Aplikace matematiky

Jiří Neuberg

# Some limit properties of the best determined terms method

*Aplikace matematiky*, Vol. 21 (1976), No. 3, 161--167

Persistent URL: http://dml.cz/dmlcz/103635

**Terms of use:**

© Institute of Mathematics AS CR, 1976

Institute of Mathematics of the Academy of Sciences of the Czech Republic provides access to digitized documents strictly for personal use. Each copy of any part of this document must contain these *Terms of use*.


This paper has been digitized, optimized for electronic delivery and stamped with digital signature within the project *DML-CZ: The Czech Digital Mathematics Library* http://project.dml.cz

# SOME LIMIT PROPERTIES OF THE BEST DETERMINED TERMS METHOD

JIŘÍ NEUBERG

(Received February 28, 1974)

A very effective method for solving Fredholm integral equations of the first kind is so called method of the best determined terms. It was proposed by R. J. Hanson [2] and further ellaborated for a certain type of problems by J. M. Varah [3]. The effectivity of the method in a natural way depends on the criteria for the choice for the approximate solutions. However appropriate criteria are still needed for applicability of the method. In this contribution we propose a general criterion for the choice of approximate solutions, analyse its properties and present some estimates of the errors for the corresponding approximate solutions. Some other criteria for the choice of approximate solutions are presented in [4,] whereina stochastic approach is used.

## 1.

Let us consider the following linear system

$$Kf = g\,, \tag{1.1}$$

where $f \in R^n$, $g \in R^m$ and $K$ is a fixed $m \times n$ matrix with real elements. In practical calculations it is quite standard that right hand side in (1.1) is not known exactly; we are given a vector $g + \delta g$, such that $\|\delta g\| \leqq \Delta$, where $\Delta$ is an a priori given bound and $\|\cdot\|$ is a suitable norm in $R^m$ .We also assume that the matrix $K$ is non well conditioned, it may be singular if $m = n$.

According to [1, p. 5] there exist unitary matrices $U$ and $V$ and a "diagonal" matrix $D = (d_{jk})$ such that

$$K = U^T D V\,,$$

where $d_{jk} = 0$ if $j \neq k$ and $d_{jj} = \sigma_j \geqq 0$, $j = 1, \ldots, r = \min(m, n)$. With no loss in generality we assume that

$$\sigma_1 \geqq \ldots \geqq \sigma_r\,.$$

Let us set

$$Vf = u\,, \quad Ug = \varphi\,.$$

Then the system (1.1) can be written as

$$Du = \varphi\,. \tag{1.2}$$

## 2.

First, we study the system (1.2). Let us denote by $\mathfrak{M}$ the set of all solutions of this system. Let $\mathfrak{M}$ be non empty $\mathfrak{M} \neq \emptyset$. It is easy to see that $\mathfrak{M}$ is closed and convex. This justifies the correctness of the following

**Definition.** A vector $u_0 \in \mathfrak{M}$ is called *normal solution to* (1.2) if

$$\|u_0\| = \min\{\|u\| : u \in \mathfrak{M}\}\,.$$

Remark. If the norm in $R^n$ is uniformly convex, then the normal solution is determined uniquely. This is the case of the euclidean norm.

In what follows the norm $\|\cdot\|$ means the euclidean norm in both spaces $R^n$ and $R^m$.

An explite form of the normal solution is given by

$$u_0 = D^+\varphi\,,$$

where $D^+$ denotes the generalized inverse matrix to $D$ (see [6, p. 1]).

Let $\Delta > 0$ be a fixed positive number and let $\delta g \in R^m$ be such that $\|\delta g\| \leqq \Delta$. We consider now the system

$$Kf = g + \delta g\,,$$

or else

$$Du = \varphi + \delta\varphi\,, \tag{2.1}$$

where $\delta\varphi = U\,\delta g$.

Defining

$$\sigma_j^+ = \begin{cases} 0 & \text{if} \quad \sigma_j = 0 \\ \dfrac{1}{\sigma_j} & \text{if} \quad \sigma_j > 0 \end{cases}$$

and

$$u_i^{(k)} = \sum_{j=1}^{k} \sigma_j^+ (\varphi_j + \delta\varphi_j)\,\delta_{ij}$$

where $\varphi = (\varphi_1, \ldots, \varphi_n)$, $\delta\varphi = (\delta\varphi_1, \ldots, \delta\varphi_n)$,

we let

$$\mathfrak{N} = \{u^{(k)} : k = 1, \ldots, r\},$$

where

$$u^{(k)} = (u_1^{(k)}, \ldots, u_n^{(k)}).$$

## 3.

Our aim is to show some limit properties and error estimates for $\|u_0 - u^{(k)}\|$, $k = 1, \ldots, r$.

Since

$$u_0 = \sum_{j=1}^{r} \sigma_j^+ \varphi_j e_j$$

and

$$u^{(k)} = \sum_{j=1}^{k} \sigma_j^+ (\varphi_j + \delta\varphi_j)\, e_j,$$

where $e_j = (0, \ldots, 1, \ldots, 0)$, we easily derive that

$$\|u_0 - u^{(k)}\|^2 = \sum_{j=1}^{k} (\sigma_j^+ \delta\varphi_j)^2 + \sum_{j=k+1}^{r} (\sigma_j^+ \varphi_j)^2 . \tag{3.1}$$

**Lemma 1.** *Let $T > 0$ and $\Delta \in (0, T]$.*

*Then $\|u_0 - u^{(s)}\|^2 \leqq \Delta^2 \sum_{j=1}^{s} (\sigma_j^+)^2$ and, consequently,*

$$\lim_{\Delta \to 0} u^{(s)} = u_0 ,$$

*where $s$ is such that $\varphi_s \neq 0$, while $\varphi_{s+1} = \ldots = \varphi_m = 0$.*

Proof. Since $U$ is unitary, we have that $\|\delta\varphi\| = \|\delta g\| \leqq \Delta$.
Then (3.1) implies that

$$\|u_0 - u^{(s)}\|^2 = \sum_{j=1}^{s} (\sigma_j^+ \delta\varphi_j)^2 \leqq \Delta^2 \sum_{j=1}^{s} (\sigma_j^+)^2 .$$

One, however, meets the situation, where

$$\min \{\|u_0 - u^{(k)}\| : k = 1, \ldots, r\} = \|u - u^{(l)}\| ,$$

with some $l$ essentially smaller than $r$, as usual.

Because of our assumption concerning the sequence $\sigma_j$ the first summand in the right hand side in (3.1) is an nondecreasing function of $k$ and the second one – a nonincreasing function of $k$ respectively. Another weakness of the approximations

$u^{(s)}$ is that this approximation is unadequate if the admissible error bound is relatively large. Thus, some other way has to be found how to get more suitable approximation to the normal solution.

Let $\varrho(k) = \|u_0 - u^{(k)}\|$, $k = 1, \ldots, r$. As we already mentioned, this function assumes its minimal value at certain positive integer $l$, the determining of it is a difficult problem, because the vectors $\varphi$ and $\delta\varphi$ are unknown in general. To avoiding this difficulty we proceed as follows.

We define $\gamma$ as a function of a discrete variable $k = 1, \ldots, r + 1$ as

$$\gamma(k) = \|Du^{(k)} - (\varphi + \delta\varphi)\|$$

if $k = 1, \ldots, r$ and

$$\gamma(r + 1) = \|Du_0 - (\varphi + \delta\varphi)\| .$$

**Lemma 2.** *The function $\gamma$ is a nonincreasing function for $k = 1, \ldots, r$ and $\gamma(r + 1) = \|\delta\varphi\|$.*

Proof. Let $1 \leqq k \leqq r - 1$. Then

$$\begin{aligned} \gamma(k + 1) &= \|Du^{(k+1)} - \varphi - \delta\varphi\| = \\ &= \|DD^{(k+1)}(\varphi + \delta\varphi) - (\varphi + \delta\varphi)\| \leqq \\ &\leqq \|DD^{(k)}(\varphi + \delta\varphi) - (\varphi + \delta\varphi)\| = \gamma(k) , \end{aligned}$$

where the $n \times m$ matrix $D^{(k)} = (d_{jt}^{(k)})$ defined as follows:

$$d_{jt}^{(k)} = 0 \quad \text{for} \quad j \neq t \text{ and } j = t > k$$

and

$$d_{jj}^{(k)} = \sigma_j^+ \quad \text{for} \quad j = 1, \ldots, k .$$

The remaining part of the lemma is an immediate consequence of the relation $Du_0 = \varphi$. The lemma is completely prooved.

Since an upper bound for $\|\delta\varphi\|$ is available ($\|\delta\varphi\| \leqq \Delta$) and since the equality $\gamma(r + 1) = \Delta$ cannot be excluded in general, it is quite suitable to choose an index $k(\Delta) \in \{1, \ldots, r\}$ such that $\gamma(k(\Delta)) \leqq \Delta$ and such that $\gamma(p) \leqq \Delta$, $p \neq k(\Delta)$, $p \in \{1, \ldots, r\}$, implies that $p > k(\Delta)$.

**Definition.** The vector $u^{(k(\Delta))}$, where $k(\Delta)$ is defined above, is called *the solution of the first kind* with respect to the system (2.1).

An important property of the solution of the first kind is described in the following.

**Theorem 1.** *Let $u_0$ be the normal solution to (1.2) and $u^{(k(\Delta))}$ the solution of the first kind to (2.1), then the following limit relation holds:*

$$\lim_{\Delta \to 0+} u^{(k(\Delta))} = u_0 .$$

Proof. Obviously, $\gamma(k(\Delta)) = \|Du^{(k(\Delta))} - \varphi - \delta\varphi\|$ and thus,

$$\|Du^{(k(\Delta))} - Du_0\| = \|(Du^{(k(\Delta))} - \varphi - \delta\varphi) - (Du_0 - \varphi - \delta\varphi)\| \leqq$$

$$\leqq \Delta + \|\delta\varphi\| \leqq 2\Delta .$$

Further,

$$Du^{(k(\Delta))} - Du_0 = (I^{(k(\Delta))} - I)\,\varphi + I^{(k(\Delta))}\,\delta\varphi ,$$

where the elements $i_{jt}$ of the $m \times m$ matrix $I^{(k\Delta())}$ are defined as follows:

$$i_{jt} = 0 \quad \text{for} \quad j \neq t, j, t = 1, \ldots, m, \quad \text{and} \quad j = t > k(\Delta) ,$$

and

$$i_{jj} = 1 \quad \text{for} \quad j = 1, \ldots, k(\Delta) .$$

It follows that

$$\|(I - I^{(k(\Delta))})\,\varphi\| \leqq \|Du^{(k(\Delta))} - Du_0\| + \|I^{(k(\Delta))}\delta\varphi\| \leqq 3\Delta . \tag{3.2}$$

Let us set $k_* = \liminf_{\Delta \to 0+} k(\Delta)$, and $k^* = \limsup_{\Delta \to 0+} k(\Delta)$. Let $s \in \{1, \ldots, r\}$ fulfil the conditions $\varphi_s \neq 0$, and $\varphi_{s+1} = \ldots = \varphi_m = 0$. Then we have that

$$k_* = s = k^* . \tag{3.3}$$

To prove (3.3) we assume that the contrary holds. Thus, let $k_* \neq s$. As first, let $k_* < s$. The quantity $R(\delta) = \inf\{k(\Delta) : 0 \leqq \Delta \leqq \delta\}$ has the property that

$$\|(I - I^{(R(\delta))})\,\varphi\| \leqq 3\delta ,$$

the last result being implied by (3.2). Thus,

$$\lim_{\delta \to 0+} \|(I - I^{(R(\delta))})\,\varphi\| = 0 .$$

Since, obviously, $\lim_{\delta \to 0+} R(\delta) = k_*$ and according to our assumption

$$0 < \|\varphi_s\| \leqq \|(I - I^{(k_*)})\,\varphi\| ,$$

we obtain a contradiction. Thus, $k_* \geqq s$. Further, it follows that

$$\lim_{\Delta \to 0+} \|Du^{(s)} - \varphi\| = \lim_{\Delta \to 0+} \|D[D^{(s)}\varphi] - \varphi\| = 0$$

because of the solvability of the system. Consequently, $k_* \leqq s$, and finally $k_* = s$. In the same manner, one can show the validity of the relation $k^* = s$. It follows that there exist the limit $\lim_{\Delta \to 0+} k(\Delta)$ and

$$\lim_{\Delta \to 0+} k(\Delta) = s .$$

This implies that

$$\lim_{\Delta \to 0+} u^{(k(\Delta))} = \lim_{\Delta \to 0+} u^{(s)}$$

and according to lemma 1, the validity of the assertion in our Theorem. This completes the proof.

We shall derive some error estimates for the solution of the first kind.

**Theorem 2.** *For the normal solution $u_0$ and the solution of the first kind $u^{(k(\Delta))}$ the following relations hold*:

$$\|u^{(k(\Delta))} - u_0\|^2 \leqq \max\{\varphi_j^2 : j = 1, \ldots, m\} \sum_{j=k(\Delta)+1}^{r} (\sigma_j^+)^2 + \\ + \Delta^2 \sum_{j=1}^{k(\Delta)} (\sigma_j^+)^2 .$$

Proof. The validity of the relation shown is a consequence of (3.1).

This estimate is worth while whenever the data concerning the vector $\varphi$ are availables. In the opposite case we have to accept the following less satisfactory result.

**Theorem 3.** *With the same notation as in Theorem* 2 *we have*

$$\|u_0 - u^{(k(\Delta))}\| \leqq \Big(\sum_{j=k(\Delta)+1}^{r} (\sigma_j^+)^2 (\varphi_j + \delta\varphi_j)^2\Big)^{1/2} + \\ + \Delta\Big(\sum_{j=1}^{r} (\sigma_j^+)^2\Big)^{1/2} . \tag{3.4}$$

Proof. We easily verify that

$$\|u^{(k(\Delta))} - u_0\| = \Big\|u^{(k(\Delta))} - \sum_{j=1}^{r} \sigma_j^+ \delta\varphi_j e_j - u_0 + \sum_{j=1}^{r} \sigma_j^+ \delta\varphi_j e_j\Big\| \leqq \\ \leqq \Big\| \sum_{j=k(\Delta)+1}^{r} \sigma_j^+ (\varphi_j + \delta\varphi_j) e_j\Big\| + \Big\| \sum_{j=1}^{r} \sigma_j^+ \delta\varphi_j e_j\Big\| ,$$

and thus, (3.4).

Remark. Since $f = V^T u$, where $V$ is a unitary matrix, similar results as those shown for the systems (1.2) and (2.1), are valid for systems (1.1) and

$$Kf = g + \delta g \tag{3.5}$$

as well, it one modifies the definitions of the concept of a normal solution and a solution of the first kind to (1.1) and (3.5).

The proposed criterion has already been tested on some examples of certain inverse problems of the Spectroscopic diagnostics of thin plane sources. The results will be published elsewhere, see [5].

*References*

[1] *G. E. Forsythe, C. B. Moler:* Computer Solution of Linear Algebraic Systems, Prentice Hall, Englewood Clifs, New Jersey 1967.

[2] *R. J. Hanson:* A numerical method for solving Fredholm integral equations of the first kind using singular values, SIAM J. Numer. Anal., Vol. *8* (1970), 616–622.

[3] *J. M. Varah:* On the numerical solution of ill-conditioned linear systems with applications to ill-posed problems, SIAM J. Numer. Anal., Vol. *10* (1973), 257–267.

[4] *J. Cífka:* The method of the best determined terms, to appear.

[5] *J. Hekela:* Inverse pomocí metody nejlépe určených termů, to appear in Bull. Astr. Inst. ČSAV.

[6] *T. L. Boullion, P. L. Odell:* Generalised Inverse Matrices, John Wiley and Sons, London, 1971.

## Souhrn

## NĚKTERÉ LIMITNÍ VLASTNOSTI METODY NEJLÉPE URČENÝCH TERMŮ

JIŘÍ NEUBERG

Tento článek se zabývá vlastnostmi jednoho výběrového kritéria pro metodu nejlépe určených termů (BDT). Řešení úlohy $Kx = y + \varepsilon$, kde $K$ je matice $m \times n$ (špatně podmíněná), $x \in R^n$, $y, \varepsilon \in R^m$, přičemž $\sum_{i=1}^{m} \varepsilon_i^2 \leqq \Delta^2$, kde $\Delta > 0$ je daná konstanta, jest obtížné. Metodou BDT stanovíme posloupnost vektorů $x^{(1)}, \ldots, x^{(\min(m,n))}$, z nichž pak následujícím výběrovým kritériem určíme aproximaci normálního řešení soustavy $Kx = y$. Tato aproximace $x^{(k)}$ jest definována:

(i) $\|Kx^{(k)} - (y + \varepsilon)\|^2 \leqq \Delta^2$

(ii) Je-li $\|Kx^{(j)} - (y + \varepsilon)\|^2 \leqq \Delta^2$ pak $j \geqq k$.

*Author's address: Jiří Neuberg*, Matematicko-fyzikální fakulta KU, Malostranské n. 25, 118 00 Praha 1.