Domains of unicity

Vilmos Totik*

In honor of Lawrence Zalcman

Abstract

The Gale-Nikaido theorem claims that if the Jacobian of a mapping $\mathbf{F}$ is a P-matrix at every point of K and K is a closed rectangular region in $\mathbf{R}^n$, then $\mathbf{F}$ is globally univalent on K. Under the more severe condition that the (symmetric part of the) Jacobian is positive definite on K, the same conclusion is valid on any closed convex set K. In this paper it is shown that the closed rectangular regions are the only ones for which the Gale-Nikaido theorem is true. In a similar fashion, it is shown that the positive definiteness of the Jacobian implies unicity only on (closed) convex sets.

1 Introduction

It is well known if $\mathbf{F} = (F_i(x_1, \dots, x_n))_{i=1}^n$ is a differentiable mapping from a subset K of $\mathbf{R}^n$ into $\mathbf{R}^n$ and if the Jacobian $(\partial F_i/\partial x_j)$ of $\mathbf{F}$ does not vanish at a point, then $\mathbf{F}$ is univalent (1-to-1) in a neighborhood of that point. Global univalence is more subtle, and the mere vanishing of the Jacobian at every point of K is not sufficient. Gale and Nikaido [4] proved in 1965 that if the Jacobian of $\mathbf{F}$ is a P-matrix at every point of K (meaning that all of its principal minors are positive) and K is a closed rectangular region[1] $\prod[a_i, b_i]$, ($a_i < b_i$ for all i), then $\mathbf{F}$ is injective. Under the more severe condition that the (symmetric part of the) Jacobian is positive definite on K, the same conclusion is valid on any convex set K, see [2], [4] and [5]. The problem if the Gale-Nikaido theorem is true on any convex set has been mentioned several times in the book [6], but counterexamples were given later in [1] and [7].

In this note we address the question on what domains are the aforementioned two unicity theorems true. We are going to show that the closed rectangular regions are the only ones for which the Gale-Nikaido theorem is true, which

*Supported by NSF grant DMS 1564541

[1] This terminology follows the original paper [4]. A more correct notion would be "closed rectangular parallelepiped".

makes that result quite a peculiar one. In a similar fashion, we shall show that positive definiteness of the Jacobian implies unicity only on (closed) convex sets.

Let $K \subset \mathbf{R}^n$, $n \geq 2$, be a compact set. In what follows we shall consider continuously differentiable maps $\mathbf{F}$ from K to $\mathbf{R}^n$, and in order not to worry about the notion of the partial derivatives at arbitrary points of K, we shall assume without mentioning that $\mathbf{F}$ is defined on a neighborhood of K.

Recall that a not necessarily symmetric (real) square matrix is called a P-matrix if all of its principal submatrices (obtained by deleting some rows and the corresponding columns) have positive determinant. See [3, Section 5.5] or [4] for properties of P-matrices. Recall also that a symmetric square matrix A is positive definite if $\mathbf{x}^* A\mathbf{x} > 0$ for all non-zero vectors $\mathbf{x}$, where $\cdot^*$ denotes transposition. By Sylvester's criterion this happens if and only if all leading principal submatrices of A have positive determinants (the $m \times m$ principal submatrix of A is the one that lies in the first m rows and first m columns). In general, a not necessarily symmetric square matrix A is called positive definite if $\mathbf{x}^* A\mathbf{x} > 0$ for all non-zero vectors $\mathbf{x}$. This is the case precisely if its symmetric part $\frac{1}{2}(A + A^*)$ is positive definite.

With these notations the two global unicity theorems above can be stated as follows, where continuous partial derivatives of $\mathbf{F}$ are assumed.

Theorem A *If $K \subset \mathbf{R}^n$ is a closed rectangular region $\prod[a_i, b_i]$, ($a_i \leq b_i$ for all i) and the Jacobian of a C^1 mapping $\mathbf{F} : K \to \mathbf{R}^n$ is a P-matrix at every point of K, then $\mathbf{F}$ is univalent on K.*

Note that K may be a degenerated rectangular region.

Theorem B *If $K \subset \mathbf{R}^n$ is a closed convex set and the Jacobian of a C^1 mapping $\mathbf{F} : K \to \mathbf{R}^n$ is positive definite at every point of K, then $\mathbf{F}$ is univalent on K.*

Here again, K may have empty interior.

In this paper we show that the following converses hold.

Theorem 1 *Let $K \subset \mathbf{R}^n$ be a non-empty compact set with the property that any C^1 mapping $\mathbf{F} : K \to \mathbf{R}^n$ for which the Jacobian is a P-matrix at every point of K, is univalent on K. Then K is a closed rectangular region.*

Theorem 2 *Let $K \subset \mathbf{R}^n$ be a non-empty compact set with the property that any C^1 mapping $\mathbf{F} : K \to \mathbf{R}^n$ for which the Jacobian is positive definite at every point of K, is univalent on K. Then K is convex.*

It is clear that Theorems A and B imply their variant for open rectangular regions and for open convex sets, respectively. However, this is not the case for Theorem 2; at the end of the paper we shall show a non-convex open set with the property that every $\mathbf{F}$ is univalent on K for which the Jacobian is positive definite at every point of K.

2 Proof of Theorem 1

First of all note that K must be connected. Indeed, in the opposite case $K = K_1 \cup K_2$, where K_1, K_2 are disjoint non-empty closed sets. If $P_1 \in K_1$ and $P_2 \in K_2$, then the mapping which is $\mathrm{Id}-P_1$ on K_1 and is $\mathrm{Id}-P_2$ on K_2 (where Id is the identity mapping) shows that K does not have the property set forth in the theorem.

For $P = (\beta_i), Q = (\gamma_i) \in \mathbf{R}^n$ we set $a_i = \min(\beta_i, \gamma_i)$, $b_i = \max(\beta_i, \gamma_i)$ and define the closed rectangular region

$$T(P,Q) := \{(\alpha_i) \,|\, a_i \le \alpha_i \le b_i,\ i = 1, \dots, n\}.$$

The segment $\overline{PQ}$ is one of the diagonals of $T(P,Q)$, and the dimension of $T(P,Q)$ equals the number of those i for which $\beta_i \ne \gamma_i$.

Let K be as in the theorem, and let m be the maximal dimension of the rectangular regions $T(P,Q)$ for $P, Q \in K$. If $m = 0$ then K is a singleton, so assume $m \ge 1$. Then there are points $P, Q \in K$ such that m of the corresponding coordinates of P and Q are different, but any two points in K have at most m different coordinates. We fix these P, Q and write $P = (\beta_i), Q = (\gamma_i)$. Since simultaneous permutation of the rows and the corresponding columns of a P-matrix results in a P-matrix again, we may assume without loss of generality that $\beta_1 \ne \gamma_1, \dots, \beta_m \ne \gamma_m$, but $\beta_{m+1} = \gamma_{m+1}, \dots, \beta_n = \gamma_n$. Set, as before, $a_i = \min(\gamma_i, \beta_i)$, $b_i = \max(\beta_i, \gamma_i)$. Then $a_i < b_i$ for $i \le m$ and $a_i = b_i = \beta_i$ for $i > m$.

If $m = 1$ then it is immediate that K lies on the line $\{(x_i) \,|\, x_i = \beta_i \text{ for } i > 1\}$. Since K is also connected, it is a segment on that line, and the theorem is true in this case. Therefore, in what follows we may assume that $m \ge 2$.

Claim 1. $T(P,Q) \subseteq K$.

Suppose this is not the case. Then there is a point $R = (\delta_i)$ in $T(P,Q) \setminus K$, and clearly $\delta_i = a_i = b_i = \beta_i$ for $i > m$, while $\delta_i \in [a_i, b_i]$ for $i \le m$, and these last intervals are non-degenerate. Since K is closed, a neighborhood of R is disjoint from K, and by changing all $\delta_i \in [a_i, b_i]$, $1 \le i \le m$, a little, we may assume that $\delta_i \in (a_i, b_i)$ for $1 \le i \le m$. Select a $\tau > 0$ such that the (closed) ball about R of radius $n\tau$ is disjoint from K, and at the same time $[\delta_i - \tau, \delta_i + \tau] \subset (a_i, b_i)$ for all $1 \le i \le m$. Note that there is no $S = (\alpha_i) \in K$ such that $\alpha_i \in [\delta_i - \tau, \delta_i + \tau]$ for all $1 \le i \le m$. Indeed, this is clear if $\alpha_{m+1} = \beta_{m+1}, \dots, \alpha_n = \beta_n$, for then S is closer to R than $n\tau$. On the other hand, if there is a $j > m$ for which $\alpha_j \ne \beta_j$, then $m+1$ coordinates of P and S are different (the first m and the j-th one), so, by the definition of m and by $P \in K$, we cannot have $S \in K$.

For each $1 \le i \le m$ select a continuously differentiable function g_i with the property that

(1) $g_i'(t) = 0$ if $t \notin [\delta_i - \tau, \delta_i + \tau]$, $1 \le i \le m$,

(2) $g_i(\gamma_i) = -\gamma_{i-1}$, $g_i(\beta_i) = -\beta_{i-1}$, $2 \le i \le m$,

(3) $g_1(\gamma_1) = -\gamma_m$, $g_1(\beta_1) = -\beta_m$.

Since β_i, γ_i lie in different components of $\mathbf{R} \setminus [\delta_i - \tau, \delta_i + \tau]$, that is possible.

With these g_i define

$$\mathbf{F}(x_1, \ldots, x_n) = (x_1 + g_2(x_2), x_2 + g_3(x_3), \cdots, x_{m-1} + g_m(x_m), x_m + g_1(x_1), x_{m+1}, x_{m+2}, \cdots, x_n).$$

(When $m = n$, the coordinates $x_{m+1}, x_{m+2}, \cdots, x_n$ are not needed.) For this mapping we have $\mathbf{F}(P) = \mathbf{F}(Q) = (0, \cdots, 0, \beta_{m+1}, \ldots, \beta_n)$, so $\mathbf{F}$ is not univalent. On the other hand, we shall show below that the Jacobian of $\mathbf{F}$ is a P-matrix at every point of K. However, this contradicts the assumed property of K, and this contradiction proves the claim.

The Jacobian of F is

$$\begin{pmatrix} 1 & g_2'(x_2) & & & & & & & \\ & 1 & g_3'(x_3) & & & & & & \\ & & 1 & & & & & & \\ & & & \ddots & \ddots & & & & \\ & & & & 1 & g_m'(x_m) & & & \\ g_1'(x_1) & & & & & 1 & & & \\ & & & & & & 1 & & \\ & & & & & & & \ddots & \\ & & & & & & & & 1 \end{pmatrix},$$

where we showed only the (possibly) non-zero entries of the Jacobian. Note that at each point of K at least one of the off-diagonal entries (i.e. at least one of $g_1'(x_1), \ldots, g_m'(x_m)$) is zero. Indeed, if $(x_i) \in K$, then, according to what we have said before, there is an $1 \le i \le m$ such that $x_i \notin [\delta_i - \tau, \delta_i + \tau]$, and then $g_i'(x_i) = 0$ by the choice of the function g_i.

Thus, it is sufficient to show that any matrix of the form

$$\mathcal{M} = \begin{pmatrix} 1 & u_2 & & & & & & & \\ & 1 & u_3 & & & & & & \\ & & 1 & & & & & & \\ & & & \ddots & \ddots & & & & \\ & & & & 1 & u_m & & & \\ u_1 & & & & & 1 & & & \\ & & & & & & 1 & & \\ & & & & & & & \ddots & \\ & & & & & & & & 1 \end{pmatrix}$$

with the side-condition that at least one of the u_i's is zero, is a P-matrix. Indeed, this follows from the two facts:

(i) any principal submatrix of $\mathcal{M}$ is of the same form (with m replaced by $m-k$ if k of the first m rows and columns are deleted from $\mathcal{M}$),

(ii) the determinant of $\mathcal{M}$ is 1.

It is sufficient to prove (i) for the case when one row and the corresponding column is deleted from $\mathcal{M}$, for we can iterate this special case. If the j-th row and column are deleted and $j > m$, then the claim is clear. If $j = 1$, then we get an upper triangular matrix, while if $2 \leq j \leq m$, then the j-th column of the obtained matrix (which is otherwise of the form as $\mathcal{M}$ but with m replaced by $m-1$) contains only zeros except for the single 1 in the diagonal, so the side-condition that at least one of the u_i's is zero is preserved.

Finally, (ii) is immediate, for the determinant of $\mathcal{M}$ is $1+(-1)^{m+1}\prod_{i=1}^{m} u_i = 1$, as can be seen by expanding the determinant according to the first column.

With this the proof of Claim 1 is complete.

Claim 2. K lies in the affine subspace $L := \{(x_i) \,|\, x_i = \beta_i$ for $i > m\}$.

Recall that (β_i) is the point P that was chosen after the definition of the number m.

The claim is immediate, for if there was a point $S = (\alpha_i) \in K$ outside L, then we could select in $T(P,Q)$ a point $R = (\theta_i)$ with $\theta_i \neq \alpha_i$ for $i \leq m$ (recall that $T(P,Q) = \prod_{i=1}^{n}[a_i, b_i]$ with $a_i < b_i$ for $i \leq m$). But then R and S would be two points in K the coordinates of which differ for at least $m+1$ indices (for the first m ones and for the j-th index for which $m < j \leq n$ and $\alpha_j \neq \beta_j$), which is not possible by the choice of m.

Seeing that all points of K have as their i-th coordinate β_i for all $i > m$, for simpler notations in what follows we shall suppress those coordinates, which amounts the same as setting $m = n$.

Claim 3. If all the m coordinates of the points $P', Q' \in K$ are different, then $T(P', Q') \subset K$.

Indeed, just follow the proof given for Claim 1 by replacing P and Q by P' and Q'.

Claim 4. If T is the smallest closed rectangular region that contains K (which is the intersection of all such closed regions), then $K = T$.

Let $T = \prod_{i=1}^{m}[A_i, B_i]$. Then $A_i < B_i$ for all $1 \leq i \leq m$ (recall that $T(P,Q) \subseteq K \subseteq T$). Now $K \subseteq T$, and if we show that $(A_i) \in K$ and $(B_i) \in K$, then $T = T((A_i),(B_i)) \subseteq K$ by Claim 3, hence $K = T$ will follow.

We shall prove that $(B_i) \in K$, the proof of $(A_i) \in K$ is similar. We shall show by induction on $k \leq m$ that K has a point M_k of the form $M_k = (B_1, \ldots, B_k, \alpha_{k+1}, \ldots, \alpha_m)$, and then $(B_i) \in K$ follows by setting $k = m$.

Let $L_j = \{(x_i) \,|\, x_j = B_j\}$ be the ($(m-1)$-dimensional) hyperplane of those points that have j-th coordinate equal to B_j. By the definition of T we have $L_j \cap K \neq \emptyset$ for all $1 \leq j \leq m$, and for $j = 1$ this proves the existence of M_1.

Suppose now that $M_k = (B_1, \dots, B_k, \alpha_{k+1}, \dots, \alpha_m) \in K$ exists for some $k < m$. If $\alpha_{k+1} = B_{k+1}$, then we can set $M_{k+1} = M_k$. Hence, we may assume that $\alpha_{k+1} < B_{k+1}$. Let $R \in K \cap L_{k+1}$, and choose a point $S \in T(P,Q)$ (where $T(P,Q)$ is the closed rectangular region considered in Claim 1) such that S and R have different coordinates (this is possible, since $T(P,Q)$ is the product of non-degenerate intervals). Then, by Claim 3, we have $T(R,S) \subset K$, and $T(R,S) \cap L_{k+1}$ is a non-empty $(m-1)$-dimensional closed rectangular region lying in L_{k+1} (note that $R \in T(R,S) \cap L_{k+1}$). So there is a point $N_{k+1} \in T(R,S) \cap L_{k+1} \subset K$ such that M_k and N_{k+1} have different coordinates. This is so because only the $(k+1)$-st coordinate of a generic point N_{k+1} from $T(R,S) \cap L_{k+1}$ is fixed to be B_{k+1} – the other coordinates can vary in some non-degenerate intervals –, and we have assumed that the $(k+1)$-st coordinate α_{k+1} of M_k is smaller than B_{k+1}. Now Claim 3 asserts that $T(M_k, N_{k+1}) \subseteq K$, and the right upper corner of $T(M_k, N_{k+1})$ is suitable as M_{k+1}, for its i-th coordinate is the maximum of the i-th coordinates of M_k and N_{k+1}, and that is B_i for all $i \le k+1$.

With this the proof of Claim 4 is complete, and Theorem 1 follows.

■

3 Proof of Theorem 2

For every large M we construct an auxiliary mapping $\mathbf{F}_M : \mathbf{R}^n \to \mathbf{R}^n$ for which the Jacobian is positive definite on a large part of $\mathbf{R}^n$. The mapping $\mathbf{F}_M$ will be the gradient of the function

$$\Phi_M(x_1, \dots, x_n) = \Big((x_1 - 1)^2 + M \sum_{i=2}^n x_i^2\Big)\Big((x_1 + 1)^2 + M \sum_{i=2}^n x_i^2\Big),$$

i.e.

$$\mathbf{F}_M(x_1, \dots, x_n) = \Big(4x_1^3 - 4x_1 + 4x_1 M \sum_{i=2}^n x_i^2, \dots, 4M(x_1^2+1)x_j + 4M^2 x_j \sum_{i=2}^n x_i^2, \dots\Big),$$

where the generic term is for $j = 2, \dots, n$. Then the Jacobian of $\mathbf{F}_M$ is the Hessian

$$H_M = \left(\frac{\partial^2 \Phi_M}{\partial x_i \partial x_j}\right)_{i,j=1}^n.$$

First we prove

Proposition 3 *The Jacobian H_M is positive definite outside the set*

$$E_M = \left[-1 + \frac{1}{128}, 1 - \frac{1}{128}\right] \times \left\{(x_2, \dots, x_n) \,\middle|\, \sum_{i=2}^n x_i^2 \le \frac{1}{M}\right\}. \tag{1}$$

Proof. By Sylvester's theorem we need to show that the principal submatrices of H_M have positive determinant. The $m \times m$ principal submatrix $H(m)$ of H_M is

$$\begin{pmatrix} h_{1,1} & 8Mx_1x_2 & \cdots & 8Mx_1x_j & \cdots & 8Mx_1x_m \\ 8Mx_1x_2 & h_{2,2} & \cdots & 8M^2x_2x_j & \cdots & 8M^2x_2x_m \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 8Mx_1x_{j-1} & 8M^2x_2x_{j-1} & \cdots & 8M^2x_{j-1}x_j & \cdots & 8M^2x_{j-1}x_m \\ 8Mx_1x_j & 8M^2x_2x_j & \cdots & h_{j,j} & \cdots & 8M^2x_jx_m \\ 8Mx_1x_{j+1} & 8M^2x_2x_{j+1} & \cdots & 8M^2x_jx_{j+1} & \cdots & 8M^2x_{j+1}x_m \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 8Mx_1x_m & 8M^2x_2x_m & \cdots & 8M^2x_{j-1}x_{m+1} & \cdots & h_{m,m} \end{pmatrix},$$

where the diagonal elements are:

$$h_{1,1} = 12x_1^2 - 4 + 4M\sum_{i=2}^{n} x_i^2,$$

and for $j \geq 2$

$$h_{j,j} = 4M(x_1^2 + 1) + 8M^2x_j^2 + 4M^2\sum_{i=2}^{n} x_i^2.$$

Even though the positivity of $\det(H(m))$ can be shown using standard row and column operators, some care has to be exercised since $\det(H(m))$ is not (cannot) be positive on the whole $\mathbf{R}^n$, so we give some details.

First assume that none of the numbers x_i, $1 \leq i \leq n$ is zero.

Divide the j-th row and j-th column of $H(m)$ by x_j for all $1 \leq j \leq m$. We obtain a matrix $A = (a_{i,j})$ for which the determinant is of the same sign as the determinant of $H(m)$, so it is sufficient to consider A, which is of the form

$$\begin{pmatrix} a_{1,1} & 8M & \cdots & 8M & 8M & 8M & \cdots & 8M \\ 8M & a_{2,2} & \cdots & 8M^2 & 8M^2 & 8M^2 & \cdots & 8M^2 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 8M & 8M^2 & \cdots & a_{j-1,j-1} & 8M^2 & 8M^2 & \cdots & 8M^2 \\ 8M & 8M^2 & \cdots & 8M^2 & a_{j,j} & 8M^2 & \cdots & 8M^2 \\ 8M & 8M^2 & \cdots & 8M^2 & 8M^2 & a_{j+1,j+1} & \cdots & 8M^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 8M & 8M^2 & \cdots & 8M^2 & 8M^2 & 8M^2 & \cdots & a_{m,m} \end{pmatrix},$$

where now

$$a_{1,1} = 12 - \frac{4}{x_1^2} + \frac{4M}{x_1^2}\sum_{i=2}^{n} x_i^2,$$

and

$$a_{j,j} = 8M^2 + 4M\frac{x_1^2+1}{x_j^2} + 4M^2\frac{\sum_{i=2}^n x_i^2}{x_j^2}, \qquad j \geq 2.$$

Subtract the last row from rows $2, 3, \ldots, (m-1)$ to obtain the matrix $B = (b_{i,j})$ of the form

$$\begin{pmatrix} b_{1,1} & 8M & \cdots & 8M & 8M & 8M & \cdots & 8M \\ 0 & b_{2,2} & \cdots & 0 & 0 & 0 & \cdots & b_{2,m} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & b_{j-1,j-1} & 0 & 0 & \cdots & b_{j-1,m} \\ 0 & 0 & \cdots & 0 & b_{j,j} & 0 & \cdots & b_{j,m} \\ 0 & 0 & \cdots & 0 & 0 & b_{j+1,j+1} & \cdots & b_{j+1,m} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 8M & 8M^2 & \cdots & 8M^2 & 8M^2 & 8M^2 & \cdots & b_{m,m} \end{pmatrix}.$$

The off-diagonal entries in B are zero except for those in the first and last rows and in the last column. In the first row all off-diagonal elements are $8M$, in the last row they are $8M, 8M^2, 8M^2, \cdots, 8M^2$, respectively, and the $b_{j,m}$, $2 \leq j \leq m-1$, element in the last column is

$$b_{j,m} = -4M\frac{x_1^2+1}{x_m^2} - 4M^2\frac{\sum_{i=2}^n x_i^2}{x_m^2} \leq -4M^2. \tag{2}$$

Finally, the diagonal entries are

$$b_{1,1} = a_{1,1} = 12 - \frac{4}{x_1^2} + \frac{4M}{x_1^2}\sum_{i=2}^n x_i^2, \tag{3}$$

$$b_{j,j} = 4M\frac{x_1^2+1}{x_j^2} + 4M^2\frac{\sum_{i=2}^n x_i^2}{x_j^2} \geq 4M^2, \qquad 2 \leq j \leq m-1, \tag{4}$$

and

$$b_{m,m} = a_{m,m} = 8M^2 + 4M\frac{x_1^2+1}{x_m^2} + 4M^2\frac{\sum_{i=2}^n x_i^2}{x_m^2}. \tag{5}$$

If $\sum_{i=2}^n x_i^2 \geq 1/M$, then $b_{1,1} \geq 12$. Now subtract $(8M/b_{1,1})$-times the first column of B from the j-th column for all $2 \leq j \leq m$ to get the matrix $C = (c_{i,j})$. For it $c_{1,1} = b_{1,1} \geq 12$, and this is the only non-zero element in the first row. In the last row of C the j-th element is

$$c_{m,j} = 8M^2 - 8M(8M/b_{1,1}) \geq 8M^2 - 8M(8M/12) = 8M^2/3 > 0$$

for $2 \leq j \leq m-1$, while

$$\begin{aligned} c_{m,m} &= 8M^2 + 4M\frac{x_1^2+1}{x_m^2} + 4M^2\frac{\sum_{i=2}^n x_i^2}{x_m^2} - 8M(8M/b_{1,1}) \\ &\geq 8M^2 - 8M(8M/12) = 8M^2/3 > 0, \end{aligned}$$

and of course, in the last column we have $c_{j,m} = b_{j,m} \le -4M^2$ for $2 \le j \le m-1$. Thus, if $C_{1,1}$ is the matrix that we obtain from C by deleting the first row and first column, then $C_{1,1}$ is an $(m-1) \times (m-1)$ matrix of the form

$$Q = \begin{pmatrix} + & & & & - \\ & + & & & - \\ & & \ddots & & \vdots \\ & & & + & - \\ + & + & \cdots & + & + \end{pmatrix}, \tag{6}$$

where $+$ indicates a positive element and $-$ indicates a negative element, and all other elements are zero. Every such Q has positive determinant – just eliminate the off-diagonal elements in the last row by subtracting appropriate multiples of the first $(m-1)$ rows from the last row to get an upper diagonal matrix with positive diagonal elements.

This proves that $\det(C) > 0$, and hence also $\det(H(m)) > 0$ if $\sum_{i=2}^n x_i^2 \ge 1/M$.

Next, assume that $x_1^2 \ge 1 - \varepsilon$ with $\varepsilon = 1/64$. We distinguish two cases.

Case 1. $\sum_{i=2}^n x_i^2 \ge 4\varepsilon/M$. In this case

$$b_{1,1} = c_{1,1} \ge 12 + \frac{-4 + 16\varepsilon}{1-\varepsilon} \ge 8 + 4\varepsilon$$

(see (3)). Hence in the matrix C in the last row we have

$$c_{m,j} = 8M^2 - 8M(8M/b_{1,1}) \ge 8M^2 - 8M(8M/(8+4\varepsilon)) > 0$$

for $2 \le j \le m-1$, so $C_{1,1}$ is again of the form (6), and we get the positivity of $\det(C) = \det(B) = \det(A)$ as before.

Case 2. $\sum_{i=2}^n x_i^2 < 4\varepsilon/M$ (still assuming $x_1^2 \ge 1-\varepsilon$), which implies $x_i^2 < 4\varepsilon/M$ for all $i \ge 2$. In this case (3) yields $b_{1,1} \ge 4$, while (4) and (5) give

$$b_{j,j} \ge \frac{M^2}{\varepsilon}, \qquad 2 \le j \le m.$$

Now subtract $(b_{m,j}/b_{j,j})$-times the j-th row of B from its last row for all $1 \le j \le m-1$ to get the matrix $D = (d_{i,j})$. D is upper diagonal with diagonal entries $d_{j,j} = b_{j,j} > 0$ for $1 \le j \le m-1$ and (see also (2))

$$d_{m,m} = b_{m,m} - \sum_{j=1}^{m-1} \frac{b_{m,j}}{b_{j,j}} b_{j,m} \ge b_{m,m} - \frac{b_{m,1}}{b_{1,1}} b_{1,m} \ge b_{m,m} - \frac{8M}{4} 8M > \frac{M^2}{\varepsilon} - 16M^2 > 0.$$

Thus, D, and hence also the matrices B and A have positive determinants also in this case.

In summary, if none of the x_j is zero and either $\sum_{j=2}^n x_j^2 \geq 1/M$ or if $x_1^2 \geq 1 - 1/64$, then $\det(H(m)) > 0$, which proves the positivity of $\det(H(m))$ outside the set E_M.

Finally, consider the case when $(x_1, \ldots, x_n) \notin E_M$ but $\prod_{j=1}^n x_j = 0$. If an x_j, $2 \leq j \leq n$, is zero, then in the matrix H_M the j-th row and j-th column is zero except for the positive diagonal element $h_{j,j} \geq 4M$ in them. For $2 \leq j \leq m$ in this case by expanding the determinant $H(m)$ according to the j-th row (during which the contribution of the non-zero element $h_{j,j}$ is positive), we can just omit that variable during the analysis of the determinant of $H(m)$. Thus, we may assume that $x_2 \cdot \cdots \cdot x_n \neq 0$. But then necessarily $x_1 = 0$ and $\sum_{j=2}^n x_j^2 > 1/M$. In this case the first row and first column of H_M is zero except for the entry

$$h_{1,1} = 12x_1^2 - 4 + 4M \sum_{i=2}^n x_i^2 = -4 + 4M \sum_{i=2}^n x_i^2 > 0,$$

and then the preceding proof works with the modification that in creating the matrix A we do not divide by x_1 (but do divide with all other x_j).

■

After these preparations the proof of Theorem 2 is immediate.

Proof of Theorem 2. Suppose K is not convex. Then there are points $P', Q' \in K$ such that the segment connecting P' and Q' has a point R that lies outside K. Since K is compact, if $P, Q \in K$ are the two closest points to R on that segment such that R lies on the segment PQ, then this latter segment PQ lies outside K except for its endpoints. We can apply a translation, dilation and rotation (orthogonal transformation) to get a $T : \mathbf{R}^n \to \mathbf{R}^n$ which maps P into the point $(-1, 0, \ldots, 0)$ and Q into $(1, 0, \ldots, 0)$. Since these operations do not change the positive definiteness of a Jacobian, we may consider instead of K the set $T(K)$, and instead of the mapping $\mathbf{F}$ the mapping $T \circ \mathbf{F} \circ T^{-1}$ of $T(K)$ into $\mathbf{R}^n$.

Thus, we may assume that $(-1, 0, \ldots, 0)$ and $(1, 0, \ldots, 0)$ are in K, but no other point on the segment connecting these points lies in K. But then, using again the compactness of K, there is an $M > 0$ such that the set E_M from (1) lies outside K. So $\mathbf{F}_M$ is a C^1 mapping that has positive Jacobian at every point of K. But $\mathbf{F}_M(-1, 0, \ldots, 0) = (0, \ldots, 0) = \mathbf{F}_M(1, 0, \ldots, 0)$, hence $\mathbf{F}_M$ is not univalent in K. Since this contradicts the assumption in Theorem 2, the proof is complete.

■

We have already mentioned that Theorem B implies its variant for open convex sets. But in that form the converse is not true, for there are non-convex open sets on which every mapping with positive definite Jacobian is univalent.

Example 1. Let $K = (-1,1)^n \setminus \{(0,\dots,0)\}$ (where $n \geq 2$). We claim that even though K is not convex, every C^1 mapping $\mathbf{F}$ on K with positive definite Jacobian is univalent. To prove that, let $\mathbf{x}, \mathbf{y}$ be two distinct points in K, and consider the function

$$g(t) = (\mathbf{y} - \mathbf{x})^* \mathbf{F}(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))$$

(where we consider the vectors as column vectors and the product is dot product). If the segment connecting $\mathbf{x}$ and $\mathbf{y}$ does not pass through the origin, then g is defined for all $t \in [0,1]$. Its derivative is

$$g'(t) = (\mathbf{y} - \mathbf{x})^* J(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))(\mathbf{y} - \mathbf{x}),$$

where J denotes the Jacobian of $\mathbf{F}$. So, by the assumed positive definiteness of J, this is positive for all $t \in [0,1]$, hence $g(1) > g(0)$. In particular, $\mathbf{F}(\mathbf{x})$ and $\mathbf{F}(\mathbf{y})$ must be different.

If the origin lies on the segment connecting $\mathbf{x}$ and $\mathbf{y}$, then g is not defined for some $t \in (0,1)$, but it is defined for all $t \in [0,\alpha]$ with some $\alpha > 0$. Let $\mathbf{z} = \mathbf{x} + \alpha(\mathbf{y} - \mathbf{x})$. As we have just seen,

$$(\mathbf{z} - \mathbf{x})^* \mathbf{F}(\mathbf{z}) - (\mathbf{z} - \mathbf{x})^* \mathbf{F}(\mathbf{x}) := b > 0 \tag{7}$$

with some $b > 0$. Apply a small translation so that the origin does not lie on the translation of the segment connecting $\mathbf{x}$ and $\mathbf{y}$, and let $\mathbf{x}', \mathbf{y}', \mathbf{z}'$ be the images of $\mathbf{x}, \mathbf{y}, \mathbf{z}$ under this translation. As above, we get

$$(\mathbf{y}' - \mathbf{z}')^* \mathbf{F}(\mathbf{y}') - (\mathbf{y}' - \mathbf{x}')^* \mathbf{F}(\mathbf{z}') > 0,$$

and since $\mathbf{y}' - \mathbf{z}'$ is a positive constant multiple of $\mathbf{z} - \mathbf{x}$, this is the same as

$$(\mathbf{z} - \mathbf{x})^* \mathbf{F}(\mathbf{y}') - (\mathbf{z} - \mathbf{x})^* \mathbf{F}(\mathbf{z}') > 0. \tag{8}$$

Finally, if the translation is small, then we have

$$(\mathbf{z} - \mathbf{x})^* \mathbf{F}(\mathbf{z}') - (\mathbf{z} - \mathbf{x})^* \mathbf{F}(\mathbf{z}) > -\frac{b}{2},$$

and

$$(\mathbf{z} - \mathbf{x})^* \mathbf{F}(\mathbf{y}) - (\mathbf{z} - \mathbf{x})^* \mathbf{F}(\mathbf{y}') > -\frac{b}{2}.$$

If we add together the last two inequalities and (7) and (8), then we obtain

$$(\mathbf{z} - \mathbf{x})^* \mathbf{F}(\mathbf{y}) - (\mathbf{z} - \mathbf{x})^* \mathbf{F}(\mathbf{x}) > 0,$$

which proves that $\mathbf{F}(\mathbf{x})$ and $\mathbf{F}(\mathbf{y})$ are different.

A similar proof works if $K = (-1,1)^n \setminus K_0$, where K_0 is any compact set which is disjoint from a dense set of segments (i.e. for every segment with

endpoints in $(-1,1)^n$ there is arbitrarily close to it another such segment which is disjoint from K_0). Note that for a Cantor-type set K_0 such a K is very far from being convex. But its closure is convex, and this is the only thing one can claim for an open connected K on which every mapping with positive Jacobian is univalent (the proof that in such a case the closure of K must be convex follows the proof of Theorem 2).

Acknowledgement. The author has learned from B. Nagy the problem raised in the book [6] if the Gale-Nikaido theorem is true on convex sets.

References

[1] V. A. Aleksandrov, On the fundamental Gale-Nikaido-Inada theorem on the injectivity of mappings. (Russian) *Sibirsk. Mat. Zh.*, **35**(1994),715–718, translation in *Siberian Math. J.*, **35**(1994), 637–639.

[2] A. M. Fomin, On a sufficient condition for the homeomorphism of a continuous differentiable mapping. (Russian) *Uspehi Matem. Nauk (N.S.)*, **4**(1949). 198-199.

[3] M. Fiedler, *Special matrices and their applications in numerical mathematics*. Second edition. Dover Publications, Inc., Mineola, NY, 2008.

[4] D. Gale and H. Nikaido, The Jacobian matrix and global univalence of mappings. *Math. Ann.*, **159**(1965), 81-93.

[5] A. D. Myskis and. A. Ya. Bunt, On a sufficient condition for homeomorphism of a continuously differentiable mapping. (Russian) *Uspehi Mat. Nauk (N.S.)*, **10**(1955), 139-142.

[6] T. Parthasarathy, *On global univalence theorems*, Lecture Notes in Mathematics, **977**. Springer-Verlag, Berlin-New York, 1983. viii+106 pp.

[7] T. Parthasarathy and G. Ravindran, Completely mixed games and global univalence in convex regions. *Optimization, design of experiments and graph theory (Bombay, 1986)*, 417-423, Indian Inst. Tech., Bombay, 1988.

MTA-SZTE Analysis and Stochastics Research Group
Bolyai Institute, University of Szeged
Szeged, Aradi v. tere 1, 6720, Hungary

and

Department of Mathematics and Statistics, University of South Florida
4202 E. Fowler Ave, CMC342, Tampa, FL 33620-5700, USA

totik@mail.usf.edu