# Fast Enclosure of Matrix Polynomials*

Shinya Miyajima
Faculty of Engineering, Gifu University, 1-1 Yanagido,
Gifu-shi, Gifu 501-1193, Japan
miyajima@gifu-u.ac.jp

**Abstract**

A method for enclosing matrix polynomials is proposed. This method is applicable when the matrix is diagonalizable and supplies an interval matrix including the matrix polynomial. The computational cost of this method does not depend on the degree of the polynomial. Hence the proposed method is faster than previous approaches when the degree is large. Numerical results show the properties of this method.

**Keywords:** numerical enclosure, matrix polynomial, eigen-decomposition
**AMS subject classifications:** 15A16, 15A23, 65G30

## 1 Introduction

In this paper, we are concerned with the accuracy of a numerically computed matrix polynomial

$$F(X) = c_p X^p + \cdots + c_1 X + c_0 I, \quad c_0, \ldots, c_p \in \mathbb{C},\ X \in \mathbb{C}^{n\times n}, \tag{1}$$

where $c_0, \ldots, c_p$ and $X$ are given, and $I$ is the $n \times n$ identity matrix. The evaluation of the matrix polynomial (1) is required in many methods for computing matrix functions (see [4]).

We consider in this paper the methods for enclosing $F(X)$. The obvious approach is to invoke Horner's method (e.g. [4]) with interval arithmetic. The approach by Paterson and Stockmeyer [8] can also be applied for this purpose by incorporating interval arithmetic. These approaches require computational costs depending on $p$.

The purpose of this paper is to propose a method for enclosing $F(X)$. This method is applicable when $X$ is diagonalizable and supplies an interval matrix including $F(X)$. In this method, eigen-decomposition of $X$, one matrix inversion, and $\mathcal{O}(1)$ dense matrix multiplications are executed. Thus the computational cost of this method does not depend on $p$, so that the method is faster than the above approaches when $p$ is large. The case of large $p$ occurs for example when the matrix exponential is computed via Taylor series without scaling and squaring (see [7]).

*Submitted: April 20, 2012; Revised: June 22, 2012 and November 1, 2012; Accepted: January 26, 2013.

In this paper, we always assume that the eigenvectors of $X$ are well conditioned. The situation that this is not the case would be a natural extension of our work. We leave this as an open challenge.

This paper is organized as follows: In Section 2, theory for enclosing $F(X)$ utilizing the eigen-decomposition is established, and the method based on the theory is proposed. In Section 3, numerical results are reported to show the property of the proposed method. Finally Section 4 summarizes the results in this paper and highlights possible extensions and future work.

# 2 Enclosure Theory

In this section, we present theory for enclosing (1) utilizing the eigen-decomposition of $X$ and propose the method based on the theory which does not require computational cost depending on $p$. Throughout this paper, let $I$ be the $n \times n$ identity matrix. For $M \in \mathbb{C}^{n\times n}$, $M_{ij}$ and $M_{:j}$ denote the $(i,j)$ element and the $j$-th column of $M$, respectively, $|M| := (|M_{ij}|)$ and $M^T := (M_{ji})$. For $M, N \in \mathbb{C}^{n\times n}$, $M \leq N$ means that $M_{ij} \leq N_{ij}$ follows for all $i$ and $j$. For $d_1, \ldots, d_n \in \mathbb{C}$, $\mathrm{diag}(d_1, \ldots, d_n)$ denotes the diagonal matrix whose diagonal elements are $d_1, \ldots, d_n$. Let $e := (1, \ldots, 1)^T \in \mathbb{R}^n$ and $E := (e, \ldots, e) \in \mathbb{R}^{n\times n}$. For $F_c \in \mathbb{C}^{n\times n}$ and $F_r \in \mathbb{R}^{n\times n}$, where all elements in $F_r$ are nonnegative, the notation $<F_c, F_r>$ denotes the interval matrix whose center and radius are $F_c$ and $F_r$, respectively.

We cite Lemmas 1 and 2, and present Lemma 3 which are utilized in the proof of Lemma 4.

**Lemma 1 (E.g. Meyer [5])** *For $S \in \mathbb{C}^{n\times n}$ and $1 \leq p \leq \infty$, if $\|S\|_p < 1$, $I - S$ is nonsingular.*

**Lemma 2 (Miyajima [6])** *Let $S, F \in \mathbb{C}^{n\times n}$ be given and $D_F := \mathrm{diag}(\|F_{:1}\|_\infty, \ldots, \|F_{:n}\|_\infty)$. If $\|S\|_\infty < 1$, it follows that*

$$|(I-S)^{-1}F| \leq |F| + \frac{1}{1-\|S\|_\infty}|S|ED_F.$$

Lemma 3 is a modification of Lemma 2 suited for *enclosing* $(I-S)^{-1}F$.

**Lemma 3** *Let $S$, $F$ and $D_F$ be as in Lemma 2. If $\|S\|_\infty < 1$, it holds that*

$$(I-S)^{-1}F \in \left\langle F, \frac{1}{1-\|S\|_\infty}|S|ED_F \right\rangle.$$

**Proof.** The Neumann series (e.g. [5, Chapter 7]) gives

$$\begin{aligned}(I-S)^{-1}F &= (I+S+S^2+\cdots)F = F + (S+S^2+\cdots)F \\ &\in <F, |(S+S^2+\cdots)F|> \\ &\subseteq <F, (|S|+|S|^2+\cdots)(|F|_{:1}, \ldots, |F|_{:n})>. \end{aligned} \tag{2}$$

For $i = 1, \ldots, n$, it holds from $\|S\|_\infty < 1$ that

$$\begin{aligned}(|S|+|S|^2+\cdots)|F|_{:i} &= |S||F|_{:i} + |S|(|S||F|_{:i}) + \cdots \\ &\leq \||F|_{:i}\|_\infty |S|e + \||S||F|_{:i}\|_\infty |S|e + \cdots \\ &\leq \|F_{:i}\|_\infty |S|e + \|S\|_\infty \|F_{:i}\|_\infty |S|e + \cdots \\ &= \|F_{:i}\|_\infty (1 + \|S\|_\infty + \|S\|_\infty^2 + \cdots)|S|e \\ &= \frac{\|F_{:i}\|_\infty}{1-\|S\|_\infty}|S|e. \end{aligned}$$

This and (2) yield

$$\begin{aligned}(I-S)^{-1}F &\in \left\langle F, \frac{1}{1-\|S\|_\infty}(\|F_{:1}\|_\infty|S|e,\ldots,\|F_{:n}\|_\infty|S|e)\right\rangle \\ &= \left\langle F, \frac{1}{1-\|S\|_\infty}(|S|e,\ldots,|S|e)D_F\right\rangle \\ &= \left\langle F, \frac{1}{1-\|S\|_\infty}|S|ED_F\right\rangle. \qquad \square\end{aligned}$$

Assume, as a result of numerical computation, we have an $n\times n$ complex diagonal matrix $D$ and an $n\times n$ complex matrix $V$ such that $XV\approx VD$. Let $W$ be an approximate inverse of $V$. From Lemmas 1, 2 and 3, we obtain the following lemma:

**Lemma 4** *Let $D$, $V$, $W\in\mathbb{C}^{n\times n}$ be given, $F(X)$ be as in (1), $R:=W(XV-VD)$, $S:=I-WV$, $D_R:=\mathrm{diag}(\|R_{:1}\|_\infty,\ldots,\|R_{:n}\|_\infty)$, $Q:=|R|+|S|ED_R/(1-\|S\|_\infty)$, $D_W:=\mathrm{diag}(\|W_{:1}\|_\infty,\ldots,\|W_{:n}\|_\infty)$ and $Y:=|S|ED_W/(1-\|S\|_\infty)$. If $\|S\|_\infty<1$, then $V$ and $W$ are nonsingular, and $F(X)\in\hat{\boldsymbol{U}}$ follows, where $\hat{\boldsymbol{U}}$ is the result of the interval arithmetic evaluation*

$$V((\cdots(c_p<D,Q>+c_{p-1}I)<D,Q>+\cdots+c_1I)<D,Q>+c_0I)<W,Y>.$$

**Proof.** The inequality $\|S\|_\infty<1$ and Lemma 1 give that $V$ and $W$ are nonsingular. It holds from Lemmas 2 and 3 that

$$|V^{-1}(XV-VD)| = |(I-S)^{-1}R|\le Q, \tag{3}$$
$$V^{-1} = (I-S)^{-1}W\in<W,Y>. \tag{4}$$

From (3), we have

$$\begin{aligned}V^{-1}XV &= D+V^{-1}XV-D=D+V^{-1}(XV-VD)\\ &\in <D,|V^{-1}(XV-VD)|>\\ &\subseteq <D,Q>.\end{aligned} \tag{5}$$

We finally obtain

$$\begin{aligned}F(X) &= V(c_pV^{-1}X^pV+\cdots+c_1V^{-1}XV+c_0I)V^{-1}\\ &= V(c_p(V^{-1}XV)^p+\cdots+c_1V^{-1}XV+c_0I)V^{-1}\\ &= V((\cdots(c_pV^{-1}XV+c_{p-1}I)V^{-1}XV+\cdots+c_1I)V^{-1}XV+c_0I)V^{-1}.\end{aligned}$$

This, (4) and (5) prove $F(X)\in\hat{\boldsymbol{U}}$. $\square$

We formulate and prove Theorem 1 for developing the proposed method.

**Theorem 1** *Let $F(X)$ be as in (1), $D$, $V$, $W$, $S$, $Q$ and $Y$ be as in Lemma 4 and $t:=(\max_j Q_{1j},\ldots,\max_j Q_{nj})^T$. Assume $\|S\|_\infty<1$ and define*

$$\begin{aligned}U_{\mathrm{mid}}^{(p-1)} &:= c_pD+c_{p-1}I,\quad U_{\mathrm{rad}}^{(p-1)}:=|c_p|Q,\\ U_{\mathrm{mid}}^{(k)} &:= U_{\mathrm{mid}}^{(k+1)}D+c_kI,\\ U_{\mathrm{rad}}^{(k)} &:= |U_{\mathrm{mid}}^{(k+1)}|Q+U_{\mathrm{rad}}^{(k+1)}|D|+(U_{\mathrm{rad}}^{(k+1)}t,\ldots,U_{\mathrm{rad}}^{(k+1)}t),\ k=p-2,\ldots,0.\end{aligned}$$

*Let $\boldsymbol{U}$ be the result of the interval arithmetic evaluation $V<U_{\mathrm{mid}}^{(0)},U_{\mathrm{rad}}^{(0)}><W,Y>$. Then $F(X)\in\boldsymbol{U}$ holds.*

**Proof.** Let $\hat{\boldsymbol{U}}$ be as in Lemma 4, $\hat{U}_{\rm rad}^{(p-1)} := U_{\rm rad}^{(p-1)}$ and $\hat{U}_{\rm rad}^{(k)} := |U_{\rm mid}^{(k+1)}|Q + \hat{U}_{\rm rad}^{(k+1)}|D| + \hat{U}_{\rm rad}^{(k+1)}Q$. Observe that $\hat{U}_{\rm rad}^{(k)} \leq U_{\rm rad}^{(k)}$ hold for all $k$, since $Q \leq (t, \ldots, t)$. From the definition of the center-radius interval arithmetic (e.g. [1]), $\hat{\boldsymbol{U}}$ coincides with the result of the interval arithmetic evaluation $V<U_{\rm mid}^{(0)}, \hat{U}_{\rm rad}^{(0)}><W, Y>$. This coincidence and $\hat{U}_{\rm rad}^{(k)} \leq U_{\rm rad}^{(k)}$ for all $k$ yield $\hat{\boldsymbol{U}} \subseteq \boldsymbol{U}$. This inclusion and Lemma 4 completes the proof. □

The proposed method computes $\boldsymbol{U}$ using directed roundings. Since $D$ is diagonal, $U_{\rm mid}^{(k)}$ is also diagonal. Thus the computations of $U_{\rm mid}^{(k)}$ and $U_{\rm rad}^{(k)}$ require $\mathcal{O}(n^2)$ operations for each $k$, so that the computation of $U_{\rm rad}^{(0)}$ involves $\mathcal{O}(pn^2)$ operations. The parts in the method where $\mathcal{O}(n^3)$ operations are required are as follows [1]:

- the approximate eigen-decomposition of $X$ to obtain $D$ and $V$
- the approximate inversion of $V$ to obtain $W$
- dense matrix multiplications within $R$, $S$ and $V<U_{\rm mid}^{(0)}, U_{\rm rad}^{(0)}><W, Y>$

From the above, the computational cost of the proposed method does not depend on $p$. Hence we can expect that this method is faster than the approaches discussed in Section 1 when $p$ is large.

# 3 Numerical Results

In this section, we report numerical results to show the properties of the proposed method and performances of our implementation. Let $F(X)$ be as in (1), and $V$ and $Y$ be as in Lemma 4. We used a computer with Intel Xeon 2.66GHz Dual CPU, 4.00GB RAM and MATLAB 7.5 with Intel Math Kernel Library and IEEE 754 double precision. The compared methods are as follows:

H: Horner's method with interval arithmetic

M: The method based on Theorem 1

PS: The function `polyvalm_ps` in the matrix function toolbox [2] (Paterson and Stockmeyer's method) with interval arithmetic

In the method M, the eigen-decomposition and inversion are executed via MATLAB function `eig` and `inv`, respectively.

Let an interval matrix $\boldsymbol{F}$ includes $F(X)$, and $\mathrm{mid}(\boldsymbol{F}_{ij})$ and $\mathrm{rad}(\boldsymbol{F}_{ij})$ be the center and radius of $\boldsymbol{F}_{ij}$, respectively. In order to assess the quality of the enclosures, we define the relative radii

$$\xi_{ij} := \frac{\mathrm{rad}(\boldsymbol{F}_{ij})}{|\mathrm{mid}(\boldsymbol{F}_{ij})| + \mathrm{rad}(\boldsymbol{F}_{ij})}, \quad i, j = 1, \ldots, n.$$

We can regard $-\log_{10} \xi_{ij}$ as the number of correct significant decimal digits, since it roughly corresponds to the number of digits to which the upper and the lower bounds coincide, i.e., the number of significant digits we know to be correct for every entry. We define maximum relative radius MRR and average relative radius ARR as

$$\mathrm{MRR} := \max_{i,j} \xi_{ij} \quad \text{and} \quad \mathrm{ARR} := \left( \prod_{i,j} \xi_{ij} \right)^{\frac{1}{n^2}},$$

[1] If $p = \mathcal{O}(n)$, the computation of $U_{\rm rad}^{(0)}$ will be listed in the items above.

respectively. Hence $-\log_{10}$MRR and $-\log_{10}$ARR represent the minimum and arithmetic mean of the correct digits, respectively. For nonsingular $M \in \mathbb{C}^{n\times n}$, define the condition number $\kappa(M) := \|M\|_2\|M^{-1}\|_2$.

## 3.1 Example 1

In this example, we observe computing times of the methods for various $n$ and $p$. Consider (1) where the vector corresponding to $c_0, \ldots, c_p$ and $X$ are generated by the following MATLAB code:

```
c = (randn(p+1,1)+i*randn(p+1,1)) ./ factorial(0:p)';
X = randn(n) + i*randn(n);  X = X/norm(X);
```

We set $c_0, \ldots, c_p$ such that $(c_0, \ldots, c_p)^T =$ `c`. The function `randn` generates a matrix whose elements are normally distributed pseudo random numbers. The code `factorial(0:p)` gives the vector $(0!, \ldots, p!)$. Table 1 displays the computing times of the methods for various $n$ and $p$.

Table 1: Computing times (sec) in Section 3.1

| $n$ | $p$ | H | M | PS |
|---|---|---|---|---|
| 500 | 10 | 5.296 | 6.248 | 4.739 |
| 500 | 50 | 28.07 | 7.188 | 16.25 |
| 500 | 100 | 56.56 | 8.412 | 28.72 |
| 1000 | 10 | 33.96 | 39.66 | 28.15 |
| 1000 | 50 | 182.1 | 43.86 | 88.37 |
| 1000 | 100 | 369.2 | 49.05 | 150.0 |
| 1500 | 10 | 104.0 | 119.9 | 82.10 |
| 1500 | 50 | 561.2 | 129.6 | 243.2 |
| 1500 | 100 | 1133 | 141.2 | 401.4 |

It can be seen from Table 1 that the computing times of M scarcely increased even when $p$ increased, and M was faster than the other methods when $p$ was large. This result coincides with the discussion in Section 2.

## 3.2 Example 2

In this example, we observe how the magnitudes of the radii change when $\kappa(V)$ increases. Consider (1) where $p = 50$, $c_0, \ldots, c_p$ are obtained similarly to Section 3.1, and $X$ is generated by

```
V = gallery('randsvd',100,cnd);
X = V*diag(randn(n,1)+i*randn(n,1))*inv(V);  X = X/norm(X);
```

We used the Higham's test matrix `randsvd` [3]. Then $\kappa(V) \approx$ `cnd` holds approximately when `cnd` is not large. Table 2 displays MRR and ARR given by the methods for various `cnd`.

We can confirm from Table 2 that MRR and ARR by M increased as `cnd` increased. One of the reason is that the each entry of $Y$ increased as `cnd` increased.

Table 2: Obtained radii in Section 3.2

| cnd | H | | M | | PS | |
|---|---|---|---|---|---|---|
| | MRR | ARR | MRR | ARR | MRR | ARR |
| 1e+0 | 7.2e–12 | 5.9e–14 | 1.7e–9 | 1.7e–11 | 1.1e–12 | 1.7e–14 |
| 1e+2 | 1.2e–11 | 1.3e–14 | 6.9e–7 | 9.3e–10 | 6.7e–13 | 2.5e–15 |
| 1e+4 | 1.9e–11 | 9.8e–15 | 1.6e–2 | 2.0e–6 | 3.9e–12 | 2.2e–15 |
| 1e+6 | 1.2e–9 | 1.1e–14 | 1.0e+0 | 1.3e–2 | 1.9e–10 | 2.1e–15 |

# 4 Conclusion

In this paper, we proposed a method for enclosing matrix polynomial (1), and reported numerical results to show the properties of this method. The method was faster than the approaches discussed in Section 1 when $p$ was large, and not ineffective when eigenvectors of $X$ were not ill-conditioned. By modifying this algorithm slightly, enclosing $F(X)$ where $c_0, \ldots, c_p$ and/or $X$ are intervals is also possible. Our future work will be to develop a method which is effective even when the eigenvectors are ill-conditioned.

# Acknowledgements

The author wish to thank the referees for valuable comments. This research was partially supported by Grant-in-Aid for Scientific Research (C) (23560066, 2011–2015) from the Ministry of Education, Science, Sports and Culture of Japan.

# References

[1] H.–R. Arndt. On the interval systems $[x] = [A][x] + [b]$ and the powers of interval matrices in complex interval arithmetics. *Reliab. Comput.*, 13:245–259, 2007.

[2] N.J. Higham. The matrix function toolbox. `http://www.maths.manchester.ac.uk/~higham/mftoolbox/`.

[3] N.J. Higham. *Accuracy and Stability of Numerical Algorithms, second ed.* SIAM Publications, Philadelphia, 2002.

[4] N.J. Higham. *Functions of Matrices: Theory and Computation.* SIAM Publications, Philadelphia, 2008.

[5] C.D. Meyer. *Matrix Analysis and Applied Linear Algebra.* SIAM Publications, Philadelphia, 2000.

[6] S. Miyajima. Fast enclosure for solutions of sylvester equations. submitted for publication, 2011.

[7] C. Moler and C. Van Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Rev.*, 45:3–49, 2006.

[8] M.S. Paterson and L.J. Stockmeyer. On the number of nonscalar multiplications necessary to evaluate polynomials. *SIAM J. Comput.*, 2:60–66, 1973.