§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 bzgl. V^o_{N-1} ausgehend, eine Folge besserer Approximationen $(E(a_n))_{n \in \mathbb{N}}$ liefert, falls E nicht beste Approximation bzgl. V_N ist; zur Konstruktion einer Minimallösung bzgl. V_N wird mit diesem Algorithmus sukzessive die Minimallösung bzgl. V^o_1, V^o_2, ... ermittelt.

Vereinbarung:
In diesem Abschnitt wird die Parametrisierung von V^o_N nach Definition 1.2 benutzt und es ist zweckmäßig, die folgende Schreibweise einzuführen:

Für $u=\begin{bmatrix} u_1 \\ \vdots \\ u_m \end{bmatrix} \in \mathbb{R}^m$ und $v=\begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} \in \mathbb{R}^n$ mit $m \geq n$ sei u+v gegeben durch

$$u+v := \begin{bmatrix} u_1+v_1 \\ \vdots \\ u_n+v_n \\ u_{n+1} \\ \vdots \\ u_m \end{bmatrix} \in \mathbb{R}^m$$

Der Algorithmus

Mit $n \in \mathbb{N}$ sei $E(a_n) \in V^o_N$ nicht Minimallösung für f bzgl. V_N. Es sei d_n die Dimension des Gradientenraumes W_n von $E(a_n)$ und $F_n = f - E(a_n)$; ferner seien $J \in \mathbb{N}$ mit $J \geq 2N+1$ und $z \in (o, \frac{1}{2})$ fest vorgegeben.

Schritt 1: Bestimmung einer Teilmenge X von I

Man untersuche, ob (6.1) erfüllt ist:

(6.1) Die Menge $\{x \in I \,|\, (1-z)\cdot\|F_n\|_I \leq |F_n(x)|\}$ besteht aus j Teilintervallen mit $j \leq J$ und in jedem Teilintervall ist das Maximum von $|F_n|$ eindeutig bestimmt; die Maximalpunkte von $|F_n|$ in diesen Teilintervallen seien die Punkte x_k, $1 \leq k \leq j$.

Fall 1: Es ist (6.1) erfüllt.

Die Menge der Maximalpunkte sei X_1 :

$$X_1=\{x_k \mid 1 \leq k \leq j\} .$$

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 (zur Bestimmung der Vorzeichenintervalle gelte sign(o)=+1). In jedem dieser Vorzeichenintervalle wird ein Punkt - günstig ist ein lokales Extremum von F_n - ausgezeichnet und 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 X_2' und X_3' mit $X_2' \subseteq X_2$ und $X_3' \subseteq X_3$ so, daß für

$$X:= X_1 \cup X_2' \cup X_3'$$

gilt:
$$d_n+1 \leq |X| \leq J .$$

Fall 2: Es ist (6.1) nicht erfüllt.

Hier wird definiert: $X:= I$.

Schritt 2:

Man bestimme die beste Approximation e_n an F_n auf X bzgl. W_n:

Für $E(a_n,x)=\sum_{i=1}^{N} a_{i,n}e^{t_{i,n}x}$ ist $r_n=(r_{1,n},\ldots,r_{d_n,n})\in \mathbb{R}^{d_n}$

mit $e_n(x)=\sum_{i=1}^{N} r_{i,n}e^{t_{i,n}x} + \sum_{i=N+1}^{d_n} r_{i,n}a_{i-N,n}xe^{t_{i-N,n}x}$ und

$$\|F_n-e_n\|_I \leq \|F_n-e\|_I \quad \text{für } e\in W_n \text{ gesucht.}$$

Die Verbesserung für $E(a_n)$ sei $\varepsilon_n=\|F_n\|_I-\|F_n-e_n\|_X$. Mit den Bezeichnungen von Definition 3.1 gilt für $d_n= 2N$:

$$e_n=(r_n,\operatorname{grad}(E(a_n))) .$$

Schritt 3:

Man setze $c:=1$ und bestimme durch fortlaufendes Halbieren von c ein $c_n\in (o,1]$, so daß gilt:

$$\|f-E(a_n+c_nr_n)\|_I \leq \|f-E(a_n)\|_I - \frac{1}{3}c_n\varepsilon_n$$

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

$$a_{n+1} :=a_n+c_nr_n .$$

Bemerkung:

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. 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 nur ein lineares Gleichungssystem zu lösen.
2. Obwohl X von F_n abhängt, wird dies bei der Bezeichnung nicht berücksichtigt, da diese Tatsache für die weiteren Ausführungen nicht benötigt wird.

Es ist nun die Frage zu beantworten, unter welchen Voraussetzungen der Algorithmus durchführbar ist und damit bessere Approximationen liefert; hierzu der folgende Hilfssatz:

<u>Hilfssatz 6.1</u>

Ist $E(a_n)$ nicht beste Approximation an f auf I, dann gilt

$$\varepsilon_{n1} \geq \varepsilon_{n2} > o ,$$

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 Punktmenge, 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 bzgl. W_n bezeichnet.

Es folgt $||F_n-e_{n2}||_I \geq ||F_n-e_{n2}||_X \geq ||F_n-e_{n1}||_X$ und damit

$$\varepsilon_{n1} \geq \varepsilon_{n2} .$$

Nach Voraussetzung besitzt F_n in I keine Alternante der Länge d_n+1 und die beste Approximation e_{n2} an F_n bzgl. W_n auf I approximiert daher besser als die Nullfunktion:

$$||F_n-e_{n2}||_I < ||F_n||_I$$

Dies ergibt $\varepsilon_{n2} > o$, womit der Rest der Behauptung gezeigt ist.

<u>Satz 6.1</u> (Braess, [3])

Ist $E(a_n) \in V_N^o$ nicht Minimallösung für f auf I bzgl. V_N^o, dann gibt es ein C>o, so daß für $c \in (o,C)$ erfüllt ist:

$$||f-E(a_n+cr_n)||_I \leqq ||f-E(a_n)||_I - \frac{1}{3} c\varepsilon_n$$

Beweis:
Nach Voraussetzung besitzt F_n in I keine Alternante der Länge d_n+1 und nach Hilfssatz 6.1 gilt (für beide Fälle) $\varepsilon_n>o$.

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 §4) gibt es ein C>o, mit

$$||E(a_n(c))-E(a_n)-ce_n||_I \leqq \frac{2}{3}c\varepsilon_n \quad \text{für } c \in (o,C) \,. \tag{6.2}$$

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 $c \in (o,C)$:

$$\begin{aligned} ||f-E(a_n(c))||_I &\leqq ||E(a_n)+ce_n-f||_I + \frac{2}{3} c\varepsilon_n = \\ &= ||cE(a_n)+ce_n-cf+(1-c)(E(a_n)-f)||_I + \frac{2}{3} c\varepsilon_n \leqq \\ &\leqq c||F_n-e_n||_I + (1-c)||F_n||_I + \frac{2}{3} c\varepsilon_n = \\ &= -c\varepsilon_n + ||F_n||_I + \frac{2}{3} c\varepsilon_n \leqq ||F_n||_I - \frac{1}{3} c\varepsilon_n \end{aligned}$$

Damit ist die Behauptung für die Durchführung des Algorithmus nach Fall 2 gezeigt.

Zum Fall 1: Hier gilt also $|X| \leqq J$.

Es sei $X^+:=\{x \in X | F_n(x)=||F_n||_I\}$ und $X^-:=\{x \in X | F_n(x)=-||F_n||_I\}$.

Für $x \in X^+$ erhält man $o<\varepsilon_n \leqq ||F_n||_I-(F_n(x)-e_n(x))=e_n(x)$ und entsprechend für $x \in X^-$ $o>-\varepsilon_n \geqq -||F_n||_I-(F_n(x)- e_n(x))=e_n(x)$.

Es gibt somit eine offene Umgebung U von $X^+ \cup X^-$ in I, so daß $|e_n(x)| \geqq \frac{2}{3} \varepsilon_n>o$ 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 $C_1>o$, so daß für $c \in (o,C_1)$ gilt: $F(c,x)>o$ für $x \in U^+$ und $F(c,x)<o$ für $x \in U^-$.

Ferner gibt es ein $C_2>o$ mit $|E(a_n(c),x)-E(a_n,x)-ce_n(x)| \leq \frac{1}{3} c\varepsilon_n$ für $c\in(o,C_2)$ und $x\in I$. Damit folgt für $x\in U\cap U^+$ und $c\in(o,\min\{C_1,C_2\})$

$$F(c,x)-F_n(x)=E(a_n,x)-E(a_n(c),x)\leq -ce_n(x)+\frac{1}{3}c\varepsilon_n \leq -\frac{2}{3}c\varepsilon_n+\frac{1}{3}c\varepsilon_n = -\frac{1}{3}c\varepsilon_n \text{ , also}$$

$$o<f(x)-E(a_n(c),x) \leq ||F_n||_I - \frac{1}{3}c\varepsilon_n .$$

Für $x\in U\cap U^-$ erhält man analog für $c\in(o,\min\{C_1,C_2\})$

$$-F(c,x)+F_n(x)=E(a_n(c),x)-E(a_n,x)\leq ce_n(x)+\frac{1}{3}c\varepsilon_n \leq -\frac{2}{3}c\varepsilon_n+\frac{1}{3}c\varepsilon_n = -\frac{1}{3}c\varepsilon_n \text{ , also}$$

$$-F(c,x)=|f(x)-E(a_n(c),x)|\leq ||F_n||_I-\frac{1}{3}c\varepsilon_n$$

Die Menge $V=(U^+\cap U)\cup(U^-\cap U)$ ist offen in I; $I_o=I-V$ ist kompakt und es gilt für $c_o=o$:

$$||F(c_o)||_{I_o}-||F_n||_I+\frac{1}{3}c_o\varepsilon_n=||F_n||_{I_o}-||F_n||_I<o.$$

Dies gilt wegen Stetigkeit für eine Umgebung von c_o:

Es gibt ein $C_o>o$ mit $||F(c)||_{I_o}\leq ||F_n||_I-\frac{1}{3}c\varepsilon_n$ für $c\in(o,C_o)$.

Mit $o<C:=\min\{C_o,C_1,C_2\}$ hat man somit für $c\in(o,C)$ erhalten:

$$|f(x)-E(a_n(c),x)|\leq ||F_n||_I-\frac{1}{3}c\varepsilon_n \qquad x\in I .$$

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.

Hilfssatz 6.2 (Braess, [3])

$(E(a_n))_{n\in\mathbb{N}}\subseteq V_N^o$ konvergiere auf I gleichmäßig gegen $E(a_o)\in V_N^o-V_{N-1}$. Es sei e_n bzw. e die nach Schritt 2 berechnete beste Approximation an F_n bzw. $F:=f-E(a_o)$ bzgl. W_n bzw. W auf X=I; (W ist der Gradientenraum von $E(a_o)$); 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:

$$||E(a_n)||_I \leqq K<\infty \quad \text{und } E(a_n,x)=\sum_{i=1}^{N} a_{i,n} e^{t_{i,n}x} \in V_N^o - V_{N-1} \quad \text{für } n\in\mathbb{N}.$$

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

$$\lim_{n\to\infty} t_{i,n} = t_i \in \mathbb{R} \qquad n\in J$$

(6.4)

$$\text{und} \qquad E(a_o,x)=\sum_{i=1}^{N} a_{o,i} e^{t_i x}$$

Weiter gilt für $n\in\mathbb{N}$

$$e_n(x)=\sum_{i=1}^{N}(b_{i,n}+d_{i,n}x)e^{t_{i,n}x} \quad \in V_{2N}$$

und $(e_n)_{n\in\mathbb{N}}$ ist auf $I=[u,v]$ gleichmäßig beschränkt:

$$||e_n||_I \leqq ||e_n-F_n||_I + ||F_n||_I \leqq 2||F_n||_I \leqq 2(K+||f||_I) \quad .$$

Nach Satz 2.3 und Satz 2.5 gibt es daher eine im Innern von I gleichmäßig konvergente Teilfolge $(e_n)_{n\in J_1}$, $J_1\subseteq J$, und auf jedem abgeschlossenen Teilintervall von (u,v) gilt

$$\lim_{n\to\infty} e_n = e_o \quad , n\in J_1,$$

mit $e_o\in V_{2N}$; wegen (6.4) folgt nach Satz 2.9 und Bemerkung 2.2 die gleichmäßige Konvergenz von $(e_n)_{n\in J_1}$ gegen e_o auf ganz I.

Aus $||F_n-e_n||_I \leqq ||F_n-e||_I$ für $n\in\mathbb{N}$ folgt

$$||F-e_o||_I = \lim_{n\to\infty}||F_n-e_n||_I \leqq \lim_{n\to\infty}||F_n-e||_I = ||F-e||_I \quad , n\in J_1.$$

Die Eindeutigkeit der Minimallösung bei linearer Approximation und $e_o\in W$, $e\in W$ ergeben: $\quad e_o=e \quad$.

Es ist noch zu zeigen, daß die ganze Folge $(e_n)_{n\in\mathbb{N}}$ gleichmäßig gegen e konvergiert:

Annahme: Es gibt ein $s>o$, 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

(6.5) $$|e_{n_j}(x_{n_j})-e(x_{n_j})|>s \qquad j\in\mathbb{N}$$

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 (6.5) . Damit ist die Annahme widerlegt und die Behauptung vollständig gezeigt.

<u>Satz 6.2</u> (Braess,[3])

Ergibt der Algorithmus eine Teilfolge $(E(a_n))_{n\in J}\subseteq V_N^o$, $J\subseteq\mathbb{N}$, die auf I gleichmäßig gegen ein Element $E(a_o)\in V_N^o-V_{N-1}$ konvergiert, dann konvergiert die ganze Folge $(E(a_n))_{n\in\mathbb{N}}$ gleichmäßig auf I und $E(a_o)$ ist die Minimallösung für f bzgl. 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 $F_o:=f-E(a_o)$.

Annahme: $E(a_o)$ ist nicht Minimallösung für f bzgl. V_N .

Wendet man den Algorithmus auf $E(a_o)$ nach Fall 2 (also mit X=I) an, dann erhält man nach Hilfssatz 6.1:

$$\varepsilon = ||F_o||_I-||F_o-(r_o,grad(E(a_o)))||_X>o.$$

($e_o=(r_o,grad(E(a_o)))$ ist die beste Approximation an F_o bezüglich des Gradientenraums von $E(a_o)$)

Ist bei der Konstruktion von $(E(a_n))_{n\in J}$ stets gemäß Fall 2 verfahren worden, dann folgt nach Hilfssatz 6.2

$$\lim_{n\to\infty}\varepsilon_n=\varepsilon \quad \text{für } n\in J$$

und es gibt damit ein $n_o\in\mathbb{N}$, so daß für $n\in J$ und $n\geqq n_o$ gilt:

(6.6) $$\varepsilon_n \geqq \frac{1}{2}\varepsilon$$

Nach Hilfssatz 6.1 gilt (6.6) auch, falls ε_n gemäß Fall 1 berechnet wird.

Fallunterscheidung:

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 $\mathrm{grd}(E(a_n))=N$)

(6.7) $\lim_{n\to\infty} a_n = a_o$ und

(6.8) $\lim_{n\to\infty} r_n = r_o$ (nach Hilfssatz 6.2)

folgt.

Auf Grund der stetigen Differenzierbarkeit der Exponentialsummen gibt es ein $M<\infty$ mit

$$||E(a_n(c))-E(a_n)-ce_n||_I \leqq c^2 M \quad \text{für } n\in J,\ |c|\leqq 1;$$

$a_n(c)$ ist definiert wie in Satz 6.1 .

Folglich gibt es ein C mit $o<C\leqq 1$, so daß mit (6.6) für $c\in(o,C)$ erfüllt ist:

$$||E(a_n(c))-E(a_n)-ce_n||_I \leqq \frac{1}{3}c\varepsilon \leqq \frac{2}{3}c\varepsilon_n \ ,\ n\in J,\ n\geqq n_o.$$

Wie in Satz 6.1 gezeigt, folgt daraus $c_n \geqq \frac{C}{2}$, $n\in J$, $n\geqq n_o$;

man erhält somit für $n\in J$, $n\geqq n_o$:

$$||E(a_{n+1})-E(a_n)||_I \geqq ||F_n||_I - ||F_{n+1}||_I \geqq \frac{1}{3}c_n\varepsilon_n \geqq \frac{1}{3}\frac{C}{2}\frac{\varepsilon}{2} > o$$

Dies steht im Widerspruch zur gleichmäßigen Konvergenz der Folge.

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:

$$||F_n||_I - ||F_{n+1}||_I \geqq \frac{1}{3}z\varepsilon_n \geqq \frac{1}{3}z\frac{1}{2}\varepsilon > o \quad n\in J_1,\ n\geqq n_o.$$

Damit ist gezeigt, daß $E(a_o)$ Minimallösung ist.

Damit konvergiert die ganze Folge $(E(a_n))_{n\in\mathbb{N}}$ gleichmäßig gegen $E(a_o)$:

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-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 $E\in V_N-V_{N-1}$.
E ist aber, wie gezeigt, Minimallösung und aus der Eindeutigkeit der besten Approximation folgt $E=E(a_o)$.

<u>Zur Anwendung des Algorithmus:</u>

a. Approximation bzgl. V_N^+
Um die Voraussetzungen von Satz 6.2 immer erfüllen zu können, geht man bei der Konstruktion der besten Approximation bzgl. V_N^+ induktiv vor:
Die beste Approximation bzgl. $V_o^+=V_o$ ist die Nullfunktion; die Minimallösung $E(a)$ für f bzgl. V_{N-1}^+ sei bekannt und nicht die beste Approximation bzgl. 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 $V_N^+-V_{N-1}$.
Die Iteration beginnt man mit

$$E(a_1,x)=E(a,x)+a_{N,1}e^{t_{N,1}x} ,$$

wobei $a_{N,1}=o$ 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 $E(a_2)\in V_N^+-V_{N-1}$ nach Satz 6.1; man erhält so eine Folge $(E(a_n))_{n\in\mathbb{N}}$, die in $V=\{E \mid E\in V_N^+-V_{N-1},\ ||f-E||_I\leq ||f-E(a_2)||_I< ||f-E(a)||_I\}$

enthalten ist.

Falls $q:= (\max_{x\in I} f(x) + \min_{x\in I} f(x))/2 >o$ gilt, ist die beste Approximation an f auf I bzgl. V_1^+ in V_1-V_o enthalten, und man kann die Iteration in V_1^+ sofort mit $E(a_1,x)=q_1\in V_1-V_o$ 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 bzgl. 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 bzgl. V_N^+.

b.
Liegt die beste Approximation bzgl. V_N^o (sofern sie existiert) nicht in V_N^+, dann ist im allgemeinen die Voraussetzung von Satz 6.2 nicht erfüllt und man hat eine
<u>Modifikation des Verfahrens</u> durchzuführen, um den Algorithmus auch hier noch sinnvoll anzuwenden:
Es sei also $E(a)\in V_{N-1}^o$ die Minimallösung bzgl. V_{N-1}^o an f und die beste Approximation bzgl. V_N^o, 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 $a_{i,1}=o$ für $K<i\leq N$

und K=grad(E(a)) gilt und ferner $t_{i,1}\in\mathbb{R}$, $K<i\leq N$, nicht im Spektrum von E(a) enthalten ist und $t_{i,1}<t_{i+1,1}$ ist.
Man erhält eine bessere Approximation $E(a_2)\in V_N^o$ nach Satz 6.1 mit dem Vorzeichenvektor $S=\mathrm{sign}(E(a_2))$. 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 $\mathrm{sign}(E(a_n))=S$ 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 $E(a_o)\in V_N^o-V_{N-1}$ gleichmäßig konvergente Teilfolge, dann ist $E(a_o)$ Minimallösung, da diese Zusatzforderung am Beweis von Satz 6.2 wegen (6.7) und (6.8) nichts ändert: C>o 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 bzgl. V_N^o , Elemente von $V_N - V_N^o$ sind (Korollar 4.1) geht man nach Braess vor, wie folgt:

Erhält man bei der Iteration in $V_N^o(S)$ 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.

Bemerkung:

Die Ausführungen zur Approximation bzgl. V_N^+ mit diesem Verfahren gelten analog auch für die Approximation bzgl. V_N^- .

6.2 Das Newtonsche Iterationsverfahren

Das Newtonsche Iterationsverfahren, das allgemein für nichtlineare Approximation in [11] beschrieben ist, läßt sich nach den Ergebnissen von §3 insbesondere bei der Approximation durch Exponentialsummen anwenden.
Es sei $E(a) \in V_N^o - V_{N-1}$ die beste Approximation an $f \in C(I)$ auf $I=[u,v]$ bzgl. 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 $\{E \in V_N^o \mid \|f-E\|_I < \inf_{w \in V_{N-1}} \|f-w\|_I\}$,

da diese den Grad N besitzen. Es existiert also eine Umgebung von $E(a)$ in V_N^o , so daß die Elemente dieser Umgebung stets einen Gradientenraum der Dimension $2N$ besitzen.

Es sei $E(a_o) \in V_N^o$ eine Näherung für $E(a)$, so daß es in I $2N+1$ Punkte $x_{1,k}$, $o \leq k \leq 2N$, gibt mit

$$(6.1o) \qquad o \neq \operatorname{sign}(F_o(x_{1,k})) \neq \operatorname{sign}(F_o(x_{1,k+1})) \neq o \qquad o \leq k \leq 2N-1;$$

dabei sei $F_o:=f-E(a_o)$. In den Punkten $x_{1,k}$ nehme $|F_o|$ möglichst große Werte an.
Nach Hilfssatz 3.1 gilt $\min_{o \leq k \leq 2N} |F_o(x_{1,k})| \leq \inf_{E \in V_N} \|f-E\|_I$.
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:

$$(6.11) \qquad F_1(x_{1,k}) = (-1)^k r_1 \qquad o \leq k \leq 2N .$$

Existiert dieser Parametervektor, dann ist $E(a_1)$ die beste Approximation an f auf $\{x_{1,k} \mid o \leq k \leq 2N\}$ 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

$$g_k(a_1,r_1):=E(a_1,x_{1,k})-f(x_{1,k})+(-1)^k r_1=o \qquad o \leq k \leq 2N .$$

Demnach ist (a_1,r_1) eine Nullstelle von

$$G(b,q):=\begin{bmatrix} g_o(b,q) \\ \vdots \\ g_{2N}(b,q) \end{bmatrix} \quad , \quad (b,q)\in \mathbb{R}^{2N+1}, \ b\in\mathbb{R}^{2N} \ .$$

Verwendet man das Newton-Verfahren (im $\mathbb{R}^{2N+1}$) mit dem Startwert $(b_1,q_1)=(a_o,q_1)$, q_1 geht in die Berechnung nicht ein,

so erhält man zunächst formal für $n\in\mathbb{N}$:

$$J_G(b_n,q_n)\cdot(b_{n+1},q_{n+1})^t = J_G(b_n,q_n)\cdot(b_n,q_n)^t - G(b_n,q_n) \ ;$$

J_G ist die Jacobi-Matrix von G.

Es folgt also explizit für $o\leq k\leq 2N$:

$$\sum_{j=1}^{2N} b_{n+1,j}D_jE(b_n,x_{1,k}) + (-1)^k q_{n+1} =$$

$$= \sum_{j=1}^{2N} b_{n,j}D_jE(b_n,x_{1,k}) + (-1)^k q_n - E(b_n,x_{1,k}) + f(x_{1,k}) - (-1)^k q_n$$

$D_jE(b)$ ist die partielle Ableitung von $E(b)$ nach b_j mit $b=(b_1,\dots,b_{2N})$; ferner gilt $b_n=(b_{n,1},\dots,b_{n,2N})$.

Es folgt damit für $o\leq k\leq 2N$:

$$\sum_{j=1}^{2N}(b_{n+1,j}-b_{n,j})D_jE(b_n,x_{1,k})+(-1)^k q_{n+1}=f(x_{1,k})-E(b_n,x_{1,k})$$

Hat der Gradientenraum von $E(a_o)$ die Dimension 2N, dann ist für n=1 dieses Gleichungssystem eindeutig lösbar und man erhält den Vektor $b_2-b_1=b_2-a_o\in\mathbb{R}^{2N}$. Ist $E(a_o)$ 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 (6.11) für numerische Zwecke hinreichend gut.

Im allgemeinen ist dies jedoch nicht der Fall (man vergleiche hierzu §8); man hat dann das Newton-Verfahren mindestens soweit durchzuführen, daß für ein $n\in\mathbb{N}$

$$\operatorname{sign}(f(x_{1,k})-E(b_n,x_{1,k}))\neq\operatorname{sign}(f(x_{1,k+1})-E(b_n,x_{1,k+1})) \ ,$$

$o\leq k<2N$, erfüllt ist.

Da auf diskreten Punktmengen nicht allgemein eine beste

Approximation bzgl. V_N^o 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 $E(a_o)$, dann verfährt man mit $E(a_1)$, falls $E(a) \neq E(a_1)$ ist, wie mit $E(a_o)$ und erhält so $E(a_2)$, wenn die entsprechenden Voraussetzungen erfüllt sind, usw.

Da Konvergenzaussagen für dieses Verfahren allgemein 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 6.2 ist hier nur eine reelle Nullstelle (etwa nach Newton-Raphson) zu ermitteln.
Die Nullfunktion ist genau dann beste Approximation an f auf I bzgl. V_1, wenn f in I eine Alternante der Länge zwei besitzt; man kann also unmittelbar die Länge der besten Approximation an f bzgl. V_1 angeben und es sei daher o.B.d.A. die Minimallösung E_o für f bzgl. V_1 Element von V_1-V_o.
In I besitzt also $f-E_o$ eine Alternante der Länge drei; man wird daher versuchen, durch Konstruktion der Minimallösung E für f auf einer Menge $X \subsetneq I$ mit $|X|=3$ die beste Approximation E_o auf I zu ermitteln. Da die Existenz von E für beliebiges X nicht gewährleistet ist, werden die Schwierigkeiten sichtbar, die bei diesem Vorgehen ,ähnlich wie in 6.2 , auftreten können:

<u>Beispiel 6.1</u>

$X=\{o, \frac{1}{2}, 1\}$, $f(o)=o=f(\frac{1}{2})$, $f(1)=1$.
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 $(E_n)_{n\in\mathbb{N}} \subsetneq V_1$ eine Minimalfolge für f auf X bzgl. V_1, denn es gilt

$$\lim_{n\to\infty} ||f-E_n||_X=o$$

Es gibt jedoch keine beste Approximation an f auf X bzgl. V_1.

Es sei $X=\{x_1,x_2,x_3\} \subsetneq I$ mit $x_1<x_2<x_3$ so gegeben, daß für f auf X die Minimalösung $E(x)=ae^{sx} \in V_1$ existiert und weiter $a \neq o$ erfüllt ist. Es gilt also

$$(6.12) \qquad ae^{sx_i}+(-1)^i r = f(x_i) \qquad 1 \leqq i \leqq 3 \; ;$$

die Minimalabweichung von f auf X sei also $|r|$, $r \in \mathbb{R}$.

Nach Voraussetzung gilt $f(x_3)+f(x_2)\neq 0$, so daß man durch Elimination von r und a das äquivalente Gleichungssystem (6.13) erhält:

$$-r=f(x_3)-ae^{sx_3}$$

$$a=\frac{f(x_3)+f(x_2)}{e^{sx_3}+e^{sx_2}} \tag{6.13}$$

$$\frac{e^{sx_3}-e^{sx_1}}{e^{sx_3}+e^{sx_2}}=\frac{f(x_3)-f(x_1)}{f(x_3)+f(x_2)}$$

Mit $F(t):=\dfrac{e^{tx_3}-e^{tx_1}}{e^{tx_3}+e^{tx_2}}$ für $t\in\mathbb{R}$ und $w:=\dfrac{f(x_3)-f(x_1)}{f(x_3)+f(x_2)}$

ist die Lösung des Gleichungssystems (6.12) zurückgeführt auf die Bestimmung einer reellen Nullstelle von $F(t)-w$.

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 $t\in\mathbb{R}$

$$\frac{dF(t)}{dt}=\frac{(x_3-x_1)e^{t(x_1+x_3)}+(x_3-x_2)e^{t(x_3+x_2)}+(x_2-x_1)e^{t(x_1+x_2)}}{(e^{tx_3}+e^{tx_2})^2}>0.$$

$F(t)$ 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 $\lim\limits_{t\to-\infty}F(t) = -\infty$. $F(t)$ nimmt also für $t\in\mathbb{R}$ jeden Wert aus $(-\infty, +1)$ genau einmal an. Damit ist der Satz gezeigt.

Bemerkung 6.1

Gilt $\operatorname{sign}(f(x)) \geq o$ ($\leq o$) für $x \in I$, ändert also f in I sein Vorzeichen nicht, und ist f in I ferner entweder streng monoton wachsend oder streng monoton fallend, dann gilt

$$\frac{f(x_3)-f(x_1)}{f(x_3)+f(x_2)} < 1$$

für $x_i \in I$, $1 \leq i \leq 3$, mit $x_1 < x_2 < x_3$.

Damit läßt sich der folgende Algorithmus formulieren:
Die Minimallösung E_o für f auf I bzgl. V_1 sei nicht in V_o enthalten, E_o ist also nicht die Nullfunktion.

1. Zur Ermittlung einer Startfunktion $E_1(x) = a_1 e^{t_1 x}$, bestimme man in I drei Punkte $x_{1,i}$, $1 \leq i \leq 3$, 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 zwei besitzt; nach Voraussetzung ist dies möglich.

   Gilt ferner $w_1 := \dfrac{f(x_{1,3})-f(x_{1,1})}{f(x_{1,3})+f(x_{1,2})} < 1$, dann ist t_1 die reelle Nullstelle von

   $$F_1(t)-w_1 := \frac{e^{tx_{1,3}}-e^{tx_{1,1}}}{e^{tx_{1,3}}+e^{tx_{1,2}}}-w_1 \quad .$$

   Man erhält a_1 aus $\qquad a_1 := \dfrac{f(x_{1,3})+f(x_{1,2})}{e^{t_1 x_{1,3}}+e^{t_1 x_{1,2}}}$.

   Die Punkte $x_{1,i}$, $1 \leq i \leq 3$, seien so gewählt, daß

   $$|r_1| = |f(x_{1,3})-a_1 e^{t_1 x_{1,3}}| > o$$

   gilt; dies ist notwendig für die Durchführung der folgenden Iteration.

2. Es sei $n \geq 2$; 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 $x_{n,i} \in I$, $1 \leq i \leq 3$, so daß

   $$o \neq \operatorname{sign}(f(x_{n,i})-E_{n-1}(x_{n,i})) = -\operatorname{sign}(f(x_{n,i+1})-E_{n-1}(x_{n,i+1})),\ i=1,2$$

   erfüllt ist und weiterhin $f-E_{n-1}$ in $x_{n,i}$ für $1 \leq i \leq 3$ dem Betrage nach einen möglichst großen Wert annimmt; nach Konstruktion von E_{n-1} ist dies möglich.

Gilt weiter

$$w_n := \frac{f(x_{n,3})-f(x_{n,1})}{f(x_{n,3})+f(x_{n,2})} < 1 \text{ , dann}$$

ist t_n die reelle Nullstelle von

$$F_n(t)-w_n := \frac{e^{tx_{n,3}}-e^{tx_{n,1}}}{e^{tx_{n,3}}+e^{tx_{n,2}}} - w_n \;;$$

$$\text{weiter ist } a_n := \frac{f(x_{n,3})+f(x_{n,2})}{e^{t_n x_{n,3}}+e^{t_n x_{n,2}}} \quad .$$

Bemerkung:

1. Für jedes Element f der in Bemerkung 6.1 beschriebenen Funktionenklasse ist das Verfahren stets durchführbar: Zunächst besitzt f nach Voraussetzung in keiner Punktmenge $X \subseteq I$ eine Alternante der Länge zwei; nimmt man an, daß bei der Ermittlung einer Startfunktion für jede Menge $X \subseteq I$, $|X|=3$, stets $|r_1|=o$ 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 $E_1(x_i)=f(x_i)$, $1 \leq i \leq 3$ und für jede Menge $\{x_1,x_2,x\}$, $x \in I$, $x_1 \neq x \neq x_2$, ein $E_x \in V_1$ mit $E_x(x_i)=f(x_i)$, $i=1,2$ und $E_x(x)=f(x)$, woraus $E_x=E_1$ folgt. Man erhält also bereits mit der Startfunktion die gewünschte Lösung.
Weitergehende Untersuchungen , etwa Aussagen zur Konvergenz waren mir hier nicht möglich. Es stellt sich die Frage, inwieweit man dieses Verfahren für die Approximation in V_N mit $N \geq 2$ 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; eventuell 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.
Zur numerischen Durchführung des Algorithmus, insbesondere zur Bestimmung der Nullstellen t_n beachte man §9; die dort angegebenen Beispiele lassen dieses Verfahren als brauchbar erscheinen.

2. Die Idee zur Lösung des Gleichungssystems (6.12) 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. lo (1962).