High-order finite-difference methods for Poisson's equation

van Linde, Hendrik Jan

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
1971

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
IV. THE THIRD BOUNDARY VALUE PROBLEM.

§ 1. An upper bound for the discretization error.

We shall now approximate the third boundary value problem (1.5), which we here again formulate as

\[-\Delta u = f \quad \text{in } R\]

\[\frac{\partial u}{\partial n} + au = g \quad \text{on } C_1\]

\[u = g_1 \quad \text{on } C_2\]

by a finite-difference analogue, using the operators derived in chapters II and III. Consider the approximation

\[-\Delta_h U(P) = f^*(P) \quad P \in R^*_h + C^*_1 h + C^*_2 h\]

\[\frac{\delta n U(P)}{\delta n} + a(P)U(P) = g^*(P) \quad P \in C^*_1 h\]

\[U(P) = g_1(P) \quad P \in C^*_2 h\]

with \(\Delta_h\) defined by (1.6) for \(P \in R^*_h\), by (1.10) for \(P \in C^*_2 h\), and by the appropriate operator defined in chapter III for \(P \in C^*_1 h\). \(\delta_n\) is defined by (2.33). The sets \(R^*_h, C^*_i h\) and \(C^*_i h, i = 1, 2\), are as in chapter I. We have already mentioned the division of \(C^*_h\) in two sets \(C^*_i h\) and \(C^*_2 h\), but have not yet discussed it in detail. Points lying in \(C^*_h\) have a part of the boundary lying inside their nine-point molecule. A point \(P \in C^*_i h\) will be in \(C^*_i h\), if this part of the boundary does entirely belong to \(C^*_i\), for \(i = 1, 2\). If the part of the boundary lying inside the nine-point molecule does not belong exclusively to \(C^*_1\) or \(C^*_2\), the corresponding point \(P\) is in \(C^*_2 h\) if \(C^*_2\) cuts a main axis or if \(C^*_2\) cuts the boundary of the molecule, while the main axes are entirely in \(R\), and otherwise in \(C^*_1 h\).
The functions $f^*(P)$ and $g^*(P)$ in (4.2) are defined as

\[
\begin{align*}
    f^*(P) &= f(P) + \frac{h^2}{12} \Delta f(P) \quad P \in R_h \\
    f^*(P) &= f(P) - hF(P) \quad P \in C^*_h \\
    g^*(P) &= g(P) + F_1(P) \quad P \in C^*_2h
\end{align*}
\]

(4.3)

with $hF(P) = F(\Delta x, \Delta y)u(P)$ (see (3.3)) and $F_1(P)$ a known function of $f$, $g$ and their derivatives defined by considering

\[
\delta_n u(P) + \alpha(P)u(P) - g(P) - F_1(P) \leq k_1h^3
\]

(4.4) as an equivalent notation for (2.34).

The matrix of the system (4.2) is of positive type provided

\[
\sum_{i=1}^4 a_i \left[ x_i y_i a_i(P) + \frac{1}{6} (3x_i^2 y_i + y_i^3) a_{88}(P) \right] + \alpha(P) \geq 0
\]

(4.5) is true for all $P \in C^*_h$. Since we have stipulated that $\alpha$ is bounded away from zero (4.5) can always be satisfied for $h$ chosen sufficiently small.

This matrix then possesses a non-negative inverse (theorems 1 and 2 of chapter I). The general idea behind the following proof concerning the magnitude of the discretization error has been taken from Bramble and Hubbard [5]; since we necessarily work with a different discrete Green's function the whole detailed proof has to be given again for this case. As we shall see in chapter V the use of a different and more complicated Green's function may severely complicate the proof of corresponding theorems. We now introduce the discrete Green's function $G_h(P, Q)$ for the region $R$ under consideration, defined by

\[
\begin{align*}
    -\delta_h G_h(P, Q) &= h^{-2} \delta(P, Q) \quad P \in R_h + C^*_1h + C^*_2h \\
    \delta_n G_h(P, Q) + \alpha(Q) G_h(P, Q) &= h^{-7} \delta(P, Q) \quad P \in C^*_h
\end{align*}
\]

(4.6)
\[(4.6)\]
\[G_h(P, Q) = h^{-7} \delta(P, Q) \quad P \in C_{2h}\]

with \(Q \in R_{h} + C_{1h} + C_{2h} + C_{1h} + C_{2h}\). The symbol \(\delta(P, Q)\) denotes the Kronecker-delta. Here and in the following chapters we suppose the operators \(A_h\) and \(\delta_n\) to be working on the first parameter. Clearly \(G_h(P, Q)\) is non-negative, being the inverse of the coefficient matrix of (4.2), multiplied by a non-negative diagonal matrix.

We now have, for any mesh-function \(V\):

\[
V(P) = h^2 \sum_{Q \in R_{h} + C_{1h} + C_{2h}} G_h(P, Q) [-\Delta_h V(Q)] + \sum_{Q \in C_{1h}} G_h(P, Q) [\delta_n V(Q) + \alpha(Q) V(Q)]
\]

\[(4.7)\]

\[
+ \sum_{Q \in C_{2h}} G_h(P, Q) V(Q)
\]

which follows from the fact that the coefficient matrix of (4.2) is non-singular.

It must be pointed out that we have used a definition slightly different from the one used by Bramble and Hubbard [5]. This difference consists of the inclusion of the factors \(h^{-7}\) in the second and third lines of (4.6). The reason for this is, that (4.7) is now more in agreement with the continuous representation of the solution of (4.1) by means of kernel functions:

\[
u(P) = \int f G_1(P, Q) f(Q) d\sigma
\]

\[+ \int_{C_1} G_2(P, Q) g(Q) ds + \int_{C_2} G_3(P, Q) g_1(Q) ds\]

We first take \(V(P) = 1\) in (4.7), which yields
We now suppose that a function $\phi \in C^3(R)$ exists satisfying

$$-\Delta \phi \geq 1 \quad \text{in } R$$

(4.9)

$$\frac{\partial \phi}{\partial t} + a\phi \geq 1 \quad \text{on } C_1$$

Then for sufficiently small $h$

$$-\Delta_h \phi(P) \geq \frac{1}{2} \quad \text{in } C_1$$

$$\delta_h \phi(P) + a(P)\phi(P) \geq \frac{1}{2} \quad \text{in } C_1$$

If we now take $V(P) = \phi(P)$ in (4.7) we obtain

$$h^2 \sum_{Q \in C_2h} G_h(P, Q) + h \sum_{Q \in C_2h} G_h(P, Q) \leq 4|\phi|_M$$

(4.10)

$$Q \in R + C_1h + C_2h \quad Q \in C_1h$$

with $|\phi|_M = \max_{P \in R + C_1h + C_2h} |\phi(P)|$

We now introduce the sets $C_1^{**}$ and $C_2^{**}$, the subsets of $C_1^*$ and $C_2^*$ where $\Delta_h$ is represented by the ordinary five-point formula.

We now define a function $W(P)$ by: $W(P) = 0$ on $C_1$, $W(P) = 1$ in $R_h$, in $C_1^{**}$ and $C_2^{**}$, and in those points of $(C_1^{**} \cup C_2^{**}) - (C_1^{**} \cup C_2^{**})$ which do not belong to a star, the centre of which is in $C_2^{**}$. In the points of $(C_1^{**} \cup C_2^{**}) - (C_1^{**} \cup C_2^{**})$ which belong to a star, the centre of which is in $C_2^{**}$, $W(P) = \frac{7}{8}$.

We can then show

$$-\Delta_h W(P) \geq \frac{1}{4} - 2 \quad \text{in } C_1^{**} + C_2^{**}$$

$$-\Delta_h W(P) \geq 0 \quad \text{in } R_h$$

We shall illustrate this by an example; in the configuration given in fig. 4.1 the point $A$ corresponds to case II in
chapter III. The function $\tilde{W}$ has, in the relevant points, the values indicated in the figure.

![Figure 4.1](image)

Using these values in (3.6), we can easily verify the above inequality for the point $A$. The inequality follows for all other points in $C_{1h}^* + C_{2h}^*$ in a similar way, while the proof for points in $R_h$ is trivial.

Taking $V(P) = W(P)$ in (4.7) we have

$$7 \geq h^2 \sum_{Q \in C_{1h}^* + C_{2h}^*} G^a_h(P, Q) \left[ -\Delta_h W(Q) \right] + h \sum_{Q \in C_{1h}^*} G^a_h(P, Q) \left[ \delta_h W(Q) + \alpha(Q) W(Q) \right]$$

or

$$\sum_{Q \in C_{1h}^* + C_{2h}^*} G^a_h(P, Q) \leq 4 + 4 \max_{Q \in C_{1h}^*} \left[ \sum_{i=1}^{4} a_i(Q) \right] h \sum_{Q \in C_{1h}^*} G^a_h(P, Q)$$

We also have (see (2.24)):

$$1 = \sum_{i=1}^{4} a_i(y_i + \frac{1}{6}(3x^2_i y_i - y_i^3)(\alpha + K))$$

$$4 \leq \sum_{i=1}^{4} a_i \left[ \min(y_i + \frac{1}{6}(3x^2_i y_i - y_i^3)(\alpha + K)) \right]$$
for sufficiently small $\epsilon$, for any $P \in C_{1h}^\epsilon$.

We now have

\[
\frac{4}{\sum_{i=1}^{d} a_i} \leq \frac{2h - 1}{3h}
\]

and this yields, with (4.10) and (4.11)

\[
(4.13) \quad \epsilon \Delta G_h(P, Q) \leq 4(h + \frac{3}{3} |\phi|_M)
\]

We shall derive a sharper bound for $Q \in C_{2h}^\epsilon$. Take $W(P) = 0$ on $C_h$, $W(P) = 1$ everywhere in $R_{C_{2h}^\epsilon}^C C_{1h}^\epsilon C_{2h}^\epsilon$ and in those points of $C_{2h}^\epsilon C_{2h}^\epsilon$ which do not belong to a five-point star, the centre of which is in $C_{2h}^\epsilon$. In the points of $C_{2h}^\epsilon C_{2h}^\epsilon$ which belong to a five-point star, the centre of which is in $C_{2h}^\epsilon$, $W(P) = \frac{7}{8}$.

We then have

\[
\epsilon \Delta W(P) \geq \frac{2h - 1}{4h} P \in C_{2h}^\epsilon
\]

\[
\epsilon \Delta W(P) \geq 0 P \in R_{C_{2h}^\epsilon}^C C_{1h}^\epsilon
\]

$V(P) = W(P)$ in (4.7) then yields

\[
(4.14) \quad \sum_{Q \in C_{2h}^\epsilon} G_h(P, Q) \leq 4
\]

We can now formulate the following important theorem:

**Theorem 1.** Let $u \in C^5(\bar{R})$ be the solution of (4.1) and suppose that a function $\phi$ satisfying (4.9) exists. Then we have

\[
(4.15) \quad \max_{P} |\epsilon(P)| \leq k h^3
\]

with $\epsilon(P) = u(P) - U(P), P \in R_{C_{2h}^\epsilon}^C C_{1h}^\epsilon C_{2h}^\epsilon C_{1h}^\epsilon C_{2h}^\epsilon, U$ being the solution of (4.2). The constant $k$ used in (4.15) depends only
on \( u \) and \( \phi \) but not on \( h \).

**Proof.** In (4.7) take \( V(P) = \varepsilon(P) \), then

\[
\varepsilon(P) = h^2 \sum_{Q \in R} G_h(P, Q)[-\Delta_h \varepsilon(Q)] + h \sum_{Q \in R} G_h(P, Q)[\delta_n \varepsilon(Q) + \alpha(Q) \varepsilon(Q)]
\]

\[
e(P) = h^2 \sum_{Q \in R} G_h(P, Q) [-\Delta_h \varepsilon(Q)] + h \sum_{Q \in R} G_h(P, Q) [\delta_n \varepsilon(Q) + \alpha(Q) \varepsilon(Q)]
\]

\[
Q \in \mathcal{C}_h + \mathcal{C}_1h + \mathcal{C}_2h
\]

Since \( G_h(P, Q) \geq 0 \), we have

\[
|\varepsilon(P)| \leq [h^2 \sum_{Q \in R} G_h(P, Q)]. \max_{Q \in \mathcal{R}_h} |\Delta_h \varepsilon(Q)|
\]

\[
+ [h \sum_{Q \in \mathcal{R}_h} G_h(P, Q)]. \max_{Q \in \mathcal{R}_h} |h \Delta_h \varepsilon(Q)|
\]

\[
+ \left[ \sum_{Q \in \mathcal{R}_h} G_h(P, Q) \right]. \max_{Q \in \mathcal{R}_h} |h^2 \Delta_h \varepsilon(Q)|
\]

\[
+ \left[ h \sum_{Q \in \mathcal{R}_h} G_h(P, Q) \right]. \max_{Q \in \mathcal{R}_h} |\delta_n \varepsilon(Q) + \alpha(Q) \varepsilon(Q)|
\]

(4.16)

This immediately yields (4.15) using (4.10), (4.13) and (4.14) together with (1.9), (3.3), (1.11) and (2.34).

**Emphasis** must be put on the fact that theorem 1 is true for all operators that have truncation errors of the same order in \( h \) as the operators we have used. For special regions \( R \) operators may be used which are much simpler than those used above. The operators of chapters II and III belong to a class of admissible operators.

§ 2. An example.

Consider the region \( R \) given in fig. 4.2. The arc between \((0,1)\) and \((1,0)\) shall be \( C_1 \), the remaining part of the boundary \( C_2 \). We shall approximate the function \( u \) for which
\[-\Delta u = \frac{-4}{(x^2+y^2+1)^2} \quad \text{in } R\]

\[
\begin{align*}
\frac{\partial u}{\partial n} + u &= 1 + \ln 2 &\text{on } C_1 \\
u &= \ln(x^2+y^2+1) &\text{on } C_2
\end{align*}
\]

Fig. 4.2

The exact solution is \(u(x,y) = \ln(x^2+y^2+1)\). The choices made for this problem give our formulas a somewhat simpler character. For instance \(g, \alpha\) and \(K\) are now constant on \(C_1\).

The system (2.24) can now be written as

\[
\begin{align*}
\sum a_i \{y_i + \frac{1}{3}(3x_i y_i - y_i^3)\} &= 1 \\
\sum a_i \{x_i + 2x_i y_i\} &= 0 \\
\sum a_i \{(x_i^2 - y_i^2) + \frac{4}{3}(3x_i^2 y_i - y_i^3)\} &= 0 \\
\sum a_i \{x_i^3 - 3x_i y_i^2\} &= 0
\end{align*}
\]

For (2.33) we have

\[
\frac{\delta}{\delta n} V(P) = \sum_{i=1}^{4} a_i \{V(P) - V(P_i)\}
\]

and for (2.34)

\[
\left| \frac{\delta}{\delta n} u(P) + u(P) - \left\{ 1 + \ln 2 + \sum_{i=1}^{4} a_i \left[ \frac{-y_i^2}{2} - \frac{y_i^3}{2} + \frac{x_i^2 y_i}{2} \right] \right\} \right| \leq k_1 h^2
\]
The problem thus formulated was solved by the use of a square net with mesh width \( h = 1/12 \) and \( h = 1/24 \) respectively, to inspect the decrease in the error. An approximation as in (4.2) was used. The resulting system of linear equations was solved by Gauss-Seidel iteration. The convergence of this process was rather slow and it was experimentally found that successive point overrelaxation could not be used for the chosen point order. This order, the seemingly most natural one, consisted of taking all inner net-points first, row by row, from left to right and from bottom to top, and finally the boundary points on \( C_1 \) starting from \((0,1)\).

For \( h = 1/12 \) the problem was solved twice, once starting from \( u = 1 \) in the whole field and once starting from the exact solution of (4.17). These results completely agreed. For \( h = 1/24 \) only the second set of starting values was used. For \( h = 1/12 \) a maximum error of \( 687_{10}^{-5} \) was found, compared with \( 89_{10}^{-5} \) for \( h = 1/24 \), which is in complete agreement with the theory, since according to (4.15) the error should decrease as \( h^3 \).

These results, as those of all other numerical examples, were obtained, with ALGOL 60-programs written by the author, on the Telefunken TR 4 computer (32K of 48 bit words) of the Rekencentrum der Rijksuniversiteit Groningen, of which computing centre the author is a staff member.