Operators and Matrices
Volume 8, Number 2 (2014), 389–398
doi:10.7153/oam-08-19

ON ONE PROPERTY OF INVERSES OF NONLINEAR OPERATORS ASSOCIATED WITH M–MATRICES

YURIY V. SHLAPAK

(*Communicated by R. A. Brualdi*)

Abstract. In this paper we show that a certain class of nonlinear operators associated with M-matrices behaves similarly to M-matrices in the sense that their inverse operators map the cone of positive vectors of $\mathbb{R}^n$ to itself. It is also proven that a certain iteration process can be used to find the values of these inverse operators at any point within the cone of positive vectors. Some results of computational experiments based on this iteration process are presented and discussed.

1. Introduction

We will consider operators from $\mathbb{R}^n$ to $\mathbb{R}^n$ of the structure

$$M(x) = Ax + F(x)$$

where A is an M-matrix and $F(x)$ is a nonlinear operator from $\mathbb{R}^n$ to $\mathbb{R}^n$. The main result of this paper will be the statement that if $F(x)$ satisfies certain reasonable conditions then M^{-1} maps the cone of positive vectors to itself. It will also be shown that an iteration sequence can be constructed that will converge to the value of M^{-1} at any point of the cone of positive vectors. On an intuitive level it means that if we introduce some nonlinear perturbations to an operator associated with an M-matrix, we obtain a nonlinear operator that behaves similarly to an M-matrix when it comes to some properties of its inverse. Namely, any positive matrix maps the cone of positive vectors to itself and so does this inverse operator.

This problem was brought to the attention of the author of this paper by the late Israel Koltracht, and the techniques used in this paper are similar to techniques used in some of Israel Koltracht's papers [1], [2], [3]. The problem considered in [1]–[3] was to find a positive eigenvector of a nonlinear perturbation of an M-matrix, while the problem considered in this paper is basically about finding the inverse operator of a nonlinear perturbation of an M-matrix on the cone of positive vectors. Applying the techniques of [1]–[3] to a very different problem required some substantial changes in the proof and entirely different conditions for the existence of a solution (the conditions $\lambda > \mu$ and (2) from [1] are replaced by the conditions (4) and (5) in this paper).

Mathematics subject classification (2010): Primary: 15B48; Secondary: 15A24, 15A60.
Keywords and phrases: M-Matrix, perturbations, nonlinear operator, inverse operator, cone of positive vectors, monotone fixed point Theorem, iterative process, computational experiment.

Paper OaM-08-19

We use the term positive matrix to refer to a matrix whose all entries are positive numbers. We use the term M-matrix to refer to a positive stable matrix whose non-diagonal entries are all nonpositive [4]. In this paper we will rely heavily on the fact that if A is an M-matrix then $(A+cI)^{-1}$ is a positive matrix for any $c>0$. This fact is easy to prove because if A is an M-matrix then $A+cI$ (for $c>0$) is also an M-matrix, and the inverse of any M-matrix is a positive matrix [4].

We can reduce the problem of finding the inverse of the operator $M(x)$ on the cone of positive vectors to a certain nonlinear vector equation. Namely, we consider the following problem: find a vector $x\in\mathbb{R}^n$ that satisfies the equation

$$Ax+F(x)=B \tag{1}$$

where A is an M-matrix of the size $n\times n$, $F(x)$ is a vector function from $\mathbb{R}^n$ to $\mathbb{R}^n$ that depends on the vector x, and $B=[b_1,b_2,...,b_n]^T$ is a positive column vector in $\mathbb{R}^n$, i.e. $b_i>0$ for every $i=1,...,n$.

We can reformulate the problem of finding a solution of equation (1) into a problem of finding a fixed point of a certain transformation from $\mathbb{R}^n$ to $\mathbb{R}^n$, namely:

$$x=S(x),\ \ S(x)=(cI+A)^{-1}(B+cx-F(x)) \tag{2}$$

where $c>0$ is any positive constant. We will use the symbol p to denote the Perron vector of the matrix A^{-1} and use the symbol μ to denote the smallest (by absolute value) eigenvalue of the matrix A (since A is an M-matrix, μ is always real and positive), so that $Ap=\mu p$ and $A^{-1}p=\mu^{-1}p$. In order to prove the main result of this article, we will need the Monotone Fixed Point Theorem [5] applied in the context of our problem.

THEOREM 1.1. (Monotone Fixed Point Theorem applied to $\mathbb{R}^n$) *Consider a space of vectors of $\mathbb{R}^n$ with the partial order relation $<$ defined in the following way: for any two vectors $x=[x_1,...,x_n]^T\in\mathbb{R}^n$ and $y=[y_1,...,y_n]^T\in\mathbb{R}^n$ we will say that x is smaller than y (denoted as $x<y$) if $x_i\leqslant y_i$ for all $i=1,...,n$ and $x_i<y_i$ for at least one i.*

If $x\in\mathbb{R}^n$, $y\in\mathbb{R}^n$ and $x<y$ we can define the interval $[x,y]\subset\mathbb{R}^n$ in the following way: we will say that $t\in[x,y]$ if and only if $x\leqslant t\leqslant y$.

Let $y\in\mathbb{R}^n$ and $z\in\mathbb{R}^n$ are such that $y<z$ and let transformation S: $\mathbb{R}^n\to\mathbb{R}^n$ be defined and continuous on the interval $[y,z]$ and suppose the following conditions are satisfied:

1) $y<S(y)<z$
2) $y<S(z)<z$
3) $y\leqslant x_1<x_2\leqslant z$ *implies* $y<S(x_1)<S(x_2)<z$

Then

a) *the fixed point iteration $x_k=S(x_{k-1})$ with $x_0=y$ converges in $\mathbb{R}^n$: $x_k\to x_*$, $S(x_*)=x_*$, $y<x_*<z$*

b) the fixed point iteration $x_k = S(x_{k-1})$ *with* $x_0 = z$ *converges in* $\mathbb{R}^n$*:* $x_k \to x^*$, $S(x^*) = x^*$, $y < x^* < z$

c) if x *is a fixed point of* S *in* $[y,z]$ *then* $x_* \leqslant x \leqslant x^*$

d) S *has a unique fixed point in* $[y,z]$ *if and only if* $x_* = x^*$

2. Iteration process and its convergence to the positive solution

Now we are ready to prove the main result of this article. It will state the existence of a positive solution of equation (1) for a positive b and convergence of a certain iteration process to this solution.

THEOREM 2.1. *Suppose in the equation (1)* A *is an M-matrix,* $B \in \mathbb{R}^n$ *is a positive vector,* μ *is the smallest (by absolute value) eigenvalue of the matrix* A *(*μ *is always a real positive number), and* $p = [p_1, ..., p_n]^T$ *is a positive eigenvector that corresponds to* μ*. Let*

$$F(x) = \begin{bmatrix} f_1(x_1) \\ f_2(x_2) \\ \vdots \\ f_n(x_n) \end{bmatrix} \tag{3}$$

be such that for every $i = 1, ..., n$ *the components* $f_i : [0, +\infty) \to \mathbb{R}$ *are* $C^1[0, +\infty)$ *functions satisfying the condition*

$$\lim_{t \to 0} f_i(t) = d_i \leqslant 0 \tag{4}$$

and there exist real numbers e_i *and* t_i *such that for every* $t > t_i$ *the following condition holds*

$$\frac{f_i(t)}{t} > e_i > -\mu. \tag{5}$$

Then (1) has a positive solution.

If, in addition to the conditions given above, for every $i = 1, ..., n$ *the following condition holds:*

$$\frac{f_i(s)}{s} < \frac{f_i(t)}{t} \quad \text{whenever} \quad 0 < s < t, \tag{6}$$

then a positive solution is unique and there exists a vector x_0 *such that the sequence* $x_{n+1} = S(x_n)$ *(where* $S(x)$ *is defined as in (2)) converges to the unique positive solution of (1). Moreover, there exist positive numbers* α_1 *and* α_2 *(where* $\alpha_1 < \alpha_2$*) such that for any* $\alpha \in (0, \alpha_1) \cup (\alpha_2, \infty)$ *the vector* $x_0 = \alpha p$ *will generate the sequence* $x_{n+1} = S(x_n)$ *that will converge to the unique positive solution of (1).*

Before we start the proof of Theorem 2.1, we can point out that the condition (5) is not restrictive. The condition (5) means that for the large values of x the graph of $f_i(x)$ stays above a straight line $y = -\mu x$. Any function that has a limit on infinity that is not

equal to the negative infinity satisfies this condition. Any continuous periodic function also satisfies this condition.

The proof will be based on the ideas that are similar to the ideas of the proof given in [1] and will be based on showing that the conditions of the Monotone Fixed Point Theorem given in [5] are satisfied for the transformation $S(x)$. It will guarantee the existence of positive solution and its uniqueness.

Proof. Now we start the proof of Theorem 2.1.

First of all, the condition (4) guarantees that there exists $\alpha_1 > 0$ such that for any $\beta_1 \in (0, \alpha_1)$ we have $b_i - \mu\beta_1 p_i > f_i(\beta_1 p_i)$ for any $i = 1, ..., n$. To show that, it is enough to take $t = \beta_1 p_i$ and note that $\lim_{t\to 0}(b_i - \mu t) = b_i > 0$ but $\lim_{t\to 0} f_i(t) = d_i \leqslant 0$.

Similarly, the condition (5) guarantees that there exists α_2 (we can always select it so that $\alpha_2 > \alpha_1$) such that for any $\beta_2 \in (\alpha_2, \infty)$ we have $b_i - \mu\beta_2 p_i < f_i(\beta_2 p_i)$ for any $i = 1, ..., n$. To show that, we can take $t = \beta_2 p_i$, rewrite the inequality that we want to prove as $\frac{b_i}{t} - \mu < \frac{f_i(t)}{t}$ and note that $\lim_{t\to\infty}\left(\frac{b_i}{t}\right) = 0$.

We can always choose a positive number c such that

$$c > \max_{1\leqslant i\leqslant n}\left(\sup_{\beta_1 p_i \leqslant t \leqslant \beta_2 p_i} |f_i'(t)|\right) \tag{7}$$

and then we reformulate the problem of finding the solution of equation (1) into a problem of finding a fixed point of the transformation $S : \mathbb{R}^n \to \mathbb{R}^n$, where S is defined as in (2), i.e.

$$S(x) = (cI + A)^{-1}(B + cx - F(x)).$$

To prove the existence of a solution of (1), it is enough to show that $S(x)$ satisfies the conditions of the Monotone Fixed Point Theorem given above for $y = \beta_1 p$ and $z = \beta_2 p$.

First of all, we want to show now that the condition (1) of the Monotone Fixed Point Theorem is satisfied, i.e. $y < S(y) < z$. In the proof below we will use the fact that $p = (A + cI)^{-1}(c + \mu)p$. We will also use the fact that if $c > 0$ and A is an M-matrix then $A + cI$ is also an M-matrix.

Since $(A + cI)^{-1}u > 0$ whenever $u > 0$ it suffices to show that

$$(c + \mu)(\beta_1 p) < B + c\beta_1 p - F(\beta_1 p) < (c + \mu)(\beta_2 p) \tag{8}$$

We start from proving the left part of this double inequality. The argument below is valid for $i = 1, ..., n$. From $b_i - \mu\beta_1 p_i > f_i(\beta_1 p_i)$ we have $b_i - \mu\beta_1 p_i - f_i(\beta_1 p_i) > 0$ and so $(c + \mu)(\beta_1 p_i) + b_i - \mu\beta_1 p_i - f_i(\beta_1 p_i) > (c + \mu)(\beta_1 p_i)$ and after canceling out the term $\mu\beta_1 p_i$ in the left part we will get $b_i + c\beta_1 p_i - f_i(\beta_1 p_i) > (c + \mu)(\beta_1 p_i)$, which is exactly the componentwise notation of the left part of the double inequality (8).

Now let us prove the right part of the double inequality (8).

By our choice of β_2 for any $i = 1, ..., n$ we have $f_i(\beta_2 p_i) > b_i - \mu\beta_2 p_i$. It can be rewritten as $f_i(\beta_1 p_i) + (f_i(\beta_2 p_i) - f_i(\beta_1 p_i)) > b_i - \mu\beta_2 p_i$ or, if we estimate the change in f by its derivative multiplied by change in argument, and use (7), we will obtain that $f_i(\beta_1 p_i) + c(\beta_2 p_i - \beta_1 p_i) > b_i - \mu\beta_2 p_i$ and after moving some terms into the left part

we get $f_i(\beta_1 p_i) - c\beta_1 p_i - b_i > -\mu\beta_2 p_i - c\beta_2 p_i$ If we multiply it by -1 we will get $(\mu + c)\beta_2 p_i > b_i + c\beta_1 p_i - f_i(\beta_1 p_i)$, which is exactly the componentwise notation of the right part of equality (8).

We also need to prove that the second condition of the Monotone Fixed Point Theorem is satisfied, namely, $y < S(z) < z$. Due to the fact that $(A + cI)^{-1}$ is a positive matrix it suffices to show only that

$$(c + \mu)(\beta_1 p) < B + c\beta_2 p - F(\beta_2 p) < (c + \mu)(\beta_2 p) \tag{9}$$

Now we will show that the right part of this double inequality holds. The argument below is valid for $i = 1, ..., n$. From $b_i - \mu\beta_2 p_i < f_i(\beta_2 p_i)$ we have $b_i - \mu\beta_2 p_i - f_i(\beta_2 p_i) < 0$ and so we can write $(c+\mu)(\beta_2 p_i) + b_i - \mu\beta_2 p_i - f_i(\beta_2 p_i) < (c+\mu)(\beta_2 p_i)$ and then after canceling out the term $\mu\beta_2 p_i$ in the left part we get $b_i + c\beta_2 p_i - f_i(\beta_2 p_i) < (c + \mu)(\beta_2 p_i)$, which is exactly the componentwise notation of the right part of the double inequality (9).

Now, let us prove the left part of the double inequality (9). From our choice of β_1 we have $f_i(\beta_1 p_i) < b_i - \mu\beta_1 p_i$ which can be written as $f_i(\beta_2 p_i) - (f_i(\beta_2 p_i) - f_i(\beta_1 p_i)) < b_i - \mu\beta_1 p_i$ or, if we estimate the the change in f by its derivative multiplied by the change in argument, and use (7), we will get $f_i(\beta_2 p_i) - c(\beta_2 p_i - \beta_1 p_i) < b_i - \mu\beta_1 p_i$. and after some simplification it will become $f_i(\beta_2 p_i) - c\beta_2 p_i - b_i < -(\mu + c)(\beta_1 p_i)$. Finally, if we multiply it by -1 we will get $(\mu + c)(\beta_1 p_i) < b_i + c\beta_2 p_i - f_i(\beta_2 p_i)$, which is exactly the componentwise notation of the left part of the double inequality (9).

Now we have to show that if $\beta_1 p \leqslant x_1 < x_2 \leqslant \beta_2 p$ then $S(x_1) < S(x_2)$. It will guarantee that the condition (3) of the Monotone Fixed Point Theorem is satisfied. We can write $S(x_2) - S(x_1)$ as $(A + cI)^{-1}((B + cx_2 - F(x_2)) - (B + cx_1 + F(x_1)))$. Due to the fact that $(A + cI)^{-1}$ is a positive matrix it suffices to show only that $(B + cx_2 - F(x_2)) - (B + cx_1 + F(x_1)) > 0$. It can be shown by some simple algebraic transformations and estimating the change in f by the maximum of its derivative multiplied by the change in argument and then by the use of (7). We have (for i-th component) that $b_i + cx_{2i} - f_i(x_{2i}) - b_i - cx_{1i} - f_i(x_{1i}) = c(x_{2i} - x_{1i}) - (f_i(x_{2i}) - f_i(x_{1i})) > c(x_{2i} - x_{1i}) - c(x_{2i} - x_{1i}) = 0$. So $\beta_1 p \leqslant x_1 < x_2 \leqslant \beta_2 p$ implies $S(x_1) < S(x_2)$.

We have checked that the conditions (1)–(3) of the Monotone Fixed Point Theorem are satisfied. It guarantees that there exists at least one fixed point of the transformation defined by (2) that will also be a solution of the equation (1). The Monotone Fixed Point Theorem also implies that if we choose $x_0 = \beta_1 p$ or $x_0 = \beta_2 p$ then the sequence $x_{n+1} = S(x_n)$ will converge to a fixed point of the transformation $S(x)$ (which will also be a solution of equation (1)). And finally, it states that these fixed point(s) of $S(x)$ will lie between $\beta_1 p$ and $\beta_2 p$, which guarantees the positivity of all components of the solution of (1).

Now we need to prove the uniqueness of the positive solution under the conditions given in the Theorem 2.1. Suppose now that the condition (6), i.e.

$$\frac{f_i(s)}{s} < \frac{f_i(t)}{t} \quad \text{whenever} \quad 0 < s < t$$

is satisfied. We will show that in this case for any two positive solutions x^* and x_* it must be $x^* = x_*$.

Since both x^* and x_* are solutions of our equation, we have

$$Ax^* + F(x^*) = B$$

$$Ax_* + F(x_*) = B$$

By multiplying the first equation by x_*^T and the second equation by x^{*T}, and subtracting the second equation from the first one we will see that

$$x_*^T F(x^*) - x^{*T} F(x_*) = x_*^T B - x^{*T} B$$

or, written in terms of vector components on the left side, that

$$\sum_{i=1}^{n} (f_i(x_i^*)x_{*i} - f_i(x_{*i})x_i^*) = = \sum_{i=1}^{n} \left(x_{*i}x_i^* \left(\frac{f_i(x_i^*)}{x_i^*} - \frac{f_i(x_{i*})}{x_{i*}} \right) \right)$$
$$= (x_*^T - x^{*T})B$$

Since x^*, x_* and B are always positive vectors and $x^* > x_*$, we can use the condition (6) and conclude that the right part of the expression above is always nonpositive while the left part of the expression above is always nonnegative. It may happen only when both of them are equal to zero. But since B is positive, the right hand side is equal to zero only when $x^* = x_*$. So the uniqueness of the positive solution of (1) under the condition (6) is proved.

The proof of Theorem 2.1 is complete. □

Before we proceed to the results of the numerical experiments, we should point out that Theorem 2.1 holds when A is any square matrix such that $(cI + A)^{-1}$ is a positive matrix for some value of c that satisfies the condition (7). So, generally speaking, A does not have to be necessarily an M-matrix. Some matrices related to discretization of operator of the second derivative using orthogonal polynomials are not M-matrices, but for them we can still find c such that (7) holds and $(cI + A)^{-1}$ is a positive matrix. In practical computations, it can be checked directly by computing $(cI + A)^{-1}$. For example, we are using the Legendre polynomials to discretize operator of the second derivative (with zero boundary conditions) at collocation points by a matrix and the iteration process presented in the proof of Theorem 2.1 still works very well in our numerical experiments. The details of such discretization by Legendre polynomials are given in [6], [3].

The second important fact to be aware of is that both the proof of the Theorem 2.1 and the iteration process introduced in the proof of it are still fully valid if we consider a two-dimensional version of this theorem, namely when x is not a vector but a square matrix, A is a Stieltjes matrix and Ax is replaced with $Ax + xA^T$. We can view it as a discrete analogue of the Laplace operator. We run a few computational experiments in two dimensions using the discretization by Legendre polynomials at collocation points that was mentioned before, and we present some of the results in the next section.

3. Some results of computational experiments

We did a few numerical experiments in order to evaluate the performance of the iteration procedure for solving the equation (1) that was described above. The matrix A was a 199×199 matrix of the three-point central finite difference approximation of the second derivative multiplied by -1, i.e.

$$A = \frac{1}{h^2} \cdot \begin{bmatrix} 2 & -1 & & & 0 \\ -1 & 2 & -1 & & \\ & -1 & \ddots & \ddots & \\ & & \ddots & 2 & -1 \\ 0 & & & -1 & 2 \end{bmatrix} \tag{10}$$

Table 1: *Convergence results for the first numerical experiment.*

Iterations	*Residue* (2-Norm)
1	2.0134
2	0.4198
5	0.0028
10	$6.4150 \cdot 10^{-7}$
15	$1.5206 \cdot 10^{-10}$

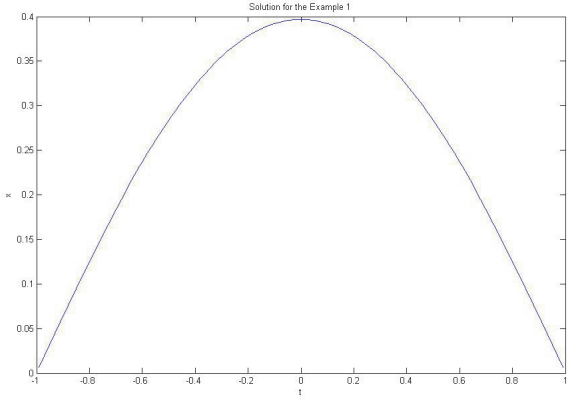

Figure 1: *The solution for the first numerical experiment.*

In the first numerical experiment we have solved the finite difference approximation of the boundary value problem $-(x(t))'' + (x(t))^3 = 1 - t^2$ on the interval $[-1, 1]$ with zero boundary conditions $x(-1) = x(1) = 0$ using a uniform mesh of 199 points. In this case the operator of the second derivative was approximated by the matrix (10)

with $h = 0.01$ while all other functions were approximated by their values at the mesh points. We used the iteration procedure $x_{n+1} = S(x_n)$ with $S(x)$ defined as in (2) with $c = 1$ and the initial iteration x_0 was chosen to be the Perron vector of the matrix A^{-1}. The results of this numerical experiment are given below.

In the second numerical experiment we have solved the finite difference approximation of the boundary value problem $-(x(t))'' + (x(t))^2 = 10e^t \cos(\frac{\pi t}{2})$ on the interval $[-1,1]$ with zero boundary conditions $x(-1) = x(1) = 0$ using a uniform mesh of 199 points. As in the previous example, the operator of the second derivative was approximated by the matrix (10) with $h = 0.01$ while all other functions were approximated by their values at the mesh points. We used the iteration procedure $x_{n+1} = S(x_n)$ (where the transformation $S(x)$ was defined as in (2)) with $c = 10$ and the initial iteration x_0 was chosen to be the Perron vector of the matrix A^{-1}. The results of this numerical experiment are given below.

Table 2: *Convergence results for the second numerical experiment.*

Iterations	*Residue* (2-norm)
1	77.9216
2	48.3652
10	0.2187
20	$1.7290 \cdot 10^{-4}$
30	$1.3633 \cdot 10^{-7}$
40	$2.9634 \cdot 10^{-10}$

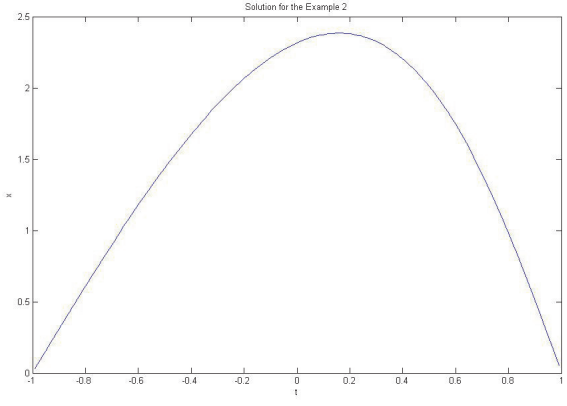

Figure 2: *The solution for the second numerical experiment.*

We also run some computational experiments for a two-dimensional version of (1), that is for $Ax + xA^T + F(x) = B$. In that case, the iteration process (2) has to be slightly modified, but the idea stays the same as for the one-dimensional case. We

used two-dimensional discretization by Legendre polynomials at collocation points and techniques described in [3], [6]. For example, in the third numerical experiment we found a solution $u(x,y)$ to a discretized version of a boundary value problem $-\Delta u + u^3 = 2e^{-\frac{1}{4}\sqrt{x^2+y^2}}$ on the square $[-10,10]\times[10,10]$ with zero boundary conditions on the boundary using a mesh of 64×64 points. We use $c=2$ in the iteration process (2). The results of this numerical experiment are given below.

Table 3: *Convergence results for the third numerical experiment.*

Iterations	*Residue* (2-norm)
1	19.0366
2	12.0016
10	0.2574
20	0.0025
30	$4.2151\cdot 10^{-5}$
40	$9.7700\cdot 10^{-7}$
50	$2.3584\cdot 10^{-8}$

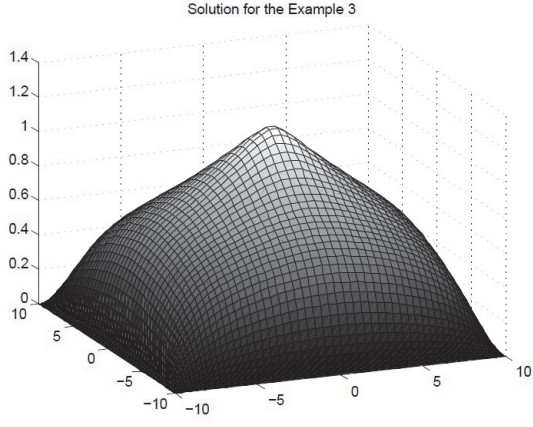

Figure 3: *The solution for the third numerical experiment.*

Our numerical experiments consistently showed that increasing c slowed down convergence of the iteration process. However, c cannot be made too small due to the lower bound (7). The typical choice of c is our numerical experiments was in the interval between 1 and 100.

REFERENCES

[1] Y. S. CHOI, I. KOLTRACHT, P. J. MCKENNA, *A generalization of the Perron-Frobenius theorem for non-linear perturbations of Stieltjes matrices*, Contemporary Mathematics, Volume 281, 2001, pp. 325–330.

[2] Y. S. CHOI, I. KOLTRACHT, P. J. MCKENNA, *On eigen-structure of a nonlinear map in R^n*, Linear Algebra and its Applications, **399** (2005), pp. 141–155.

[3] Y. S. CHOI, J. JAVANAINEN, I. KOLTRACHT, M. KOSTRUN, P. J. MCKENNA, N. SAVYTSKA, *A fast algorithm for the solution of the time-independent Gross-Pitaevskii equation*, Journal of Computational Physics, **190** (2003), pp. 1–21.

[4] R. A. HORN, C. A. JOHNSON, *Topics in Matrix Analysis*, Cambridge University Press, 1994.

[5] L. V. KANTOROVICH, B. Z. VULIKH, A. G. PINSKER, *Functional analysis in Semi-ordered Spaces*, (in Russian Language), Moscow, GosIzdat Technico-Teoreticheskoi Literatury, 1950.

[6] D. GOTTLIEB, S. A. ORSZAG, *Numerical Analysis of Spectral Methods: Theory and Applications*, (CBMS-NSF Regional Conference Series in Applied Mathematics). Society for Industrial Mathematics, 1987.

(Received February 16, 2012)

Yuriy V. Shlapak
University of Wisconsin-Marshfield/Wood County
2000 West 5th Street, Marshfield, WI 54449
e-mail: yuriy.shlapak@uwc.edu