MATHEMATICS OF COMPUTATION, VOLUME 32, NUMBER 143
JULY 1978, PAGES 725–749

Some Results on the Global Inversion of Bilinear and Quadratic Isoparametric Finite Element Transformations*

By A. E. Frey, C. A. Hall and T. A. Porsching

Abstract. This paper contains sufficient conditions under which a map whose domain is a compact set is a bijection onto a given set. Relative to certain isoparametric finite element maps, one set of conditions involves the nonvanishing of the Jacobian; another the notion of overspill. An algorithm based on elimination is given for the numerical inversion of these maps.

1. Introduction. Let $S = \{(r, s) | 0 \leqslant r, s \leqslant 1\}$, and let $\bar{x}$: $\partial S \longrightarrow R^2$ be a continuous transformation of the boundary of S into the plane. Then the transformation, $\bar{T}$: $S \longrightarrow R^2$, given by

$$\begin{aligned}\bar{T}(r, s) = {} & (1 - r)\bar{x}(0, s) + r\bar{x}(1, s) + (1 - s)\bar{x}(r, 0) \\ & + s\bar{x}(r, 1) - (1 - s)(1 - r)\bar{x}(0, 0) - (1 - s)r\bar{x}(1, 0) \\ & - s(1 - r)\bar{x}(0, 1) - sr\bar{x}(1, 1),\end{aligned} \tag{1}$$

has the property that $\bar{T}(\partial S) = \bar{x}(\partial S)$. Thus, given the four curves $\bar{x}(0, s)$, $\bar{x}(1, s)$, $0 \leqslant s \leqslant 1$; $\bar{x}(r, 0)$, $\bar{x}(r, 1)$, $0 \leqslant r \leqslant 1$, the transformation (1) maps the boundary of the unit square onto these and "fills in" the remaining points from the interior of S. As such, (1) represents an interpolation formula and indeed has been termed a "transfinite bilinearly blended" interpolation formula by Gordon and Hall [5].

In this paper we investigate conditions under which the mapping (1) is a bijection from S to a closed, bounded set E having $\bar{x}(\partial S)$ as its boundary. In particular, we consider the cases when the curves $\bar{x}(0, s)$, $\bar{x}(1, s)$, $\bar{x}(r, 0)$ and $\bar{x}(r, 1)$ are either four straight line segments specified by the four nodes (points) $\bar{x}(i, j)$, $i, j = 0, 1$, or four parabolic arcs specified by the eight nodes $\bar{x}(i, j)$, $\bar{x}(\frac{1}{2}, j)$, $\bar{x}(i, \frac{1}{2})$, $i, j = 0, 1$. Then (1) reduces, respectively, to the well-known *bilinear* or *quadratic* isoparametric transformations of finite element analyses, and E is known as the four- or eight-node isoparametric element [3], [10] (see Figure 1).

Considerations concerning the bijective nature of isoparametric transformations are important from both the theoretical and practical points of view. For instance, the numerical solution of boundary value problems by finite element techniques employing isoparametric elements requires the evaluation of certain integrals by means of

---

Received December 6, 1976; revised August 8, 1977.
AMS (MOS) subject classifications (1970). Primary 65N30, 65H10.
*This research supported by the Air Force Office of Scientific Research under Contract No. F44620-76-C-0104.

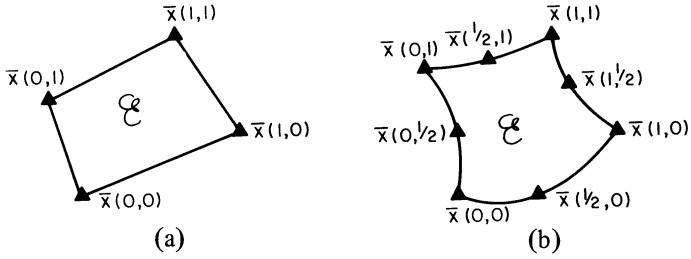


FIGURE 1

(a) *4-node isoparametric element.* (b) *8-node isoparametric element*

the change of variables defined by $\bar{T}$. Thus, knowledge of the bijectivity of $\bar{T}$ is necessary to insure that this change of variables is in fact proper. Furthermore, after the isoparametric finite element solution has been found, the actual inversion of (1) is necessary to obtain values of the dependent variables, such as stress, at prescribed points of E. Therefore, in addition to establishing the a priori existence of an inverse of $\bar{T}$, it is also useful to have an algorithm for its pointwise inversion.

In the next section of this paper, we recall an early theorem of de la Vallée Poussin, relating the bijectivity of a smooth transformation of a compact domain to the nonvanishing of its Jacobian. We then use this result to establish computable sufficient conditions for: (a) the bilinear transformations, (b) a special class of quadratic transformations called semi-rectangles, and (c) other general quadratic transformations.

The notion of "no overspill" is introduced in Section 3 and is shown to be a necessary and sufficient condition for a certain subclass of the quadratic transformations to be bijections. Finally, in Section 4 we develop an elimination algorithm for the numerical inversion of the bilinear and quadratic transformations, and illustrate its effectiveness by several examples.

2. The Jacobian and Global Invertibility. Clearly, if $\bar{x}\colon \partial S \to R^2$ is not an injection, then $\bar{T}\colon S \to R^2$ as defined by (1) cannot be a bijection to any set having $\bar{x}(\partial S)$ as its boundary. Therefore, we state the following fundamental

Boundary Hypothesis: *The continuous transformation* $\bar{x}\colon \partial S \to R^2$ *is an injection.*

This condition is obviously equivalent to hypothesizing that $\bar{x}(\partial S)$ is a simple closed curve. Under the boundary hypothesis, we know from the Jordan Curve Theorem that $\bar{x}(\partial S)$ partitions the plane into two disjoint, open, connected sets and forms their common boundary. Furthermore, only one of these sets is bounded and in the sequel *it is the closure of this bounded set that we take as the set* E.

THEOREM 1. *Let* $\bar{T}$, *as defined by* (1), *be a continuously differentiable transformation on an open set* $T \supset S$. *If the boundary hypothesis holds, and if the Jacobian of* $\bar{T}$ *does not vanish on* T, *then* $\bar{T}$ *is a bijection from* S to E.

Proof. The theorem is essentially a rewording of a result of de la Vallée Poussin [9, p. 355], and the reader is referred to this reference for the details of the proof.

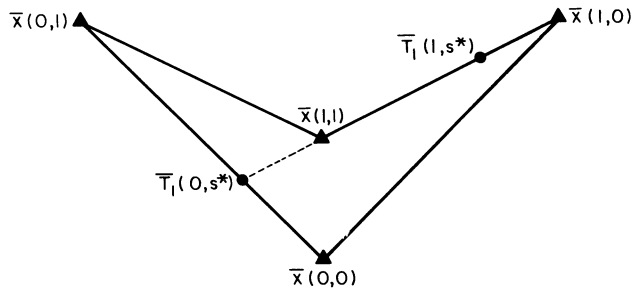


FIGURE 2
Nonconvex 4-node element E

We remark that de la Vallée Poussin's theorem is also cited without proof in [1] and [4]. Q.E.D.

2.1. *Bilinear Transformations. The bilinear isoparametric transformation* results from (1) when the four nodes $\bar{x}(i, j)$, $i, j = 0, 1$, are given and $\bar{x}(\partial S)$ is defined by

$$\bar{x}(0, s) \equiv (1 - s)\bar{x}(0, 0) + s\bar{x}(0, 1),$$
$$\bar{x}(1, s) \equiv (1 - s)\bar{x}(1, 0) + s\bar{x}(1, 1),$$
$$\bar{x}(r, 0) \equiv (1 - r)\bar{x}(0, 0) + r\bar{x}(1, 0),$$
$$\bar{x}(r, 1) \equiv (1 - r)\bar{x}(0, 1) + r\bar{x}(1, 1).$$

In this case, if we denote the left side of (1) by $\bar{T}_1(r, s)$, it follows that

$$(2)\quad \bar{T}_1(r, s) = (1 - r)(1 - s)\bar{x}(0, 0) + r(1 - s)\bar{x}(1, 0) + rs\bar{x}(1, 1) + (1 - r)s\bar{x}(0, 1).$$

THEOREM 2. *Consider the transformation* $\bar{T}_1$ *and assume that the boundary hypothesis holds. Then the following conditions are equivalent*:

(i) *The four-node isoparametric element* E *is convex.*

(ii) *The Jacobian of* $\bar{T}_1$ *is positive at the four vertices of* S (*i.e.* $(r, s) = (i, j)$, $i, j = 0, 1$).

(iii) $\bar{T}_1$ *is a bijection from* S *to* E.

Proof. That (i) implies (ii) is shown by Strang and Fix [8, p. 157] (see also Ciarlet and Raviart [2]). To show that (ii) implies (iii) we note that $\bar{T}_1$ is continuously differentiable in R^2. Thus the only hypothesis of Theorem 1 that requires verification is the nonvanishing of the Jacobian of $\bar{T}_1$ in some open set containing S. By continuity, it is sufficient to have the Jacobian nonzero in S. But we find by direct computation that this determinant is a linear function of r and s and so, if (ii) holds, is in fact positive in S.

To prove that (iii) implies (i) we suppose that (i) does not hold, and for definiteness assume that the reentrant corner of E is at node $\bar{x}(1,1)$ as shown in Figure 2. From (2) it follows that the image of any coordinate line s = constant in S is a straight line segment whose endpoints lie on the sides $\bar{x}(0, s)$ and $\bar{x}(1, s)$, $0 \leqslant s \leqslant 1$. Moreover, as s varies continuously from 0 to 1, these endpoints move in a continuous, strictly monotone manner from $\bar{x}(0, 0)$ to $\bar{x}(0, 1)$ and from $\bar{x}(1, 0)$ to $\bar{x}(1, 1)$. Thus

for some s^*, $0 < s^* < 1$, $T_1(r, s^*) \cap \{\bar{x}(1, s), 0 \leqslant s \leqslant 1\}$ is a nondegenerate line segment. Since this segment is also the image of a portion of the line $r = 1$, $0 \leqslant S \leqslant 1$, $\bar{T}_1$ cannot be an injection on S. Q.E.D.

2.2. *Quadratic Transformations.* Now suppose that the eight nodes $\bar{x}(i, j)$, $\bar{x}(\frac{1}{2}, j)$, $\bar{x}(i, \frac{1}{2})$, $i, j = 0, 1$, are given; cf. Figure 1(b). We define $\bar{x}(\partial S)$ by

$$
\begin{aligned}
&\bar{x}(0, s) \equiv 2(s - \tfrac{1}{2})(s - 1)\bar{x}(0, 0) - 4s(s - 1)\bar{x}(0, \tfrac{1}{2}) + 2s(s - \tfrac{1}{2})\bar{x}(0, 1),\\
&\bar{x}(1, s) \equiv 2(s - \tfrac{1}{2})(s - 1)\bar{x}(1, 0) - 4s(s - 1)\bar{x}(1, \tfrac{1}{2}) + 2s(s - \tfrac{1}{2})\bar{x}(1, 1),\\
&\bar{x}(r, 0) \equiv 2(r - \tfrac{1}{2})(r - 1)\bar{x}(0, 0) - 4r(r - 1)\bar{x}(\tfrac{1}{2}, 0) + 2r(r - \tfrac{1}{2})\bar{x}(1, 0),\\
&\bar{x}(r, 1) \equiv 2(r - \tfrac{1}{2})(r - 1)\bar{x}(0, 1) - 4r(r - 1)\bar{x}(\tfrac{1}{2}, 1) + 2r(r - \tfrac{1}{2})\bar{x}(1, 1).
\end{aligned} \tag{3}
$$

When this is used in conjunction with (1), the resulting transformation, which we denote by $\bar{T}_2(r, s)$, is called the 8-node *quadratic isoparametric transformation.* Of course, Theorem 1 again applies. However, we have been unable to find an analogue of Theorem 2 relating bijectivity directly to an obvious geometric property of the set E.

2.2.1. *Semirectangles.* Let $\bar{P}_i = (x_i, y_i)$, $i = 1, \ldots, 8$, denote the given nodes, where $\bar{x}(0, 0) = \bar{P}_1$, $\bar{x}(1, 0) = \bar{P}_2$, $\bar{x}(1, 1) = \bar{P}_3$, $\bar{x}(0, 1) = \bar{P}_4$, $\bar{x}(\frac{1}{2}, 0) = \bar{P}_5$, $\bar{x}(1, \frac{1}{2}) = \bar{P}_6$, $\bar{x}(\frac{1}{2}, 1) = \bar{P}_7$ and $\bar{x}(0, \frac{1}{2}) = \bar{P}_8$. We consider a special class of quadratic isoparametric transformations obtained by requiring that the boundary transformation $\bar{x}$ satisfy:

(a) $\bar{x}(r, 0) = (x_2 r, 0)$,

(b) $\bar{x}(0, s) = (0, y_4 s)$,

(c) under componentwise ordering, $\bar{x}(r, 1) \geqslant (0, \epsilon)$ and $\bar{x}(1, s) \geqslant (\epsilon, 0)$ for some $\epsilon > 0$,

(d) $x_7 = x_3/2$ and $y_6 = y_3/2$.

If (a)–(d) and the boundary hypothesis hold, we call the set E a *semirectangle.* Figure 3(a) shows a typical semirectangular element. In this case, the components $x(r, s)$, $y(r, s)$ of $\bar{T}_2$ assume the form

$$x(r, s) = (\alpha_0 s^2 + \alpha_1 s + \alpha_2)r,$$

$$y(r, s) = (\beta_0 r^2 + \beta_1 r + \beta_2)s,$$

where

$$\alpha_0 = 4\left(\frac{x_2 + x_3}{2} - x_6\right), \qquad \alpha_1 = x_3 - x_2 - \alpha_0, \qquad \alpha_2 = x_2,$$

$$\beta_0 = 4\left(\frac{y_3 + y_4}{2} - y_7\right), \qquad \beta_1 = y_3 - y_4 - \beta_0, \qquad \beta_2 = y_4.$$

Thus, if $\det J$ denotes the Jacobian of $\bar{T}_2$,

$$\det J = (\beta_0 r^2 + \beta_1 r + \beta_2)(\alpha_0 s^2 + \alpha_1 s + \alpha_2) - rs(2\beta_0 r + \beta_1)(2\alpha_0 s + \alpha_1).$$

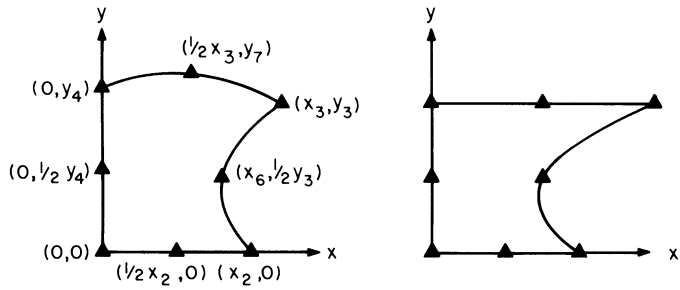


FIGURE 3
Semirectangles

Since $\det J > 0$ when $r = s = 0$, $\det J \neq 0$ on S if and only if

$$\max_{(r,s)\in S} f(r)g(s) < 1, \tag{4}$$

where

$$f(r) = \frac{r(2\beta_0 r + \beta_1)}{\beta_0 r^2 + \beta_1 r + \beta_2}, \qquad g(s) = \frac{s(2\alpha_0 s + \alpha_1)}{\alpha_0 s^2 + \alpha_1 s + \alpha_2}.$$

Note that by (c) both denominators are positive on S. Now

$$\max_{(r,s)\in S} f(r)g(s) = \max(M_f M_g, M_f m_g, m_f M_g, m_f m_g), \tag{5}$$

where $M_f = \max_{0\leqslant r\leqslant 1} f(r)$, $m_f = \min_{0\leqslant r\leqslant 1} f(r)$, etc. So in any given instance it is a straightforward exercise to compute the right side of (5) and test (4) (interior critical points of f and g are solutions of simple quadratic equations, e.g. $\beta_0\beta_1 r^2 + 4\beta_0\beta_2 r + \beta_1\beta_2 = 0$).

There are a number of conditions which imply the validity of (4). We are content to note that the simplest of these occurs when either f or g vanishes identically, that is, when the semirectangle has two parallel sides (Figure 3(b)). Thus, $\bar{T}_2$ is a bijection from S to any such semirectangle.

2.2.2. *Perturbations.* Let the convex quadrilateral Q have vertices and side midpoints $\bar{Q}_i$, $i = 1, \ldots, 8$, shown in Figure 4. Then the associated quadratic transformation defined by (1) and (3) and the nodes $\bar{Q}_i$, $i = 1, \ldots, 8$, is in fact bilinear and by Theorem 2 has a positive Jacobian on S. A nondegenerate quadratic transformation may be obtained by perturbing the midside nodes from ∂Q. By continuity, the Jacobian remains positive for all sufficiently small perturbations. In the remainder of this section, we develop bounds on the size of perturbations which guarantee that the associated transformation has a positive Jacobian on S.

Suppose that the transformation $\bar{T}_2$ is defined by the nodes $\bar{P}_i = (x_i, y_i)$, where $\bar{P}_i = \bar{Q}_i$, $i = 1, \ldots, 4$, $\bar{P}_i = \bar{Q}_i + \bar{\eta}_i$, $i = 5, \ldots, 8$. We consider the class of perturbations $\bar{\eta}_i$ for which $\bar{\eta}_5 = (0, \eta_5)$, $\bar{\eta}_6 = (\eta_6, 0)$, $\bar{\eta}_7 = (0, \eta_7)$, $\bar{\eta}_8 = (\eta_8, 0)$, and we assume that $x_1 < x_2$, $x_4 < x_3$, $y_1 < y_4$, $y_2 < y_3$. See Figure 4.

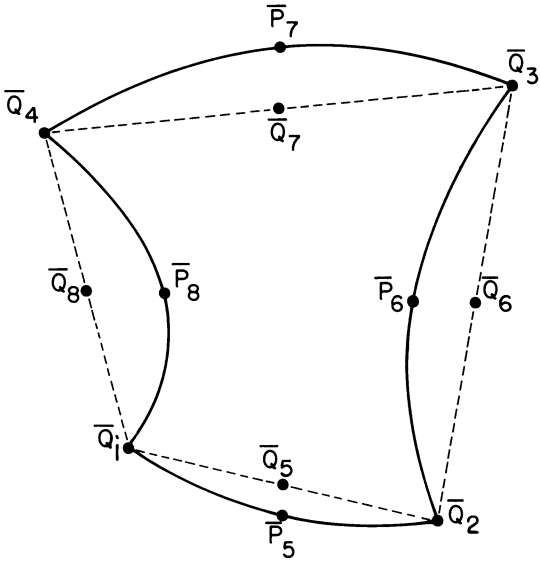


FIGURE 4

Quadratic element obtained by perturbations

Under these assumptions, the components $x(r, s), y(r, s)$ of $\overline{T}_2$ may be written

$$x(r, s) = (1 - r)x(0, s) + rx(1, s), \qquad y(r, s) = (1 - s)y(r, 0) + sy(r, 1).$$

Hence, for its Jacobian $\det J$, we have

$$\det J = \det \begin{bmatrix} x(1, s) - x(0, s) & (1 - r)x_s(0, s) + rx_s(1, s) \\ (1 - s)y_r(r, 0) + sy_r(r, 1) & y(r, 1) - y(r, 0) \end{bmatrix}.$$

It is convenient now to let $\eta_i = \epsilon_i/4$, $i = 5, 6, 7, 8$,

$\bar{\epsilon} = (\epsilon_5, \epsilon_6, \epsilon_7, \epsilon_8)^T$ and $x_{ij} \equiv x_i - x_j$, $y_{ij} \equiv y_i - y_j$, $i, j = 1, 2, 3, 4$.

By direct calculation

$$\det J(\bar{\epsilon}) = A_0 + \sum_{i=5}^{8} A_i\epsilon_i + \sum_{5 \leqslant i < j \leqslant 8} A_{ij}\epsilon_i\epsilon_j,$$

where

$$\begin{aligned}
A_0 &= \det J(\bar{0}) = [(1 - s)x_{21} + sx_{34}]\,[(1 - r)y_{41} + ry_{32}] \\
&\quad - [(1 - s)y_{21} + sy_{34}]\,[(1 - r)x_{41} + rx_{32}], \\
A_5 &= x_{21}(r^2 - r) - (x_{41} - x_{32})s(r^2 - r) - (1 - 2r)(1 - s)[x_{41}(1 - r) + x_{32}r], \\
A_6 &= -y_{32}(s^2 - s) - (y_{21} - y_{34})\,(1 - r)(s^2 - s) - r(1 - 2s)[y_{21}(1 - s) + y_{34}s], \\
A_7 &= -x_{34}(r^2 - r) - (x_{41} - x_{32})(r^2 - r)(1 - s) - (1 - 2r)s[x_{41}(1 - r) + x_{32}r], \\
A_8 &= y_{41}(s^2 - s) - (y_{21} - y_{34})r(s^2 - s) - (1 - r)(1 - 2s)[y_{21}(1 - s) + y_{34}s], \\
A_{56} &= -rs(1 - r)(1 - s) + (1 - 2r)(1 - s)(1 - 2s)r, \\
A_{58} &= rs(1 - r)(1 - s) - (1 - 2r)(1 - s)(1 - 2s)(1 - r), \\
A_{67} &= rs(1 - r)(1 - s) - (1 - 2r)s(1 - 2s)r, \\
A_{78} &= -rs(1 - r)(1 - s) - (1 - 2r)s(1 - 2s)(1 - r),
\end{aligned}$$

and the other $A_{ij} = 0$.

Clearly, $\det J(\bar{\epsilon}) > 0$ on S if

$$\left| \sum_{i=5}^{8} A_i \epsilon_i + \sum_{5 \leqslant i < j \leqslant 8} A_{ij} \epsilon_i \epsilon_j \right| < \det J(\bar{0})$$

for $(r, s) \in S$. But this holds with $\|\bar{\epsilon}\| = \max |\epsilon_i|$ and

$$m \leqslant \min_S \det J(\bar{0}), \quad B_1 \geqslant \max_S \sum_{i=5}^{8} |A_i|, \quad B_2 \geqslant \max_S \sum_{5 \leqslant i < j \leqslant 8} |A_{ij}|$$

if we have

$$B_1 \|\bar{\epsilon}\| + B_2 \|\bar{\epsilon}\|^2 < m,$$

or

$$\|\bar{\epsilon}\| < \frac{-B_1 + \sqrt{B_1^2 + 4B_2 m}}{2B_2}.$$

To obtain constants B_1 and B_2, we first note that after a somewhat tedious exercise, one can show

$$\max_S \sum_{i,j} |A_{ij}| = 1;$$

and moreover, the maximum occurs at each of the four corners of S. Hence, we set $B_2 = 1$.

We next observe that A_5 is linear in s, so $|A_5|$ is maximized over S when $s = 0$ or $s = 1$. It then follows that

$$|A_5| \leqslant |x_{21}|/4 + \max\{|x_{41}|, |x_{32}|\}.$$

Similarly,

$$|A_6| \leqslant |y_{32}|/4 + \max\{|y_{21}|, |y_{34}|\},$$
$$|A_7| \leqslant |x_{34}|/4 + \max\{|x_{41}|, |x_{32}|\},$$
$$|A_8| \leqslant |y_{41}|/4 + \max\{|y_{21}|, |y_{34}|\}.$$

Therefore, we can take

$$B_1 = [|x_{21}| + |y_{32}| + |x_{34}| + |y_{41}|]/4 + 2 \max\{|x_{41}|, |x_{32}|\} + 2 \max\{|y_{21}|, |y_{34}|\}. \tag{6a}$$

Finally, since $\det J(\bar{0})$ attains its minimum at a corner, we have

$$m = \min\{y_{41}x_{21} - x_{41}y_{21}, y_{32}x_{21} - x_{32}y_{21}, y_{41}x_{34} - x_{41}y_{34}, y_{32}x_{34} - x_{32}y_{34}\}. \tag{6b}$$

We summarize all of this as follows: *If an 8-node element E is obtained from a convex quadrilateral Q by perturbations of its midside nodes in the manner shown in Figure 4, and if*

$$|\eta_i| \leqslant \frac{-B_1 + \sqrt{B_1^2 + 4m}}{8}, \qquad i = 5, \ldots, 8, \tag{7}$$

where B_1 and m are given by (6a) *and* (6b), *then the Jacobian of the associated quadratic transformation $\bar{T}_2$ is positive on S. Furthermore, if the boundary hypothesis holds, then $\bar{T}_2$ is a bijection from S to E.*

Suppose, for example, that Q is defined by the nodes $\bar{Q}_1 = (0, 0)$, $\bar{Q}_2 = (1, 0)$, $\bar{Q}_3 = (1.3, 1.2)$, $\bar{Q}_4 = (-.2, 1)$. Then $B_1 = 2.175$, $m = 1$; and therefore, $\det J(\epsilon) > 0$ for all $(r, s) \in S$ if $|\eta_i| < .09747$. Figure 5a shows the element obtained when $\eta_i = -.097$, $i = 5, 6, 7, 8$.

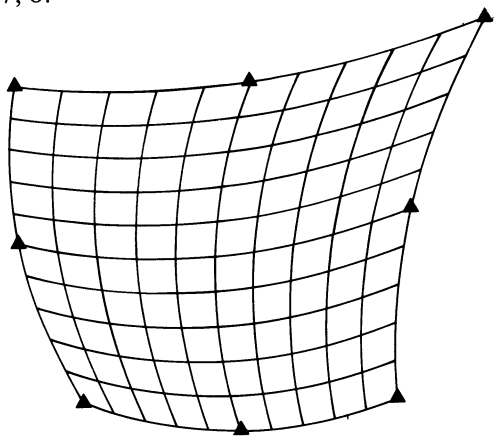


FIGURE 5a

Perturbed element with the associated invertible transformation $\bar{T}_2$. The images of $s = i/10$ and $r = i/10$, $1 \leqslant i \leqslant 10$, are also displayed

It can be shown that choosing $\eta_i = -0.27$, $i = 5, 6, 7, 8$ yields a noninvertible map. A more dramatic example of noninvertibility is shown in Figure 5b. This map results from the choice $\eta_5 = \eta_6 = -\eta_7 = .5$ and $\eta_8 = .1$.

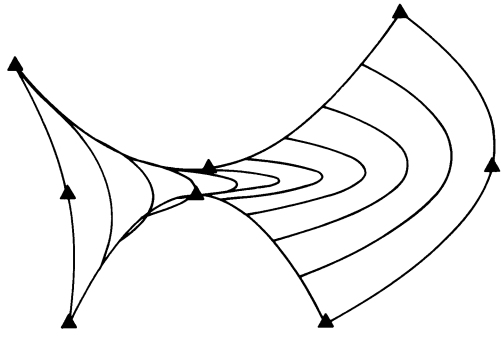


FIGURE 5b

Perturbations too large in magnitude. The associated transformation $\bar{T}_2$ is noninvertible

3. The No Overspill Property and Global Invertibility. The term *overspill* has been used to describe instances when $\bar{T}$: $S \longrightarrow R^2$ is such that $\bar{T}(S)$ properly contains E, a specified set (element), [10]. When such is the case, the image of a constant coordinate line, for example $s = s^*$ originates at one boundary curve of E, say at $\bar{x}(0, s^*)$ extends "beyond" $\bar{x}(1, s)$ and returns by design to terminate at $\bar{x}(1, s^*)$; it overspills the set E. In other words, the image of some constant coordinate line intersects ∂E in more than two points. Formally, we say that the transformation $\bar{T}$ of (1) has the *no overspill property* if $\bar{T}\bar{z} \notin \bar{x}(\partial S)$ when $\bar{z} \in S^\circ$, where S° denotes the interior of S.

LEMMA 1. *Let the boundary hypothesis hold, and let $\bar{T}$ have the no overspill property. Then $\bar{T}(S) \subseteq E$.*

Proof. Let $[\bar{a}, \bar{b}]$ be any line segment in S°. By the continuity of $\bar{T}$ and the no overspill property, there exists an $\epsilon > 0$ such that $\text{dist}(\bar{T}\bar{z}, \partial E) > \epsilon$ for all $\bar{z} \in [\bar{a}\,\bar{b}]$, and moreover, $\bar{T}\bar{z}_0 \in E^\circ$ for some $\bar{z}_0 \in S^\circ$, i.e. $\bar{T}$ cannot be an inversion in ∂E. Suppose that for some $\bar{z} \in S^\circ$, $\bar{T}\bar{z} \notin E$. Since $[\bar{z}, \bar{z}_0] \in S^\circ$, we can apply a bisection argument to the segment $[\bar{z}, \bar{z}_0]$ to deduce that for any $\delta > 0$, there exists $\bar{z}_n, \bar{z}_{n+1}$ in $[\bar{z}, \bar{z}_0]$ such that $|\bar{z}_n - \bar{z}_{n+1}| < \delta$, $\bar{T}\bar{z}_{n+1} \notin E$, and $\bar{T}\bar{z}_n \in E$. But, $|\bar{T}\bar{z}_n - \bar{T}\bar{z}_{n+1}| > \text{dist}(\bar{T}\bar{z}_n, \partial E) > \epsilon$, which contradicts the continuity of $\bar{T}$. Q.E.D.

It is clear that no overspill is a necessary condition for $\bar{T}$ to be a bijection of S to an element E having $\bar{x}(\partial S)$ as its boundary. It is also sufficient for a large subclass of the quadratic isoparametric transformation $\bar{T}_2$ defined by (3). To define this subclass, we begin with a result concerning a parametrized curve.

LEMMA 2. *Consider the curve*

$$\bar{z}(t) = 2(t - \tfrac{1}{2})(t - 1)\bar{Q}_1 + 4t(1 - t)\bar{Q}_2 + 2t(t - \tfrac{1}{2})\bar{Q}_3, \qquad 0 \leqslant t \leqslant 1,$$

where $\bar{Q}_i = (x_i, y_i)$ are three given noncollinear points in the (x, y) plane. Then $\bar{z}(t)$ is an arc of a parabola with axis parallel to the y-coordinate axis if and only if $x_2 = (x_1 + x_3)/2$, $x_1 \neq x_3$.

Proof. The x-coordinate of $\bar{z}(t)$ is linear in t if and only if the hypothesis holds. Q.E.D.

As in Section 2, we denote the eight nodes appearing in (3) by $\bar{P}_i = (x_i, y_i)$, $i = 1, \ldots, 8$. Now suppose that they satisfy (cf. Figure 4)

Assumption A1.

$$\begin{aligned} x_5 &= \tfrac{1}{2}(x_1 + x_2), \quad x_7 = \tfrac{1}{2}(x_3 + x_4), \quad x_1 \neq x_2, \qquad x_4 \neq x_3, \\ y_6 &= \tfrac{1}{2}(y_2 + y_3), \quad y_8 = \tfrac{1}{2}(y_1 + y_4), \quad y_1 \neq y_4, \qquad y_2 \neq y_3, \end{aligned}$$

and let the boundary hypothesis hold. According to Lemma 2, the boundary of E consists of four parabolic arcs, two of which have the generic functional form $y = f(x) \equiv ax^2 + bx + c$, and two of which have the form $x = g(y) \equiv Ay^2 + By + C$. See Figure 6. As the following lemma shows, this is also true of the images of the r and s coordinate lines.

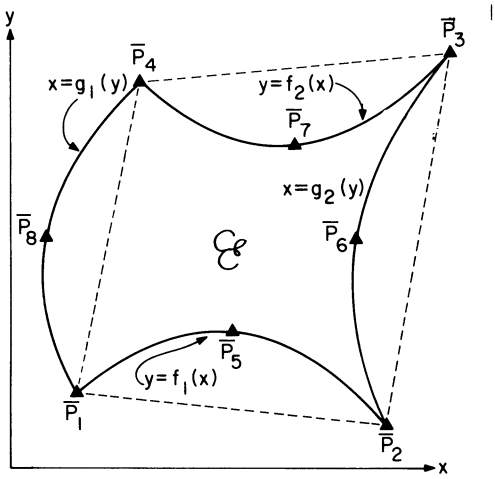


FIGURE 6

An 8-node element with parabolic boundary segments

LEMMA 3. *Under Assumption* A1, *the curve* $\bar{T}_2(r, s^*) \equiv (x(r, s^*), y(r, s^*))$ *is a parabola* $y = f(x)$ *with axis parallel to the y-axis for each fixed* $s = s^*$, *and* $\bar{T}_2(r^*, s)$ *is a parabola* $x = g(y)$ *with axis parallel to the x-axis for each fixed* $r = r^*$.

Proof. We consider only the case $s = s^*$ since the $r = r^*$ case follows a similar argument. From the formula for $\bar{T}_2(r, s)$ we have that the coordinate

$$\begin{aligned} x(r, s^*) = {} & (1 - s^*)[(x_2 - x_1)r + x_1] + s^*[(x_3 - x_4)r + x_4] \\ & + \tfrac{1}{2}(1 - r)s^*(s^* - 1)[4x_1 - 8x_8 + 4x_4] \\ & + \tfrac{1}{2}rs^*(s^* - 1)[4x_2 - 8x_6 + 4x_3] \end{aligned}$$

is linear in r. Hence, we can solve for r and substitute into $y(r, s^*)$ to get a quadratic in x. Q.E.D.

The final assumption that we need to define the subclass concerns the rates of change of the tangents of a typical pair of parabolas $y = f(x)$ and $x = g(y)$ appearing in Lemma 3. Specifically, we assume that

$$\max_{(x,f(x))\in E} |f''(x)| < \min_{(x,g^{-1}(x))\in E} |[g^{-1}(x)]''|.$$

In terms of the transformation $\bar{T}_2(r, s) \equiv (x(r, s), y(r, s))$, this becomes

Assumption A2.

$$\max_{0\leqslant r,s\leqslant 1} \left| \frac{\partial^2 y}{\partial r^2} \left(\frac{\partial x}{\partial r} \right)^{-3} \right| < \min_{0\leqslant r,s\leqslant 1} \left| \frac{\partial^2 x}{\partial s^2} \left(\frac{\partial x}{\partial s} \right)^{-3} \right|.$$

Note that $\partial^2 y/\partial r^2$ and $\partial^2 x/\partial s^2$ are respectively independent of r and s. Moreover, if Assumption A1 holds, then $\partial x/\partial r$ also does not depend on r, and the left side of the above inequality is a function of s only. We will discuss the implications of this assumption in more detail later. Now, however, we prove the main result of this section.

THEOREM 3. *Let the transformation* $\bar{T}_2$ *be defined by* (1) *and* (3). *Assume that the boundary hypothesis holds and that Assumptions* A1 *and* A2 *are true. If* $\bar{T}_2$ *has the no overspill property, then it is a bijection from* S *to* E.

Proof. We first show that $\bar{T}_2$ is an injection on S. Suppose, on the contrary, that $\bar{T}_2(r_1, s_1) = \bar{T}_2(r_2, s_2)$ for $r_1 < r_2$. Then the boundary hypothesis and the no overspill assumptions imply that the curve $\bar{T}_2(r, s_1)$ intersects the curve $\bar{T}_2(r_2, s)$ at the two distinct points $\bar{T}_2(r_i, s_i) \in E^\circ$, $i = 1, 2$. Moreover, $\bar{T}_2(r, s_1)$, $0 < r < 1$, and $\bar{T}_2(r_2, s)$, $0 < s < 1$, are in E°.

By Lemma 3 we can reparametrize the curves $\bar{T}_2(r, s_1)$, $0 \leqslant r \leqslant 1$, and $\bar{T}_2(r_2, s)$, $0 \leqslant s \leqslant 1$, as the parabolic arcs $y = f(x)$, $x_0 \leqslant x \leqslant x_1$, and $x = g(y)$, $y_0 \leqslant y \leqslant y_1$. Suppose that $g''(y) > 0$, i.e., the parabola opens to the right. (The case $g''(y) < 0$ is even simpler.) Let $\phi_+(x)$ and $\phi_-(x)$ be, respectively, the increasing and decreasing functions which are inverse to $g(y)$. Since the curves $x = g(y)$ and $y = f(x)$ are assumed to intersect twice, there is an x^* such that either $f(x^*) = \phi_+(x^*)$, $0 < \phi'_+(x^*) \leqslant f'(x^*)$, or $f(x^*) = \phi_-(x^*)$, $f'(x^*) \leqslant \phi'_-(x^*) < 0$. But A2 requires that in the first case $\phi''_+(x) < f''(x)$, $x^* \leqslant x \leqslant \min(x_1, g(y_2))$, and in the second case $f''(x) < \phi''_-(x)$, $x^* \leqslant x \leqslant \min(x_1, g(y_1))$. Hence, in each case we can show that either $\bar{T}_2(1, s_1)$ lies on the same side of $x = g(y)$ as $\bar{T}_2(0, s_1)$ or the no overspill assumption is violated. These contradictions then establish that $\bar{T}_2$ is an injection on S.

To prove that $\bar{T}_2$ is onto E, let $\bar{P}$ be any point in the interior of E. Since we have just shown that $\bar{T}_2$ is an injection on S, the curve $\bar{T}_2(r, s_1)$, $s_1 = \frac{1}{2}$, divides E into two subsets, $E(0, \frac{1}{2})$ and $E(\frac{1}{2}, 1)$, each having a simple closed curve as its boundary. If $\bar{P} \in \bar{T}_2(r, \frac{1}{2})$, the proof is complete. Otherwise $\bar{P}$ is in the interior of either $E(0, \frac{1}{2})$ or $E(\frac{1}{2}, 1)$, and we can repeat the subdivision process by using either $\bar{T}_2(r, s_2)$, $s_2 = \frac{1}{4}$, or $\bar{T}_2(r, s_2)$, $s_2 = \frac{3}{4}$. Continuing in this way, we generate a convergent (possibly finite) sequence $s_n \longrightarrow s^*$ such that $\bar{P} \in T_2(r, s^*)$. Q.E.D.

The above proof of the surjectivity portion of Theorem 3 is essentially constructive in nature. However, a more general result can be obtained in a nonconstructive manner by the use of degree theory. For the definition of the *degree* of a transformation, as well as the fundamental properties of degree, the reader is referred to [7, Chapter 6].

THEOREM 4. *Let* $\bar{T}$, *as defined by* (1), *be a continuously differentiable transformation on an open set* $T \supset S$, *and let the boundary hypothesis hold. Let there be a point* (r_0, s_0) *in* S° *satisfying the following three conditions*: (i) $\bar{T}(r_0, s_0) \in E^\circ$; (ii) $\bar{T}(r, s) = \bar{T}(r_0, s_0)$ *implies that* $r = r_0$, $s = s_0$; (iii) *the Jacobian of* $\bar{T}$ *is not zero at* (r_0, s_0). *Then* $E \subseteq \bar{T}(S)$.

Proof. Let $\bar{P}_0 = \bar{T}(r_0, s_0)$, and denote the degree of $\bar{T}$ at any point $\bar{P} \notin \bar{T}(\partial S)$ with respect to S° by $\deg(\bar{T}, S^\circ, \bar{P})$. Then by the hypotheses $\deg(\bar{T}, S^\circ, \bar{P}_0) = \pm 1$. Furthermore, if $\bar{P}$ is any point in E°, then there is a continuous curve lying in E° with $\bar{P}_0$ and $\bar{P}$ as its endpoints. It follows from the properties of the degree [7, p. 158] that $\deg(\bar{T}, S^\circ, \bar{P}) = \deg(\bar{T}, S^\circ, \bar{P}_0) = \pm 1$, and, hence, the system $\bar{T}(r, s) = \bar{P}$ has a solution in S°. Q.E.D.

The surjectivity part of Theorem 3 now follows from the fact that the Jacobian

of $\bar{T}_2$ cannot vanish at every point in S°. Therefore, any point where the Jacobian is nonzero will serve as the point (r_0, s_0) in Theorem 4 once it is known that $\bar{T}_2$ is an injection on S.

Some remarks on Assumption A2 are now in order. In the first place, if A1 holds, then we find by direct computation, using (1) and (3), that

$$\begin{aligned}
\partial x/\partial r &= x(1, s) - x(0, s),\\
\partial^2 y/\partial r^2 &= 4[(1 - s)(y_1 - 2y_5 + y_2) + s(y_4 - 2y_7 + y_3)],\\
\partial x/\partial s &= (1 - r)[s(4x_1 - 8x_8 + 4x_4) - (3x_1 - 4x_8 + x_4)]\\
&\quad + r[s(4x_2 - 8x_6 + 4x_3) - (3x_2 - 4x_6 - x_3)],\\
\partial^2 x/\partial s^2 &= 4[(1 - r)(x_1 - 2x_8 + x_4) + r(x_2 - 2x_6 + x_3)].
\end{aligned}$$

Therefore, for A2 to hold, it is necessary that $x_1 - 2x_8 + x_4$ and $x_2 - 2x_6 + x_3$ have the same sign. That is, the parabolas $\bar{T}_2(0, s)$ and $\bar{T}_2(1, s)$ should both open to the right or left (see Figure 6). When this is the case, it is easy to see that the right side of the inequality in A2 is bounded below by

$$m \equiv \frac{4 \min(|x_1 - 2x_8 + x_4|, |x_2 - 2x_6 + x_3|)}{d^3}, \tag{8}$$

where

$$d = \max(|4x_6 - 3x_2 - x_3|, |4x_6 - 3x_3 - x_2|, |4x_8 - 3x_1 - x_4|, |4x_8 - 3x_4 - x_1|),$$

and the left side is bounded above by

$$M \equiv \frac{4 \max(|y_1 - 2y_5 + y_2|, |y_4 - 2y_7 + y_3|)}{\left[\min\limits_{0 \leq s \leq 1} |x(1, s) - x(0, s)|\right]^3}. \tag{9}$$

Note that in any given case, it is a simple matter to obtain the quantities m and M, the denominator in M giving rise to an elementary minimization problem via (3). Clearly, A2 holds if $M < m$, which is certainly the case if $\bar{T}_2(r, 0)$ and $\bar{T}_2(r, 1)$ are straight line segments since then $M = 0$.

As a final remark on A2, we note that in Theorem 3, it may be replaced by *Assumption* A2′.

$$\max_{0 \leq r,s \leq 1} \left| \frac{\partial^2 x}{\partial s^2} \left(\frac{\partial y}{\partial s}\right)^{-3} \right| < \min_{0 \leq r,s \leq 1} \left| \frac{\partial^2 y}{\partial r^2} \left(\frac{\partial y}{\partial r}\right)^{-3} \right|.$$

In view of Theorem 3, we now seek conditions which guarantee that quadratic transformations $\bar{T}_2$ satisfying A1 will have the no overspill property. Consider the element E in Figure 4. If both coordinates of the four midside nodes were averaged, E would be the straight sided quadrilateral Q with vertices $\bar{Q}_i$, $i = 1, 2, 3, 4$, as indicated by the dotted lines in Figure 4. In this case $\bar{T}_2 \equiv \bar{T}_1$ and Theorem 2 applies. The element E differs from Q by perturbations η_6 and η_8 in the x-coordinate of $\bar{Q}_6$ and $\bar{Q}_8$, and perturbations η_5 and η_7 in the y-coordinate of $\bar{Q}_5$ and $\bar{Q}_7$. The question is: How large can these perturbations be without producing overspill? Before answering

this question we find it necessary to make a further assumption.

Assumption A3. Assume that for each $r = r^*$ (resp. $s = s^*$) the straight line segment

$$(10)\qquad (1-s)\bar{x}(r^*, 0) + s\bar{x}(r^*, 1), \qquad 0 \leqslant s \leqslant 1,$$
$$(\text{resp. } (1-r)\bar{x}(0, s^*) + r\bar{x}(1, s^*), \ 0 \leqslant r \leqslant 1)$$

intersects each of the boundary curves $\bar{x}(r, 0)$ and $\bar{x}(r, 1)$ (resp. $\bar{x}(0, s)$ and $\bar{x}(1, s)$, $i = 1, 2$, once and only once.

If η_6 and η_8 (resp. η_5 and η_7) are zero, then A3 guarantees that $\bar{T}_2$ is one-to-one since $\bar{T}_2$ is then just a "railing" of the curves $\bar{x}(r, 0)$ and $\bar{x}(r, 1)$, (resp. $\bar{x}(0, s)$ and $\bar{x}(1, s)$). That is,

$$(11)\qquad \bar{T}_2(r, s) = \bar{P}_s(r, s) \equiv (1-s)\bar{x}(r, 0) + s\bar{x}(r, 1)$$
$$(\text{resp. } \bar{T}_2(r, s) = \bar{P}_r(r, s) \equiv (1-r)\bar{x}(s, 0) + r\bar{x}(s, 1)).$$

Conversely, if $\bar{P}_r$ and $\bar{P}_s$ are invertible for $0 \leqslant r, s \leqslant 1$, then Assumption A3 holds, hence we have

LEMMA 4. *Assumption* A3 *holds if and only if* $\bar{P}_r$ *and* $\bar{P}_s$ *are injections on* S.

The importance of Lemma 4 is that the validity of A3 can be established computationally by investigating the invertibility of $\bar{P}_r$ and $\bar{P}_s$. By Theorem 1, we need only establish the nonvanishing of their Jacobians. But, the Jacobian of $\bar{P}_s$ is by direct calculation of the form $(1-s)q_1(r) + sq_2(r)$, $0 \leqslant s \leqslant 1$, where the $q_i(r)$, $i = 1, 2$, are quadratics in r. The nonvanishing of the Jacobian is then established by checking if $q_1(r)q_2(r) > 0$ for $0 \leqslant r \leqslant 1$.

We are now ready to determine bounds on the perturbations η_5, η_6, η_7 and η_8 so as to guarantee that $\bar{T}_2$ has the no overspill property.

THEOREM 5. *Assume that the boundary hypothesis,* A1 *and* A3 *hold and, referring to Figure* 6, *let*

$$x_L = \min_{0\leqslant s\leqslant 1} x(0, s), \quad x_R = \max_{0\leqslant s\leqslant 1} x(1, s),$$
$$y_T = \max_{0\leqslant r\leqslant 1} y(r, 1), \quad y_B = \min_{0\leqslant r\leqslant 1} y(r, 0),$$
$$M_x = \max\left\{ \max_{x_L\leqslant x\leqslant x_R} |df_1/dx|, \max_{x_L\leqslant x\leqslant x_R} |df_2/dx| \right\},$$
$$M_y = \max\left\{ \max_{y_B\leqslant y\leqslant y_T} |dg_1/dy| \max_{y_B\leqslant y\leqslant y_T} |dg_2/dy| \right\},$$
$$S_x = \max_{0\leqslant r\leqslant 1} |x(r, 1) - x(r, 0)|, \quad S_y = \max_{0\leqslant s\leqslant 1} |y(1, s) - y(0, s)|,$$
$$H_x = \min_{0\leqslant s\leqslant 1} (x(1, s) - x(0, s)), \quad H_y = \min_{0\leqslant r\leqslant 1} (y(r, 1) - y(r, 0)),$$
$$y_5 = (y_1 + y_2)/2 + \eta_5, \qquad y_7 = (y_3 + y_4)/2 + \eta_7,$$
$$x_6 = (x_2 + x_3)/2 + \eta_6, \quad \textit{and} \quad x_8 = (x_1 + x_4)/2 + \eta_8.$$

If

$$\eta_x \equiv \max\{|\eta_6|, |\eta_8|\} \leqslant \tfrac{1}{4}\{H_y/M_x - S_x\},$$
$$\eta_y \equiv \max\{|\eta_5|, |\eta_7|\} \leqslant \tfrac{1}{4}\{H_x/M_y - S_y\}, \tag{12}$$

then $\bar{T}_2$ *has the no overspill property.*

Proof. We first show that the intermediate curve $\bar{T}_2(r, s^*)$, for any fixed $0 < s^* < 1$, does not intersect $y = f_2(x)$. By Lemma 3, $\bar{T}_2(r, s^*)$ is a parabola $y = f_3(x)$.

Fix $r = r^*$. Then by (1) and A1, the y-coordinate of $\bar{T}_2(r^*, s)$ is linear in s

$$y(r^*, s) = (1 - s)y(r^*, 0) + sy(r^*, 1).$$

We recognize this as also being the linear parametrization of the y-coordinate of the line $\overline{EF}$ in Figure 7, or equivalently the y-coordinate of the mapping $\bar{P}_s$ in (11). But for $\eta_x \equiv 0$ the hypotheses imply that the line $\overline{EF}$ is contained in E. We now increase η_x, and $\bar{T}_2(r^*, s^*)$ moves off of the line $\overline{EF}$ to a point A. We want to restrict the x-coordinate $x(r^*, s^*)$ so that A is on the same side of $y = f_2(x)$ as B. But such will be the case if

$$\frac{|\overline{EB}|}{|\overline{AB}|} > \frac{|\overline{EH}|}{|\overline{AB}|} = \left|\frac{df_2}{dx}(\xi)\right| \quad \text{for some } a \leqslant \xi \leqslant b. \tag{13}$$

Now it may happen that $a < \min\{x_1, x_4\}$, but the monotonicity of $x(r, s^*)$ implies that $x_L \leqslant x(0, s^*) \leqslant a$. Hence (13) holds if

$$|\overline{AB}| < |\overline{EB}|/M_x.$$

This in turn yields

$$|AB| < \frac{|y(r^*, 1) - y(r^*, s^*)|}{M_x} = \frac{(1 - s^*)(y(r^*, 1) - y(r^*, 0))}{M_x},$$

which is true if

$$|AB| < (1 - s^*)H_y/M_x.$$

Now, from the formula for $x(r^*, s)$ we have

$$\begin{aligned} |AB| &= |x(r^*, 1) - x(r^*, s^*)| \\ &= (1 - s^*)|x(r^*, 1) - x(r^*, 0) - 4s^*((1 - r)\eta_8 + r\eta_6)| \\ &\leqslant (1 - s^*)[S_x + 4\eta_x], \end{aligned}$$

which is bounded by $(1 - s^*)H_y/M_x$ if η_x satisfies (12).

By similar arguments, the curve $y = f_3(x)$ does not intersect $y = f_1(x)$. In this case (12) will guarantee that $|\overline{JA}|/|\overline{JF}| > |\overline{IF}|/|\overline{JF}|$.

Now if some intermediate parabola $x = g(y)$ intersects $y = f_2(x)$ twice, then there is an s^*, $0 < s^* < 1$, such that $\bar{T}_2(r, s^*)$ also intersects $y = f_2(x)$. But we have shown this is impossible, and so for each r^*, $0 < r^* < 1$, $\bar{T}_2(r^*, s)$ intersects ∂E only for $s = 0, 1$.

Interchanging the roles of x and y, we show that no intermediate curve can intersect $x = g_i(y)$, $i = 0, 1$. Q.E.D.

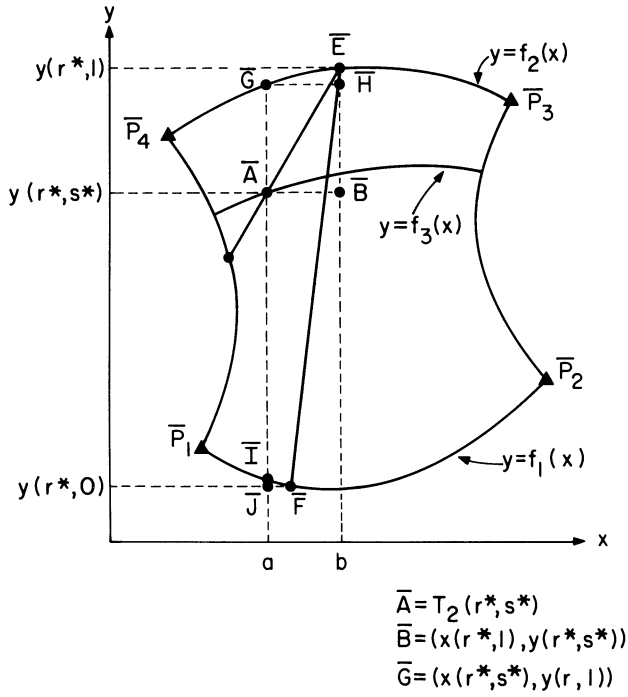


FIGURE 7

Intermediate parabola does not intersect $\bar{T}_2(r, 1)$

Remarks. Note that $S_x = \max\{|x_4 - x_1|, |x_3 - x_2|\}$ since $x(r, 1)$ and $x(r, 0)$ are linear in r. Similarly, $S_y = \max\{|y_1 - y_2|, |y_4 - y_3|\}$. S_x and S_y are measures of the "skewness" of the quadrilateral Q and, hence, the element E. If M_x or M_y is zero, then we interpret the bounds in (12) to be arbitrarily large.

To illustrate how one might use the bounds in (12), consider the following example: Let $\bar{P}_1$: (0, 0), $\bar{P}_2$: (6, 1), $\bar{P}_3$: (5, 5), and $\bar{P}_4$: (0, 4) be the corner nodes and $\bar{P}_6$: (5, 3) and $\bar{P}_8$: (−0.5, 2) be two of the midside nodes of a given element. We consider how the straight lines $\overline{P_1P_2}$ and $\overline{P_3P_4}$ can be deformed into parabolas so as to guarantee that the element does not have the overspill property. A class of such parabolas is described in terms of the perturbations η_5 and η_7 of the y-coordinates of the midside nodes $\bar{P}_5 = (3, 0.5 + \eta_5)$ and $\bar{P}_7 = (2.5, 4.5 + \eta_7)$. Note $\eta_5 = \eta_7 = 0$ corresponds to $\bar{T}_2(r, s) = \bar{P}_r(r, s)$, which is one-to-one. We use Theorem 5 to bound $|\eta_5|$ and $|\eta_7|$ as follows:

1. Compute

$$S_x = \max\{|x_4 - x_1|, |x_3 - x_2|\} = 1,$$

$$S_y = \max\{|y_1 - y_2|, |y_4 - y_3|\} = 1.$$

2. By direct calculation, using (3),

$$H_x = \min_{0 \leq s \leq 1} |x(1, s) - x(0, s)| = 5.$$

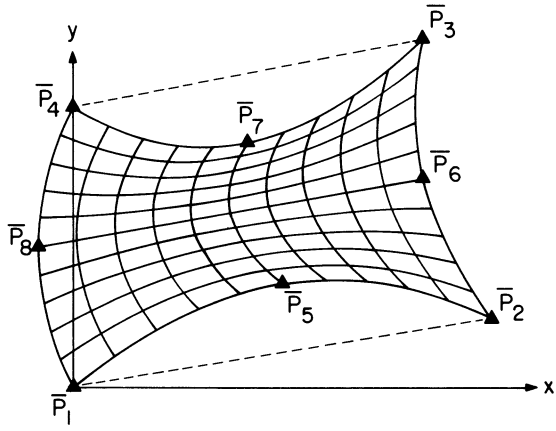


FIGURE 8

A one-to-one isoparametric mapping $\bar{T}_2$

3. We choose $\eta_7 < 0$ and $\eta_5 > 0$, so $y_B = 0$ and $y_T = 5$. By the chain rule

$$\max_{0 \leqslant y \leqslant 5} |dg_1/dy| = .75, \qquad \max_{0 \leqslant y \leqslant 5} |dg_2/dy| = 1.0$$

so $M_y = 1$.

From (12) $\bar{T}_2$ has the no overspill property if η_5 and η_7 are chosen in magnitude less than 1.0. Figure 8 illustrates this extreme case where $\bar{P}_7 = (2.5, 3.5)$ and $\bar{P}_5 = (3.0, 1.5)$.

Referring back to (8) and (9), we see that $m = 4/27$ and $M = 8/125 < m$. Hence, A2 holds and by Theorem 3, for the element in Figure 8, $\bar{T}_2$ is a bijection from S to E.

4. Inversion by Elimination. Calculation of the stiffness matrices involved in the finite element method does *not* require the inversion of any associated isoparametric transformations [10, Chapter 8]. However, computation of displacements or stresses at points (other than nodes) in the x-y coordinate system does in general require numerical inversion of the transformation. For example, suppose that a quadratic transformation $\bar{T}_2$ is used and that stresses along a specific line are desired (cf. Figure 9). Since the basis functions are in terms of the generalized coordinates (r, s), if $\bar{P}$ is given on l, we first must find its preimage $\bar{Q}$ in S. The basis functions or their derivatives are then evaluated at $\bar{Q}$.

Another application where inversion of an isoparametric transformation (or determination of preimages) is of importance is mesh generation. Refining a given mesh about a point $\bar{P}$ can be facilitated by being able to work directly with the (r, s) system once the preimage of $\bar{P}$ has been determined.

In this section, we describe an algorithm based on elimination for the pointwise numerical inversion of the quadratic transformation $\bar{T}_2$. Since the bilinear transformation $\bar{T}_1$ is a degenerate case of $\bar{T}_2$, our method also inverts $\bar{T}_1$.

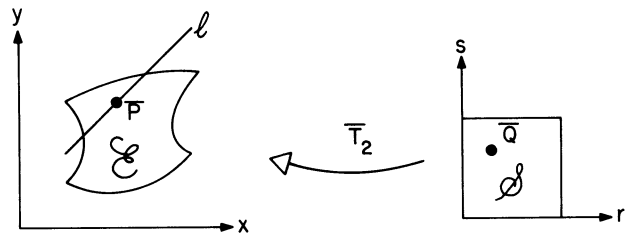


FIGURE 9
The preimage $\bar{Q}$ of $\bar{P}$

A straightforward but tedious calculation shows that $\bar{T}_2$ defined by (1) and (3) can be written as

$$\bar{T}_2(r, s) = \begin{pmatrix} x(r, s) \\ y(r, s) \end{pmatrix} = \begin{pmatrix} a_0(s) \\ b_0(s) \end{pmatrix} r^2 + \begin{pmatrix} a_1(s) \\ b_1(s) \end{pmatrix} r + \begin{pmatrix} A_2(s) \\ B_2(s) \end{pmatrix}, \tag{15}$$

where

$$a_0(s) = \alpha_{00}s + \alpha_{01}, \quad a_1(s) = \alpha_{10}s^2 + \alpha_{11}s + \alpha_{12}, \quad A_2(s) = \alpha_{20}s^2 + \alpha_{21}s + \alpha'_{22},$$

$$b_0(s) = \beta_{00}s + \beta_{01}, \quad b_1(s) = \beta_{10}s^2 + \beta_{11}s + \beta_{12}, \quad B_2(s) = \beta_{20}s^2 + \beta_{21}s + \beta'_{22},$$

and where, with $\bar{P}_i = (x_i, y_i)$, $i = 1, 2, \ldots, 8$, as in Section 2,

$$\begin{aligned}
\alpha_{00} &= -2x_1 - 2x_2 + 2x_3 + 2x_4 + 4x_5 - 4x_7, \\
\alpha_{01} &= 2x_1 + 2x_2 - 4x_5, \\
\alpha_{10} &= -2x_1 + 2x_2 + 2x_3 - 2x_4 - 4x_6 + 4x_8, \\
\alpha_{11} &= 5x_1 - x_2 - 3x_3 - x_4 - 4x_5 + 4x_6 + 4x_7 - 4x_8, \\
\alpha_{12} &= -3x_1 - x_2 + 4x_5, \\
\alpha_{20} &= 2x_1 + 2x_4 - 4x_8, \\
\alpha_{21} &= -3x_1 - x_4 + 4x_8, \\
\alpha'_{22} &= x_1.
\end{aligned}$$

Similar expressions for the β's can be obtained by replacing the x_i's by y_i's in each of the above equations.

If we let $a_2 = A_2 - x$, $b_2 = B_2 - y$, we see that determining inverse images of a point $(x, y) \in E$ under the mapping $\bar{T}_2$ is equivalent to finding the roots, r and s, of the two simultaneous bivariate polynomial equations

$$0 = a_0r^2 + a_1r + a_2, \qquad 0 = b_0r^2 + b_1r + b_2 \tag{16}$$

with the a's and b's defined as above. System (16) may be solved by the method of elimination which we now briefly discuss.

Our discussion is based on Householder's elegant presentation [6]. Consider the seemingly simpler problem of determining all of the common zeros of the two univari-

ate polynomials

$$f(r) = a_0 r^n + a_1 r^{n-1} + \cdots + a_n, \qquad g(r) = b_0 r^m + b_1 r^{m-1} + \cdots + b_m.$$

This is clearly equivalent to finding the roots of the greatest common divisor (g.c.d.) of f and g. One way to generate the g.c.d. is by the use of *bigradients*. There are two classes of bigradients. To define the first class, let $a_i = 0$ if $i > n$, $b_i = 0$ if $i > m$. Then the bigradient $\delta\binom{(a)_i}{(b)_j}$ is the $i + j$ order determinant

$$\delta\begin{pmatrix}(a)_i\\(b)_j\end{pmatrix} = \det \begin{vmatrix} a_0 & a_1 & \cdots & a_{i+j-1} \\ 0 & a_0 & \cdots & a_{i+j-2} \\ & & \cdots & \\ 0 \cdots 0 & a_0 & \cdots & a_j \\ \hline 0 \cdots 0 & b_0 & \cdots & b_i \\ & & \cdots & \\ 0 & b_0 & \cdots & b_{i+j-2} \\ b_0 & b_1 & \cdots & b_{i+j-1} \end{vmatrix},$$

where there are i rows of a's and j rows of b's. The second class of bigradients consists of polynomials in r and is defined by the relation

$$\delta\begin{pmatrix}(f)_i\\(g)_j\end{pmatrix} = \det \begin{vmatrix} a_0 & a_1 & \cdots & a_{i+j-2} r^{j-1} f \\ 0 & a_0 & \cdots & a_{i+j-3} r^{j-2} f \\ & & \cdots & \\ 0 \cdots 0 & a_0 & \cdots & a_{j-1} r^0 f \\ 0 \cdots 0 & b_0 & \cdots & b_{i-1} r^0 g \\ & & \cdots & \\ 0 & b_0 & \cdots & b_{i+j-3} r^{j-2} g \\ b_0 & & \cdots & b_{i+j-2} r^{j-1} g \end{vmatrix}.$$

Although bigradients are defined for any $i, j \geqslant 1$, we are concerned with those for which $i = m - k$, $j = n - k$, $k = 0, 1, \ldots, \min(m, n)$. The relationship of these particular bigradients to the g.c.d. is given by the following two lemmas.

LEMMA 5.

$$\delta\begin{pmatrix}(f)_{m-k}\\(g)_{n-k}\end{pmatrix} = r^k \delta\begin{pmatrix}(a)_{m-k}\\(b)_{n-k}\end{pmatrix} + O(r^{k-1}).$$

LEMMA 6. *The g.c.d. of f and g is $\delta\binom{(f)_{m-k}}{(g)_{n-k}}$ if and only if*

$$0 = \delta\begin{pmatrix}(a)_m\\(b)_n\end{pmatrix} = \delta\begin{pmatrix}(a)_{m-1}\\(b)_{n-1}\end{pmatrix} = \cdots = \delta\begin{pmatrix}(a)_{m-k+1}\\(b)_{n-k+1}\end{pmatrix} \neq \delta\begin{pmatrix}(a)_{m-k}\\(b)_{n-k}\end{pmatrix}.$$

Both lemmas are proven in [6] and as pointed out by Householder, Lemma 6 goes back to Trudi (1862). From these two lemmas, we can draw the following conclusions:

(1) f and g have a common root if and only if the degree of the g.c.d. is greater than or equal to 1, i.e., if and only if $\delta\binom{(a)_m}{(b)_n} = 0$. This particular bigradient is also known as *Sylvester's determinant*, [6] or the *resultant* of f and g. We note that it is independent of r.

(2) The common roots of (16) are exactly those of

$$\delta\begin{pmatrix}(a)_m\\(b)_n\end{pmatrix} = 0 \quad \text{and} \quad \delta\begin{pmatrix}(f)_{m-k}\\(g)_{n-k}\end{pmatrix} = 0,$$

where k is given by Lemma 6. This trivial observation is the key to the method of elimination.

(3) In most cases the value of k in Lemma 6 is unity, whence the g.c.d. is linear in r.

Now consider the system arising from the pointwise inversion of $\bar{T}_2$. From what has been said above, (r, s) is a solution of (16) if and only if s is a zero of

$$\delta\begin{pmatrix}(a)_2\\(b)_2\end{pmatrix} = \det\begin{bmatrix} a_0 & a_1 & a_2 & 0\\ 0 & a_0 & a_1 & a_2\\ 0 & b_0 & b_1 & b_2\\ b_0 & b_1 & b_2 & 0\end{bmatrix} \tag{17}$$

$$= \det\begin{bmatrix} a_0b_1 - a_1b_0 & a_0b_2 - a_2b_0\\ a_0b_2 - a_2b_0 & a_1b_2 - a_2b_1\end{bmatrix} = 0.$$

Furthermore, if

$$\delta\begin{pmatrix}(a)_1\\(b)_1\end{pmatrix} = \det\begin{bmatrix} a_0 & a_1\\ b_0 & b_1\end{bmatrix} \neq 0, \tag{18}$$

then the following equation yields the desired values of r,

$$\delta\begin{pmatrix}(f)_1\\(g)_1\end{pmatrix} = \det\begin{bmatrix} a_0 & f\\ b_0 & g\end{bmatrix} = \det\begin{bmatrix} a_0 & a_1r + a_2\\ b_0 & b_1r + b_2\end{bmatrix} = 0. \tag{19}$$

However, we have

$$a_0b_1 - a_1b_0 = (\alpha_{00}\beta_{10} - \beta_{00}\alpha_{10})s^3 + (\alpha_{00}\beta_{11} + \alpha_{01}\beta_{10} - \beta_{00}\alpha_{11} - \beta_{01}\alpha_{10})s^2$$
$$+ (\alpha_{00}\beta_{12} + \alpha_{01}\beta_{11} - \beta_{00}\alpha_{12} - \beta_{01}\alpha_{11})s + (\alpha_{01}\beta_{12} - \beta_{01}\alpha_{12}),$$

$$a_1b_2 - a_2b_1 = (\alpha_{10}\beta_{20} - \beta_{10}\alpha_{20})s^4 + (\alpha_{10}\beta_{21} + \alpha_{11}\beta_{20} - \beta_{10}\alpha_{21} - \beta_{11}\alpha_{20})s^3$$
$$+ (\alpha_{10}\beta_{22} + \alpha_{11}\beta_{21} + \alpha_{12}\beta_{20} - \beta_{10}\alpha_{22} - \beta_{11}\alpha_{21} - \beta_{12}\alpha_{20})s^2$$
$$+ (\alpha_{11}\beta_{22} + \alpha_{12}\beta_{21} - \beta_{11}\alpha_{22} - \beta_{12}\alpha_{21})s + (\alpha_{12}\beta_{22} - \beta_{12}\alpha_{22}),$$
$$a_0b_2 - a_2b_0 = (\alpha_{00}\beta_{20} - \beta_{00}\alpha_{20})s^3 + (\alpha_{00}\beta_{21} + \alpha_{01}\beta_{20} - \beta_{00}\alpha_{21} - \beta_{01}\alpha_{20})s^2$$
$$+ (\alpha_{00}\beta_{22} + \alpha_{01}\beta_{21} - \beta_{00}\alpha_{22} - \beta_{01}\alpha_{21})s + (\alpha_{01}\beta_{22} - \beta_{01}\alpha_{22}),$$

where $\alpha_{22} = x_1 - x$, $\beta_{22} = y_1 - y$. Equation (17) is seen to be at most a seventh degree polynomial in s, while (19) is linear in r if (18) holds.

It should be pointed out that the above procedure for solving system (16) is based on the implicit assumption that $a_0(s) \cdot b_0(s) \not\equiv 0$. In many situations, however, either $a_0(s) \equiv 0$ or $b_0(s) \equiv 0$. In such cases the above procedure fails, but then, at least one of the equations in (16) is at most linear in r. Let us suppose, for example, that $a_0(s) \not\equiv 0$, $b_0(s) \equiv 0$, and $b_1(s) \not\equiv 0$. Then system (16) reduces to

$$0 = a_0(s)r^2 + a_1(s)r + a_2(s)$$
$$0 = b_1(s)r + b_2(s).$$

The second of these two equations is linear in r, so

$$r = \frac{-b_2(s)}{b_1(s)}. \tag{20}$$

We can substitute (20) into the first equation of (16) to get

$$a_0(s)\left(\frac{-b_2(s)}{b_1(s)}\right)^2 + a_1(s)\left(\frac{-b_2(s)}{b_1(s)}\right) + a_2(s) = 0$$

or

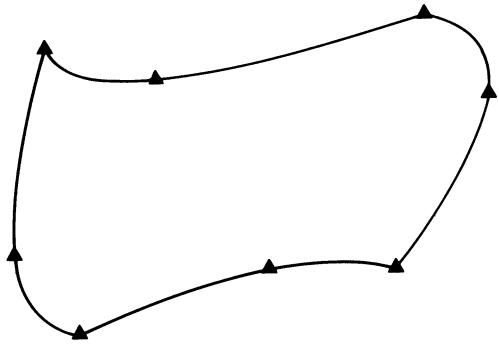
$$a_0(s)[b_2(s)]^2 - a_1(s)b_1(s)b_2(s) + a_2(s)[b_1(s)]^2 = 0. \tag{21}$$

Equation (21) is at most a sixth degree polynomial in the variable s. Thus the solutions of (16) can be found by solving the triangular system consisting of Eqs. (20) and (21). If more of the a's and b's vanish, then the similar procedures can be employed to solve the even simpler resulting system.

We finally note that system (16) may have many solutions, and although we are usually concerned with those solutions $(r, s) \in S$, it is sometimes useful to know the solutions of (16) which lie both inside and outside the unit square (see Example 4).

We conclude this section with some examples illustrating the capabilities of the elimination algorithm.

Example 1. In this example the points $\bar{P}_1, \bar{P}_2, \ldots, \bar{P}_8$ are chosen as shown in Figure 10. The elimination algorithm is used to find the preimages of the points (.5, .5), (.1, .3), (1., .2), (.9, .8), and (.2, .7), all lying in the region E. The preimages of these points are, respectively, (.5220, .5233), (0.2210, 0.4358), (1., 0.), (.8215, .7663) and (.4117, .8822). Figure 11 shows the location of these points.



8 NODE 2-D ELEMENT

(0.00 , 0.00)
(1.00 , 0.20)
(1.10 , 1.00)
(- 0.10 , 0.90)
(0.60 , 0.20)
(1.30 , 0.75)
(0.25 , 0.80)
(- 0.20 , 0.25)

FIGURE 10. *Example* 1

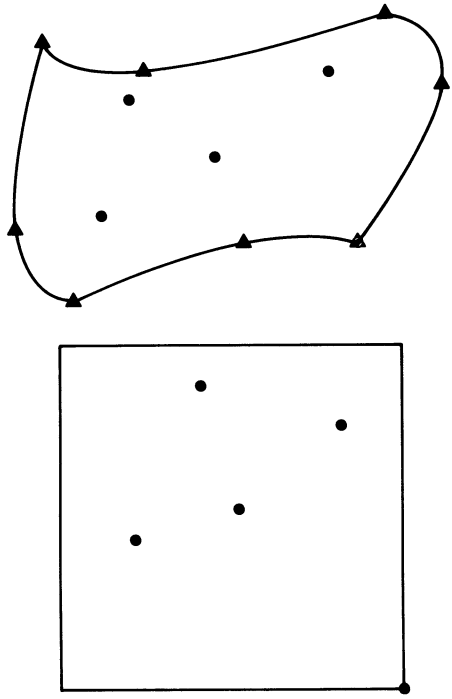
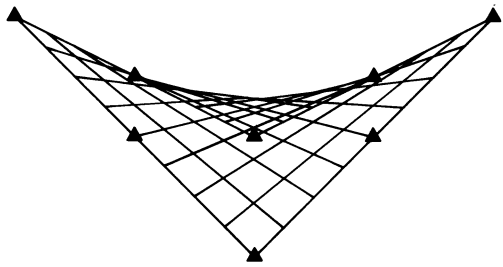


FIGURE 11. *Example* 1



8 NODE 2-D ELEMENT

(0.00 , 0.00)
(2.00 , 2.00)
(0.00 , 1.00)
(- 2.00 , 2.00)
(1.00 , 1.00)
(1.00 , 1.50)
(- 1.00 , 1.50)
(- 1.00 , 1.00)

FIGURE 12. *Example* 2

As a particular example of the techniques described in this section, we list the equations used in determining the preimage of (.5, .5). First of all, system (16) becomes

$$f(r, s) = (1.4s - .4)r^2 + (-1.6s^2 + .4s + 1.4)r + (.6s^2 - .7s - .5) = 0,$$

$$g(r, s) = (s - .4)r^2 + (-1.4s^2 + .3s + .6)r + (.8s^2 + .1s - .5) = 0.$$

Obviously, $a_0(s)b_0(s) \not\equiv 0$ so we determine the resultant (17)

$$.15840s^7 + .116s^6 - .9012s^5 + .0922s^4 + .0066s^3 - .1564s^2 + .3712s - .128 = 0.$$

This equation has the following real zeros

$$s_1 = -2.8030, \quad s_2 = 1.9835, \quad s_3 = -0.9062, \quad s_4 = 0.5233, \quad s_5 = 0.5011,$$

plus two complex zeros.

Now applying these five roots to Eqs. (18) and (19) we get corresponding to each s_i, the following r_i, $i = 1, 2, \ldots, 5$,

$$r_1 = 0.4357, \quad r_2 = 1.6015, \quad r_3 = -0.7015, \quad r_4 = 0.5220, \quad r_5 = -4.4916.$$

The only (r_i, s_i), $i = 1, 2, \ldots, 5$, in S is (0.5220, 0.5233) and so this is the only admissible preimage of the point (.5, .5) under the mapping $\bar{T}_2$.

It required 1.34 sec. of DEC-10 CPU time to determine these five preimages.

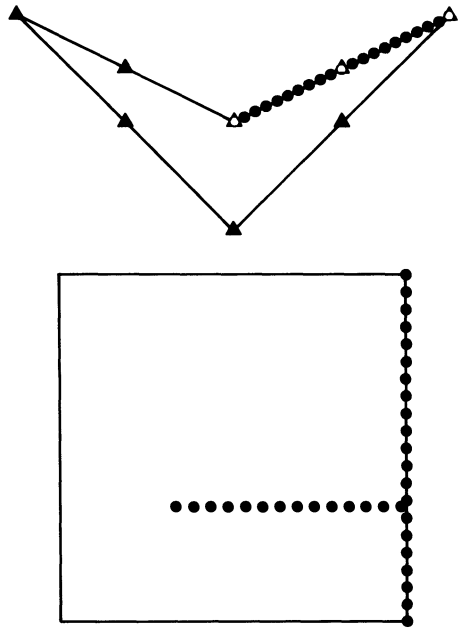
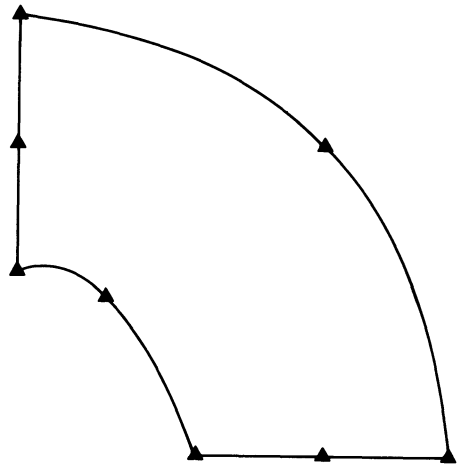


FIGURE 13. *Example* 2



8 NODE 2-D ELEMENT

(0.00 , 1.00)
(1.00 , 0.00)
(2.41 , 0.00)
(0.00 , 2.41)
(0.50 , 0.87)
(1.71 , 0.00)
(1.71 , 1.71)
(0.00 , 1.71)

FIGURE 14. *Example* 3

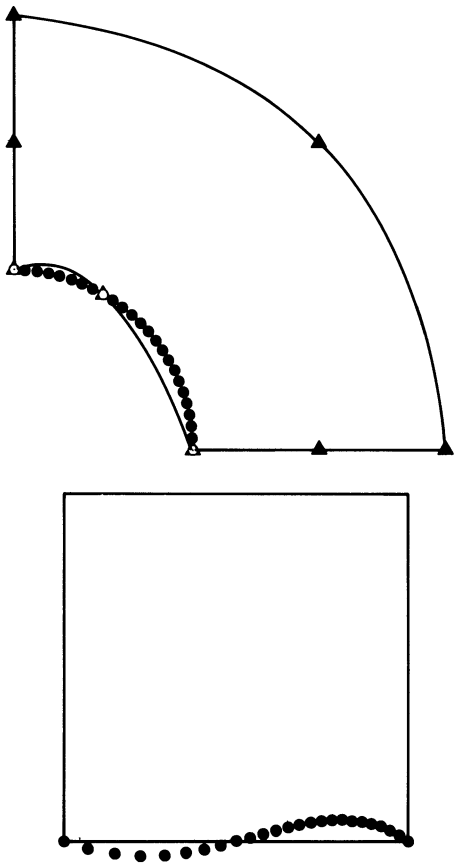


FIGURE 15. *Example* 3

Example 2. In Theorem 2 it was proven that the bilinear transformation $\bar{T}_1(r, s)$ is a bijection if and only if E is convex. Figure 12 illustrates the typical situation when E is *not* convex. Note that there is overspill in the neighborhood of the node where the interior angle exceeds π. Furthermore, Figure 13 shows that, as demonstrated in the proof of Theorem 2, the preimage of the boundary segment $\bar{P}_2\bar{P}_3$ is a boundary segment of S plus a portion of its interior.

Example 3. This example illustrates the necessity of carefully selecting the points $\bar{P}_i$, $i = 1, 2, \ldots, 8$, so that the region E determined by those points closely approximates the original region under consideration. The region we wish to approximate in this example is a quarter annulus with inner radius 1 and outer radius $1 + \sqrt{2}$. If the eight nodes are chosen with the following polar coordinates

$$\bar{P}_1 = (1, \pi/2), \qquad \bar{P}_5 = (1, \pi/3),$$
$$\bar{P}_2 = (1, 0), \qquad \bar{P}_6 = (1 + \sqrt{2}/2, 0),$$
$$\bar{P}_3 = (1 + \sqrt{2}, \pi/2), \qquad \bar{P}_7 = (1 + \sqrt{2}, \pi/4),$$
$$\bar{P}_4 = (1 + \sqrt{2}, 0), \qquad \bar{P}_8 = (1 + \sqrt{2}/2, \pi/2),$$

then the resulting region E is shown in Figure 14. Note that our choice of $\bar{P}_5$ produces a poor parabolic approximation to the circular arc of radius 1, but the choice of $\bar{P}_7$ gives us an excellent approximation to the circular arc of radius $1 + \sqrt{2}$.

We now consider 25 equally spaced points along the quarter circle of radius one and compute their preimages under $\bar{T}_2$. Figure 15 shows the location of these 25 points and their preimages. Note that since seven of the points on the circular arc lie outside the region E determined by the parabolic arc, the preimages of these seven points lie outside the unit square S.

Department of Mathematics and Statistics
University of Pittsburgh
Pittsburgh, Pennsylvania 15260

1. R. C. BUCK, *Advanced Calculus*, McGraw-Hill, New York, 1956.

2. P. G. CIARLET & P. A. RAVIART, "Interpolation theory over curved elements with applications to finite element methods," *Comput. Methods Appl. Mech. Engrg.*, v. 1, 1972, pp. 217–249.

3. I. ERGATOUDIS, B. IRONS & O. ZIENKIEWICZ, "Curved isoparametric, "quadrilateral" elements for finite element analysis," *Internat. J. Solids and Structures*, v. 4, 1968, pp. 31–42.

4. W. FULKS, *Advanced Calculus*, Wiley, New York, 1961.

5. W. J. GORDON & C. A. HALL, "Transfinite element methods: Blending function interpolation over arbitrary curved element domains," *Numer. Math.*, v. 21, 1973, pp. 109–129.

6. A. S. HOUSEHOLDER, "Bigradients and the problem of Routh and Hurwitz," *SIAM Rev.*, v. 10, 1968, pp. 56–66.

7. J. M. ORTEGA & W. C. RHEINBOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

8. W. G. STRANG & G. J. FIX, *An Analysis of the Finite Element Method*, Prentice-Hall, Englewood Cliffs, N. J., 1973.

9. C. J. de la VALLÉE POUSSIN, *Cours d'Analyse Infinitesimale*, vol. 1, Gauthier-Villars, Paris, 1926.

10. O. C. ZIENKIEWICZ, *The Finite Element Method in Engineering Science*, McGraw-Hill, London, 1971.