On the positivity of matrix-vector products

Zoltán Horváth

Department of Mathematics, Széchenyi István University, 1 Egyetem square, Győr H-9026, Hungary

Received 1 September 2003; accepted 27 March 2004

Submitted by J.M. Pena

Abstract

In this paper we examine the positivity of $Rv$ where $R \in \mathbb{R}^{N \times N}$, $v \in \mathbb{R}^N$, $v \geq 0$ with $R = r(\tau A)$, $r$ is a given (rational) function, $A \in \mathbb{R}^{N \times N}$ and $\tau \in (0, \infty)$. Here we mean by positivity the ordering w.r.t. an arbitrary order cone, which includes the classical entrywise positivity of vectors. Since the requirement $R \geq 0$ leads to very severe restrictions on $r$ and $\tau$ we construct a positive cone $P = P(A)$ and determine $\tau^* = \tau^*(r, P)$ such that $r(\tau A)P \subset P$ for all $\tau \in [0, \tau^*]$. Finally we give an example arising from applications to partial differential equations where our results explain actual computations much better than the general theory on $R \geq 0$.

© 2004 Elsevier Inc. All rights reserved.

AMS classification: 15A48; 47H07; 65L05

Keywords: Positive matrices; Invariant cone; Positivity; Numerical methods

1. Motivation and the outline of the results

Linear recursions of type

$$u_{n+1} = Ru_n, \quad u_0 \in \mathbb{R}^N, \quad R \in \mathbb{R}^{N \times N} \text{ given} \quad (1)$$

arise frequently in the applications of the linear algebra. For example, (1) approximates the solution of the initial value problem for linear differential equations

$$U'(t) = AU(t) \forall t \geq 0, \quad U(0) = u_0 \in \mathbb{R}^N, \quad A \in \mathbb{R}^{N \times N} \text{ given} \quad (2)$$

in the sense that $u_n \approx U(n\tau)$ whenever $r$ is a rational function approximating the exponential function $\exp$ and

---

E-mail address: horvathz@sze.hu (Z. Horváth).
\( R = r(\tau A) \) with a given rational function \( r \) and \( \tau \in (0, \infty) \),
\[
R = r(\tau A) \quad \text{with a given rational function } r \quad \text{and} \quad \tau \in (0, \infty),
\]
see [1]. We know that \( U(t) \geq 0 \) for all \( t \) (in this section \( \geq \) is meant entrywise) whenever \( -A \) is an M-matrix and \( u_0 \geq 0 \), see e.g. [2]. In these cases \( u_n \geq 0 \) for all \( n \) and all \( u_0 \geq 0 \) is required, see [3] and references therein, which is equivalent to \( R \geq 0 \). However, in general \( R = r(\tau A) \geq 0 \) only under severe conditions found by Bolley and Crouzeix, see [4], namely
\[
r(\tau A) \geq 0 \quad \text{for all } A \quad \text{M-matrix with } A_{ii} \geq -\alpha \quad \text{iff} \quad \tau \leq \frac{\rho_r}{\alpha},
\]
where
\[
\rho_r := \sup \left\{ \rho \geq 0 \mid \frac{d^2 r(\xi)}{d\xi^2} \geq 0 \forall i = 0, 1, \ldots, \forall \xi \in [-\rho, 0] \right\}
\]
is the absolute monotonicity radius of \( r \). However, for many rational functions of great practical interest \( \rho_r = -\infty \), for example for \( r(z) = \frac{1}{1 - z + \frac{z^2}{2}} \), although numerical experiments show that in practice we have \( r(\tau A)v \geq 0 \) with large \( \tau \) and a sort of “smooth” \( v \) positive vectors. Another disadvantage of the application of (4) to investigate positivity of (1) with a particular \( u_0 \) is that \( \alpha \) can be large, as is the case when (2) stems from semi-discretization of partial differential equations (see Example 10). Hence, even if \( \rho_r > 0 \), we get positivity of \( u_n \) by (4) with small \( \tau \) only. We remark that \( \rho_r = \infty \) is possible only with 1st order methods (i.e. when \( r(z) = 1 + z + cz^2 + O(z^3) \) with \( c \neq 1/2 \) when \( z \to 0 \)), see [4], otherwise \( \rho_r \) is typically less than 3 if it is positive at all.

The main goal of the present paper is to determine \( P \), a large class of positive vectors depending on \( A \) and a parameter \( \tau^* \) such that
\[
r(\tau A)v \geq 0 \quad \forall v \in P, \quad \forall \tau \leq \tau^*.
\]
We shall see that \( \tau^* = \frac{\delta_r}{\lambda_1} \) is suitable where \( \delta_r \) is usually much larger than \( \rho_r \) and \( \lambda_1 \) is the largest eigenvalue of \( A \), which is supposed to be corresponding to a strictly positive eigenvector. Considering \( r(z) = 1/(1 - z + \frac{z^2}{2}) \) as an example of 2nd order method we can see that \( r(\tau A)P \subset P \) for all positive \( \tau \), i.e. \( \tau^* = \infty \).

2. Construction of a positive cone

In this section we derive a condition (see Lemma 3) which is sufficient for positivity of vectors. An application of this condition enables us to construct a positive cone.

First we introduce the notations and define the notions we shall use in this paper. For a deeper overview consult e.g. [5, Chapter 1] or [6, Chapter 7].

**Definition 1.** Throughout the paper, let be fixed a positive integer \( N \), \( V := \mathbb{R}^N \), a non-trivial order cone in \( V \) denoted by \( \mathcal{K}(\mathcal{K} \neq \{0\}, V) \) and \( \geq, \gg \) the orderings on \( V \) w.r.t. \( \mathcal{K} \), i.e. \( x \geq y \) iff \( x - y \in \mathcal{K} \) and \( x \gg y \) iff \( x - y \in \text{int}(\mathcal{K}) \). Further, we write \( x > y \) if \( x \geq y \) and \( x \neq y \). (See Definition 7.1. in [6].)
As an example we can consider \( \mathcal{K} = [0, \infty)^N \), the positive orthant in \( \mathbb{R}^N \); hence \( \geq \) is meant entrywise.

**Lemma 2.** Assume that \( s_1 \in V \) is fixed such that \( s_1 \gg 0 \). Then \( \exists! \sigma : V \setminus \{0\} \to (0, \infty) \) such that \( \forall v \in V \setminus \{0\} \) we have \( \sigma(v)s_1 \pm v \geq 0 \) and \( \sigma(v) \) is the minimum of the real numbers with this property.

**Proof.** Let \( \alpha_{s_1} \) denote the function (see, e.g. [6, p. 293])

\[
\alpha_{s_1}(w) := \max\{\alpha > 0 \mid s_1 + tw \geq 0 \forall t \in [0, \alpha]\}
\]

for all \( w \gg 0 \).

In case of \( v > 0 \) or \( -v > 0 \) let \( \sigma(v) := 1/\alpha_{s_1}(-v) \) or \( \sigma(v) := 1/\alpha_{s_1}(v) \), respectively. Else let \( \sigma(v) := 1/\max(\alpha_{s_1}(v), \alpha_{s_1}(-v)) \). This \( \sigma \) function clearly satisfies the requirements. \( \square \)

**Lemma 3.** Let \( \{s_k\} \) be a basis of \( V \), \( s_1 \gg 0 \), \( \sigma_k := \sigma(s_k) \) with the \( \sigma \) function defined in Lemma 2. Suppose that

\[
v = \sum_k \eta_k s_k \in V \quad \text{such that} \quad \eta_1 \geq 0 \quad \text{and} \quad \sum_k \sigma_k |\eta_k| \leq 2\eta_1.
\]

Then \( v \gg 0 \).

**Proof.** Let \( \rho_k \) be non-negative real numbers such that \( \rho_k > 0 \) iff \( \eta_k \neq 0 \) and \( \sum_{k=2}^N \rho_k = 1 \). Then we have

\[
v = \eta_1 s_1 + \sum_{k=2}^N \eta_k s_k = \eta_1 \sum_{k=2, \eta_k \neq 0}^N \rho_k \left( s_1 + \frac{\eta_k}{\rho_k \eta_1} s_k \right).
\]

Here the terms in the sum are \( \geq 0 \) whenever, by the definition of \( \sigma \), \( |\frac{\eta_k}{\rho_k \eta_1}| \leq \frac{1}{\sigma_1} \), i.e. \( \rho_k \geq \frac{\sigma_1 |\eta_k|}{\eta_1} \) for all \( k \). For the existence of such a sequence \( \{\rho_k\} \) it is sufficient that \( 1 = \sum_{k=2}^N \rho_k \geq \sum_{k=2}^N \frac{\sigma_1 |\eta_k|}{\eta_1} \), which is equivalent to \( \sum_k \sigma_k |\eta_k| \leq 2\eta_1 \) (observe that \( \sigma_1 = \sigma(s_1) = 1 \)). \( \square \)

**Example 4.** Let \( \mathcal{K} = [0, \infty)^N \), i.e. \( x \geq y \) iff \( x_i \geq y_i \) for all \( i \). Further, let \( h := 1/(N + 1) \) and \( s_{k,i} := \sqrt{2} \sin(k\pi hi) \) for all \( k = 1, \ldots, N \), \( i = 1, \ldots, N \). Then \( \{s_k\} \) is an orthonormal basis (w.r.t. the scalar product \( \langle x, y \rangle := h \sum_i x_i y_i \)) and \( s_1 \gg 0 \).

One can check that \( \forall k : k s_1 \pm s_k \geq 0 \) entrywise, hence \( \sigma_k \leq k \).

For this situation Lemma 3 asserts that the inequality for the Fourier coefficients \( \sum_k |\eta_k| \leq 2\eta_1 \) implies the positivity of \( \sum_k \eta_k s_k \).

Now we are in a position to present the construction of the set of positive vectors, a positive cone we announced previously.

**Definition 5.** Let \( \{s_k\} \) be a given basis of \( V \), \( \mathcal{K} \) an order generating cone in \( V \) and suppose \( s_1 \gg 0 \). Let us consider \( \sigma_k \) defined in Lemma 3.
Then we define
\[
\|v\|_P : V \to [0, \infty), \quad \|v\|_P := \sum k \sigma_k |\eta_k| \quad \text{whenever } v = \sum k \eta_k s_k
\]
and
\[
\mathcal{P} := \{v \in V | \|v\|_P \leq 2 \eta_1\}.
\]

**Lemma 6.** \(\|.\|_P\) is a norm and \(\mathcal{P}\) is a positive cone in \(V\) (the latter means that \(\mathcal{P}\) is a cone and \(\mathcal{P} \subset \mathcal{K}\) for any \(\mathcal{K}\) and \(\{s_k\}\) with \(s_1 \gg 0\). Further, \(\mathcal{P}\) is generating, i.e. \(\mathcal{P} - \mathcal{P} := \{x - y | x, y \in \mathcal{P}\} = V\).

**Proof.** It is straightforward to check that the axioms of the norm and that of the cone are satisfied with \(\|\cdot\|_P\) and \(\mathcal{P}\), respectively. \(\mathcal{P} \subset \mathcal{K}\) follows from Lemma 3. Finally, for any \(v \in V\) we have \(v = (v + \sigma(v)s_1) - \sigma(v)s_1\) and both terms in the subtraction belong to \(\mathcal{P}\). \(\square\)

### 3. The invariance of the cone \(\mathcal{P}\)

In this section we consider the construction of cones presented in the previous section and fit it to matrices such that the cone be invariant w.r.t. the given matrix.

**Theorem 7.** Let \(R \in \mathbb{R}^{N \times N}\) have a complete system of real eigenvectors, which is denoted by \(\{s_k\}_{k=1}^N\). Further \(Rs_k = \lambda_k s_k\). Let \(s_1 \gg 0\) and \(\mathcal{P}\) be constructed according to Definition 5.

Then \(\lambda_1 \geq |\lambda_k|\) for all \(k\) implies \(R\mathcal{P} \subset \mathcal{P}\).

Moreover, \(\lambda_1 > \mu_2 := \max_{1<k<N} |\lambda_k|\) implies that for each \(v = \sum k \eta_k s_k\) with \(\eta_1 > 0\) there exist a non-negative integer \(m_0\) such that \(R^m v \in \mathcal{P}\) for all \(m \geq m_0\).

**Proof.** Suppose \(v = \sum k \eta_k s_k \in \mathcal{P}\), i.e. \(\|v\|_P \leq 2 \eta_1\). Since \(Rv = \sum \lambda_k \eta_k s_k\), we have
\[
\|Rv\|_P = \sum k \sigma_k |\lambda_k \eta_k| \leq \max_k |\lambda_k| \|v\|_P \leq \lambda_1 2 \eta_1,
\]
which was to be proven for the first statement of the theorem. Moreover, \(R^m v = \sum k \lambda_k^m \eta_k s_k \in \mathcal{P}\) iff \(\|R^m v\|_P \leq 2 \lambda_1^m \eta_1\). Since
\[
\|R^m v\|_P = \sum k \sigma_k |\lambda_k^m| \eta_k \leq \lambda_1^m \eta_1 + \mu_2^m \sum_{k \geq 2} \sigma_k |\eta_k|
\]
\[= \lambda_1^m \eta_1 + \mu_2^m (\|v\|_P - \eta_1),\]
we deduce that \(R^m v \in \mathcal{P}\) whenever \(\lambda_1 \mu_2^m \geq \frac{1}{\eta_1} - 1\). Therefore, in case of \(\eta_1 > 0\) and \(\lambda_1 > \mu_2\).
Theorem 9. Let formulate the conditions of positivity in terms of \( \tau \) mating the exponential function, \( r(\tau) \in [0, \infty)^N \), because then \( 0 < \langle v, s_1 \rangle = \eta_1 \).

As we stated in Section 1, in many applications \( R \) is a given function of an underlying matrix \( A \), for example \( R = r(\tau A) \) where \( r \) is a (rational) function approximating the exponential function, \( \tau \in (0, \infty) \). In these situations it is more natural to formulate the conditions of positivity in terms of \( r, \tau \) and \( A \).

**Theorem 9.** Let \( A \in \mathbb{R}^{N \times N} \) have a complete system of real eigenvectors denoted by \( \{s_k\}_{k=1}^N \) and \( As_k = \lambda_k s_k \) with \( 0 > \lambda_1 \geq \lambda_2 \geq \cdots \). Let \( s_1 > 0 \) and \( \mathcal{P} \) be constructed according to Definition 5. Suppose that \( \delta_r \in (0, \infty) \) has the following properties: \( r \) is non-negative and strictly increasing on \([-\delta_r, 0]\) and \( r(-\delta_r) \geq |r(\xi)| \) for all \( \xi \in (-\infty, -\delta_r) \).

Then \( r(\tau A)v \in \mathcal{P} \) for all \( v \in \mathcal{P} \) whenever \( \tau \leq \frac{\delta_r}{\lambda_1} \).

**Proof.** One can check easily that Theorem 7 applies here for \( R = r(\tau A) \).

**Example 10.** The following problem arises from semi-discretization of the heat equation. Let \( h := 1/(N + 1) \), \( A := h^{-2} \text{tridiag}(1, -2, 1) \in \mathbb{R}^{N \times N} \). Then one can check that the \( s_k \) vectors defined in Example 4 are the eigenvectors of \( A \) with eigenvalues \( \lambda_k = -4h^{-2} \sin(k\pi h/2) \). Hence for an application of (4) we need \( \alpha = 2/h^2 \); further \( -\lambda_1 = \pi^2 \sin((\pi h/2)/((\pi h/2))^2 \leq \pi^2 \) for all \( N \).

Thus, for a given \( r \) we have, by (4), \( r(\tau A) \geq 0 \) whenever \( \tau \leq h^2 \rho_r/2 \) and, by Theorem 9, \( r(\tau A) \in \mathcal{P} \) whenever \( \tau \leq \delta_r/\pi^2 \). There is a significant difference between these conditions on \( \tau \). Namely, \( \rho_r = \infty \) can happen only to first order methods (i.e. when \( r(z) = \exp(z) + O(z) \)), but \( \delta_r = \infty \) even for higher order methods. For example \( r(z) = 1/(1 - z + z^2/2) = \exp(z) + O(z^3) \) and \( \delta_r = \infty \) (while \( \rho_r = \infty \)). Moreover, if \( \rho_r, \delta_r > 0 \) for a certain \( r \), the threshold of \( \tau \) in the first case, \( h^2 \rho_r/2 \) is very small if \( N \) is large while that of in the second case, \( \delta_r/\pi^2 \) is constant, i.e. independent of \( N \).

We should remark here that Theorem 9 ensures positivity of the \( u_n \) values in (1) only in the case when \( u_0 \in \mathcal{P} \). We can see as a non-trivial example, that \( u_0 \in \mathcal{P} \) with \( u_{0,i} = ih(1 - ih) (i = 1, \ldots, N) \) in Example 10. Hence now \( u_n \geq 0 \) for all \( n \) and all \( \tau > 0 \), although the general theory on \( R \geq 0 \) does not apply to this situation. However, in many situations of practical interest it seems to be difficult to check whether a given \( u_0 \) belongs to \( \mathcal{P} \), even if the eigenvectors are explicitly known. We may expect only that if the initial vector is smooth in the sense that \( \eta_k \ll \eta_1 \) for
$k \geq k_0$ (implying $\|u_0\|_P/\eta_1$ of moderate size) then $u_m \geq 0$ for all $m \geq m_0$ with $m_0$ of moderate size (c.f. (6)).

Acknowledgements

The author thanks the unknown referee for his/her valuable comments on the presentation of an earlier version of this paper.

References