SOME ERROR ESTIMATES FOR THE BOX METHOD *

RANDOLPH E. BANK† AND DONALD J. ROSE‡

Abstract. We define and analyze several variants of the box method for discretizing elliptic boundary value problems in the plane. Our estimates show the error to be comparable to a standard Galerkin finite element method using piecewise linear polynomials.

Key words. box method, error bounds, piecewise linear triangular finite elements

AMS subject classifications. 65M15, 65M60

1. Introduction. In this work we derive some error estimates for the box method applied to self-adjoint, positive definite elliptic boundary value problems in regions of the plane. The "classical" box method, as described in, say, Varga [9], is a finite difference approximation to an integral formulation of the problem after applying Green's Theorem. In the present work, the box method is generalized and cast as a Galerkin procedure more in the spirit of the finite element method. In this Galerkin formulation, the boxes are constructed as a dual mesh of an underlying triangular grid. The classical box method is one specific instance of this more general formulation.

In the classical box method, the difference equations are formulated on the dual mesh of boxes without explicit reference to the underlying triangular mesh. On the other hand, in our Galerkin procedure, the trial space consists of continuous piecewise linear polynomials on the triangular mesh, while the test space consists of piecewise constants on the dual box mesh.

The main result of this paper is that, under reasonable hypotheses, the solution generated by the box method, u_B, is of comparable accuracy to the solution u_L generated by the standard Ritz-Galerkin procedure using piecewise linear finite elements. In particular, if u of the true solution and $\|\!| \cdot |\!\|$ denotes the energy norm, we have, for Poisson's equation

$$|\!|\!| u - u_L |\!|\!| \leq |\!|\!| u - u_B |\!|\!| \leq C |\!|\!| u - u_L |\!|\!|. \tag{1.1}$$

Our proof of (1.1) follows the strategy of Babuška and Aziz outlined in [2], in that we show the bilinear form associated with the box method satisfies and inf-sup condition. Because this bilinear form involves certain line integrals of u, an additional assumption on u (inequality (2.4)), beyond that imposed by the standard finite element method, is required. For more general self-adjoint elliptic equations with zero-order terms there is an additional term on the right hand side of (1.1). Inequality (1.1) is true for general irregular and nonuniform meshes, which are required to satisfy a shape regularity property for the triangular elements. Results having the same flavor, but treating different methods and using different proof techniques, have been shown for one-dimensional problems and for tensor product-like meshes for higher-dimensional problems by Kreiss *et al.* [5] and Manteuffel and White [6]. See also

*Received by the editors September 18, 1984; accepted for publication (in revised form) May 30, 1986.

†Department of Mathematics, University of California at San Diego, La Jolla, California 92093. The work of this author was supported in part by the Office of Naval Research under contract N00014-82K-0197.

‡Department of Computer Science, Duke University , Durham, North Carolina 27706. The work of this author was supported in part by the Office of Naval Research under contract N00014-85K-0487.

Tikhonov and Samarskii [8]. In related work, Nakata, Weiser and Wheeler have shown certain block centered finite difference schemes on rectangular meshes are equivalent to a mixed finite element procedure [7].

In the proof of (1.1) it is revealed that the linear system for the box method and standard linear finite elements are strikingly similar. This is widely known for the special case of the Laplacian on a uniform square $n \times n$ mesh, where both methods reduce to the standard 5-point centered finite difference approximation. In fact, the box method and linear finite elements *always* produce the same matrix for the Laplacian when the standard basis functions are used. We first observed this when using the box method for the system of elliptic partial differential equations used in modeling semiconductor devices [4, 3], and this served as motivation for the present theoretical investigation.

This observation has significant practical as well as theoretical importance. For example, in assembling the sparse linear system, the traditional finite element approach has been to assemble matrices and right-hand sides triangle by triangle, while the traditional finite difference approach has been to carry out the assembly process equation by equation. Armed with the knowledge that the end result will be the same, one can shift freely between both viewpoints, choosing those algorithms best suited to exploit parallelism and other facets of the machine architecture.

The remainder of this paper is organized as follows: In Section 2, we establish notation and prove some preliminary lemmas. In Section 3, we analyze the box method for Poisson's equation; and in Section 4, we consider the box method applied to more general elliptic equations.

2. Preliminaries. Let Ω be a bounded polygonal region in $\mathbb{R}^2$ with boundary $\partial\Omega$. Let $\mathcal{T}$ denote a triangulation of Ω. We require triangles $t \in \mathcal{T}$ to be shape regular but do not require the mesh as a whole to be quasi-uniform.

In particular, for $t \in \mathcal{T}$, let h_t denote the diameter of the circumscribing circle for t and k_t denote the diameter of the inscribing circle. We assume there is a positive constant δ_0 such that

$$\delta_0 \leq \frac{k_t}{h_t} \qquad \text{for all } t \in \mathcal{T}. \tag{2.1}$$

Let v_i, $1 \leq i \leq n$ denote the vertices of the triangulation. With each vertex v_i we associate a region Ω_i consisting of those triangles $t \in \mathcal{T}$ which have v_i as a vertex. (See Figure 2.1.)

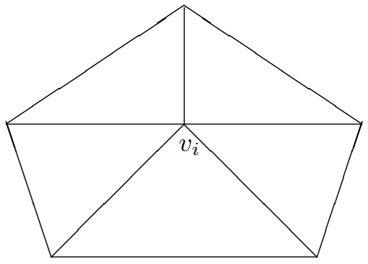

FIG. 2.1. Ω_i

We now construct a dual mesh $\mathcal{B}$ for $\mathcal{T}$. The elements in the dual mesh are called boxes, and are constructed as follows (see Figure 2.2): for each triangle $t \in \mathcal{T}$, select a distinguished point $p \in \bar{t}$. Connect p by straight-line segments (e_1, e_2, e_3, in Figure 2.2) to the edge midpoints of t (m_1, m_2, m_3, in Figure 2.2). This partitions t into three subregions (with areas a_1, a_2, a_3).

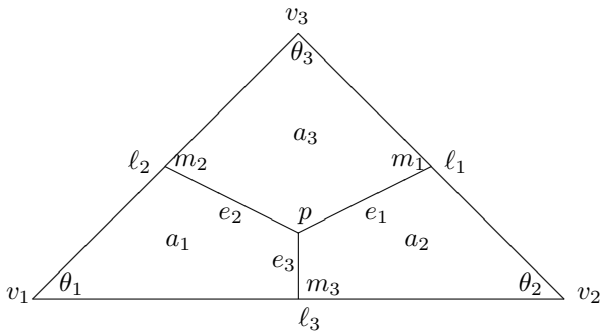

FIG. 2.2. $t \in \mathcal{T}$

With each vertex v_i, we will associate a box $b_i \in \mathcal{B}$, $b_i \subset \Omega_i$, which consists of the union of the subregions in Ω_i which have v_i as a corner (see Figure 2.3). Boxes need not be convex, but are star-shaped. Also, since the boundary on b_i must pass through the midpoints of the triangles sides, (2.1) implies there exists a positive constant $\delta_1 = \delta_1(\delta_0)$ such that

$$\delta_1^{-1} \max_{t \cap \Omega_i \neq \emptyset} h_t \leq h_{b_i} \leq \delta_1 \min_{t \cap \Omega_i \neq \emptyset} h_t \tag{2.2}$$

where $h_{b_i} = \operatorname{diam}(b_i)$.

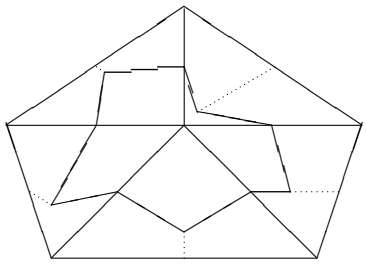

FIG. 2.3. $b_i \subset \Omega_i$

For most of our results, we do not require p to lie in the strict interior of t; p can be on ∂t and even coincide with a vertex. This could result in some boxes having zero area but nontrivial perimeters. In other cases, notably Lemma 2.2 and Theorem 4.1 that follow, we require boxes to have nontrivial area. There we assume there exists a constant $\alpha > 0$ such that

$$\alpha \leq \frac{|b_i|}{|\Omega_i|} \qquad \text{for all } b_i \in \mathcal{B}. \tag{2.3}$$

In the classical box method, p is chosen as the intersection of the perpendicular bisectors of the three edges. This choice requires $\max \theta_i \leq \pi/2$ in order that $p \in \bar{t}$. A second natural choice for p is the barycenter, in which case $a_1 = a_2 = a_3 = |t|/3$.

Let E denote the set of edges generated by this process. Three (unique) members of E can be associated with each $t \in \mathcal{T}$. A box $b \in \mathcal{B}$ corresponding to a boundary vertex (on lying on $\partial\Omega$) has a boundary with nontrivial intersection with $\partial\Omega$. However, the notation ∂b will refer only to that part of the actual boundary consisting of edges in E.

We now define the function spaces which are relevant to our analysis. Let $\mathcal{L}^2(R)$ and $\mathcal{H}^1(R)$ denote the usual Sobolev spaces equipped with the norms

$$\|u\|^2_{\mathcal{L}^2(R)} = \int_R u^2 \, dx, \qquad \|u\|^2_{\mathcal{H}^1(R)} = \|\nabla u\|^2_{\mathcal{L}^2(R)} + \|u\|^2_{\mathcal{L}^2(R)}$$

where $\|\nabla u\|^2_{\mathcal{L}^2(R)} = \int_R \nabla u \cdot \nabla u \, dx$. Other Sobolev spaces are not used. Let $\mathcal{P}^1(\mathcal{T}) \subset \mathcal{H}^1(\Omega)$ denote the space of C^0 piecewise linear polynomials associated with $\mathcal{T}$. We will denote by $\mathcal{S}^0(\mathcal{B})$ the space of piecewise $\mathcal{H}^1$ functions with respect to $\mathcal{B}$.

$$\mathcal{S}^0(\mathcal{B}) = \{v \in \mathcal{L}^2(\Omega) | v \in \mathcal{H}^1(b) \text{ for all } b \in \mathcal{B}\}.$$

Let $e \in E$ and $u \in \mathcal{S}^0(\mathcal{B})$. We denote the jump in u across e at $x \in e$ by

$$u_J(x) = u(x + 0) - u(x - 0)$$

where $u(x \pm 0)$ are the two limit values of $u(x)$ along the normal directions for e. (The normal can have either sign, but once chosen, will be used consistently.) We let $\mathcal{P}^0 \subset \mathcal{S}^0(\mathcal{B})$ denote the space of discontinuous piecewise constants with respect to the boxes.

We let $\mathcal{S}^1(\mathcal{T}) \subset \mathcal{H}^1(\Omega)$ denote the class of functions $u \in \mathcal{H}^1(\Omega)$ which satisfy

$$\left(\sum_{t \in \mathcal{T}} h_t^2 \|\Delta u\|^2_{\mathcal{L}^2(t)} \right)^{1/2} \leq \gamma \inf_{\chi \in \mathcal{P}^1(\mathcal{T})} \|\nabla(u - \chi)\|_{\mathcal{L}^2(\Omega)} \tag{2.4}$$

for $\gamma(\delta_0)$ independent of $\max h_t$. Generally speaking, $\mathcal{S}^1(\mathcal{T})$ is that set of functions in $\mathcal{H}^1(\Omega)$ which can be well approximated by piecewise linear polynomials on $\mathcal{T}$. Finally, we denote by $\mathcal{H}^1_0(\Omega)$, $\mathcal{P}^1_0(\mathcal{T})$, $\mathcal{S}^1_0(\mathcal{T})$, $\mathcal{P}^0_0(\mathcal{B})$, and $\mathcal{S}^0_0(\mathcal{B})$ those subsets of $\mathcal{H}^1(\Omega)$, $\mathcal{P}^1(\mathcal{T})$, $\mathcal{S}^1(\mathcal{T})$, $\mathcal{P}^0(\mathcal{B})$, and $\mathcal{S}^0(\mathcal{B})$, respectively, whose elements are zero on $\partial\Omega$.

There is a natural correspondence between the spaces $\mathcal{P}^1(\mathcal{T})$ and $\mathcal{P}^0(\mathcal{B})$ (which have equal dimension) that we shall exploit. Let $\{\phi_i\}$ denote the usual nodal basis for $\mathcal{P}^1(\mathcal{T})$, satisfying

$$\phi_i(v_j) = \delta_{ij}$$

where v_j is a vertex in the triangulation. Note support$(\phi_i) = \bar{\Omega}_i$. Let $\{\bar{\phi}_i\}$ denote the basis for $\mathcal{P}^0(\mathcal{B})$ consisting of the characteristic functions for b_i.

$$\bar{\phi}_i(x) = \begin{cases} 1 & x \in b_i, \\ 0 & \text{otherwise.} \end{cases}$$

Let $\chi \in \mathcal{P}^1(\mathcal{T})$ be expressed in terms of $\{\phi_i\}$ as

$$\chi = \sum_i a_i \phi_i(x).$$

We associate (via and invertible map G) with χ and element $\bar{\chi} \in \mathcal{P}^0(\mathcal{B})$ defined by

$$G(\chi) = \bar{\chi} = \sum_i a_i \bar{\phi}_i(x).$$

Note χ and $\bar{\chi}$ have common values at the vertices of $\mathcal{T}$.

LEMMA 2.1. *Let $u \in \mathcal{P}^1(\mathcal{T})$, $\bar{u} = G(u) \in \mathcal{P}^0(\mathcal{B})$. Then there exists a constant $C_0 = C_0(\delta_0)$ such that*

$$C_0^{-1} \|\nabla u\|_{\mathcal{L}^2(\Omega)} \leq \left(\sum_{e \in E} |\bar{u}_J|^2 \right)^{1/2} \leq C_0 \|\nabla u\|_{\mathcal{L}^2(\Omega)}. \tag{2.5}$$

Proof. Let $t \in \mathcal{T}$ have side lengths ℓ_i $(1 \leq i \leq 3)$ as in Figure 2.2, and let u and $\bar{u}$ have node values u_i, $1 \leq i \leq 3$ at the vertices of t. Let

$$d_i = \frac{\ell_j \ell_k \cos \theta_i}{4|t|} \tag{2.6}$$

where (i, j, k) are cyclic permutations of $(1, 2, 3)$. Then by a direct computation the element stiffness matrix for t is

$$K_t = \begin{pmatrix} d_2 + d_3 & -d_3 & -d_2 \\ -d_3 & d_3 + d_1 & -d_1 \\ -d_2 & -d_1 & d_1 + d_2 \end{pmatrix}$$

and

$$\|\nabla u\|^2_{\mathcal{L}^2(t)} = d_1(u_2 - u_3)^2 + d_2(u_3 - u_1)^2 + d_3(u_1 - u_2)^2. \tag{2.7}$$

Furthermore

$$\sum_1^3 |\bar{u}_J|^2 = (u_2 - u_3)^2 + (u_3 - u_1)^2 + (u_1 - u_2)^2 \tag{2.8}$$

with corresponding "jump" matrix

$$\bar{K}_t = \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}.$$

Both (2.7) and (2.8) are zero when u is constant on t. We may thus restrict attention to any two-dimensional subspace not containing the constant function (for example, vectors of the form $(u_1\, u_2\, 0)^T$) and directly compute the minimum and maximum of the generalized Rayleigh quotient involving K_t and $\bar{K}_t$. From this computation we find

$$C_- \|\nabla u\|^2_{\mathcal{L}^2(t)} \leq \sum_1^3 |\bar{u}_J|^2 \leq C_+ \|\nabla u\|^2_{\mathcal{L}^2(t)} \tag{2.9}$$

where

$$C_\pm = \{d_1 + d_2 + d_3 \pm 2^{-1/2}((d_1 - d_2)^2 + (d_2 - d_3)^2 + (d_3 - d_1)^2)^{1/2}/3.$$

Note $C_\pm = C_\pm(\delta_0) > 0$. The lemma now follows by summing (2.9) over $t \in \mathcal{T}$. ∎

LEMMA 2.2. *Let $u \in \mathcal{P}^1(\mathcal{T})$, $\bar{u} = G(u) \in \mathcal{P}^0(\mathcal{B})$. Then there exists a constant $C_1 = C_1(\delta_0)$ such that*

$$\|\bar{u}\|_{\mathcal{L}^2(\Omega)} \le C_1 \|u\|_{\mathcal{L}^2(\Omega)}. \tag{2.10}$$

Furthermore, if (2.3) is satisfied then there exists a constant $C_2 = C_2(\delta_0, \Omega)$ such that

$$\|u\|_{\mathcal{L}^2(\Omega)} \le C_2 \|\bar{u}\|_{\mathcal{L}^2(\Omega)}. \tag{2.11}$$

Proof. The proof is analogous to that of Lemma 2.1. Let $t \in \mathcal{T}$ be as in Figure 2.2. Then

$$\begin{aligned} \|u\|^2_{\mathcal{L}^2(t)} &= \frac{|t|}{12}\{u_1^2 + u_2^2 + u_3^2 + (u_1 + u_2 + u_3)^2\}, \\ \|\bar{u}\|^2_{\mathcal{L}^2(t)} &= a_1 u_1^2 + a_2 u_2^2 + a_3 u_3^2 \end{aligned} \tag{2.12}$$

Inequality (2.10) follows immediately from (2.12) by summing over $t \in \mathcal{T}$. Inequality (2.11) is only modestly more complicated to show (since some a_i may be zero even if (2.3) is satisfied). ∎

LEMMA 2.3. *Let $u \in \mathcal{P}^1(\mathcal{T})$, $v \in \mathcal{P}^1(\mathcal{T})$, $\bar{v} = G(v) \in \mathcal{P}^0(\mathcal{B})$. Then*

$$-\sum_{b \in \mathcal{B}} \int_{\partial b} \frac{\partial u}{\partial n} \bar{v}\, ds = \int_\Omega \nabla u \cdot \nabla v\, dx \tag{2.13}$$

where n is the outward pointing normal.

Proof. It is sufficient to show (2.13) for $v = \phi_i$, $\bar{v} = \bar{\phi}_i$, where ϕ_i and $\bar{\phi}_i$ are the basis functions for $\mathcal{P}^1(\mathcal{T})$ and $\mathcal{P}^0(\mathcal{B})$ defined above. For this choice (2.13) reduces to

$$-\int_{\partial b_i} \frac{\partial u}{\partial n}\, ds = \int_{\Omega_i} \nabla u \cdot \nabla \phi_i\, dx \tag{2.14}$$

This equality will be shown triangle by triangle for $t \in \mathcal{T}$ such that $t \subseteq \Omega_i$; each such triangle contains two edges $e \in E$ lying on ∂b_i. Without loss of generality, assume that vertex 1 in Figure 2.2 corresponds to vertex i. Then

$$\begin{aligned} \int_{\Omega_i \cap t} \nabla u \cdot \nabla \phi_i\, dx &= \begin{pmatrix} 1 & 0 & 0 \end{pmatrix} K_t \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} \\ &= d_2(u_1 - u_3) + d_3(u_1 - u_2). \end{aligned}$$

Next, note that since u is a linear polynomial on t, $\Delta u = 0$ in t, and by Green's theorem

$$\int_{\partial b_i} \frac{\partial u}{\partial n}\, ds$$

depends only on the endpoints of the integration path (in this case, midpoints m_2 and m_3 in Figure 2.2) and not on the location of p. One can choose a simple integration path (the straight line connecting m_2 and m_3, or the perpendicular bisectors of sides 2 and 3 of t, for example) and directly compute

$$\begin{aligned} -\int_{\partial b_i} \frac{\partial u}{\partial n}\, ds &= d_2(u_1 - u_3) + d_3(u_1 - u_2) \\ &= \int_t \nabla u \cdot \nabla \phi_i\, dx, \end{aligned}$$

which completes the proof. □

3. The box method for Poisson's equation. In this section we consider the model problem

$$-\Delta u = f \quad \text{in } \Omega, \qquad u = 0 \quad \text{on } \partial\Omega, \tag{3.1}$$

where $f \in \mathcal{L}^2(\Omega)$. The standard weak form of (3.1) is: Find $u \in \mathcal{H}_0^1(\Omega)$ such that

$$\begin{aligned} &a(u,v) = (f,v) \qquad \text{for all } v \in \mathcal{H}_0^1(\Omega), \\ &a(u,v) = \int_\Omega \nabla u \cdot \nabla v \, dx, \\ &(f,v) = \int_\Omega f v \, dx. \end{aligned} \tag{3.2}$$

The energy norm $\|\!| \cdot |\!\|$ is defined for $u \in \mathcal{H}_0^1(\Omega)$ by

$$|\!|\!| u |\!|\!|^2 = a(u,u) = \|\nabla u\|^2_{\mathcal{L}^2(\Omega)}. \tag{3.3}$$

The standard finite element approximation using the space $\mathcal{P}_0^1(\mathcal{T})$ is given by: Find $u_L \in \mathcal{P}_0^1(\mathcal{T})$ such that

$$a(u_L, v) = (f,v) \qquad \text{for all } v \in \mathcal{P}_0^1(\mathcal{T}). \tag{3.4}$$

The solution u_L is known to be the best approximation of the solution u of (3.2) in the energy norm, i.e.,

$$|\!|\!| u - u_L |\!|\!| = \inf_{\chi \in \mathcal{P}_0^1(\mathcal{T})} |\!|\!| u - \chi |\!|\!|. \tag{3.5}$$

In order to formulate the box method we must generalize (3.2). Let $v \in \mathcal{S}^0(\mathcal{B})$; multiply (3.1) by v and integrate by parts to obtain

$$a(u,v) + \bar{a}(u,v) = (f,v) \tag{3.6}$$

where $a(u,v)$ is interpreted as the "broken" inner product

$$a(u,v) = \sum_{b \in \mathcal{B}} \int_b \nabla u \cdot \nabla v \, dx \tag{3.7}$$

and the bilinear form $\bar{a}(u,v)$ is given by

$$\bar{a}(u,v) = -\sum_{b \in \mathcal{B}} \int_{\partial b} \frac{\partial u}{\partial n} v \, ds. \tag{3.8}$$

The box method is defined by: Find $u_B \in \mathcal{P}_0^1(\mathcal{T})$ such that

$$\bar{a}(u_B, \bar{v}) = (f, \bar{v}) \qquad \text{for all } \bar{v} \in \mathcal{P}_0^0(\mathcal{B}). \tag{3.9}$$

For $\bar{v} \in \mathcal{P}_0^0(\mathcal{B})$, we define the "energy norm" $|\!|\!| \bar{v} |\!|\!|$ by setting $|\!|\!| \bar{v} |\!|\!| = |\!|\!| v |\!|\!|$ $v = G^{-1}(\bar{v})$. By Lemma 2.1, $|\!|\!| \bar{v} |\!|\!|$ is comparable to$(\sum_e |\bar{v}_J|^2)^{1/2}$. The main result of this section is the following theorem.

THEOREM 3.1. *Let u, u_L, u_B be defined by (3.2), (3.4), and (3.9). Assume (2.1) and $u \in \mathcal{S}_0^1(\mathcal{T})$. Then*

$$\|\!|u - u_B|\!\| \leq C \inf_{\chi \in \mathcal{P}_0^1(\mathcal{T})} \|\!|u - \chi|\!\| \tag{3.10}$$

where $C = C(\delta_0, \gamma)$.

Proof. Our proof follows the general strategy given by Babuška and Aziz in [2], in that we prove the bilinear form (3.8) satisfies a discrete inf $-$ sup condition. However, our bound is somewhat weaker in that it (necessarily) depends on the constant γ.

From (3.6) and (3.9)

$$\bar{a}(u - u_B, \bar{v}) = 0 \quad \text{for all } \bar{v} \in \mathcal{P}_0^0(\mathcal{B}). \tag{3.11}$$

Let $\chi \in \mathcal{P}_0^1(\mathcal{T})$ and let $v = G^{-1}(\bar{v})$. Then from (3.11) and Lemma 2.3,

$$\begin{aligned} \sup_{\bar{v} \in \mathcal{P}_0^0(\mathcal{B})} \frac{\bar{a}(u - \chi, \bar{v})}{\|\!|\bar{v}|\!\|} &= \sup_{\bar{v} \in \mathcal{P}_0^0(\mathcal{B})} \frac{\bar{a}(u_B - \chi, \bar{v})}{\|\!|\bar{v}|\!\|} \\ &= \sup_{v \in \mathcal{P}_0^1(\mathcal{T})} \frac{a(u_B - \chi, v)}{\|\!|v|\!\|} \\ &\geq \|\!|u_B - \chi|\!\|. \end{aligned} \tag{3.12}$$

On the other hand, we have

$$\begin{aligned} |\bar{a}(u - \chi, \bar{v})| &= \left| \sum_e \bar{v}_J \int_e \frac{\partial (u - \chi)}{\partial n} \, ds \right| \\ &\leq C \|\!|\bar{v}|\!\| \left(\sum_e |e| \int_e \left(\frac{\partial (u - \chi)}{\partial n} \right)^2 \, ds \right)^{1/2}. \end{aligned} \tag{3.13}$$

We consider the last term in (3.13) on a triangle by triangle basis. By (2.1)-(2.2) $|e_i| \leq C h_t$ if $e_i \in t$. Using standard trace inequalities [1], and noting $\Delta \chi = 0$ in t, we have

$$\begin{aligned} \left| \sum_{i=1}^3 h_t \int_{e_i} \left(\frac{\partial (u - \chi)}{\partial n} \right)^2 \, ds \right| &\leq C h_t \left\{ h_t \|\Delta (u - \chi)\|^2_{\mathcal{L}^2(t)} + h_t^{-1} \|\nabla (u - \chi)\|^2_{\mathcal{L}^2(t)} \right\} \\ &\leq C \left\{ h_t^2 \|\Delta u\|^2_{\mathcal{L}^2(t)} + \|\nabla (u - \chi)\|^2_{\mathcal{L}^2(t)} \right\}. \end{aligned} \tag{3.14}$$

Summing (3.14) over $t \in \mathcal{T}$, then using (2.4) and (3.13) we have

$$|\bar{a}(u - \chi, \bar{v})| \leq C(\delta_0, \gamma) \|\!|\bar{v}|\!\| \|\!|u - \chi|\!\|. \tag{3.15}$$

Combining (3.12) and (3.15) we prove the result, since

$$\|\!|u - u_B|\!\| \leq \|\!|u - \chi|\!\| + \|\!|u_B - \chi|\!\|$$

□

COROLLARY 3.2.

$$\|\!|u - u_L|\!\| \leq \|\!|u - u_B|\!\| \leq C \|\!|u - u_L|\!\| \tag{3.16}$$

where C is given by Theorem 3.1.

Proof. Take $\chi = u_B$ in (3.5) and $\chi = u_L$ in (3.10). □

Theorem 3.1 and its corollary show the box method and the standard finite element method for piecewise linear polynomials yield solutions of comparable accuracy, at least in the energy norm. If we use the standard nodal basis for $\mathcal{P}_0^1(\mathcal{T})$, then (3.4) corresponds to the linear system

$$AU = F$$

where

$$\begin{aligned} A_{ij} &= a(\phi_j, \phi_i), \\ F_i &= (f, \phi_i) = \int_{\Omega_i} f\phi_i \, dx, \\ u_L &= \sum_i U_i \phi_i(x). \end{aligned}$$

The box method leads to the linear system

$$\bar{A}\bar{U} = \bar{F}$$

where, by Lemma 2.3

$$\begin{aligned} \bar{A}_{ij} &= \bar{a}(\phi_j, \bar{\phi}_i) = a(\phi_j, \phi_i) = A_{ij} \\ \bar{F}_i &= (f, \bar{\phi}_i) = \int_{b_i} f \, dx, \\ u_B &= \sum_i \bar{U}_i \phi_i(x). \end{aligned}$$

The stiffness matrices for both linear systems are identical and the solution vectors are point values for u_L and u_B at the vertices. The only difference is the right-hand side, and in an average sense they are close, For particular box methods (e.g. choosing p in Figure 2.2 to be the barycenter) one can give "natural" interpretations of $\bar{F}_i$ in terms of quadrature rule approximations of F_i. From this point of view, the results of Theorem 3.1 seem quite reasonable.

4. The box method for self-adjoint problems. In this section, we consider the more general self-adjoint elliptic boundary value problem of the form

$$-\nabla \cdot a\nabla u + \sigma u = f \quad \text{in } \Omega, \qquad u = 0 \quad \text{on } \partial\Omega, \tag{4.1}$$

where a (respectively σ) is a smooth positive (nonnegative) function and

$$\underline{a} \le a(x) \le \bar{a}, \qquad 0 \le \sigma(x) \le \bar{\sigma} \quad \text{for } x \in \bar{\Omega}.$$

The weak form of (4.1) is Find $u \in \mathcal{H}_0^1(\Omega)$ such that

$$\begin{aligned} a(u, v) &= (f, v) \qquad \text{for all } v \in \mathcal{H}_0^1(\Omega), \\ a(u, v) &= \int_\Omega a\nabla u \cdot \nabla v + \sigma uv \, dx. \end{aligned} \tag{4.2}$$

Withe this definition for $a(\cdot, \cdot)$ the energy norm $|\!|\!| u |\!|\!|$ is given by

$$|\!|\!| u |\!|\!|^2 = a(u, u),$$

and the finite element solution satisfies (3.4)-(3.5). In analogy with (3.9) the box method for (4.1) is defined by: Find $u_B \in \mathcal{P}_0^1(\mathcal{T})$ such that

$$\bar{a}(u_B, \bar{v}) + (\sigma \bar{u}_B, \bar{v}) = (f, \bar{v}) \qquad \text{for all } \bar{v} \in \mathcal{P}_0^0(\mathcal{B}), \tag{4.3}$$

$$\bar{a}(u_B, \bar{v}) = -\sum_{b \in \mathcal{B}} \int_{\partial b} a \frac{\partial u_B}{\partial n} \bar{v}\, ds. \tag{4.4}$$

A natural generalization of the Galerkin formulation of the box method would use u_B instead of $\bar{u}_B$ in the second term on the left-hand side of (4.3). However, in the classical formulation of the box method, the σu term in (4.1) would normally be discretized using a diagonal matrix. The use of $\bar{u}_B$ allows this when the basis functions $\{\phi_i\}$ for $\mathcal{P}_0^1(\mathcal{T})$ and $\{\bar{\phi}_i\}$ for $\mathcal{P}_0^0(\mathcal{B})$ are used in constructing the stiffness matrix. Using u_B in this term would result in a nondiagonal (and generally nonsymmetric) matrix.

To analyze the error, we begin with the analogue of (3.11):

$$\bar{a}(u - u_B, \bar{v}) + (\sigma(u - \bar{u}_B), \bar{v}) = 0 \qquad \text{for all } \bar{v} \in \mathcal{P}_0^0(\mathcal{B}). \tag{4.5}$$

Let $\chi \in \mathcal{P}_0^1(\mathcal{T})$; then. from (4.5) we have

$$\bar{a}(u - \chi, \bar{v}) + (\sigma(u - \bar{\chi}), \bar{v}) = \bar{a}(u_B - \chi, \bar{v}) + (\sigma(\bar{u}_B - \bar{\chi}), \bar{v}) \tag{4.6}$$

where $\bar{\chi} = G(\chi)$.

We consider first the right-hand side of (4.6). In the standard basis, the first term corresponds to a scalar product of the form $W^T A V$ where A is symmetric and positive definite and W, V are vectors containing the point values of $u_B - \chi$ and $\bar{v}$, respectively, at the vertices. This matrix is not generally equal to the stiffness matrix for the piecewise linear finite element approximation of the term $\int_\Omega a \nabla u \cdot \nabla v\, dx$ but it is close. In particular, both matrices are comparable to the matrices of section 3 for $a(x) \equiv 1$ and those matrices are equal by Lemma 2.3. In analogy with (2.13), we have for $t \subset \Omega_i$, $w \in \mathcal{P}_0^1(\mathcal{T})$

$$-\int_{\partial b_i \cap t} a \frac{\partial w}{\partial n} \bar{\phi}_i\, ds = \eta(a) \int_t a \nabla w \cdot \nabla \phi_i\, dx.$$

Note $\eta(a) = 1 + O(h_t)$ for smooth a. Thus, using Lemmas 2.1-2.3, we obtain the analogue of (3.12)

$$\sup_{\bar{v} \in \mathcal{P}_0^0(\mathcal{B})} \frac{\bar{a}(u_B - \chi, \bar{v}) + (\sigma(\bar{u}_B - \bar{\chi}), \bar{v})}{|\!|\!| \bar{v} |\!|\!|} \geq C |\!|\!| u_B - \chi |\!|\!| \tag{4.7}$$

where $C = C(\underline{a}, \bar{a}, \bar{\sigma}, \alpha, \delta_0)$ and $|\!|\!| \bar{v} |\!|\!| = |\!|\!| G^{-1}(\bar{v}) |\!|\!|$.

The left-hand side of (4.6) is treated analogously to the corresponding term in (3.12). Using Lemmas 2.1 and 2.2 and (2.4), we obtain

$$\bar{a}(u - \chi, \bar{v}) + (\sigma(u - \bar{\chi}), \bar{v}) \leq C \left\{ |\!|\!| u - \chi |\!|\!| |\!|\!| \bar{v} |\!|\!| + \| u - \bar{\chi} \|_{\mathcal{L}^2(\Omega)} \| \bar{v} \|_{\mathcal{L}^2(\Omega)} \right\} \tag{4.8}$$

where $C = C(\underline{a}, \bar{a}, \bar{\sigma}, \alpha, \delta_0)$. Since $\| \bar{v} \|_{\mathcal{L}^2(\Omega)} \leq C |\!|\!| \bar{v} |\!|\!|$, we have the next theorem.

THEOREM 4.1. *Let u, u_L, u_B be defined by (4.2), (3.4), and (4.3). Assume (2.1), (2.3) and $u \in \mathcal{S}_0^1(\mathcal{T})$. Then*

$$|\!|\!| u - u_B |\!|\!| \leq C \inf_{\chi \in \mathcal{P}_0^1(\mathcal{T})} \{ |\!|\!| u - \chi |\!|\!| + \| u - \bar{\chi} \|_{\mathcal{L}^2(\Omega)} \}$$

where $C = C(\underline{a}, \bar{a}, \bar{\sigma}, \alpha, \delta_0)$ *and* $\bar{\chi} = G(\chi)$.

COROLLARY 4.2.

$$\|\!|u - u_L|\!\| \leq \|\!|u - u_B|\!\| \leq C \left(\|\!|u - u_L|\!\| + \|u - \bar{u}_L\|_{\mathcal{L}^2(\Omega)} \right)$$

where C *is as in Theorem 4.1.*

Acknowledgement. We are indebted to Professor Ivo Babuška of the University of Maryland for several helpful and enlightening discussions during the preparation of this manuscript.

REFERENCES

[1] S. AGMON, *Lectures on Elliptic Boundary Value Problems*, Van Nostrand, New York, 1965.
[2] A. K. AZIZ AND I. BABUŠKA, *Part I, survey lectures on the mathematical foundations of the finite element method*, in The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations, Academic Press, New York, 1972, pp. 1–362.
[3] R. E. BANK, W. FICHTNER, AND D. J. ROSE, *Numerical methods for semiconductor device simulation*, SIAM J. Sci. and Stat. Computing, 4 (1983), pp. 416–435.
[4] R. E. BANK AND D. J. ROSE, *Discretization and multilevel solution techniques for nonlinear elliptic systems*, in Elliptic Problem Solvers II, (G. Birkhoff and A. Schoenstadt, eds.), Academic Press, New York, 1984, pp. 493–505.
[5] H.-O. KRIESS, T. A. MANTEUFFEL, B. SWARTZ, B. WENDROFF, AND A. B. WHITE, *Supra-convergent schemes on irregular grids*, tech. rep., Los Alamos National Laboratory, New Mexico, 1983.
[6] T. A. MANTEUFFEL AND A. B. WHITE, *The numerical solution of second-order boundary value problems on nonuniform meshes*, tech. rep., Los Alamos National Laboratory, New Mexico, 1983.
[7] M. NAKATA, A. WEISER, AND M. F. WHEELER, *Some superconvergence results for mixed finite element methods for elliptic problems on rectangular domains*, tech. rep., Rice University and Exxon Production Research Co., Houston, Texas, 1984.
[8] A. N. TIKHONOV AND A. A. SAMARSKII, *Homogeneous difference shemes on nonuniform nets*, Zh. Vychist. Mat. i. Fiz. V1, (1962), p. (English translation).
[9] R. S. VARGA, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1962.