum_{j,k=0}^{N-1} e^{jhs_n}y_j e^{khs_n}y_k = \\
&= \sum_{n=1}^{\infty} c_n s^{s_n a}(\sum_{j=0}^{N-1} e^{js_n h}y_j)^2\ ,
\end{aligned}$$

so daß $\sum_{j,k=0}^{N-1} f_{j+k} y_j y_k > 0$ gilt, falls nicht $y_i = 0$, $0 \leqq j \leqq N-1$ erfüllt ist (wegen $c_i > 0$ für $1 \leqq i \leqq N$).
Ebenso folgt die positive Definitheit der zweiten quadratischen Form. Man beachte hierzu die Beispiele von Kapitel 8.

Beispiel 5.1 zeigt, daß die Bestimmung von $\det F_N$ keine Aussage über die Durchführbarkeit des Verfahrens zuläßt:

Beispiel 5.1
Es sei $N = 2$, $X = \{0, 1, 2, 3\}$.

1. $f_0 = 1$, $f_1 = e$, $f_2 = e^2$, $f_3 = e^3$.
Es gilt

$$\det F_2 = \det \begin{pmatrix} e & 1 \\ e^2 & e \end{pmatrix} = 0 \ ,$$

jedoch lassen sich Lösungen von

$$F_2 \begin{pmatrix} q_1 \\ q_2 \end{pmatrix} = - \begin{pmatrix} e^2 \\ e^3 \end{pmatrix}$$

angeben: Die Lösungsmenge des Systems ist gegeben durch:

$$\{ \begin{pmatrix} -r \\ -e^{2+er} \end{pmatrix} \mid r \in R \} \ .$$

Es sei nun $r \in \mathbb{R}$, $r \neq e$, fest gewählt:
Damit folgt $Q(x) = (-e^2 + er)x^2 - rx + 1$; die Nullstellen von Q sind $z_1 = e^{-1}$ und $z_2 = (r-e)^{-1}$. Damit folgt $E_1 = e$ und $E_2 = r - e$. Es gilt $L_{N-1} = L_1 = 1$ und daher $p_1 = -r + e$ und $p_0 = 1$.
Die Gleichung (5.3) lautet hier

$$1 + (e - r)x = A_1 + A_2 + x(e - r)A_1 - A_2 ex$$

Dies ergibt: $A_2 = 1 - A_1$ und $e - r = (e - r)A_1 - (1 - A_1)e$

$$2e - r = A_1(2e - r) \Longrightarrow A_1 = 1 \ , \ A_2 = 0$$

Von Bedeutung ist also nur die Bestimmung von t_1: $t_1 = \ln E_1 = 1$. Man erhält $E(x) = e^x$; dies ist die Lösung des Interpolationsproblems.

2. $f_0 = 1$, $f_1 = -1$, $f_2 = 1$, $f_3 = 1$.
Hierfür gibt es keine Lösung $E \in V_2$ mit $E(i) = f_i$,$1 \leqq i \leqq 3$, da $E \not\equiv 0$ höchstens eine reelle Nullstelle besitzt. Die Gleichungen

$$\begin{aligned} -q_1 + q_2 &= -1 \\ q_1 - q_2 &= -1 \end{aligned}$$

sind unverträglich und es gilt

$$\det F_2 = \det \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} = 0$$

5.2 Konstruktion einer Näherung nach Rice

Es sei in $I = [a, b]$ die äquidistante Punktmenge

$$X = \{x_i \mid x_i \in I, x_i = a + ih, 1 \leqq i \leqq 2N + 1, h > 0\}$$

gegeben.

Unter der Annahme, daß es eine beste Approximation $E \in V_N^0 \setminus V_{N-1}$ an $f \in C(I)$ gibt, wird ein Verfahren zur Bestimmung von E nach Rice, [16], hergeleitet.

Man kann hoffen, auf diesem Wege Näherungen für die beste Approximation an f bezüglich V_N^0 zu erhalten, da diese, sofern sie existiert und die Länge N besitzt, in I eine Alternante der Länge 2N+1 hat.

Es sei also $E(x) = \sum_{i=1}^{N} a_i e^{t_i x} \in V_N^0 \setminus V_{N-1}$ die beste Approximation an f auf X; es gilt daher

$$E(x_i) = f(x_i) + (-1)^i r \qquad 1 \leqq i \leqq 2N + 1 \tag{5.11}$$

$| r |$ ist also die Minimalabweichung von f auf X; eliminiert man r durch Addition aufeinanderfolgender Gleichungen, so erhält man das Gleichungssystem

$$\sum_{i=1}^{N} a_i e^{t_i a}(e^{t_i h} + 1)e^{t_i j h} = f(x_j) + f(x_{j+1}) \qquad 1 \leqq j \leqq 2N \tag{5.12}$$

Die Bestimmung der Koeffizienten und Frequenzen von E aus (5.12) wird nun nach Cesàro, [4], zurückgeführt auf die Ermittlung von Nullstellen eines Polynoms und die anschließende Lösung eines linearen Gleichungssystems Hierzu werden die folgenden Bezeichnungen eingeführt:

$$\begin{aligned} f_j &:= f(x_j) + f(x_{j+1}) && 1 \leqq j \leqq 2N \\ A_i &:= a_i e^{t_i a}(e^{t_i h} + 1), \quad E_i := e^{t_i h} && 1 \leqq i \leqq N \end{aligned}$$

Damit erhält man aus (5.12)

$$\sum_{i=1}^{N} A_i E_i^j = f_j \qquad 1 \leqq i \leqq 2N$$

und da nach Voraussetzung $A_i \neq 0$ für $1 \leqq i \leqq N$ erfüllt ist, gilt für $1 \leqq k \leqq N$ (wegen der linearen Abhängigkeit der ersten Spalte):

$$\det \begin{pmatrix} f_k & E_1^k & \dots & E_N^k \\ f_{k+1} & E_1^{k+1} & \dots & E_N^{k+1} \\ \vdots & \vdots & \dots & \vdots \\ f_{k+N} & E_1^{k+N} & \dots & E_N^{k+N} \end{pmatrix} = 0$$

Wegen $E_i \neq 0$, $1 \leqq i \leqq N$, folgt daher

$$\det \begin{pmatrix} f_k & 1 & \dots & 1 \\ f_{k+1} & E_1^1 & \dots & E_N^1 \\ \vdots & \vdots & \dots & \vdots \\ f_{k+N} & E_1^N & \dots & E_N^N \end{pmatrix} = 0 \qquad 1 \leqq k \leqq N \tag{5.13}$$

Entwickelt man die Determinanten in (5.13) nach der ersten Spalte und ist D_i für $0 \leqq i \leqq N$ die Adjunkte von f_{k+i}, so gilt (da D_i unabhängig ist von k):

$$\sum_{i=0}^{N} f_{k+i} D_i = 0 \qquad 1 \leqq k \leqq N \tag{5.14}$$

Ebenso erhält man

$$\sum_{i=0}^{N} E_j^i D_i = 0 \qquad 1 \leqq j \leqq N \tag{5.15}$$

da in den entsprechenden Matrizen zwei Spalten übereinstimmen.
Aus $E_i < E_{i+1}$ folgt

$$D_0 \neq 0 \qquad D_N \neq 0 \tag{5.16}$$

Je eine Gleichung aus dem System (5.15) ergibt zusammen mit (5.14) ein lineares, homogenes Gleichungssystem mit den Unbekannten D_i, $0 \leqq i \leqq N$:

Es sei $j \in \{1, \dots, N\}$:

$$\begin{aligned} \sum_{i=0}^{N} E_j^i D_i &= 0 \\ \sum_{i=0}^{N} f_{k+i} D_i &= 01 \qquad \leqq k \leqq N \end{aligned} \tag{5.17}$$

Wegen (5.16) folgt hieraus für $1 \leqq j \leqq N$:

$$\det \begin{pmatrix} 1 & E_j^1 & \dots & E_j^N \\ f_1 & f_2 & \dots & f_{N+1} \\ f_2 & f_3 & \dots & f_{N+2} \\ \vdots & \vdots & \dots & \vdots \\ f_N & f_{N+1} & \dots & f_{2N} \end{pmatrix} = 0$$

Dies bedeutet, daß das Polynom

$$P(x) := \det \begin{pmatrix} x^0 & x^1 & \dots & x^N \\ f_1 & f_2 & \dots & f_{N+1} \\ f_2 & f_3 & \dots & f_{N+2} \\ \vdots & \vdots & \dots & \vdots \\ f_N & f_{N+1} & \dots & f_{2N} \end{pmatrix}$$

die N Nullstellen $E_i = e^{t_i h}, 1 \leqq i \leqq N$, besitzt.

Dieses Ergebnis legt folgendes Vorgehen zur Ermittlung der Parameter einer Näherung E nahe:

Man bestimmt die Nullstellen von P; besitzt P die N Nullstellen E_i, $1 \leqq i \leqq N$, und sind diese reell und weiterhin einfach und positiv, dann erhält man die N reellen Frequenzen

$$t_i = h^{-1} \ln E_i \ , \qquad 1 \leqq i \leqq N$$

Mit diesen Werten löst man das lineare Gleichungssystem (5.11) oder (5.12), so daß damit die restlichen Parameter bestimmt sind.

Bemerkung:
Die Lösung von (5.11) ist vorteilhaft, falls $\| f - E \|_X$ von Bedeutung ist, etwa zur Bestimmung einer unteren Schranke für $\| f - E \|_I$.
Da die Koeffizienten von P sich aus der Berechnung von Determinanten ergeben, können hier Rundungsfehler das Ergebnis besonders leicht beeinflussen.

5.3 Konstruktion einer Näherung nach Willers

In [23] behandelt Willers ein Verfahren zur "Annäherung" von Funktionen durch Elemente von V_N^0, falls m Funktionswerte (Messwerte) mit $m \geqq 2N$ gegeben sind.

Für die Konstruktion einer Näherung aus V_N^0 für eine gegebene Funktion $f \in C(I)$, $I = [a, b]$, bietet sich also mit diesem Verfahren die Möglichkeit, die Zahl der Funktionswerte, die in die Rechnung eingehen, zu variieren; mit $m = 2N$ erhält man das in Abschnitt 5.1 beschriebene Verfahren.

Es sei also $m \geqq 2N$ und $X = \{x_i \mid x_i = a + ih, h > 0, 1 \leqq i \leqq m\} \subseteq I$ gegeben. Nimmt man an daß

$$E(x_j) = f(x_j) \qquad 1 \leqq j \leqq m$$

mit $E(x) = \sum_{i=1}^{N} a_i e^{t_i x} \in V_N^0 \setminus V_{N-1}$ erfüllt ist, so erhält man mit $A_i := a_i e^{t_i a}$, $E_i := e^{t_i h}$ für $1 \leqq i \leqq N$ und $f_j := f(x_j)$ für $1 \leqq j \leqq m$:

$$\sum_{i=1}^{N} A_i E_i^j = f_j \qquad 1 \leqq j \leqq m \tag{5.18}$$

Durch $Q(x) := \sum_{j=0}^{N} q_j x^j := \prod_{j=1}^{N} (x - E_j)$ sei q_i für $0 \leqq i \leqq N$ gegeben. Es gilt also

$$Q(E_j) = 0 \qquad 1 \leqq j \leqq N \tag{5.19}$$

Es sei nun $k \in \{1, \ldots, m - N\}$ fest und man betrachte in (5.18) die $N + 1$ Geichungen

$$\sum_{i=1}^{N} A_i E_i^{j+k} = f_{j+k} \qquad 0 \leqq j \leqq N \tag{5.20}$$

Multipliziert man für $0 \leqq j \leqq N$ die (j+1)-te Gleichung von (5.20) auf beiden Seiten mit q_j, so erhält man mit (5.19) durch anschließende Addition der N+1 Gleichungen:

$$\sum_{j=0}^{N} f_{j+k} q_j = \sum_{j=0}^{N} (q_j \sum_{i=1}^{N} (A_i E_i^{k+j}) = \sum_{i=1}^{N} (A_i E_i^k \sum_{j=0}^{N} q_i E_i^j) = 0$$

Da diese Beziehung für $1 \leqq k \leqq m - N$ gilt, erhält man auf diese Weise ein lineares Gleichungssystem mit $m - N$ Gleichungen und N Unbekannten q_i, $0 \leqq i \leqq N - 1$, wegen $q_N = 1$:

$$\sum_{j=0}^{N-1} f_{j+k} q_j = -f_{N+k} \qquad 1 \leqq k \leqq m - N \tag{5.21}$$

Es wird daher folgendes *Verfahren* vorgeschlagen:

Für $m = 2N$ bestimme man die Unbekannten q_i, $0 \leqq i \leqq N-1$, durch Lösen des Gleichungssystems (5.21), falls dieses lösbar ist.

Für $m > 2N$ bestimme man q_i, $0 \leqq i \leqq N - 1$, durch Anwendung der Methode der kleinsten Quadrate.

Hat man so Koeffizienten q_i, $0 \leqq i \leqq N - 1$, gefunden, so bestimmt man die Nullstellen E_i, $1 \leqq i \leqq N$ von $\sum_{i=0}^{N} q_i x^i$ mit $q_N = 1$.
Gilt $E_i \in \mathbb{R}$ und $E_i > 0$ für $1 \leqq i \leqq N$, dann erhält man die N Frequenzen

$$t_i = h^{-1} \ln E_i \qquad 1 \leqq i \leqq N$$

Sind die Nullstellen des Polynoms einfach, gilt also $t_i \neq t_j$ für $i \neq j$, dann bestimmt man a_i für $1 \leqq i \leqq N$ durch Lösung eines linearen Gleichungssystems, das aus N Gleichungen von (5.18) besteht.

5.4 Numerische Beispiele

Zur numerischen Erprobung der drei beschriebenen Verfahren sind mit der CD 3300 des Rechenzentrums der Universität Erlangen-Nürnberg Berechnungen durch geführt worden; die Programme sind in FORTRAN IV geschrieben und es ist stets mit doppelter Genauigkeit gerechnet worden. Alle Zahlenangaben sind daher gerundete Größen.
Anmerkung (Februar 2023):
Diese Berechnungen wurden nochmals mit MEINARDUS [27] durchgeführt.

In den folgenden Beispielen werden die Bezeichnungen benutzt, wie sie bei der Herleitung der Verfahren eingeführt worden sind.

Für $1 \leqq N \leqq 4$ treten mit $f(x) = \dfrac{1}{x+1}$, $f(x) = \sqrt{x}$ in $I = [0,1]$ oder der Riemannschen Zetafunktion in $I = [2,3]$, $I = [2,4]$ keine Schwierigkeiten auf: Es wurden keine Punktmenge X gefunden, so daß hier ein Verfahren gescheitert wäre. Die Güte der jeweiligen Näherung E, d.h. $\| f - E \|_I$, hängt wesentlich von der Menge X ab und man wird daher im allgemeinen, um gute Näherungen zu erhalten, verschiedene Punktmengen X verwenden. In den Zeichnungen sind stets die jeweiligen Fehlerfunktionen $f - E$ dargestellt.

Beispiel 5.2
Es sei $f(x) = \sqrt{x}$, $I = [0,1]$ und N=2.

Zu bestimmen ist also eine Näherung $E(x) = a_1 e^{t_1 x} + a_2 e^{t_2 x}$; zur Beurteilung der folgenden Ergebnisse beachte man die Resultate von [26], §7.

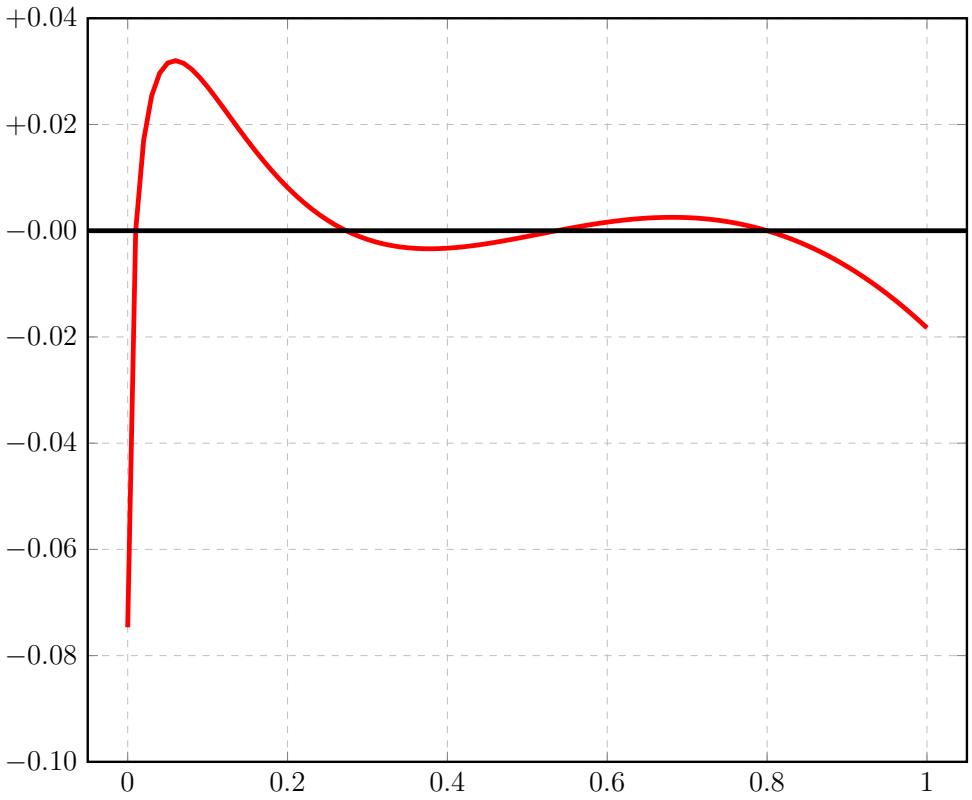
Eine günstige Näherung liefert das **Verfahren von Meinardus**, Abschnitt 5.1, mit den Punkten $x_i = 0.01 + ih$, $0 \leqq i \leqq 3$, wobei $x_3 = 0.800$ und damit $h = 0.2633$ ist. Man erhält die Polynomkoeffizienten

```
q2=+0.345 112 153 , q1=-1.467 229 623 , q0=+1.000 000 000,
p1=+0.376 089 942 , p0=+1.000 000 000 .
```

Die Nullstellen von Q sind 3.398 958 380 und 0.852 499 044 .
Es ergeben sich die Parameter

```
a1=-0.483 373 717 , t1=-4.646 084 908
a2= 0.558 037 443 , t2= 0.606 012 117
```

Zeichnung 5.1 stellt die Fehlerfunktion dar. Die Norm der Fehlerfunktion ist 0.07466.

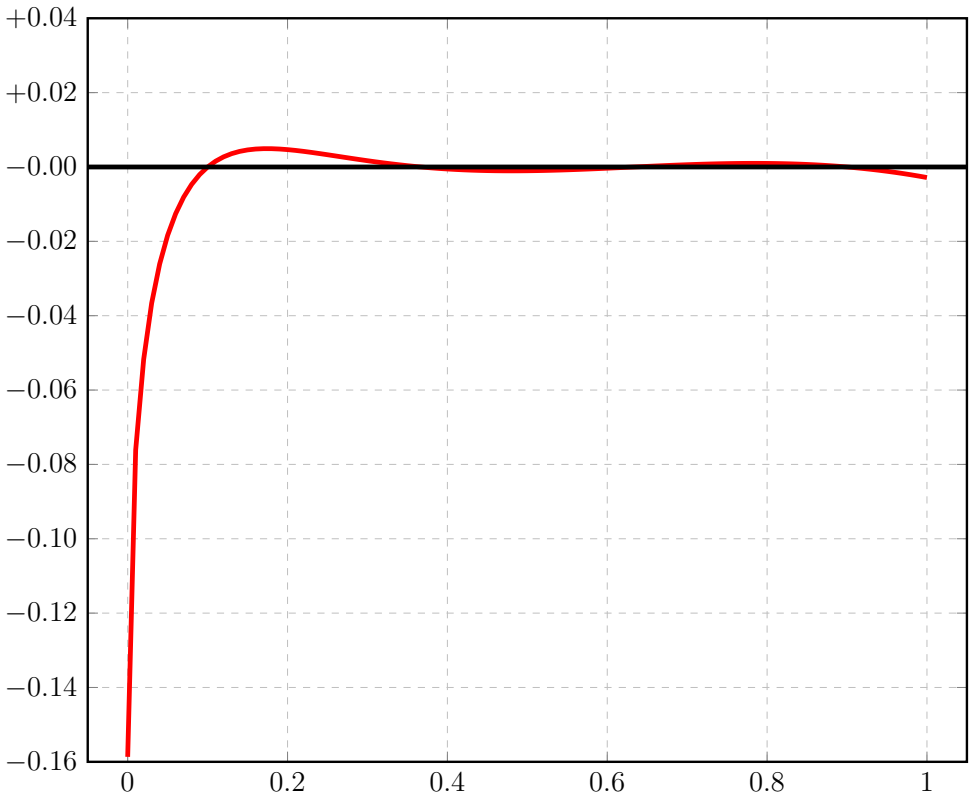


Zeichnung 5.1

Mit $x_0 = 0.1$, $x_3 = 0.9$, $h = 0.2666$ erhält man

```
a1=-0.496 967 379 , t1=-2.961 059 561
a2=+0.655 627 088 , t2=+0.450 327 955
```

Zeichnung 5.2 stellt die Fehlerfunktion dar. Die Norm der Fehlerfunktion auf I ist hier 0.15866, womit die Abhängigkeit von X demonstriert sei.



Zeichnung 5.2

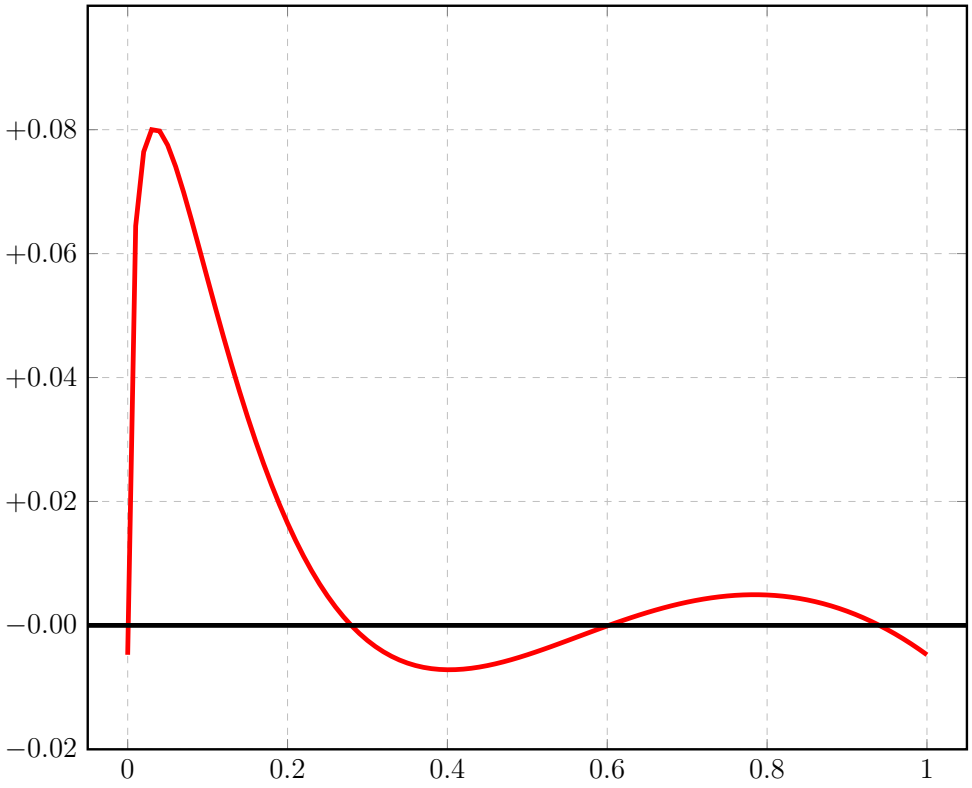
Das **Verfahren von Rice**, Abschnitt 5.2, ergibt mit $x_i = \dfrac{i-1}{4}$, $1 \leqq i \leqq 5$, das Polynom $P(x) = p_2x^2 + p_1x + p_0$ mit

```
p2=-0.670 540 689 , p1=+0.965 925 826 , p0=-0.222 252 952
```

und den Nullstellen 1.153 063 231, 0.287 454 610. Man erhält so die Parameter

```
a1=-0.565 826 602 , t1=-4.986 761 225
a2=+0.570 571 292 , t2=+0.569 688 323
```

Zeichnung 5.3 stellt die Fehlerfunktion dar.



Zeichnung 5.3

Für $1 \leqq i \leqq 5$ gilt

$$|\sqrt{x} - a_1 e^{t_1 x_i} - a_2 e^{t_2 x_i}| = 0.004\ 744\ 688$$

Damit ist gezeigt, daß, obwohl $x_1 = 0.0$ und $x_5 = 1.0$ Alternantenpunkte der Fehlerfunktion für die beste Approximation an f bezüglich V_2 auf I sind, die Näherung nach Rice, die ja die beste Approximation auf $\{x_i \mid 1 \leqq i \leqq 5\}$ für F ist, in x_1 und x_5 wesentlich von der Minimallösung für f bezüglich V_2 verschieden sein kann.

Mit dem **Verfahren von Willers**, Abschnitt 5.3, kann man oft bessere Näherungen für die Frequenzen erhalten; es werden m Punkte

$$x_i = 0.0 + (i-1) * h \ , \ 1 \leqq i \leqq m \text{ mit } x_m = 1.0$$

verwendet.

m = 10:
Es ist also $h = \frac{1}{9}$; man errechnet die Koeffizienten

`q2=+1.000 000 000 , q1=-1.456 251 707 , q0=+0.406 468 037`

und damit die Nullstellen `1.079 834 854` und `0.376 414 853` .
Die Frequenzen sind also `t1=-8.793 522 892 , t2=+0.691 273 052` .

Durch Lösung des Systems

$$\sum_{i=1}^{2} a_i e^{t_i x_k} = \sqrt{x_k} \qquad k = 2, 3$$

folgt

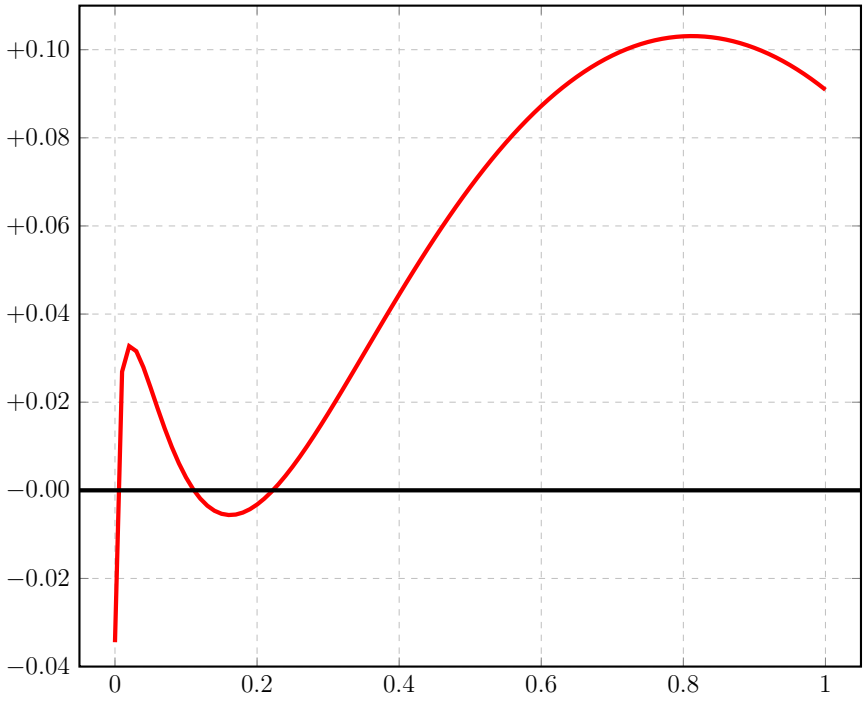
`a1=-0.420 954 177 , a2=+0.455 428 511`

Zeichnung 5.4 stellt die Fehlerfunktion dar.

Durch Vergrößern von m erhält man nicht unbedingt bessere Werte:

m	t1	t2
22	-15.928 838	+0.799 146
32	-21.251 001	+0.838 908

Anmerkung (Februar 2023):
Nach Neu-Implementierung des Verfahrens von Willers werden Zeichnungen der zugehörigen Fehlerfunktionen der Vollständigkeit halber nachgetragen.



Zeichnung 5.4

Schwierigkeiten ergeben sich mit $X \subseteq [-1, +1]$ für die Funktionen

- $f(x) = |x|$ bei $N \geqq 3$
- $f(x) = \begin{cases} -5 & : x \leqq -0.5 \\ 10x & : x \in (-0.5, +0.5) \\ +5 & : x \geqq +0.5 \end{cases}$ bei $N \geqq 2$
- $f(x) = \begin{cases} 2 + x & : x \leqq 0 \\ 2 - x & : x \geqq 0 \end{cases}$ bei $N \geqq 2$

Für die letzte Funktion wird dies nachfolgend in Beispiel 5.3 ausgeführt.

Beispiel 5.3

$I = [-1, +1]$, $N = 2$, $f(x) = \begin{cases} 2 + x & : x \leqq 0 \\ 2 - x & : x \geqq 0 \end{cases}$

Bei der **Berechnung nach Meinardus** von Abschnitt 5.1 erhält man für Q meist komplexe Nullstellen, wie die folgende Tabelle zeigt[12]:

		Die Nullstellen von Q	
x_0	x_3	*Realteil* $\Re$	*Imaginärteile* $\Im$
-1.0	1.0	0.8000	$\pm$0.6000
-0.9	0.9	0.8235	$\pm$0.5673
-0.8	0.8	0.8462	$\pm$0.5329
-0.7	0.7	0.8679	$\pm$0.4967
-0.5	0.5	0.9091	$\pm$0.4166
-0.8	0.7	0.7282	$\pm$0.4225
-0.7	0.8	1.0274	$\pm$0.5962
-0.5	0.4	0.6986	$\pm$0.2149

Mit $\mathbf{x_0 = -0.5}$ und $\mathbf{x_3 = +0.35}$ jedoch ist das Verfahren durchführbar und man erhält die Polynomkoeffizienten

```
q2=+2.837 462 834 , q1=-3.470 763 132 , q0=+1.000 000 000
p1=-3.422 811 364 , p0=+1.500 000 000
```

und damit die Nullstellen `0.758 641 874` und `0.464 550 581` .

Es folgt damit

```
a1=+2.820 549 263 , t1=+0.974 913 359
a2=-0.898 915 427 , t2=+2.705 946 470
```

Zeichnung 5.5 stellt die Fehlerfunktion dar:

- auf Intervall I
- auf Teilintervall $[-1.0, +0.4]$.

Die Minimalabweichung für f in I bezüglich V_1 ist 0.5 [13]: diese Näherung approximiert also schlechter als die beste Approximation bezüglich V_1.

[12] x_0/x_3 ist der erste/letzte der äquidistanten Punkte

[13] die beste Approximation ist die konstante Funktion $E(x) \equiv 1.5 \in V_1$

Weitere Ergebnisse:

- Mit $\mathbf{x_0 = -0.5}$, $\mathbf{x_3 = +0.3}$ erhält man

  ```
  a1=+2.138 865 180 , t1=-0.692 894 768
  a2=-0.185 691 049 , t2=+5.381 257 439
  ```

  Zeichnung 5.6 stellt die Fehlerfunktion dar:

  - auf Intervall I
  - auf Teilintervall $[-0.2, +0.2]$.

  Die Norm der Fehlerfunktion ist größer als 37.0 .

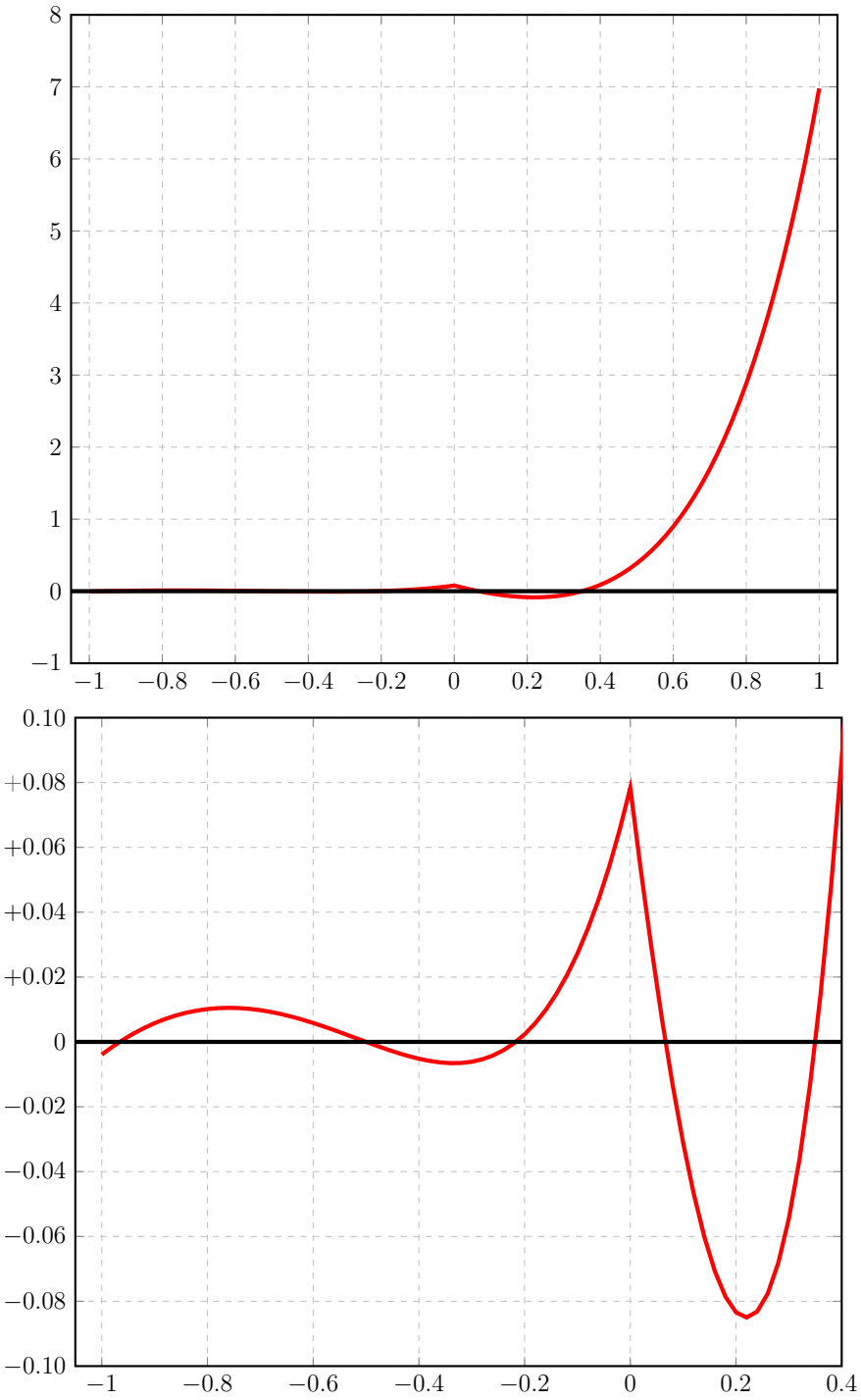
- Mit $\mathbf{x_0 = -0.5}$, $\mathbf{x_3 = +0.25}$ erhält man

  ```
  a1=+2.050 428 834 , t1= +0.624 759 329
  a2=-0.050 428 834 , t2=+10.207 441 475
  ```

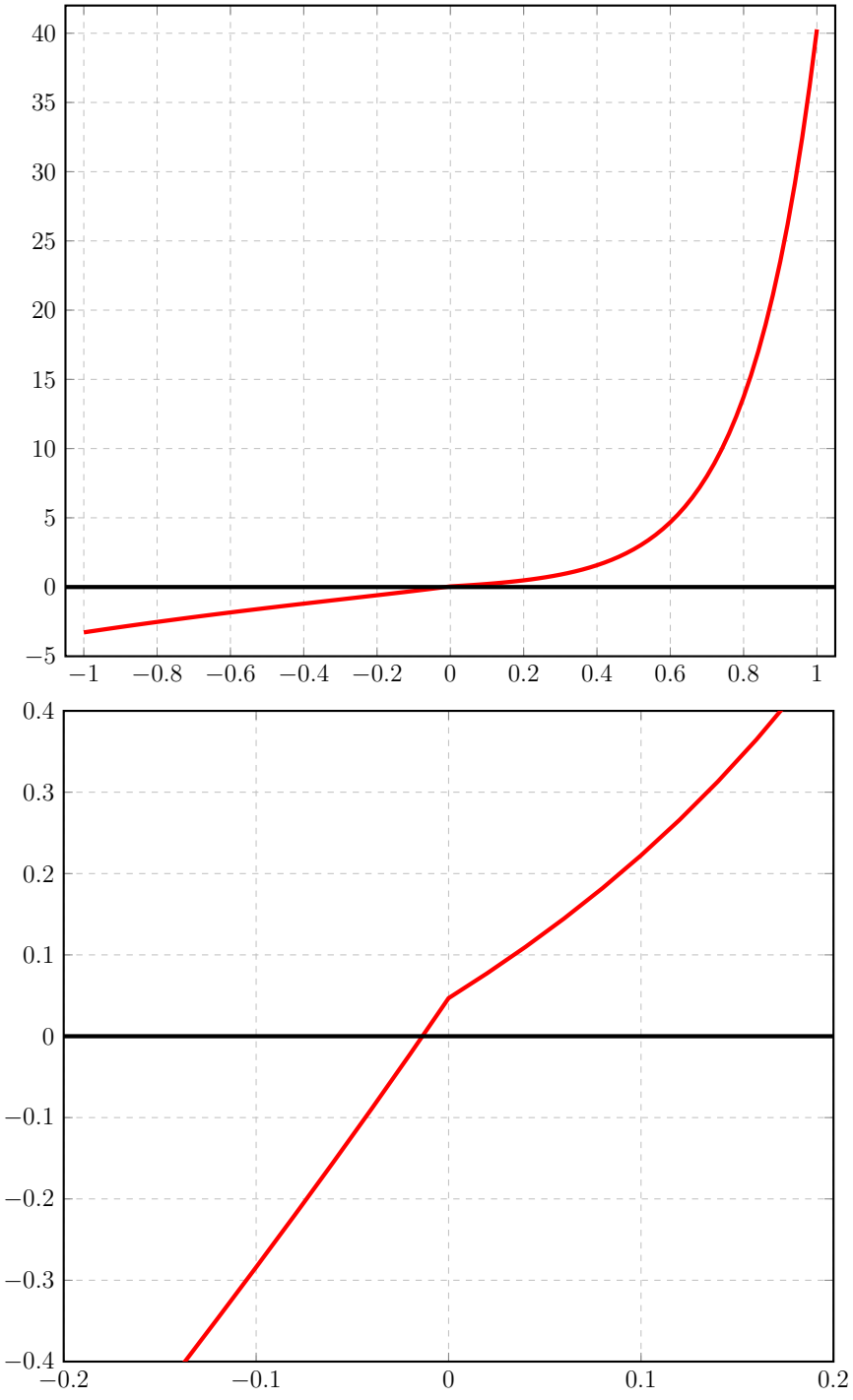
  Zeichnung 5.7 stellt die Fehlerfunktion dar

  - auf Intervall I
  - auf Teilintervall $[-1.05, +0.3]$.

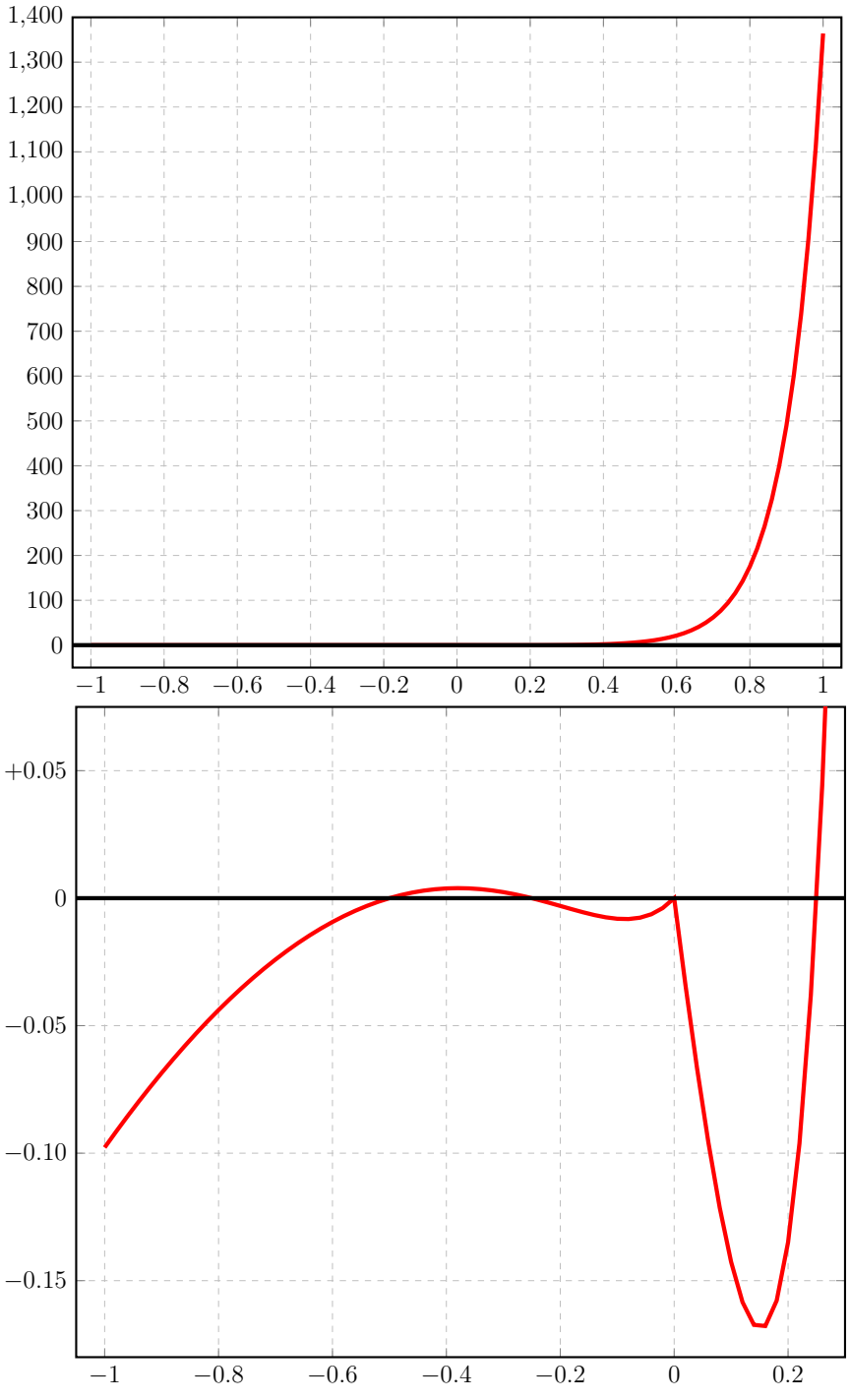
Für diese drei Fehlerfunktionen ist $x = 1.0$ der Extremalpunkt.



Zeichnung 5.5



Zeichnung 5.6



Zeichnung 5.7

Bei der Ermittlung einer **Näherung nach Rice**, Abschnitt 5.2, treten die gleichen Schwierigkeiten auf, da auch P meist komplexe Nullstellen besitzt:

		Die Nullstellen von P	
x_1	x_5	*Realteil* $\Re$	*Imaginärteile* $\Im$
-1.0	1.0	0.8571	$\pm$0.5151
-0.9	0.9	0.8732	$\pm$0.4873
-0.8	0.8	0.8888	$\pm$0.4581
-0.5	0.5	0.9333	$\pm$0.3590

Für $x_1 = -0.5$ und $x_5 = -0.25$ erhält man das Polynom $P(x) = \sum_{i=0}^{2} p_i x^i$ mit

```
p2=-0.539 062 500 , p1=-1.828 125 000 , p0=-1.398 437 500
```

und den Nullstellen 2.225 778 005, 1.165 526 343 . Damit folgt

```
a1=+2.314 520 833 , t1=+0.816 921 503
a2=-0.376 853 596 , t2=+4.267 234 771
```

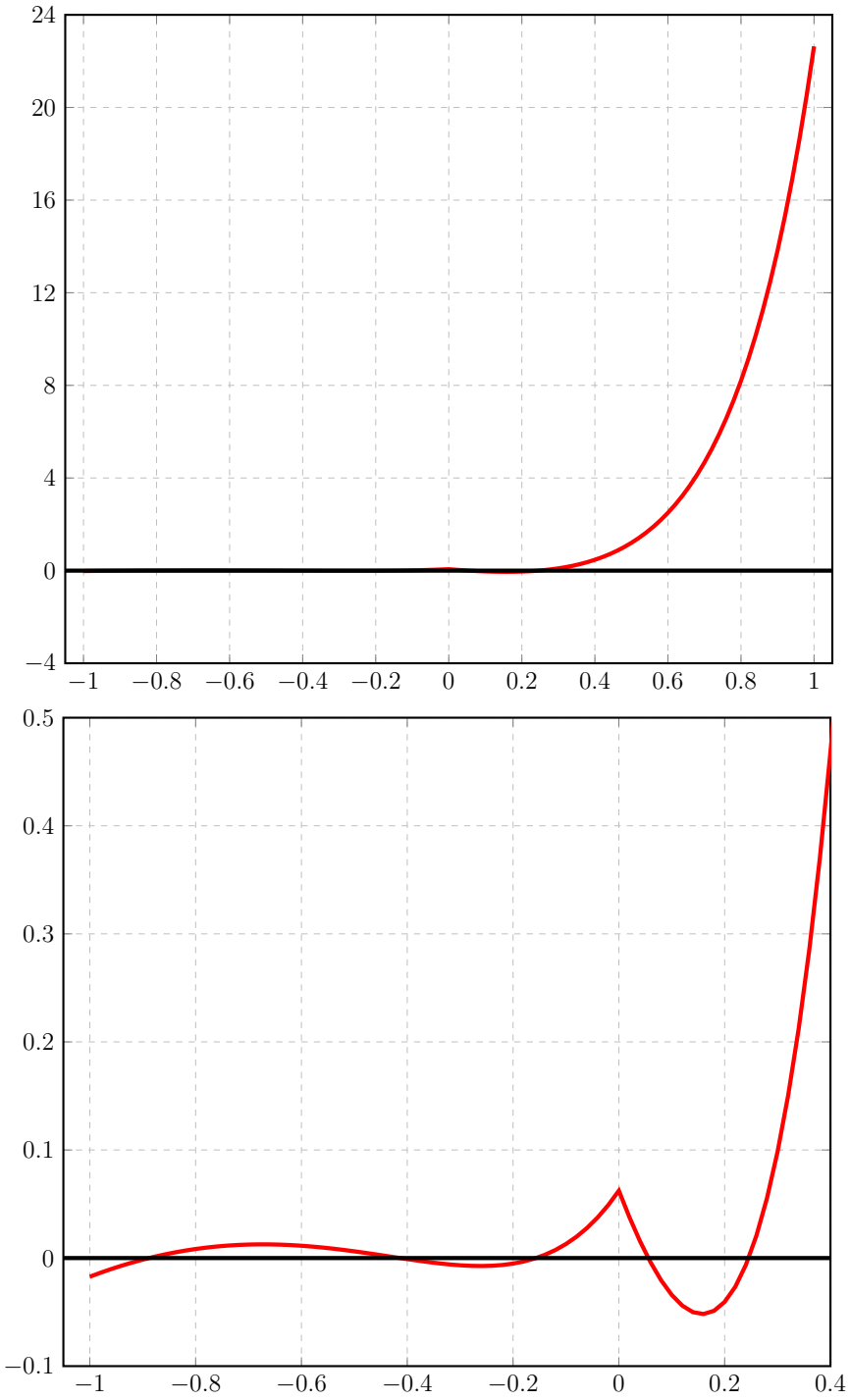
Zeichnung 5.8 stellt die Fehlerfunktion dar:

- auf Intervall I
- auf Teilintervall $[-1.0, +0.4]$.

Auch die Berechnung einer **Näherung nach Willers**, Abschnitt 5.3, mit verschiedenen Werten für m führt auf komplexe Nullstellen; es sei $x_1 = -1.0$ und $x_m = -1.0$:

	Die Nullstellen von Q	
m	*Realteil* $\Re$	*Imaginärteile* $\Im$
10	0.9329	$\pm$0.1892
22	0.9740	$\pm$0.0831
32	0.9828	$\pm$0.0567

Die hier auftretenden Schwierigkeiten dürften darin begründet sein, daß es für f in [-1,+1] nach Satz 3.3 zwei Minimallösungen E_1, E_2 bezüglich V_2 gibt und daher keine beste Approximation bezüglich V_2^0 existiert; es liegt also nahe zu vermuten, daß $f - E_i$, $i = 1, 2$, in I keine Alternante der Länge 5 besitzt, was sich besonders auf die Verfahren von Meinardus, Abschnitt 5.1, und Willers, Abschnitt 5.2, auswirkt.



Zeichnung 5.8

Beispiel 5.4 Näherungen für die Riemannsche Zetafunktion [14]

Es werden Näherungen für die Riemannsche Zetafunktion nach dem **Verfahren von Meinardus**, Abschnitt 5.1, für $1 \leqq N \leqq 4$ jeweils auf den Intervallen $[2,3]$, $[2,4]$ angegeben.[15]

Für $I = [2,3]$, $N = 1$ erhält man mit $\mathbf{x_0} = \mathbf{2.333}$, $\mathbf{x_1} = \mathbf{2.666}$ die Parameter

```
a1=+2.793 905 734 , t1=-0.291 505 518
```

Zeichnung 5.9 stellt die Fehlerfunktion dar [16].

Zeichnung 5.9

[14] Dieses Beispiel ist eine Erweiterung der Diplomarbeit [26], §5.

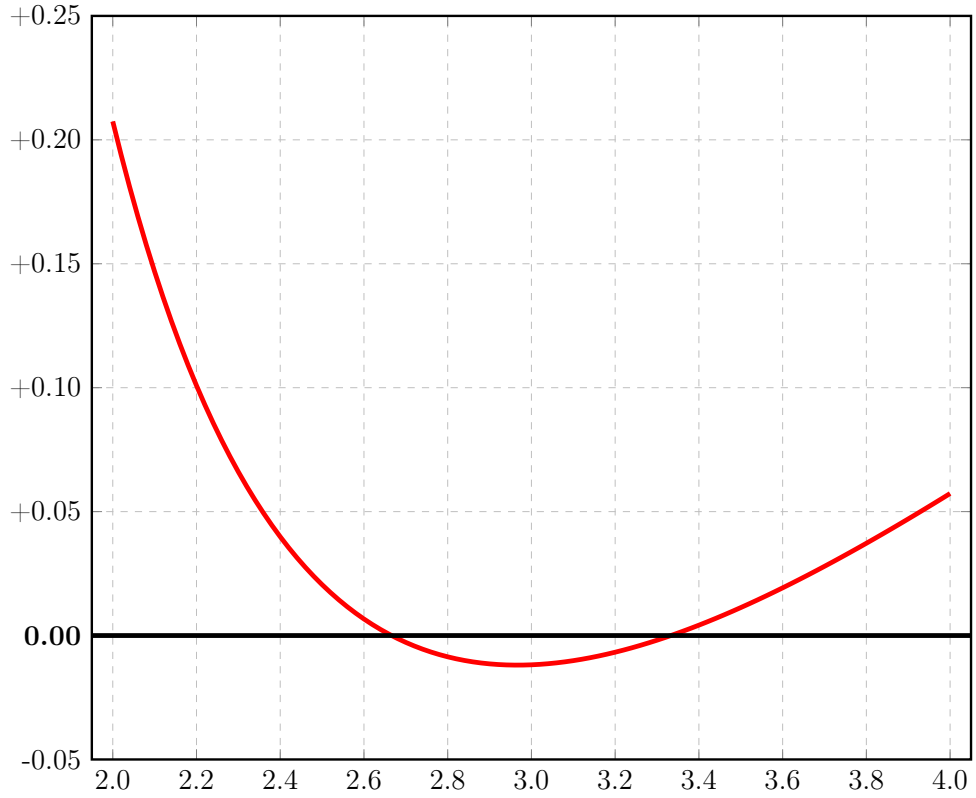
[15] Es sind dies die Startfunktionen für die Berechnungen von Minimallösungen für die Zetafunktion von [26], §8

[16] Zeichnung 48 von [26], §8

Für $I = [2, 4]$, $N = 1$ erhält man mit $\mathbf{x_0} = \mathbf{2.666}$, $\mathbf{x_1} = \mathbf{3.333}$ die Parameter

```
a1=+2.015 963 567 , t1=-0.169 096 199
```

Zeichnung 5.10 stellt die Fehlerfunktion dar [17].



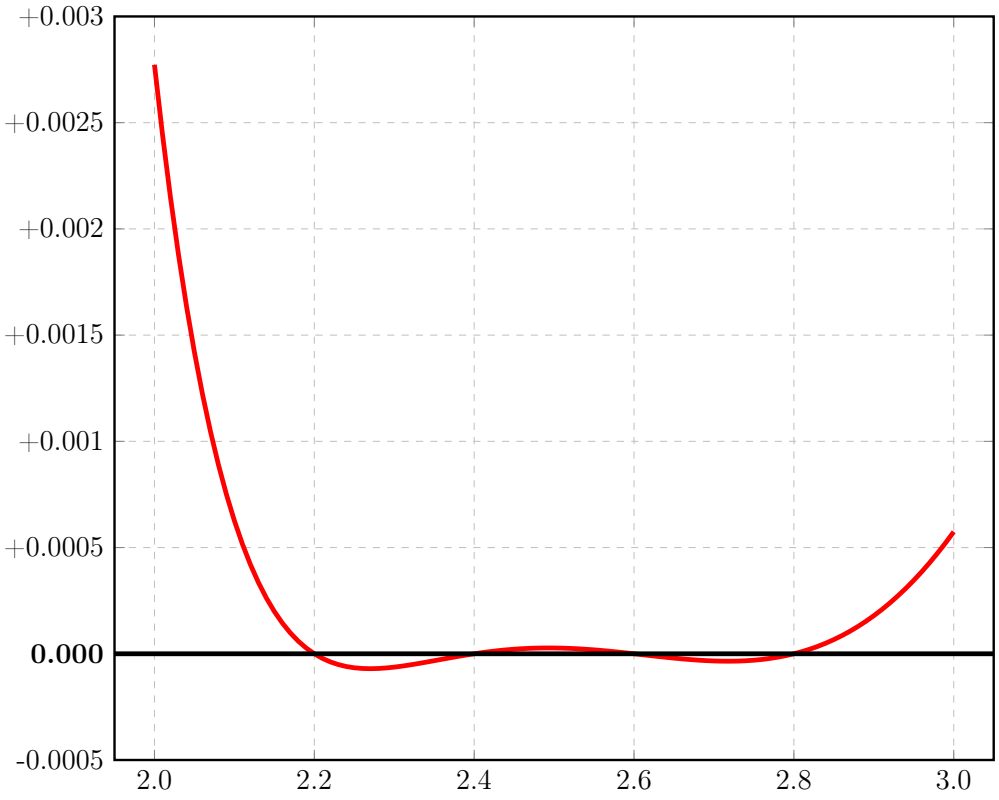
Zeichnung 5.10

[17] Zeichnung 52 von [26], §8

Für $I = [2, 3]$, $N = 2$ erhält man mit $\mathbf{x_0} = \mathbf{2.200}$, $\mathbf{x_3} = \mathbf{2.800}$ die Parameter

```
a1= +1.525 165 278 , t1=-0.092 521 466
a2=+24.876 184 782 , t2=-2.097 844 336
```

Zeichnung 5.11 stellt die Fehlerfunktion dar [18].



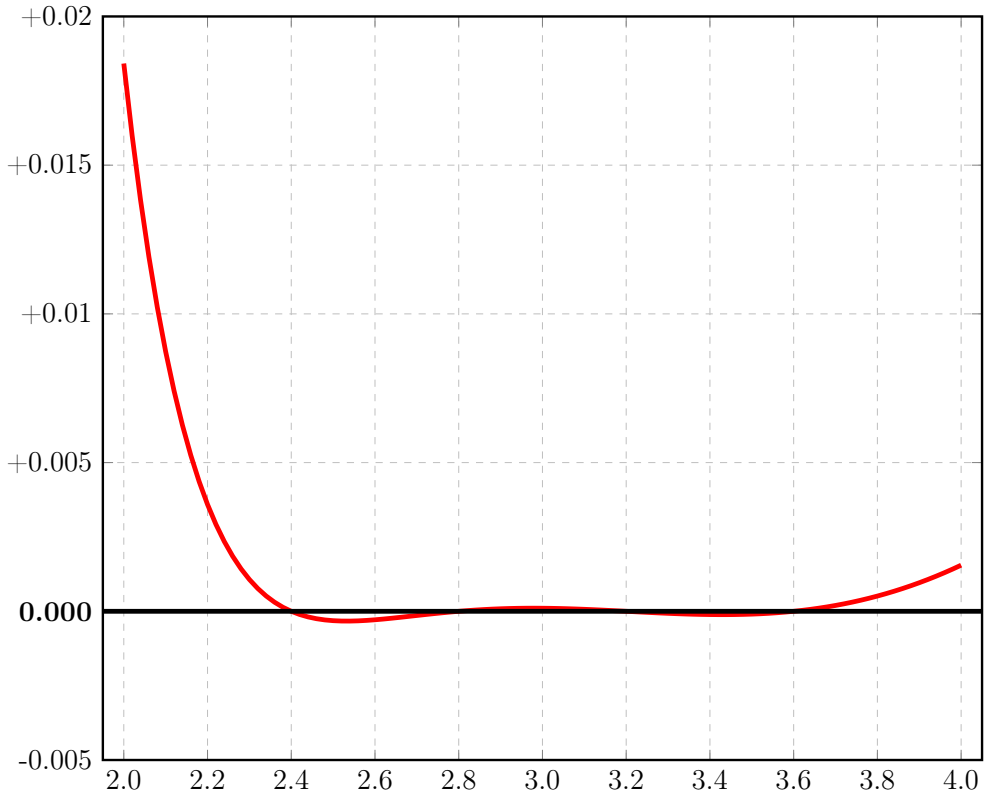
Zeichnung 5.11

[18] Zeichnung 56 von [26], §8

Für $I = [2, 4]$, $N = 2$ erhält man mit $\mathbf{x_0} = \mathbf{2.400}$, $\mathbf{x_3} = \mathbf{3.600}$ die Parameter

```
a1= +1.262 562 853 , t1=-0.043 145 667
a2=+11.959 298 968 , t2=-1.620 042 255
```

Zeichnung 5.12 stellt die Fehlerfunktion dar [19].



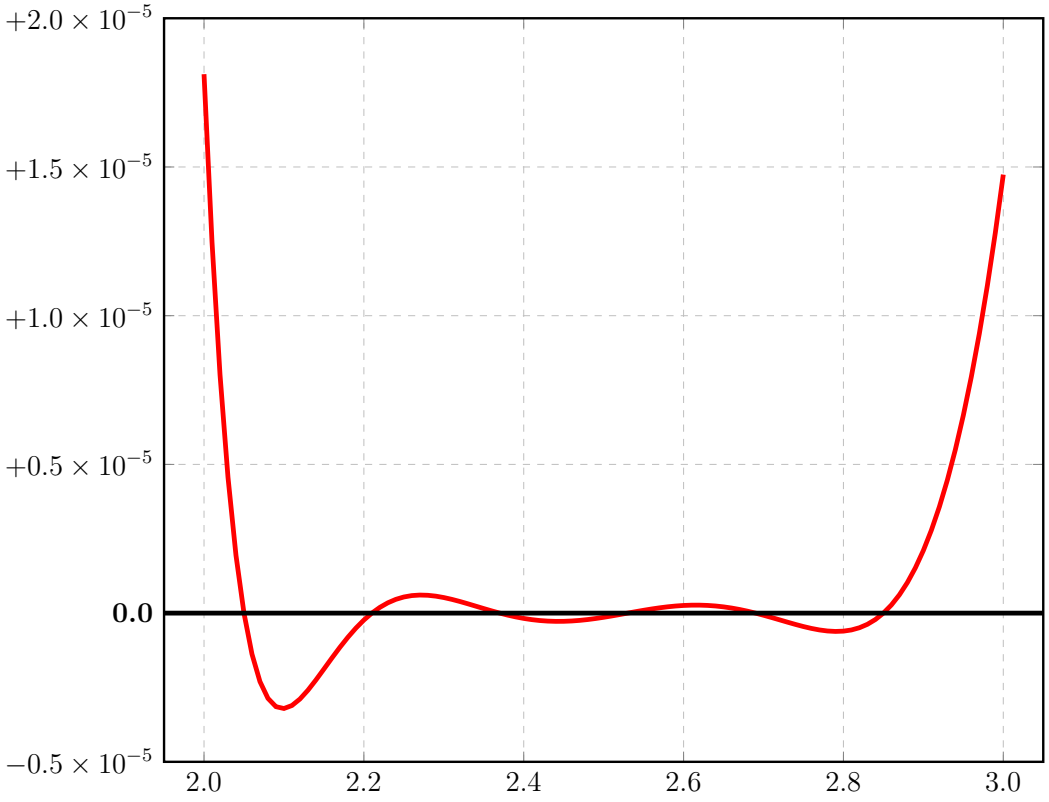
Zeichnung 5.12

[19] Zeichnung 62 von [26], §8

Für $I = [2, 3]$, $N = 3$ erhält man mit $\mathbf{x_0} = \mathbf{2.050}$, $\mathbf{x_5} = \mathbf{2.850}$ die Parameter

```
a1=  +1.236 018 240 , t1=-0.039 532 424
a2=  +8.421 439 207 , t2=-1.466 241 490
a3=+301.455 919 776 , t3=-4.310 972 926
```

Zeichnung 5.13 stellt die Fehlerfunktion dar [20].



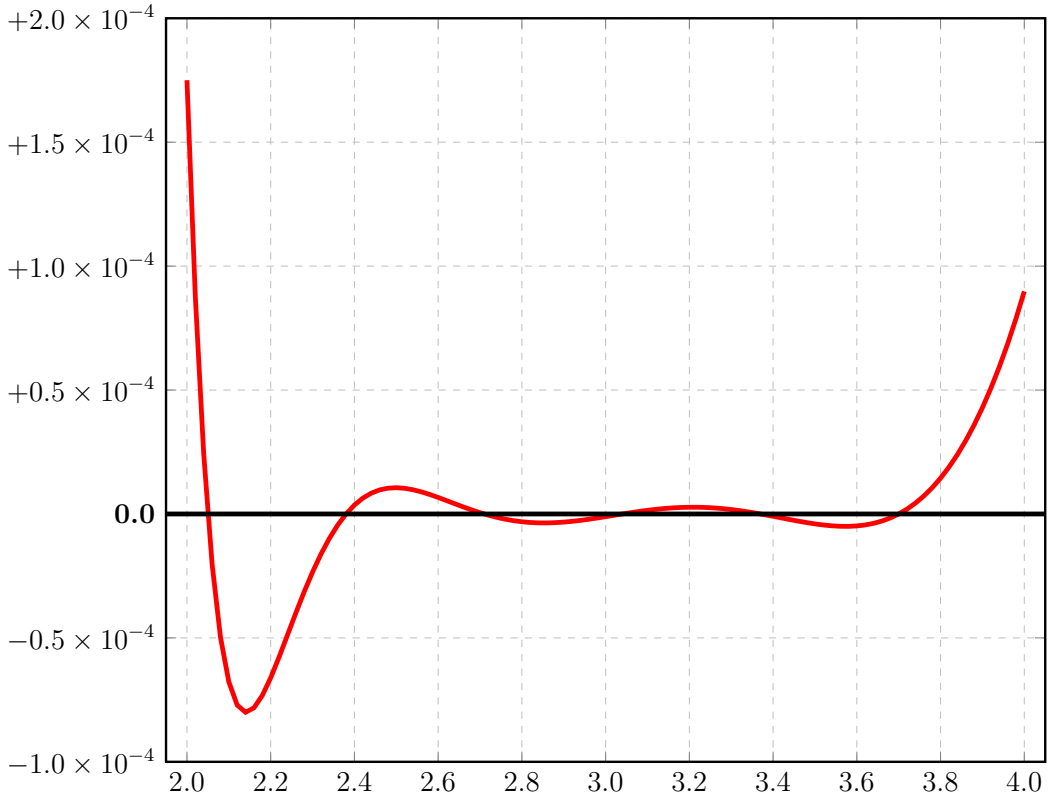
Zeichnung 5.13

[20] Zeichnung 69 von [26], §8

Für $I = [2, 4]$, $N = 3$ erhält man mit $\mathbf{x_0} = \mathbf{2.050}$, $\mathbf{x_5} = \mathbf{3.700}$ die Parameter

```
a1=  +1.112 198 403 , t1=-0.017 032 105
a2=  +5.076 875 370 , t2=-1.191 695 146
a3=+113.053 478 781 , t3=-3.507 624 270
```

Zeichnung 5.14 stellt die Fehlerfunktion dar [21].



Zeichnung 5.14

[21] Zeichnung 75 von [26], §8

Für $I = [2, 3]$, $N = 4$ erhält man mit $\mathbf{x_0} = \mathbf{2.100}$, $\mathbf{x_7} = \mathbf{2.800}$ die Parameter

```
a1=   +1.108 917 301 , t1=-0.016 846 093
a2=   +4.516 500 699 , t2=-1.151 831 800
a3=  +56.338 107 956 , t3=-3.097 712 958
a4=+3183.222 936 114 , t4=-6.533 155 909
```

Zeichnung 5.15 stellt die Fehlerfunktion dar:

- auf Intervall I
- auf Teilintervall $[+2.1, +2.8]$.

(Entsprechen Zeichnungen 82, 82A von [26], §8)

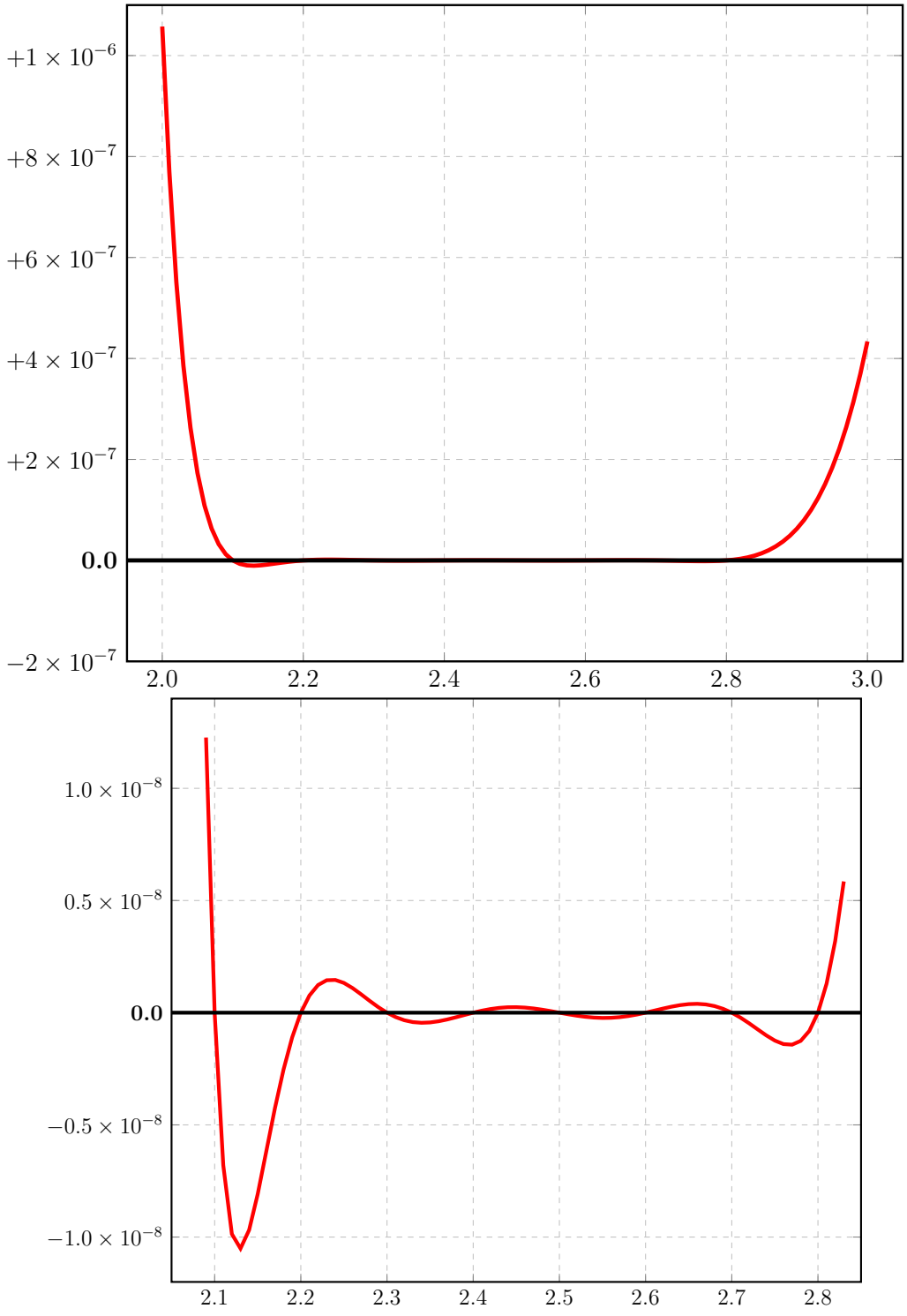
Für $I = [2, 4]$, $N = 4$ erhält man mit $\mathbf{x_0} = \mathbf{2.050}$, $\mathbf{x_7} = \mathbf{3.600}$ die Parameter

```
a1=   +1.048 520 425 , t1=-0.006 758 602
a2=   +3.067 446 915 , t2=-0.980 260 258
a3=  +27.094 443 001 , t3=-2.561 189 494
a4= +928.206 585 131 , t4=-5.450 993 513
```

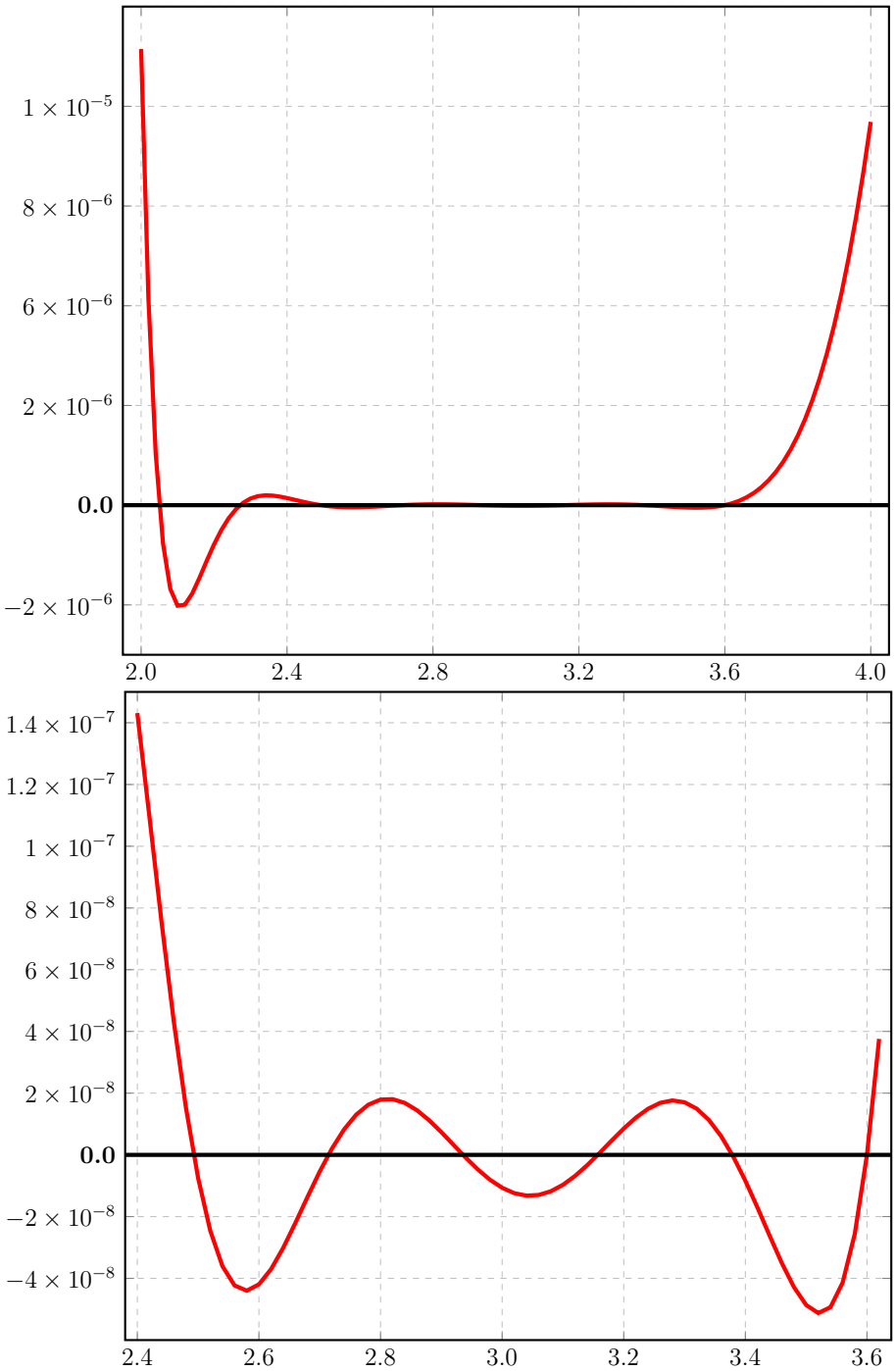
Zeichnung 5.16 stellt die Fehlerfunktion dar:

- auf Intervall I
- auf Teilintervall $[+2.4, +3.6]$.

(Entsprechen Zeichnungen 91, 91A von [26], §8)



Zeichnung 5.15



Zeichnung 5.16

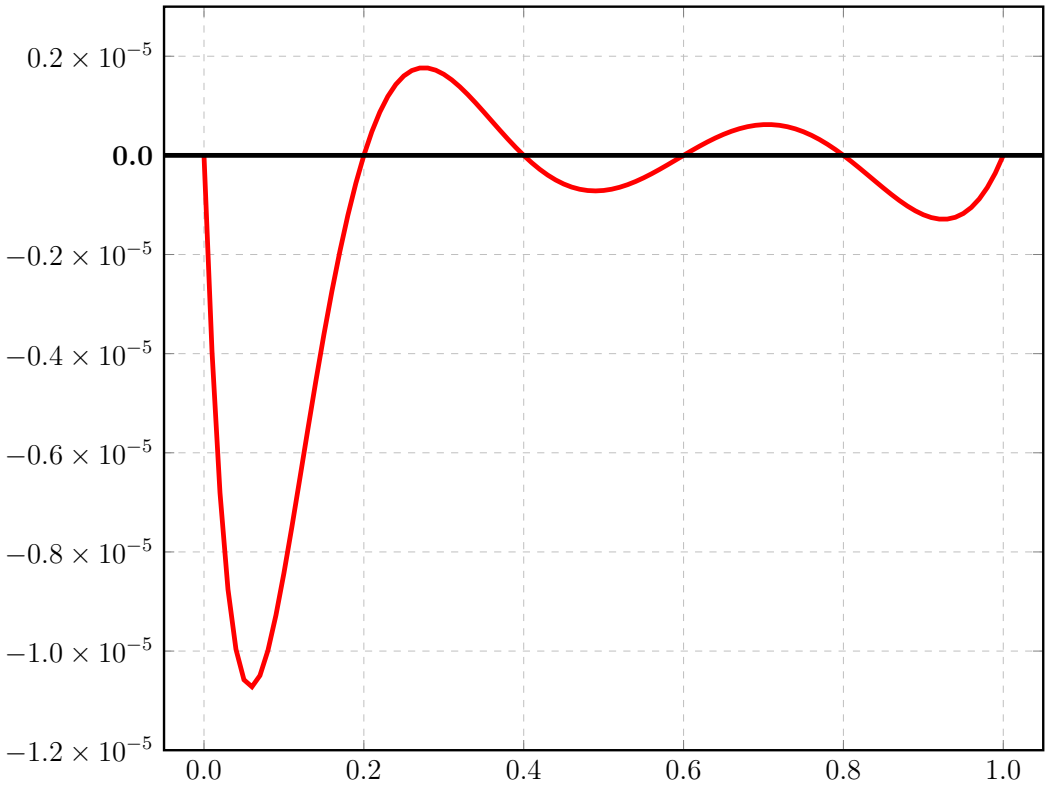
Beispiel 5.5 Näherungen für $\mathbf{f(x) = \frac{1}{1+x}}$ [22]

Es werden zwei Näherungen für $f(x) = \frac{1}{1+x}$ nach dem **Verfahren von Meinardus** von Abschnitt 5.1 für $N = 3$ auf $I = [0, 1]$ mit unterschiedlichen Werten für x_0, x_5 angegeben.[23]

Mit $\mathbf{x_0 = 0.0}$, $\mathbf{x_5 = 1.0}$ erhält man die Parameter

```
a1=+0.546 271 421 , t1=-0.277 576 311
a2=+0.401 383 367 , t2=-1.546 985 019
a3=+0.052 345 210 , t3=-4.336 157 075
```

Zeichnung 5.17 stellt die Fehlerfunktion dar [24].



Zeichnung 5.17

[22] Dieses Beispiel ist eine Erweiterung der Diplomarbeit [26], §5.

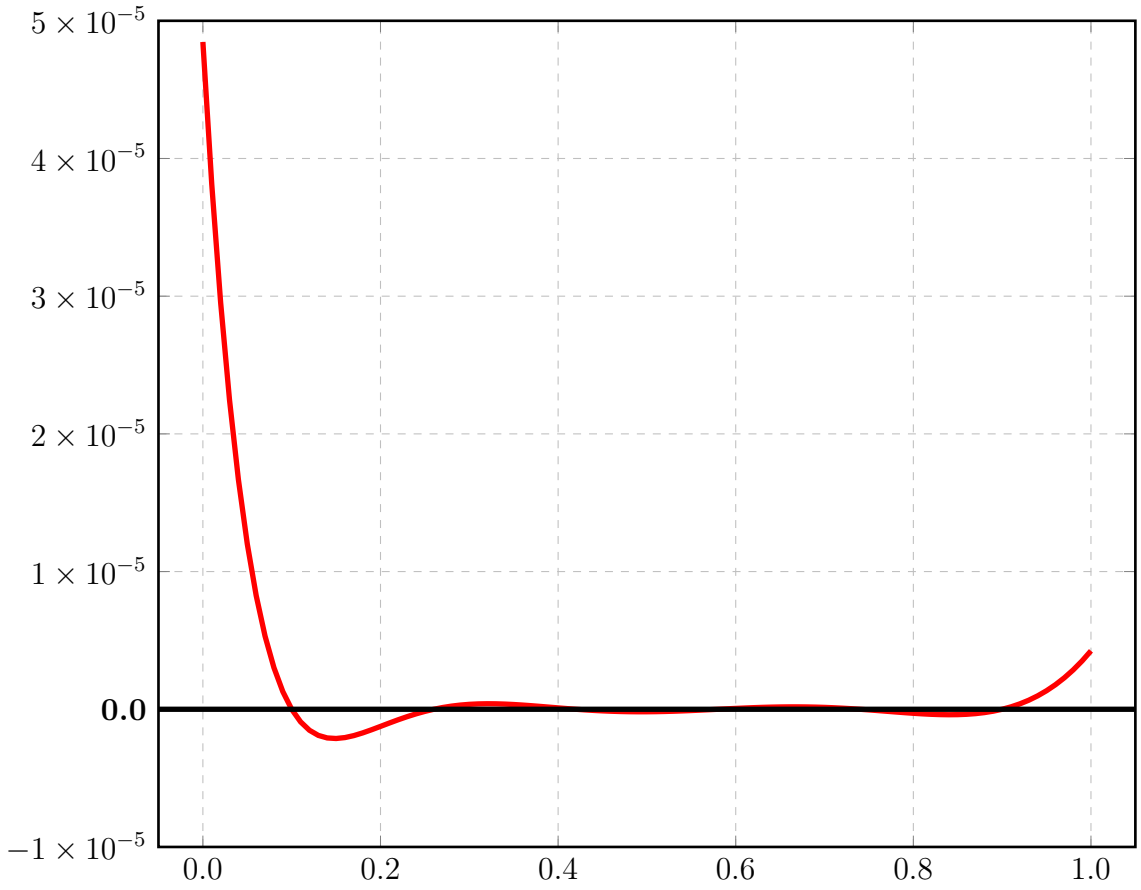
[23] Es sind dies die Startfunktionen für die Berechnungen von Minimallösungen in [26], §7

[24] Zeichnung 27 von [26], §7

Mit $\mathbf{x_0} = \mathbf{0.1}$, $\mathbf{x_5} = \mathbf{0.9}$ erhält man die Parameter

```
a1=+0.545 644 214 , t1=-0.277 434 496
a2=+0.400 482 842 , t2=-1.540 603 343
a3=+0.053 824 480 , t3=-4.282 493 872
```

Zeichnung 5.18 stellt die Fehlerfunktion dar [25].



Zeichnung 5.18

[25] Zeichnung 36 von [26], §7

6 Iterationsverfahren

6.1 Konstruktion von Minimallösungen nach Braess

Es sei $I \subseteq \mathbb{R}$ ein kompaktes Intervall, $N \in \mathbb{N}$, $f \in C(I)$.
In [3] behandelt Braess einen Algorithmus, der, von einer Minimallösung E bezüglich V_{N-1}^0 ausgehend, eine Folge besserer Approximationen $((E(a_n))_{n\in\mathbb{N}}$ liefert, falls E nicht die beste Approximation bezüglich V_N ist; zur Konstruktion einer Minimallösung bezüglich V_N wird mit diesem Algorithmus sukzessive die Minimallösung bezüglich V_1^0, V_2^0, ... ermittelt.

Vereinbarung:
In diesem Abschnitt wird die Parametrisierung von V_N^0 nach Definition 1.2 benutzt und es ist zweckmäßig, die folgende Schreibweise einzuführen:

Für $u = \begin{pmatrix} u_1 \\ \vdots \\ u_m \end{pmatrix} \in \mathbb{R}^m$ und $v = \begin{pmatrix} v_1 \\ \vdots \\ u_n \end{pmatrix} \in \mathbb{R}^n$ mit $m \geqq n$ sei $u+v$ gegeben durch

$$u + v := \begin{pmatrix} u_1 + v_1 \\ \vdots \\ u_n + v_n \\ \vdots \\ u_{m-1} \\ u_m \end{pmatrix} \in \mathbb{R}^m$$

6.1.1 Der Algorithmus

Mit $n \in \mathbb{N}$ sei $E(a_n) \in V_N^0$ nicht Minimallösung für f bezüglich V_N.

Es sei
- d_n die Dimension des Gradientenraumes[26] W_n von $E(a_n)$
- $F_n := f - E(a_n)$.

Ferner seien $J \in \mathbb{N}$ mit $J \geqq 2N+1$ und $z \in (0, \frac{1}{2})$ fest vorgegeben.

[26] s. Definition 3.1 a,b

Schritt 1: Bestimmung einer Teilmenge X von I
Man untersuche, ob (6.1) erfüllt ist:

(6.1) Die Menge $\{x \in I \mid (1-z) \parallel F_n \parallel_I \leqq \mid F_n(x) \mid\}$ besteht aus j Teilintervallen mit $j \leqq J$ und in jedem Teilintervall ist das Maximum von $\mid F_n \mid$ eindeutig bestimmt. Die Maximalpunkte von $\mid F_n \mid$ in diesen Teilintervallen seien die Punkte $x_k,\ 1 \leqq k \leqq j$.

Fall 1: Es ist (6.1) erfüllt.

Die Menge der Maximalpunkte sei $X_1 := \{x_k \mid 1 \leqq k \leqq j\}$.

Weiter zerlege man I in *Vorzeichenintervalle*:
Dies sind Teilintervalle von I, in denen F_n das Vorzeichen nicht wechselt, im Innern benachbarter Teilintervalle jedoch verschiedenes Vorzeichen besitzt. Abweichend von Definition 1.7 gelte hier zur Bestimmung der Vorzeichenintervalle $sign(0) = +1$.
In jedem dieser Vorzeichenintervalle wird ein Punkt - günstig ist eine lokales Extremum von F_n - ausgezeichnet, die somit gegebene Punktmenge sei X_2.
X_3 sei eine beliebige Teilmenge von I mit $|X_3| = d_n + 1$.

Man bestimme nun Teilmengen $\overline{X_2}$ und $\overline{X_3}$ mit $\overline{X_2} \subseteq X_2$ und $\overline{X_3} \subseteq X_3$ so, daß für $\mathbf{X} := \mathbf{X_1} \cup \overline{\mathbf{X_2}} \cup \overline{\mathbf{X_3}}$
gilt: $d_n + 1 \leqq |X| \leqq J$

Fall 2: Es ist (6.1) nicht erfüllt.

Hier wird definiert: $\mathbf{X} := \mathbf{I}$

Schritt 2:
Man bestimme die *beste Approximation* e_n *an* $F_n = f - E(a_n)$ *auf X bezüglich* W_n, d.h.: für $E(a_n, x) = \sum_{i=1}^{N} a_{i,n} e^{t_{i,n} x}$ ist gesucht

$$r_n = (r_{1,n}, \ldots, r_{d_n,n}) \in \mathbb{R}^{d_n}$$

mit

- $e_n(x) = \sum_{i=1}^{N} r_{i,n} e^{t_{i,n} x} + \sum_{i=N+1}^{d_n} r_{i,n} a_{i-N,n} x e^{t_{i-N,n} x}$ und
- $\parallel F_n - e_n \parallel_I \leqq \parallel F_n - e \parallel_I$ für $e \in W_n$

Mit den Bezeichnungen von Definition 3.1 gilt für $d_n = 2N$:

$$e_n = (r_n, grad(E(a_n)))$$

Die *Verbesserung* für $E(a_n)$ sei $\epsilon_n := \| F_n \|_I - \| F_n - e_n \|_X$.

Schritt 3:
Man setze $c := 1$ und bestimme durch fortlaufendes Halbieren von c ein $c_n \in (0, 1]$ so, daß gilt:

$$\| f - E(a_n + c_n r_n) \|_I \leqq \| f - E(a_n) \|_I - \frac{1}{3} c_n \epsilon_n$$

Liegt Fall 1 vor und gilt $c_n < z$, dann werden Schritt 2 und Schritt 3 mit $X = I$ durchgeführt.

Der Parametervektor a_{n+1} und damit $E(a_{n+1})$ ist gegeben durch

$$a_{n+1} := a_n + c_n r_n$$

Bemerkungen:

1. Die Durchführung von Schritt 2 erfordert die Lösung eines linearen Approximationsproblems; e_n existiert damit und ist stets eindeutig bestimmt, da W_n die Haarsche Bedingung erfüllt.
2. Wie die folgenden Untersuchungen zeigen, wird die Fallunterscheidung nach Schritt 1 durchgeführt, um durch Lösung eines diskreten, linearen Approximationsproblems den Rechenaufwand möglichst gering zu halten. Speziell für $d_n+1 = J$ ist hier ein lineares Gleichungssystem zu lösen.
3. Obwohl X von F_n abhängt, wird dies bei der Bezeichnung nicht berücksichtigt, da diese Tatsache für die weiteren Ausführungen nicht benötigt wird.

6.1.2 Zur Theorie des Verfahrens

Es ist nun die Frage zu beantworten, unter welchen Voraussetzungen der Algorithmus durchführbar ist und damit bessere Approximationen liefert; hierzu der folgende Hilfssatz:

Hilfsatz 6.1
Ist $E(a_n)$ nicht beste Approximation an f auf I, dann gilt

$$\epsilon_{n1} \geqq \epsilon_{n2} > 0 \ ,$$

wobei ϵ_{n1} (ϵ_{n2}) die nach Fall 1 (Fall 2) ermittelte Verbesserung für $E(a_n)$ ist.

Beweis:
Es sei $X \subseteq I$ eine nach Fall 1 ermittelte Punktemenge. Mit e_{n1} sei die beste Approximation an F_n auf X und mit e_{n2} die beste Approximation an F_n auf I jeweils bezüglich W_n bezeichnet.

Es folgt $\| F_n - e_{n2} \|_I \geqq \| F_n - e_{n2} \|_X \geqq \| F_n - e_{n1} \|_X$ und damit $\epsilon_{n1} \geqq \epsilon_{n2}$.

Nach Voraussetzung besitzt F_n in I keine Alternante der Länge $d_n + 1$ und die beste Approximation e_{n2} an F_n bezüglich W_n auf I approximiert daher besser als die Nullfunktion:

$$\| F_n - e_{n2} \|_I < \| F_n \|_I$$

Dies ergibt $e_{n2} > 0$, womit der Rest der Behauptung gezeigt ist.

Satz 6.1 (Braess, [3])
Ist $E(a_n) \in V_N^0$ nicht Minimallösung für f auf I bezüglich V_N^0, dann gibt es ein $C > 0$, so daß für $c \in (0, C)$ erfüllt ist:

$$\| f - E(a_n + cr_n) \|_I \leqq \| f - E(a_n) \|_I - \frac{1}{3} c\epsilon_n$$

Beweis:
Nach Voraussetzung besitzt F_n in I keine Alternante der Länge $d_n + 1$ und nach Hilfssatz 6.1 gilt (für beide Fälle) $\epsilon_n > 0$.

Es sei $a_n(c) := a_n + cr_n$ und $F(c, x) = f(x) - E(a_n(c), x)$, $c \in \mathbb{R}$.

Wegen $\| E(a_n(c)) - E(a_n) - ce_n \|_I = o(|c|)$ (man vergleiche hierzu Kapitel 4) gibt es ein $C > 0$, mit

$$\| E(a_n(c)) - E(a_n) - ce_n \|_I \leqq \frac{2}{3} c\epsilon_n \text{ für } c \in (0, C) \ . \tag{6.2}$$

Weiter gilt

$$\| E(a_n(c)) - f \|_I \leqq \| E(a_n) + ce_n - f \|_I + \| E(a_n(c)) - E(a_n) - ce_n \|_I \ .$$

Für die Berechnung nach Fall 2 folgt aus (6.2) für $c \in (0, C)$:

$$\begin{aligned} \| f - E(a_n(c)) \|_I &\leqq \| E(a_n) + ce_n - f \|_I + \frac{2}{3}c\epsilon_n = \\ &= \| cE(a_n) + ce_n - cf + (1-c)(E(a_n) - f) \|_I + \frac{2}{3}c\epsilon_n \leqq \\ &\leqq c \| F_n - e_n \|_I + (1-c) \| F_n \|_I + \frac{2}{3}c\epsilon_n = \\ &= -c\epsilon_n + \| F_n \|_I + \frac{2}{3}c\epsilon_n \leqq \| F_n \|_I - \frac{1}{3}c\epsilon_n \end{aligned}$$

Damit ist die Behauptung für die Durchführung des Algorithmus nach *Fall 2* gezeigt.

Zu *Fall 1*: Hier gilt also $|X| \leqq J$.

Es sei

$$\begin{aligned} X^+ &:= \{x \in X \mid F_n(x) = \| F_n \|_I\} \\ X^- &:= \{x \in X \mid F_n(x) = - \| F_n \|_I\} \end{aligned}$$

Für $x \in X^+$ erhält man $0 < \epsilon_n \leqq \| F_n \|_I - (F_n(x) - e_n(x)) = e_n(x)$ und entsprechend für $x \in X^-$ $0 > -\epsilon_n \geqq - \| F_n \|_I - (F_n(x) - e_n(x)) = e_n(x)$.

Es gibt somit eine offene Umgebung U von $X^+ \cup X^-$ in I, so daß $|e_n(x)| \leqq \frac{2}{3}\epsilon_n > 0$ für $x \in U$ erfüllt ist.
Da F(c,x) in c und x stetig ist, gibt es eine offene Umgebung U^+ von X^+ und U^- von X^- in I und ein $C_1 > 0$, so daß für $c \in (0, C_1)$ gilt: $F(c,x) > 0$ für $x \in U^+$ und $F(c,x) < 0$ für $x \in U^-$.

Ferner gibt es ein $C_2 > 0$ mit $|E(a_n(c), x) - E(a_n, x) - ce_n(x)| \leqq \frac{1}{3}c\epsilon_n$ für $c \in (0, C_2)$ und $x \in I$. Damit folgt für $x \in U \cap U^+$ und $c \in (0, \min\{C_1, C_2\})$

$$\begin{aligned} F(c,x) - F_n(x) &= E(a_n, x) - E(a_n(c), x) \leqq \\ &\leqq -ce_n(x) + \frac{1}{3}c\epsilon_n \leqq -\frac{2}{3}c\epsilon_n + \frac{1}{3}c\epsilon_n = -\frac{1}{3}c\epsilon_n \ , \end{aligned}$$

also

$$0 < f(x) - E(a_n(c), x) \leqq \| F_n \|_I - \frac{1}{3}c\epsilon_n \ .$$

Für $x \in U \cap U^-$ erhält man analog für $c \in (0, \min\{C_1, C_2\})$

$$\begin{aligned} -F(c,x) + F_n(x) &= E(a_n(c),x) - E(a_n,x) \leqq \\ &\leqq ce_n(x) + \frac{1}{3}c\epsilon_n \leqq -\frac{2}{3}c\epsilon_n + \frac{1}{3}c\epsilon_n = -\frac{1}{3}c\epsilon_n \ , \end{aligned}$$

also

$$-F(c,x) = |f(x) - E(a_n(c),x)| \leqq \parallel F_n \parallel_I - \frac{1}{3}c\epsilon_n \ .$$

Die Menge $V = (U^+ \cap U) \cup (U^- \cap U)$ ist offen in I; $I_0 = I \setminus V$ ist kompakt und es gilt für $c_0 = 0$:

$$\parallel F(c_0) \parallel_{I_0} - \parallel F_n \parallel_I + \frac{1}{3}c_0\epsilon_n = \parallel F_n \parallel_{I_0} - \parallel F_n \parallel_I < 0 \ .$$

Dies gilt wegen Stetigkeit für eine Umgebung von c_0:
Es gibt ein $C_0 > 0$ mit $\parallel F(c) \parallel_{I_0} \leqq \parallel F_n \parallel_I - \frac{1}{3}c\epsilon_n$ für $c \in (0, C_0)$. Mit $0 < C := \min\{C_0, C_1, C_2, \}$ hat man somit für $c \in (0, C)$ erhalten:

$$|f(x) - E(a_n(c),x)| \leqq \parallel F_n \parallel_I - \frac{1}{3}c\epsilon_n \qquad x \in I \ .$$

Damit ist auch für Fall 1 die Behauptung bewiesen,

Nach Satz 6.1 ergibt also der Algorithmus eine bessere Approximation als $E(a_n)$, falls dies möglich ist.

Hilfsatz 6.2 (Braess, [3])
$(E(a_n))_{n\in\mathbb{N}} \subseteq V_N^0$ konvergiere auf I gleichmäßig gegen $E(a_0) \in V_N^0 \setminus V_{N-1}$. Es sei e_n bzw. e die nach Schritt 2 berechnete beste Approximation an F_n bzw. $F := f - E(a_0)$ bezüglich W_n bzw. W auf $X = I$; (W ist der Gradientenraum von $E(a_0)$); es liege also stets Fall 2 vor.

Behauptung: $(e_n)_{n\in\mathbb{N}}$ konvergiert auf I gleichmäßig gegen e.

Beweis:
Es kann o.B.d.A. angenommen werden:

$$\parallel E(a_n) \parallel_I \leqq K < \infty \text{ und } E(a_n,x) = \sum_{i=1}^{N} a_{i,n}e^{t_{i,n}x} \in V_N^0 \setminus V_{N-1} \text{ für } n \in \mathbb{N}$$

Nach Korollar 2.3 sind die Frequenzen der Folge beschränkt; nach Satz 2.5

gibt es eine Teilfolge $(E(a_n))_{n\in J}$, $J \subseteq \mathbb{N}$ mit

$$\lim_{n\to\infty} t_{i,n} = t_i \in \mathbb{R} \qquad n \in J$$

(6.3) und

$$E(a_0, x) = \sum_{i=1}^{N} a_{0,i} e^{t_{i,n}x}$$

Weiter gilt für $n \in \mathbb{N}$

$$e_{in}(x) = \sum_{i=1}^{N} (b_{i,n} + d_{i,n}x) e^{t_{i,n}x} \in V_{2N}$$

und $(e_n)_{n\in\mathbb{N}}$ ist auf $I = [u, v]$ gleichmäßig beschränkt:

$$\| e_n \|_I \leqq \| e_n - F_n \|_I + \| F_n \|_I \leqq 2 \| F_n \|_I \leqq 2(K + \| f \|_I)$$

Nach Satz 2.3 und Satz 2.5 gibt es daher eine im Innern von I gleichmäßig konvergente Teilfolge $(c_n)_{n\in J_1}$, $J_1 \subseteq J$, und auf jedem abgeschlossenen Teilintervall von (u,v) gilt

$$\lim_{n\to\infty} e_n = e_0 \ , \ n \in J_1$$

mit $e_0 \in V_{2N}$; wegen (6.4) folgt nach Satz 2.9 und Bemerkung 2.2 die gleichmäßige Konvergenz von $(e_n)_{n\in J_1}$ gegen e_0 auf ganz I.
Aus $\| F_n - e_n \|_I \leqq \| F_n - e \|_I$ für $n \in \mathbb{N}$ folgt

$$\| F - e_0 \|_I = \lim_{n\to\infty} \| F_n - e_n \|_I \leqq \lim_{n\to\infty} \| F_n - e \|_I = \| F - e \|_I \ , \ n \in J_1$$

Die Eindeutigkeit der Minimallösung bei linearer Approximation und $e_0 \in W$, $e \in W$ ergeben: $e_0 = e$.

Es ist noch zu zeigen, daß die ganze Folge $(e_n)_{nN}$ gleichmäßig gegen e konvergiert:
Annahme: Es gibt ein $s > 0$, so daß es für jedes $n \in \mathbb{N}$ ein $m_n \geqq n$ und ein $x_{m_n} \in I$ gibt mit $|e_{m_n}(x_{m_n}) - e(x_{m_n})| > s$.

I ist kompakt, daher gibt es in $(x_{m_n})_{n\in\mathbb{N}}$ eine konvergente Folge $(x_{n_j})_{j\in\mathbb{N}}$ mit

(6.4) $$|e_{n_j}(x_{n_j}) - e(x_{n_j})| > s \qquad j \in \mathbb{N}$$

Für $(e_{n_j})_{j\in\mathbb{N}}$ folgt wie oben für $(e_n)_{n\in\mathbb{N}}$ die Existenz einer auf ganz I gleichmäßig gegen e konvergenten Teilfolge und dies steht im Widerspruch zu (6.4). Damit ist die Annahme widerlegt und die Behauptung vollständig gezeigt.

Satz 6.2 (Braess, [3])
Ergibt der Algorithmus eine Teilfolge $(E(a_n))_{n\in J} \subseteq V_N^0$, $J \subseteq \mathbb{N}$, die auf I gleichmäßig gegen ein Element $E(a_0) \in V_N^0 \setminus V_{N-1}$ konvergiert, dann konvergiert die ganze Folge $(E(a_n))_{n\in\mathbb{N}}$ gleichmäßig auf I und $E(a_0)$ ist die Minimallösung für f bezüglich V_N auf I.

Beweis:
Es kann nach Voraussetzung o.B.d.A. $grd(E(a_n)) = N$ für $n \in \mathbb{N}$ angenommen werden.

Es sei $F_0 := f - E(a_0)$.

Annahme: $E(a_0)$ ist nicht Minimallösung für f bezüglich V_N.
Wendet man den Algorithmus auf $E(a_0)$ nach Fall 2 (also mit $X = I$) an, dann erhält man nach Hilfssatz 6.1:

$$\epsilon = \| F_0 \|_I - \| F_0 - (r_0, grad(E(a_0))) \|_X > 0$$

($e_0 = (r_0, grad(E(a_0)))$ ist die beste Approximation an F_0 bezüglich des Gradientenraums von $E(a_0)$.)
Ist bei der Konstruktion von $(E(a_n))_{n\in\mathbb{N}}$ stets gemäß Fall 2 verfahren worden, dann folgt nach Hilfssatz 6.2

$$\lim_{n\to\infty} \epsilon_n = \epsilon \text{ für } n \in J$$

und es gibt damit ein $n_0 \in \mathbb{N}$, so daß für $n \in J$ und $n \geqq n_0$ gilt:

$$\epsilon_n \geqq \frac{1}{2} < \epsilon \tag{6.5}$$

Nach Hilfssatz 6.1 gilt (6.5) auch, falls ϵ_n gemäß Fall 1 berechnet wird.

Fallunterscheidung:
Fall a: Bei der Teilfolge $(E(a_n))_{n\in J}$ sei **stets nach Fall 2** verfahren worden. Aus der gleichmäßigen Konvergenz ergibt sich wieder die Beschränktheit der Frequenzen von $(E(a_n))_{n\in J}$ (man vergleiche hierzu (6.4) im Beweis zu Hilfssatz 6.2), woraus nach Satz 2.8 für $n \in J$ (wegen $grd(E(a_n)) = N$)

$$\lim_{n\to\infty} a_n = a_0 \qquad \text{und} \tag{6.6}$$

$$\lim_{n\to\infty} r_n = r_0 \qquad \text{(nach Hilfssatz 6.2)} \tag{6.7}$$

folgt.

Auf Grund der stetigen Differenzierbarkeit der Exponentialsummen gibt es ein $M < \infty$ mit

$$\| E(a_n(c)) - E(a_n) - ce_n \|_I \leqq c^2 M \text{ für } n \in J \text{ , } |c| \leqq 1 \text{ ;}$$

$a_n(c)$ ist definiert wie in Satz 6.1.
Folglich gibt es ein C mit $0 < C \leqq 1$, so daß mit (6.5) für $c \in (0, C)$ erfüllt ist:

$$\| E(a_n(c)) - E(a_n) - ce_n \|_I \leqq \frac{1}{3}c\epsilon \leqq \frac{2}{3}c\epsilon_n \ , \ n \in J \ \ , \ n \geqq n_0 \ .$$

Wie in Satz 6.1 gezeigt, folgt daraus $c_n \geqq \frac{C}{2}$, $n \in J$, $n \geqq n_0$; man erhält somit für $n \in J$, $n \geqq n_0$:

$$\| E(a_{n+1}) - E(a_n) \|_I \geqq \| F_n \|_I - \| F_{n+1} \|_I \geqq \frac{1}{3}c_n\epsilon_n \geqq \frac{1}{3}\frac{C}{2}\frac{\epsilon}{2} > 0$$

Dies steht im Widerspruch zur gleichmäßigen Konvergenz der Folge.

Fall b: Es gebe eine Teilfolge$(E(a_n))_{n\in J_1}$, $J_1 \subseteq J$, bei der **gemäß Fall 1** verfahren wird.
Wegen $c_n \geqq z$ folgt hier im Widerspruch zur gleichmäßigen Konvergenz

$$\| F_n \|_I - \| F_{n+1} \|_I \geqq \frac{1}{3}z\epsilon_n \geqq \frac{1}{3}z\frac{1}{2}\epsilon > 0 \qquad n \in J_1 \ , \ n \geqq n_0 \ .$$

Damit konvergiert die ganze Folge $(E(a_n))_{n\in\mathbb{N}}$ gleichmäßig gegen $E(a_0)$:
Es sei $(E(a_n))_{n\in J_2}$, $J_2 \subseteq \mathbb{N}$, eine beliebige (unendliche) Teilfolge. Da $(E(a_n))_{n\in J_2}$ eine Minimalfolge ist, ist sie gleichmäßig auf I beschränkt; es gibt also eine im Innern von $I = [u, v]$ gleichmäßig konvergente Teilfolge, die auf jedem abgeschlossenen Intervall in (u, v) gleichmäßig gegen ein Element von $V_N \setminus V_{N-1}$ konvergiert, da nach Voraussetzung

$$\inf_{E\in V_{N-1}} \| f - E \|_I > \| f - E(a_{n)} \|_I$$

für hinreichend großes $n \in J_2$ gilt. Somit konvergiert diese Teilfolge auf I gleichmäßig gegen ein $E \in V_N \setminus V_{N-1}$.
E ist aber, wie gezeigt, Minimallösung und aus der Eindeutigkeit der besten Approximation folgt $E = E(a_0)$.

6.1.3 Zur Anwendung des Algorithmus

Zur Konstruktion bester Approximationen geht man nach Braess vor wie folgt, vgl. [3], Kapitel 14 (p. 267 ff):

a. Approximation bezüglich V_N^+
Um die Voraussetzungen von Satz 6.2 immer erfüllen zu können, geht man bei der Konstruktion der besten Approximation bezüglich V_N^+ induktiv vor:
Die beste Approximation bezüglich $V_0^+ = V_0$ ist die Nullfunktion;
die Minimallösung $E(a)$ für f bezüglich V_{N-1}^+ sei bekannt und nicht die beste Approximation bezüglich V_N^+. Jede Alternante von $f - E(a)$ in I hat also höchstens die Länge $2N-1$ und jede bessere Approximation in V_N ist nach Korollar 3.3 Element von $V_N^+ \setminus V_{N-1}$.

Die Iteration beginnt man mit

$$E(a_1, x) = E(a, x) + a_{N,1} e^{t_{N,1} x} \ ,$$

wobei $a_{N,1} = 0$ und $t_{N,1} \in \mathbb{R}$ nicht im Spektrum von $E(a)$ enthalten ist. Es gilt also $d_1 = 2N-1$ und die Durchführung des Algorithmus ergibt eine bessere Approximation $E(a_2) \in V_N^+ \setminus V_{N-1}$ nach Satz 6.1; man erhält so eine Folge $(E(a_n))_{n\in\mathbb{N}}$, die in

$$V = \{E \mid E \in V_N^+ \setminus V_{N-1},\ \| f - E \|_I \leqq \| f - E(a_2) \|_I < \| f - E(a) \|_I \}$$

enthalten ist.

Falls $q := \dfrac{\max\limits_{x\in I} f(x) + \min\limits_{x\in I} f(x)}{2} > 0$ gilt, ist die beste Approximation an f auf I bezüglich V_1^+ in $V_1 \setminus V_0$ enthalten, und man kann die Iteration in V_1^+ sofort mit $E(a_1, x) = q \in V_1 \setminus V_0$ beginnen.

Nach Satz 2.3, Satz 2.5 und Satz 2.9 gibt es eine auf I gleichmäßig konvergente Teilfolge, deren Grenzfunktion nach Satz 2.7 in V enthalten ist.
Bei der Konstruktion der besten Approximation bezüglich V_N^+ ist auf diese Weise die Voraussetzung von Satz 6.2 erfüllt und die bei der Iteration ermittelte Folge $(E(a_n))_{n\in\mathbb{N}}$ konvergiert gleichmäßig gegen die beste Approximation an f bezüglich V_N^+.

Bemerkung:
Diese Ausführungen zur Approximation bezüglich V_N^+ mit diesem Verfahren gelten analog auch für die Approximation bezüglich V_N^-.

b. Approximation bezüglich V_N^0 ohne Vorzeichenbeschränkung
Liegt die beste Approximation bezüglich V_N^0 (sofern sie existiert) nicht in V_N^+, dann ist im allgemeinen die Voraussetzung von Satz 6.2 nicht erfüllt[27] und man hat eine *Modifikation des Verfahrens*[28] durchzuführen, um den Algorithmus auch hier noch sinnvoll anzuwenden.

Es sei also $E(a) \in V_{N-1}^0$ die Minimallösung bezüglich V_{N-1}^0 an f und die beste Approximation bezüglich V_N^0, deren Existenz angenommen wird, sei nicht in V_N^+ oder V_N^- enthalten; bei induktivem Vorgehen kann man mit Satz 3.4 überprüfen, ob diese Voraussetzung erfüllt ist.

Wie oben beginnt man mit

$$E(a_1, x) = E(a, x) + \sum_{i=K+1}^{N} a_{i,1} e^{t_{i,1} x} \ ,$$

wobei $a_{i,1} = 0$ für $K < i \leqq N$ und $K = grad(E(a))$ gilt und ferner $t_{i,1} \in \mathbb{R}$, $K < i \leqq N$, nicht im Spektrum von $E(a)$ enthalten ist und $t_{i,1} < t_{i+1,1}$ ist. Man erhält eine bessere Approximation $E(a_2) \in V_N^0$ nach Satz 6.1 mit dem Vorzeichenvektor $S = sign(E(a_2))$. Die Modifikation besteht nun darin, daß bei der Konstruktion einer Folge besserer Approximationen $(E(a_n))_{n\in\mathbb{N}}$ im Schritt 3 des Algorithmus der Faktor c_n für $n \in \mathbb{N}$ so klein gewählt wird, daß stets $sign(E(a_n)) = S$ gilt. Dies ist nach Satz 6.1 möglich; in der Regel ist diese Zusatzforderung bereits bei der Bestimmung von c_n nach Schritt 3 erfüllt.

Erhält man auf diese Weise eine gegen $E(a_0) \in V_N^0 \setminus V_{N-1}$ gleichmäßig konvergente Teilfolge, dann ist $E(a_0)$ Minimallösung, da diese Zusatzforderung am Beweis von Satz 6.2 wegen (6.6) und (6.7) nichts ändert: $C > 0$ kann hinreichend klein gewählt werden.

Für $N \geqq 2$ tritt dieser Fall im allgemeinen jedoch nicht ein, da es nach Bemerkung 4.1 und Satz 4.3 verschiedene Vorzeichenklassen mit besseren Approximationen als $E(a)$ und damit mehrere lokale Minima gibt. Da die lokalen Minima, abgesehen von der besten Approximation bezüglich V_N^0, Elemente von $V_N \setminus V_N^0$ sind (Korollar 4.1), geht man nach Braess vor wie folgt:

[27] "wenn die beste Exponentialsumme sowohl positive als auch negative Faktoren enthält", [3], Kap 14, p. 268

[28] in [3], Kap 14, p. 269: "Erweiterung des Algorithmus"

Erhält man bei der Iteration von $V_N^0(S)$ zwei oder mehrere Folgen von Frequenzen, die gegen einen gemeinsamen Grenzwert konvergieren, so breche man die Iteration ab und bestimme durch geeignete Wahl der Frequenzen $t_{i,1}$, $K < i \leqq N$, nach Hilfssatz 4.1 und Satz 4.2 eine neue Startfunktion $E(a_1')$, die in einer anderen Vorzeichenklasse enthalten ist, und führe hier eine neue Iteration durch.

Für $K = N - 1$ gibt Bemerkung 4.1 die Zahl der Vorzeichenklassen an, in denen auf diese Weise der Algorithmus angewandt werden kann. Es ist allerdings nicht gesichert, daß so sämtliche Vorzeichenklassen mit besseren Approximationen erreicht werden und ob nicht mehr lokale Minima existieren, als Satz 4.3 angibt.

6.2 Das Newtonsche Iterationsverfahren

Das Newtonsche Iterationsverfahren, das allgemein für nichtlineare Approximation in [11] beschrieben ist, läßt sich nach den Ergebnissen von Kapitel 3 insbesondere bei der Approximation durch Exponentialsummen anwenden.

Es sei $E(a) \in V_N^0 \setminus V_{N-1}$ die beste Approximation an $f \in C(I)$ auf $I = [u, v]$ bezüglich V_N; E(a) ist also nach Satz 3.1 eindeutig bestimmt und $F := f - E(a)$ besitzt in I eine Alternante der Länge 2N+1.
Der Gradientenraum von E(a) hat die Dimension 2N und dies gilt für alle Elemente von $\{E \in V_N^0 \mid \| f - E \|_I < \inf_{w \in V_{N-1}} \| f - w \|_I\}$, da diese den Grad N besitzen. Es existiert also eine Umgebung von E(a) in V_N^0, so daß die Elemente dieser Umgebung stets einen Gradientenraum der Dimension 2N besitzen.

Es sei $E(a_0) \in V_N^0$ eine Näherung für E(a), so daß es in I 2N+1 Punkte $x_{1,k}$, $0 \leqq k \leqq 2N$, gibt mit

$$0 \neq sign(F_0(x_{1,k})) \neq sign(F_0(x_{1,k+1})) \neq 0 \quad 0 \leqq k \leqq 2N-1 \; ; \tag{6.8}$$

dabei sei $F_0 := f - E(a_0)$. In den Punkten $x_{1,k}$ nehme $|F_0|$ möglichst große Werte an.

Nach Hilfssatz 3.1 gilt $\min_{0 \leqq k \leqq 2N} |F_0(x_{1,k})| \leqq \inf_{E \in V_N} \| f - E \|_I$.
In Analogie zum Remez-Verfahren wird man daher versuchen, einen Parametervektor $a_1 \in \mathbb{R}^{2N}$ so zu bestimmen, daß mit $F_1 := f - E(a_1)$ gilt:

$$F_1(x_{1,k}) = (-1)^k r_1 \qquad 0 \leqq k \leqq 2N \tag{6.9}$$

Existiert dieser Parametervektor, dann ist $E(a_1)$ die beste Approximation an f auf $\{x_{1,k} \mid 0 \leqq k \leqq 2N\}$ und $|r_1|$ eine weitere untere Abschätzung für die Minimalabweichung von f auf I. Gesucht ist also $(a_1, r_1) \in \mathbb{R}^{2N+1}$ mit

$$g_k(a_1, r_1) := E(a_1, x_{1,k}) - f(x_{1,k}) + (-1)^k r_1 = 0 \quad 0 \leqq k \leqq 2N$$

Demnach ist (a_1, r_1) eine Nullstelle von

$$G(b, q) := \begin{pmatrix} g_0(b, q) \\ \vdots \\ g_{2N}(b, q) \end{pmatrix}, \; (b, q) \in \mathbb{R}^{2N+1}, \; b \in \mathbb{R}^{2N} \; ,$$

Verwendet man das Newton-Verfahren (im $\mathbb{R}^{2N+1}$) mit dem Startwert $(b_1, q_1) = (a_0, q_1)$, q_1 geht in die Berechnung nicht ein, so erhält man zunächst formal für $n \in \mathbb{N}$:

$$J_G(b_n, q_n) \times (b_{n+1}, q_{n+1})^t = J_G(b_n, q_n) \times (b_n, q_n)^t - G(b_n, q_n) \; ;$$

J_G ist die Jacobi-Matrix von G.

Es folgt also explizit für $0 \leqq k \leqq 2N$:

$$\begin{aligned}\sum_{j=1}^{2N} b_{n+1,j} D_j E(b_n, x_{1,k}) + (-1)^k q_{n+1} \quad &= \\ = \sum_{j=1}^{2N} b_{n,j} D_j E(b_n, x_{1,k}) + (-1)^k q_n \quad &- \quad E(b_n, x_{1,k}) + f(x_{1,k}) - (-1)^k q_n\end{aligned}$$

$D_j E(b)$ ist die partielle Ableitung von E(b) nach b_j mit $b = (b_1, \ldots, b_{2N})$; ferner gilt $b_n = (b_{n,1}, \ldots, b_{n,2N})$. Es folgt damit für $0 \leqq k \leqq 2N$:

$$\sum_{j=1}^{2N} (b_{n+1,j} - b_{n,j}) D_j E(b_n, x_{1,k}) + (-1)^k q_{n+1} = f(x_{1,k}) - E(b_n, x_{1,k})$$

Hat der Gradientenraum von $E(a_0)$ die Dimension 2N, dann ist für n=1 dieses Gleichungssystem eindeutig lösbar und man erhält den Vektor $b_2 - b_1 = b_2 - a_0 \in \mathbb{R}^{2N}$. Ist $E(a_0)$ weiterhin eine hinreichend gute Näherung für E(a), dann kann $a_1 = b_2$ gesetzt werden, d.h. $E(a_1) = E(b_2)$ erfüllt (6.9) für numerische Zwecke hinreichend gut.

Im allgemeinen ist dies jedoch nicht der Fall (man vergleiche hierzu Kapitel 8); man hat dann das Newton-Verfahren mindestens soweit durchzuführen, daß für ein $n \in \mathbb{N}$

$$sign(f(x_{1,k}) - E(b_n, x_{1,k})) \neq sign(f(x_{1,k+1}) - E(b_n, x_{1,k+1})) \ , \ 0 \leqq k < 2N$$

erfüllt ist.

Da auf diskreten Punktmengen nicht allgemein eine beste Approximation bezüglich V_N^0 existiert, ist nicht gesichert, daß dies stets möglich ist. Es ist jedenfalls erforderlich, daß der Gradientenraum der bei der Newton-Iteration berechneten Elemente $E(b_n)$ die Dimension 2N hat.

Ist eine Funktion $E(a_1)$ ermittelt worden und approximiert diese f besser als $E(a_0)$, dann verfährt man mit $E(a_1)$, falls $E(a) \neq E(a_1)$ ist, wie mit $E(a_0)$ und erhält so $E(a_2)$, wenn die entsprechenden Voraussetzungen erfüllt sind, usw.

Da Konvergenzaussagen für dieses Verfahren nicht möglich sind, wird man nach jedem Schritt überprüfen, ob man eine bessere Approximation erhalten hat.

6.3 Approximation bzgl. V_1 durch Bestimmung von reellen Nullstellen

Gegeben sei das Intervall I und $f \in C(I)$.

Zur Bestimmung der besten Approximation an f auf I bzgl. V_1 wird ein Verfahren beschrieben, das eine Übertragung des Remez-Algorithmus darstellt; im Gegensatz zum Newton-Verfahren von Abschnitt 6.2 ist hier nur eine reelle Nullstelle, etwa nach Newton-Raphson, zu ermitteln.

Ohne Beschränkung der Allgemeinheit sei die Minimallösung E für f bzgl. V_1, sofern sie existiert, Element von $V_1 \setminus V_0$. [29]

In I besitzt also $f-E$ eine Alternante der Länge 3; es liegt daher nahe zu versuchen, durch Konstruktion der Minimallösung E_1 für f auf einer Menge $X \subseteq I$ mit $|X| = 3$ die beste Approximation E auf I zu ermitteln.

Es sei $X = \{x_1, x_2, x_3\} \subseteq I$ mit $x_1 < x_2 < x_3$ so gegeben, daß für f auf X die Minimallösung $E(x) = ae^{sx} \in V_1$ existiert und weiter $a \neq 0$ erfüllt ist. Es gilt also

$$ae^{sx_i} + (-1)^i r = f(x_i) \quad 1 \leq i \leq 3 \ ; \tag{6.10}$$

die Minimalabweichung von f auf X sei also $|r|$, $r \in \mathbb{R}$.

Nach Voraussetzung gilt

$$f(x_3) + f(x_2) = a(e^{sx_3} + e^{sx_2}) \neq 0 \ ,$$

so daß man durch Elimination von r und a aus (6.10) das äquivalente Gleichungssystem (6.11) erhält (vgl. Anhang A):

$$\begin{aligned} -r &= f(x_3) - ae^{sx_3} \\ a &= \frac{f(x_3) + f(x_2)}{e^{sx_3} + e^{sx_2}} \\ \frac{e^{sx_3} - e^{sx_1}}{e^{sx_3} + e^{sx_2}} &= \frac{f(x_3) - f(x_1)}{f(x_3) + f(x_2)} \end{aligned} \tag{6.11}$$

Setzt man

$$F(t) := \frac{e^{tx_3} - e^{tx_1}}{e^{tx_3} + e^{tx_2}} \quad \text{für } t \in \mathbb{R} \ , \quad w := \frac{f(x_3) - f(x_1)}{f(x_3) + f(x_2)}$$

so ist die Lösung des Gleichungssystems 6.10 zurückgeführt auf die Bestimmung einer reellen Nullstelle von $F(t)-w$:

[29] V_0 enthält nur die Nullfunktion, s. Bemerkung 1.1, Punkt 3. Nach Korollar 3.1b ist die Nullfunktion 0 genau dann die beste Approximation an F auf I bezüglich V_1, wenn $f-0$, also die Funktion f selbst, in I eine Alternante der Länge 2 besitzt.

- die Frequenz s ergibt sich aus $F(s) = w$
- der Koeffizient a ergibt sich aus der zweiten Gleichung von (6.11):
  $a = \dfrac{f(x_3) + f(x_2)}{e^{sx_3} + e^{sx_2}}$

Eine Charakterisierung der Nullstellen von $F(t) - w$ liefert Satz 6.3:

Satz 6.3 Es gibt genau dann ein $s \in \mathbb{R}$ mit $F(s) = w$, wenn $w < 1$ erfüllt ist, und s ist dann eindeutig bestimmt.

Beweis:
Wegen $x_1 < x_2 < x_3$ gilt für $t \in \mathbb{R}$

$$\frac{\mathrm{d}F(t)}{\mathrm{d}t} = \frac{(x_3 - x_1)e^{t(x_1+x_3)} + (x_3 - x_2)e^{t(x_3+x_2)} + (x_2 - x_1)e^{t(x_1+x_2)}}{(e^{tx_3} + e^{tx_2)^2}} > 0$$

$F(t)$ ist also in $\mathbb{R}$ streng monoton wachsend.
Weiter erhält man

$$\lim_{t\to\infty} F(t) = \lim_{t\to\infty} \left(\frac{1}{1 + e^{t(x_2-x_3)}} - \frac{1}{e^{t(x_3-x_1)} + e^{t(x_2-x_1)}}\right) = 1$$

und analog

$$\lim_{t\to-\infty} F(t) = -\infty$$

$F(t)$ nimmt also für $t \in \mathbb{R}$ jeden Wert aus $(-\infty, +1)$ genau einmal an.
Damit ist der Satz gezeigt.

Bemerkung 6.1 gibt eine nicht-leere Menge von Funktionen an, für die das Gleichungssystem 6.10 lösbar ist:

Bemerkung 6.1
Genügt $f \in C(I)$ den beiden Eigenschaften

- $sign(f(x) \geqq 0$ oder $sign(f(x)) \leqq 0$ für $x \in I$
  (d.h.: f ändert in I sein Vorzeichen nicht)
- f ist in I entweder streng monoton wachsend oder streng monoton fallend

dann gilt für $x_i \in I$, $1 \leqq i \leqq 3$, mit $x_1 < x_2 < x_3$

$$\frac{f(x_3) - f(x_1)}{f(x_3) + f(x_2)} < 1$$

Damit läßt sich der folgende **Algorithmus** formulieren für den ("normalen") Fall, daß die Minimallösung E für f auf I bezüglich V_1 nicht in V_0 enthalten ist.

Schritt 1: Ermittlung einer Startfunktion

Zur Ermittlung einer Startfunktion $E_1(x) = a_1 e^{t_1 x}$ bestimme man drei Punkte $x_{1,i} \in I$, $1 \leqq i \leqq 3$, mit $x_{1,1} < x_{1,2} < x_{1,3}$, so daß f in $\{x_{1,1}, x_{1,2}, x_{1,3}\}$ keine Alternante der Länge 2 besitzt; nach Voraussetzung ist dies möglich.
Mit

$$w_1 = \frac{f(x_{1,3}) - f(x_{1,1})}{f(x_{1,3}) + f(x_{1,2})} < 1$$

ist t_1 die reelle Nullstelle von

$$F_1(t) - w_1 = \frac{e^{tx_{1,3}} - e^{tx_{1,1}}}{e^{tx_{1,3}} + e^{tx_{1,2}}} - w_1$$

Man erhält a_1 aus

$$a_1 = \frac{f(x_{1,3}) + f(x_{1,2})}{e^{t_1 x_{1,3}} + e^{t_1 x_{1,2}}}$$

Falls für die Punkte $x_{1,i}$, $1 \leqq i \leqq 3$,

$$|r_1| \;=\; |f(x_{1,3}) - a_1 e^{t_1 x_{1,i}}| \;>\; 0$$

erfüllt ist, wird mit Schritt 2 fortgefahren.
Andernfalls sind neue Punkte $x_{1,i}$ zu so wählen, daß mit diesen $|r_1| > 0$ gilt. Nach Bemerkung 6.2 ist dies möglich.

Schritt 2: Iterative Berechnung besserer Approximationen

Ausgehend von der in Schritt 1 ermittelten Startfunktion E_1 werden iterativ bessere Approximationen E_2, E_3, E_4, ... errechnet. Es sei also $n \geqq 2$.

Ist $E_{n-1}(x) = a_{n-1} e^{t_{n-1} x}$ nicht bereits die beste Approximation für f auf I, dann bestimme man Punkte $x_{n,i} \in I$, $1 \leqq i \leqq 3$, so daß

- $0 \neq sign(f(x_{n,i}) - E_{n-1}(x_{n,i})) = -sign(f(x_{n,i+1}) - E_{n-1}(x_{n,i+1}))$ für $i \in \{1, 2\}$ erfüllt ist und weiterhin
- $f - E_{n-1}$ in $x_{n,i}$ für $1 \leqq i \leqq 3$ dem Betrage nach einen möglichst großen Wert annimmt

Nach Konstruktion von E_{n-1} ist dies möglich.
Mit

$$w_n := \frac{f(x_{n,3}) - f(x_{n,1})}{f(x_{n,3}) + f(x_{n,2})} < 1$$

ist t_n die reelle Nullstelle von

$$F_n(t) - w_n \;:=\; \frac{e^{tx_{n,3}} - e^{tx_{n,1}}}{e^{tx_{n,3}} + e^{tx_{n,2}}} - w_n$$

Man erhält a_n aus

$$a_n = \frac{f(x_{n,3}) + f(x_{n,2})}{e^{t_n x_{n,3}} + e^{tx_{n,2}}}$$

Ist $E_n(x) = a_n e^{t_n x}$ nicht die beste Approximation, dann wird dieser Schritt mit n+1 erneut durchgeführt.
Falls $\| f - E_n \|_I \geqq \| f - E_{n-1} \|_I$ gilt, wird das Verfahren beendet.

Bemerkung 6.2
Für jedes Element f der in Bemerkung 6.1 beschriebenen Funktionenklasse ist das Verfahren stets durchführbar:
Zunächst besitzt f nach Voraussetzung in keiner Punktemenge $X \subseteq I$ eine Alternante der Länge zwei.
Nimmt man an, daß bei der Ermittlung einer Startfunktion für jede Menge $X \subseteq I$, $| X | = 3$, stets $| r_1 | = 0$ gilt, dann folgt mit Satz 1.1: $f \in V_1$:
denn dann gibt es für $\{x_1, x_2, x_3\} \subseteq I$, $x_1 < x_2 < x_3$, ein $E_1 \in V_1$ mit $E_1(x_i) = f(x_i)$, $1 \leqq i \leqq 3$, und für jede Menge $\{x_1, x_2, \overline{x}\}$, $\overline{x} \in I$, $x_1 \neq \overline{x} \neq x_2$, ein $E_{\overline{x}} \in V_1$ mit $E_{\overline{x}}(x_i) = f(x_i)$, $i = 1, 2$ und $E_{\overline{x}}(\overline{x}) = f(\overline{x})$, woraus $E_{\overline{x}} = E_1 = f$ folgt.

Die Existenz einer Minimallösung $E \in V_1$ für beliebiges X ist aber nicht gewährleistet wie Beispiel 6.1 zeigt. Es können somit auch hier Schwierigkeiten, ähnlich wie beim Newtonsche Iterationsverfahren von Abschnitt 6.2, auftreten.

Beispiel 6.1

$$X = \{0, \frac{1}{2}, 1\},\ f(0) = f(\frac{1}{2}) = 0,\ f(1) = 1$$

Es sei $a_n = e^{-n}$, $t_n = n$ für $n \in \mathbb{N}$; mit $E_n(x) = a_n e^{t_n x}$ ist $(E_n)_{n \in \mathbb{N}} \subseteq V_1$ eine Minimalfolge für f auf X bzgl. V_1, denn es gilt

$$\lim_{n \to \infty} \| f - E_n \|_X = 0$$

Es gibt jedoch keine beste Approximation an f auf X bzgl. V_1.

Bemerkungen:

1. Weitergehende Untersuchungen, etwa Aussagen zur Konvergenz waren mir hier nicht möglich.

2. Es stellt sich die Frage, inwieweit man dieses Verfahren für die Approximation in V_N mit $N \geqq 2$ verallgemeinern kann und somit, im Gegensatz zum Newtonschen Iterationsverfahren, nur die Frequenzen von Näherungen durch Bestimmung von Nullstellen ermittelt und dann die linearen Parameter direkt berechnet; eventell kann man auf diesem Wege wie hier für $N = 1$ Aussagen über die Existenz bester Approximationen auf diskreten Punktmengen X mit $|X| = 2N + 1$ erhalten.

3. Zur numerischen Durchführung des Algorithmus, insbesondere zur Bestimmung der Nullstellen t_n beachte man Kapitel 9; die dort angegebenen Beispiele lassen das Verfahren als brauchbar erscheinen.

4. Die Idee zur Lösung des Gleichungssystems 6.10 geht zurück auf H.J. Maehly, Numerical Solution of a Certain Transcendental Equation Involving Exponentials (A Remark on a Paper of J.R.Rice), J.Soc.Indust.Appl.Math., Vol. 10 (1962).

Teil III

Numerische Experimente und Erprobung der Iterationsverfahren

7 Beispiele zum Verfahren von Braess

Für den Textteil von §7 von [26] wird auf Anhang B verwiesen.

In den Abschnitten 7.1, 7.2 und 7.3 von [26] wird das Verfahren von Braess für $f(x) = \sqrt{x}$ auf dem Intervall $I = [0, 1]$ mit N=1 und N=2 durchgeführt:

- *In Abschnitt 7.1 erfolgt die Approximation bezüglich V_1 nach Fall 1 und Fall 2*
- *In Abschnitt 7.2 werden Startfunktionen für die Iteration in V_2 berechnet, ebenfalls nach Fall 1 und Fall 2*
- *In Abschnitt 7.3 wird die Konstruktion einer Minimallösung für f bezüglich V_2 durchgeführt.*

In Abschnitt 7.4 von [26] wird in das Konvergenzverhalten des Verfahrens von Braess anhand der Funktion $f(x) = (1 + x)^{-1}$ in V_3 behandelt.

Anmerkung:
Diese Berechnungen wurden auch mit dem Programm BRAESS [28] durchgeführt, stets nach Fall 2, s. [29]

7.1 Approximation von $\mathbf{f(x) = \sqrt{x}}$ in [0,1] mit N = 1 nach Braess

Siehe §7.1 von [26]

7.2 Konstruktion von Startfunktionen

Siehe §7.2 von [26]

7.3 Zur Konstruktion der Minimallösung für $\mathbf{f(x) = \sqrt{x}}$ in [0,1], $\mathbf{N = 2}$

Siehe §7.3 von [26]

7.4 Zur Konvergenz des Algorithmus mit $\mathbf{f(x) = \frac{1}{1+x}}$ und $\mathbf{N = 3}$

Siehe §7.4 von [26]

8 Approximationen für die Riemannsche Zetafunktion nach dem Newtonschen Iterationsverfahren

Zur Erprobung des Newtonschen Iterationsverfahrens von Abschnitt 6.2 werden für die Riemannsche Zetafunktion

$$\zeta(x) = \sum_{n=1}^{\infty} \frac{1}{n^x}$$

in den Intervallen [2, 3] und [2, 4] Approximationen aus V_N für $1 \leqq N \leqq 4$ ermittelt.

Die Startfunktionen, die die Voraussetzung (6.8) erfüllen, sind nach dem Verfahren von Meinardus berechnet worden; es werden jeweils die Punkte x_0 und x_{2N-1} mit h angegeben.
Aus den gleichen Gründen wie in [26], §7 wird auch hier die Approximation auf einer diskreten Punktmenge X durchgeführt:

$$X = \{2 + i \times \frac{1}{256} \mid 0 \leqq i \leqq 256 \times d\}$$

Dabei ist $d = 1$, falls auf [2, 3], und $d = 2$, falls auf [2, 4] approximiert wird.

Für jeden Iterationsschritt werden die Punkte $x_0, \ldots, x_{2N}$ aus X (ohne zu runden) angegeben, mit denen das Newtonsche Verfahren zur Lösung von (6.9) durchgeführt wird; in dieen Punkten hat die durch die Startfunktion bzw. die durch die im vorausgehenden Iterationsschritt ermittelte Näherung F gegebenen Fehlerfunktion $F = \zeta - E$ eine lokales Extremum für $x \in X$. Nach jedem Schritt mit dem Newton-Verfahren werden die neu ermittelten Parameter a_i, t_i für $1 \leqq i \leqq N$ und die Werte F_i der entsprechenden Fehlerfunktion angegeben:

$$F_i = \zeta(x_i) - \sum_{j=1}^{N} a_j e^{t_j x_i} \quad 0 \leqq i \leqq 2N$$

Dabei wird zur Durchfhrung eines jeden Iterationsschrittes nach Newton verfahren, bis erfüllt ist:

$$\frac{\max_{0 \leqq j \leqq 2N} |F_j|}{\min_{0 \leqq j \leqq 2N} |F_j|} \leqq 1.2$$

Eine derartige Vorschrift ist wesentlich, wie die Approximationen für $N \geqq 2$ zeigen.

Wie in [26], §7 wird im Anschluß an jede Iteration eine Auswertung der letzten Fehlerfunktion (für das ganze Intevall) zur Abschätzung der Minimalabweichung und der erreichten Genauigkeit angegeben. Es sind dabei die Abschätzungen (8.3) zu berücksichtigen: für die Nullstellen und die lokalen Extrema im Innern des betreffenden Intevalles werden in Klammern eine obere Schranke für den Betrag des Fehlers angegeben. Abschätzungen für die Genauigkeit der Funktionswerte sind durch (8.3) gegeben.

Numerische Auswertung von $\zeta(\mathbf{x})$

Die Auswertung von $\zeta(x) = \sum_{n=1}^{\infty} \frac{1}{n^x}$ für $x > 1$ wird durch Anwendung der Eulerschen Summenformel nach K.Knopp, [10], vorgenommen; alle Seitenangaben beziehen sich auf [10].

Es gilt für $k \in \mathbb{N}$ ([10], p. 549):

$$\begin{aligned} \zeta(x) &= \frac{1}{x-1} + \frac{1}{2} + \sum_{i=1}^{k} \left(\frac{B_{2i}}{(2i)!} \prod_{j=0}^{2i-2} (x+j) \right) - R_k(x) \ , \quad x > 1 \ , \text{ mit} \\ R_k(x) &= \prod_{i=0}^{2k} (x+i) \int_1^{\infty} \frac{P_{2k+1}(t)}{t^{x+2k+1}} dt \ . \end{aligned}$$

Die Größen B_i sind die Bernoulli-Zahlen ([10], p. 185)

$$B_0 = 1,\ B_1 = \frac{1}{2},\ B_2 = \frac{1}{6},\ B_3 = B_5 = B_7 = \ldots = 0,\ B_4 = \frac{1}{30},\ B_6 = \frac{1}{42},\ \ldots \ ;$$

Die Bernoulli-Polynome P_{2k} und P_{2k+1} sind gegeben durch ([10], p. 541)

$$\begin{aligned} P_{2k}(t) &= (-1)^{k-1} \sum_{n=1}^{\infty} \frac{2 \cos 2n\pi t}{(2n\pi)^{2k}} \\ & \qquad\qquad\qquad\qquad k \in \mathbb{N} \qquad (8.1) \\ P_{2k+1}(t) &= (-1)^{k-1} \sum_{n=1}^{\infty} \frac{2 \sin 2n\pi t}{(2n\pi)^{2k+1}} \end{aligned}$$

Es wird $k = 3$ verwendet:

$$R_3(x) = \prod_{i=0}^{6} (x+i) \left[\int_1^{10} \frac{P_7(t)}{t^{x+7}} dt + \int_{10}^{\infty} \frac{P_7(t)}{t^{x+7}} dt \right] \qquad (8.2)$$

Mit $P_7(t) = \frac{1}{7!}\sum_{i=0}^{7}\binom{7}{i}B_i x^{7-i}$ ([10], p. 542) erhält man nach umfangreicher, jedoch elementarer Rechnung

$$\begin{aligned}\int_1^{10}\frac{P_7(t)}{t^{x+7}}dt &= \sum_{n=1}^{9}\int_n^{n+1}\frac{P_7(t)}{t^{x+7}}dt \\ &= -\sum_{n=1}^{9}\frac{(n+1)^{-x}}{\prod_{i=0}^{6}(x+i)} + \frac{1-10^{-x-5}}{7!6(x+6)(x+5)} + \frac{10^{-x-3}-1}{6!\prod_{i=3}^{6}(x+i)} + \\ &\quad + \frac{1-10^{-x-1}}{12\prod_{i=1}^{6}(x+i)} + \frac{10^{-x}-1}{2\prod_{i=0}^{6}(x+i)} + \frac{1-10^{-x+1}}{(x-1)\prod_{i=0}^{6}(x+i)}\end{aligned}$$

Zur Abschätzung von

$$R(x) := \prod_{i=0}^{6}(x+i)\int_{10}^{\infty}\frac{P_7(t)}{t^{x+7}}dt$$

wird $|P_k(t)| \leqq \frac{2}{(2\pi)^k}\sum_{n=1}^{\infty}\frac{1}{n^k}$, $k \in \mathbb{N}$, benutzt, was aus (8.1) folgt.

Mit $\sum_{n=1}^{\infty}\frac{1}{n^{2k}} = (-1)^{k-1}\frac{B_{2k}(2\pi)^{2k}}{2(2k)!}$ ([10], p. 245) für $k \in \mathbb{N}$ erhält man speziell für P_{2k+1}:

$$|P_{2k+1}(t)| \leqq \frac{2}{(2\pi)^{2k+1}}\sum_{n=1}^{\infty}\frac{1}{n^{2k+1}} < \frac{2}{(2\pi)^{2k+1}}\sum_{n=1}^{\infty}\frac{1}{n^{2k}} \leqq \frac{1}{2\pi}\frac{B_{2k}}{(2k)!} \qquad k \in \mathbb{N}$$

Es folgt:

$$\begin{aligned}|R(x)| < \prod_{i=0}^{6}(x+i)\frac{B_6}{2\pi 6!}\int_{10}^{\infty}t^{-x-7}dt &= \prod_{i=0}^{6}(x+i)\frac{B_6}{2\pi 6!}\frac{1}{x+6}10^{-x-6} = \\ &= \frac{B_6}{2\pi 6!}\prod_{i=0}^{5}(x+i)10^{-x-6}\end{aligned}$$

$|R(x)|$ ist im Intervall $(1.1072, \infty)$ monoton fallend; spezielle Werte:

$$|R(x)| < \begin{cases} 2.7\times 10^{-10} & : x \geqq 2.0 \\ 1.8\times 10^{-10} & : x \geqq 2.5 \\ 1.1\times 10^{-10} & : x \geqq 3.0 \\ 0.6\times 10^{-10} & : x \geqq 3.5 \\ 3.2\times 10^{-11} & : x \geqq 4.0 \end{cases} \tag{8.3}$$

Mit (8.2) erhält man nach einigen Zwischenrechnungen:

$$\begin{aligned}\zeta(x) &= \frac{1}{x-1}+\frac{1}{2}+\frac{B_2}{2!}x+\frac{B_4}{4!}x(x+1)(x+2)+\frac{B_6}{6!}x(x+1)(x+2)(x+3)(x+4)- \\ &\quad -\prod_{i=0}^{6}(x+i)\int_1^{10}\frac{P_7(t)}{t^{x+7}}dt-R(x)= \\ &= 1+\sum_{n=2}^{9}\frac{1}{n^x}+\frac{x(x+1)(x+2)[(x+3)(x+4)-4200]}{30240\times 10^{x+5}}+ \\ &\quad +\frac{1}{10^{x+1}}(\frac{x}{12}+\frac{5x+95}{x-1})-R(x)\ , \qquad x>1\ .\end{aligned}$$

Für die folgenden Berechnungen von $\zeta(x)$, $x \in [2,4]$, wird $R(x)$ vernachlässigt. Abschätzungen für den hierdurch entstehenden Fehler sind also mit (8.3) gegeben.

Die Berechnungen: Siehe §8 von [26]

9 Zur Approximation bezüglich V_1 nach dem Verfahren von Abschnitt 6.3

Mit dem Verfahren von Abschnitt 6.3 ist die Approximation von $f(x) = \sqrt{x}$ auf [0,1] und der Riemannschen Zetafunktion ζ auf [2,3] und [2,4] durchgeführt worden; Man beachte hierzu Bemerkung 6.1.

Zur Konstruktion einer Startfunktion auf Intervall [a,b] werden die Punkte $x_1 = a$, $x_2 = \frac{a+b}{2}$, $x_3 = b$ verwendet.
Die Frequenzen t_n werden iterativ nach Newton-Raphson ermittelt: der Startwert für die Bestimmung von t_1 ist stets 0, für $n \geqq 2$ wird t_{n-1} als Startwert für die iterative Bestimmung von t_n verwendet.

Die Approximationen werden wie oben auf diskreten Punktmengen X durchgeführt. Die Auswertungen der Fehlerfunktionen erfolgen wieder für das gesamte Intervall.

In den Abschnitten 9.2 und 9.3 erfolgt die Auswertung der Riemannschen Zetafunktion wie in Kapitel 8; entsprechend gelten die Bemerkungen über die Genauigkeit der angegebenen Werte.
Für den Vergleich der Ergebnisse mit denen von Kapitel 8 beachte man, daß X in den Abschnitten 9.2 und 9.3 weniger Punkte enthält als in Kapitel 8.

Anmerkung (Februar 2023):
Diese Berechnungen wurden nochmals mit EXPFJP [30] durchgeführt.

9.1 Approximation von $f(x) = \sqrt{x}$ auf [0,1]

Es ist $X = \{\frac{i}{100} \mid 0 \leqq i \leqq 100\}$.

1. Iterationsschritt:

$x_i = 0.0 \qquad x_2 = 0.5 \qquad x_3 = 1.0$

5 Schritte nach Newton-Raphson ergeben:

`t1=+1.762 747 174 038` ; Fehler: $\pm\ 10^{-12}$

Es folgt:

`a1=0.207 106 781 186`

Zeichung 9.1 zeigt die Fehlerfunktion.

2. Iterationsschritt:

$x_i = 0.0$ $x_2 = 0.42$ $x_3 = 1.0$

3 Schritte nach Newton-Raphson ergeben:

`t2=+1.751 452 157 150` ; Fehler: $\pm\ 10^{-12}$

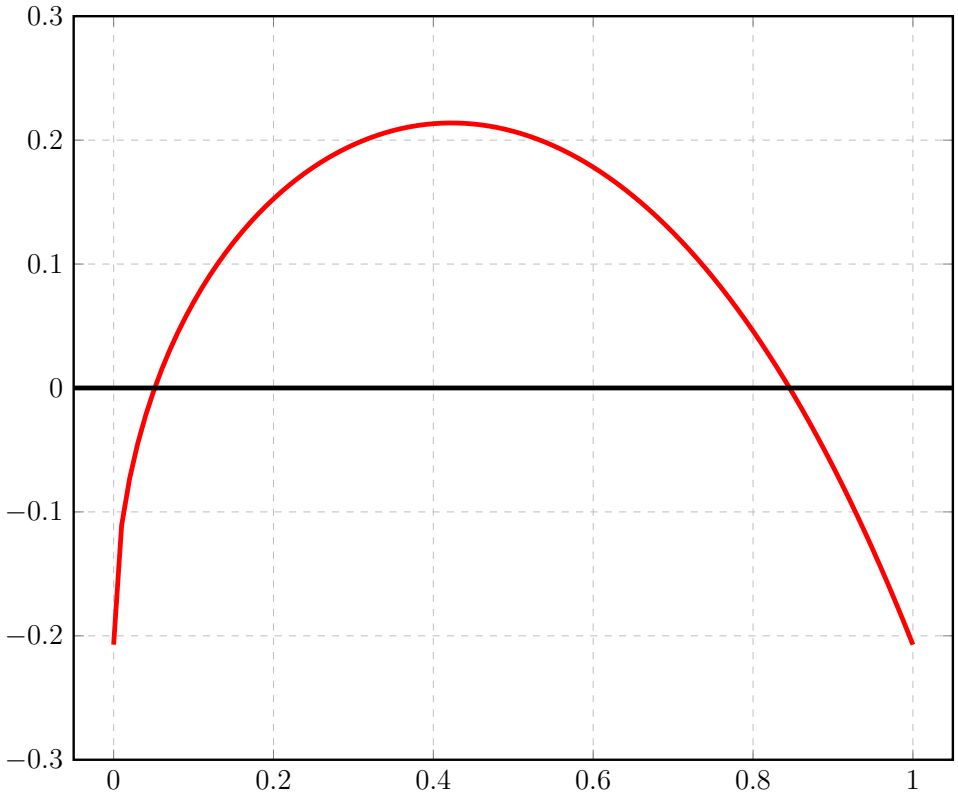
Es folgt:

`a2=0.209 953 239 189`

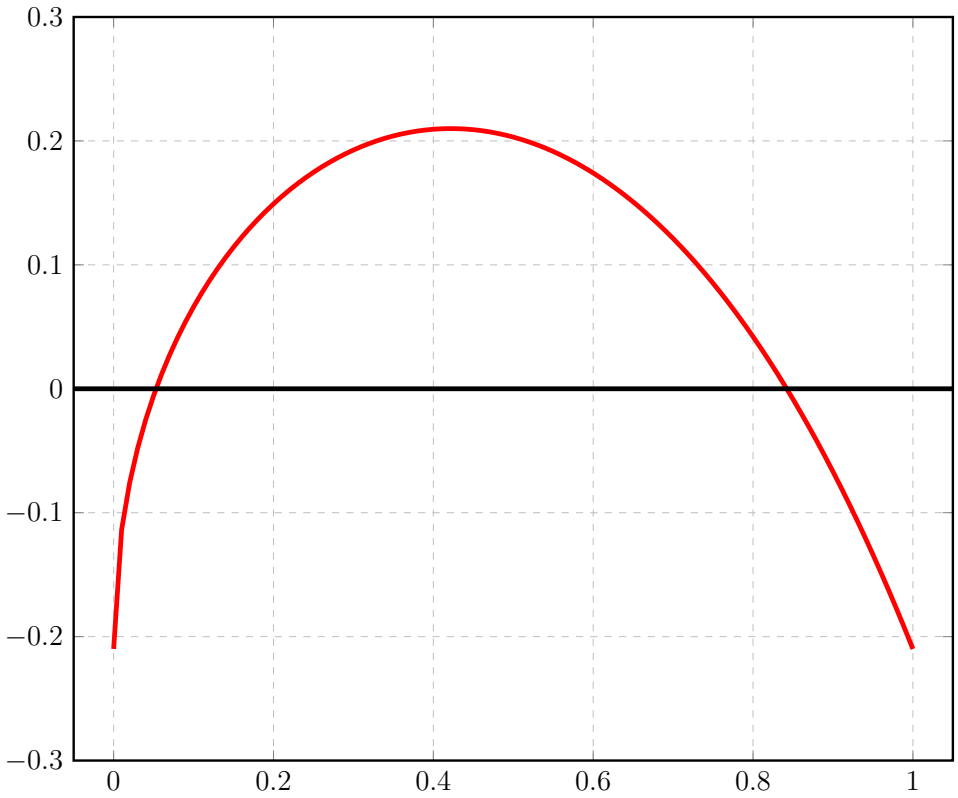
Zeichung 9.2 zeigt die Fehlerfunktion.

Ende der Iteration

Auswertung der Fehlerfunktion		
Die Nullstellen	Lokale Extrema und Funktionswerte	
	0.0	-0.2099532392
0.0530896736		
	0.4218434150	+0.2099570814
0.8421382898		
	1.0	-0.2099532392



Zeichnung 9.1



Zeichnung 9.2

9.2 Approximation von $f(x) = \zeta(x)$ auf [2,3]

Es ist $X = \{2.0 + \frac{i}{128} \mid 0 \leqq i \leqq 128\}$.

1. Iterationsschritt:

$x_i = 2.0 \qquad x_2 = 2.5 \qquad x_3 = 3.0$

4 Schritte nach Newton-Raphson ergeben:

`t1=-0.321 034 683 610` ; Fehler: $\pm\ 10^{-12}$

Es folgt:

`a1=+3.064 974 872 707`

Zeichung 9.3 zeigt die Fehlerfunktion.

2. Iterationsschritt:

$x_i = 2.0$ $x_2 = 2.390625$ $x_3 = 3.0$

3 Schritte nach Newton-Raphson ergeben:

`t2=-0.321 380 719 310` ; Fehler: $\pm\ 10^{-12}$

Es folgt:

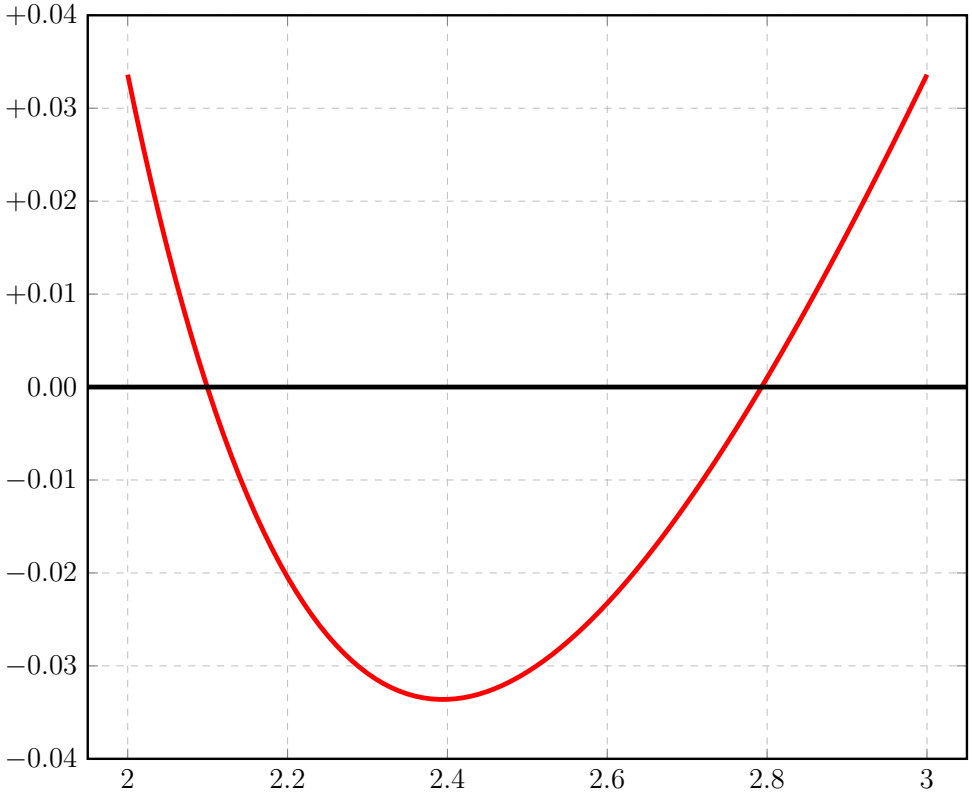
`a2=+3.064 296 199 406`

Zeichung 9.4 zeigt die Fehlerfunktion.

Ende der Iteration

Auswertung der Fehlerfunktion		
Die Nullstellen	Lokale Extrema und Funktionswerte	
2.09944560	2.0	+0.033 609 618 743
2.79300401	2.394200(3.4e-5)	-0.033 613 341 238
	3.0	+0.033 609 618 743

Zeichnung 9.3



Zeichnung 9.4

9.3 Approximation von $f(x) = \zeta(x)$ auf [2,4]

Es ist $X = \{2.0 + \frac{i}{128} \mid 0 \leqq i \leqq 256\}$.

1. Iterationsschritt:

$x_i = 2.0 \qquad x_2 = 3.0 \qquad x_3 = 4.0$

4 Schritte nach Newton-Raphson ergeben:

`t1=-0.220 167 923 889` ; Fehler: $\pm\ 10^{-12}$

Es folgt:

`a1=+2.453 432 894 140`

Zeichung 9.5 zeigt die Fehlerfunktion.

2. Iterationsschritt:

$x_i = 2.0 \qquad x_2 = 2.6796875 \qquad x_3 = 4.0$

3 Schritte nach Newton-Raphson ergeben:

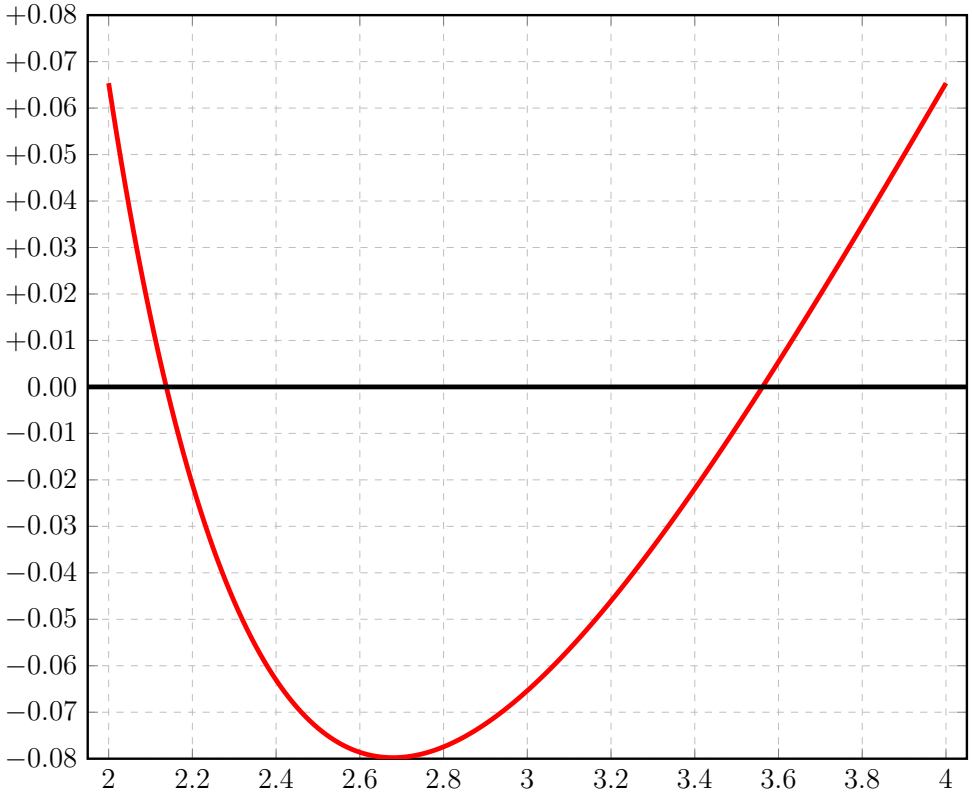
`t2=-0.221 422 241 102` ; Fehler: $\pm\ 10^{-12}$

Es folgt:

```
a2=+2.448 506 467 119
```
Zeichung 9.6 zeigt die Fehlerfunktion.

Ende der Iteration

Auswertung der Fehlerfunktion		
Die Nullstellen	Lokale Extrema und Funktionswerte	
	2.0	+0.072 485 893 085
2.15753518		
	2.678848(3.2e-5)	-0.072 486 014 988
3.50969112		
	4.0	+0.072 485 893 085



Zeichnung 9.5

Zeichnung 9.6

10 Das Lokale Kolmogoroff-Kriterium bei der Konstruktion besserer Approximationen

Es sei $f \in C(I)$ und $E(s) \in V_N$ gegeben.
Im Beweis von Theorem 89 in [12] (Satz 2.10 in [12]) wird gezeigt:

Gibt es ein $b \in \mathbb{R}^{2Nmit}$ mit $(b, grad(E(a, x_k))(f(x_k) - E(a, x_k)) > 0$ für $1 \leqq k \leqq m$, wobei $\{x_k \mid 1 \leqq k \leqq m\}$ die Menge der Extremalpunkte von $f - E(a)$ in I ist, dann gibt es ein $C > 0$, so dass

$$\| f - E(a + cb) \|_I < \| f - E(a) \|_I$$

für $c \in (0, C)$ ist.

Es sei $E(a, x) = \sum_{i=1}^{N} a_i e^{t_i x}$. Gibt es in V_N^0 bessere Approximationen als E(a) für f, so wird zur Konstruktion besserer Approximationen folgender Ansatz nahegelegt, falls $f - E(a)$ endlich viele Extremalpunkte in I besitzt:

1. Man löse die lineare Optimierungsaufgabe

$$(10.1)\ (f(x_k) - E(a, x_k)) \sum_{i=1}^{N} (b_i + b_{N+i} a_i x_k) e^{t_i x_k} - \mu \geqq 0 \quad 1 \leqq k \leqq m$$

$$(10.2) \qquad \begin{aligned} b_i &\geqq -1 \\ &\qquad 1 \leqq i \leqq 2N \\ b_i &\leqq +1 \end{aligned}$$

$$\mu = \text{Max!}$$

mit den Unbekannen μ, b_i, $1 \leqq i \leqq 2N$; dabei besitzt $f - E(a)$ in I genau m Extremalpunkte x_k, $1 \leqq k \leqq m < \infty$.

2. Erhält man eine Lösung mit $\mu > 0$, dann setze man $c = 1$ und bestimme durch sukzessives Halbieren ein $c > 0$ mit

$$\| f - E(a + cb) \|_I < \| f - E(a) \|_I \ ,$$

wobei $b = (b_1, \ldots, b_{2N})$ ist.

Hierzu ein **Beispiel**:
Es ist $N = 3$, $f(x) = \frac{1}{1+x}$, $I = [0, 1]$; die Approximation wird auf
$X = \{\frac{i}{100} \mid 0 \leqq i \leqq 100\}$ durchgeführt[30].

Die Ausgangsfunktion $E(a)$ ist gegeben durch die Parameter

```
a1=+0.053 824 480 , t1=-4.282 493 872
a2=+0.400 482 842 , t2=-1.540 603 343
a3=+0.545 644 214 , t3=-0.277 434 496
```

Zeichnung 10.1 stellt die Fehlerfunktion dar.

1. Iterationsschritt:
Es ist hier $m = 1$ und $x_1 = 0.0$; als Lösung der Optimierungsaufgabe erhält man $b_1 = b_2 = b_3 = 1.0$, $b_4 = b_5 = b_6 = 0.0$, $\mu = 0.000145$.
Eine bessere Approximation erhält man mit $c = 2^{-17}$:

```
a1=+0.053 832 109 , t1=-4.282 493 872
a2=+0.400 490 472 , t2=-1.540 603 343
a3=+0.545 651 843 , t3=-0.277 434 496
```

Zeichnung 10.2 stellt die Fehlerfunktion dar.

2. Iterationsschritt:
Es wird hier $m = 2$ und $x_1 = 0.0$, $x_2 = 0.13$ verwendet. Als Lösung der Optimierungsaufgabe erhält man $b_1 = b_2 = b_3 = 1.0$, $b_4 = b_5 = b_6 = 0.0$, $\mu = 0.000077$. Eine bessere Approximation erhält man mit $c = 2^{-20}$:

```
a1=+0.053 833 063 , t1=-4.282 493 872
a2=+0.400 491 426 , t2=-1.540 603 343
a3=+0.545 652 797 , t3=-0.277 434 496
```

Zeichnung 10.3 stellt die Fehlerfunktion dar.

Bei Fortsetzung der Iteration erhält man keine wesentlich bessere Approximation mehr: c muss von Schritt zu Schritt kleiner gewählt werden und als Lösung der Optimierungsaufgabe erhält man $b_1 = b_2 = b_3 = 1.0$, $b_4 = b_5 = b_6 = 0.0$; an den Frequenzen wird also nichts geändert. Dies gilt auch dann, wenn man m größer wählt.

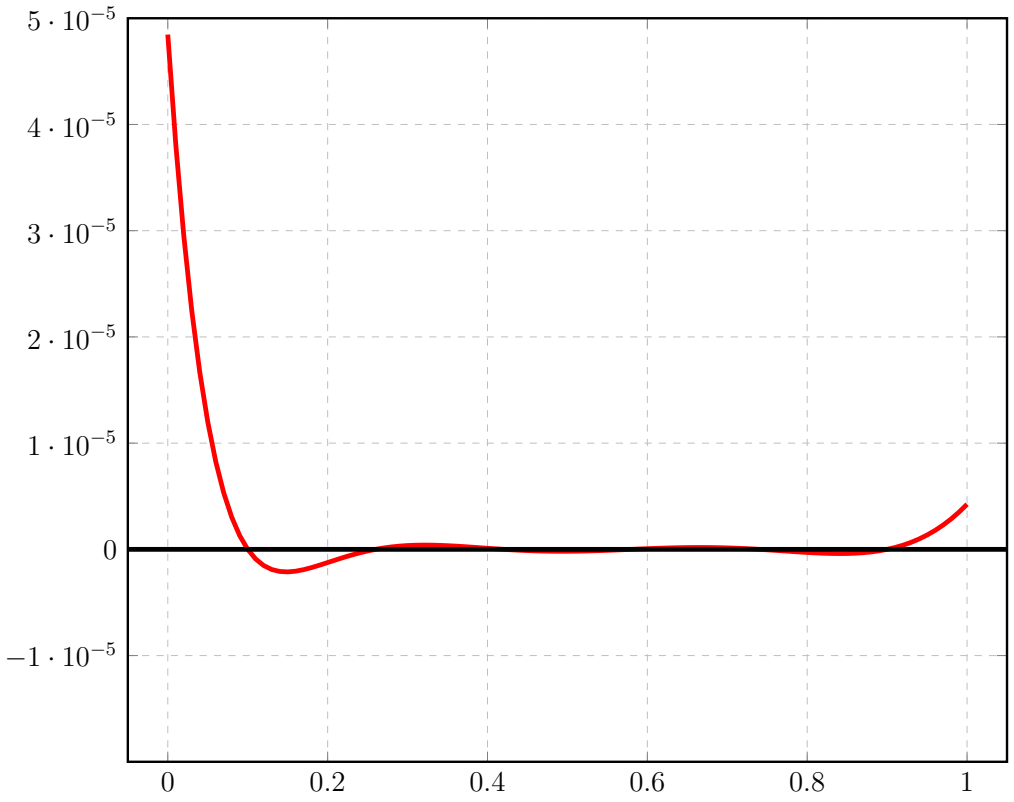
Es sei $E(a)$ die hier gegebene Startfunktion, $E(a')$ die mit dem achten Iterationsschritt von 7,4A gegebene Lösung und $(b_1, \ldots, b_6) = a' - a$; die b_i erfüllen

[30] Die Optimierungsaufgabe ist nach dem Simplex-Algorithmus mit dem FORTRAN-Unterprogramm SIMPX des Rechenzentrums gelöst worden

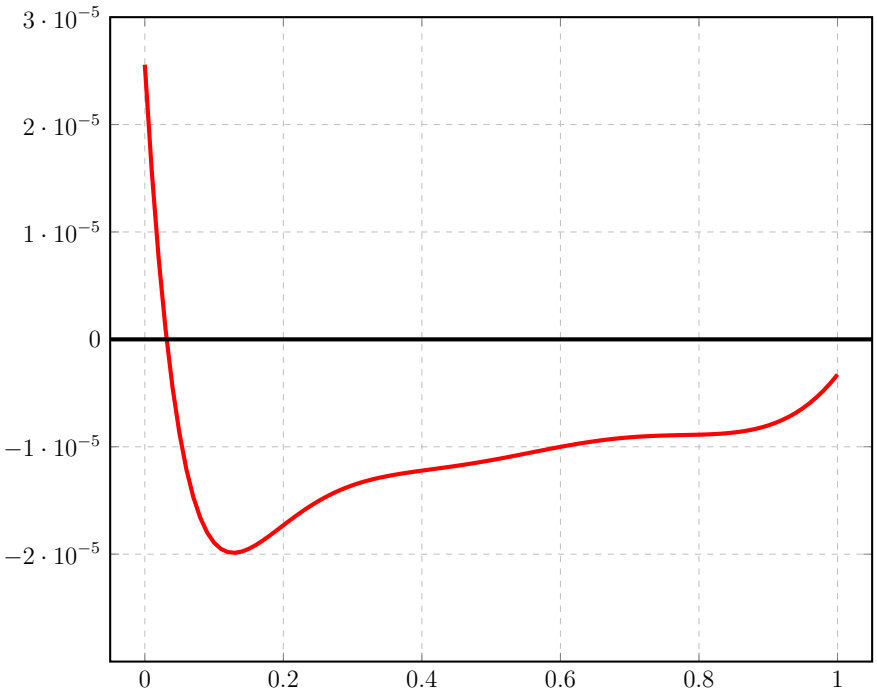
die Restriktionen (10.2) und mit diesen Werten folgt aus (10.1) für μ:

$$\mu \leqq 2.4 \times 10^{-9}$$

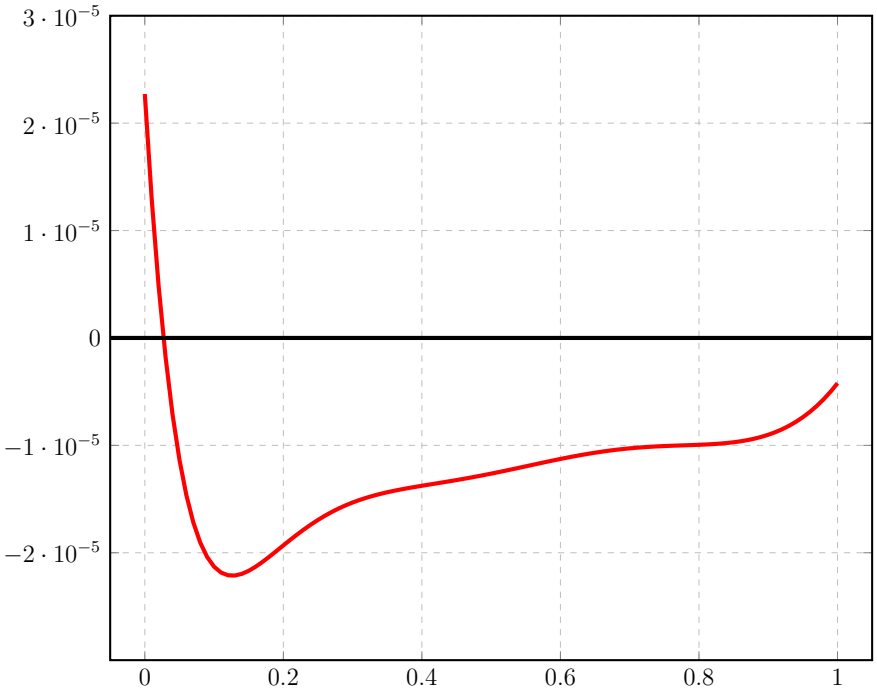
Aus diesen Ergebnissen kann man folgern, dass im allgemeinen das Verfahren in dieser Form für die praktische Ermittlung von Minimallösungen wenig brauchbar ist, da hierfür μ nicht minimal sein muss.



Zeichnung 10.1



Zeichnung 10.2



Zeichnung 10.3

Literatur

[1] D. Braess.
Approximation mit Exponentialsummen.
Computing 2 (1967), 309-321

[2] D. Braess.
Über die Vorzeichenstruktur der Exponentialsummen.
J.Approximation Theory 3 (1970), 101-113

[3] D. Braess.
Die Konstruktion der Tschebyscheff-Approximierenden bei der Anpassung mit Exponentialsummen.
J.Approximation Theory 3 (1970), 261-273

[4] E. Cesàro.
Elementares Lehrbuch der algebraischen Analysis und der Infinitesimalrechnung.
Leipzig, 1904, pp 77-79

[5] F. Erwe.
Gewöhnliche Differentialgleichungen.
Bibliographisches Institut, Mannheim.

[6] H. Grauert - W. Fischer.
Differential- und Integralrechnung II.
Springer-Verlag, Berlin-Heidelberg-New York, 1968.

[7] E. Kamke.
Differentialgleichungen I.
Leipzig, 1964

[8] S. Karlin - W. J. Studden.
Tchebycheff Systems: With Applications in Analysis and Statistics.
Interscience Publ., New York-London-Sydney, 1966

[9] L. G. Kelly.
Handbook of Numerical Methods and Applications.
Addison-Wesely Publishing Company, 1967

[10] K. Knopp.
Theorie und Anwendung der Unendlichen Reihen.
Springer-Verlag, Berlin-Göttingen-Heidelberg-New York, 1964

[11] G. Meinardus u. D. Schwedt.
Nichtlineare Approximationen.
Arch.Rat.Mech.Anal. 17 (1967), 297-326

[12] G. Meinardus.
Approximation of Functions: Theory and Numerical Methods.
Springer-Verlag, Berlin-Heidelberg-New York, 1967

[13] G. Meinardus.
Die unter [13] zitierten Angaben gehen auf *mündliche Mitteilungen* zurück.

[14] S. Ramanujan.
Note on a Set of Simultaneous Equations. In: Collected Papers of Srinivasa Ramanujan. Chelsea Publishing Company, New York, 1962

[15] J. R. Rice.
The Approximation of Functions, Vol. II.
Addison-Wesely Publishing Company, 1969

[16] John R. Rice.
Chebyshev Approximation by Exponentials.
J.Soc.Indust.Appl.Math. Vol. 10, No. 1 (March 1962), 149-161

[17] E. Schmidt.
Zur Kompaktheit bei Exponentialsummen.
J.Approximation Theory 3 (1970), 445-454

[18] E. Schmidt.
Normalität und Stetigkeit bei der Tschebyscheff-Approximation mit Exponentialsummen.
Dissertation, Münster 1968

[19] J. F. Steffenson.
Interpolation.
Chelsea Publishing Company, New York, 1965

[20] H. Werner.
Der Existenzsatz für das Tschebyscheffsche Approximationsproblem mit Exponentialsummen.
In: Funktionalanalytische Methoden der numerischen Mathematik, Vortragsauszüge der Tagung über funktionalanalytische Methoden der numerischen Mathematik vom 19. bis 25. November 1967 im Mathematischen Forschungsinstitut Oberwolfach, herausgegeben von L. Collatz und H. Unger.
Birkhäuser Verlag Basel und Stuttgart, 1969

[21] H. Werner.
Tschebyscheff-Approximationsproblem with Sums of Exponentials.
In: Approximation Theory, Proceedings of a Symposium held at Lancaster, July 1969. Edited by A. Talbot.
Academic Press, London and New York, 1970

[22] H. Werner - R. Schaback.
Praktische Mathematik II.
Springer-Verlag, Berlin-Heidelberg-New York, 1972

[23] F. A. Willers.
Methoden der praktischen Analysis.
Berlin, 1971

[24] B. Brosowski.
Nicht-lineare Tschebyscheff-Approximation.
Bibliographisches Institut, Mannheim, 1968

[25] F. R. Gantmacher.
Matrizenrechnung II.
Berlin, 1959

[26] F. J. Polster.
Exponentielle Approximation: Theorie und Numerische Verfahren.
Diplomarbeit, FAU Erlangen, 1973

[27] F. J. Polster.
MEINARDUS: Ein Programm zur Exponentiellen Approximation bzgl. V_N.
Januar 2023

[28] F. J. Polster.
BRAESS: Ein Programm zur Exponentiellen Approximation bzgl. V_N.
Januar 2023

[29] F. J. Polster.
BRAESS: Berechnungen
Januar 2023

[30] F. J. Polster.
EXPFJP: Ein Programm zur Exponentiellen Approximation bzgl. V_1.
Januar 2023

Anhang A:
Herleitung der Gleichungen (6.11)

1. Aus der Gleichung (6.10) mit $i = 3$ folgt unmittelbar

$$-r = f(x_3) - ae^{sx_3}$$

2. Ersetzen von r in der Gleichung (6.10) mit $i = 2$:

$$\begin{aligned} ae^{sx_2} &= f(x_2) - r = f(x_2) + f(x_3) - ae^{sx_3} \\ \Rightarrow \quad ae^{sx_3} + ae^{sx_2} &= f(x_3) + f(x_2) \\ \Rightarrow \quad a &= \frac{f(x_3) + f(x_2)}{e^{sx_3} + e^{sx_2}} \end{aligned}$$

3. Ersetzen von r und a in der Gleichung (6.10) mit $i = 1$:

$$\begin{aligned} ae^{sx_1} - r &= f(x_1) \\ \frac{f(x_3) + f(x_2)}{e^{sx_3} + e^{sx_2}} e^{sx_1} + f(x_3) - ae^{sx_3} &= f(x_1) \\ \frac{f(x_3) + f(x_2)}{e^{sx_3} + e^{sx_2}} e^{sx_1} + f(x_3) - \frac{f(x_3) + f(x_2)}{e^{sx_3} + e^{sx_2}} e^{sx_3} &= f(x_1) \\ \frac{e^{sx_1}(f(x_3) + f(x_2)) - e^{sx_3}(f(x_3) + f(x_2)}{e^{sx_3} + e^{sx_2}} &= f(x_1) - f(x_3) \\ \frac{(e^{sx_1} - e^{sx_3})(f(x_3) + f(x_2))}{e^{sx_3} + e^{sx_2}} &= f(x_1) - f(x_3) \\ \Rightarrow (-1) \times \frac{(e^{sx_1} - e^{sx_3})}{e^{sx_3} + e^{sx_2}} &= \frac{f(x_1) - f(x_3)}{f(x_3) + f(x_2} \times (-1) \\ \text{also:} \qquad \frac{(e^{sx_3} - e^{sx_1})}{e^{sx_3} + e^{sx_2}} &= \frac{f(x_3) - f(x_1)}{f(x_3) + f(x_2} \end{aligned}$$

Anhang B:
§7 der Diplomarbeit (LaTeX-Version)[31]

Vorbemerkung:
Zur technischen Durchführung der folgenden Berechnungen gelten die Angaben von Abschnitt 5.4; da hier direkt Listen des Rechenzentrums verwendet werden, sind die Zahlen in Gleitkommadarstellung angegeben. Wie in Abschnitt 5.4 werden in den Zeichnungen die jeweiligen Fehlerfunktionen dargestellt; weiter werden die Bezeichnungen, wie sie zur Herleitung der Verfahren benutzt worden sind, beibehalten.

§7 Beispiele zum Verfahren von Braess

Bemerkungen, Erläuterungen

1. Es ist hier stets $I = [0, 1]$.

2. Der Algorithmus von Braess wird, falls nicht anders angegeben, gemäß Fall 2 durchgeführt; die lineare Approximation nach Schritt 2 des Verfahrens wird mit dem *allgemeinen Remez-Algorithmus* (s. z.B. [12], Abschnitt 7.1, p. 107) ermittelt.

   Um dabei nicht die Differenzierbarkeit der hier zur Demonstration benutzten Funktionen f zu verwenden, wird die Approximation auf den Punkten
   $$x_i = \frac{i}{100}, \quad 0 \leqq i \leqq 100$$
   durchgeführt; der hierbei entstehende Diskretisierungsfehler kann vernachlässigt werden. Es wird so nach Remez iteriert, daß

   $$\frac{\max\limits_{0 \leqq i \leqq 100} |F(x_i) - e'(x_i)|}{\max\limits_{0 \leqq i \leqq 100} |F(x_i) - e(x_i)|} \leqq 1.01 \tag{7.1}$$

   erfüllt ist, wobei e die beste lineare Approximation an F auf x_i, $0 \leqq i \leqq 100$, und e' eine durch die Remez-Iteration gegebene Näherung für e sei. F ist die nach Schritt 2 des Braess-Verfahrens zu approximierende Fehlerfunktion.

[31] dies ist der Textteil von [26], §7, inkl. der einleitenden "Vorbemerkung" von "Teil III"

Zu jedem Iterationsschritt werden zunächst die Parameter r_i der linearen Approximation $\sum_{i=1}^{N}(r_i + r_{N+i}a_i x)\mathrm{e}^{t_i x}$ angegeben, wobei $\sum_{i=1}^{N} a_i \mathrm{e}^{t_i x}$ die entsprechende Startfunktion bzw. die im vorhergehenden Iterationsschritt ermittelte Näherung ist.

3. Der nach Schritt 3 des Braess-Verfahrens ermittelte Faktor, mit dem eine bessere Approximation berechnet wird, ist mit C bezeichnet und wird als Bruch angegeben, so daß dem Nenner leicht entnommen werden kann, wie oft zur Durchführung von Schritt 3 die Norm der Fehlerfunktion bestimmt werden muß; dann folgen als Resultat des Iterationsschrittes die Parameter der somit gegebenen besseren Approximation, die wieder mit $a_i, t_i, 1 \leqq i \leqq N$, bezeichnet werden. (Bei der Indizierung wird der Iterationsschritt nicht angegeben, da keine Verwechslungen zu befürchten sind.)

4. An jede Iteration schließt sich eine Auswertung der Fehlerfunktion $f-E$ auf ganz I an, wobei E die im letzten Iterationsschritt ermittelte Näherung sei; es werden die Nullstellen und lokalen Extrema von $f - E$ mit den dazugehörigen Funktionswerten angegeben, so daß eine Abschätzung der Minimalabweichung von f möglich ist.

Die Experimente

In den Abschnitten 7.1, 7.2 und 7.3 wird der Algorithmus für $f(x) = \sqrt{x}$ mit N=1 und N=2 durchgeführt, Abschnitt 7.4 behandelt das Konvergenzverhalten dieses Verfahrens anhand $f(x) = (1 + x)^{-1}$ und $N = 3$.

In **Abschnitt 7.1** erfolgt zunächst die Approximation bezüglich V_1 nach Fall 2. Während die Iteration von 7.1A wie oben beschrieben durchgeführt ist, wird in 7.1B die Remez-Iteration bereits beendet, wenn der Quotient von (7.1) kleiner als 2 ist. Für die Rechnung bedeutet dies, daß in jedem Iterationsschritt von 7.1B nur 2-mal, in den beiden ersten Schritten von 7.1A jedoch 4-mal und in den folgenden 3-mal nach Remez iteriert wird.
In 7.1B ist daher der Rechenaufwand beachtlich kleiner als in 7.1A, nach dem dritten Iterationsschritt ist aber mit den so bestimmten Parametern r_1 und r_2 keine Verbesserung nach Schritt 3 mehr möglich; andererseits dürfte für viele Zwecke die erreichte Genauigkeit genügen.

In den Abschnitten 7.2 und 7.3 werden zur Approximation von f bezüglich V_2 die Berechnungen des Algorithmus sowohl für Fall 2 als auch Fall 1 durchgeführt:
In **Abschnitt 7.2** werden Startfunktionen für die Iteration in V_2 angegeben, wobei aus Platzgründen die Parameter der linearen Approximation nicht aufgeführt werden.
Wie in §6 beschrieben, geht man zur Ermittlung von Startfunktionen aus von

$$a_1 e^{t_1 x} + a_2 e^{t_2 x}, \qquad a_2 = 0, t_2 \neq t_1 \ ,$$

wobei a_1, t_1 die im 5-ten Iterationsschritt von 7.1A ermittelten Parameter sind.
In **7.2A** sind die Ergebnisse für die Berechnung nach Fall 2 angegeben.
Für die Berechnung von **7.2B** nach Fall 1 ist

$$X = \{\ 0.0,\ 0.3,\ 0.415,\ 0.6,\ 1.0\ \}$$

benutzt worden; man beachte hierzu Zeichnung 10.
Es sind die Startfunktionen für $t_2 \in \mathbb{N}$ mit $-15 \leqq t_2 \leqq 14$ angegeben:

- für $t_2 \leqq 1$ liegen die Startfunktionen in $V_2(-,+)$
- für $t_2 \geqq 2$ liegen die Startfunktionen in $V_2(+,-)$,

was auch aus §4 folgt.
Die Norm der mit diesen Startfunktionen gegebenen Fehlerfunktionen ist auf ganz I ermittelt worden; sowohl in 7.2A wie in 7.2B erhält man bei $t_2 = -10$ ein Minimum für die Norm der Fehlerfunktion.

In **Abschnitt 7.3** wird die Konstruktion einer Minimallösung für f bezüglich V_2 durchgeführt.
Die Ergebnisse von Abschnitt 7.2 legen es nahe, Startfunktionen mit $t_2 \leqq 1.0$ zu verwenden; die Iteration von 7.3A bzw. 7.3B geht von der Startfunktion mit $t_2 = -5.0$ von 7.2A bzw. 7.2B aus:

- in 7.3A erfolgt die Berechnung nach Fall 2
- in 7.3B wird gemäß Fall 1 mit $|X| = 5$ verfahren, so daß man die lineare Approximation nach Schritt 2 durch Lösung eines linearen Gleichungssystems erhält.

Es zeigt sich, daß bei Verwendung der Startfunktionen mit

$$t_2 \in \{-7.0, -9.0, -11.0, -13.0, -15.0\}$$

ebenfalls 5 Iterationsschritte nötig sind, um die Genauigkeit zu erhalten, wie sie in 7.3A bzw. 7.3B erzielt worden ist; die Iterationen, die von den

Startfunktionen mit $t_2 = -9.0$ oder $t_2 = -11.0$ ausgehen, konvergieren also nicht schneller, was man (nachträglich) wohl erwarten würde.

Verwendet man wie in 7.3C und 7.3D die Startfunktionen mit $t_2 = 0.0$, so muß der Faktor C zur Verbesserung so klein gewählt werden, daß die Folgen schlecht konvergieren; dies ist unabhängig davon, ob nach Fall 2 wie in 7.3C oder nach Fall 1 wie in 7.3D gerechnet wird. Die Iteration von 7.3C ergibt erst mit dem 25. Iterationsschritt ene Lösung mit der Genauigkeit, wie sie in 7.3A gegeben ist.

Geht man von den Startfunktionen mit $t_2 \geqq 2.0$ aus, so erhält man Iterationen, die äußerst langsam konvergieren. Die Startfunktion mit $t_2 = 7.0$ in 7.2A ergibt (nach Fall 2 berechnet) nach 180 Iterationsschritten eine Näherung mit den Parametern

$$\begin{aligned} a_1 &= +0.191596 \qquad & t_1 &= +2.514113 \\ a_2 &= -0.007498 \qquad & t_2 &= +5.072941 \end{aligned}$$

(Zeichnung 26)

In jedem Iterationsschritt ist der Faktor C kleiner als $\frac{1}{4096}$ [32]. Da man hier nur vermuten kann, daß die beiden Frequenzen bei Fortsetzung der Iteration gegen einen gemeinsamen Grenzwert konvergieren, werden die Schwierigkeiten deutlich, die die Anwendung des Algorithmus bei der Approximation ohne Vorzeichenbeschränkung mit sich bringen kann.

Die Iterationen in V_2 verlaufen bereits bei Bestimmung von C nach Schritt 3 stets in nur einer Vorzeichenklasse von V_2.

Diese Ergebnisse lassen es sinnvoll erscheinen, bei der Konstruktion einer Minimallösung nach Braess im allgemeinen mehrere Startfunktionen zu konstruieren und zunächst mehrere Iterationen nebeneinander durchzuführen. So kann man hoffen, eine Iteration zu erhalten, die mit vertretbarem Aufwand eine Lösung liefert. Hätte man sich bei der Approximation von f bezüglich V_2 in$V_2(-,+)$ auf die Iteration von 7.3C und in $V_2(+,-)$ auf die oben angegebene Iteration beschränkt, so wäre trotz ernormen Rechenaufwands lange ungeklärt geblieben, in welcher Vorzeichenklasse die Lösung enthalten ist.

[32] $\frac{1}{4096} = 0.0002441$, $4096 = 2^{-12}$

In **Abschnitt 7.4** wird in das Konvergenzverhalten dieses Verfahrens anhand der Funktion $f(x) = (1+x)^{-1}$ mit $N = 3$ behandelt:
Für die Iterationen in 7.4A und 7.4B werden die Parameter der Startfunktionen mit dem Verfahren von Meinardus unter Verwendung der angegebenen Punktemengen ermittelt; es genügt die Angabe von x_0 und x_5. Die Startfunktion von 7.4A sei E_1, die von 7.4B sei E_2: es gilt also (vgl. Zeichnungen 27, 36)

$$4 \parallel f - E_2 \parallel_I < \parallel f - E_1 \parallel_I$$

Dennoch benötigt man von E_2 ausgehend doppelt soviele Iterationsschritte wie in 7.3A, um die gleiche Genauigkeit zu erhalten.

Neben der Konstruktion der besten linearen Approximation nach Schritt 2 erfordert die Bestimmung des Faktor C in Schritt 3 den größten Teil des Rechenaufwandes. Beachtet man dies, dann wird an diesen Beispielen die Bedeutung der Startfunktion deutlich. Es zeigt sich, daß in der Regel der Faktor C im Verlauf der Iteration wächst, falls eine Minimallösung ermittelt wird. Um den Rechenaufwand möglichst klein zu halten, wird daher zur Ermittlung des Faktors C vorgeschlagen:

- Bei Beginn der Iteration verfahre man wie in Abschnitt 6.1 beschrieben. Bei den weiteren Iterationsschritten gehe man vom Faktor C des vorhergehenden Iterationsschrittes aus und prüfe, ob man hiermit eine bessere Approximation erhält.
- Gilt dies, dann bestimme man durch sukzessives Verdoppeln den gesuchten Faktor, den man nach Schritt 3 von 1 ausgehend durch (häufigeres) Halbieren erhält.
- Andernfalls hat man mit diesem Faktor nach Schritt 3 zu verfahren.

Die Iterationen werden dann abgebrochen, wenn die Parameter r_i der linearen Approximation einen kleineren Betrag als 5×10^{-10} haben. Quadratische Konvergenz ist dabei nur bei den letzten Iterationsschritten zu beobachten.

Diese Beispiele zeigen weiter, daß die Konvergenzgeschwindigkeit der ermittelten Folgen im allgemeinen unabhängig davon ist, ob nach Fall 1 oder Fall 2 verfahren wird.