

9. Tschebyscheff-Approximation: Numerik

Der Remez-Algorithmus.

Der meist verwendete Algorithmus zur Lösung von Approximationsaufgaben bezüglich der Tschebyscheff-Norm ist nach dem russischen Mathematiker Evgeny Yakovlevich Remez (1896–1975) benannt.

Wir gehen wieder von der folgenden Standard-Situation aus: $[a, b] \subset \mathbb{R}$ sei ein kompaktes Intervall, $R := C[a, b]$ sei ausgestattet mit der Maximumsnorm, V sei ein $(n+1)$ -dimensionaler Haarscher Teilraum von R .

Ferner sei $B \subset [a, b]$ eine kompakte Teilmenge, die wenigstens $(n+2)$ Punkte enthält. Schließlich sei $f \in C(B)$.

Wir beginnen mit einer Vorüberlegung, die sich aus dem Beweis des Satzes von de la Vallée-Pouissin ergibt. Es sei $M = \{t_0 < \dots < t_{n+1}\}$ eine $(n+2)$ elementige Teilmenge von B , die als Schätzung für eine Alternante, vgl. (8.19), zu interpretieren ist. Die Bestapproximation von f bezüglich V **auf** M lässt sich dann folgendermaßen ermitteln.

a) Man bestimme $\lambda_0, \dots, \lambda_{n+1}$ aus dem homogenen linearen Gleichungssystem

$$\begin{pmatrix} h_0(t_0) & \dots & h_0(t_{n+1}) \\ \vdots & & \vdots \\ h_n(t_0) & \dots & h_n(t_{n+1}) \end{pmatrix} \begin{pmatrix} \lambda_0 \\ \vdots \\ \lambda_{n+1} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}. \quad (9.1)$$

Skalierung: $\sum_{k=0}^{n+1} |\lambda_k| = 1, \quad \lambda_0 > 0.$

b) Man bestimme die Koeffizienten a_k in der Darstellung $p = \sum a_k h_k$ sowie die Minimalabweichung μ aus den folgenden Bedingungen, vgl. (8.19) und (8.25),

$$f(t_j) - \sum_{k=0}^n a_k h_k(t_j) = \mu \operatorname{sign} \lambda_j, \quad j = 0, \dots, n+1.$$

Dies als lineares Gleichungssystem geschrieben ergibt

$$\begin{pmatrix} h_0(t_0) & \dots & h_n(t_0) & \operatorname{sign} \lambda_0 \\ h_0(t_1) & \dots & h_n(t_1) & \operatorname{sign} \lambda_1 \\ \vdots & & \vdots & \vdots \\ h_0(t_{n+1}) & \dots & h_n(t_{n+1}) & \operatorname{sign} \lambda_{n+1} \end{pmatrix} \begin{pmatrix} a_0 \\ \vdots \\ a_n \\ \mu \end{pmatrix} = \begin{pmatrix} f(t_0) \\ \vdots \\ f(t_n) \\ f(t_{n+1}) \end{pmatrix}. \quad (9.2)$$

Das lineare Gleichungssystem (9.2) ist aufgrund der Voraussetzungen stets eindeutig lösbar (Übungsaufgabe). Ferner folgt aus dem Beweis zum Satz von de la Vallée-Pouissin, dass die λ_k sämtlich von Null verschieden sind und im Vorzeichen alternieren - hierzu wird die Haarsche Bedingung für V als Teilmenge von $C[a, b]$ benötigt.

Zusammen mit der Skalierung $\lambda_0 > 0$ ergibt sich somit

$$\text{sign } \lambda_j = (-1)^j, \quad j = 0, \dots, n+1. \quad (9.3)$$

Damit lässt sich das lineare Gleichungssystem (9.2) ohne explizite Kenntnis der λ_j lösen, nämlich vermöge

$$\begin{pmatrix} h_0(t_0) & \dots & h_n(t_0) & 1 \\ h_0(t_1) & \dots & h_n(t_1) & -1 \\ \vdots & & \vdots & \vdots \\ h_0(t_{n+1}) & \dots & h_n(t_{n+1}) & (-1)^{n+1} \end{pmatrix} \begin{pmatrix} a_0 \\ \vdots \\ a_n \\ \mu \end{pmatrix} = \begin{pmatrix} f(t_0) \\ \vdots \\ f(t_n) \\ f(t_{n+1}) \end{pmatrix}. \quad (9.4)$$

Ablauf des Remez-Algorithmus.

I. Man arbeitet mit einer Schätzung $M_\nu = \{t_0^{(\nu)} < \dots < t_{n+1}^{(\nu)}\} \subset B$ für die Alternante. Dabei bezeichnet ν den Iterationsindex, die Menge M_ν heißt auch ν -te Referenz. Man löst nun das lineare Gleichungssystem (9.4) und bestimmt damit

a) die Bestapproximation $p^{(\nu)} = \sum_{j=0}^n a_j^{(\nu)} h_j \in V$ von f auf M_ν und

b) die zugehörige Minimalabweichung $\mu = \mu_\nu$.

Bemerkung (9.5)

Man beachte auch (8.25) für die Konstruktion der Bestapproximation $p^{(\nu)}$ mittels Interpolationstechniken.

II. Im nächsten Schritt bestimmen wir den tatsächlichen Approximationsfehler in Bezug auf die Menge B

$$\delta_\nu := \max\{|(f - p^{(\nu)})(t)| : t \in B\} \quad (9.6)$$

Gilt $|\mu_\nu| = \delta_\nu$, so ist $p^{(\nu)}$ nach (8.18) die gesuchte Bestapproximation. Für die numerische Realisierung ersetzen wir diese Relation durch das *Abbruchkriterium*

$$\delta_\nu - |\mu_\nu| \leq \text{TOL} \cdot |\mu_\nu|. \quad (9.7)$$

Dabei bezeichne TOL eine vorgegebene (relative) Toleranzschranke.

III. Iterationsschritt:

Ist das obige Abbruchkriterium nicht erfüllt, so ist eine neue (bessere) Referenz $M_{\nu+1}$ zu bestimmen. Wir verlangen dabei, dass die folgenden Konvergenz erzeugenden Bedingungen erfüllt sind

$$\begin{aligned}
\text{(a)} \quad & |(f - p^{(\nu)})(t_j^{(\nu+1)})| \geq |\mu_\nu|, \quad \forall j, \\
\text{(b)} \quad & \exists j_0 : |(f - p^{(\nu)})(t_{j_0}^{(\nu+1)})| \geq (|\mu_\nu| + \delta_\nu)/2, \\
\text{(c)} \quad & \text{sign}[(f - p^{(\nu)})(t_j^{(\nu+1)})] = \sigma(-1)^j, \quad \sigma \in \{\pm 1\}.
\end{aligned} \tag{9.8}$$

Im Folgende zeigen wir, dass sich aus diesen drei Eigenschaften tatsächlich die Konvergenz des Verfahrens folgt.

Satz (9.9)

Unter der Annahme $\delta_\nu > |\mu_\nu|$ folgt mit (9.8): $|\mu_{\nu+1}| > |\mu_\nu|$.

Beweis:

Für das maximale lineare Funktional $\ell_{\nu+1} \in V^\perp$ auf $M_{\nu+1}$ gilt

$$\begin{aligned}
|\mu_{\nu+1}| &= |\ell_{\nu+1}(f)| = \left| \sum_{j=0}^{n+1} \lambda_j^{(\nu+1)} f(t_j^{(\nu+1)}) \right| \\
&= \left| \sum_{j=0}^{n+1} \lambda_j^{(\nu+1)} (f(t_j^{(\nu+1)}) - p^{(\nu)}(t_j^{(\nu+1)})) \right|
\end{aligned}$$

Da die $\lambda_j^{(\nu+1)}$ nicht verschwinden und alternierendes Vorzeichen haben und nach (9.8) (c) auch die Fehlerfunktion $e^{(\nu)}$ auf der neuen Referenz $M_{\nu+1}$ alterniert, ergibt sich

$$|\mu_{\nu+1}| = \sum_{j=0}^{n+1} |\lambda_j^{(\nu+1)}| |f(t_j^{(\nu+1)}) - p^{(\nu)}(t_j^{(\nu+1)})|.$$

Wegen $\sum_j |\lambda_j^{(\nu+1)}| = 1$ folgt nun mit (9.8) (a) und (b)

$$\begin{aligned}
|\mu_{\nu+1}| - |\mu_\nu| &= \sum_{j=0}^{n+1} |\lambda_j^{(\nu+1)}| (|f(t_j^{(\nu+1)}) - p^{(\nu)}(t_j^{(\nu+1)})| - |\mu_\nu|) \\
&\geq |\lambda_{j_0}^{(\nu+1)}| ((|\mu_\nu| + \delta_\nu)/2 - |\mu_\nu|) \\
&= (1/2) |\lambda_{j_0}^{(\nu+1)}| (\delta_\nu - |\mu_\nu|) > 0 \quad \square
\end{aligned}$$

Satz (9.10)

Die zu einer beliebigen Referenz $M = \{t_0 < t_1 < \dots < t_{n+1}\} \subset B$ gemäß (9.1) bestimmten λ_j sind nicht nur sämtlich von Null verschieden und alternierend, sondern sogar gleichmäßig von Null weg beschränkt, d.h. zu jedem $c > 0$ existiert $d = d(f) > 0$, so dass für alle Referenzen M gilt

$$\left| \sum_{j=0}^{n+1} \lambda_j(M) f(t_j) \right| \geq c \Rightarrow \forall j : |\lambda_j(M)| \geq d$$

Beweis:

Würde die Behauptung nicht gelten, so gäbe es ein $c > 0$ und eine Folge von Referenzen $M_\nu = \{t_0^{(\nu)} < \dots < t_{n+1}^{(\nu)}\} \subset B$ mit $|\sum \lambda_j^{(\nu)} f(t_j^{(\nu)})| \geq c$, $\forall \nu$, so dass für ein $j_0 \in \{0, \dots, n+1\}$ gilt $\lambda_{j_0}^{(\nu)} \rightarrow 0$ ($\nu \rightarrow \infty$). Da B kompakt ist und $\sum |\lambda_j^{(\nu)}| = 1$ lassen sich konvergente Teilfolgen von $(\lambda_j^{(\nu)})$ und $(t_j^{(\nu)})$ finden mit $t_j^{(\nu)} \rightarrow t_j$, $\nu \rightarrow \infty$ und $\lambda_j^{(\nu)} \rightarrow \lambda_j$, $\nu \rightarrow \infty$. Dabei ist natürlich $\lambda_{j_0} = 0$.

Sei nun $q \in V$ durch die Interpolationsbedingungen $q(t_j) = f(t_j)$ für $j \neq j_0$ bestimmt. Dann folgt

$$\begin{aligned} \sum_{j=0}^{n+1} \lambda_j^{(\nu)} f(t_j^{(\nu)}) &= \sum_{j=0}^{n+1} \lambda_j^{(\nu)} (f - q)(t_j^{(\nu)}) \\ &\rightarrow \sum_{j=0}^{n+1} \lambda_j (f - q)(t_j) = 0 \quad (\nu \rightarrow \infty) \end{aligned}$$

Widerspruch zur Annahme $|\sum \lambda_j^{(\nu)} f(t_j^{(\nu)})| \geq c > 0$. □

Satz (9.11) (Konvergenz des Remez-Algorithmus)

Erfüllt eine Folge von Referenzen $M_\nu = \{t_0^{(\nu)} < \dots < t_{n+1}^{(\nu)}\} \subset B$, $\nu \in \mathbb{N}_0$, die Konvergenz erzeugenden Bedingungen (9.8), so konvergieren die zugehörigen $p^{(\nu)} \in V$ auf B gleichmäßig gegen die Bestapproximation p^* von f bezüglich V . Diese Konvergenzaussage gilt unabhängig von der Startreferenz M_0 .

Beweis:

(i) Nach (9.9) wächst die Folge der $|\mu_\nu|$ streng monoton. Daher folgt insbesondere

$$\sum_{j=0}^{n+1} \lambda_j^{(\nu)} f(t_j^{(\nu)}) = |\mu_\nu| \geq |\mu_1| > 0.$$

Nach (9.10) und dem Beweis zu (9.9) folgt hiermit

$$|\mu_{\nu+1}| - |\mu_\nu| \geq \frac{1}{2} |\lambda_{j_0}^{(\nu+1)}| (\delta_\nu - |\mu_\nu|) \geq \frac{d}{2} (\delta_\nu - |\mu_\nu|), \quad (9.12)$$

wobei o.E.d.A. $d \in]0, 2[$ gewählt werden kann. Damit ergibt sich

$$\begin{aligned} d_V(f) - |\mu_{\nu+1}| &\leq d_V(f) - |\mu_\nu| - \frac{d}{2} (\delta_\nu - |\mu_\nu|) \\ &\leq (1 - d/2) [d_V(f) - |\mu_\nu|] \end{aligned}$$

Mit $q := (1 - d/2) \in]0, 1[$ liefert die obige Abschätzung

$$0 \leq (d_V(f) - |\mu_{\nu+1}|) \leq q^\nu (d_V(f) - |\mu_0|) \rightarrow 0 \quad (\nu \rightarrow \infty)$$

und damit

$$\lim_{k \rightarrow \infty} |\mu_\nu| = d_V(f). \quad (9.13)$$

(ii) Aus (9.12) folgt analog zur obigen Abschätzung

$$\begin{aligned} 0 &\leq \delta_\nu - d_V(f) \leq \frac{2}{d} (|\mu_{\nu+1}| - |\mu_\nu|) + |\mu_\nu| - d_V(f) \\ &= \frac{2}{d} (|\mu_{\nu+1}| - d_V(f)) - \left(\frac{2}{d} - 1\right) (|\mu_\nu| - d_V(f)) \\ &\leq \left(\frac{2}{d} - 1\right) [d_V(f) - |\mu_\nu|]. \end{aligned}$$

Hieraus folgt: Es gibt eine Konstante $C > 0$ mit

$$0 \leq \|f - p^{(\nu)}\| - d_V(f) \leq C q^\nu. \quad (9.14)$$

(iii) Aufgrund der strikten Eindeutigkeit der Bestapproximation, vgl. (8.15), gilt mit einem $\gamma = \gamma(f) > 0$ für alle $q \in V$:

$$\|f - q\| \geq \|f - p^*\| + \gamma \|q - p^*\|,$$

wobei p^* die Bestapproximation von f aus V bezeichnet. Setzt man $q = p^{(\nu)}$, so folgt mit (9.14)

$$\gamma \|p^{(\nu)} - p^*\| \leq \|f - p^{(\nu)}\| - \|f - p^*\| \leq C q^\nu.$$

Damit ist gezeigt

$$\|p^{(\nu)} - p^*\| \leq \frac{C}{\gamma} q^\nu, \quad \text{mit } 0 < q < 1. \quad \square$$

Bemerkung (9.15)

I. Allg. müssen die Referenzen M_ν nicht notwendigerweise konvergieren; sie sind ja i. Allg. auch nicht eindeutig bestimmt. Aus Kompaktheitsgründen gibt es jedoch stets eine Teilfolge der (M_ν) , die gegen eine Alternante konvergiert.

Der kritische Punkt des Remez-Verfahrens ist es nunmehr, bei vorliegender Referenz M_ν , ein Verfahren zur Konstruktion von $M_{\nu+1}$ anzugeben, so dass die Konvergenz

erzeugenden Eigenschaften (9.8) erfüllt sind. Prinzipiell sind dabei zwei Verfahren in Gebrauch, die als *Einzelaustausch* bzw. *Simultanaustausch* bezeichnet werden.

IV. Einzelaustausch: Man bestimme einen Punkt $\tau \in [a, b]$ mit

$$|(f - p^{(\nu)})(\tau)| \geq \frac{1}{2}(\delta_\nu + |\mu_\nu|).$$

Hierzu könnte man beispielsweise einen Punkt $\tau \in [a, b]$ bestimmen, an dem $|(f - p^{(\nu)})|$ näherungsweise maximal wird. Dies könnte durch numerische Bestimmung einer Nullstelle der Ableitung $(f - p^{(\nu)})'$ mittels Bisektion oder mittels Newton-Verfahren geschehen, oder (einfacher) durch Absuchen auf einem festen, feinen Gitter des Intervalls $[a, b]$.

Austauschregeln (9.16) ($e_\nu := f - p^{(\nu)}$)

(a) Falls $t_{j_0}^{(\nu)} < \tau < t_{j_0+1}^{(\nu)}$:

$$\begin{aligned} t_{j_0}^{(\nu+1)} &:= \tau, \quad t_j^{(\nu+1)} := t_j^{(\nu)} \quad (j \neq j_0), \quad \text{für } \text{sign } e_\nu(t_{j_0}^{(\nu)}) = \text{sign } e_\nu(\tau), \\ t_{j_0+1}^{(\nu+1)} &:= \tau, \quad t_j^{(\nu+1)} := t_j^{(\nu)} \quad (j \neq j_0 + 1), \quad \text{für } \text{sign } e_\nu(t_{j_0+1}^{(\nu)}) = \text{sign } e_\nu(\tau), \end{aligned}$$

(b) Falls $\tau < t_0^{(\nu)}$:

$$\begin{aligned} t_0^{(\nu+1)} &:= \tau, \quad t_j^{(\nu+1)} := t_j^{(\nu)} \quad (j \neq 0), \quad \text{für } \text{sign } e_\nu(t_0^{(\nu)}) = \text{sign } e_\nu(\tau), \\ t_0^{(\nu+1)} &:= \tau, \quad t_j^{(\nu+1)} := t_{j-1}^{(\nu)} \quad (j \neq 0), \quad \text{für } \text{sign } e_\nu(t_0^{(\nu)}) \neq \text{sign } e_\nu(\tau), \end{aligned}$$

(c) Falls $\tau > t_{n+1}^{(\nu)}$:

$$\begin{aligned} t_{n+1}^{(\nu+1)} &:= \tau, \quad t_j^{(\nu+1)} := t_j^{(\nu)} \quad (j \leq n), \quad \text{für } \text{sign } e_\nu(t_{n+1}^{(\nu)}) = \text{sign } e_\nu(\tau), \\ t_{n+1}^{(\nu+1)} &:= \tau, \quad t_j^{(\nu+1)} := t_{j+1}^{(\nu)} \quad (j \leq n), \quad \text{für } \text{sign } e_\nu(t_{n+1}^{(\nu)}) \neq \text{sign } e_\nu(\tau). \end{aligned}$$

Beispiel (9.17)

Zu approximieren sei

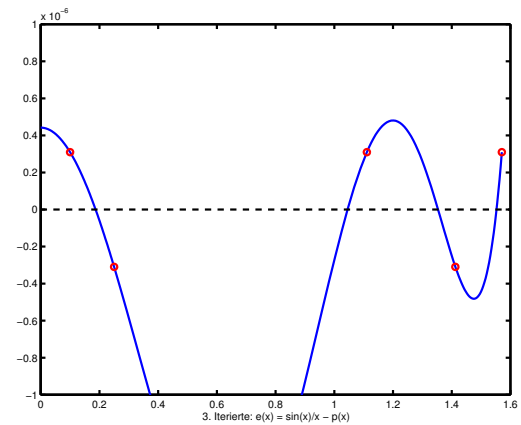
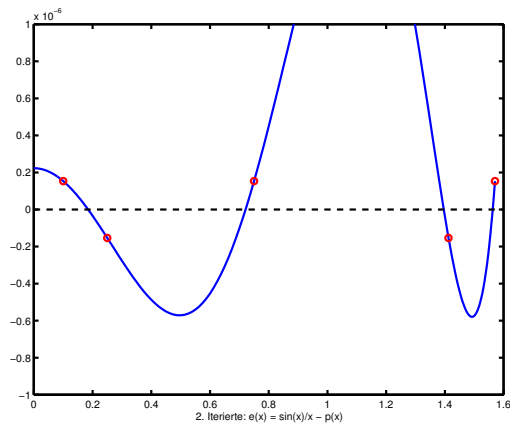
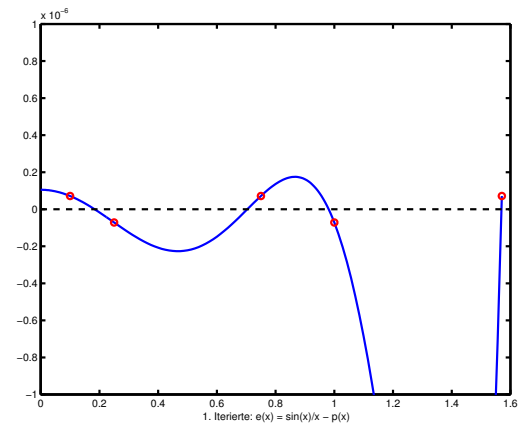
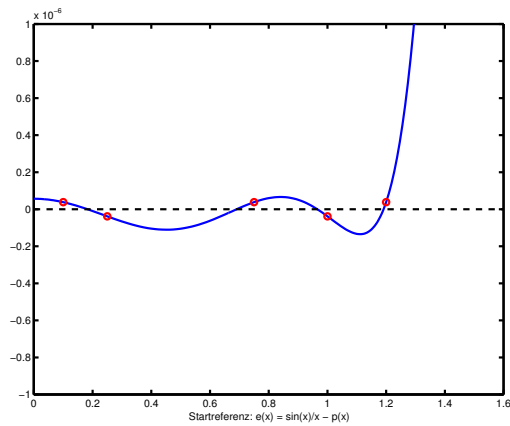
$$f(t) = \frac{\sin t}{t}, \quad 0 \leq t \leq \pi/2$$

durch ein gerades Polynom $p(t) = a_0 + a_1 t^2 + a_2 t^4 + a_3 t^6$.

In der folgenden Tabelle sind die Referenzen der ersten Iterationsschritte des Algorithmus angegeben

iter	t_0	t_1	t_2	t_3	t_4
0	$0.10000e+00$	$0.25000e+00$	$0.75000e+00$	$0.10000e+01$	$0.12000e+01$
1	$0.10000e+00$	$0.25000e+00$	$0.75000e+00$	$0.10000e+01$	$0.15708e+01$
2	$0.10000e+00$	$0.25000e+00$	$0.75000e+00$	$0.14125e+01$	$0.15708e+01$
3	$0.10000e+00$	$0.25000e+00$	$0.11109e+01$	$0.14125e+01$	$0.15708e+01$
4	$0.10000e+00$	$0.63943e+00$	$0.11109e+01$	$0.14125e+01$	$0.15708e+01$
5	$0.00000e+00$	$0.63943e+00$	$0.11109e+01$	$0.14125e+01$	$0.15708e+01$
6	$0.00000e+00$	$0.63943e+00$	$0.11109e+01$	$0.14499e+01$	$0.15708e+01$
10	$0.00000e+00$	$0.59967e+00$	$0.11091e+01$	$0.14506e+01$	$0.15708e+01$

Für die Koeffizienten der Bestapproximation erhält die folgenden Näherungen $a_0 = 0.99999e+00$, $a_1 = -0.16666e+00$, $a_2 = 0.83132e-02$ und $a_3 = -0.18524e-03$. Für die Minimalabweichung ergibt sich $\delta = 0.75439e-06$.



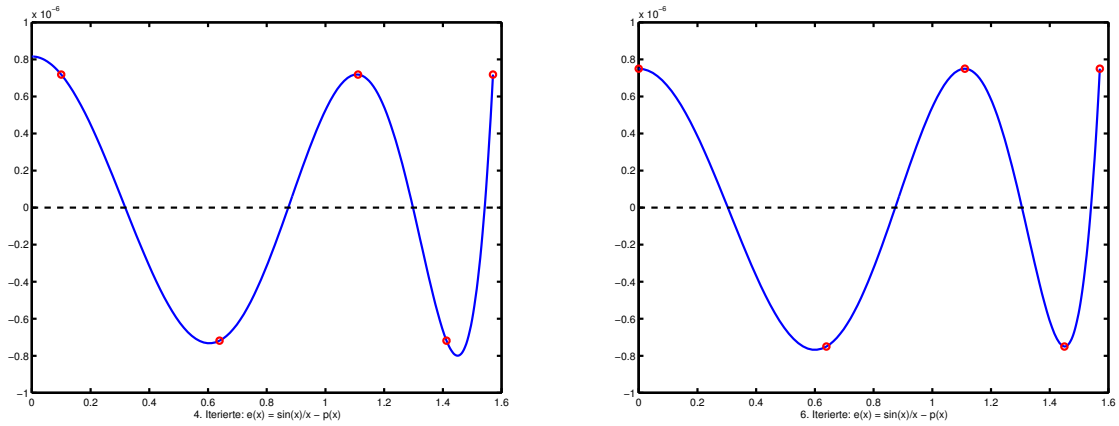


Abb. 9.1: Die ersten Iterationen des Remez-Algorithmus für Beispiel (9.17)

V. Simultanaustausch:

Hierbei zerlegt man das Intervall $[a, b]$ in mindestens $(n+2)$ Teilintervalle, in denen die aktuelle Fehlerfunktion $e_\nu = f - p^{(\nu)}$ abwechselnd nur nichtnegativ bzw. nichtpositiv ist. Man bestimmt dann näherungsweise die Maxima bzw. Minima der Fehlerfunktion in diesen Teilintervallen und wählt hieraus $(n+2)$ Punkte als neue Referenz aus.

Eine Variante dieses Verfahren ergibt sich durch die Bestimmung der lokalen Maxima/Minima der Fehlerfunktion mittels des Newton-Verfahrens für die Ableitung e'_ν . Als Startwerte wählt man die Punkte aus der Referenz M_ν und iteriert wie folgt

$$\begin{aligned} t_0^{(\nu+1)} &= a, \quad t_{n+1}^{(\nu+1)} = b, \quad \text{falls dieses bekannt ist,} \\ t_j^{(\nu+1)} &= t_j^{(\nu)} - \frac{e'_\nu(t_j^{(\nu)})}{e''_\nu(t_j^{(\nu)})}, \quad j = 1, \dots, n \end{aligned} \quad (9.18)$$

Bemerkungen (9.19)

- Der Remez-Algorithmus mit Simultanaustausch ist i. Allg. schneller als der mit Einzelaustausch. Allerdings ist die Technik zur Sicherung der globalen Konvergenz mühsamer.
- Beim Einzelaustausch lassen sich zur Lösung des linearen Gleichungssystems (9.4) so genannte *update-Techniken* verwenden.
- Für den Remez-Algorithmus mit Simultanaustausch lässt sich unter gewissen Zusatzvoraussetzungen (u.a. die C^2 -Eigenschaft von f, h_j) die *quadratische Konvergenz* des Verfahrens zeigen, d.h.

$$\exists C > 0 : \quad [d_V(f) - |\mu_{\nu+1}|] \leq C [d_V(f) - |\mu_\nu|]^2$$

- Die Wahl der Startreferenz M_0 ist wegen der globalen Konvergenz i. Allg. nicht sehr kritisch; man kann M_0 beispielsweise über eine L_2 -Approximation – z.B. eine Tschebyscheff-Entwicklung – erhalten.

Das Newton–Verfahren.

Als Alternative zum Remez–Algorithmus bietet sich an, das nichtlineare Gleichungssystem (8.20) des Alternantensatzes mit einer geeigneten Variante des Newton–Verfahrens zu lösen. Gehören etwa beide Randpunkte des Intervalls $[a, b]$ zur Alternante (ein hinreichendes Kriterium gibt Satz (8.24) an), so liefert der Alternantensatz das folgende nichtlineare Gleichungssystem

$$\begin{aligned} \sum_{j=0}^n a_j h_j(t_k) - f(t_k) + (-1)^k \mu &= 0, \quad k = 0, \dots, n+1, \\ \sum_{j=0}^n a_j h'_j(t_k) - f'(t_k) &= 0, \quad k = 1, \dots, n. \end{aligned} \quad (9.20)$$

Dies sind $(2n + 2)$ Gleichungen in den $(2n + 2)$ Unbekannten $(a_0, \dots, a_n, t_1, \dots, t_n, \mu)$. Bezeichnet man mit $e(t) := f(t) - \sum a_j h_j(t)$ wieder die Fehlerfunktion, so lautet die zugehörige Newton–Gleichung, d.h. das lineare Gleichungssystem zur Berechnung der Newton–Korrekturen Δa_j , Δt_k und $\Delta \mu$:

$$\begin{aligned} \sum_{j=0}^n h_j(t_k) \Delta a_j - e'(t_k) \Delta t_k + (-1)^k \Delta \mu &= e(t_k) - (-1)^k \mu, \\ &k = 0, 1, \dots, n+1 \\ \sum_{j=0}^n h'_j(t_k) \Delta a_j - e''(t_k) \Delta t_k &= e'(t_k), \\ &k = 1, \dots, n. \end{aligned} \quad (9.21)$$

Dabei ist $\Delta t_0 := \Delta t_{n+1} := 0$, $a_j^{(\nu+1)} := a_j^{(\nu)} + \Delta a_j$, $t_k^{(\nu+1)} := t_k^{(\nu)} + \Delta t_k$ und $\mu^{(\nu+1)} := \mu^{(\nu)} + \Delta \mu$.

Schreibt man dieses lineare Gleichungssystem auf die Unbekannten $a_j^{(\nu+1)}$, Δt_k und $\mu^{(\nu+1)}$ um (dies sollte man nicht machen, wenn man Dämpfungsstrategien verwenden möchte), so ergibt sich mit $t_k = t_k^{(\nu)}$

$$\begin{aligned} \sum_{j=0}^n h_j(t_k) a_j^{(\nu+1)} - e'(t_k) \Delta t_k + (-1)^k \mu^{(\nu+1)} &= f(t_k), \\ &k = 0, 1, \dots, n+1 \\ \sum_{j=0}^n h'_j(t_k) a_j^{(\nu+1)} - e''(t_k) \Delta t_k &= f'(t_k), \\ &k = 1, \dots, n. \end{aligned} \quad (9.22)$$

Man vergleiche diese Relationen mit den entsprechenden Gleichungen (9.4) und (9.18) des Remez–Verfahrens.

Zusammenhang zur Linearen Optimierung.

Wir betrachten eine diskrete Approximationsaufgabe im Tschebyscheffschen Sinn. Dazu sei $B \subset [a, b]$ eine endliche Menge, $B = \{t_1, \dots, t_m\}$ mit $\#B = m > n + 2$.

Wir können ferner davon ausgehen, dass i. Allg. $m \gg n$ gelten wird. Wieder sei $V = V_n$ ein $(n+1)$ dimensionaler linearer Teilraum von $C(B)$ mit Basis $(h_0, \dots, h_n)$.

Die *Approximationsaufgabe* lautet: Man bestimme $(a_0, \dots, a_n) \in \mathbb{R}^{n+1}$, so dass

$$I(a_0, \dots, a_n) := \max\left\{ \left| f(t_k) - \sum_{j=0}^n a_j h_j(t_k) \right| : k = 1, \dots, m \right\} \quad (9.23)$$

minimal wird.

Diese Approximationsaufgabe lässt sich nun unmittelbar in eine lineare Optimierungsaufgabe transformieren. Dazu definieren wir

$$\delta := \max\left\{ \left| f(t_k) - \sum_{j=0}^n a_j h_j(t_k) \right| : k = 1, \dots, m \right\} \quad (9.24)$$

(9.23) ist dann äquivalent zur *linearen Optimierungsaufgabe*:

Bestimme $(a_0, \dots, a_n, \delta) \in \mathbb{R}^{n+2}$, so dass $J := \delta$ minimal wird unter den Nebenbedingungen

$$\begin{aligned} \sum_{j=0}^n a_j h_j(t_k) - \delta &\leq f(t_k) \\ - \sum_{j=0}^n a_j h_j(t_k) - \delta &\leq -f(t_k), \quad k = 1, \dots, m \end{aligned} \quad (9.25)$$

Um die Standardformulierung einer linearen Optimierungsaufgabe zu erhalten führen wir die folgenden Definitionen ein

$$\mathbf{A}^T := \begin{pmatrix} h_0(t_1) & \dots & h_n(t_1) & -1 \\ \vdots & & \vdots & \vdots \\ h_0(t_m) & \dots & h_n(t_m) & -1 \\ -h_0(t_1) & \dots & -h_n(t_1) & -1 \\ \vdots & & \vdots & \vdots \\ -h_0(t_m) & \dots & -h_n(t_m) & -1 \end{pmatrix} \in \mathbb{R}^{(2m, n+2)} \quad (9.26)$$

$$\mathbf{b}^T := [f(t_1) \dots f(t_m), -f(t_1) \dots -f(t_m)] \in \mathbb{R}^{2m}$$

$$\mathbf{z}^T := [a_0 \dots a_n, \delta] \in \mathbb{R}^{n+2}$$

$$\mathbf{c}^T := [0 \dots 0, -1] \in \mathbb{R}^{n+2}.$$

Damit lautet die lineare Optimierungsaufgabe schließlich

$$\text{Maximiere } J_D(\mathbf{z}) = \mathbf{c}^T \mathbf{z}; \quad \text{Nebenbedingungen: } \mathbf{A}^T \mathbf{z} \leq \mathbf{b}. \quad (9.27)$$

Bemerkungen (9.28)

Die obige Darstellung (9.27) heißt auch **Dualform** einer linearen Optimierungsaufgabe (vgl. Numerik-Vorlesung bzw. Opfer: Numerische Mathematik). Die Zielfunktion $J_D(\mathbf{z})$ ist auf der zulässigen Menge $Z := \{\mathbf{z} : \mathbf{A}^T \mathbf{z} \leq \mathbf{b}\}$ nach oben beschränkt (durch $-\delta = 0$). Ferner ist Z nichtleer, da jeder Punkt $\mathbf{z} = (\mathbf{0}, \delta)^T$ mit $\delta \geq \|f\|_\infty$ zulässig ist. Nach der Theorie der linearen Optimierungsaufgaben (vgl. J. Werner: Numerische Mathematik 2; Satz 2.4) folgt hieraus, dass (9.27) wenigstens eine Lösung $\mathbf{z}^*$ besitzt.

Da i. Allg. $m \gg n$ gelten wird, empfiehlt es sich nicht, die Ungleichungen $\mathbf{A}^T \mathbf{z} \leq \mathbf{b}$ durch Einführung von Schlupfvariablen in Gleichungen zu transformieren. Vielmehr ist es i. Allg. vorteilhaft, anstelle des dualen Problems (9.27) das zugehörige *primale Problem* zu lösen (etwa mit dem Simplexverfahren). Das primale lineare Optimierungsproblem in Normalform lautet

$$\text{Minimiere } J_P(\mathbf{y}) = \mathbf{b}^T \mathbf{y}; \quad \text{Nebenbedingungen: } \mathbf{A} \mathbf{y} = \mathbf{c}, \quad \mathbf{y} \geq \mathbf{0}. \quad (9.29)$$

Dabei ist $\mathbf{y} \in \mathbb{R}^{2m}$.

Schreibt man diese lineare Optimierungsaufgabe mittels der Definitionen (9.26) wieder explizit auf, so ergibt sich das **primale Optimierungsproblem**

$$\text{Minimiere} \quad J_P(\mathbf{y}) = \sum_{k=1}^m f(t_k) (y_k - y_{m+k}) \quad (9.30)$$

unter den Nebenbedingungen

$$\begin{aligned} \sum_{k=1}^m h_j(t_k) (y_k - y_{m+k}) &= 0, \quad j = 0, \dots, n \\ \sum_{k=1}^{2m} y_k &= 1, \quad y_k \geq 0, \quad k = 1, \dots, 2m. \end{aligned} \quad (9.31)$$

Bemerkungen (9.32)

a) Die Normierung $\sum y_k = 1$ lässt sich abschwächen zu $\sum y_k \leq 1$. Gilt nämlich für einen zulässigen Punkt $\mathbf{y}$ des relaxierten Problems $\sum y_k < 1$, so erhöhe man irgendein y_{k_0} und das zugehörige y_{m+k_0} um den gleichen Wert $(1 - \sum y_k)/2 > 0$. Der neue Vektor $\tilde{\mathbf{y}}$ ist dann zulässig für (9.30) bei gleichem Wert der Zielfunktion.

b) Weiterhin lässt sich für das relaxierte lineare Optimierungsproblem durch den Übergang

$$\begin{pmatrix} y_k \\ y_{m+k} \end{pmatrix} \rightarrow \begin{pmatrix} y_k - \min\{y_k, y_{m+k}\} \\ y_{m+k} - \min\{y_k, y_{m+k}\} \end{pmatrix} \geq \mathbf{0}$$

erreichen, dass die *Komplementaritätsbedingung* $y_k y_{m+k} = 0$ erfüllt ist.

Relaxiertes Primales Optimierungsproblem

$$\text{Minimiere} \quad J_P(\mathbf{y}) = \sum_{k=1}^m f(t_k) (y_k - y_{m+k}) \quad (9.33)$$

unter den Nebenbedingungen

$$\begin{aligned} \sum_{k=1}^m h_j(t_k) (y_k - y_{m+k}) &= 0, \quad j = 0, \dots, n \\ \sum_{k=1}^{2m} y_k &\leq 1, \quad y_k y_{m+k} = 0, \quad k = 1, \dots, m \\ y_k &\geq 0, \quad k = 1, \dots, 2m. \end{aligned} \quad (9.34)$$

Ist $\mathbf{y}$ zulässiger Basisvektor des primalen Optimierungsproblems, so gibt es $I = \{k_0, \dots, k_{n+1}\} \subset \{1, \dots, m\}$ mit $y_k = y_{m+k} = 0$ für alle $k \notin I$. Unter Beachtung der Komplementaritätsbedingung sei nun

$$\begin{aligned} \lambda_i &:= y_{k_i} - y_{m+k_i}, \\ |\lambda_i| &= y_{k_i} + y_{m+k_i}, \quad i = 0, \dots, n+1. \end{aligned} \quad (9.35)$$

Damit findet man

$$\begin{aligned} J_P &= \sum_{i=0}^{n+1} f(t_{k_i}) \lambda_i \\ \sum_{i=0}^{n+1} h_j(t_{k_i}) \lambda_i &= 0, \quad j = 0, \dots, n \\ \sum_{i=0}^{n+1} |\lambda_i| &= 1. \end{aligned} \quad (9.36)$$

Sei $\mathbf{y}$ nun *optimale Basislösung*, $\mathbf{z} = (a_0, \dots, a_n, \delta)^T$ sei Lösung des dualen Problems (9.27). Wegen des Dualitätssatzes, vgl. Vorlesung über Numerik, gilt dann

$$J_D = J_P = \mathbf{b}^T \mathbf{y} = \mathbf{c}^T \mathbf{z}$$

Damit folgt

$$\begin{aligned} 0 &= \mathbf{y}^T (\mathbf{b} - \mathbf{A}^T \mathbf{z}) \\ &= \sum_{k=1}^m y_k [f(t_k) - \sum_{j=0}^n a_j h_j(t_k) + \delta] \\ &\quad + \sum_{k=1}^m y_{m+k} [-f(t_k) + \sum_{j=0}^n a_j h_j(t_k) + \delta] \end{aligned}$$

Hier sind alle Summanden nichtnegativ, so dass sich die folgenden Komplementaritätsbedingungen ergeben

$$\begin{aligned} y_k [f(t_k) - \sum_{j=0}^n a_j h_j(t_k) + \delta] &= 0, \\ y_{m+k} [-f(t_k) + \sum_{j=0}^n a_j h_j(t_k) + \delta] &= 0. \end{aligned} \quad (9.37)$$

Zusammen mit (9.35) und (9.36) folgt nun

$$\begin{aligned} k \notin I &\Rightarrow y_k = y_{m+k} = 0, \\ k_i \in I, \quad \lambda_i > 0 &\Rightarrow f(t_{k_i}) - \sum_{j=0}^n a_j h_j(t_{k_i}) = -\delta, \\ k_i \in I, \quad \lambda_i < 0 &\Rightarrow f(t_{k_i}) - \sum_{j=0}^n a_j h_j(t_{k_i}) = \delta. \end{aligned}$$

Damit lässt sich (9.37) als ein lineares Gleichungssystem zur Berechnung von $\mathbf{z} = (a_0, \dots, a_n, \delta)^T$ ansehen. Nach Herleitung ist dieses Gleichungssystem stets lösbar; die Eindeutigkeit der Lösung ist jedoch nur für $\lambda_k \neq 0$, für alle $k \in I$, gewährleistet.

Wir fassen das Ergebnis zusammen

Satz (9.38)

a) $p = \sum a_j h_j$ ist genau dann Bestapproximation von f aus V auf $B = \{t_1, \dots, t_m\}$, wenn es Punkte $t_{k_0} < \dots < t_{k_{r+1}} \in B$ gibt, $0 \leq r \leq n$ und $\lambda_i \neq 0$ mit

$$\begin{aligned} f(t_{k_i}) - \sum_{j=0}^n a_j h_j(t_{k_i}) &= \text{sign} \lambda_i \|f - p\|_\infty, \quad i = 0, \dots, r+1, \\ \sum_{i=0}^{r+1} \lambda_i h_j(t_{k_i}) &= 0, \quad j = 0, \dots, n, \\ \sum_{i=0}^{r+1} |\lambda_i| &= 1. \end{aligned}$$

b) Ist $\mathbf{y}$ optimale Basislösung zu (9.29), so ist das lineare Gleichungssystem (9.37) für $\mathbf{z} = (a_0, \dots, a_n, \delta)^T$ lösbar. Jede Lösung liefert eine Bestapproximation von f aus V auf B .

Bemerkung (9.39)

Genügt V der Haarschen Bedingung, so ist jede (zulässige) Basislösung des primalen Problems nichtentartet, d.h. $\forall i = 0, \dots, n+1 : y_{k_i} \neq 0$. Ferner lässt sich analog zum Beweis des Satzes von de la Vallée, Pouissin folgern, dass die λ_i nicht verschwinden und alternieren.

Beispiel (9.40) (aus J. Werner: Numerische Mathematik 2)

Zu minimieren sei $\max\{|e^{t_j} - p(t_j)| : j\}$ über $p \in \Pi_4$ auf dem Gitter $t_j = (j-1)/10, j = 1, \dots, 11$.

Das zugehörige (primale) lineare Optimierungsproblem hat die folgende Form

$$\text{Minimiere } J_P = \mathbf{b}^T \mathbf{y}, \quad \mathbf{y} \in \mathbb{R}^{22}$$

unter den Nebenbedingungen

$$\mathbf{A}\mathbf{y} = \mathbf{c}, \quad \mathbf{y} \geq \mathbf{0}.$$

Dabei ist

$$\mathbf{A}^T := \begin{pmatrix} 1 & t_1 & \dots & t_1^4 & -1 \\ \vdots & & & \vdots & \vdots \\ 1 & t_{11} & \dots & t_{11}^4 & -1 \\ -1 & -t_1 & \dots & -t_1^4 & -1 \\ \vdots & & & \vdots & \vdots \\ -1 & -t_{11} & \dots & -t_{11}^4 & -1 \end{pmatrix} \in \mathbb{R}^{(22,6)}$$

$$\mathbf{b}^T := [e^{t_1} \dots e^{t_{11}}, -e^{t_1} \dots -e^{t_{11}}] \in \mathbb{R}^{2m}$$

$$\mathbf{c}^T := [0 \dots 0, -1] \in \mathbb{R}^6.$$

Die numerische Lösung des primalen Problems mit Hilfe der MATLAB Routine **linprog** ergibt eine optimale nichtentartete Basislösung mit den Basisindizes $J_B = (1, 5, 10, 13, 19, 22)$.

Das zugehörige lineare Gleichungssystem (9.37) mit diesen Indizes liefert schließlich die Lösung

$$\mathbf{a} \approx \begin{pmatrix} 1.000026 \\ 0.998714 \\ 0.510077 \\ 0.139716 \\ 0.069722 \end{pmatrix}, \quad \delta \approx 2.602631e-05.$$

