Numerische Mathematik 12, 139–145 (1968)

Instability and Non-Monotonicity Phenomena in Discretizations to Boundary-Value Problems*

HANS J. STETTER

Received May 26, 1968

Dedicated to ROBERT SAUER on the occasion of his 70th birthday

Introduction

When the differential operator $-d^2/dt^2$, with homogeneous boundary conditions on both ends of a finite interval I of the t-axis, is replaced by the second difference quotient on a finite mesh in I the resulting finite-dimensional operator has a number of agreeable properties: (a) its inverse is a positive operator, (b) this inverse operator is bounded independently of the refinement of the mesh, (c) the "difference" between the inverses of the infinitesimal and the finite operator tends to zero with the refinement of the mesh. These properties are typical for a number of well-known discretizations of boundary-value problems (see e.g. [1]). With discretizations of *initial*-value problems, properties (b) and (c) are commonly called "stability" and "strong stability" (see e.g. [2]), resp., and play an important role in the analysis of discretization methods. With boundary-value problems, they are hardly ever explicitly discussed but rather taken for granted.

In this paper we will show that there are natural *consistent* discretizations of the above differential operator which do not possess all or any of the properties (a) to (c) (naturally, not (b) implies not (c)). The investigations were stimulated by the observation that with initial-value problems there are many classes of consistent discretizations which are not strongly stable or not stable at all while for boundary-value problems no such cases seem to have been reported. Also it seemed to be an open question whether properties (a) and (c) are correlated.

For the interval $I=[0,1]$, the inverse operator G to $-d^2/dt^2$, with $y(0)=y(1)=0$, is given by $Gf:=\int_0^1 g(t,\tau)\,f(\tau)\,d\tau$,

$$g(t,\tau):=\begin{cases}(1-t)\,\tau, & \tau\leqq t,\\ t(1-\tau), & \tau\geqq t.\end{cases}$$

Let $I_n:=\{\nu/n,\ \nu=0\ (1)\ n\}$ be the mesh in I; a function $I_n\to R$ is naturally represented by an element of R^{n-1} if we restrict ourselves to functions satisfying the above boundary conditions, mappings between such functions are represented by square matrices of order $n-1$. The finite approximation to $-d^2/dt^2$ generated

* The research in this paper has been sponsored in part by the United States Government under Contract 61(052)-960.

by the second difference quotient on I_n is represented by $n^2 A$ where

$$A := \begin{pmatrix} 2 & -1 & 0 & \dots & & \\ -1 & 2 & -1 & 0 & & 0 \\ 0 & -1 & 2 & -1 & 0 & \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}, \quad \text{with } A^{-1} =: (\gamma_{\mu\nu}),$$

$$\gamma_{\mu\nu} := \begin{cases} \frac{(n-\mu)\,\nu}{n}, & \nu \leqq \mu, \\ \frac{\mu\,(n-\nu)}{n}, & \nu \geqq \mu. \end{cases} \tag{0.1}$$

Thus the inverse operator represented by the matrix $\frac{1}{n^2} A^{-1}$ is positive, is bounded by $\frac{1}{8}$ in maximum norm independently of n, and $\lim_{n\to\infty} \varrho\left(\frac{1}{n^2} A^{-1}\right) = \frac{1}{\pi^2}$. Furthermore, if we define a discretization G_n of G as the matrix $\frac{1}{n}\left(g\left(\frac{\mu}{n}, \frac{\nu}{n}\right)\right)$ we have $\frac{1}{n^2} A^{-1} = G_n$ which is the non plus ultra of strong stability.

We will now restrict ourselves to the consideration of the operator $-d^2/dt^2$ on the function space $\{y: I \to R : y \in C^{(3)}[0, 1], y(0) = y''(0) = y(1) = y''(1) = 0\}$.[1] Then each member of the class P of finite-difference operators represented by

$$n^2 (A + \alpha_1 A^2 + \cdots + \alpha_m A^{m+1}) =: n^2 p(A), \quad m \geqq 1, \tag{0.2}$$

is a consistent approximation to $-d^2/dt^2$ on I_n. Each of the matrices $p(A)$ is a symmetric band matrix with $m+1$ non-vanishing codiagonals. It is within this simple class of operators that we will find a variety of phenomena which are in striking contrast to the tame behavior of A for $n \to \infty$.

I. Stability with Respect to Euclidean Norm

Let

$$d(A) := p(A)^{-1} - A^{-1}.$$

We call a discretization from class P represented by the matrix $n^2 p(A)$

stable $\quad \frac{1}{n^2} \|p(A)^{-1}\| = O(1)$

if $\quad$ as $n \to \infty$.

strongly stable $\quad \frac{1}{n^2} \|d(A)\| = o(1)$

As the orders of the matrices involved tend to infinity with $n \to \infty$ stability properties may depend on the choice of norms[2].

It is well-known that the spectrum of A consists of the $n-1$ values

$$0 < \lambda_\nu := 4 \sin^2 \frac{\nu}{n} \frac{\pi}{2} < 4, \quad \nu = 1\,(1)\,n-1. \tag{1.1}$$

Since $p(A)^{-1}$ as well as $d(A)$ are rational functions of A their spectral radii are easily established from (1.1). Due to the normality of A the behavior of these

[1] The restrictions are not essential: The conditions on y'' can always be met by a simple transformation, the condition on the differentiability of y is reasonable when one considers higher order approximations.

[2] Matrix norms will always be the l.u.b. norms associated with the given vector norms.

spectral radii for $n \to \infty$ displays the stability and strong stability, resp., of our discretization operator with respect to the Euclidean norm.

Theorem 1. Let $p(x) > 0$, $x \in (0, 4)$, and

$$p(4-x) = \sum_{\mu=0}^{m+1} c_\mu (4-x)^\mu. \tag{1.2}$$

W.r.t. the Euclidean norm the discretization is

strongly stable		$c_0 \neq 0$,
stable but not strongly stable	if	$c_0 = 0$, $c_1 \neq 0$,
unstable with $\|n^2 p(A)^{-1}\|_E = O(n^{2s-2})$		$c_0 = \cdots = c_{s-1} = 0$, $c_s \neq 0$.

Proof. Follows immediately from (1.1) and

$$\|p(A)^{-1}\|_E = \max_\nu \frac{1}{p(\lambda_\nu)},$$

$$\|d(A)\|_E = \max_\nu \left| \frac{1}{p(\lambda_\nu)} - \frac{1}{\lambda_\nu} \right|$$

since $p(x) = x + O(x^2)$ according to (0.2).

For $\xi \in (0, 4)$ we define

$$\arccos(\xi/2 - 1) =: \varphi(\xi) =: r(\xi)\pi. \tag{1.3}$$

When p has a zero ξ in $(0, 4)$ the behavior of $p(A)^{-1}$ and $d(A)$ is dependent upon $r(\xi)$.

Theorem 2. Let $p(x) = \sum_{\mu=1}^{m+1} c_\mu (x-\xi)^\mu$, $\xi \in (0, 4)$, $p(x) \neq 0$ for $x \neq \xi$, $x \in (0, 4]$. If $r(\xi)$ is a rational number, $r(\xi) = k/l$, k, l relatively prime, $p(A)$ is singular for $n = ql$, q integer. For n from any infinitely increasing sequence of integers not containing multiples of l, w.r.t. the Euclidean norm the discretization is

strongly stable		$c_1 \neq 0$,
stable but not strongly stable	if	$c_1 = 0$, $c_2 \neq 0$,
unstable with $\|n^2 p(A)^{-1}\|_E = O(n^{s-2})$		$c_1 = \cdots = c_{s-1} = 0$, $c_s \neq 0$.

If $r(\xi)$ is irrational, the discretization is never strongly stable w.r.t. the Euclidean norm for n from an arbitrary increasing sequence of integers. It is stable for any such sequence if $r(\xi)$ is a quadratic irrationality and ξ a simple zero, otherwise it is generally unstable.

Proof. a) $r(\xi) = k/l$: For $n = ql$, q integer, $p(\lambda_{qk}) = 0$. For $n \neq ql$, $|r(\xi) - \nu/n| \geqq \frac{1}{ln}$ for the closest rational approximation with denominator n which implies $|p(\lambda_\nu)^{-1}| \leqq \left| \frac{nl}{\sin \varphi(\xi)} \right|^s$ by (1.3).

b) $r(\xi)$ irrational: By a well-known theorem on continued fractions (see e.g. [3]) there are infinitely many integers n and associated numerators ν such that $|r(\xi) - \nu/n| < 1/n^2$. Hence $|1/p(\lambda_\nu)| > \frac{n^2}{|\sin \varphi(\xi)|}$ for these n so that $\lim_{n\to\infty} \frac{1}{n^2} \|d(A)\|_E$ does not exist. For quadratically irrational $r(\xi)$, there exists a constant C such that $|r(\xi) - \nu/n| > C/n^2$ for the closest rational approximation with denominator n (see e.g. [3]) which implies stability if $p'(\xi) \neq 0$. For other types of irrationalities no such inequalities can be established.

II. Stability with Respect to the Maximum Norm

Since we are interested in the behavior for $n \to \infty$ the various norms for R^n are *not* equivalent but we have

Lemma 1 (e.g. [4]). For an arbitrary $n \times n$-matrix A

$$\|A\|_{\max} \leqq \sqrt{n}\, \|A\|_E, \tag{2.1}$$

the equality being attainable.

Furthermore, for normal A and rational $f(A)$, we have the natural relation

$$\|f(A)\|_{\max} \geqq \varrho(f(A)) = \|f(A)\|_E. \tag{2.2}$$

Hence, all results of sect. I of the type $\|\ldots\|_E = O(n^s)$, $|s| \geqq 1$, can be translated into maximum norm results at least qualitatively. However, the important assertions of the type $\|\ldots\|_E = O(1)$ cannot be claimed w.r.t. the maximum norm. We will therefore prove that the results of sect. I carry over to the maximum norm *without* change in the powers of n involved, at least for a more restricted class of matrices $p(A)$.

Let $p(x) =: x\,\bar{p}(x)$ with $\bar{p}(0) = 1$ by (0.2). Then we have from

$$\begin{aligned} p(A)^{-1} &= [\bar{p}(A)\,A]^{-1} = A^{-1} + d(A), \\ A^{-1} &= (A^{-1} + d(A))\,\bar{p}(A), \\ d(A) &= A^{-1}(I - \bar{p}(A))\,\bar{p}(A)^{-1}. \end{aligned} \tag{2.4}$$

As we will need explicit estimates of the elements of $d(A)$ we will now consider only quadratic polynomials $p(A) = A + \alpha A^2$, $\alpha \neq 0$. For these

$$d(A) = -\alpha(I + \alpha A)^{-1}$$

and the elements $\delta_{\mu\nu}$ of $d(A)$ can easily be calculated from the recursion

$$-\alpha\,\delta_{\mu,\nu-1} + (1+2\alpha)\,\delta_{\mu\nu} - \alpha\,\delta_{\mu,\nu+1} = \begin{cases} 0, & \nu < \mu \quad \text{and} \quad \nu > \mu, \\ -\alpha, & \nu = \mu, \end{cases}$$

with boundary values $\delta_{\mu 0} = \delta_{\mu n} = 0$. This yields for $\alpha \neq -\frac{1}{4}$ $(\alpha \neq 0)$:

$$\delta_{\mu\nu} = \begin{cases} -\dfrac{(z^{n-\mu} - z^{-(n-\mu)})\,(z^{\nu} - z^{-\nu})}{(z^n - z^{-n})\,(z - z^{-1})}, & \nu \leqq \mu, \\ -\dfrac{(z^{\mu} - z^{-\mu})\,(z^{n-\nu} - z^{-(n-\nu)})}{(z^n - z^{-n})\,(z - z^{-1})}, & \nu \geqq \mu, \end{cases} \tag{2.5}$$

where z is one of the zeros of $z^2 - \dfrac{1+2\alpha}{\alpha}\, z + 1$.

For quadratic $p(x)$, $\alpha > -\frac{1}{4}$ implies $p(x) > 0$ in $(0, 4]$ while for $\alpha \leqq -\frac{1}{4}$ $p(x)$ has a simple zero at $\xi = -\dfrac{1}{\alpha} \in (0, 4]$. For $\alpha < -\dfrac{1}{4}$ we can rewrite (2.5) with the aid of (1.3):

$$\delta_{\mu\nu} = \begin{cases} -\dfrac{\sin(n-\mu)\,\varphi \sin \nu\varphi}{\sin n\varphi \sin \varphi}, & \nu \leqq \mu, \\ -\dfrac{\sin \mu\varphi \sin(n-\nu)\,\varphi}{\sin n\varphi \sin \varphi}, & \nu \geqq \mu. \end{cases} \tag{2.6}$$

Theorem 3. For $p(A)=A+\alpha A^2$

$$\frac{1}{n^2}\|d(A)\|_{\max}=\begin{cases} O\left(\frac{1}{n^2}\right) & \text{for } \alpha>-\frac{1}{4}, \\ O\left(\frac{1}{n}\right) & \text{for } \alpha<-\frac{1}{4} \text{ with } r(\xi)=k/l \text{ rational}, \\ & \quad n\neq q\,l, \quad q \text{ integer}, \\ O(1) & \text{for } \alpha=-\frac{1}{4} \text{ and for } \alpha<-\frac{1}{4} \text{ with } r(\xi) \\ & \quad \text{quadratically irrational.} \end{cases}$$

For $\alpha<-\frac{1}{4}$, the order of $\frac{1}{n^2}\|d(A)\|_{\max}$ may equal any positive power of n along certain infinite sequences corresponding to the type of irrationality (other than quadratic) of $r(\xi)$.

Proof. a) For $\alpha>-\frac{1}{4}$, in (2.5) we may assume $|z|>1$ without restriction of generality. Summation of (2.5) yields:

$$\sum_{\nu=1}^{n-1}|\delta_{\mu\nu}|=\frac{|z|}{|(|z|-1)^2}\,\frac{|z|^n-|z|^{-n}-(|z|^{\mu}-|z|^{-\mu})-(|z|^{n-\mu}-|z|^{-(n-\mu)})}{|z|^n-|z|^{-n}}$$

$$\leqq\frac{|z|}{(|z|-1)^2}\left[1-\frac{2}{|z|^{n/2}+|z|^{-n/2}}\right]\leqq\frac{|z|}{(|z|-1)^2}\ \text{independent of } n.$$

b) For $\alpha<-\frac{1}{4}$, $\|d(A)\|_{\max}$ obviously depends on the smallness of $\sin n\varphi$ and hence again on the number theoretic properties of $r(\xi)$. The assertions follow through the same arguments as in the proof of Theorem 2.

c) For $\alpha=-\frac{1}{4}$, it is easy to see that

$$\delta_{\mu\nu}:=(-1)^{\mu+\nu}\gamma_{\mu\nu}, \tag{2.7}$$

hence

$$\frac{1}{n^2}\|d(A)\|_{\max}=\frac{1}{n^2}\|A^{-1}\|_{\max}=\frac{1}{8}\neq 0.$$

Theorem 3 suggests that *all* $O(n^s)$ assertions may carry over from sect. I without change also for general $p(A)$. One further evidence for this conjecture is provided by the consideration of a double root of $p(x)$ at $x=4$:

It is easy to see that $p(x)=x-x^2/2+x^3/16$ satisfies the assumptions of Theorem 1 with $c_0=c_1=0$, $c_2\neq 0$. Application of (2.4) yields

$$d(A)=\left(\frac{I}{2}-\frac{A}{16}\right)\left(I-\frac{A}{2}+\frac{A^2}{16}\right)^{-1} \tag{2.8}$$

and by inspection one finds

$$\bar{p}(A)=I-\frac{A}{2}+\frac{A^2}{16}=\frac{1}{16}|A^2| \tag{2.9}$$

where $|(\alpha_{\mu\nu})|:=(|\alpha_{\mu\nu}|)$. Due to the distribution of minus signs along alternate codiagonals in A and A^2 this implies

$$|\bar{p}(A)^{-1}|=16A^{-2},$$

$$\|\bar{p}(A)^{-1}\|_{\max}=16\|A^{-2}\|_{\max}=O(n^4),\quad \frac{1}{n^2}\|d(A)\|_{\max}\leqq O(n^2).$$

But due to (2.2) and Theorem 1, the equality sign has to hold.

III. Monotonicity

All discretizations of $-d^2/dt^2$ considered in the literature so far turned out to share the inverse positivity of $n^2 A$ (see e.g. [1, 5]). Since all these discretizations have a positive spectrum one might conjecture that a consistent finite difference approximation to $-d^2/dt^2$ with positive spectrum has a positive inverse. Also since all these discretizations are strongly stable one might conjecture that this property implies the positivity of the inverse.

We will show that both conjectures are false but that within the class of approximations $A+\alpha A^2$ the combination of positivity of the spectrum *and* strong stability is necessary and sufficient for the positivity of the inverse. This suggests the conjecture that this may be the correct criterion for any consistent discretization of $-d^2/dt^2$ to have a positive inverse.

Theorem 4. Let $p(A)=A+\alpha A^2$. Then

$$p(A)^{-1}\begin{cases} >0 & \text{for } \alpha > -\frac{1}{4}, \\ \geqq 0 & \text{for } \alpha = -\frac{1}{4}, \\ \text{contains negative elements} & \text{for } \alpha < -\frac{1}{4}. \end{cases}$$

Proof. a) For $-\frac{1}{4}<\alpha\leqq 0$,

$$A+\alpha A^2=\begin{pmatrix} 2+5\alpha & -(1+4\alpha) & \alpha & 0 & & \\ -(1+4\alpha) & 2+6\alpha & -(1+4\alpha) & \alpha & & 0 \\ \alpha & -(1+4\alpha) & 2+6\alpha & -(1+4\alpha) & \alpha & \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}$$

is of positive type (i.e. satisfies the weak row-sum criterion). For $\alpha>0$, we use a regular splitting of $p(A)$ (see e.g. [1]):

$$A+\alpha A^2=\alpha B^2-\frac{1}{4\alpha}I \qquad \text{with}$$

$$B:=A+\frac{1}{2\alpha}I, \quad B^{-1}>0. \tag{3.1}$$

According to Theorem 1.1.3 of [6] it remains to prove $\varrho\left(\frac{1}{4\alpha^2}B^{-2}\right)<1$. But (3.1) implies this inequality due to (1.1).

b) For $\alpha=-\frac{1}{4}$, (2.7) implies

$$(p(A)^{-1})_{\mu\nu}=\begin{cases} 2\gamma_{\mu\nu}>0 & \text{for } \mu+\nu \text{ even}, \\ 0 & \text{for } \mu+\nu \text{ odd}. \end{cases}$$

c) For $\alpha<-\frac{1}{4}$, compare $\delta_{1\nu}=-\frac{\sin(n-\nu)\varphi}{\sin n\varphi}$ and $\gamma_{1\nu}=\frac{n-\nu}{n}<1$. For sufficiently large[1] n, there are always values of ν such that $\delta_{1\nu}\leqq -1$ which implies $(p(A)^{-1})_{1\nu}=\gamma_{1\nu}+\delta_{1\nu}<0$.

Thus $p(A)^{-1}>0$ fails equally for the case $\alpha=-\frac{1}{4}$ with positive spectrum but no strong stability and for cases $\alpha<-\frac{1}{4}$ with strong stability (see Theorems 2 and 3) but some negative eigenvalues $p(\lambda_\nu)$.

[1] Actually one has to exclude only a few trivial cases.

While only vanishing but no negative elements are present in $(A-\frac{1}{4}A^2)^{-1}$ our example $p(A)=A-\frac{1}{2}A^2+\frac{1}{16}A^3$ represents a consistent approximation to $-d^2/dt^2$ with $p(\lambda_\nu)>0$, $\nu=1(1)n-1$, and negative elements in $p(A)^{-1}$ as can be easily established from the representation (2.8) with the help of (2.9).

IV. Further Phenomena, Conclusions

A further unusual phenomenon which can be shown to occur with some discretizations of d^2/dt^2 is that the stability properties may be changed (usually improved) by the addition of terms which are discretizations of "lower order terms" in the corresponding differential operator: Discretizations which are genuinely unstable for $-y''=f(t)$ may be stable for $-y''+gy=f(t)$, $g>0$.

Also it is evident from Theorem 1 that an instability due to a multiple zero of $p(x)$ at $x=4$, e.g., may be "overcome" by a sufficiently high order of consistency since the divergence of $p(A)^{-1}$ is only proportional to a *power* of n.

When all the phenomena displayed in this paper have received little or no attention so far, this seems to be due to the fact that in the numerical solution of boundary-value problems by discrete variable methods the traditional view point has been somewhat different from that with initial-value problems: With boundary-value problems the basic discretization has usually been chosen in a natural manner (e.g., from variational principles) while much emphasis has been placed on how to solve the generated system of equations economically. With initial-value problems, on the other hand, multitudes of different discretizations have been suggested which gave rise to the analysis of various instability phenomena. The results reported in this paper show that similarly unexpected phenomena may occur with discretizations of boundary-value problems when one analyzes the behavior of *classes* of consistent approximations containing also members which would normally not be considered seriously.

In conclusion, it should be remarked that most of the effects reported have been verified in a number of numerical experiments. Further experiments and analytic considerations are under way to clarify more fully the abnormalities which may occur with discretizations of boundary-value problems.

References

1. Varga, R. S.: Matrix iterative analysis. Englewood Cliffs: Prentice-Hall, Inc. 1962.
2. Stetter, H. J.: A study of strong and weak stability in discretization algorithms. J. SIAM Num. Anal. **2**, 265–280 (1965).
3. Khintchine, A. Ya.: Kettenbrüche. Leipzig: Teubner Verlagsges. 1956 (Übersetzung aus dem Russischen).
4. Stone, B. J.: Best possible ratios of certain matrix norms. Stanford University Tech. Report, 1961.
5. Bramble, J. H., and B. E. Hubbard: On a finite difference analogue of an elliptic boundary problem which is neither diagonally dominant nor of non-negative type. J. Math. Physics **43**, 117–132 (1964).
6. Price, H. S.: Monotone and oscillation matrices applied to finite difference approximations. Doctoral Thesis, Case Inst. of Technology, 1965.

Prof. Dr. Hans J. Stetter
Institut für Numerische Mathematik
der TH Wien
A-1040 Wien, Karlsplatz 13, Österreich