II. NUMERISCHE VERFAHREN

§5 Konstruktion von Näherungen

Es sei I das Intervall $[a,b] \subseteq \mathbb{R}$ 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 $X=\{x_i \mid o \leq i \leq 2N-1\} \subseteq I$; gesucht ist eine Funktion $E(x)=\sum_{i=1}^{N} a_i e^{t_i x} \in V_N^o$ mit

$$E(x_i)=f(x_i) \qquad o \leq i \leq 2N-1 \tag{5.1}$$

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

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

Setzt man für $1 \leq j \leq N$ $\quad A_j:=a_j e^{t_j a}$, $E_j:=e^{t_j h}$ und

$f_j:=f(x_j)$ für $o \leq j \leq 2N-1$,

so erhält man aus (5.1)

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

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 $x \in U$,

$U=\{x \in \mathbb{R} \mid \max_{1 \leq j \leq N} |E_j x|<1\}$, die Darstellung

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

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

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

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

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

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

Es gilt somit für $x \in U$

$$\sum_{i=o}^{N-1} p_i x^i = \left(\sum_{i=o}^{N} q_i x^i \right) \left(\sum_{i=o}^{\infty} x^i f_i \right) = \sum_{i=o}^{\infty} x^i \sum_{\substack{k+j=i \\ k \geq o \\ o \leq j \leq N}} q_j f_k$$

und Koeffizientenvergleich ergibt:

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

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

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

$$p_o = f_o$$

$$(5.5) \qquad \sum_{j=1}^{i} q_j f_{i-j} - p_i = -f_i \qquad 1 \leq i \leq N-1$$

$$(5.6) \qquad \sum_{j=1}^{N} q_j f_{i-j} = -f_i \qquad N \leq i \leq 2N-1$$

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:

$$L_{N-1} = \begin{bmatrix} f_o & & & \mathcal{O} \\ f_1 & f_o & & \\ f_2 & f_1 & \ddots & \\ \vdots & \vdots & & \ddots \\ f_{N-2} & f_{N-3} & \cdots\cdots & f_o \end{bmatrix}$$

$$F_N = \begin{bmatrix} f_{N-1} & f_{N-2} & \cdots & f_o \\ f_N & f_{N-1} & \cdots & f_1 \\ \vdots & \vdots & & \vdots \\ f_{2N-2} & f_{2N-3} & \cdots & f_{N-1} \end{bmatrix}$$

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

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

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

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

$$E_i = z_i^{-1} \qquad 1 \leq i \leq N \,. \tag{5.7}$$

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

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

2. Die Koeffizienten A_i und a_i

Die Koeffizienten p_i, $1 \leq i \leq N-1$, erhält man aus

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

Mit $p_o = f_o$ ergeben sich nunmehr aus (5.3) und (5.7) durch Koeffizientenvergleich die Größen A_i, $1 \leq i \leq N$; es genügt dazu,

daß Q genau N reelle Nullstellen besitzt und so E_i, $1 \leq i \leq N$, 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 $t_i \in \mathbb{R}$, $1 \leq i \leq N$, ermittelt, so gilt

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

Bemerkung:

1. Sind die Frequenzen t_i, $1 \leq i \leq N$, bekannt, dann können die Koeffizienten a_i, $1 \leq i \leq N$, 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
$$\{x_i \mid o \leq i \leq 2N,\ x_i = x_o + ih\}$$
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, $o \leq i \leq 2N-1$, falls $f_i = E(x_i)$ mit $E \in V_N^o$ 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.

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}, $1 \leq j \leq N$; E_j ist für $1 \leq j \leq N$ wegen $q_o = 1$ definiert.

Behauptung:

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

Beweis:

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

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

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

Damit ist die Behauptung gezeigt.

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

Hilfssatz 5.1

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

Beweis:

Nach Voraussetzung gilt

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

wegen $z^n \neq o$ folgt daraus die Behauptung.

Satz 5.2

Es gelte $f_i = E(x_i) = \sum_{j=1}^{n} a_j e^{t_j x_i}$ für $o \leq i \leq 2N-1$ mit grd(E) $= n \leq N$;

A_i und E_i sind für $1 \leq i \leq n$ definiert wie oben. Es sei $z_i \in \mathbb{R}$ mit $z_i \neq o$ für $n+1 \leq i \leq N$ beliebig gewählt.
Die Koeffizienten q_i, $o \leq i \leq N$, seien gegeben durch

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

Behauptung:

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

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

Beweis:

Das Polynom $\sum_{i=o}^{N} q_i x^i$ besitzt nach Voraussetzung die n Nullstellen E_i^{-1}, $1\leq i\leq n$, (E_i^{-1} ist wegen $E_i=e^{t_i h}$ stets definiert).

Nach Hilfssatz 5.1 gilt für $1\leq k\leq n$:

$$o=\sum_{i=o}^{N} q_{N-i}E_k^i=\sum_{i=1}^{N} q_i E_k^{N-i} + q_o E_k^N \quad ; \text{ wegen } q_o=1 \text{ folgt für } 1\leq k\leq n:$$

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

Für $1\leq j\leq N$ gilt hiermit $\sum_{i=1}^{N} f_{N+j-1-i}q_i =$

$$= \sum_{i=1}^{N}\Big(\sum_{k=1}^{n} A_k E_k^{N+j-1-i}\Big)q_i = \sum_{k=1}^{n}\Big(\sum_{i=1}^{N} q_i E_k^{N-i}\Big)A_k E_k^{j-1} =$$

$$= \sum_{k=1}^{n} A_k E_k^{j-1}(-E_k^N) = -f_{N+j-1} \quad ;$$

damit ist (5.8) gezeigt.

Es sei nun $n=N$; es gilt also $\sum_{i=o}^{N} q_i x^i = \prod_{i=1}^{N}(1-E_i x)$ und

es sei $\begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_N \end{bmatrix}=v$ eine zweite Lösung von (5.8); für $1\leq j\leq N$ folgt

also

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

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

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

Somit erhält man:

$$\det \begin{bmatrix} A_1E_1^0 & A_2E_2^0 & \cdots & A_NE_N^0 \\ A_1E_1^1 & A_2E_2^1 & \cdots & A_NE_N^1 \\ \vdots & \vdots & & \vdots \\ A_1E_1^{N-1} & A_2E_2^{N-1} & \cdots & A_NE_N^{N-1} \end{bmatrix} =$$

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

da eine Vandermondesche Determinante vorliegt und die Ungleichungen (5.1o) bestehen. Das Gleichungssystem (5.9) ist also eindeutig lösbar und es gilt daher für $1 \leq k \leq N$

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

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

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

$$\sum_{i=o}^{N} v_i x^i = o \qquad 1 \leq k \leq N .$$

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

$$q_i = v_i , \qquad 1 \leq i \leq N .$$

Bemerkung:

Nimmt man an, daß $f_i = E(x_i)$ für $o \leq i \leq 2N-1$ mit $E \in V_N^o - V_{N-1}$ 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 $E \in V_N^o$ richtig ist.

Hilfssatz 5.2

Für $n \in \mathbb{N}$ und $t \in \mathbb{R}$ gilt mit $o \leq m < n$:

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

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 $r > n$:

$$\frac{1}{2\pi i} \int_{K_r} (t-z)^m \prod_{j=o}^{n} (z-j)^{-1} dz = \sum_{k=o}^{n} \operatorname{Res}_{z=k} (t-z)^m \prod_{j=o}^{n} (z-j)^{-1} =$$

$$= \sum_{k=o}^{n} \lim_{\substack{z \to k \\ z \neq k}} (z-k)(t-z)^m \prod_{j=o}^{n} (z-j)^{-1} = \sum_{k=o}^{n} (t-k)^m \prod_{\substack{j=o \\ j \neq k}}^{n} (k-j)^{-1} =$$

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

Es sei $A(z) := \sum_{j=o}^{m} a_j z^j := (t-z)^m$ und $B(z) := \sum_{j=o}^{n+1} b_j z^j := \prod_{j=o}^{n} (z-j)$.

Es gilt $a_m = (-1)^m$, $b_{n+1} = 1$ und man erhält für $z \in \mathbb{C}$, $z \neq o$:

$$|A(z)| \leq |z|^m \left(1 + \sum_{j=o}^{m-1} |a_j| |z|^{j-m}\right)$$

$$|B(z)| \geq |z|^{n+1} \left(1 - \sum_{i=o}^{n} |b_i| |z|^{i-n-1}\right)$$

Wählt man daher $r_o \in \mathbb{R}$, $r_o > n$, hinreichend groß, so gilt $|A(z)| \leq 2|z|^m$ und $|B(z)| \geq \frac{1}{2}|z|^{n+1}$ für $z \in \mathbb{C}$ mit $|z| \geq r_o$.

Daraus folgt mit $r \geq r_o$: $\left| \int_{K_r} \frac{A(z)}{B(z)} dz \right| \leq 8\pi r^{m-n}$.

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

$$\frac{1}{n!} \left| \sum_{k=o}^{n} (-1)^k (t-k)^m \binom{n}{k} \right| \leq 4r^{-1} ,$$

woraus die Behauptung folgt.

Bemerkung:
Diesen Hilfssatz erhält man auch aus der Beziehung

$$\Delta^n(o,1,\ldots,n)g=\frac{1}{n!}\sum_{k=o}^{n}(-1)^{n-k}\binom{n}{k}g(k) \quad \text{(Willers, [23], S.73)}$$

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

Satz 5.3

Es sei $N\geq 2$ und $f_i=E(x_i)$ mit $E(x)=(\sum_{i=o}^{N-1}a_ix^i)e^{tx}\in V_N-V_{N-1}$

für $o\leq i\leq 2N-1$; ferner sei mit $E_1=e^{th}$ für $o\leq i\leq N$ q_i gegeben

durch $\sum_{i=o}^{N}q_ix^i=(1-E_1x)^N$.

Behauptung:

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

Beweis:
Mit dem Binomischen Lehrsatz erhält man

$$\sum_{i=o}^{N}q_ix^i=\sum_{i=o}^{N}(-1)^i\binom{N}{i}E_1^ix^i$$

und daraus durch Koeffizientenvergleich

$$q_i=(-1)^i\binom{N}{i}E_1^i \quad , \; o\leq i\leq N;$$

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

$$\sum_{i=1}^{N}f_{N+j-1-i}q_i=A\sum_{i=1}^{N}(\sum_{k=o}^{N-1}a_kx_{N+j-1-i}^k)E_1^{N+j-1-i}q_i=$$

$$=AE_1^{N+j-1}\sum_{k=o}^{N-1}a_k(\sum_{i=1}^{N}E_1^{-i}q_ix_{N+j-1-i}^k)=$$

$$=AE_1^{N+j-1}\sum_{k=o}^{N-1}a_k(\sum_{i=1}^{N}(-1)^i\binom{N}{i}x_{N+j-1-i}^k)$$

Mit $o \leq k < N$ erhält man durch Anwendung von Hilfssatz 5.2 :

$$\sum_{i=1}^{N} (-1)^i \binom{N}{i} x_{N+j-1-i}^k = \sum_{i=1}^{N} (-1)^i \binom{N}{i} \sum_{m=o}^{k} \binom{k}{m} a^{k-m} (N+j-1-i)^m h^m =$$

$$= \sum_{m=o}^{k} h^m \binom{k}{m} a^{k-m} \sum_{i=1}^{N} (-1)^i \binom{N}{i} ((N+j-1)-i)^m =$$

$$= - \sum_{m=o}^{k} \binom{k}{m} a^{k-m} (N+j-1)^m h^m = -x_{N+j-1}^k$$

Damit gilt $\sum_{i=1}^{N} f_{N+j-1-i} q_i = A E_1^{N+j-1} \sum_{k=o}^{N-1} a_k (-x_{N+j-1}^k) = -f_{N+j-1}$,

was zu zeigen war.

Bemerkung:

Versucht man (5.2) mit $A_i > o$ für $1 \leq i \leq N$ zu lösen ($o < E_i < E_{i+1}$ 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 $\sum_{j,k=o}^{N-1} f_{j+k} y_j y_k$, $\sum_{j,k=o}^{N-1} f_{j+k+1} y_j y_k$ positiv definit sind.

Läßt sich f als Dirichletsche Reihe mit positiven Koeffizienten darstellen, gilt also

$$f(x) = \sum_{n=1}^{\infty} c_n e^{s_n x} \quad \text{mit } c_n, s_n \in \mathbb{R},\ c_i > o \text{ für } 1 \leq i \leq N \text{ und } c_n \geq o,\ n \in \mathbb{N},$$

dann ist (5.2) lösbar, wie in [11] gezeigt ist:

Es gilt $f_i > o$ für $o \leq i \leq 2N-1$; mit $f_{j+k} = \sum_{n=1}^{\infty} c_n e^{s_n a} e^{(j+k)hs_n}$ gilt

$$\sum_{j,k=o}^{N-1} f_{j+k} y_j y_k = \sum_{n=1}^{\infty} c_n e^{s_n a} \sum_{j,k=o}^{N-1} e^{jhs_n} y_j e^{khs_n} y_k =$$

$$= \sum_{n=1}^{\infty} c_n e^{s_n a} \Big(\sum_{j=o}^{N-1} e^{js_n h} y_j \Big)^2 \text{ , so daß } \sum_{j,k=o}^{N-1} f_{j+k} y_j y_k > o \text{ gilt,}$$

falls nicht $y_i = o$, $o \leq j \leq N-1$, erfüllt ist (wegen $c_i > o$ für $1 \leq i \leq N$). Ebenso folgt die positive Definitheit der zweiten quadratischen Form. Man beachte hierzu die Beispiele von §8.

Es wird nun gezeigt, daß die Bestimmung von det F_N keine Aussage über die Durchführbarkeit des Verfahrens zuläßt:

Beispiel 5.1

Es sei N=2, X={o,1,2,3}.

1. $f_o=1$, $f_1=e$, $f_2=e^2$, $f_3=e^3$

Es gilt $\det F_2 = \det\begin{bmatrix} e & 1 \\ e^2 & e \end{bmatrix} = o$, jedoch lassen sich Lösungen von $F_2 \begin{bmatrix} q_1 \\ q_2 \end{bmatrix} = -\begin{bmatrix} e^2 \\ e^3 \end{bmatrix}$ angeben:

Die Lösungsmenge des Systems ist gegeben durch

$$\{\begin{bmatrix} -r \\ -e^2+er \end{bmatrix} \mid r\in\mathbb{R}\}.$$

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 $p_o=1$.
Die Gleichung (5.3) lautet hier:

$$1+(e-r)x = A_1+A_2+x(e-r)A_1-A_2ex \quad .$$

Dies ergibt: $A_2=1-A_1$ und $e-r=(e-r)A_1-(1-A_1)e$

$$2e-r=A_1(2e-r) \quad \Longrightarrow \quad A_1=1, \quad A_2=o$$

Von Bedeutung ist also nur die Bestimmtung von t_1:

$$t_1= \ln E_1 =1 \quad .$$

Man erhält $E(x)=e^x$; dies ist die Lösung des Interpolationsproblems.

2. $f_o=1$, $f_1=-1$, $f_2=1$, $f_3=1$.

Hierfür gibt es keine Lösung $E\in V_2$ mit $E(i)=f_i$, $o\leq i\leq 3$, da $E\not\equiv o$ 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

$$\det F_2 = \det\begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} = o$$

5.2 Konstruktion einer Näherung nach Rice

Es sei in $I=[a,b]$ die äquidistante Punktmenge

$$X=\{x_i \mid x_i \in I,\ x_i=a+ih,\ 1\leq i\leq 2N+1,\ h>0\}$$

gegeben.

Unter der Annahme, daß es eine beste Approximation $E\in V_N^o-V_{N-1}$ an $f\in C(I)$ bzgl. V_N^o auf X 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 auf I bzgl. V_N^o zu erhalten, da diese, sofern sie existiert und die Länge N besitzt, in I eine Alternante der Länge 2N+1 hat.

Es sei also $E(x)=\sum_{i=1}^{N} a_i e^{t_i x}\in V_N^o-V_{N-1}$ die beste Approximation an f auf X; es gilt daher

$$E(x_i)=f(x_i)+(-1)^i r \qquad 1\leq i\leq 2N+1 \tag{5.11}$$

$|r|$ ist also die Minimalabweichung von f auf X; eliminiert man r durch Addition aufeinanderfolgender Gleichungen, so erhält man das Gleichungssystem

$$\sum_{i=1}^{N} a_i e^{t_i a}\left(e^{t_i h}+1\right)e^{t_i jh} = f(x_j)+f(x_{j+1}) \qquad 1\leq j\leq 2N\ . \tag{5.12}$$

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 anschließenden Lösung eines linearen Gleichungssystems. Hierzu werden die folgenden Bezeichnungen eingeführt:

$$f_j:=f(x_j)+f(x_{j+1}) \qquad 1\leq j\leq 2N$$

$$A_i:=a_i e^{t_i a}\left(e^{t_i h}+1\right),\ E_i:=e^{t_i h} \qquad 1\leq i\leq N$$

Damit erhält man aus (5.12)

$$\sum_{i=1}^{N} A_i E_i^j = f_j \qquad 1\leq j\leq 2N$$

und da nach Voraussetzung $A_i\neq o$ für $1\leq i\leq N$ erfüllt ist, gilt

für $1 \leq k \leq N$ (wegen linearer Abhängigkeit der ersten Spalte):

$$\det \begin{bmatrix} f_k & E_1^k & \cdots\cdots & E_N^k \\ f_{k+1} & E_1^{k+1} & \cdots\cdots & E_N^{k+1} \\ \vdots & \vdots & & \vdots \\ f_{k+N} & E_1^{k+N} & \cdots\cdots & E_N^{k+N} \end{bmatrix} = o$$

Wegen $E_i \neq o$, $1 \leq i \leq N$, folgt daher

$$\det \begin{bmatrix} f_k & 1 & \cdots\cdots & 1 \\ f_{k+1} & E_1^1 & \cdots\cdots & E_N^1 \\ \vdots & \vdots & & \vdots \\ f_{k+N} & E_1^N & \cdots\cdots & E_N^N \end{bmatrix} = o \qquad 1 \leq k \leq N \,. \tag{5.13}$$

Entwickelt man die Determinanten in (5.13) nach der ersten Spalte und ist D_i für $o \leq i \leq N$ die Adjunkte von f_{k+i}, so gilt (da D_i unabhängig ist von k):

$$\sum_{i=o}^{N} f_{k+i} D_i = o \qquad 1 \leq k \leq N \,. \tag{5.14}$$

Ebenso erhält man

$$\sum_{i=o}^{N} E_j^i D_i = o \qquad 1 \leq j \leq N \,, \tag{5.15}$$

da in den entsprechenden Matrizen zwei Spalten übereinstimmen. Aus $E_i < E_{i+1}$ folgt

$$D_o \neq o, \; D_N \neq o \tag{5.16}$$

Je eine Gleichung aus dem System (5.15) ergibt zusammen mit (5.14) eine lineares, homogenes Gleichungssystem mit den Unbekannten D_i, $o \leq i \leq N$:
Es sei $j \in \{1, \ldots N\}$:

$$\begin{aligned} &\sum_{i=o}^{N} E_j^i D_i = o \\ &\sum_{i=o}^{N} f_{k+i} D_i = o \qquad 1 \leq k \leq N \end{aligned} \tag{5.17}$$

Wegen (5.16) folgt hieraus für $1 \leq j \leq N$:

$$\det \begin{bmatrix} 1 & E_j^1 & \cdots\cdots & E_j^N \\ f_1 & f_2 & \cdots\cdots & f_{N+1} \\ f_2 & f_3 & \cdots\cdots & f_{N+2} \\ \vdots & \vdots & & \vdots \\ f_N & f_{N+1} & \cdots\cdots & f_{2N} \end{bmatrix} = 0$$

Dies bedeutet, daß das Polynom

$$P(x) := \det \begin{bmatrix} x^0 & x^1 & \cdots\cdots & x^N \\ f_1 & f_2 & \cdots\cdots & f_{N+1} \\ \vdots & \vdots & & \vdots \\ f_N & f_{N+1} & \cdots\cdots & f_{2N} \end{bmatrix}$$

die N Nullstellen $E_i = e^{t_i h}$, $1 \leq i \leq N$, 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, $1 \leq i \leq N$, und sind diese reell und weiterhin einfach und positiv, dann erhält man die N reellen Frequenzen

$$t_i = h^{-1} \ln E_i \quad , \quad 1 \leq i \leq N \; .$$

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^o, falls m Funktionswerte (Messwerte) mit $m \geq 2N$ gegeben sind.
Für die Konstruktion einer Näherung aus V_N^o 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 5.1 beschriebene Verfahren.
Es sei also $m \geq 2N$ und $X=\{x_i \mid x_i=a+ih,\ h>o,\ 1 \leq i \leq m\} \subseteq I$ gegeben.

Nimmt man an, daß

$$E(x_j) = f(x_j) \qquad 1 \leq j \leq m$$

mit $E(x)=\sum_{i=1}^{N} a_i e^{t_i x} \in V_N^o - V_{N-1}$ erfüllt ist, so erhält man

mit $A_i := a_i e^{t_i a}$, $E_i := e^{t_i h}$ für $1 \leq i \leq N$ und $f_j := f(x_j)$ für $1 \leq j \leq m$:

$$\sum_{i=1}^{N} A_i E_i^j = f_j \qquad 1 \leq j \leq m \tag{5.18}$$

Durch $Q(x) := \sum_{j=o}^{N} q_j x^j := \prod_{j=1}^{N} (x-E_j)$ sei q_i für $o \leq i \leq N$ gegeben;

Es gilt also

$$Q(E_j) = o \qquad 1 \leq j \leq N \tag{5.19}$$

Es sei nun $k \in \{1,\ldots,m-N\}$ fest und man betrachte in (5.18) die N+1 Gleichungen

$$\sum_{i=1}^{N} A_i E_i^{j+k} = f_{j+k} \qquad o \leq j \leq N \tag{5.2o}$$

Multipliziert man für $o \leq j \leq N$ die (j+1)-te Gleichung von (5.2o) auf beiden Seiten mit q_j, so erhält man mit (5.19) durch anschließende Addition der N+1 Gleichungen:

$$\sum_{j=o}^{N} f_{j+k} q_j = \sum_{j=o}^{N} \left(q_j \sum_{i=1}^{N} (A_i E_i^{k+j}\right) = \sum_{i=1}^{N} \left(A_i E_i^k \sum_{j=o}^{N} q_j E_i^j\right) = o$$

Da diese Beziehung für $1 \leq k \leq m-N$ gilt, erhält man auf diese Weise eine lineares Gleichungssystem mit m-N Gleichungen und N Unbekannten q_i, $o \leq i \leq N-1$, wegen $q_N=1$:

$$\sum_{j=o}^{N-1} f_{j+k} q_j = -f_{N+k} \qquad 1 \leq k \leq m-N \tag{5.21}$$

Es wird daher folgendes Verfahren vorgeschlagen:

Für m=2N bestimme man die Unbekannten q_i, $o \leq i \leq N-1$, durch Lösen des Gleichungssystems (5.21), falls dieses lösbar ist.
Für m>2N bestimme man q_i, $o \leq i \leq N-1$, durch Anwendung der Methode der kleinsten Quadrate.
Hat man so Koeffizienten q_i, $o \leq i \leq N-1$, gefunden, so bestimmt man die Nullstellen E_i, $1 \leq i \leq N$, von $\sum_{i=o}^{N} q_i x^i$ mit $q_N=1$.

Gilt $E_i \in \mathbb{R}$ und $E_i > o$ für $1 \leq i \leq N$, dann erhält man die N Frequenzen $t_i = h^{-1} \ln E_i$, $1 \leq i \leq N$.

Sind die Nullstellen des Polynoms einfach, gilt also $t_i \neq t_j$ für $i \neq j$, dann bestimmt man a_i für $1 \leq i \leq N$ 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 durchgeführt worden; die Programme sind in FORTRAN IV geschrieben und es ist stets mit doppelter Genauigkeit gerechnet worden. Alle Zahlenangaben sind daher gerundete Größen. Die Zeichnungen sind mit dem Plotter des Rechenzentrums und den dort vorhandenen Plotterroutinen angefertigt worden.

In den folgenden Beispielen werden die Bezeichnungen benutzt, wie sie bei der Herleitung der Verfahren eingeführt worden sind.

Für $1 \leq N \leq 4$ treten mit $f(x)=\frac{1}{x+1}$, $f(x)=\sqrt{x}$ in $I=[0,1]$ oder der Riemannschen Zetafunktion in $I=[2,3]$, $I=[2,4]$ keine Schwierigkeiten auf: Es wurde 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 §7.

Eine günstige Näherung liefert das Verfahren von 5.1 mit den Punkten $x_i = 0.01 + ih$, $0 \leq i \leq 3$, wobei $x_3=0.800$ und damit $h=0.2633$ ist. Man erhält die Polynomkoeffizienten

$q_2=0.345\ 112\ 153$, $q_2=-1.467\ 229\ 623$, $q_0=1.000\ 000\ 000$,

$p_1=0.376\ 089\ 942$, $p_0= 1.000\ 000\ 000$.

Die Nullstellen von Q sind 3.398 958 380 und 0.852 499 044.

Es ergeben sich die Parameter

$a_1=-0.483\ 373\ 717$ $\quad t_1=-4.646\ 084\ 908$

$a_2= 0.558\ 037\ 443$ $\quad t_2= 0.606\ 012\ 117$ (Zeichnung 1)

Die Norm der Fehlerfunktion ist 0.07466.

Mit x_o=o.1 , x_3=o.9, h=o.2666 erhält man

a_1=-o.496 967 379 t_1=-2.961 o59 561
a_2= o.655 627 o88 t_2= o.45o 327 955 (Zeichnung 1A)

Die Norm der Fehlerfunktion auf I ist hier o.15866, womit die Abhängigkeit von X demonstriert sei.

Das Verfahren von Rice ergibt mit $x_i=\frac{i-1}{4}$, $1\leq i\leq 5$, das

Polynom $P(x)=p_2x^2+p_1x+p_o$ mit

p_2=-o.67o 54o 689 , p_1=o.965 925 826 , p_o=-o.222 252 952

und den Nullstellen 1.153 o63 231 , o.287 454 61o .
Man erhält so die Parameter

a_1=-o.565 826 6o2 t_1=-4.986 761 225
a_2= o.57o 571 292 t_2= o.569 688 323 (Zeichnung 2)

Für $1\leq i\leq 5$ gilt:

$$|\sqrt{x_i}-a_1e^{t_1x_i}-a_2e^{t_2x_i}|=\text{o.oo4 744 688}$$

Damit ist gezeigt, daß, obwohl x_1=o.o und x_5=1.o Alternantenpunkte der Fehlerfunktion für die beste Approximation an f bzgl. V_2 auf I sind, die Näherung nach Rice, die ja die beste Approximation auf $\{x_i \mid 1\leq i\leq 5\}$ für f ist, in x_1 und x_5 wesentlich von der Minimallösung für f bzgl. V_2 verschieden sein kann
Mit dem Verfahren von Willers kann man oft bessere Näherungen für die Frequenzen erhalten; es werden m Punkte x_i=o.o+(i-1)h mit $1\leq i\leq m$ und x_m=1.o verwendet:
m=1o: Es ist also $h=\frac{1}{9}$; man errechnet die Koeffizienten

q_2=1.ooo ooo ooo , q_1=-1.456 251 7o7 , q_o=o.4o6 468 o37

und damit die Nullstellen 1.o79 834 854 und o.376 416 853 .
Die Frequenzen sind also: t_1=-8.793 522 892
t_2= o.691 273 o52

Durch Lösung des Systems

$$\sum_{i=1}^{2} a_i e^{t_i x_k} = \sqrt{x_k} \qquad k=2,3$$

folgt a_1=-o.42o 954 177
a_2= o.455 428 511 (Zeichnung 3)

Durch Vergrößern von m erhält man nicht unbedingt bessere Werte:

m=22: t_1=-15.928 838 t_2=o.799 146

m=32: t_1=-21.251 ool t_2=o.838 9o8

Schwierigkeiten ergeben sich mit $X \subseteq [-1,1]$ für $f(x)=|x|$ bei $N \geq 3$

und $f(x)=\begin{cases} -5 & : x \leq -o.5 \\ lox & : x \in (-o.5, o.5) \\ +5 & : x \geq o.5 \end{cases}$ oder $f(x)=\begin{cases} 2+x & : x \leq o \\ 2-x & : x \geq o \end{cases}$

für $N \geq 2$:

Beispiel 5.3

$I=[-1,1]$, $N=2$, $f(x)=\begin{cases} 2+x & : x \leq o \\ 2-x & : x \geq o \end{cases}$.

Bei der Berechnung nach 5.1 erhält man für Q meist komplexe Nullstellen, wie die folgende Tabelle zeigt:
(es genügt die Angabe von x_o und x_3)

x_o , x_3		Die Nullstellen von Q Realteil	Imaginärteil
-1.o	1.o	o.8ooo	± o.6ooo
-o.9	o.9	o.8235	± o.5673
-o.8	o.8	o.8462	± o.5329
-o.7	o.7	o.8679	± o.4967
-o.5	o.5	o.9o91	± o.4166
-o.8	o.7	o.7282	± o.4225
-o.7	o.8	1.o274	± o.5962
-o.5	o.4	o.6986	± o.2149

Mit x_o=-o.5 und x_3=o.35 jedoch ist das Verfahren durchführbar:
Man erhält die Polynomkoeffizienten
q_2= 2.837 462 834 q_1=-3.47o 763 132 q_o=1.ooo ooo ooo
p_1=-3.422 811 364 p_o= 1.5oo ooo ooo und damit
die Nullstellen o.758 641 874 und o.464 55o 581 .

Es folgt damit
a_1= 2.82o 549 263
a_2=-o.893 915 427

t_1=o.974 913 359
t_2=2.7o5 946 47o (Zeichnung 4)

Die Minimalabweichung für f in I bzgl. V_1 ist o.5 (die beste Approximation ist die konstante Funktion $E(x)\equiv 1.5 \in V_1$); die gegebene "Näherung" approximiert also schlechter als die beste Approximation bzgl. V_1 .

Weitere Ergebnisse:

Mit x_o=-o.5, x_3=o.3 erhält man

a_1= 2.138 865 18o $\qquad$ t_1=o.692 894 768

a_2=-o.185 691 o49 $\qquad$ t_2=5.381 257 439

Die Norm der Fehlerfunktion ist größer als 37.o .

Mit x_o=-o.5, x_3=o.25 erhält man

a_1= 2.o5o 428 834 $\qquad$ t_1= o.624 759 329

a_2=-o.o5o 428 834 $\qquad$ t_2=1o.2o7 441 475

Die Norm der Fehlerfunktion ist größer als 1365.o;
für diese drei Fehlerfunktionen ist x=1.o der Extremalpunkt.

Bei der Ermittlung einer Näherung nach Rice treten die gleichen Schwierigkeiten auf, da auch P meist komplexe Nullstellen besitzt:

	Die Nullstellen von P	
x_1 , x_5	Realteil	Imaginärteil
-1.o 1.o	o.8571	±o.5151
-o.9 o.9	o.8732	±o.4873
-o.8 o.8	o.8888	±o.4581
-o.5 o.5	o.9333	±o.359o

Für x_1=-o.5 und x_5=o.25 erhält man das Polynom $P(x)=\sum_{i=o}^{2} p_i x^i$

mit p_2=-o.539 o62 5oo, p_1=1.828 125 ooo, p_o=-1.398 437 5oo

und den Nullstellen 2.225 778 oo5, 1.165 526 343 .

Damit folgt

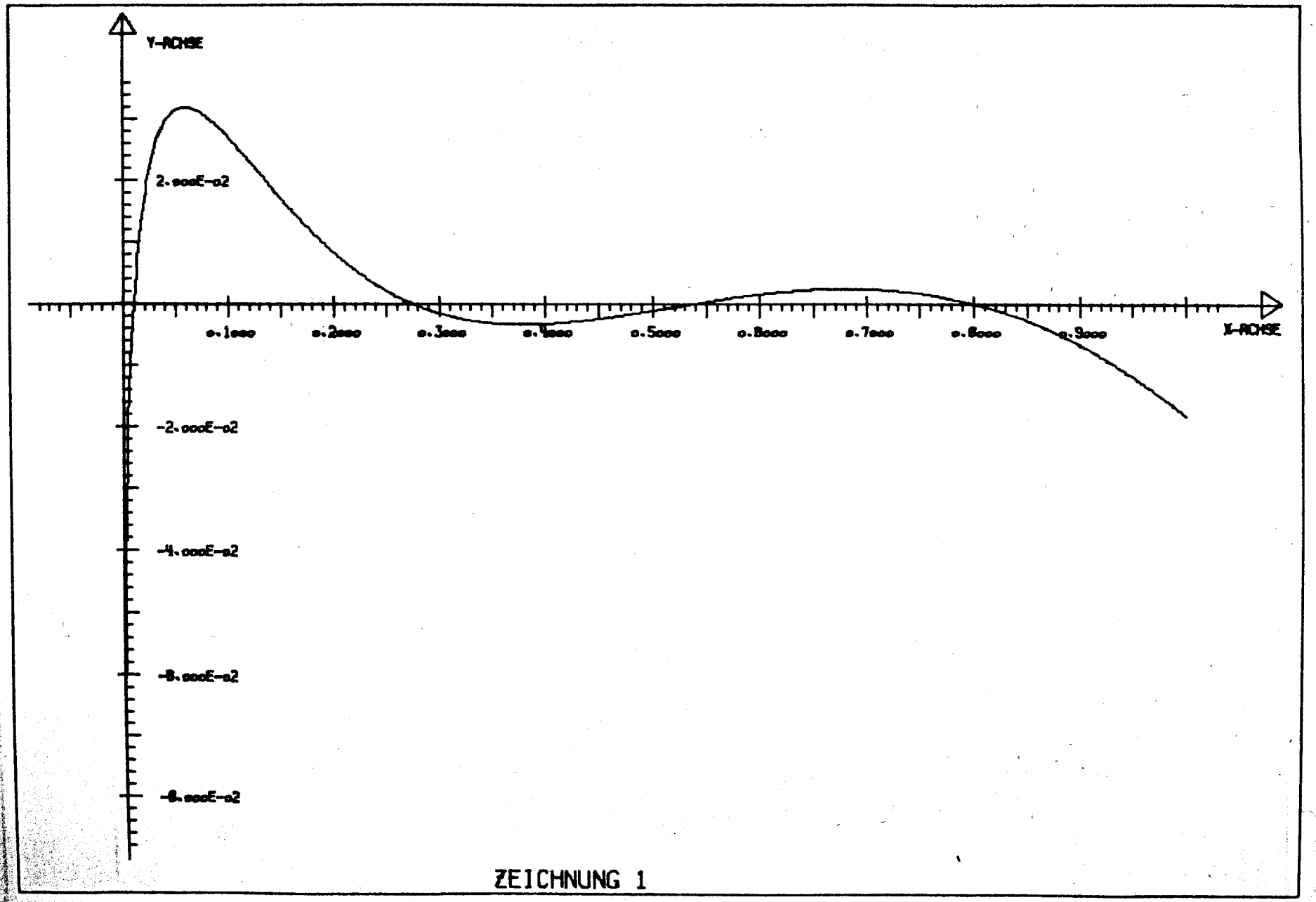
a_1= 2.314 52o 833 $\qquad$ t_1=o.816 921 5o3

a_2=-o.376 853 596 $\qquad$ t_2=4.267 234 771 $\qquad$ (Zeichnung 5)

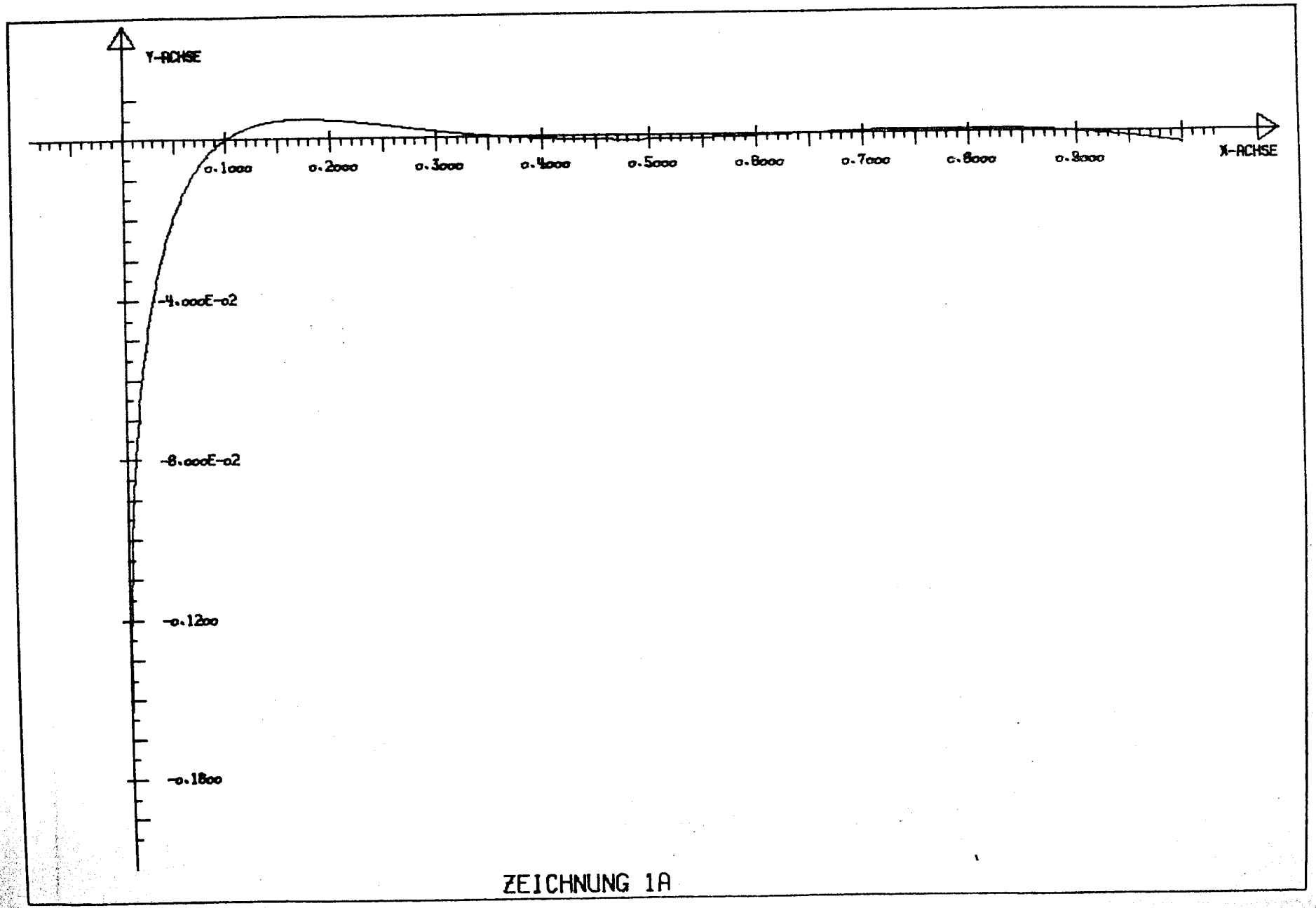
Auch die Berechnung einer Näherung nach 5.3 mit verschiedenen Werten für m führt auf komplexe Nullstellen; es sei $x_1=-1.0$ und $x_m=1.0$:

m	Die Nullstellen von Q	
	Realteil	Imaginärteil
10	0.9329	$\pm$0.1892
22	0.9740	$\pm$0.0831
32	0.9828	$\pm$0.0567

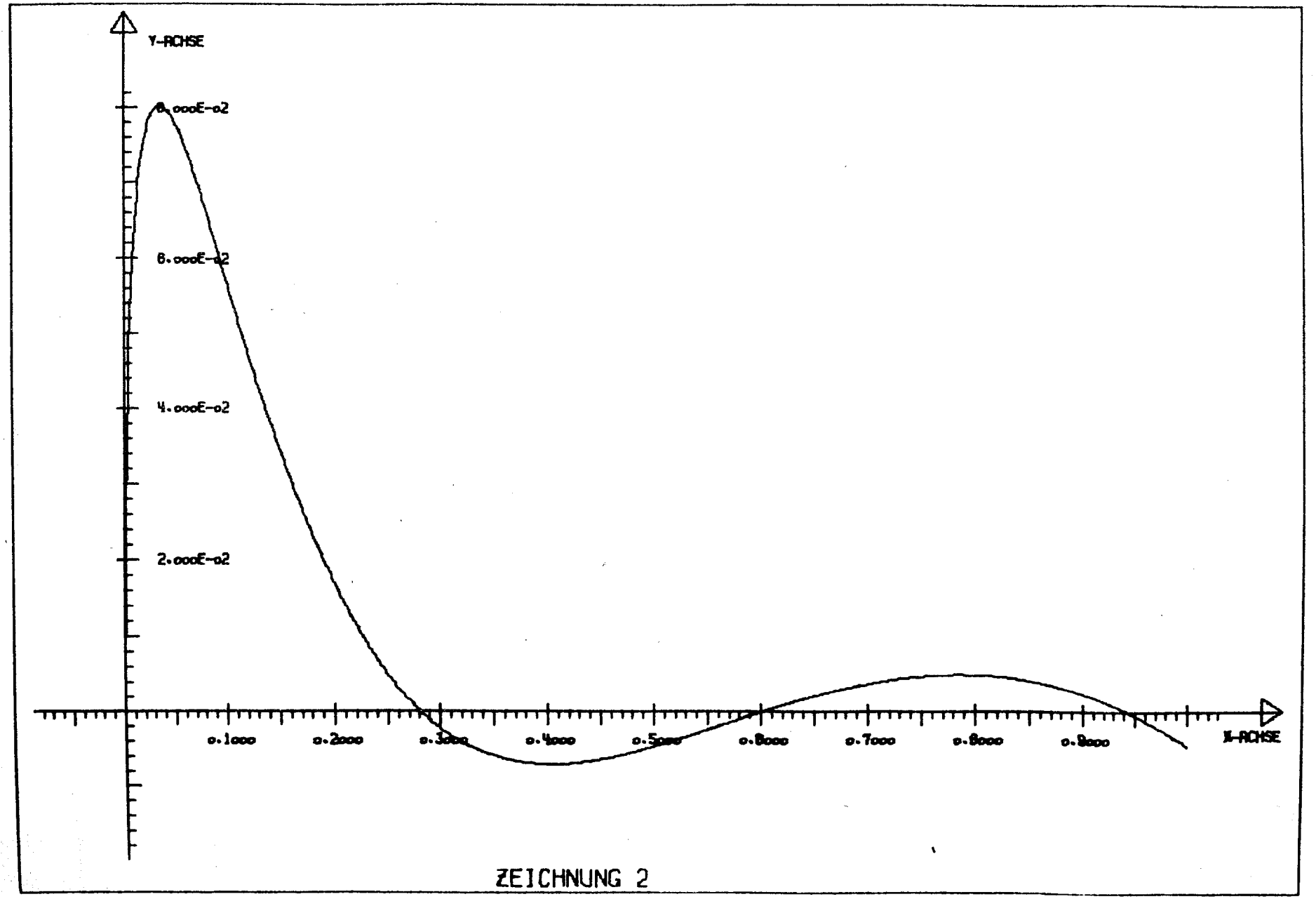
Die hier auftretenden Schwierigkeiten dürften darin begründet sein, daß es für f in $[-1,1]$ nach Satz 3.3 zwei Minimallösungen E_1, E_2 bzgl. V_2 gibt und daher keine beste Approximation bzgl. V_2^o existiert; es liegt also nahe zu vermuten, daß $f-E_i$, $i=1,2$, in I keine Alternante der Länge 5 besitzt, was sich besonders auf die Verfahren von 5.1 und 5.2 auswirkt. Zur Anwendung des Verfahrens von 5.1 auf $f(x)=\frac{1}{1+x}$ bzw. die Riemannsche Zetafunktion sei auf §7 bzw. §8 verwiesen.


Y-ACHSE
2.000E-02
-2.000E-02
-4.000E-02
-6.000E-02
-8.000E-02
0.1000
0.2000
0.3000
0.4000
0.5000
0.6000
0.7000
0.8000
0.9000
X-ACHSE

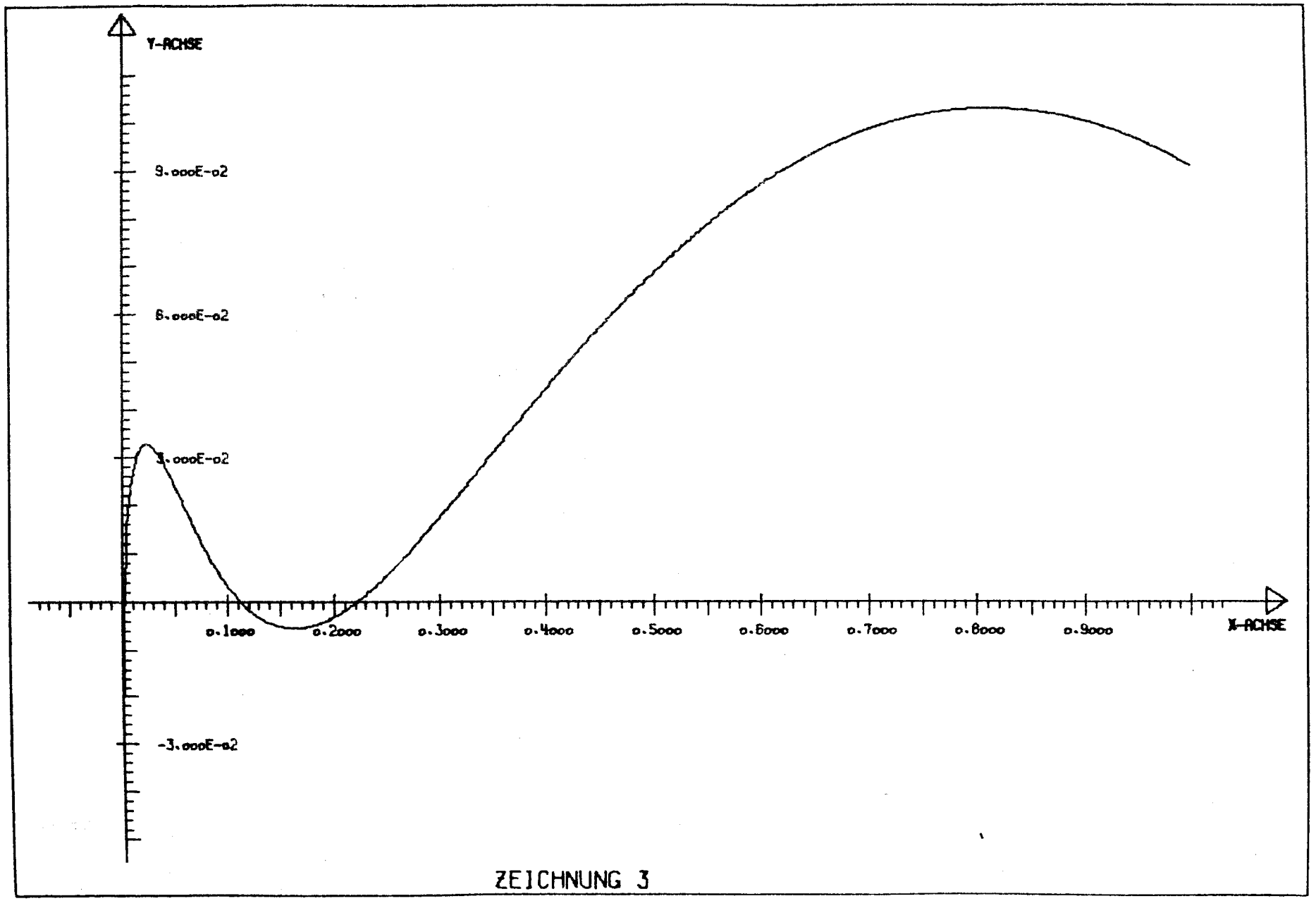
ZEICHNUNG 1



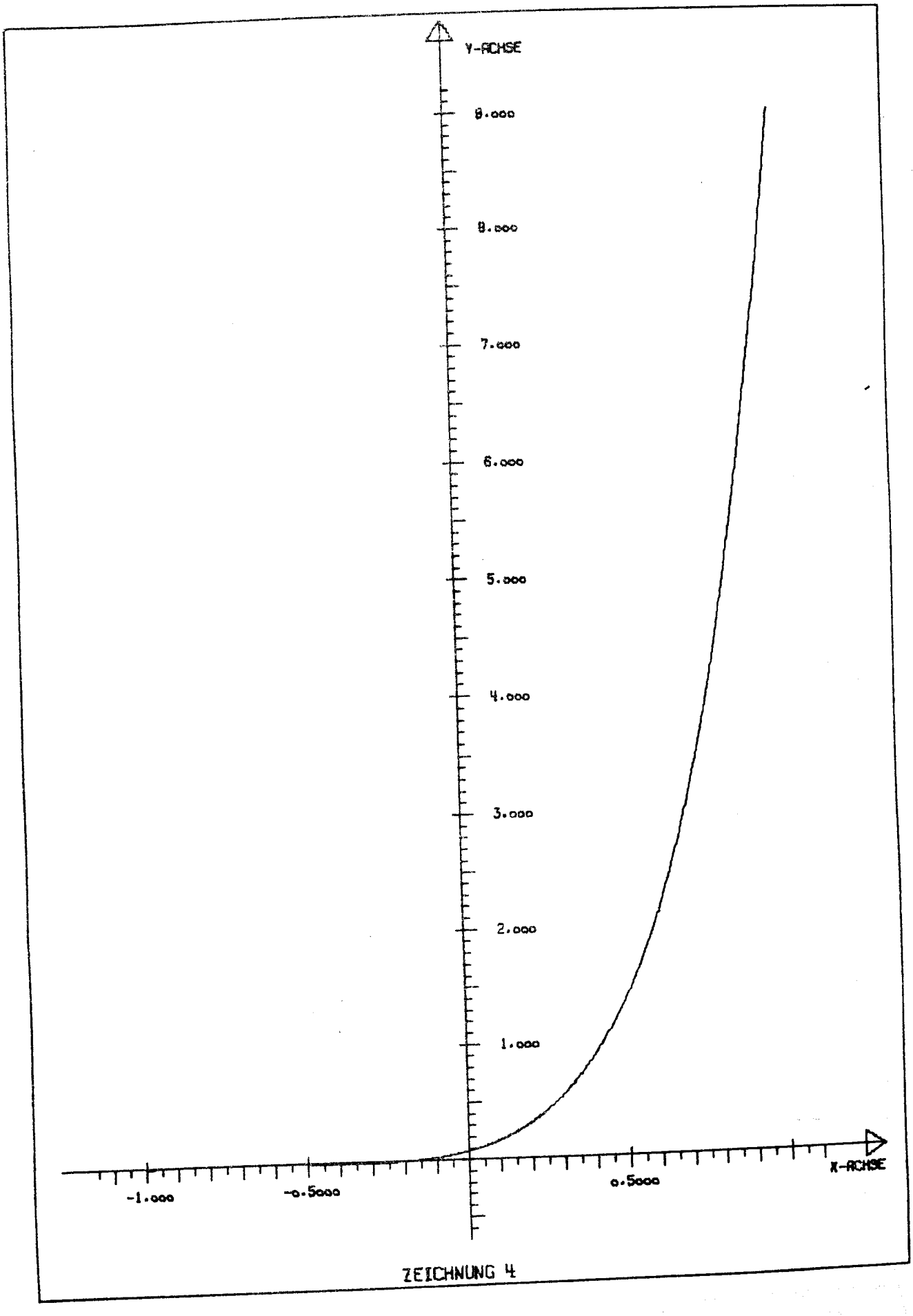
ZEICHNUNG 1A


Y-ACHSE
8.000E-02
6.000E-02
4.000E-02
2.000E-02
0.1000
0.2000
0.3000
0.4000
0.5000
0.6000
0.7000
0.8000
0.9000
X-ACHSE

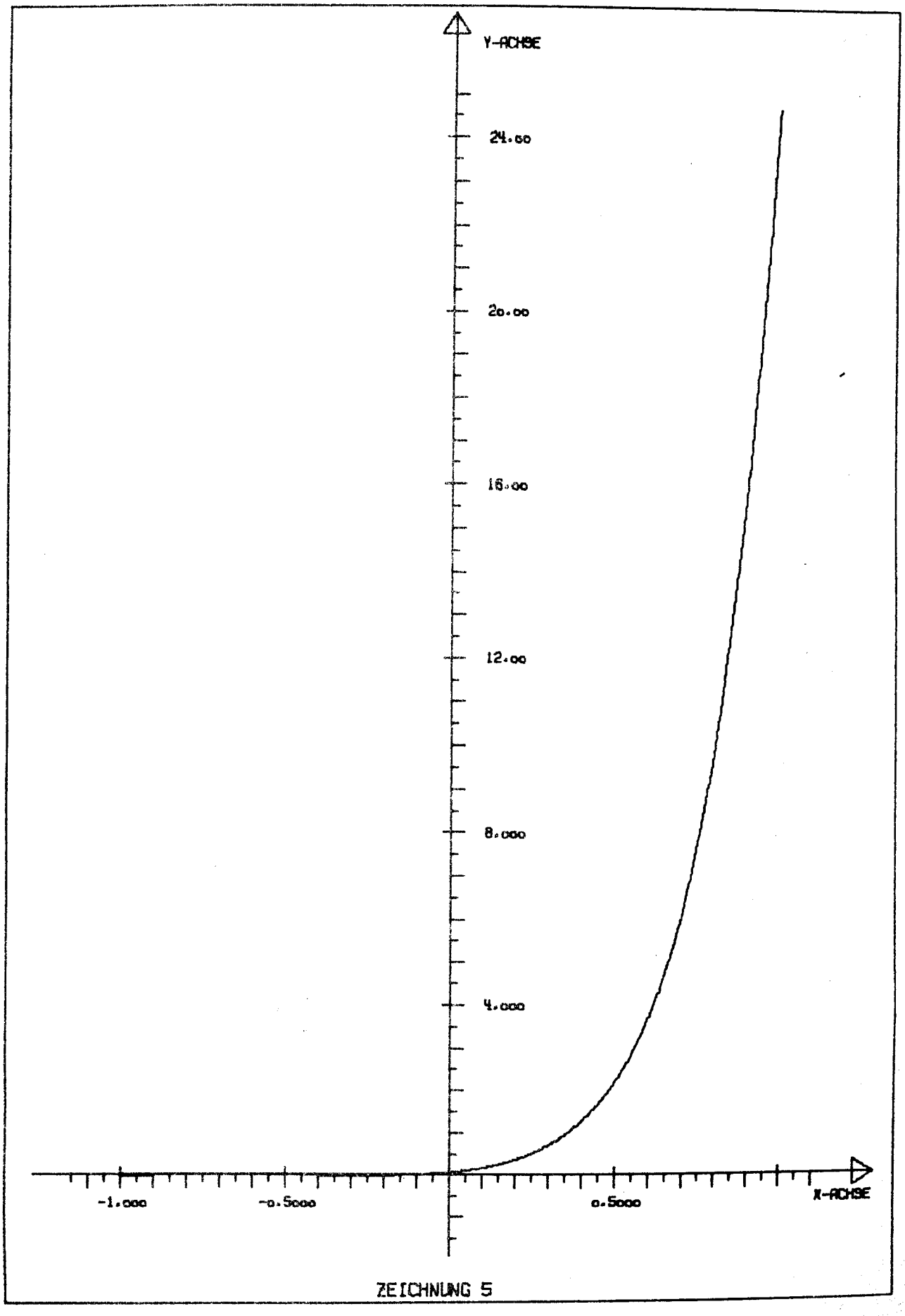
ZEICHNUNG 2


Y-ACHSE
9.000E-02
6.000E-02
3.000E-02
-3.000E-02
0.1000
0.2000
0.3000
0.4000
0.5000
0.6000
0.7000
0.8000
0.9000
X-ACHSE

ZEICHNUNG 3


Y-ACHSE
9.000
8.000
7.000
6.000
5.000
4.000
3.000
2.000
1.000
-1.000
-0.5000
0.5000
X-ACHSE

ZEICHNUNG 4


Y-ACHSE
24.00
20.00
16.00
12.00
8.000
4.000
-1.000
-0.5000
0.5000
X-ACHSE

ZEICHNUNG 5