Beiträge zur Numerischen Mathematik 7

Beiträge zur Numerischen Mathematik 7

Herausgegeben von
Frieder Kuhnert und Jochen W. Schmidt

R. Oldenbourg Verlag München Wien 1979

CIP-Kurztitelaufnahme der Deutschen Bibliothek

Beiträge zur Numerischen Mathematik / hrsg. von
Frieder Kuhnert u. Jochen W. Schmidt. — München,
Wien: Oldenbourg.
NE: Kuhnert, Frieder [Hrsg.]
7. — 1. Aufl. — 1979.
ISBN 3-486-22411-5

Printed in the German Democratic Republic
Lizenz-Nr. 206 · 435/116/78
Gesamtherstellung: VEB Druckhaus „Maxim Gorki", Altenburg
ISBN: 3-486-22411-5

Inhalt

Beiträge zur Numerischen Mathematik
7 (1979), 7–12

Error estimations for conjugate direction methods

JÓZSEF ABAFFY and AURÉL GALÁNTAI

This paper gives an error estimation for the class of conjugate direction methods when the objective function is quadratical.

1. Introduction

It is an important problem of the unconstrained function minimization to minimize quadratical functions of the form

$$f(\boldsymbol{x}) = \frac{1}{2}\,\boldsymbol{x}^{\mathsf{T}} A\boldsymbol{x} + \boldsymbol{b}^{\mathsf{T}}\boldsymbol{x} + c \qquad (\boldsymbol{b}, \boldsymbol{x} \in R^n, c \in R^1) \tag{1.1}$$

where A is an $n \times n$ positive definite and symmetrical matrix.

The most important methods are based on the computation of conjugate directions ([1–5]). As it is known, the conjugate directions corresponding to the matrix A are defined as a linearly independent system of vectors $\{\boldsymbol{d}_i\}_{i=1}^n$ for which

$$\boldsymbol{d}_i^{\mathsf{T}} A \boldsymbol{d}_j = \delta_{ij} \cdot c_{ij} \qquad (i, j = 1, 2, \ldots, n), \tag{1.2}$$

where δ_{ij} denotes the Kronecker symbol and $c_{ij} \neq 0$ are constant.

The conjugate direction methods then have the following general form:

Let $\boldsymbol{x}_1 \in R^n$ be arbitrary and

$$\boldsymbol{x}_{i+1} + \boldsymbol{x}_i + \alpha_i \boldsymbol{d}_i \qquad (i = 1, 2, \ldots, n), \tag{1.3}$$

where α_i is defined by

$$\alpha_i = -\boldsymbol{g}(\boldsymbol{x}_i)^{\mathsf{T}}\, \boldsymbol{d}_i / \boldsymbol{d}_i^{\mathsf{T}} A \boldsymbol{d}_i \tag{1.4}$$

and $\boldsymbol{g}(\boldsymbol{x}_i) = \operatorname{grad} f(\boldsymbol{x}_i)$.

In this scheme the methods differ from each others only in the computation way of conjugate directions $\{\boldsymbol{d}_i\}_{i=1}^n$.

Without loss of generality we can assume that

$$\|\boldsymbol{d}_i\| = 1 \qquad (i = 1, 2, \ldots, n). \tag{1.5}$$

Note that the number α_i for non-quadratical functions is defined by the relation

$$f(\boldsymbol{x}_{i+1}) = \min_{\alpha \in R} f(\boldsymbol{x}_i + \alpha \boldsymbol{d}_i) \tag{1.6}$$

which is equivalent to (1.4) for the quadratical functions.

In two special cases the error of the above scheme was estimated by M. SACHET and S. KAHNE [4] for quadratical functions. The first case supposes error only in one step of the process and exact computations in the further steps. In the second case it is assumed that the errors in the computation of α_i and $\boldsymbol{d}_i$ are independent of each others. In this way a simple estimation of the error was obtained.

In this paper we give a general error estimation (that means the errors of the different steps may depend on each others) which is obviously sharper than the result of SACHET and KAHNE.

2. A closed form of the error for quadratical functions

This section gives a closed form for the error propagation.

We denote the perturbed conjugate directions by $\boldsymbol{p}_i$ $(i = 1, \ldots, n;\ \boldsymbol{p}_i \neq \boldsymbol{0})$. The computed minimumpoints in the linear subspace of the vectors $\boldsymbol{d}_1, \ldots, \boldsymbol{d}_i$ are denoted by $\boldsymbol{y}_i$ $(i = 1, \ldots, n+1)$. Similarly, β_i denotes the computed value of α_i $(i = 1, \ldots, n)$. Let $\boldsymbol{\delta}_i = \boldsymbol{p}_i - \boldsymbol{d}_i$ be the error of direction $\boldsymbol{d}_i$ $(i = 1, \ldots, n)$ and let us decompose β_i in the form

$$\beta_i = \alpha_i + \gamma_i + \varepsilon_i \qquad (i = 1, \ldots, n), \tag{2.1}$$

where

$$\gamma_i = -\boldsymbol{g}^{\mathsf{T}}(\boldsymbol{y}_i)\, \boldsymbol{p}_i / \boldsymbol{p}_i^{\mathsf{T}} A \boldsymbol{p}_i - \alpha_i \qquad (i = 1, \ldots, n) \tag{2.2}$$

is the accumulated error and ε_i is the round-off error of the computer actually used.

With the above notations we have

Theorem 1. *Consider the process* (1.3)–(1.4) *for the quadratical function* $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^{\mathsf{T}} A \boldsymbol{x} + \boldsymbol{b}^{\mathsf{T}} \boldsymbol{x} + c$ *and assume that the first* k *steps are performed exactly i.e.*

$$\beta_i = \alpha_i, \qquad \boldsymbol{p}_i = \boldsymbol{d}_i, \qquad \boldsymbol{y}_i = \boldsymbol{x}_i \qquad (i = 1, \ldots, k;\ k < n). \tag{2.3}$$

Then

$$\boldsymbol{y}_j = \boldsymbol{x}_{k+1} + \sum_{i=k+1}^{j-1} \beta_i \boldsymbol{p}_i \qquad (j = k+1, \ldots, n+1) \tag{2.4}$$

and for the accumulated error

$$\gamma_i = \left[\alpha_i \boldsymbol{\delta}_i^{\mathsf{T}} A \boldsymbol{\delta}_i - 2\alpha_i \boldsymbol{\delta}_i^{\mathsf{T}} A \boldsymbol{d}_i - \sum_{j=k+1}^{i-1} \beta_j \boldsymbol{\delta}_j^{\mathsf{T}} A \boldsymbol{d}_j - \boldsymbol{g}^{\mathsf{T}}(\boldsymbol{y}_i)\, \boldsymbol{\delta}_i \right] \Big/ \boldsymbol{p}_i^{\mathsf{T}} A \boldsymbol{p}_i \tag{2.5}$$

$(i = k+1, \ldots, n)$ *holds.*

Proof. Since the first relation is obvious we prove only (2.5) by induction corresponding to i. It is noted that (2.3) implies $\boldsymbol{y}_{k+1} = \boldsymbol{x}_{k+1}$. Using this fact and the relation

$$\boldsymbol{d}_i^{\mathsf{T}} A \boldsymbol{d}_i = \boldsymbol{p}_i^{\mathsf{T}} A \boldsymbol{p}_i - 2\boldsymbol{\delta}_i^{\mathsf{T}} A \boldsymbol{p}_i + \boldsymbol{\delta}_i^{\mathsf{T}} A \boldsymbol{\delta}_i \tag{2.6}$$

we have

$$\begin{aligned} -\boldsymbol{g}^{\mathsf{T}}(\boldsymbol{y}_{k+1})\, \boldsymbol{p}_{k+1} &= -\boldsymbol{g}^{\mathsf{T}}(\boldsymbol{x}_{k+1})\, (\boldsymbol{d}_{k+1} + \boldsymbol{\delta}_{k+1}) \\ &= \alpha_{k+1} \boldsymbol{d}_{k-1}^{\mathsf{T}} A \boldsymbol{d}_{k+1} - \boldsymbol{g}^{\mathsf{T}}(\boldsymbol{x}_{k+1})\, \boldsymbol{\delta}_{k+1} \\ &= \alpha_{k+1} \boldsymbol{p}_{k-1}^{\mathsf{T}} A \boldsymbol{p}_{k+1} + [\alpha_{k+1} \boldsymbol{\delta}_{k-1}^{\mathsf{T}} A \boldsymbol{\delta}_{k+1} - 2\alpha_{k+1} \boldsymbol{\delta}_{k-1}^{\mathsf{T}} A \boldsymbol{p}_{k+1} \\ &\quad - \boldsymbol{g}^{\mathsf{T}}(\boldsymbol{y}_{k+1})\, \boldsymbol{\delta}_{k+1}] \\ &= (\alpha_{k+1} + \gamma_{k+1})\, \boldsymbol{p}_{k+1}^{\mathsf{T}} A \boldsymbol{p}_{k+1} \end{aligned}$$

which gives the requested form of γ_{k+1}. Assume the truth of (2.5) for $i\,(1 \leqq i < n)$. Then we have

$$\begin{aligned} \gamma_{i+1} = -\alpha_{i+1} - [\boldsymbol{g}^{\mathsf{T}}(\boldsymbol{x}_{k+1})\, \boldsymbol{d}_{i+1} + \sum_{j=k+1}^{i} \beta_j (\boldsymbol{d}_j + \boldsymbol{\delta}_j)^{\mathsf{T}} A \boldsymbol{d}_{i+1} \\ + \boldsymbol{g}^{\mathsf{T}}(\boldsymbol{y}_{i+1})\, \boldsymbol{\delta}_{i+1}] / \boldsymbol{p}_{i-1}^{\mathsf{T}} A \boldsymbol{p}_{i+1}. \end{aligned}$$

Now using the relation

$$\boldsymbol{g}^{\mathsf{T}}(\boldsymbol{x}_{i+1})\, \boldsymbol{d}_{i+1} = [\boldsymbol{g}^{\mathsf{T}}(\boldsymbol{x}_{k+1}) + \sum_{j=k+1}^{i} \alpha_j \boldsymbol{d}_j^{\mathsf{T}} A]\, \boldsymbol{d}_{i+1} = \boldsymbol{g}^{\mathsf{T}}(\boldsymbol{x}_{k+1})\, \boldsymbol{d}_{i+1}$$

and the relation (2.6) for $i + 1$ we have

$$\begin{aligned} \gamma_{i+1} = -\alpha_{i+1} + \big[\alpha_{i+1} \boldsymbol{p}_{i-1}^{\mathsf{T}} A \boldsymbol{p}_{i+1} + \big(\alpha_{i+1} \boldsymbol{\delta}_{i-1}^{\mathsf{T}} A \boldsymbol{\delta}_{i+1} \\ - 2\alpha_{i+1} \boldsymbol{\delta}_{i-1}^{\mathsf{T}} A \boldsymbol{p}_{i+1} - \sum_{j=k+1}^{i} \beta_j \boldsymbol{\delta}_j^{\mathsf{T}} A \boldsymbol{d}_{i+1} - \boldsymbol{g}^{\mathsf{T}}(\boldsymbol{y}_{i+1})\, \boldsymbol{\delta}_{i+1}\big)\big] / \boldsymbol{p}_{i-1}^{\mathsf{T}} A \boldsymbol{p}_{i+1} \end{aligned}$$

which also implies (2.5) for γ_{i+1}, q.e.d.

3. A priori bound for the error propagation

In this section we give a bound for the error $\|\boldsymbol{y}_j - \boldsymbol{x}_j\|$ using the result of the previous section. The estimation needs a priori informations on the spectrum of A $(\lambda_{\max}, \lambda_{\min})$ and on the errors of conjugate directions actually used.

We need to prove the following lemma:

Lemma 1. *For the quadratical function* (1.1) *and the sequence* (1.3) *the inequality*

$$\|\boldsymbol{g}(\boldsymbol{x}_i)\| \leqq \|\boldsymbol{g}(\boldsymbol{x}_1)\| \sqrt{\lambda_{\max}} / \sqrt{\lambda_{\min}} \qquad (i = 1, \ldots, n+1) \tag{3.1}$$

holds.

Proof. The principal axis transformation keeps the norm of the gradient vector, so it is enough to prove (3.1) for the function of form

$$f^*(\boldsymbol{x}) = \sum_{i=1}^{n} \lambda_i x_i^2 \qquad (\lambda_1 \geqq \lambda_2 \geqq \cdots \geqq \lambda_n > 0) \tag{3.2}$$

with the gradient

$$\boldsymbol{g}^*(\boldsymbol{x}) = (2\lambda_1 x_1, \ldots, 2\lambda_n x_n)^{\mathsf{T}}. \tag{3.3}$$

consider the level set

$$S(\boldsymbol{z}) = \{\boldsymbol{x} \in R^n \mid f^*(\boldsymbol{x}) \leqq f^*(\boldsymbol{z})\} \qquad (\boldsymbol{z} \in R^n) \tag{3.4}$$

and look for $\max \left(\boldsymbol{g}^*(\boldsymbol{x})\right)^{\mathsf{T}} \boldsymbol{g}^*(\boldsymbol{x})$ in the set $S(\boldsymbol{z})$. Using the inequality $f^*(\boldsymbol{x}) \leqq K$ $\left(K = f^*(\boldsymbol{z})\right)$ we have

$$\begin{aligned}
\left(\boldsymbol{g}^*(\boldsymbol{x})\right)^{\mathsf{T}} \boldsymbol{g}^*(\boldsymbol{x}) = 4 \sum_{i=1}^{n} \lambda_i^2 x_i^2 &\leqq 4\lambda_1 \left(K - \sum_{i=2}^{n} \lambda_i x_i^2 \right) + 4 \sum_{i=2}^{n} \lambda_i^2 x_i^2 \\
&= 4\lambda_1 K + 4 \sum_{i=2}^{n} \lambda_i (\lambda_i - \lambda_1)\, x_i^2 \\
&\leqq 4\lambda_1 K \leqq 4 \frac{\lambda_1}{\lambda_n} \sum_{i=1}^{n} \lambda_i \lambda_n z_i^2 \leqq \frac{\lambda_1}{\lambda_n}\, 4 \sum_{i=1}^{n} \lambda_i^2 z_i^2 = \frac{\lambda_1}{\lambda_n} \|\boldsymbol{g}^*(\boldsymbol{z})\|^2
\end{aligned}$$

which implies

$$\|\boldsymbol{g}^*(\boldsymbol{x})\| \leqq \|\boldsymbol{g}^*(\boldsymbol{z})\| \sqrt{\lambda_{\max}} / \sqrt{\lambda_{\min}} \qquad \left(\boldsymbol{x} \in S(\boldsymbol{z})\right).$$

For the conjugate direction method (1.3)–(1.4) holds $f(\boldsymbol{x}_{i+1}) \leqq f(\boldsymbol{x}_i)$ $(i = 1, \ldots, n+1)$, so the process cannot step out of the level set $S(\boldsymbol{x}_1)$. This obviously implies (3.1). q.e.d.

It is noted that Lemma 1 also implies

$$\|\boldsymbol{g}(\boldsymbol{y}_k)\| \leqq \|\boldsymbol{g}(\boldsymbol{x}_1)\| \sqrt{\lambda_{\max}} / \sqrt{\lambda_{\min}} \qquad (k > 1) \tag{3.5}$$

under the fairly formal assumption $f(\boldsymbol{y}_k) \leqq f(\boldsymbol{x}_1)$ $(k > 1)$, which holds for $|\varepsilon_i|$, $\|\boldsymbol{\delta}_i\|$ small enough.

Another consequence of Lemma 1 is the inequality

$$|\alpha_i| \leqq \alpha = \|\boldsymbol{g}(\boldsymbol{x}_1)\| \sqrt{\lambda_{\max}} / \left(\sqrt{\lambda_{\min}}\right)^3 \tag{3.6}$$

for $i = 1, \ldots, n$.

Let $\varepsilon = \max_i |\varepsilon_i|$ be the maximal round-off error and similarly, let $\delta = \max_i \|\boldsymbol{\delta}_i\|$ be the maximal error of the conjugate directions. We assume that

$$\|\boldsymbol{p}_i\| = 1 \qquad (i = 1, \ldots, n) \tag{3.7}$$

also holds for the computed conjugate directions.

We then have

Theorem 2. *Consider the process* (1.3)—(1.4) *for the quadratical function* $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^{\mathsf{T}}A\boldsymbol{x} + \boldsymbol{b}^{\mathsf{T}}\boldsymbol{x} + c$ *and assume that the first* k *steps are performed exactly* ($\beta_i = \alpha_i$, $\boldsymbol{p}_i = \boldsymbol{d}_i$, $\boldsymbol{y}_i = \boldsymbol{x}_i$; $i = 1, \ldots, k$, $k < n$) *and*

$$f(\boldsymbol{y}_i) \leqq f(\boldsymbol{x}_1) \qquad (i = 1, \ldots, n+1) \tag{3.8}$$

is satisfied. Then

$$\|\boldsymbol{y}_j - \boldsymbol{x}_j\| \leqq (j-k-1)\left[\varepsilon + \frac{\alpha\delta}{\lambda_{\min}}\left(\delta\lambda_{\max} + 2\lambda_{\min} + \frac{j-k+2}{2}\|A\|\right)\right] \tag{3.9}$$

holds for $j \geqq k+1$.

Proof. From Lemma 1 and the relation $\boldsymbol{x}_j = \boldsymbol{x}_{k+1} + \sum_{i=k+1}^{j-1} \alpha_i \boldsymbol{d}_i$ it follows that

$$\|\boldsymbol{y}_j - \boldsymbol{x}_j\| \leqq \left\| \sum_{i=k+1}^{j-1} [\alpha_i \boldsymbol{\delta}_i + (\gamma_i + \varepsilon_i)\,\boldsymbol{p}_i] \right\| \leqq \sum_{i=k+1}^{j-1} (\alpha\delta + \varepsilon + |\gamma_i|). \tag{3.10}$$

So we need to estimate $|\gamma_i|$ using the inequalities

$$|\gamma_i| \leqq \left[\alpha\boldsymbol{\delta}_i^{\mathsf{T}}A\boldsymbol{\delta}_i + 2\alpha\delta\,\|A\| + \|\boldsymbol{g}^{\mathsf{T}}(\boldsymbol{y}_i)\|\,\delta + \sum_{j=k+1}^{i-1} |\beta_j|\,\delta\,\|A\|\right] / \boldsymbol{p}_i^{\mathsf{T}}A\boldsymbol{p}_i$$

and

$$|\beta_i| = |\boldsymbol{g}^{\mathsf{T}}(\boldsymbol{y}_i)\,\boldsymbol{p}_i / \boldsymbol{p}_i^{\mathsf{T}}A\boldsymbol{p}_i| \leqq \|\boldsymbol{g}(\boldsymbol{x}_1)\|\,\sqrt{\lambda_{\max}} / \left(\sqrt{\lambda_{\min}}\right)^3$$

as well as the inequality

$$\lambda_{\min}\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x} \leqq \boldsymbol{x}^{\mathsf{T}}A\boldsymbol{x} \leqq \lambda_{\max}\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x}. \tag{3.11}$$

Thus we get the bound

$$|\gamma_i| \leqq \alpha\delta[\delta\lambda_{\max} + \lambda_{\min} + (i-k+1)\,\|A\|]/\lambda_{\min}. \tag{3.12}$$

Hence, a simple calculation yields the bound (3.9), q.e.d.

This result means that the error propagation is linear in the term of the maximal error of the conjugate directions. The bound can be decreased if it is possible to choose the conjugate directions such that

$$\lambda_{\min} \leqq \boldsymbol{d}_i^{\mathsf{T}}A\boldsymbol{d}_i < \boldsymbol{d}_{i+1}^{\mathsf{T}}A\boldsymbol{d}_{i+1} \qquad (i = 1, \ldots, n-1). \tag{3.13}$$

However a considerable improvement can be obtained only from the suitable choice of the initial value $\boldsymbol{x}_1$ according to the condition

$$\|\boldsymbol{g}(\boldsymbol{x}_1)\| = \|A\boldsymbol{x}_1 + \boldsymbol{b}\| \leqq \|\boldsymbol{b}\|. \tag{3.14}$$

At last we remark that using Gerschgorin type estimates for the spectrum of A the dependence on $\lambda_{\max}$ and $\lambda_{\min}$ can be omitted in the bound (3.9). The simple estimate $\lambda_{\max} \leqq \|A\|$ yields the independence of $\lambda_{\max}$.

References

[1] ABAFFY, J., A new three parameter class of Quasi-Newton methods, Alkalmazott Matematikai Lapok (to be published).
[2] ABAFFY, J., and F. SLOBODA, General algorithms for generating conjugate directions and their applications, in: Algoritmy vo vypoctovej technike, Slovenska vedechotechnisa spolicnost (1975), p. 97—108.
[3] FLETCHER, R., and C. M. REEVES, Function minimisation by conjugate gradients, Computer J. 7 (1964), 149.
[4] SACHET, M., and S. KAHNE, Error analysis in conjugate direction methods (to be published).
[5] WALSH, G. R., Methods of optimization, John Wiley, London 1975.
[6] PSENICNY, B. N., and YU. N. DANILIN, Numerical methods of extremal problems (Russian), Nauka, Moscow 1975.

Manuskripteingang: 14. 2. 1977

VERFASSER:

Dr. JÓZSEF ABAFFY, Automation and Computer Institute of the Hungarian Academy of Sciences, Budapest
Dr. AURÉL GALÁNTAI, Department of Numerical Methods and Computer Science of the Eötvös Loránd University Budapest

Beiträge zur Numerischen Mathematik
7 (1979), 13–18

Eine elementare Realisierung des lexikographischen Simplexverfahrens

Horst Bialy

1. Einleitung

Es gibt heute mehrere Zusatzmaßnahmen, die die Endlichkeit des Simplexverfahrens erzwingen. Am bekanntesten sind die Störungsmethode von Charnes [1] und die Ausnutzung der Vektorlexikographie von Dantzig, Orden und Wolfe [2]. Sinn beider Methoden ist die Verhinderung von Basiszyklen. Eine von Jerke in [3] für das duale Simplexverfahren vorgeschlagene und von Schiebel, Terno und Unger in [4] noch vereinfachte lexikographische Vorgehensweise läßt sich in analoger Weise auch für das primale Simplexverfahren angeben. Das zugehörige Pivotwahlverfahren kommt ohne wesentliche Zusatzrechnung aus, ist numerisch einfacher als die Methode [5] und kann bedenkenlos auch in einer Einführungsvorlesung vorgetragen werden.

Wir bezeichnen Vektoren bzw. Matrizen mit halbfetten Klein- bzw. Großbuchstaben. Hochgestelltes $\mathbf{T}$ ist das Transponierungszeichen.

Es sei

$$\begin{aligned} z &= d_0 + \boldsymbol{d}^{\mathbf{T}} \boldsymbol{x}_N, \\ \boldsymbol{x}_B &= \boldsymbol{b} + \boldsymbol{B}\boldsymbol{x}_N \qquad \text{mit } \boldsymbol{b} \geqq \boldsymbol{o} \end{aligned} \tag{1}$$

ein Simplextableau eines in Normalform gebrachten linearen Optimierungsproblems. Die Komponenten der Vektoren $\boldsymbol{x}_B$ bzw. $\boldsymbol{x}_N$ sind die Basis- bzw. Nichtbasisvariablen. Die durch die Nichtbasisvariablen dargestellte Zielfunktion z sei zu minimieren. Für die numerische Rechnung benutzt man (1) in der Tabellenform

	1	$\boldsymbol{x}_N{}^{\mathbf{T}}$
z	d_0	$\boldsymbol{d}^{\mathbf{T}}$
$\boldsymbol{x}_B$	$\boldsymbol{b}$	$\boldsymbol{B}$

(2)

Die Zeilen werden von 0 bis m, die Spalten von 0 bis n numeriert. $Z = \{0, 1, \ldots, m\}$ ist die Menge der Zeilenindizes, $S = \{0, 1, \ldots, n\}$ die Menge der Spaltenindizes. Im Sinne einheitlicher Transformationsformeln (3) ist die Verwendung der Tabellenmatrix

$$\boldsymbol{A} = \begin{pmatrix} d_0 & \boldsymbol{d}^{\mathbf{T}} \\ \boldsymbol{b} & \boldsymbol{B} \end{pmatrix} = (a_{ik}), \qquad i = 0(1)m, \qquad k = 0(1)n,$$

vorteilhaft. Ein Basiswechsel mit dem Pivot $p := a_{st}$, also der Austausch der Basisvariablen in der Zeile $s \in Z \setminus \{0\}$ gegen die Nichtbasisvariable in der Spalte $t \in S \setminus \{0\}$,

erfolgt bekanntlich nach den Formeln

$$\begin{aligned} a^*_{sl} &= 1/p, \\ a^*_{sk} &= a_{sk}/(-p), && k \in S \setminus \{t\}, \\ a^*_{il} &= a_{il}/p, && i \in Z \setminus \{s\}, \\ a^*_{ik} &= a_{ik} + a_{il}a^*_{sk}, && i \in Z \setminus \{s\}, \quad k \in S \setminus \{t\}. \end{aligned} \tag{3}$$

Hier und im folgenden sind die Koeffizienten der neu entstehenden Tabelle durch einen Stern gekennzeichnet.

2. Das lexikographische Simplexverfahren

Definition 1. Eine Zeile der Tabelle (2) heißt *l-positiv* (*linkspositiv*), wenn das erste von Null verschiedene Element dieser Zeile positiv ist. Die Tabelle (2) heißt *l-positiv*, wenn die Zeilen mit den Nummern 1 bis m l-positiv sind.

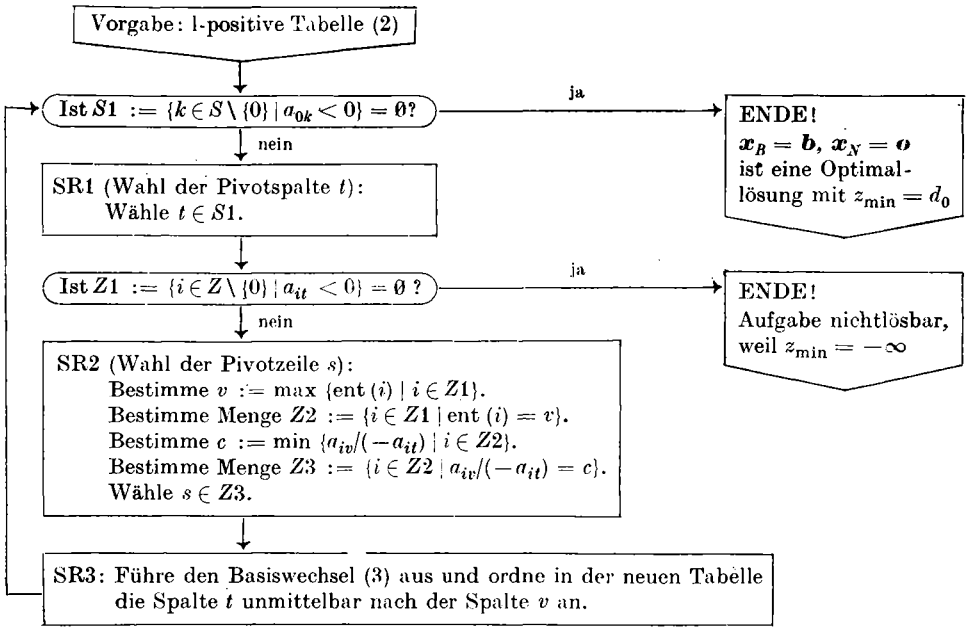
Definition 2. Tabelle (2) sei l-positiv. Dann heißt die Zahl

$$\text{ent}(i) := \min \{k \mid k \in S \wedge a_{ik} > 0\}$$

Entartungszahl der Zeile $i \in Z \setminus \{0\}$.

Die Entartungszahl gibt diejenige Stelle einer Zeile an, in der das erste von Null verschiedene Element steht.

Definition 3. *Das lexikographische Simplexverfahren* (LSV)



Die Zahl v heißt *Entartungsstufe des Basiswechsels.*

Satz 1. *Das LSV ist stets ausführbar. Nach jedem Basiswechsel entsteht wieder eine l-positive Tabelle, und es gilt* $a^*_{0v} < a_{0v}$, *wobei* v *die Entartungsstufe des Basiswechsels ist.*

Beweis. Die Ausführbarkeit der Simplexschritte SR1 bis SR3 ist klar, da $Z2$ und $Z3$ keine leeren Mengen sind, wenn $Z1$ nicht die leere Menge ist.

Wegen $Z3 \subset Z1$ gilt $p := a_{st} < 0$. Wegen $a_{sv} > 0$ ist sicher $v < t$. Die Spalten mit den Nummern $k = 0, 1, \ldots, v-1$ (das sind die ersten v Spalten!) ändern sich nicht.

Es gilt $a^*_{sv} = a_{sv}/(-p) > 0$. Außerdem ist $a^*_{sv} = c > 0$.

Für eine beliebige Zeile $i \in Z \setminus \{0, s\}$ mit der Entartungszahl $w := \mathrm{ent}(i)$ sind folgende Fälle denkbar:

F 1: $w < v$. Es folgt $a^*_{iw} = a_{iw} > 0$.

F 2.1: $w = v \wedge a_{it} \geqq 0$. Es gilt $a^*_{iw} = a^*_{iv} = a_{iv} + a_{it}a^*_{sv} \geqq a_{iv} > 0$.

F 2.2: $w = v \wedge a_{it} < 0$. Wegen $i \in Z2$ gilt

$$a^*_{iw} = a^*_{iv} = a_{iv} + a_{it}(c) \geqq a_{iv} + a_{it}\big(a_{iv}/(-a_{it})\big) = 0.$$

F 3.1: $w > v \wedge a_{it} > 0$. Dann ist $a^*_{iv} = 0 + c \cdot a_{it} > 0$ das erste von Null verschiedene Element in der neuen Zeile i.

F 3.2: $w > v \wedge a_{it} = 0$. Im Fall $t < w$ verändert sich die Zeile i nicht. Der Fall $t = w$ ist nicht möglich. Im Fall $t > w$ erfolgt lediglich die Umstellung einer Null aus Spalte t nach Spalte $v + 1$. In der neuen Tabelle gilt dann $\mathrm{ent}(i) = w + 1$.

F 3.3: $w > v \wedge a_{it} < 0$. Dieser Fall ist wegen $i \in Z1$ und der Definition von v nicht möglich.

Die im Simplexschritt SR3 vorzunehmende Spaltenvertauschung ist ausschließlich wegen der in F 2.2 nicht vermeidbaren Beziehung $a^*_{iw} = 0$ erforderlich. Aus der Fallunterscheidung folgt die l-Positivität der neuen Tabelle.

Die Ungleichung $a^*_{0v} = a_{0v} + c \cdot a_{0t} < a_{0v}$ ist wegen $a_{0t} < 0$ trivial.

Satz 2. *Eine Wiederkehr der gleichen Basis ist im LSV unmöglich. Daher ist das LSV endlich, d. h., nach endlich vielen Ausführungen der Simplexregeln SR1 bis SR3 bricht das Verfahren ab.*

Beweis. Wir nehmen an, daß nach r aufeinanderfolgenden Basiswechseln wieder die gleiche Basis auftritt. $v_1, v_2, \ldots, v_r$ seien die zu den Basiswechseln gehörigen Entartungsstufen und $v := \min(v_1, v_2, \ldots, v_r)$. Dann sind die ersten $v + 1$ Spalten während der Basiswechsel sicher nicht umgeordnet worden. Es sei x_u die zur Spalte v gehörende Nichtbasisvariable. Dann ist der z-Koeffizient von x_u nach Satz 1 kleiner geworden. Die Darstellung der Zielfunktion durch die Nichtbasisvariablen einer Basis ist aber eindeutig, unsere Annahme somit falsch. Im Verlauf des LSV ist also die Wiederkehr der gleichen Basis nicht möglich. Da jede Aufgabe nur endlich viele verschiedene Basen besitzt, muß das LSV nach endlich vielen Basiswechseln abbrechen.

3. Das Aufsuchen einer l-positiven Tabelle

Das LSV verlangt eine l-positive Ausgangstabelle, die stets auch ein Simplextableau ist. Aber die Umkehrung gilt nicht. Wir schlagen folgenden einfachen *Weg zur Aufstellung einer l-positiven Tabelle* vor.

1. *Bestimme ein Simplextableau* (2). Hierfür sind zahlreiche Methoden bekannt, die an dieser Stelle nicht erörtert werden müssen.

2. Ist dieses Simplextableau nicht l-positiv, so gehe man zur *gestörten Tabelle*

	$\mathbf{1}$	ε	$\boldsymbol{x}_N^{\mathsf{T}}$
z	d_0	M	$\boldsymbol{d}^{\mathsf{T}}$
$\boldsymbol{x}_B$	$\boldsymbol{b}$	$\boldsymbol{e}$	$\boldsymbol{B}$

$$\text{mit } \boldsymbol{e} = \begin{pmatrix} 1 \\ 1 \\ \cdots \\ 1 \end{pmatrix} \in R^m \tag{4}$$

über, die sich durch eine zusätzliche positive Spalte mit der Nichtbasisvariablen ε auszeichnet. Dabei ist $M > 0$ eine hinreichend große Zahl, die garantiert, daß die Zusatzspalte niemals Pivotspalte gemäß der Simplexregel SR 1 im LSV wird.

Da ε im Verlauf des LSV niemals Basisvariable wird, besitzen Originalproblem und gestörtes Problem die gleichen Lösungseigenschaften. Speziell ist die im LSV erzeugte Optimallösung des gestörten Problems wegen $\varepsilon = 0$ gleichzeitig Optimallösung des ungestörten Problems. Bei der numerischen Rechnung muß man M weder eintragen noch nach (3) umrechnen. Die Tabellen (2) und (4) sind äquivalent.

4. Ein Zahlenbeispiel

Das folgende Beispiel dient lediglich zur Demonstration der Regeln des LSV, beansprucht also kein Interesse als Optimierungsproblem. Die Tabellen sind links oben numeriert und die Pivots durch $\boxed{}$ gekennzeichnet. Die erste Tabelle T 1 ist bereits l-positiv. Die im LSV benötigten Daten sind ebenfalls angegeben.

T 1	1	x_1	x_2	x_3	x_4	x_5	x_6	x_7
z	2	1	0	4	−2	−2	1	1
x_8	0	0	0	2	−1	$\boxed{-1}$	0	1
x_9	3	2	1	1	0	−1	−1	−1
x_{10}	0	0	0	3	1	−1	−1	2
x_{11}	0	0	0	4	−3	−2	1	0
x_{12}	0	0	0	0	1	0	−2	3
x_{13}	0	0	0	0	0	0	1	−1

$S1 = \{4, 5\}$. Wir wählen $t = 5$. $Z1 = \{1, 2, 3, 4\}$, $v = 3$, $Z2 = \{1, 3, 4\}$, $c = 2$, $Z3 = \{1, 4\}$. Wir wählen $s = 1$. Nach dem Austausch von x_8 gegen x_5 ist die Spalte mit x_8 unmittelbar nach der Spalte mit x_3 anzuordnen. Andernfalls wäre die Zeile mit x_{11} in T 2 nicht mehr l-positiv (Beweisfall F 2.2 zu Satz 1). In den Zeilen mit x_{12} und x_{13} liegt der Fall 3.2 des Beweises zu Satz 1 vor.

T 2	1	x_1	x_2	x_3	x_8	x_4	x_6	x_7
z	2	1	0	0	2	0	1	−1
x_5	0	0	0	2	−1	−1	0	1
x_9	3	2	1	−1	1	1	−1	−2
x_{10}	0	0	0	1	1	2	−1	1
x_{11}	0	0	0	0	2	−1	1	−2
x_{12}	0	0	0	0	0	1	−2	3
x_{13}	0	0	0	0	0	0	1	**−1**

$S1 = \{7\}$. Also wird $t = 7$. $Z1 = \{2, 4, 6\}$, $v = 6$, $Z2 = Z3 = \{6\}$. Also wird $s = 6$. Da die Pivotspalte unmittelbar auf die Spalte mit der Nummer $v = 6$ folgt, unterbleibt ein Spaltenaustausch beim Übergang von T2 nach T3.

T 3	1	x_1	x_2	x_3	x_8	x_4	x_6	x_{13}
z	2	1	0	0	2	0	0	1
x_5	0	0	0	2	−1	−1	1	−1
x_9	3	2	1	−1	1	1	−3	2
x_{10}	0	0	0	1	1	2	0	−1
x_{11}	0	0	0	0	2	−1	−1	2
x_{12}	0	0	0	0	0	1	1	−3
x_7	0	0	0	0	0	0	1	−1

Wegen $S1 = \emptyset$ ist das LSV mit T3 beendet. Mit Ausnahme von $x_9 = 3$ sind sämtliche Komponenten der Optimallösung Null. Es gilt $z_{\min} = 2$.

Literatur

[1] Charnes, A., Optimality and degeneracy in linear programming, Econometrica **20** (1952), 160—170.

[2] Dantzig, G. B., A. Orden and P. Wolfe, The generalized simplex method for minimizing a linear form under linear inequality restraints, Pacific J. Math. **5** (1955), 183—195.

[3] Jerke, W., Duale Schnittverfahren der ganzzahligen linearen Optimierung unter besonderer Berücksichtigung der Entartung, Dissertation, TU Dresden 1974.

[4] Jerke, W., W. Schiebel, J. Terno und G. Unger, Ein Beitrag zur Behandlung der Entartung beim Simplexverfahren, Beitr. Numer. Math. **4** (1975), 105—114.

[5] Bialy, H., Eine elementare Methode zur Behandlung des Entartungsfalles in der linearen Optimierung, Unternehmensforschung **10** (1966), 118—123.

Zusammenfassung

In der vorliegenden Arbeit wird eine numerische Realisierung des lexikographischen Simplexverfahrens angegeben, die gegenüber dem gewöhnlichen Simplexverfahren nur wenig Zusatzrechnung erfordert.

Summary

This paper presents a numerical realization of the lexicographic simplex method, which requires only a few supplement calculationwork compared to the usual simplex method.

Résumé

Dans l'article présent on donne une procédure numérique pour l'algorithme du simplexe lexicographique, quelle en comparaison de l'algorithme du simplex usuel demande peu de calcul supplémentaire.

Резюме

В данной работе указывают численную реализацию лексикографического симплекс-метода, требующая по сравнению с обычным симплекс-методом относительно мало добавочных вычислений.

Manuskripteingang: 15. 10. 1976

VERFASSER:

Dr. HORST BIALY, Sektion Mathematik, Naturwissenschaften und Rechentechnik
der Hochschule für Verkehrswesen „Friedrich List" Dresden

Beiträge zur Numerischen Mathematik
7 (1979), 19–31

Die Anwendung der Iterierten Defektkorrektur auf das Dirichletproblem

Reinhard Frank und Jörg Hertling

1. Darstellung der Methode

In [1] und [3] wurde die Methode der Iterierten Defektkorrektur entwickelt und deren Anwendung auf Anfangswertprobleme bei gewöhnlichen Differentialgleichungen und in [2] die Anwendung auf Zweipunkt-Randwertprobleme diskutiert. Die dort angegebenen Algorithmen, die sich in numerischen Experimenten als außerordentlich effektiv erwiesen haben [6], sollen nun auf partielle Differentialgleichungen vom elliptischen Typ erweitert werden. Während für Zweipunkt-Randwertprobleme in [2] die asymptotische Theorie ($h \to 0$) entwickelt wurde, beschränken wir uns in dieser Arbeit darauf, den Algorithmus der Iterierten Defektkorrektur für gewisse elliptische Randwertaufgaben (RWA) und numerische Experimente anzugeben. Wir betrachten

$$L[u] \equiv a^{\langle 11\rangle}(x, y)\,\frac{\partial^2 u}{\partial x^2} + a^{\langle 22\rangle}(x, y)\,\frac{\partial^2 u}{\partial y^2} + a^{\langle 1\rangle}(x, y)\,\frac{\partial u}{\partial x} + a^{\langle 2\rangle}(x, y)\,\frac{\partial u}{\partial y}$$
$$= f(x, y, u) \tag{1.1}$$

auf dem Einheitsquadrat $R \equiv [0, 1] \times [0, 1]$ mit den Randbedingungen

$$\begin{aligned} &u(0, y) = r_1(y), \qquad u(1, y) = r_2(y), \qquad 0 \leqq y \leqq 1, \\ &u(x, 0) = \bar{r}_1(x), \qquad u(x, 1) = \bar{r}_2(x), \qquad 0 \leqq x \leqq 1. \end{aligned} \tag{1.2}$$

Natürlich setzen wir die Stetigkeit der Randfunktion in den Eckpunkten voraus, d. h., $r_1(0) = \bar{r}_1(0)$, $r_1(1) = \bar{r}_2(0)$, $r_2(0) = \bar{r}_1(1)$, $r_2(1) = \bar{r}_2(1)$.

Die exakte Lösung dieses Problems sei $z(x, y)$. In unserer Notation hat die Bedingung der gleichmäßigen Elliptizität die folgende Gestalt:

$$a^{\langle 11\rangle}(x, y)\,\xi^2 + a^{\langle 22\rangle}(x, y)\,\eta^2 \geqq \varkappa(\xi^2 + \eta^2), \qquad \varkappa > 0. \tag{1.3}$$

Zunächst möchten wir die Idee unserer Methode skizzieren. Dazu nehmen wir an, daß unser Problem mittels eines klassischen Differenzenschemas (Fünfpunkt-Sterne) gelöst wird. Dadurch erhält man auf dem Punktgitter

$$\mathbf{G}_h := \{(x_\nu, y_\mu);\ x_\nu = \nu h,\ y_\mu = \mu h,\ h = \frac{1}{l},\ \nu, \mu = 0(1)l\} \tag{1.4}$$

die Näherungswerte $\eta_{\nu\mu}^{[0]}$ für die exakte Lösung $z(x_\nu, y_\mu)$. Nun interpolieren wir die Näherungswerte $\eta_{\nu\mu}^{[0]}$ auf $[0, 1] \times [0, 1]$ mit der Interpolationsfunktion $P_h^{[0]}(x, y)$:

$$P_h^{[0]}(x_\nu, y_\mu) = \eta_{\nu\mu}^{[0]}, \qquad \nu, \mu = 0(1)m.$$

Der Index h deutet an, daß die von den $\eta_{\nu\mu}^{[0]}$-Werten abhängige Interpolationsfunktion damit auch von der Schrittweite h abhängt. Nun konstruieren wir ein neues Randwertproblem

$$L[u] = f(x, y, u) + L[P_h^{[0]}] - f\big(x, y, P_h^{[0]}(x, y)\big), \tag{1.5}$$

$$\begin{aligned} &u(0, y) = P_h^{[0]}(0, y), \qquad u(1, y) = P_h^{[0]}(1, y), \\ &u(x, 0) = P_h^{[0]}(x, 0), \qquad u(x, 1) = P_h^{[0]}(x, 1), \end{aligned} \tag{1.6}$$

wobei in L dieselben Koeffizientenfunktionen $a^{(11)}(x, y), \ldots$ auftreten wie in (1.1). Offensichtlich hat dieses Randwertproblem $P_h^{[0]}(x, y)$ als exakte Lösung. Die Differentialgleichung (1.5) weicht von der ursprünglichen Gleichung (1.1) um die Störung

$$d_h^{[0]}(x, y) := L[P_h^{[0]}] - f\big(x, y, P_h^{[0]}(x, y)\big) \tag{1.7}$$

ab. Gilt $P_h^{[0]}(x, y) \approx z(x, y)$, $\dfrac{\partial^k}{\partial x^k} P_h^{[0]}(x, y) \approx \dfrac{\partial^k}{\partial x^k} z(x, y)$ und $\dfrac{\partial^k}{\partial y^k} P_h^{[0]}(x, y) \approx \dfrac{\partial^k}{\partial y^k} z(x, y)$ $(k = 1, 2)$, was dadurch erreicht werden kann, daß erstens in (1.4) mit hinreichend kleinem h gearbeitet wird, so daß die Größen $\eta_{\nu\mu}^{[0]}$ hinreichend gute Approximationen von $z(x_\nu, y_\mu)$ sind, und daß zweitens $P_h^{[0]}(x, y)$ aus einer geeigneten Klasse von Interpolationsfunktionen gewählt wird, dann ist die Störung $d_h^{[0]}(x, y)$ „klein". Das bedeutet, daß das Problem (1.5), (1.6) „nahe" unserem ursprünglichen Problem (1.1), (1.2) ist; daher nennen wir ab nun (1.5), (1.6) „Nachbarproblem" (NP). Entsprechend bezeichnen wir nun (1.1), (1.2) als „Originalproblem" (OP). Obwohl wir die exakte Lösung des NPs kennen, lösen wir es mit demselben Differenzenschema auf demselben Gitter. Das liefert die Näherungswerte $\pi_{\nu\mu}^{[0]}$ für $P_h^{[0]}(x_\nu, y_\mu)$. Natürlich kennen wir jetzt den globalen Diskretisierungsfehler $\pi_{\nu\mu}^{[0]} - P_h^{[0]}(x_\nu, y_\mu)$ von (1.5), (1.6). Da das NP nahe dem Originalproblem (1.1), (1.2) ist, können wir erwarten, daß $\pi_{\nu\mu}^{[0]} - P_h^{[0]}(x_\nu, y_\mu)$ eine gute Schätzung für den unbekannten Diskretisierungsfehler $\eta_{\nu\mu}^{[0]} - z(x_\nu, y_\mu)$ von (1.1), (1.2) ist. Der eben beschriebene Gedanke geht auf ZADUNAISKY zurück, der ihn für die Anwendung von Runge-Kutta-Methoden auf Anfangswertprobleme bei gewöhnlichen Differentialgleichungen eingeführt hat und ihn durch die obigen heuristischen Argumente motiviert hat [7]. Da wir eine echte Schätzung (mit richtigem Vorzeichen) und keine Abschätzung gewonnen haben, können wir diese Schätzung dazu benutzen, um unsere erste numerische Approximation auf die folgende Art zu verbessern: Wenn wir in der Identität

$$z(x_\nu, y_\mu) = \eta_{\nu\mu}^{[0]} - \big(\eta_{\nu\mu}^{[0]} - z(x_\nu, y_\mu)\big)$$

den Ausdruck $\eta_{\nu\mu}^{[0]} - z(x_\nu, y_\mu)$ durch seine Schätzung $\pi_{\nu\mu}^{[0]} - P_h^{[0]}(x_\nu, y_\mu)$ ersetzen, erhalten wir voraussichtlich bessere Approximationen

$$\eta_{\nu\mu}^{[1]} := \eta_{\nu\mu}^{[0]} - \big(\pi_{\nu\mu}^{[0]} - P_h^{[0]}(x_\nu, y_\mu)\big) \tag{1.8}$$

für $z(x_\nu, y_\mu)$.

Es ist offensichtlich, daß diese Prozedur iteriert werden kann. Natürlich benutzt man im zweiten Schritt dieser Iteration eine neue Interpolationsfunktion $P_h^{[1]}(x, y)$, die die neuen Werte $\eta_{\nu\mu}^{[1]}$ interpoliert, und erhält

$$\eta_{\nu\mu}^{[2]} := \eta_{\nu\mu}^{[0]} - \left(\pi_{\nu\mu}^{[1]} - P_h^{[1]}(x_\nu, y_\mu)\right). \tag{1.9}$$

Weitere Iterationsschritte liefern

$$\eta_{\nu\mu}^{[j]} := \eta_{\nu\mu}^{[0]} - \left(\pi_{\nu\mu}^{[j-1]} - P_h^{[j-1]}(x_\nu, y_\mu)\right), \qquad j = 3, 4, \ldots \tag{1.10}$$

Um die Iteration (1.9) auf heuristische Art zu motivieren, beachten wir:

a) Die Werte $\eta_{\nu\mu}^{[1]}$ sind (voraussichtlich) bessere Approximationen für $z(x_\nu, y_\mu)$ als $\eta_{\nu\mu}^{[0]}$.

b) Daher ist die Funktion $P_h^{[1]}(x, y)$ (bzw. ihre Ableitungen), die die „besseren" Werte $\eta_{\nu\mu}^{[1]}$ interpoliert, der exakten Lösung (bzw. ihren Ableitungen) näher.

c) Daher läßt das neue NP (das $P_h^{[1]}(x, y)$ als exakte Lösung hat) eine noch kleinere Störung

$$d_h^{[1]}(x, y) := L[P_h^{[1]}] - f\left(x, y, P_h^{[1]}(x, y)\right)$$

erwarten. (Das neue NP ist an das OP „nähergerückt".)

d) Das bedingt voraussichtlich, daß die neue Fehlerschätzung $\pi_{\nu\mu}^{[1]} - P_h^{[1]}(x_\nu, y_\mu)$ besser ist als die erste Fehlerschätzung $\pi_{\nu\mu}^{[0]} - P_h^{[0]}(x_\nu, y_\mu)$.

In verschiedenen Arbeiten über gewöhnliche Differentialgleichungen (Anfangs- und Randwertprobleme) ist diese Methode „Iterierte Defektkorrektur" genannt und ihr asymptotisches Verhalten für $h \to 0$ analysiert worden. Es stellte sich heraus, daß für eine Basismethode der Konsistenzordnung p und für Interpolationsfunktionen aus einer geeigneten Funktionenklasse der j-te Schritt der Iteration die Ordnung $(j + 1)\,p$ hatte. Bei eindimensionalen Problemen beruhte der Beweis dieser Tatsache auf der Existenz einer asymptotischen Entwicklung des globalen Diskretisierungsfehlers (nach h-Potenzen). Dieser Beweis läßt sich nicht unmittelbar auf Probleme vom Typ (1.1), (1.2) übertragen, da bei physikalisch relevanten Problemen im allgemeinen die Existenz „hinreichend langer" asymptotischer Entwicklungen nicht gesichert ist. Auch bei glatten Daten $a^{(11)}(x, y)$, $a^{(22)}(x, y)$, $a^{(1)}(x, y)$, $a^{(2)}(x, y)$, $f(x, y, u)$, $r_1(y)$, $r_2(y)$, $\bar{r}_1(x)$, $\bar{r}_2(x)$ können nämlich in den Eckpunkten Singularitäten der höheren Ableitungen von $z(x, y)$ auftreten. Die numerischen Experimente, die wir durchgeführt haben, zeigen jedoch, daß die erzielten Resultate ähnlich gut sind wie bei eindimensionalen Problemen.

Die Konstruktion der Interpolationsfunktion und die Berechnung der Störungen $d_h^{[j]}(x_\nu, y_\mu)$ erscheint aufwendig. In Abschnitt 2 werden wir jedoch sehen, daß diese Berechnung in sehr einfacher Weise möglich ist, falls mit stückweise polynomialer Interpolation gearbeitet wird.

Ein wesentlicher Vorteil unserer Methode ist der folgende: Um das Originalproblem zu lösen (Berechnung der Werte $\eta_{\nu\mu}^{[0]}$) und um das j-te NP zu lösen (Berechnung der Werte $\pi_{\nu\mu}^{[j]}$), muß man nichtlineare Gleichungssysteme lösen. Wenn die Werte $\eta_{\nu\mu}^{[j]}$ näher an $z(x_\nu, y_\mu)$ rücken und wenn gleichzeitig die Fehlerschätzungen $\pi_{\nu\mu}^{[j]} - P_h^{[j]}(x_\nu, y_\mu)$ besser mit dem tatsächlichen Diskretisierungsfehler $\eta_{\nu\mu}^{[0]} - z(x_\nu, y_\mu)$ übereinstimmen, dann rücken die Werte $\pi_{\nu\mu}^{[j]}$ näher an unsere erste numerische

Approximation $\eta_{\nu\mu}^{[0]}$ (Abb. 1). Für die numerische Lösung des j-ten NPs, d. h. für die Lösung des entsprechenden nichtlinearen Gleichungssystems, erhalten wir auf Grund unserer letzten Bemerkung die guten Startwerte $\eta_{\nu\mu}^{[0]}$.

Dies bedeutet, daß wir bei allen Schritten der Iterierten Defektkorrektur mit demselben Startvektor arbeiten. Da die Störungen $d_h^{[j]}(x, y)$ nicht von u abhängen, verschwinden sie bei der Konstruktion der Jacobimatrix, die dem j-ten NP entspricht (Differentiation von $f(x, y, u) + d_h^{[j]}(x, y)$ nach u). Falls nun mit einem Pseudonewtonverfahren gearbeitet wird (Auswertung der Jacobimatrix nur an der Startstelle), so bedeutet dies, daß während der iterierten Defektkorrektur stets mit derselben Jacobimatrix gearbeitet wird. (Das heißt, Dreieckszerlegungen treten nur bei der Lösung des OPs auf, jedoch nicht bei der Lösung der NPs!)

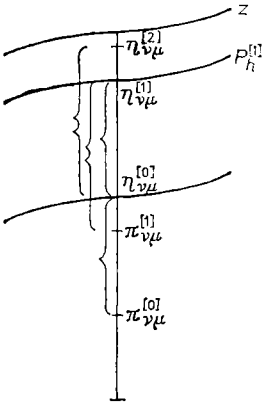


Abb. 1

Vergleichen wir unsere Methode mit einer Extrapolationsmethode, so bemerken wir, daß die Schrittweite fest bleibt und die Gleichungssysteme nicht explodieren. Das ist für mehrdimensionale Probleme ein sehr wichtiger Gesichtspunkt.

2. Algorithmische Details

Praktische Gründe, die noch später besprochen werden, legen nahe, daß für die Konstruktion unserer NPs stückweise polynomiale Interpolation in zwei Variablen sehr günstig ist. Da die NPs numerisch auf demselben Gitter gelöst werden müssen, müssen die Störungen $d_h^{[j]}(x, y)$ nur auf den Gitterpunkten berechnet werden. Daraus folgt, daß wir die Interpolationspolynome nicht explizit konstruieren müssen, sondern nur ihre ersten und zweiten partiellen Ableitungen an den Gitterpunkten.

Für die stückweise polynomiale Interpolation in zwei Variablen teilen wir das Einheitsquadrat in Teilquadrate (Abb. 2). Die Eckpunkte dieser Teilquadrate bilden das Gitter

$$\mathbf{G}_H := \left\{(x^i, y^k);\ x^i = iH,\ y^k = kH,\ H = \frac{1}{n},\ i, k = 0(1)n\right\} \tag{2.1}$$