26

APPROXIMATE SOLUTION AND ERROR BOUNDS FOR QUASILINEAR ELLIPTIC BOUNDARY VALUE PROBLEMS

by

J. B. Rosen

Technical Report #30

November 1968

APPROXIMATE SOLUTION AND ERROR BOUNDS FOR QUASILINEAR ELLIPTIC BOUNDARY VALUE PROBLEMS*

BY

J. B. ROSEN

ABSTRACT

The approximate solution of quasilinear elliptic boundary value problems by linear programming is considered with emphasis on computational aspects. The solution is obtained as a linear combination of m selected functions, with coefficients determined so as to minimize a weighted sum of the maximum error in the differential equation and on the boundary. A realistic error bound is obtained together with the solution for any value of m, and convergence as $m \rightarrow \infty$ is shown under certain conditions. A variety of linear and quasi-linear two-dimensional problems are solved and the numerical results, including error bounds, are presented and discussed.

---

* This research was supported in part by the National Science Foundation under Grant NSF GP-6070 and in part by the Mathematics Research Center, United States Army, Madison, Wisconsin, under Contract No. DA-31-124-ARO-D-462.

1. INTRODUCTION

This paper is concerned with the numerical determination of approximate solutions and error bounds for quasilinear elliptic boundary value problems. Problems are considered on a bounded domain D in ℓ-dimensional space, with boundary ∂D. Of primary interest is the following type of problem:

$$\begin{aligned} (L + g)[u] &= r \quad \text{in} \quad D \\ u &= s \quad \text{on} \quad \partial D \end{aligned} \tag{1.1}$$

where L is an elliptic differential operator and g may be a nonlinear function.

An approximate solution is assumed of the form $v_m = \sum_{i=1}^{m} \alpha_i \varphi_i(x)$, where the $\varphi_i(x)$ are appropriately selected functions and the coefficients α_i are to be determined in an optimal way. Specifically, the coefficients are determined so as to minimize a weighted sum of the maximum error δ_1 in the differential equation and the maximum boundary error δ_2. If the problem is linear this can be done by solving a single linear programming problem. For the quasilinear problem an iterative linear programming solution is required. The solution of a closely related linear program also gives an error bound for the approximate solution v_m. This error bound is based on the monotone property of the operator $(L + g)$. The approximate solution is defined (and differentiable) at all points in D, and is completely specified

by the m coefficients α_i. Determination of an error bound and convergence of v_m to the exact solution as $m \to \infty$ is also considered here for a more general class of problems.

An extensive study of the use of monotonicity to obtain estimates and bounds for partial differential equations has been carried out by Collatz and Schröder [1, 2, 3, 11, 12]. The present work considers in more detail the numerical aspects of the problem, and in particular takes full advantage of the duality theory of linear programming to get both an approximate solution and an error bound using a standard linear programming code. Essentially the same procedure is used for quasilinear problems as for linear problems, and no special numerical finite difference solution is required to obtain starting values for the quasilinear problems (see for example [11]). A rigorous error bound requires careful consideration of the relation between the maximum error over the closure $\bar{D}$ and the maximum error over a finite grid $\bar{D}_n$. In particular a bound must be known on the difference between $(L + g)[v_m]$ at any point of D and a closest point of D_n. This important problem is taken care of by imposing an appropriate Lipschitz condition on the approximation v_m, where the Lipschitz constant is determined so as to minimize the error bound. This condition is imposed by requiring that the coefficient vector α be an element of a bounded polyhedral set Ω_m, and depends crucially on the use of linear programming. The weighting of the interior and boundary errors is also determined by the linear programming solution so as to minimize the error bound obtained.

A similar approach to that described here has been used for nonlinear parabolic problems [9], and for certain linear harmonic problems with mixed boundary conditions [15], but without a detailed analysis of the effect of the finite grid and minimization of the error bound. A closely related method has been used for nonlinear two-point boundary value problems ($\ell = 1$) and is described in [10]. Other related methods are discussed in a recent survey article [8].

In Section 2 a general problem is formulated and a readily computed error bound given in Theorem 1. With appropriate assumptions on the functions φ_i, convergence of v_m to u as $m \to \infty$ is shown in Theorem 2. In Section 3 we specialize to the quasilinear elliptic problem (1.1). The linear case $g = p(x)u$, where however p may be negative, is considered first. It is shown (Theorem 4) that if a solution exists satisfying certain linear constraints, then $(L + p)$ has the desired monotone property and an error bound can be given in terms of δ_1 and δ_2. This result is extended in Theorem 5 to the case where g may be nonlinear. The only assumption made on g is that it has a continuous derivative with respect to u in D. Given any approximate solution v, we again obtain an error bound if certain linearized constraints have a solution.

The numerical solution by linear programming using a finite grid of n points is considered in Section 4. In order to simplify the calculation the same functions φ_i are used both for the approximate solution v_m and the

error bound function μ_m. For the linear problem only a single linear programming solution is required to determine each of these functions. The function μ_m is first determined by (4.5), and then v_m is determined by (4.6). As shown by Theorem 6, these solutions give the error bound (4.7) in terms of the function ρ given by (4.8). These problems are formulated so that for fixed m and n the maximum value of ρ on $\bar{D}$ is a minimum. A similar bound for the quasilinear problem is given by Theorem 7. The error bound ν appearing in both Theorem 5 and Theorem 7 can be considered as a nonlinear parameter. In general it is desired to find the smallest value of ν which satisfies the constraints. An iterative procedure to determine a suitably small value of ν is given immediately preceding Theorem 8. It is essentially a Newton type iterative solution of the two linear programs (4.5) and (4.6). Its use is justified by Theorem 8. The explicit formulation of the two linear programs using duality is given by (4.21) and (4.22). It is seen that in both cases the problem is reduced to a primal problem with $m + 2$ rows. The solution time will therefore depend primarily on the number m of functions used. Based on this formulation it follows that both the monotone property of $(L + g)$ and the error bound are consequences of a finite solution to the single linear program (4.21). Two important special cases are also considered. These occur when the functions φ_i can be chosen so as to identically satisfy either the boundary conditions or the differential equation.

In Section 5 computational results are presented and discussed. A variety of linear and quasilinear problems in two dimensions were solved

using the negative Laplacian as the elliptic operator. Included in the nonlinear functions g used were $g = -e^{-u}$, $g = \pm e^{u}$ and $g = -u^{2}$. The domains D considered were a square, a truncated square and an ellipse. The approximating functions used included polynomials, trigonometric and harmonic functions. The results obtained, including error bounds, are tabulated in Tables 1-5, and contour plots of the approximate solutions are given in Figures 1-5.

For certain cases where the derivative of g is negative, solutions to (1.1) may not even exist. No solution exists for example when $g = -u^{2}$ and r is sufficiently large, or when $g = -\tau e^{u}$ and τ is sufficiently large. The question of existence of solutions to such problems has been investigated [4, 5]. Based on these results it can be shown [6] that there are values $\hat{r}$ and $\hat{\tau}$ such that no solution exists if $r > \hat{r}$ or $\tau > \hat{\tau}$. The numerical results include approximate solutions and error bounds for values of r and τ close to $\hat{r}$ and $\hat{\tau}$. As might be expected both the solutions and bounds get large as the limiting values are approached.

It should be remarked that it is not the purpose of this work to develop a method for computing highly accurate numerical solutions. Rather, it is desired to develop an efficient computational technique which can be applied to a variety of problems taking advantage of available knowledge about the nature of the solution, and which gives an approximate solution in a convenient form together with the corresponding error bounds.

The author would like to thank his colleagues S. V. Parter and J. W. Daniel for several helpful discussions concerning this work. The computer program used was written by Dennis Kuba who also obtained the computed results given here.

2. ERROR BOUND AND CONVERGENCE

We consider a boundary value problem on a bounded, connected, open domain $D \subset R_\ell$, with boundary ∂D and closure $\overline{D} = D \cup \partial D$. Let F be a differential operator in $\overline{D}$ and B a boundary operator on ∂D, and let $r(x)$ and $s(x)$ be given functions satisfying a Lipschitz condition in $\overline{D}$ and on ∂D, respectively. In particular we assume that a constant $\lambda_0 \geq 0$ exists such that

$$\begin{aligned} |r(x_1) - r(x_2)| &\leq \lambda_0 \|x_1 - x_2\|, \quad \forall x_1, x_2 \in \overline{D} \\ |s(x_1) - s(x_2)| &\leq \lambda_0 \|x_1 - x_2\|, \quad \forall x_1, x_2 \in \partial D \end{aligned} \tag{2.1}$$

where $\|x\| = \max_i |x_i|$.

We assume that there exists a $u = u(x)$, continuous on $\overline{D}$, which satisfies

$$F[u] = r(x) \quad \text{in} \quad D \tag{2.2}$$

$$B[u] = s(x) \quad \text{on} \quad \partial D \tag{2.3}$$

We further assume that F and B have a certain monotone property with respect to the domain $\overline{D}$. Specifically let v and w be any continuous functions in $\overline{D}$ for which $F[v]$ and $F[w]$ are defined in D and $B[v]$, $B[w]$ are defined on ∂D. Then there exist constants $k_1, k_2 \geq 0$ such that

$$\|v-w\|_{\overline{D}} \leq k_1 \|F[v] - F[w]\|_D + k_2 \|B[v] - B[w]\|_{\partial D} \tag{2.4}$$

where $\|\cdot\|_X = \sup_X |\cdot|$.

An a priori bound on the magnitude of u follows immediately from (2.2), (2.3) and (2.4). Taking $v = u$ and $w = 0$, we get

$$\|u\|_{\bar{D}} \leq k_1 \|r\|_D + k_2 \|s\|_{\partial D} \tag{2.5}$$

Furthermore the solution u is unique since if both v and w satisfy (2.2) and (2.3) if follows immediately from (2.4) that $v = w$ on $\bar{D}$.

Finally we construct a finite grid $\bar{D}_n$ over $\bar{D}$, with a total of n points such that for each point $x \in \bar{D}$ there exists some point $y \in \bar{D}_n$ with $\|x - y\| \leq h(n)$, where the distance $h(n) \to 0$ as $n \to \infty$. For example with a uniform grid we have $h(n) \sim n^{-1/\ell}$. We will denote the points of $\bar{D}_n$ in D by $D_n = \bar{D}_n \cap D$ and those in ∂D by $\partial D_n = \bar{D}_n \cap \partial D$.

In order to approximate the solution $u(x)$ to (2.2) and (2.3) we consider a generalized polynomial of specified functions $\varphi_i(x)$,

$$v_m = v_m(\alpha, x) = \sum_{i=1}^{m} \alpha_i \varphi_i(x) \tag{2.6}$$

We make the assumption that each function $\varphi_i(x)$ and its derivatives satisfy a certain Lipschitz condition. Given any two positive constants λ_1 and λ_2, we say that $v \in \Lambda(\lambda_1, \lambda_2)$ if v is continuous in $\bar{D}$ and

$$\begin{aligned} |F[v](x_1) - F[v](x_2)| &\leq \lambda_1 \|x_1 - x_2\|, \ \forall x_1, x_2 \in D \\ |B[v](x_1) - B[v](x_2)| &\leq \lambda_2 \|x_1 - x_2\|, \ \forall x_1, x_2 \in \partial D \end{aligned} \tag{2.7}$$

We observe from (2.1), (2.2) and (2.3) that $u \in \Lambda(\lambda_0, \lambda_0)$. We assume that each function φ_i, $i = 1, \cdots m$, has been normalized so that $\varphi_i \in \Lambda(1, 1)$.

We now choose constants $\lambda_1, \lambda_2 > 0$, and wish to impose conditions (depending on m) on the coefficients α_i of the generalized polynomial so that for each fixed m we have $v_m \in \Lambda(\lambda_1, \lambda_2)$. Let $\alpha \in R_m$ denote the vector with components α_i. We consider a compact, convex set $\Omega_m = \Omega_m(\lambda_1, \lambda_2) \subset R_m$, such that

$$\alpha \in \Omega_m(\lambda_1, \lambda_2) \Longrightarrow v_m \in \Lambda(\lambda_1, \lambda_2) \tag{2.8}$$

Lemma

For F and B linear, a suitable choice for $\Omega_m(\lambda_1, \lambda_2)$ is given by

$$\Omega_m = \{\alpha \Big| \sum_{i=1}^{m} |\alpha_i| \leq \hat{\lambda}\}, \quad \hat{\lambda} = \min\{\lambda_1, \lambda_2\} \tag{2.9}$$

Proof:

By linearity $F[v_m] = \sum_{i=1}^{m} \alpha_i F[\varphi_i]$. Also since $\varphi_i \in \Lambda(1, 1)$, we have $|F[\varphi_i](x_1) - F[\varphi_i](x_2)| \leq \|x_1 - x_2\|$. Then if $\alpha \in \Omega_m$,

$$|F[v_m](x_1) - F[v_m](x_2)| \leq \sum_{i=1}^{m} |\alpha_i| \, |F[\varphi_i](x_1) - F[\varphi_i](x_2)|$$

$$\leq \|x_1 - x_2\| \sum_{i=1}^{m} |\alpha_i| \leq \hat{\lambda} \|x_1 - x_2\| \leq \lambda_1 \|x_1 - x_2\|$$

and similarly for $B[v_m]$. Therefore $\alpha \in \Omega_m \Longrightarrow v_m \in \Lambda(\lambda_1, \lambda_2)$. ∎

The set Ω_m will normally be defined by linear inequality constraints in R_m, so that it is normally a bounded polyhedral set.

The coefficients α_i in v_m are determined so as to minimize (over the compact, convex set Ω_m) the maximum error in $|F[v_m] - r|$ on D_n and $|B[v_m] - s|$ on ∂D_n. If F and B are linear this can usually be done by solving a single linear programming problem. If either F or B are nonlinear an iterative solution is required. In either case the following theorem gives an error bound on the approximate solution v_m.

Theorem 1.

Let $\lambda_1, \lambda_2 \geq \lambda_0$ be chosen and a corresponding set $\Omega_m(\lambda_1, \lambda_2)$ be determined. Let $n \geq m$ be selected, and let $v^*_{m,n}$ be a minimizing solution, with coefficient vector α^* and value $\delta_{m,n}$ which satisfies the relation

$$\begin{aligned} \delta_{m,n} &= k_1 \| F[v^*_{m,n}] - r \|_{D_n} + k_2 \| B[v^*_{m,n}] - s \|_{\partial D_n} \\ &\leq \min_{\alpha \in \Omega_m} k_1 \| F[v_m(\alpha)] - r \|_{D_n} + k_2 \| B[v_m(\alpha)] - s \|_{\partial D_n} \end{aligned} \tag{2.10}$$

Then

$$\| v^*_{m,n} - u \|_{\bar{D}} \leq \delta_{m,n} + 2(k_1\lambda_1 + k_2\lambda_2)\, h(n) \tag{2.11}$$

Proof.

For simplicity we let $F^*_{m,n} = F[v^*_{m,n}]$ and $B^*_{m,n} = B[v^*_{m,n}]$. Let $x_1 \in \bar{D}$ be a point at which $|F^*_{m,n} - r|$ attains its maximum, i.e., $|F^*_{m,n}(x_1) - r(x_1)| = \| F^*_{m,n} - r \|_D$. Also let $x_2 \in \partial D$ be a point at which

$|B^*_{m,n} - s|$ attains its maximum, i.e., $|B^*_{m,n}(x_2) - s(x_2)| = \|B^*_{m,n} - s\|_{\partial D}$. By the construction of the finite grid $\overline{D}_n$ there exists $y_1 \in D_n$ and $y_2 \in \partial D_n$ such that $\|x_1 - y_1\|, \|x_2 - y_2\| \le h(n)$. Since $F[u] = r$ on D and $B[u] = s$ on ∂D we have from (2.4), (2.1), (2.8), (2.7) and (2.10),

$$\|v^*_{m,n} - u\|_{\overline{D}} \le k_1 \|F^*_{m,n} - r\|_D + k_2 \|B^*_{m,n} - s\|_{\partial D}$$

$$= k_1 |F^*_{m,n}(x_1) - r(x_1)| + k_2 |B^*_{m,n}(x_2) - s(x_2)|$$

$$\le k_1 |F^*_{m,n}(y_1) - r(y_1)| + 2k_1\lambda_1 \|x_1 - y_1\| + k_2 |B^*_{m,n}(y_2) - s(y_2)| + 2k_2\lambda_2 \|x_2 - y_2\|$$

$$= k_1 \|F^*_{m,n} - r\|_{D_n} + 2k_1\lambda_1 h(n) + k_2 \|F^*_{m,n} - s\|_{\partial D_n} + 2k_2\lambda_2 h(n)$$

$$\le \delta_{m,n} + 2(k_1\lambda_1 + k_2\lambda_2)\, h(n) \qquad \blacksquare$$

We note that the minimization problem (2.10) gives the optimal set of coefficients α_i^*, $i = 1, \ldots m$ and the error bound term $\delta_{m,n}$. The numbers $k_1\lambda_1$ and $k_2\lambda_2$ are known, and $h(n)$ is also known as a function of the number n of grid points. Thus both the approximate solution $v^*_{m,n}$ and its error bound (2.11) are given by the solution to (2.10).

We now consider the question of convergence of the approximation v_m to u as $m \to \infty$. Such convergence can only occur when the approximating functions φ_i are properly chosen. We certainly require that the solution u can be approximated arbitrarily well by some linear combination of the functions φ_i, $i = 1, 2, \ldots$. Since we are determining the coefficients by minimizing the error in the differential equation we also require that the functions

simultaneously approximate those derivatives of u which occur in F and B. We therefore require the simultaneous approximation on $\overline{D}$ of u and its derivatives (up to the highest occurring in F and B) by some linear combination of the functions φ_i. A suitable such basis, for example, might consist of polynomials in the components of x. Let u_m denote an approximation of the form (2.6) with the convergence property stated above. Then there exist coefficients of u_m, say $\overline{\alpha}_{m,i}$, such that

$$\lim_{m \to \infty} \{ \| u_m - u \|_{\overline{D}} + \| F[u_m] - r \|_D + \| B[u_m] - s \|_{\partial D} \} = 0 \qquad (2.12)$$

Since we do not know u we cannot actually find the coefficients $\overline{\alpha}_{m,i}$. However we can usually determine bounds on the $\overline{\alpha}_{m,i}$ in terms of the uniform bound (2.5) on u. We denote by $\overline{\alpha}_m \in R_m$ the vector with coefficients $\overline{\alpha}_{m,i}$, and by $\alpha^* \in R_m$ the vector which attains a minimizing solution $v^*_{m,n}$ in (2.10). The convergence of $v^*_{m,n}$ to u as $m,n \to \infty$ will now be given.

Theorem 2.

For each m, choose $n \geq m$. Assume that there exist positive constants $\lambda_1, \lambda_2 \geq \lambda_0$, such that for each m we can find a compact, convex set $\Omega_m(\lambda_1, \lambda_2) \subseteq R_m$ with $\overline{\alpha}_m \in \Omega_m(\lambda_1, \lambda_2)$. Then $v^*_{m,n}$ converges uniformly on $\overline{D}$ to u as $m,n \to \infty$.

Proof.

Given any $\varepsilon > 0$, there exists by (2.12) an m_1 such that for $m \geq m_1$, we have

$$k_1 \| F[u_m] - r \|_D + k_2 \| B[u_m] - s \|_{\partial D} \leq \varepsilon$$

Since $h(n) \to 0$ as $n \to \infty$, we can choose n_1 such that $h(n) \leq \varepsilon$ for $n \geq n_1$. Since $\overline{D}_n \subset \overline{D}$ we have $\| F[u_m] - r \|_{D_n} \leq \| F[u_m] - r \|_D$ and $\| B[u_m] - s \|_{\partial D_n} \leq \| B[u_m] - s \|_{\partial D}$. Since $\overline{\alpha}_m \in \Omega_m$, it follows from (2.10) that $\| F^*_{m,n} - r \|_{D_n} \leq \| F[u_m] - r \|_{D_n}$ and $\| B^*_{m,n} - s \|_{\partial D_n} \leq \| B[u_m] - s \|_{\partial D_n}$. Therefore

$$\begin{aligned} \delta_{m,n} &= k_1 \| F^*_{m,n} - r \|_{D_n} + k_2 \| B^*_{m,n} - s \|_{\partial D_n} \\ &\leq k_1 \| F[u_m] - r \|_{D_n} + k_2 \| F[u_m] - s \|_{\partial D_n} \\ &\leq k_1 \| F[u_m] - r \|_D + k_2 \| B[u_m] - s \|_{\partial D} \leq \varepsilon \end{aligned}$$

Then from (2.11) for $m \geq m_1$ and $n \geq n_1$,

$$\| v^*_{m,n} - u \|_{\overline{D}} \leq [1 + 2(k_1\lambda_1 + k_2\lambda_2)] \varepsilon \qquad \blacksquare$$

The convergence of the approximation $v^*_{m,n}$ is of considerable theoretical importance, and also will help to determine a suitable choice of functions for the φ_i. However as a practical matter the approximation $v^*_{m,n}$ is always obtained for some fixed values of m and n, so that the error bound given by (2.11) is the most important information available

in addition to the approximation itself. Since the constants λ_0, k_1 and k_2 are determined by the problem, the parameters to be chosen which affect the bound are m, n, λ_1 and λ_2, in addition to the choice of functions φ_i. The computer time required for a given problem will depend primarily on m and to a lesser extent on n. The choice of m and n will therefore usually be determined by computer time limitations. The choice for the remaining parameters λ_1 and λ_2 should be made to give the best error bound. In order to reduce $k_1\lambda_1$ and $k_2\lambda_2$ we want λ_1 and λ_2 as small as possible. As seen from the lemma for the linear case this will in general decrease the set Ω_m over which the minimization takes place, thereby increasing the value $\delta_{m,n}$. The constraints on the coefficients are imposed to reduce the variation in v_m and its derivatives between the points of $\bar{D}_n$, but this is done in general by making it more difficult to fit the differential equation and boundary conditions at the points of $\bar{D}_n$. Thus a proper choice of λ_1 and λ_2 is important in order to balance these two competing factors in the error bound.

3. QUASILINEAR ELLIPTIC PROBLEM

The general theory and method described in the previous section will now be applied to the case where F is a quasilinear elliptic operator of the form

$$F = L + g \tag{3.1}$$

and B is the identity operator. L is a linear elliptic differential operator in $D \subset R_\ell$ and g is a (possibly nonlinear) mapping $g:u \rightarrow g[u]$ in D, that is, $g[u](x) = g(x, u(x))$. Specifically L is given by

$$L[u] = \sum_{i,j=1}^{\ell} - a_{ij}(x) \frac{\partial^2 u}{\partial x_i \partial x_j} + \sum_{i=1}^{\ell} b_i(x) \frac{\partial u}{\partial x_i} \tag{3.2}$$

where the $\ell \times \ell$ matrix of coefficients a_{ij} is positive definite for all $x \in D$. Assumptions on g will be discussed below.

It is well known (for example, see [2, 14]) that the elliptic operator L has the following monotone property on $\overline{D}$

$$\left.\begin{array}{l} L[v] \geq L[w] \quad \text{in} \quad D \\ v \geq w \quad \text{on} \quad \partial D \end{array}\right\} \Longrightarrow \quad v \geq w \quad \text{on} \quad \overline{D} \tag{3.3}$$

As shown by Collatz [2] if an operator F has this monotone property an approximate solution and error bound can be obtained. This can be done by finding two solutions v and w of the form (2.6) such that

$$\|v - w\|_{\bar{D}} = \min$$

$$\left.\begin{aligned} F[v] - r &\geq 0 \\ F[w] - r &\leq 0 \end{aligned}\right\} \quad \text{in} \quad D$$

$$\left.\begin{aligned} v - s &\geq 0 \\ w - s &\leq 0 \end{aligned}\right\} \quad \text{on} \quad \partial D \tag{3.4}$$

It then follows directly from (2.2), (2.3) and (3.3) that

$$v \geq u \geq w \quad \text{on} \quad \bar{D} \tag{3.5}$$

There are several computational difficulties with this direct use of the monotone property. First, it may be difficult to insure that the constraints of (3.4) are satisfied at all points of $\bar{D}$. Second, there are $2m$ coefficients to be determined, m for v and m for w. Third, if v and w are large relative to $\|v - w\|_{\bar{D}}$, numerical difficulties may be encountered which may even give $w > v$ at some point of $\bar{D}$.

The approach to be described avoids the difficulties mentioned above, and also gives an error bound in certain quasilinear cases where the monotone property is not known to hold directly. In particular, an error bound can be obtained even when $g' < 0$. The error bound given below is obtained using the maximum principle for the linear differential operator $F = L + g$, where $g[u] = p(x)u$, and $p(x)$ may be negative. This maximum principle is given by Protter and Weinberger [7] in their Theorem 10, and is used to get bounds on u of the type (3.5), in their Theorem 13. For our purposes this principle is most conveniently stated as a minimum principle.

Theorem 3.

Let σ satisfy

$$(L + p)[\sigma] \geq 0 \quad \text{in} \quad D \tag{3.6}$$

If there exists a function $\rho(x) > 0$ on $\overline{D}$ such that

$$(L + p)[\rho] \geq 0 \quad \text{in} \quad D \tag{3.7}$$

and if

$$\min_{x \in \overline{D}} \left(\frac{\sigma}{\rho}\right) \leq 0$$

then this minimum is attained on ∂D . ∎

It follows that the existence of such a function ρ is sufficient for $F = L + p$ to have a monotone property.

Corollary

If there exists a positive ρ on $\overline{D}$ which satisfies (3.7), then

$$\left.\begin{array}{lll} (L + p)[\sigma] \geq 0 & \text{in} & D \\ \sigma \geq 0 & \text{on} & \partial D \end{array}\right\} \Longrightarrow \quad \sigma \geq 0 \text{ on } \overline{D} \tag{3.8}$$

Proof.

Since $\sigma \geq 0$ and $\rho > 0$ on ∂D, we have $\sigma/\rho \geq 0$ on ∂D. If there were a point in D at which σ/ρ were negative this would contradict the fact that σ/ρ attains its minimum on ∂D. Therefore $\sigma/\rho \geq 0$ in D, which requires that $\sigma \geq 0$. ∎

We first give the error bound for this linear case $F = L + p$, where we assume that p is bounded below, and let $\hat{p} = \inf_D \{0, p(x)\} \leq 0$. Consider the problem of finding $\mu = \mu(x)$ such that

$$\gamma = \min$$

Subject to:

$$\begin{aligned} (L + p)[\mu] &\geq 1 \quad \text{in} \quad D \\ \gamma \geq \mu &\geq 0 \quad \text{on} \quad \overline{D} \end{aligned} \tag{3.9}$$

Theorem 4.

Let v be an approximate solution such that

$$\begin{aligned} \| (L + p)[v] - r \|_D &\leq \delta_1 \\ \| v - s \|_{\partial D} &\leq \delta_2 \end{aligned} \tag{3.10}$$

Then if a feasible solution exists to the constraints (3.9), we have

$$|v - u| \leq \rho \tag{3.11}$$

where

$$\rho = \mu^* \delta_1 + (1 - \hat{p}\mu^*)\, \delta_2 \tag{3.12}$$

and μ^* is an optimal solution to (3.9).

For $p(x) \geq 0$ on D a solution to (3.9) always exists, with $\|\mu^*\|_{\overline{D}} \leq \|\mu_1\|_{\overline{D}}$, where μ_1 solves $L[\mu] = 1$ in D and $\mu = 0$ on ∂D.

Proof

The solution μ_1 is nonnegative on $\overline{D}$ by the monotone property of L, and therefore satisfies the constraints of (3.9) for $p(x) \geq 0$. Since γ is bounded below, the existence of a function satisfying the constraints implies the existence of an optimal solution.

Now suppose that μ solves (3.9). Let $\delta_1 = \delta_2 = 1$ in (3.12). Then $\rho = 1 + (1 - \hat{p})\mu$ is positive on $\overline{D}$. Furthermore

$$(L + p)[\rho] = (1 - \hat{p})(L + p)[\mu] + p \geq 1 - \hat{p} + p > 0 \text{ in } D .$$

It therefore follows from the corollary that $L + p$ has the monotone property (3.8).

Observe that from (3.10) we have $-\delta_1 \leq (L + p)[v] - (L + p)[u] \leq \delta_1$ on D and $-\delta_2 \leq v - u \leq \delta_2$ on ∂D. Consider $\eta = u - v + \rho$. Then from (3.12)

$$\eta = u - v + \rho \geq -\delta_2 + \delta_2 \geq 0 \quad \text{on} \quad \partial D$$

Also

$$(L + p)[\eta] = (L + p)[u] - (L + p)[v] + (\delta_1 - \hat{p}\delta_2)(L + p)[\mu^*] + p\delta_2$$
$$\geq -\delta_1 + \delta_1 + (p - \hat{p})\,\delta_2 \geq 0 \quad \text{in } D$$

Then by the nonotone property of $L + p$ we have $\eta \geq 0$ on $\overline{D}$, which gives $v - \rho \leq u$. In the same way with $\eta = v + \rho - u$ we get $u \leq v + \rho$.

■

It should be noted that if p is negative and $|p|$ sufficiently large, no error bound can be obtained, since there will be no feasible solution to the constraints of (3.9). Specifically, it can be shown that no solution exists to (3.9) if $p(x) < w_1$ in D, where $w_1 < 0$ is the maximum eigenvalue of $(L + w)[\mu] = 0$ in D, $\mu = 0$ on ∂D.

We are now able to obtain a similar result for the important case where g may be nonlinear. We assume that we have an approximate solution v such that

$$\begin{aligned} \| (L + g)[v] - r \|_D &\leq \delta_1 \\ \| v - s \|_{\partial D} &\leq \delta_2 \end{aligned} \tag{3.13}$$

We assume that g has a continuous derivative with respect to u which is uniformly bounded for $x \in D$, and we let $g'[u](x) = \frac{\partial}{\partial u} g(u, x)$. Given the function $v(x)$ we can readily determine a lower bounding function $p(\xi, v, x)$ and $\hat{p}(\xi, v) \leq 0$, such that

$$p(\xi, v, x) = \min_{|v(x)-\eta| \leq \xi} g'[\eta](x) \geq \hat{p}(\xi, v) \quad \text{in} \quad D \tag{3.14}$$

For notational convenience we will use $p(\xi, v)$ or just $p(\xi)$ to represent $p(\xi, v, x)$, and $\hat{p}(\xi)$ to represent $\hat{p}(\xi, v)$, when no confusion results.

Theorem 5

If there exists a positive constant ν and a function $\mu(x) \geq 0$ on $\bar{D}$ which satisfies

$$(L + p(\nu, v))[\mu] \geq 1 \qquad \text{in} \quad D \tag{3.15}$$

and

$$\delta_2 + (\delta_1 - \hat{p}(\nu, v)\,\delta_2)\mu \leq \nu \qquad \text{on} \quad \bar{D} \tag{3.16}$$

Then

$$|v - u| \leq \rho \qquad \text{on} \quad \bar{D} \tag{3.17}$$

where

$$\rho = \delta_2 + (\delta_1 - \hat{p}(\nu, v)\delta_2)\mu \tag{3.18}$$

For $g'[\cdot] \geq 0$, a solution μ to (3.15) and (3.16) always exists and we have $\rho = \delta_2 + \delta_1\mu \leq \delta_2 + \delta_1\mu_1$.

Proof:

We have

$$g[v] - g[u] = g'[\bar{v}](v - u), \quad |\bar{v} - v| \leq |v - u| \quad \text{in} \quad D \tag{3.19}$$

Then

$$\begin{aligned}(L + g'[\bar{v}])[v + \rho - u] &= (L + g)[v] - (L + g)[u] + (L + g'[\bar{v}])[\rho] \\ &= (L + g)[v] - r + (L + g'[\bar{v}])[\rho] \quad \text{in} \quad D\end{aligned}$$

Since $p(\nu) \leq 0$, we have $\rho \geq 0$. We temporarily make the additional assumption that $g'[\cdot](x) \geq p(\nu, v, x)$, which gives $(L + g'[\bar{v}])[\rho] \geq (L + p(\nu))[\rho]$. From (3.18) and (3.15)

$$(L + p(\nu))[\rho] \geq p(\nu)\,\delta_2 + \delta_1 - \hat{p}(\nu)\,\delta_2 \geq \delta_1 \quad \text{in} \quad D$$

Then using (3.13) we have in D

$$(L+g'[\bar{v}])[v+\rho-u] \geq -\delta_1 + (L+g'[\bar{v}])[\rho] \geq -\delta_1 + (L+p(\nu))[\rho] \geq -\delta_1 + \delta_1 = 0 \tag{3.20}$$

On the boundary we have $\rho \geq \delta_2$, since $\mu \geq 0$ on $\bar{D}$. Therefore since $u = s$ on ∂D, by (3.13) we have

$$v + \rho - u \geq -\delta_2 + \delta_2 = 0 \quad \text{on} \quad \partial D \tag{3.21}$$

Now let $\delta_1 = \delta_2 = 1$ in (3.18) and consider $\rho = 1 + (1 - \hat{p}(\nu))\mu > 0$ on $\bar{D}$. We have

$$(L+g'[\bar{v}])[\rho] \geq (L+p(\nu))[\rho] \geq p(\nu) + 1 - \hat{p}(\nu) \geq 1 \quad \text{in} \quad D$$

Then by the corollary, $(L+g'[\bar{v}])$ has the monotone property (3.8). It therefore follows directly from (3.20) and (3.21) that $v+\rho-u \geq 0$ on $\bar{D}$, which gives one of the inequalities of (3.17). In a similar way we can show that $(L+g'[\bar{v}])[u+\rho-v] \geq 0$ in D and $u+\rho-v \geq 0$ on ∂D so that $u+\rho-v \geq 0$ on $\bar{D}$, which gives the other inequality of (3.17).

By (3.16) - (3.19), we have $|\bar{v}-v| \leq |v-u| \leq \rho \leq \nu$, so that $p(\nu, v, x)$ as given by (3.14) is in fact a lower bound for $g'[\bar{v}](x)$ without any additional assumptions. We see from (3.14) that $g'[\cdot] \geq 0$ gives $p(\nu) \geq 0$, and we may take $\hat{p} = 0$. Therefore the solution μ_1 of $L[\mu] = 1$ in D and $\mu = 0$ on ∂D is a feasible solution to (3.15) and (3.16) with $\nu = \delta_2 + \delta_1 \|\mu_1\|_D$. If $p(\nu) > 0$, a solution to (3.15) will exist in general with $\mu < \mu_1$, giving an improved bound ρ.

It is easy to see that Theorem 4 is a special case of Theorem 5. In the linear case with $g[u](x) = p(x)u$, we have $g'[\cdot] = p(x)$, so that (3.14) gives $p(\xi, x) = p(x)$, independent of ξ.

We also observe that in the linear case the solution μ to (3.9) is independent of v. Therefore $\rho \to 0$ as $\delta_1, \delta_2 \to 0$, provided (3.9) has a feasible solution. Thus the existence of a nonnegative solution to $(L + p)[\mu] \geq 1$ in D implies the uniqueness of the solution to the boundary value problem (1.1) with $g = pu$. Furthermore we can minimize the maximum value of the error bound ρ by choosing v so as to minimize $\gamma^* \delta_1 + (1 - \hat{p}\gamma^*)\delta_2$, a linear function of δ_1 and δ_2. The constant $\gamma^* = \|\mu\|_{\overline{D}}$ is the optimal value of γ in (3.9). The constants k_1 and k_2 in (2.4) for this case are therefore given by $k_1 = \gamma^*$ and $k_2 = 1 - \hat{p}\gamma^*$.

4. NUMERICAL METHOD

As discussed in Section 2, the numerical determination of v and ρ makes use of a finite grid $\overline{D}_n$. We only compute the error in the approximate solution v at the grid points $\overline{D}_n$. Furthermore, we wish to use this same grid for the calculation of the function μ which gives the error bound.

As in Theorem 1 we choose λ_1, λ_2 and determine a corresponding compact, convex set Ω_m. Recall that $\alpha \in \Omega_m$ implies that v_m as given by (2.6) satisfies a Lipschitz condition of the form (2.7). We also let μ be a linear combination of the same functions $\varphi_i(x)$ as are used for v_m. Specifically we let

$$\mu_m = \sum_{i=1}^{m} \beta_i \varphi_i(x) \tag{4.1}$$

and determine the coefficients β_i so as to satisfy (3.9) or (3.15) and and (3.16). If we require of the coefficient vector β that $\beta \in \Omega_m$, then μ_m satisfies the same Lipschitz condition as v_m.

We can always choose (by appropriate scaling) the functions $\varphi_i(x)$, $i = 1, \ldots m$, so that on $\overline{D}$

$$\begin{aligned} |L[\varphi_i](x_1) - L[\varphi_i](x_2)| &\leq \|x_1 - x_2\| \\ |\varphi_i(x_1) - \varphi_i(x_2)| &\leq \|x_1 - x_2\| \end{aligned} \tag{4.2}$$

For any positive constant λ we let

$$\Omega_m = \Omega_m(\lambda) = \{\alpha \mid \sum_{i=1}^{m} |\alpha_i| \leq \lambda\} \tag{4.3}$$

Considering the linear case first we let $\hat{q} = \sup_D |p(x)|$. Then $\alpha \in \Omega_m(\lambda)$ implies

$$\begin{aligned} &|(L+p)[v_m](x_1) - (L+p)[v_m](x_2)| \le (1+\hat{q})\lambda \|x_1 - x_2\| \quad \text{in } D \\ &|v_m(x_1) - v_m(x_2)| \le \lambda \|x_1 - x_2\| \quad \text{on } \partial D \end{aligned} \tag{4.4}$$

The function μ_m also satisfies the same conditions when $\beta \in \Omega_m$.

We now solve two similar linear programming problems on a finite grid $\overline{D}_n$ with a distance $h = h(n)$. It is convenient to allow the use of different grid sizes in D_n and ∂D_n. We therefore denote the distance in D_n by $h_1 = h_1(n)$ and in ∂D_n by $h_2 = h_2(n)$. First we solve

$$\min_{\beta, \gamma, \lambda} \quad \gamma$$

Subject to:

$$\begin{aligned} &(L + p)[\mu_m] - (1 + \hat{q})h_1\lambda \ge 1 \quad \text{on} \quad D_n \\ &\gamma - h_j\lambda \ge \mu_m \ge h_j\lambda, \quad j = 1 \text{ on } D_n, \quad j = 2 \text{ on } \partial D_n \\ &\sum_{i=1}^{m} |\beta_i| \le \lambda \end{aligned} \tag{4.5}$$

We denote by μ_m^* and γ^* the optimal function and bound obtained. Using the bound γ^*, we let $b_1 = \gamma^*$, $b_2 = 1 - \hat{p}\gamma^*$ and $b_3 = (1 + \hat{q})h_1\gamma^* + (1 - \hat{p}\gamma^*)h_2$, and solve

$$\min_{\alpha, \xi_1, \xi_2, \lambda} \quad b_1\xi_1 + b_2\xi_2 + b_3\lambda$$

Subject to:

$$\begin{aligned} \| (L + p)[v_m] - r \|_{D_n} &\leq \xi_1 \\ \| v_m - s \|_{\partial D_n} &\leq \xi_2 \\ \sum_{i=1}^{m} |\alpha_i| &\leq \lambda \end{aligned} \tag{4.6}$$

Let $\alpha = \alpha^*$ (and corresponding optimal approximation $v_m = v_m^*$), $\xi_1 = \xi_1^*$, $\xi_2 = \xi_2^*$ and $\lambda = \lambda^*$, be the optimal values thus obtained, and define

$$\delta_1^* = \xi_1^* + (1 + \hat{q}) h_1 \lambda^* + h_1 \lambda_0$$

$$\delta_2^* = \xi_2^* + h_2(\lambda_0 + \lambda^*)$$

Theorem 6

If the linear programming problem (4.5) has a feasible solution, then

$$|u - v_m^*| \leq \rho \quad \text{on} \quad \overline{D} \tag{4.7}$$

where

$$\rho = \delta_2^* + (\delta_1^* - \hat{p}\, \delta_2^*)\, \mu_m^* \tag{4.8}$$

Proof

Let λ_β be the optimum value of λ from (4.5). Since $\beta \in \Omega_m(\lambda_\beta)$, the extreme values of μ_m^* on $\overline{D}$ can differ from its extreme values on $\overline{D}_n$ by at most $h_1 \lambda_\beta$ on D and at most $h_2 \lambda_\beta$ on ∂D. Therefore by (4.5), $\gamma^* \geq \mu_m^* \geq 0$ on $\overline{D}$. By (4.4) and (4.5)

$$\inf_{D} (L+p)[\mu_m^*] \geq \min_{D_n} (L+p)[\mu_m^*] - (1+\hat{q})h_1\lambda_\beta \geq 1$$

so that μ_m^* satisfies (3.9). Since $\alpha \in \Omega_m(\lambda^*)$ and r, s satisfy the condition (2.1), we have from (4.6) that v_m^* satisfies the relations (3.10) with $\delta_1 = \delta_1^*$ and $\delta_2 = \delta_2^*$ as defined. Theorem 4 then gives the stated bound. ∎

In order to obtain a similar result for the quasilinear case we again choose a grid $\bar{D}_n$ and determine an approximate solution v_m and constants ξ_1^*, ξ_2^* and λ^* such that

$$\begin{aligned} \| (L+g)[v_m] - r \|_{D_n} &\leq \xi_1^* \\ \| v_m - s \|_{\partial D_n} &\leq \xi_2^* \\ \sum_{i=1}^{m} |\alpha_i| &\leq \lambda^* \end{aligned} \tag{4.9}$$

We also determine a bound $\hat{q}(\xi)$ such that

$$\hat{q}(\xi) = \hat{q}(\xi, v) \geq \| g'[\eta] \|_D \quad \text{for} \quad \| v - \eta \|_D \leq \xi \tag{4.10}$$

and define

$$\begin{aligned} \delta_1^*(\xi) &= \xi_1^* + [1 + \hat{q}(\xi, v_m)]h_1 \lambda^* + h_1\lambda_0 \\ \delta_2^* &= \xi_2^* + h_2(\lambda_0 + \lambda^*) \end{aligned}$$

Theorem 7

If there exist coefficients β_i and constants λ and ν such that

$$(L + p(\nu))[\mu_m] \geq 1 + (1 + \hat{q}(\nu))h_1\lambda \quad \text{on} \quad D_n \tag{4.11}$$

$$\left.\begin{aligned} \delta_2^* + (\delta_1^*(\nu) - \hat{p}(\nu)\delta_2^*)(\mu_m + h_j\lambda) \leq \nu \\ \mu_m \geq h_j\lambda \end{aligned}\right\} \quad \begin{aligned} j = 1 \ \text{on} \ D_n \\ j = 2 \ \text{on} \ \partial D_n \end{aligned} \tag{4.12}$$

$$\sum_{i=1}^{m} |\beta_i| \leq \lambda \tag{4.13}$$

then

$$\left.\begin{aligned} |v_m - u| \leq \rho \\ \rho = \delta_2^* + (\delta_1^* - p(\nu)\delta_2^*)\mu_m \leq \nu \end{aligned}\right\} \quad \text{on} \quad \overline{D} \tag{4.14}$$

Proof

Let λ_β now denote the optimum value of λ from (4.11) - (4.13). From the definitions (3.14) and (4.10) we have $\hat{q}(\nu, v_m) \geq \|p(\nu, v_m)\|_D$. Then since $\beta \in \Omega_m(\lambda_\beta)$, it follows from (4.11) that $\mu = \mu_m$ satisfies (3.15). Furthermore from (4.12) it follows that μ_m is nonnegative on $\overline{D}$ and satisfies (3.16) on $\overline{D}$ with $\delta_1 = \delta_1^*(\nu)$ and $\delta_2 = \delta_2^*$. Since $\alpha \in \Omega_m(\lambda^*)$ we have

$$\begin{aligned} |(L+g)[v_m](x_1) - (L+g)[v_m](x_2)| &\leq |L[v_m](x_1) - L[v_m](x_2)| + \\ &+ |g'[\overline{v}][v_m(x_1) - v_m(x_2)]| \leq \lambda^* \|x_1 - x_2\| + \\ &+ \hat{q}(\nu)\lambda^* \|x_1 - x_2\| \leq [1 + \hat{q}(\nu)]\lambda^* h_1 \end{aligned}$$

for any point $x_1 \in D$ and x_2 a closest point in D_n, and $|\tilde{v} - v_m| \le |u - v_m| \le \rho \le \nu$ on D. It then follows from (2.1) and (4.9) that $v = v_m$, satisfies (3.13) with $\delta_1 = \delta_1^*(\nu)$ and $\delta_2 = \delta_2^*$. Then by Theorem 5, the bound (4.14) holds.

Returning to consideration of the linear case we see that if we can solve the linear programming problem (4.5) then we know that $L + p$ has the desired monotone property. Furthermore for a specified grid and corresponding distance d the optimum value of the Lipschitz constant λ_β in (4.5) is determined so as to minimize the upper bound γ^* of μ_m on $\bar{D}$. In a similar way in (4.6) the optimum values ξ_1^*, ξ_2^* and λ^* are determined so as to minimize the maximum value of the error bound, given by $\nu = \delta_2^* + (\delta_1^* - \hat{p}\,\delta_2^*)\gamma^*$. Thus for a selected grid $\bar{D}_n$ and number of functions m, we determine the approximate solution v_m so as to minimize the error bound in the uniform norm. This minimization determines the optimum balance between the interior and boundary error.

The solution time for a linear program of the form (4.5) or (4.6) depends primarily on m, and may increase as m^3. Thus it is considerably faster to solve the two sequential problems (4.5) and (4.6) than the single problem (3.4) with $2m$ coefficients. The accuracy of the error function μ_m will also usually be better than that given by $v - w$ in (3.4) since the inequalities (4.5) have been normalized to unity, avoiding the possibility of numerical difficulties if v and w are large compared to $v - w$. It should also be noted that the same value of m is used in (4.5) and (4.6) only as a matter

of convenience. In fact it may be desirable to increase the number of functions in (4.6) to obtain smaller values of ξ_1^* and ξ_2^*. This can be done using any value of γ^* obtained from (4.5).

Since we are attempting in (4.6) to minimize the maximum error in $(L + p)[v_m] - r$ and $v_m - s$, the approximate solution v_m will generally be determined so as to give a Chebyshev fit. That is, the error will oscillate between its maximum and minimum of equal magnitude over the grid points of $\bar{D}_n$. As a result, the average over D of the error in the differential equation will usually be smaller than ξ_1^*, so that the true error will usually be less than ρ as given by (4.8).

The quasilinear problem requires that we solve (4.9) and (4.11) - (4.13). This could be done directly as a nonlinear programming problem, but this may lead to difficulties because the problem will usually be nonconvex. A better method seems to be an iterative solution based on the linear problems (4.5) and (4.6).

We start with an initial coefficient vector α^o and corresponding approximation v_m^o, estimates for the initial error bound ν^o, and coefficients b_i^o, $i = 1, 2, 3$. We now describe the k^{th} cycle of the iteration procedure which starts with a known coefficient vector α^{k-1}, corresponding approximation v_m^{k-1} and function p^{k-1}, and known constants ν^{k-1} and b_i^{k-1}, $i = 1, 2, 3$.

1. Define the function

$$r^{k-1} = r + p^{k-1} v_m^{k-1} - g[v_m^{k-1}] \tag{4.15}$$

and solve (4.6) using b_i^{k-1}, p^{k-1} and r^{k-1}. The optimal solution gives α^k, v_m^k, ξ_2^k and λ^k.

2. Compute

$$\xi_1^k = \| (L + g)[v_m^k] - r \|_{D_n} \tag{4.16}$$

Also compute $p^k = p(\nu^{k-1}, v_m^k)$ and $\hat{p}^k = \hat{p}(\nu^{k-1}, v_m^k)$ according to (3.14), and $\hat{q}^k = \hat{q}(\nu^{k-1}, v_m^k)$ according to (4.10). Also compute

$$\begin{aligned} \delta_1^k &= \xi_1^k + (1 + \hat{q}^k) h_1 \lambda^k + h_1 \lambda_0 \\ \delta_2^k &= \xi_2^k + h_2(\lambda_0 + \lambda^k) \end{aligned} \tag{4.17}$$

3. Solve (4.5) with $p = p^k$ and $\hat{q} = \hat{q}^k$. The optimal solution gives μ_m^k and γ^k.

4. Compute

$$\begin{aligned} b_1^k &= \gamma^k, \quad b_2^k = 1 - \hat{p}^k \gamma^k \\ b_3^k &= (1 + \hat{q}^k) h_1 \gamma^k + (1 - \hat{p}^k \gamma^k) h_2 \\ \nu^k &= \delta_2^k + (\delta_1^k - \hat{p}^k \delta_2^k) \gamma^k \end{aligned} \tag{4.18}$$

This completes the k^{th} cycle.

Theorem 8

Consider the sequence $\{\nu^k\}$, $k = 1, 2, \cdots$, generated by the iteration procedure just described. If for any k we have $\nu^k \le \nu^{k-1}$, then the error bound (4.14) holds with $v_m = v_m^k$ and

$$\rho = \delta_2^k + (\delta_1^k(\nu^k) - \hat{p}(\nu^k)\delta_2^k)\mu_m^k \le \delta_2^k + (\delta_1^k - \hat{p}^k\delta_2^k)\mu_m^k \le \nu^k \qquad (4.19)$$

Proof

To show that (4.14) holds we demonstrate that (4.11) - (4.13) are satisfied. From (3.14) we see that for a fixed function v, $p(\xi, v)$ and $\hat{p}(\xi, v)$ are monotone nonincreasing functions of ξ. From (4.10) we see that $\hat{q}(\xi, v)$ is a monotone nondecreasing function of ξ. Therefore since $\nu^k \le \nu^{k-1}$, we have $p(\nu^k, v_m^k) \ge p(\nu^{k-1}, v_m^k) = p^k$, $\hat{p}(\nu^k, v_m^k) \ge \hat{p}(\nu^{k-1}, v_m^k) = \hat{p}^k$ and $\hat{q}(\nu^k, v_m^k) \le \hat{q}(\nu^{k-1}, v_m^k) = \hat{q}^k$. Since μ_m^k satisfies (4.5) with $p = p^k$ and $\hat{q} = \hat{q}^k$, we have $(L + p(\nu^k, v_m^k))[\mu_m^k] - [1 + \hat{q}(\nu^k, v_m^k)]h_1\lambda \ge [L + p^k][\mu_m^k] - [1 + \hat{q}^k]h_1\lambda \ge 1$, so that μ_m^k satisfies (4.11) with $\nu = \nu^k$. Furthermore let

$$\delta_1^k(\xi) = \xi_1^k + [1 + \hat{q}(\xi, v_m^k)]h_1\lambda^k + h_1\lambda_0$$

so that $\delta_1^k = \delta_1^k(\nu^{k-1})$. Then we have

$$\delta_1^k(\nu^k) - \hat{p}(\nu^k)\delta_2^k \le \delta_1^k(\nu^{k-1}) - \hat{p}(\nu^{k-1})\delta_2^k = \delta_1^k - \hat{p}^k\delta_2^k$$

so that

$$\delta_2^k + (\delta_1^k(\nu^k) - \hat{p}(\nu^k)\delta_2^k)\gamma^k \le \delta_2^k + (\delta_1^k - \hat{p}^k\delta_2^k)\gamma^k = \nu^k$$

Also from (4.5), $\gamma^k \ge \mu_m^k + h_j\lambda \ge 2h_j\lambda \ge 0$ on $\overline{D}_n$, so that (4.12) is satisfied. The relation (4.13) is satisfied since it also appears in (4.5). Then by Theorem 7, the bound (4.14) holds with ρ given by (4.19).

It follows from Theorem 8 that in order to get an error bound after the least number of iterations, the initial choice ν^0 should overestimate the actual error in the initial approximation v_m^0. In general, if such a choice is made we will have $\nu^1 \leq \nu^0$ so that an error bound will be available after a single cycle. The only difficulty which may be encountered is that no solution may exist to (4.5) if too large a value of ν^0 is selected. If no better values are known, reasonable starting values are $b_1^0 = b_2^0 = 1$ and $b_3^0 = h_1 + h_2$.

The solution of (4.6) using (4.15) is a linearization about v_m^{k-1} of $(L + g)[v_m]$, so that we may consider the iterative procedure as essentially Newton's method. Assuming a solution exists, the procedure can therefore be expected to converge quadratically once the approximation gets sufficiently close. Thus we can expect a sequence of approximations converging to an approximation which gives the smallest error bound with the selected functions. In general the sequence $\{\nu^k\}$ will decrease monotonically, so that we have an improved error bound at each iteration and can terminate the process when the possible improvement in accuracy no longer justifies the additional computation required. A useful criterion for convergence is the difference

$$w^k = \xi_1^k - \| (L + p^{k-1})[v_m^k] - r^{k-1} \|_{D_n}$$

If the sequence has converged we have $v_m^k = v_m^{k-1}$, so that $w^k = 0$.

This iterative procedure was used for the numerical solution of all the quasilinear problems described in Section 5. With one exception it was possible to choose an initial approximation v_m^0 and bound ν^0 so that an

error bound was obtained within several cycles. To get an improved bound the iteration was continued until $|w^k/\xi_1^k| \le 0.01$. In most cases this required no more than 5 cycles, and the maximum number required was 9 cycles.

It should also be noted that if $g'[\cdot] \ge 0$, the linear program (4.5) needs to be solved only once. Based on Theorem 5, this is done by setting $p = \hat{q} = 0$ in (4.5) so that the error function μ_m^1 and bound γ^1 obtained are independent of v_m. Only the iterative solution of (4.6) is then required. At each cycle we have the bound $\rho = \delta_2^k + \delta_1^k \mu_m^1 \le \delta_2^k + \delta_1^k \gamma^1$, where $\delta_1^k = \xi_1^k + h(\lambda_0 + \lambda^k)$. In general the error function μ_m^1 thus obtained will not be the best possible, but the possible improvement may not be worth the additional calculation required.

There are two important special cases in which one can simplify the calculation by a proper choice of functions φ_i. The first occurs when $s = 0$ and we can find suitable functions which vanish on ∂D. We then have $v_m = 0$ on ∂D so that the boundary error $\xi_2^* = 0$ for any choice of coefficients α_i. Since $\mu_m = 0$ on ∂D we also satisfy $0 \le \mu_m \le \gamma$ on ∂D. Therefore in effect we have $\partial D_n = \partial D$ which means that we can assume we have $h_2 = 0$. Thus the second inequalities in (4.6) and (4.9) are deleted and we set $h_2 = \delta_2^* = 0$ wherever they occur. In particular we have $\rho = \delta_1^* \mu_m^*$ in (4.8) and (4.14). Even if $s \neq 0$ we may still be able to treat the problem in a similar way. If a function $\varphi_0(x)$, continuous on $\overline{D}$, can be found such that

$\varphi_0 = s$ on ∂D and $\varphi_0 \in \Lambda(\lambda_0, \lambda_0)$, then we take $\varphi_0 + v_m$ as the approximate solution and minimize the error in $(L + g)[\varphi_0 + v_m] - r$ on D_n. Since $\varphi_0 + v_m = s$ on ∂D, the error on the boundary is identically zero as before.

In the second special case we can find functions that satisfy the differential equation exactly in D. This will generally be possible only in the linear case where we can find functions φ_i, $i = 0, 1, \cdots m$, such that $(L + p)[\varphi_0] = r$ and $(L + p)[\varphi_i] = 0$ in D, and a function μ_0 which satisfies the constraints (3.9). Then $\varphi_0 + v_m$ identically satisfies the differential equation in D so that $\xi_1^* = 0$ for any choice of coefficients α_i. In this case we set $h_1 = \delta_1^* = 0$, wherever they occur. In particular we now have $\rho = (1 - \hat{p}\mu_0)\delta_2^*$ in (4.8), and the minimization (4.6) is carried out only over the boundary points so that the first inequality is deleted.

We complete this section by describing the solution of the problems (4.5) and (4.6) as standard linear programming problems. We let J_1 denote the set of points in D_n and J_2 the set of points in ∂D_n, and assume that there are n_1 points in J_1 and n_2 points in J_2, with $n_1 + n_2 = n$. We define two column vectors, $\bar{r} \in R_{n_1}$ and $\bar{s} \in R_{n_2}$ with elements $r_j = r(x_j)$, $j \in J_1$ and $s_j = s(x_j)$, $j \in J_2$. We define an $m \times n_1$ matrix H, an $m \times n_1$ matrix G_1 and an $m \times n_2$ matrix G_2 with elements given by

$$(H)_{ij} = (L + p)[\varphi_i](x_j), \qquad i = 1, \cdots m, \quad j \in J_1$$

$$(G_k)_{ij} = \varphi_i(x_j), \qquad i = 1, \cdots m, \quad j \in J_k, \quad k = 1, 2$$

We let I_m denote the $m \times m$ identity matrix and $e_1 \in R_{n_1}$, $e_2 \in R_{n_2}$ and $e_m \in R_m$ denote column vectors with each element unity (sum vectors). The transpose of a vector or matrix will be denoted here by a prime so that e_i' is a row vector. A new vector $\theta \in R_m$ is also introduced.

Considering (4.6) first, we can write it in the form

$$\min_{\alpha, \xi_1, \xi_2, \theta} \quad b_1\xi_1 + b_2\xi_2 + b_3 \sum_{i=1}^{m} \theta_i$$

Subject to:

$$\begin{aligned} -e_1\xi_1 &\le H'\alpha - \bar{r} \le e_1\xi_1 \\ -e_2\xi_2 &\le G_2'\alpha - \bar{s} \le e_2\xi_2 \\ -\theta &\le \alpha \le \theta \end{aligned} \tag{4.20}$$

We consider this to be in the unsymmetric dual form $\min_y \{b'y \mid A'y \ge c\}$. Then by the duality theory of linear programming [13] the equivalent primal problem is given by $\max_z \{c'z \mid Az = b,\ z \ge 0\}$. If the minimum problem has a feasible solution and $b'y$ is bounded below then both problems have optimal solutions and $b'y^* = c'z^*$. If the primal problem has an infinite solution then there is no feasible solution to the dual. Furthermore if we denote by B the nonsingular primal optimal basis matrix then $y^* = (B^{-1})'\bar{c}$ gives the dual optimal vector, where $\bar{c}$ is a vector of the cost coefficients corresponding to the primal optimal basis columns. Thus we can obtain the desired optimal dual solution by solving the equivalent primal problem.

The dual problem (4.20) has a total of $2m+2$ variables (α, θ, ξ_1 and ξ_2). Corresponding to each dual variable is a primal equation. The size of the primal problem can be reduced to $m+2$ equations by eliminating the equations corresponding to θ, and imposing an upper bound on some of the primal variables. We then obtain the following bounded variable primal problem

$$\max_{z_1, z_2} \left\{ c_1' z_1 \,\middle|\, \begin{matrix} A_1 z_1 + A_2 z_2 = b \\ z_1 \geq 0,\ b_3 e_m \geq z_2 \geq 0 \end{matrix} \right\} \tag{4.21}$$

where

$$A_1 = \begin{pmatrix} H & G_2 & -H & -G_2 \\ e_1' & 0 & e_1' & 0 \\ 0 & e_2' & 0 & e_2' \end{pmatrix}, \quad A_2 = \begin{pmatrix} 2I_m \\ 0 \\ 0 \end{pmatrix}, \quad b = \begin{pmatrix} b_3 e_m \\ b_1 \\ b_2 \end{pmatrix} \in R_{m+2}$$

$$c_1' = (\bar{r}' \quad \bar{s}' \quad -\bar{r}' \quad -\bar{s}'), \quad z_1 \in R_{2n}, \quad z_2 \in R_m$$

A feasible solution to (4.20) always exists ($\alpha = \theta = 0$, and ξ_1, ξ_2 sufficiently large) and the objective function is bounded below by zero. Therefore it has an optimal solution, and by duality so does (4.21). The optimal solution to this primal problem gives the desired coefficient vector α as the first m elements of the dual vector $y \in R_{m+2}$, and the bounds ξ_1^* and ξ_2^* as the last 2 elements. The $m+2$ columns in the optimal basis will be selected from the columns of A_1 and A_2. Each column of A_1 in the basis corresponds to a point of $\bar{D}_n$ at which the maximum error is attained. A column from H corresponds to a point of D_n at which the error is $-\xi_1^*$; a column from $-H$ corresponds to a point of D_n at which

the error is ξ_1^*. Similarly a column from G_2 or $-G_2$ corresponds to a point of ∂D_n at which the error is $-\xi_2^*$ or ξ_2^*. Thus the optimal solution to (4.21) not only gives the desired approximate solution and error bound, but also the points in D_n at which the error in the differential equation is a maximum and the points in ∂D_n at which the error itself is a maximum. The linearized problem is solved in exactly the same way except that the vector $\bar{r}$ now has as its elements $r_j = r^{k-1}(x_j)$, $j \in J_1$, where r^{k-1} is given by (4.15).

It is also worth noting that the vectors $\bar{r}$ and $\bar{s}$ appear only in the cost row of (4.21). Thus a sequence of problems with different functions $r(x)$ and $s(x)$ can be solved rapidly using the multiple cost row or parametric cost row feature of a linear programming code.

In a similar manner we construct a primal problem of $m+2$ rows corresponding to (4.5) considered as the unsymmetric dual.

The following primal problem is thus obtained.

$$\max_{z_1, z_2, z_3} \left\{ c_1' z_1 \;\middle|\; \begin{array}{l} A_1 z_1 + A_2 z_2 + a_3 z_3 = b \\ z_1 \geq 0, \quad z_3 e_m \geq z_2 \geq 0 \end{array} \right\} \tag{4.22}$$

where

$$A_1 = \begin{pmatrix} H & G_1 & G_2 & -G_1 & -G_2 \\ \hat{d}_1' & d_1' & d_2' & d_1' & d_2' \\ 0 & 0 & 0 & e_1' & e_2' \end{pmatrix}, \quad A_2 = \begin{pmatrix} 2I_m \\ 0 \\ 0 \end{pmatrix}, \quad a_3 = \begin{pmatrix} -e_m \\ 1 \\ 0 \end{pmatrix}, \quad b = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \in R_{m+2}$$

$$c_1' = (e_1' \quad 0 \quad 0 \quad 0 \quad 0), \quad z_1 \in R_{2n+n_1}, \quad z_2 \in R_m, \quad z_3 \in R_1$$

and for convenience we have defined $d_1' = -h_1 e_1'$, $d_2' = -h_2 e_2'$ and $\hat{d}_1' = -(1 + \hat{q}) h_1 e_1'$.

It is easily shown that a feasible solution to (4.22) always exists. If (4.5) has no feasible solution then the solution to (4.22) will be infinite. If (4.22) has a finite optimal solution then so does (4.5) and the error bound is valid. By duality the optimal solution to (4.22) gives $\gamma^* = c_1' z_1^*$. The desired coefficient vector β is given by the first m elements of the dual vector y, while the last 2 elements give the optimal value for λ and γ^* .

Simplification in (4.21) and (4.22) occurs in either of the two special cases discussed earlier. When the boundary conditions are identically satisfied, the columns containing G_2 are deleted from the matrices A_1 in (4.21) and (4.22) and the corresponding elements of c_1 deleted. The last row of A_1 and A_2 and the last element of b are deleted in (4.21) and we set $h_2 = 0$ in b_3 . In the special linear case where functions identically satisfying the homogeneous differential equation and a function μ_0 satisfying (3.9) can be found, the columns of A_1 containing H and corresponding elements of c_1 are deleted in (4.21). The next to last row of A_1 and A_2 and the element b_1 in b are also deleted and we set $h_1 = 0$ in b_3 .

The solution of (4.22) is not needed in this case since μ_0 gives all the additional information required.

If the number n of points chosen is at least 2^{ℓ} times as large as the number of variables $m+2$, it has been found empirically that the maximum errors ξ_1^* and ξ_2^* computed on $\overline{D}_n$ are in fact very close to the maximum errors on $\overline{D}$. Thus the terms in δ_1^* and δ_2^* which contain h_1 or h_2 give an overestimate of the maximum errors on $\overline{D}$. This empirical knowledge can be used to improve the error bound ρ with only a small additional amount of computation. After solving (4.21) and (4.22) and obtaining the coefficients α_i, the maximum errors (say $\hat{\xi}_1$ and $\hat{\xi}_2$) over a finer grid (with smaller distances $\hat{h}_1$ and $\hat{h}_2$) are obtained by direct evaluation. We now get new values $\hat{\delta}_1$ and $\hat{\delta}_2$ with $\hat{\xi}_1$, $\hat{\xi}_2$, $\hat{h}_1$ and $\hat{h}_2$ replacing ξ_1^*, ξ_2^*, h_1 and h_2. In general, the reduced values of h_1 and h_2 (say by a factor of 2 or more) will give $\hat{\delta}_1 < \delta_1^*$ and $\hat{\delta}_2 < \delta_2^*$, even though ξ_1 and ξ_2 may have increased somewhat. These smaller values $\hat{\delta}_1$ and $\hat{\delta}_2$ are now used in ρ to give an improved error bound. This scheme was used to obtain the error bounds for the numerical problems described in the next section.

5. COMPUTATIONAL RESULTS

In order to test the computational aspects and efficiency of the linear programming method described in the previous section it has been applied to a number of linear and quasilinear problems. All of the cases discussed here are two-dimensional ($\ell = 2$, x and y as coordinates) with the negative Laplacian as the elliptic operator $L[u] = -\nabla^2 u = -(u_{xx} + u_{yy})$. In addition to linear cases, several different nonlinear functions g were used including $g(u) = -e^{-u}$, $g(u) = \pm e^{u}$ and $g(u) = -u^2$. The domains D considered were a square, a truncated square and an ellipse. Various approximating functions φ_i were used, including polynomials, trigonometric and harmonic functions. The dependence of the error on the number of functions (up to a maximum of 45) and on the type of function used was studied. In some cases the functions were chosen so as to identically satisfy a homogeneous boundary condition, while in others they were chosen to satisfy the differential equation. The more difficult situation where the functions do not identically satisfy either the equation or the boundary condition was also considered.

We will first discuss the results for the cases where $\overline{D}$ is the unit square in the first quadrant (lower left-hand corner at origin), with the boundary ∂D given by its four sides. We wish to solve

$$\begin{aligned} -\nabla^2 u + g(u) &= r \quad \text{in} \quad D \\ u &= 0 \quad \text{on} \quad \partial D \end{aligned} \tag{5.1}$$

By symmetry we need consider only a triangular domain with $1/8$ the total area. However the symmetry about the diagonal was not incorporated in the choice of functions, but used afterwards as an additional check on accuracy. The quarter size square domain was used with the requirement that $u_x = 0$ along $x = 0.5$ and $u_y = 0$ along $y = 0.5$ replacing the conditions along $x = 1$ and $y = 1$. For the majority of these cases, polynomials were chosen which satisfied the boundary conditions identically:

$$\varphi_i(x, y) = \frac{1}{8}\sigma_p(x)\,\sigma_q(y), \quad i = 1, \cdots 45; \quad p, q = 1, \cdots 9, \quad p+q \le 10 \qquad (5.2)$$
$$\sigma_p(x) = x^p\left(1 - \frac{2p}{p+1}x\right)$$

It is not difficult to show that for $0 \le x, y \le \frac{1}{2}$,

$$\left|\nabla^2\varphi_i(x + \Delta x, y + \Delta y) - \nabla^2\varphi_i(x, y)\right| \le (p+q)^3\, 2^{-(p+q+3)}\, \|(\Delta x, \Delta y)\|$$
$$\le \|(\Delta x, \Delta y)\|$$

so that the Lipschitz condition (4.2) holds.

Since the functions φ_i satisfy the zero boundary conditions exactly we have the first special case discussed in the previous section, with grid points required only in D. The domain D included the lines $x = 0.5$ and $y = 0$. A total of 225 points on a uniform grid in the interior of the square were used, so that $h_1 \cong 0.0167$. The simplest linear case $-\nabla^2 u = 1$ was solved first using 45 functions. For this case we have $g = 0$ and $r = 1$, so that $p = \hat{p} = \hat{q} = \lambda_0 = 0$. The error function μ_m^* and approximate solution v_m were obtained from (4.22) and (4.21). The approximate solution attains its

maximum value (as does the exact solution) at the center of the unit square, so that $\|v_m\| = \|v_m\|_{\bar{D}} = v_m(0.5, 0.5)$. For this case we have $\|v_m\| = 0.0737$. The corresponding bounds obtained were $\gamma^* = 0.074$ and $\delta_1 = 0.0031$, where δ_1 here represents the $\hat{\delta}_1$ as described at the end of Section 4. The error bound here is given by $\rho = \delta_1 \mu_m^* \leq \delta_1 \gamma^* = \nu = 2.3 \times 10^{-4}$, so that $0.0735 \leq u(.5, .5) \leq 0.0739$. For this simple case we can compare with a trigonometric series solution giving the more accurate result $u(.5, .5) = 0.07366$. The relative error ρ/v_m is essentially constant for this case and has a maximum value of 0.31%. This information is tabulated in the first line of Table 1. A normalized contour plot of v_m is given in Figure 1. Each contour represents one of the curves

$$v_m(x, y) = \frac{j}{10} \|v_m\| , \quad j = 1, 2, \cdots 9$$

The rest of Table 1 gives the results for two nonlinear functions g for which $g'[\cdot] \geq 0$, so that unique solutions always exist, $p(\nu) \geq 0$ and $\tilde{p} = 0$. The first such function is $g = -\tau e^{-u}$, where τ is a parameter and $r = 0$. Based on the discussion following Theorem 8 we took $\mu_m' = \mu_m^*$ (where μ_m^* is the optimal error function for the linear case) and solved the linear program (4.21) iteratively to obtain v_m and δ_1. Note that for $\tau = 1$ the solution differs only a little from the linear case, since $e^{-u} \cong 1$.

As a result the normalized contour plot for $\tau = 1$ is almost identical to Figure 1. Skipping for the moment to the second case $g = e^u$, we find that the increased value of $\|v_m\|$ is essentially a matter of scaling for $r = 5, 10$. This is confirmed by the rather surprising fact that the normalized contour plots for these two cases differ only slightly from Figure 1. Only 21 functions were used $(p + q \leq 7)$ so that the values of δ_1 and the relative error increased by almost a factor of 10. Starting with the linear problem solution as v_m^0, no more than 4 iterations of (4.21) were required for the above cases.

The nonlinear effect becomes evident for $g = -\tau e^{-u}$ with $\tau = 10$ and for $g = e^u$ with $r = 50, 100$. This is shown in Figures 2 and 3 as a movement of the normalized contours toward the boundaries as r increases. The normalized contour plot for $g = -10e^{-u}$ differs only slightly from Figure 2. The solutions for $r = 50, 100$ were obtained first with $m = 21$, based on the iterative procedure of Section 4 with (4.21) and (4.22) used to solve (4.6) and (4.5). The best solution for $r = 10$ was used as v_m^0 for $r = 50$. The best solution for $r = 50$ was then in turn used to start the iteration for $r = 100$. In each case 4 cycles were required. The more accurate solutions with $m = 45$ were obtained in 2 additional cycles starting with the best solutions for $m = 21$. Note that increasing m from 21 to 45 decreases the bound δ_1 by a factor of almost 10. It should also be noted that the bound γ^* is decreased because of the large positive value of $g'[v_m]$ in the central

region of D, so that the relative error bound is actually less than for the linear case with the same number of functions ($m = 45$).

The second set of cases solved on the unit square with zero boundary conditions were with $g = -u^2$, so that $g'[\cdot] \leq 0$. A total of 10 such cases were solved with values of $r = 1, 10, 80, 84$. The results obtained are given in Table 2. The first 6 cases ($r = 1$) in Table 2 show the convergence as, m is increased, of the approximate solution v_m to the exact solution u. The error function and bound γ^* were obtained with $m = 10$ using the iterative procedure. The same γ^* is valid for the remaining 5 cases with larger m. In each of these cases the previous v_m was used to begin the iteration for the next larger value of m. Only one or two iterations were required in each case. We see that very roughly $\delta_1 \cong 10/m^2$, so that convergence is relatively rapid. A numerical solution to this problem using a finite difference method is quoted by Collatz [3]. The most accurate solution (mesh size = 0.025) required 91 iterations and gave $v(0.4, 0.4) = 0.0689$. The corresponding value obtained here using 45 functions was 0.0690.

The solution for $r = 10$ behaved as expected, but as r was increased to $r = 84$ more iterations were required and both δ_1 and γ^* increased as well as $\|v_m\|$. For $r = 84$ with $m = 21$ no solution could be obtained with $\nu^k \leq \nu^{k-1}$. The sequence $\{\nu^k\}$ continued to increase until no solution to (4.5) could be obtained. However by setting $\nu^k = 0$, convergence of (4.6)

using (4.15) was obtained, giving an approximate solution v_m in 9 iterations. The corresponding value $\delta_1 = 2.29$ was obtained, but no error bound could be computed. The v_m thus obtained was used to start an iterative solution with $m = 45$. An improved solution with $\delta_1 = 0.26$ allowed the error bound $\gamma^* = 1.46$ to be obtained in 3 additional cycles. The normalized contour plot for $r = 84$ is given in Figure 4. The increase in magnitude of $\|v_m\|$ is as expected. The more significant increase occurs in γ^*, which is 73 times larger than it is for $r = 100$ in Table 1. This increased magnitude shows that we are close to a function $p(\nu, v_m, x)$, as given by (3.14), for which no nonnegative solution exists to (3.15). For $r = 84$ we have $\hat{p}(\nu, v_m) = -2(\|v_m\| + \nu) = -25.4$. For comparison we note that the maximum eigenvalue of $(-\nabla^2 + w)[\mu] = 0$, is $w_1 = -2\pi^2 \cong -19.739$. When r was increased to $r = 85$, no solution could be obtained. This was reflected in the fact that the iterative solution of (4.6) using (4.15) diverged when started with v_m for $r = 84$.

Similar behavior was observed with $g = -\tau e^u$, for which we also have $g'[\cdot] \leq 0$. A total of 5 solutions and error bounds were obtained for this case using the values $\tau = 1, 5, 6.7, 6.8$. The results are summarized in Table 3. In this case also, more iterations were required and both δ_1 and γ^* increased as τ was increased to its maximum value. The value of γ^* obtained here for $\tau = 6.8$ is even larger than for $r = 84$ in Table 2. The interesting result was obtained that the normalized contour plot for

$\tau = 6.8$ differs only slightly from Figure 4. For this case we have $\hat{p}(\nu, v_m) = -\tau \exp(\|v_m\| + \nu) = -27.1$, so that a solution to the inequalities (4.11), (4.12) and (4.13) exists with $p(\nu, v_m, x)$ taking on a more negative value than it did for $g = -u^2$ with $r = 84$. This can be explained by observing that $p(\nu, v_m, x)$ has a steeper valley in the center for the exponential case as compared with the quadratic case, so that the average over D in some sense is the same. Once again no solution was obtained for larger τ, as reflected in no convergence for $\tau = 6.81$.

As mentioned in the Introduction, it can be shown theoretically [6] that no solution exists when $g = -u^2$ and $r > \hat{r} = \pi^4 \cong 98$, or when $g = -\tau e^u$ and $\tau > \hat{\tau} = 2\pi^2/e \cong 7.2$. On the basis of these computational results it appears that better bounds are given by $\hat{r} = 85$ and $\hat{\tau} = 6.81$.

As an illustration of a case in which the differential equation was satisfied exactly in D, the equation $-\nabla^2 u = 1$ in D, $u = 0$ on ∂D was solved on an ellipsoidal domain. The problem was proposed and solved by Collatz (see [2], pp. 391-2). The domain considered has its center at the origin and has a boundary ∂D consisting of the two straight-line segments $|x| \leq 1$, $y = \pm 1$, connected by two semicircles of radius 1 about the points $x = \pm 1$, $y = 0$. By symmetry only the quarter of the domain contained in the first quadrant need be considered. The function $\varphi_0 = -0.25(x^2 + y^2)$ satisfies $-\nabla^2 \varphi_0 = 1$. The functions $\varphi_i = R_e(x+iy)^{2i-2}$, $i = 1, 2, \cdots m$, satisfy $-\nabla^2 \varphi_i = 0$. Therefore $\varphi_0 + v_m$ satisfies $-\nabla^2 u = 1$ identically in D.

The error bound for this case is given by $\rho = \nu = \hat{\delta}_2$, so that we use a simplified version of (4.21) to minimize the boundary error. A total of 100 boundary points in the first quadrant were used. The results obtained as a function of m are summarized in Table 4. The first result ($m = 6$) corresponds almost exactly with that obtained by Collatz. It appears that the actual error in v_m at the origin is smaller (by a factor of approximately 10) than it is on the boundary.

The final set of results obtained were for the general quasilinear problem where neither the boundary conditions nor the differential equation could be exactly satisfied. A total of 6 such cases are presented in Table 5. For the first 3 cases the domain was again the unit square, but some functions were used which did not vanish on the boundaries. In particular, linear combinations of the following harmonic and trigonometric functions were used.

$$\Psi_0 = 0.25[0.5 - (x - 0.5)^2 - (y - 0.5)^2]$$

$$\Psi_j = \begin{cases} \mathrm{Im}[(x+iy)^j + (y+ix)^j], & j = 1, 2, 3, 5, \cdots \\ \mathrm{Re}[(x+iy)^j + (y+ix)^j], & j = 4, 8, \cdots \end{cases} \qquad (5.3)$$

$$\Psi_{jk} = \sin j\pi x \sin k\pi y + \sin k\pi x \sin j\pi y, \quad j, k \text{ odd}$$

Note that Ψ_0 and the Ψ_{jk} vanish on the boundaries while the Ψ_j do not. We also have $-\nabla^2\Psi_0 = 1$ and $\nabla^2\Psi_j = 0$, $j \geq 1$. Furthermore, all functions are symmetric about the line $y = x$. The results were obtained by the iterative procedure using (4.21) and (4.22), to get δ_1, δ_2 and γ^* as given. Since

both δ_1 and δ_2 are nonzero for these cases we have $v = \gamma^* \delta_1 + (1 - \hat{p}\gamma^*)\delta_2$. The value of m gives the total number of functions used. We can compare the first 3 cases with the results given in Tables 1 and 3 for the same cases using different functions. We see that the approximate solution is essentially the same, but that the error bounds have increased somewhat. It therefore seems best to use functions which satisfy the boundary conditions when this is possible.

The domain for the last 3 cases was a truncated unit square, with the region of the square deleted for which $x + y > 1.5$. The domain and a typical normalized contour plot are shown in Figure 5. The same functions (5.3) were again used and m represents the total number used. In this case, of course, none of them vanish along the boundary segment $x + y = 1.5$. Since we require the solution to vanish closer to the point $x = y = 0.5$, it is to be expected that $v_m(0.5, 0.5)$ will be smaller than it is in the corresponding case on the full square. As seen from the contour plot the maximum value $\|v_m\|$ is no longer attained at $x = y = 0.5$, but close to $x = y = 0.45$. We also see that the required minimization over both D_n and ∂D_n has caused a significant increase in the value of the error bound. However the approximate solutions obtained are probably better than is indicated by the error bounds. While the normalized contour plot for the last of these cases is given in Figure 5, the normalized plots for the other 2 cases differ only slightly from Figure 5.

A total of 180 points were used in the interior and on the boundary of half the truncated square, taking advantage of the symmetry about the line $y = x$. Since reasonably good initial approximations were known, at most 5 cycles were required in any of these cases. These numerical solutions were obtained using a standard linear programming code which was modified to carry out the iterative solution more efficiently. The problems were run on the University of Wisconsin CDC 3600 computer and required from 10 to 30 seconds per iteration for solution. The contour plots were generated by an auxiliary routine directly from the approximation v_m, and plotted automatically on a Calcomp plotter.

REFERENCES

1. Collatz, L., "Approximation in partial differential equations," On Numerical Approximation (R. E. Langer, Ed.), Univ. of Wisconsin, 1959. pp. 413-422.
2. Collatz, L., Functional Analysis and Numerical Mathematics, Academic Press, 1966. Chapter 3.
3. Collatz, L., "Monotonicity and related methods in nonlinear differential equations problems," Numerical Solutions of Nonlinear Differential Equations (D. Greenspan, Ed.), Wiley, 1966. pp. 65-87.
4. Fujita, H., "On the nonlinear equations $\Delta u + e^u = 0$ and $v_t = \Delta v + e^v$." To appear, Bulletin AMS.
5. Gel'fand, I. M., "Some problems in the theory of quasilinear equations," Uspehi Mat. Nauk (N.S.) 14 (1959), no. 2 (86), 87-158 (in Russian). AMS Translations, Ser. 2, Vol. 29 (1963), 295-381.
6. Parter, S. V., Private communication, Oct. 1, 1968.
7. Protter, M. H., and H. F. Weinberger, Maximum Principles in Differential Equations, Prentice Hall, 1967. Chapter 2.
8. Rabinowitz, P., "Applications of linear programming to numerical analysis," SIAM Review 10, 121-159 (1968).
9. Rosen, J. B., "Approximate computational solution of nonlinear parabolic partial differential equations by linear programming," Numerical Solutions of Nonlinear Differential Equations (D. Greenspan, Ed.), Wiley, 1966, pp. 265-296.
10. Rosen, J. B. and R. Meyer, "Solution of nonlinear two-point boundary value problems by linear programming," Mathematical Theory of Control (A. V. Balakrishnan, L. W. Neustadt, Eds.), Academic Press, 1967. pp. 71-84.

11. Schröder, J., "Estimations in nonlinear equations," Proc. of IFIP Congress 65 (W. A. Kalenich, Ed.), Spartan Books, 1965. Vol. I, pp. 187-194.

12. Schröder, J., "Operator-ungleichungen und ihr numerische anwendung bei randwertaufgaben," Numerische Mathematik 9, 149-162 (1966).

13. Simonnard, M., Linear Programming (translated by W. S. Jewell), Prentice-Hall, 1966. Chapter 5.

14. Walter, W., Differential-und Integral Ungleichungen, Springer-Verlag, 1964.

15. Whiteman, J. R., "Numerical solution of a harmonic mixed boundary value problem by linear programming," Math. Research Center Tech. Summary Report No. 857, Univ. of Wisconsin, 1968.

Table 1

Unit Square Domain

$$-\nabla^2 u + g(u) = r, \quad \text{in} \quad D$$

$$u = 0, \quad \text{on} \quad \partial D$$

Polynomial approximation (boundary conditions satisfied)

$g(u)$	τ	r	m	$v_m(.5, .5)$	δ_1	γ^*	ν	Rel. error bound
0		1	45	.0737	.0031	.074	.00023	.31%
$-\tau e^{-u}$	1	0	45	.0700	.0031	.074†	.00023	.33
"	10	0	45	.5042	.031	.074†	.0023	.46
e^u		5	21	.277	.11	.074†	.0081	2.9
"		10	21	.617	.25	.074†	.018	2.9
"		50	21	2.886	1.42	.043	.061	2.1
"		50	45	2.891	.15	.043	.0065	.23
"		100	21	4.282	2.95	.020	.059	1.4
"		100	45	4.284	.31	.020	.0062	.15

† μ_m^* for linear case used.

Table 2

Unit Square Domain

$$-\nabla^2 u + g(u) = r \ , \quad \text{in} \quad D$$

$$u = 0 \ , \quad \text{on} \quad \partial D$$

Polynomial approximation (boundary conditions satisfied)

g(u)	r	m	$v_m(.5, .5)$	δ_1	γ^*	ν	Rel. error bound
$-u^2$	1	10	.0746	.11	.075	.0082	11.0%
"	1	15	.0744	.052	.075	.0039	5.3
"	1	21	.0738	.028	.075	.0021	2.8
"	1	28	.0739	.015	.075	.0011	1.5
"	1	36	.0739	.0067	.075	.0005	0.68
"	1	45	.0740	.0040	.075	.0003	0.40
"	10	45	.7631	.032	.079	.0025	0.32
"	80	45	10.1	.245	.39	.096	0.95
"	84	21	12.1	2.29	†	†	†
"	84	45	12.3	.26	1.46	.38	3.1
"	85	21	NO SOLUTION OBTAINED				

† No solution obtained to error bound equation.

Table 3

Unit Square Domain

$$-\nabla^2 u + g(u) = 0, \quad \text{in} \quad D$$

$$u = 0, \quad \text{on} \quad \partial D$$

Polynomial approximation (boundary conditions satisfied)

g(u)	τ	m	$v_m(.5, .5)$	δ_1	γ^*	ν	Rel. error bound
$-\tau e^u$	1.00	45	.0781	.0031	.075	.00023	0.30%
"	5.00	21	.5558	.140	.125	.0174	3.1
"	5.00	45	.5571	.012	.126	.0015	0.27
"	6.70	45	1.16	.021	.55	.012	1.0
"	6.80	45	1.33	.021	2.53	.053	4.0
"	6.81	45	NO SOLUTION OBTAINED				

Table 4

Ellipsoidal Domain

$$-\nabla^2 u = 1 \quad \text{in} \quad D$$
$$u = 0 \quad \text{on} \quad \partial D$$

Harmonic function approximation

m	$v_m(0, 0)$	ν = Max. bdry. error
6	0.44240	.0022
10	0.44278	.00054
15	0.44267	.00023
20	0.44267	.00010

Table 5

Square and Truncated Square Domain

$$-\nabla^2 u + g(u) = r, \quad \text{in} \quad D$$

$$u = 0, \quad \text{on} \quad \partial D$$

Harmonic and trigonometric functions

$g(u)$	r	D	m	$v_m(.5, .5)$	δ_1	δ_2	γ^*	$\hat{p}(\nu)$	ν	Rel. error bound
$-e^u$	0	Sq.	21	.0781	.00028	.00028	.075	-1.08	.00033	0.42%
e^u	50	Sq.	21	2.878	1.34	.013	.043	0	.071	2.5
e^u	50	Sq.	31	2.890	.31	.003	.043	0	.016	0.55
$-e^u$	0	Tr.Sq.	31	.0639	.016	.0002	.070	-1.07	.0011	1.7
$-e^{-u}$	0	Tr.Sq	31	.0584	.014	.0003	.063	0	.0012	2.1
e^u	5	Tr.Sq	31	.231	.142	.0014	.060	0	.0099	4.2

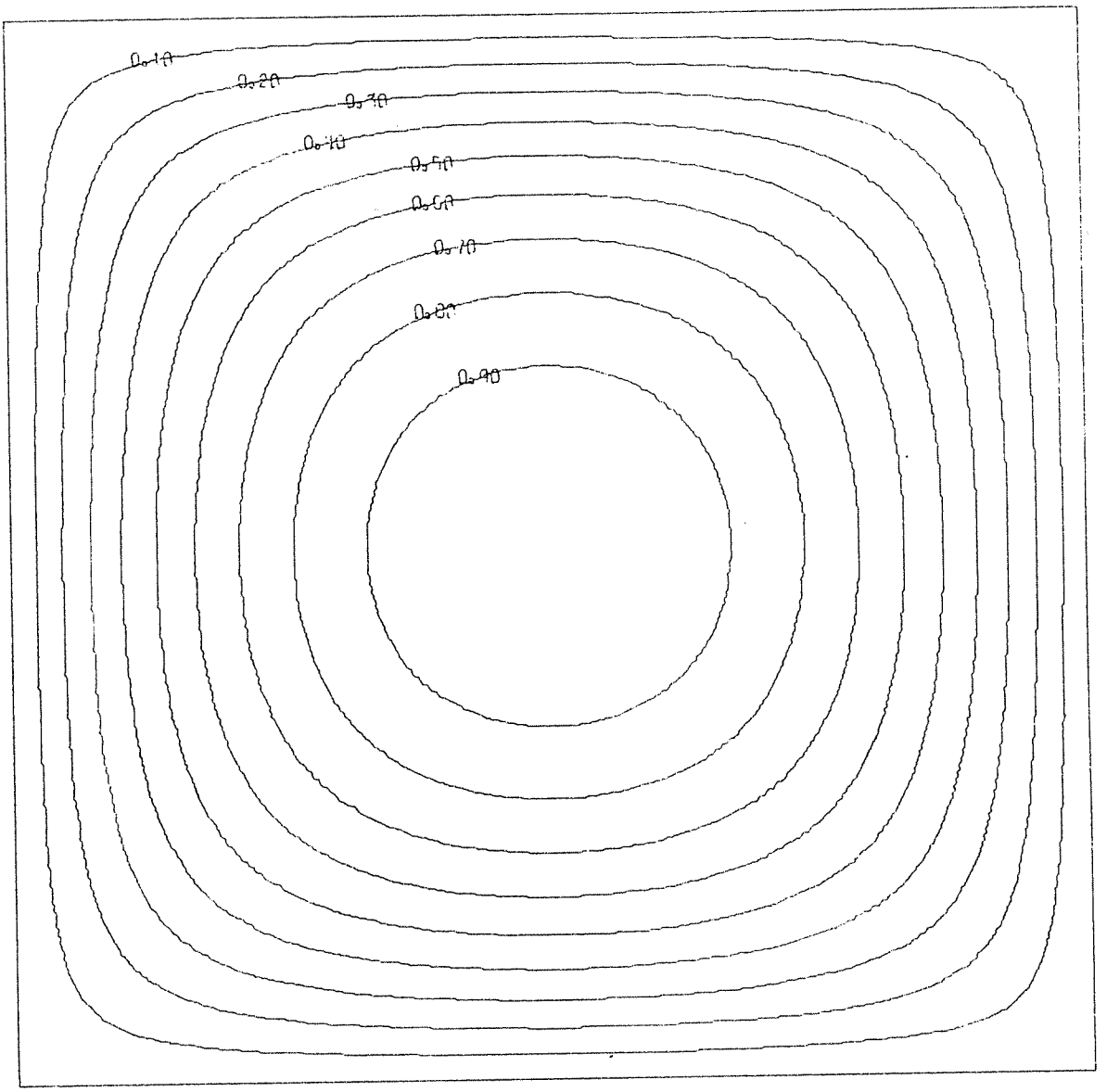

Fig. 1

Linear Elliptic Problem, $-\nabla^2 u = 1$
Polynomial Approximation (m = 45)
$\|v\| = 0.0737$

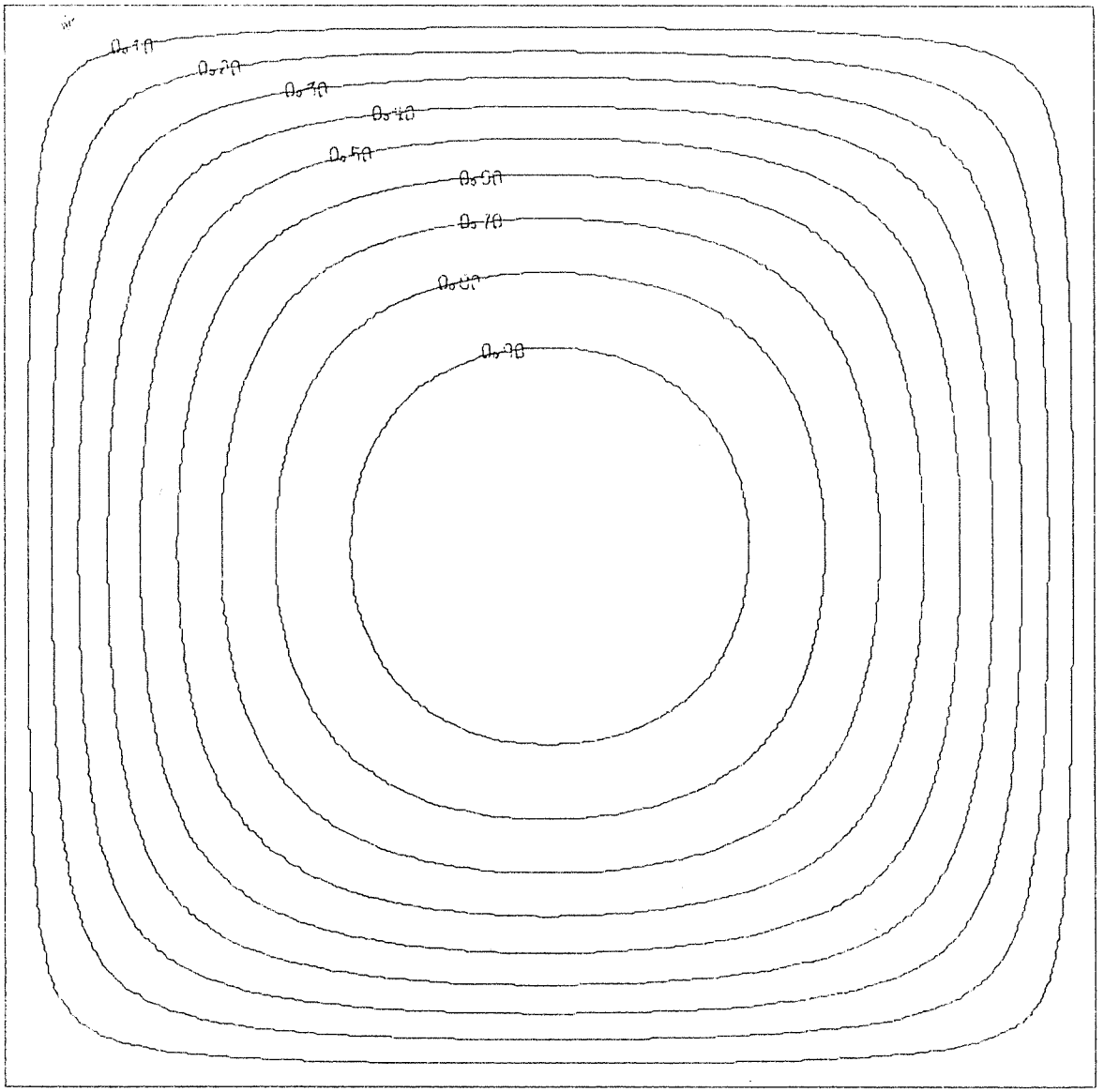

Fig. 2

Quasilinear Elliptic Problem

$-\nabla^2 u + e^u = 50$

Polynomial Approximation (m = 45)

$\|v\| = 2.891$

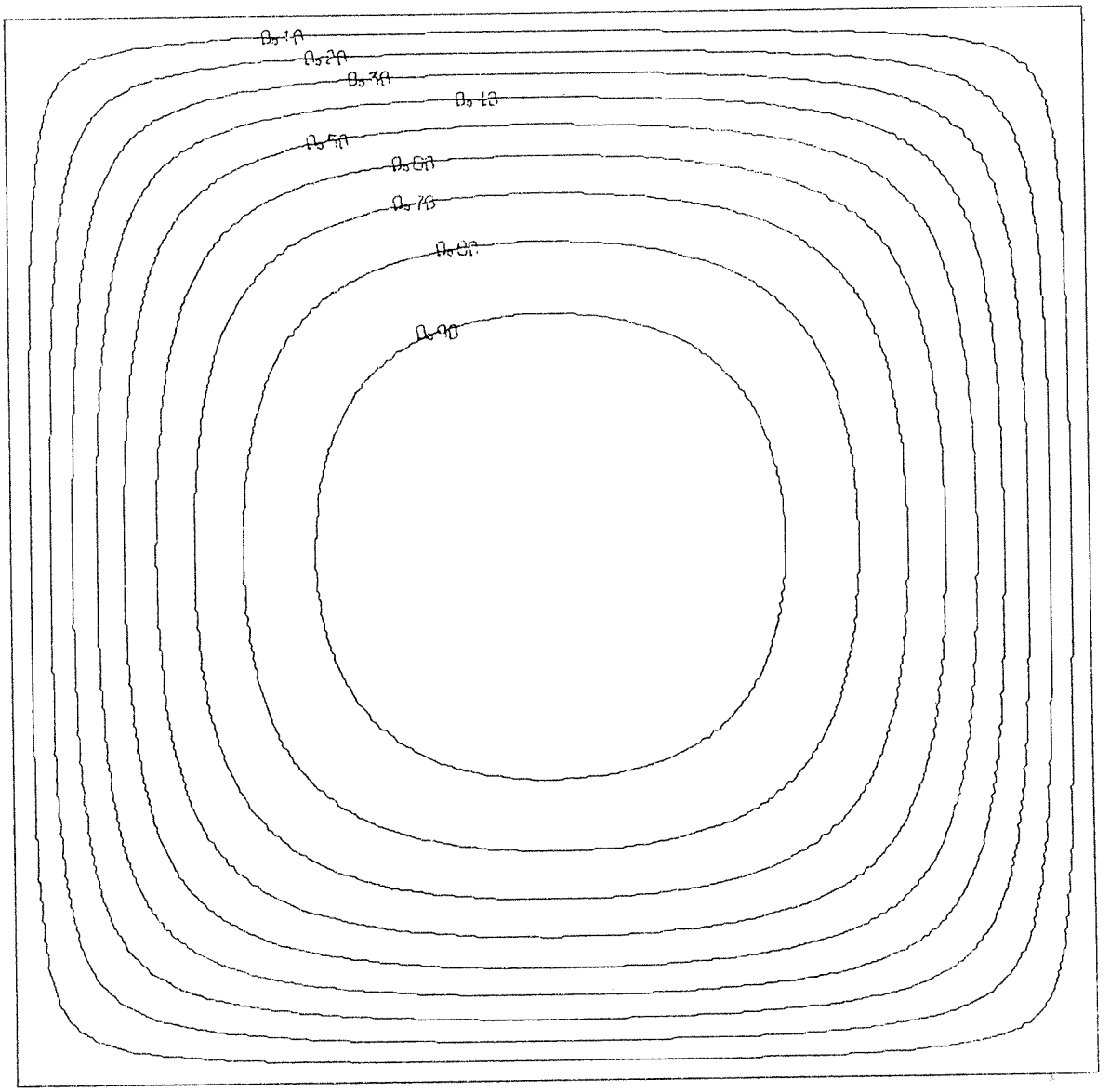

Fig. 3

Quasilinear Elliptic Problem

$-\nabla^2 u + e^u = 100$

Polynomial Approximation (m = 45)

$\|v\| = 4.284$

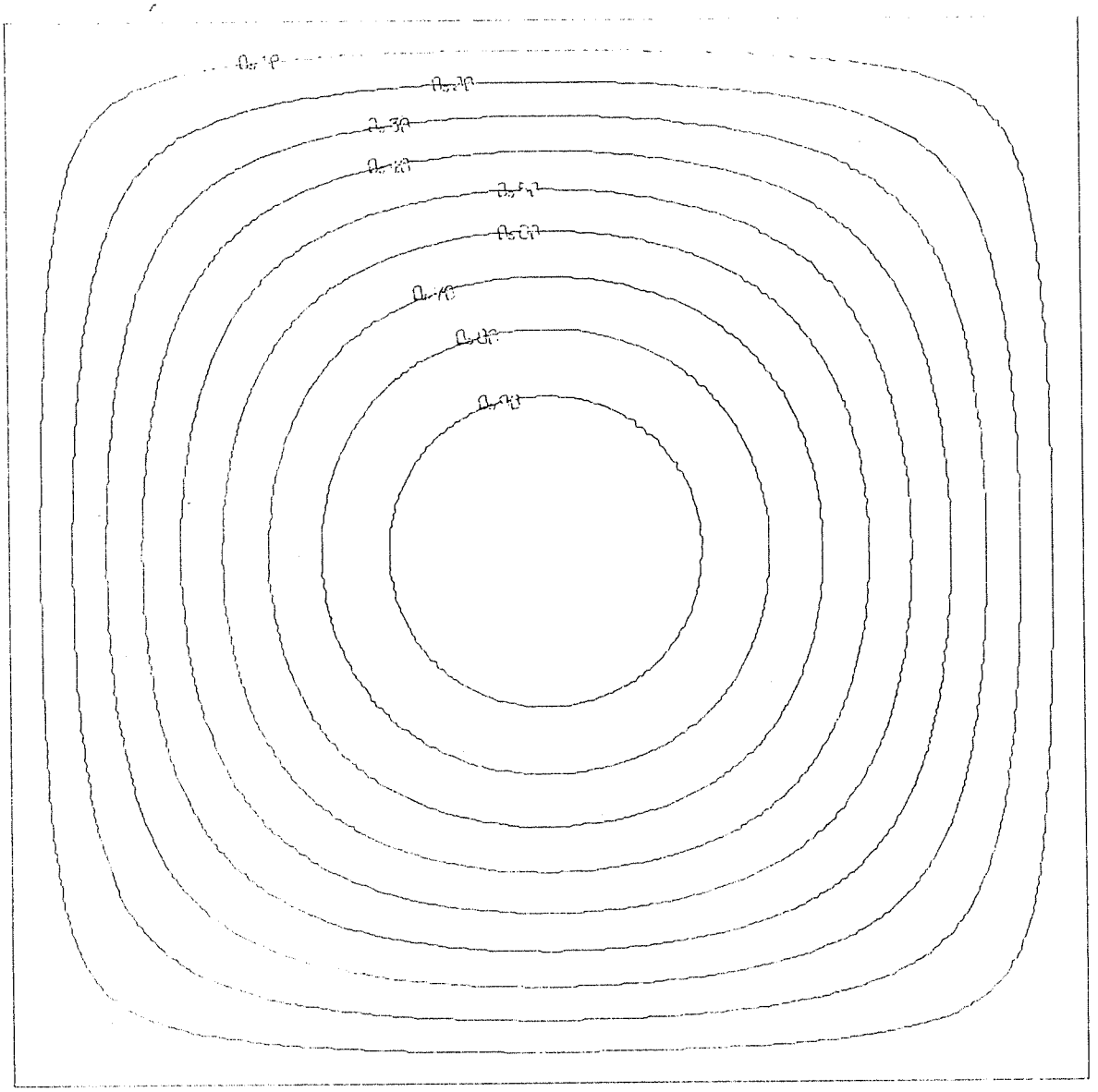

Fig. 4

Quasilinear Elliptic Problem

$-\nabla^2 u - u^2 = 84$

Polynomial Approximation (m = 45)

$\|v\| = 12.323$

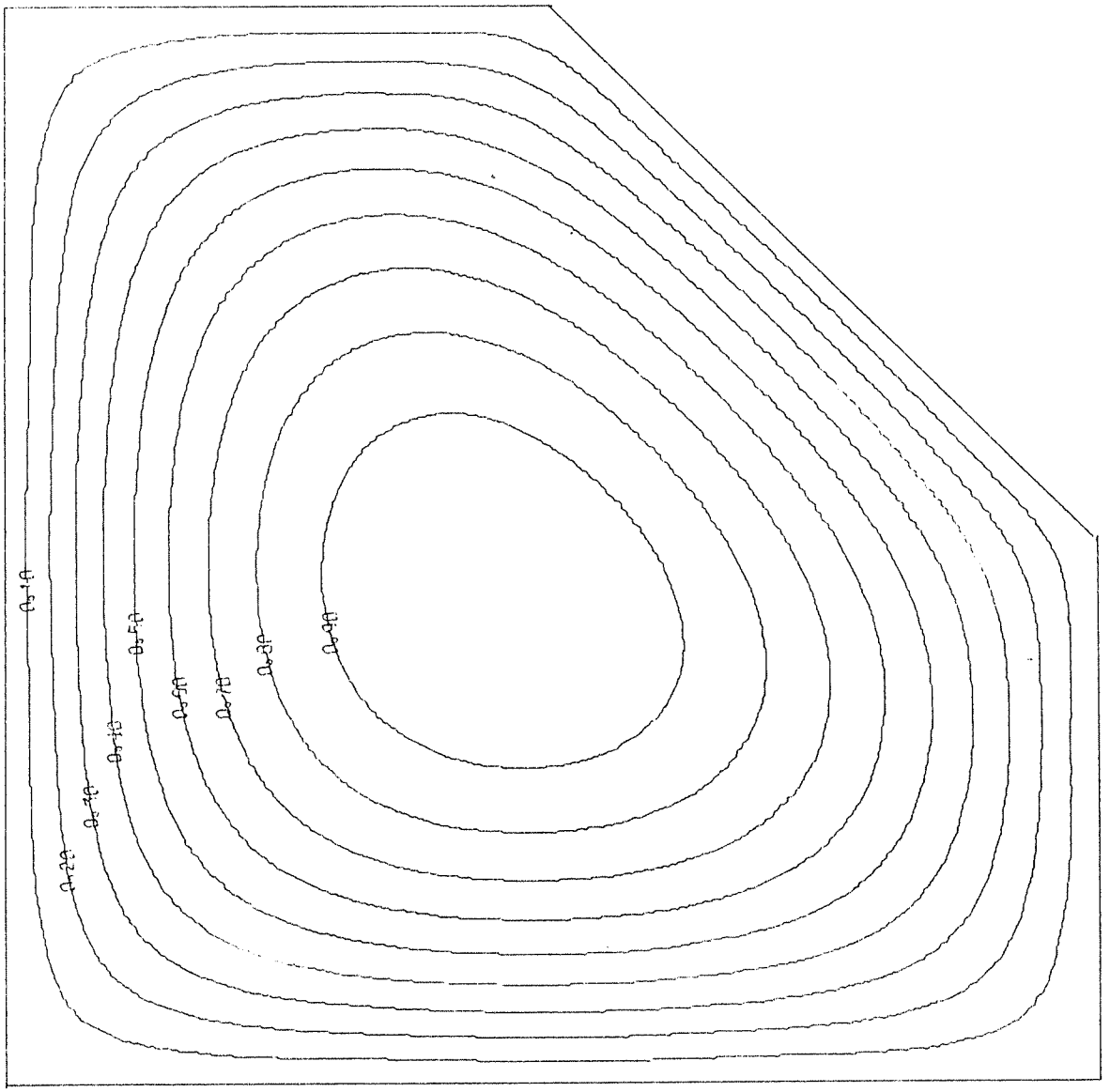

Fig. 5
Quasilinear Elliptic Problem
$-\nabla^2 u + e^u = 5$
Harmonic & Trigonometric Functions (m = 31)
$\|v\| = 0.234$