Roots of Stochastic Matrices and Fractional Matrix Powers

Lin, Lijing

2011

MIMS EPrint: **2011.9**

Manchester Institute for Mathematical Sciences
School of Mathematics

The University of Manchester

Reports available from: http://eprints.maths.manchester.ac.uk/
And by contacting: The MIMS Secretary
School of Mathematics
The University of Manchester
Manchester, M13 9PL, UK

ISSN 1749-9097

ROOTS OF STOCHASTIC MATRICES AND FRACTIONAL MATRIX POWERS

A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy
in the Faculty of Engineering and Physical Sciences

2011

Lijing Lin
School of Mathematics

Contents

List of Tables

List of Figures

The University of Manchester

Lijing Lin
Doctor of Philosophy
Roots of Stochastic Matrices and Fractional Matrix Powers
January 12, 2011

In Markov chain models in finance and healthcare a transition matrix over a certain time interval is needed but only a transition matrix over a longer time interval may be available. The problem arises of determining a stochastic pth root of a stochastic matrix (the given transition matrix). By exploiting the theory of functions of matrices, we develop results on the existence and characterization of stochastic pth roots. Our contributions include characterization of when a real matrix has a real pth root, a classification of pth roots of a possibly singular matrix, a sufficient condition for a pth root of a stochastic matrix to have unit row sums, and the identification of two classes of stochastic matrices that have stochastic pth roots for all p. We also delineate a wide variety of possible configurations as regards existence, nature (primary or nonprimary), and number of stochastic roots, and develop a necessary condition for existence of a stochastic root in terms of the spectrum of the given matrix.

On the computational side, we emphasize finding an approximate stochastic root: perturb the principal root $A^{1/p}$ or the principal logarithm $\log(A)$ to the nearest stochastic matrix or the nearest intensity matrix, respectively, if they are not valid ones; minimize the residual $\|X^p - A\|_F$ over all stochastic matrices X and also over stochastic matrices that are primary functions of A. For the first two nearness problems, the global minimizers are found in the Frobenius norm. For the last two nonlinear programming problems, we derive explicit formulae for the gradient and Hessian of the objective function $\|X^p - A\|_F^2$ and investigate Newton's method, a spectral projected gradient method (SPGM) and the sequential quadratic programming method to solve the problem as well as various matrices to start the iteration. Numerical experiments show that SPGM starting with the perturbed $A^{1/p}$ to minimize $\|X^p - A\|_F$ over all stochastic matrices is method of choice.

Finally, a new algorithm is developed for computing arbitrary real powers A^α of a matrix $A \in \mathbb{C}^{n\times n}$. The algorithm starts with a Schur decomposition, takes k square roots of the triangular factor T, evaluates an $[m/m]$ Padé approximant of $(1-x)^\alpha$ at $I - T^{1/2^k}$, and squares the result k times. The parameters k and m are chosen to minimize the cost subject to achieving double precision accuracy in the evaluation of the Padé approximant, making use of a result that bounds the error in the matrix Padé approximant by the error in the scalar Padé approximant with argument the norm of the matrix. The Padé approximant is evaluated from the continued fraction representation in bottom-up fashion, which is shown to be numerically stable. In the squaring phase the diagonal and first superdiagonal are computed from explicit formulae for $T^{\alpha/2^j}$, yielding increased accuracy. Since the basic algorithm is designed for $\alpha \in (-1, 1)$, a criterion for reducing an arbitrary real α to this range is developed, making use of bounds for the condition number of the A^α problem. How best to compute A^k for a negative integer k is also investigated. In numerical experiments the new algorithm is found to be superior in accuracy and stability to several alternatives, including the use of an eigendecomposition, a method based on the Schur–Parlett algorithm with our new algorithm applied to the diagonal blocks and approaches based on the formula $A^\alpha = \exp(\alpha \log(A))$.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement

Publications

- The material in Chapter 2 is based on the paper:

  Nicholas J. Higham and Lijing Lin. On pth roots of stochastic matrices. *Linear Algebra Appl.*, In Press, 2010. doi: 10.1016/j.laa.2010.04.007.

- The material in Chapter 4 is based on the paper:

  Nicholas J. Higham and Lijing Lin. A Schur–Padé algorithm for fractional powers of a matrix. MIMS EPrint 2010.91, Manchester Institute for Mathematical Sciences, The University of Manchester, UK, October 2010. 25 pp. Submitted to *SIAM J. Matrix Anal. Appl.*

Acknowledgements

First, I must acknowledge my immense debt of gratitude to my supervisor, Professor Nicholas J. Higham, for his excellent guidance and his essential influence on my way of thinking and writing. Nick has closely followed through the whole way of my studies as a research student, always quickly responding to inquiries, carefully reading manuscripts, generously sharing his knowledge and expertise and constantly offering valuable comments. Without him, finishing this thesis would not have been possible.

I am pleased to thank Dr Françoise Tisseur—teacher, advisor and role model, for her many helpful suggestions and continuous support over the past three years. I appreciate her time in carefully reading this work and many insightful comments concerning the thesis.

I thank those people whose advice and help have benefited me a lot during my PhD work. Many thanks go to Professor Steve Kirkland of Hamilton Institute, National University of Ireland Maynooth. I had useful discussions with him on stochastic roots problem. I appreciate his private note on the imprimitive stochastic matrices which contributes to Section 2.6.7 and his valuable comments and suggestions on the whole thesis. I thank Professor Ilse Ipsen of North Carolina State University for sending me a note on the stochastic symmetric positive semidefinite matrices which contributes to Section 2.6.8. I thank Professor Paul Van Dooren for pointing out Theorem 2.31 on our group meeting during his visit in 2008, which is later exploited in Section 2.5.2. A big thank you goes to Dr Awad Al-Mohy for many valuable suggestions on the numerical experiments in matrix fractional powers. I also had enjoyable discussions with Professor Ernesto Estrada, Professor Chun-Hua Guo, Dr Bruno Iannazzo and Professor Yangfeng Su during their visits at Manchester.

A special thank you goes to Maha Al-Ammari, for always knowing the answer, having an advice, being there and making Office 2.111 the best place to work in; to Rüdiger Borsdorf for his unfailing enthusiasm and helpful discussions on SPGM and generously sharing his MATLAB codes; to Chris Munro who always does exactly what he plans to, for useful technical discussions, proofreading my documents and introducing me to the South Manchester Parkrun.

I would like to acknowledge the financial support from the Secretary of State for Education and Science of the United Kingdom, and the School of Mathematics at the University of Manchester under the Overseas Research Students Awards Scheme (ORSAS) during the last three years. The travel support from the School of Mathematics at the University of Manchester to attend the 23rd Biennial Conference on Numerical Analysis in 2009 and the Gene Golub SIAM Summer School in 2010 is gratefully acknowledged.

Last but by no means least, for many reasons, thanks to my parents.

Dedication

To My Parents

Chapter 1

Introduction

The history of matrix functions dates back to 1858 when Cayley in his *A Memoir on the Theory of Matrices* treated the square roots of 2×2 and 3×3 matrices. Some remarkable time points in this long history are: Sylvester first stated the definition of $f(A)$ for general f via the interpolating polynomial in 1883 [126]; the first book on matrix functions was written by Schwerdtfeger and published in 1938 [116]; in the same year, Frazer, Duncan and Collar published the book *Elementary Matrices and Some Applications to Dynamics and Differential Equations* which was "the first book to treat matrices as a branch of applied mathematics" [30], [72]. For a brief history of matrix functions, we can do no better than refer the reader to [72, sec. 1.10]. Over the past 100 years, matrix functions have developed from their origin in pure mathematics into a flourishing subject of study in applied mathematics, with a growing number of applications ranging from natural science, engineering to social science. Such applications include, to name a few, differential equations, nuclear magnetic resonance and social networks; for more applications, see [72, Chap. 2]. New applications are regularly being found.

A major theme of this thesis is functions of structured matrices. The problem of computing a function of a structured matrix is of growing importance and what makes it a deep and fascinating subject is the new applications appearing and the many open questions remaining in it. This thesis is concerned with this very active area of research.

One issue involved in structured $f(A)$ problems is whether or not $f(A)$ will preserve the structure of A or, more generally, how $f(A)$ inherits structure from A (possibly with different, but related structures). Simple but not trivial examples are that, the square root function preserves the property of being unitary while the exponential function maps a skew-Hermitian matrix into a unitary matrix. However, based on a more general setting of matrix automorphism groups and the Lie algebra, more general results can be found: the square root function preserves matrix automorphism groups; the exponential map takes the Lie algebra into the corresponding Lie group. For details in the square root function and other structure preserving functions for matrix automorphism groups, see [77]. The exponential mapping on the Lie algebra is important in the numerical solution of ODEs on Lie groups by geometric integration methods. For details, see [61], [83], [84].

The other important issue is: assuming we know that A and $f(A)$ are both structured, can we exploit the structure? For example, can we by any means derive a

structure-preserving iteration to get the structured $f(A)$, in the presence of rounding and truncation errors? The potential benefits to accrue from exploiting the structure include faster and more accurate algorithms and reduced storage, and a possibly more physically meaningful solution. Take again the matrix square root function of A in an automorphism group for example, in which case a family of coupled iterations that preserve the automorphism group is derived in [77] by exploiting the matrix sign function. Methods for computing square roots of some other special classes of matrices of practical importance, including matrices close to the identity or with "large diagonal", M-matrices, H-matrices, and Hermitian positive definite matrices are investigated in [72, sec. 6.8].

We address both main issues in this thesis. Motivated by its widespread applications, our work starts with this simply stated problem: determine a stochastic root of a stochastic matrix. A stochastic matrix, also known as transition matrix in Markov models, is a square matrix with nonnegative entries and row sums equal to 1. For a time-homogeneous discrete-time Markov model in which individuals move among n states, the transition matrix $A \in \mathbb{R}^{n\times n}$ has (i, j) entry equal to the probability of transition from state i to state j over a time interval. In credit risk, for example, a transition matrix records the probabilities of a firm's transition from one credit rating to another. Often in practice, the shortest period over which a transition matrix can be estimated is one year. However, for valuation purposes, a transition matrix for a period shorter than one year is usually needed. A short term transition matrix can be obtained by computing a root of an annual transition matrix. This requires a stochastic root of a given stochastic matrix A, that is, a stochastic matrix X such that $X^p = A$, where p is typically an integer, but could be rational.

A number of questions arise: does such a root exist; if so, how can one be computed; and what kind of approximation should be used if a stochastic root does not exist. The first question about the existence of stochastic root has not previously been investigated in any depth. A quick answer is: a stochastic root of a stochastic does not always exist. In other words, the matrix pth root function does not preserve the structure of being stochastic. This is illustrated by the following example. Let $A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$. It is easy to check that there are four matrices satisfying $X^2 = A$ which are given by $\pm\frac{1}{2}\begin{bmatrix} 1+i & 1-i \\ 1-i & 1+i \end{bmatrix}$ and $\pm\frac{1}{2}\begin{bmatrix} 1-i & 1+i \\ 1+i & 1-i \end{bmatrix}$, neither of which is real, let alone stochastic. We go beyond this point and in Chapter 2, by exploiting the theory of functions of matrices, we develop results on the existence and characterization of matrix pth roots, and in particular on the existence of stochastic pth roots of stochastic matrices. Regarding the second question, various methods are available for computing matrix pth roots, based on the Schur decomposition and appropriate recurrences [57], [120], Newton or inverse Newton iterations [60], [79], Padé iterations [80], [98], or a variety of other techniques [14], [59]; see [72, Chap. 7] and [74] for surveys. However, there are currently no methods tailored to finding a stochastic root. Current approaches are based on computing *some* pth root and perturbing it to be stochastic [26], [85], [95]. We consider more computational matters as well as some popular techniques used in statistics in Chapter 3.

More generally, matrix powers A^α with a real α arise in, for example, fractional differential equations [81], discrete representations of norms corresponding to finite element discretizations of fractional Sobolev spaces [8], and the computation of geodesic-midpoints in neural networks [46]. Here, α is an arbitrary real number, not necessarily rational. In the case where α is the reciprocal of an integer p, $X = A^\alpha = A^{1/p}$ is a pth root of A. As we mentioned before, various methods are available for the pth root problem. However, none of these methods is applicable for A^α with arbitrary real α. MATLAB is capable of computing arbitrary matrix powers, which are specified with the syntax `A^t`. However, in versions up to MATLAB R2010b (the latest version at the time of writing), the computed results can be very inaccurate, as the following example shows:

```
>> A = [1 1e-8; 0 1];
>> A^0.1
ans =
     1     0
     0     1
>> expm(0.1*logm(A))
ans =
  1.0000e+000  1.0000e-009
            0  1.0000e+000
```

Here, the second evaluation, via `expm` and `logm`, produces the exact answer. The first evaluation is inaccurate because the algorithm used to compute `A^t` when t is not an integer apparently employs an eigenvalue decomposition and so cannot cope reliably with defective (as here) or "nearly" defective matrices.

The aim of our work in Chapter 4 is to devise a reliable algorithm for computing A^α for arbitrary A and α—one that, in particular, could be used by the MATLAB `mpower` function, which is the underlying function invoked by the `A^t` syntax in MATLAB. Some numerical experiments illustrating the superiority of the new algorithm over several alternatives in accuracy and stability are presented. In the rest of this chapter, we establish some of the basic definitions and properties for matrix theories and matrix functions, which will be used throughout this thesis.

1.1 Functions of matrices

We are concerned with functions mapping $\mathbb{C}^{n\times n}$ to $\mathbb{C}^{n\times n}$ that are defined in terms of an underlying scalar function f. There are various equivalent ways to define a matrix function. We give the following two definitions of $f(A)$, one by Jordan canonical form and the other by polynomial interpolation, both of which are very useful in developing the theory.

It is a standard result that any matrix $A \in \mathbb{C}^{n\times n}$ can be expressed in the Jordan canonical form

$$Z^{-1}AZ = J = \operatorname{diag}(J_1, J_2, \ldots, J_p), \tag{1.1a}$$

$$J_k = J_k(\lambda_k) = \begin{bmatrix} \lambda_k & 1 & & \\ & \lambda_k & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_k \end{bmatrix} \in \mathbb{C}^{m_k \times m_k}, \tag{1.1b}$$

where Z is nonsingular and $m_1 + m_2 + \cdots + m_p = n$. Denote by $\lambda_1, \ldots, \lambda_s$ the distinct eigenvalues of A and let n_i be the order of the largest Jordan block in which λ_i appears, which is called the *index* of λ_i. We call the function f being *defined on the spectrum* of A if the values $f^{(j)}(\lambda_i)$, $j = 0 : n_i - 1$, $i = 1 : s$ exist. We now give the definition of $f(A)$ via Jordan canonical form.

Definition 1.1 (matrix function via Jordan canonical form)**.** *Let f be defined on the spectrum of $A \in \mathbb{C}^{n\times n}$ and let A have the Jordan canonical form* (1.1)*. Then*

$$f(A) := Zf(J)Z^{-1} = Z\mathrm{diag}(f(J_k))Z^{-1}, \tag{1.2}$$

where

$$f(J_k) := \begin{bmatrix} f(\lambda_k) & f'(\lambda_k) & \cdots & \dfrac{f^{(m_k-1)}(\lambda_k)}{(m_k-1)!} \\ & f(\lambda_k) & \ddots & \vdots \\ & & \ddots & f'(\lambda_k) \\ & & & f(\lambda_k) \end{bmatrix}. \tag{1.3}$$

Some comments on Definition 1.1 are made in order. First, the definition yields an $f(A)$ that can be shown to be independent of the particular Jordan canonical form. Second, in the case of multivalued functions such as $\sqrt{t}$ and $\log t$ it is implicit that a single branch has been chosen in (1.3). Moreover, this definition yields a *primary* matrix function which requires that if an eigenvalue occurs in more than one Jordan block then the same choice of branch must be made in each block. If the latter requirement is violated then a *nonprimary* matrix function is obtained. We are mainly concerned with primary matrix functions in developing the theory while nonprimary functions are sometimes of practical importance in applications, as discussed in Chapter 2. For more about nonprimary matrix functions, see [72, sec. 1.4].

Before giving the second definition, we recall some background on polynomials at matrix argument. The *minimal polynomial* of $A \in \mathbb{C}^{n\times n}$ is defined to be the unique monic polynomial ϕ of lowest degree such that $\phi(A) = 0$. The existence and uniqueness of the minimal polynomial can be found in most textbooks on linear algebra. By considering the Jordan canonical form it is not hard to see that $\phi(t) = \prod_{i=1}^{s}(t - \lambda_i)^{n_i}$, where $\lambda_1, \ldots, \lambda_s$ are the distinct eigenvalues of A and n_i is the index of λ_i. It follows immediately that ϕ is zero on the spectrum of A. Now given any polynomial $p(t)$ and any matrix $A \in \mathbb{C}^{n\times n}$, it is obvious that $p(A)$ is defined and that $p(t)$ is defined on the spectrum of A. For polynomials p and q, $p(A) = q(A)$ if and only if p and q take the same values on the spectrum (see [72, Thm. 1.3]). Thus the matrix $p(A)$ is completely determined by the values of p on the spectrum of A. The following definition gives a way to generalize this property of polynomials to arbitrary functions and define $f(A)$ completely by the values of f on the spectrum of A.

Definition 1.2 (matrix function via Hermite interpolation)**.** *Let f be defined on the spectrum of $A \in \mathbb{C}^{n\times n}$. Then $f(A) := p(A)$, where p is the polynomial of degree less than $\sum_{i=1}^{s} n_i$ (namely the degree of the minimal polynomial) that satisfies the interpolation conditions*

$$p^{(j)}(\lambda_i) = f^{(j)}(\lambda_i), \quad j = 0 : n_i - 1,\ i = 1 : s. \tag{1.4}$$

There is a unique such p and it is known as the Hermite interpolating polynomial.

Definition 1.1 and Definition 1.2 are equivalent [72, Thm. 1.12]. One of the most important basic properties of $f(A)$ is that $f(A)$ is a polynomial in $A \in \mathbb{C}^{n\times n}$, which is immediate from Definition 1.2. Some other important properties are collected in the following theorem.

Theorem 1.3 ([72, Thm. 1.13])**.** *Let $A \in \mathbb{C}^{n\times n}$ and let f be defined on the spectrum of A. Then*

(a) *$f(A)$ commutes with A;*

(b) *$f(A^T) = f(A)^T$;*

(c) *$f(XAX^{-1}) = Xf(A)X^{-1}$;*

(d) *the eigenvalues of $f(A)$ are $f(\lambda_i)$, where the λ_i are the eigenvalues of A;*

(e) *if X commutes with A then X commutes with $f(A)$;*

(f) *if $A = (A_{ij})$ is block triangular then $F = f(A)$ is block triangular with the same block structure as A, and $F_{ii} = f(A_{ii})$;*

(g) *if $A = \operatorname{diag}(A_{11}, A_{22}, \dots, A_{mm})$ is block diagonal then*

$$f(A) = \operatorname{diag}(f(A_{11}), f(A_{22}), \dots, f(A_{mm})).$$

Proof. The proof is straightforward from Definition 1.1 and 1.2; see [72, Thm. 1.13]. □

The Taylor series is a basic tool for approximating matrix functions applicable to general functions. Before giving a theorem that guarantees the validity of a matrix Taylor series, we explain first how $f(J_k)$ in (1.3) can be obtained from Taylor series considerations. In (1.1b) write $J_k = \lambda_k I + N_k \in \mathbb{C}^{m_k\times m_k}$, where N_k is zero except for a superdiagonal of 1s. For example, for $m_k = 3$ we have

$$N_k = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \quad N_k^2 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad N_k^3 = 0.$$

In general, powering N_k causes the superdiagonal of 1s to move a diagonal at a time towards the top right-hand corner, until at the m_kth power it disappears: $N_k^{m_k} = 0$. Assume that f has a convergent Taylor series expansion

$$f(t) = f(\lambda_k) + f'(\lambda_k)(t - \lambda_k) + \cdots + \frac{f^{(j)}(\lambda_k)(t - \lambda_k)^j}{j!} + \cdots.$$

On substituting $J_k \in \mathbb{C}^{m_k \times m_k}$ for t we have the finite series

$$f(J_k) = f(\lambda_k)I + f'(\lambda_k)N_k + \cdots + \frac{f^{(j)}(\lambda_k)N_k^{m_k-1}}{j!},$$

since all powers of N_k from the m_kth onwards are zero. This expression is easily seen to agree with (1.3). A more general result is given in the following theorem.

Theorem 1.4 (convergence of matrix Taylor series)**.** *Suppose f has a Taylor series expansion*

$$f(z) = \sum_{k=0}^{\infty} a_k(z-\alpha)^k \quad \left(a_k = \frac{f^{(k)}(\alpha)}{k!}\right) \tag{1.5}$$

with radius of convergence r. If $A \in \mathbb{C}^{n\times n}$ then $f(A)$ is defined and is given by

$$f(A) = \sum_{k=0}^{\infty} a_k(A-\alpha I)^k \tag{1.6}$$

if and only if each of the distinct eigenvalues $\lambda_1, \ldots, \lambda_s$ of A satisfies one of the conditions

(a) $|\lambda_i - \alpha| < r$,

(b) $|\lambda_i - \alpha| = r$ *and the series for $f^{(n_i-1)}(\lambda)$ (where n_i is the index of λ_i) is convergent at the point $\lambda = \lambda_i$, $i = 1:s$.*

Proof. See [72, Thm. 4.7]. □

A very important issue involved in the computation of matrix functions is the conditioning. Due to the inexactness and uncertainty of the data and rounding errors from finite precision computations, the latter of which can often be interpreted as being equivalent to perturbations in the data, it is important to understand the sensitivity of $f(A)$ to perturbations in A. Sensitivity is measured by condition numbers defined as follows.

Definition 1.5. *Let $f : \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}$ be a matrix function. The relative condition number of f is defined as*

$$\operatorname{cond}(f, A) := \lim_{\epsilon\to 0} \sup_{\|E\| \le \epsilon\|A\|} \frac{\|f(A+E) - f(A)\|}{\epsilon\|f(A)\|}, \tag{1.7}$$

where the norm is any matrix norm.

To obtain explicit expressions for $\operatorname{cond}(f, A)$, we need an appropriate notion of derivative for matrix functions. The *Fréchet derivative* of a matrix function $f : \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}$ at a point $A \in \mathbb{C}^{n\times n}$ is a linear mapping

$$\begin{array}{ccc} \mathbb{C}^{n\times n} & \overset{L_f(A)}{\longrightarrow} & \mathbb{C}^{n\times n} \\ E & \mapsto & L_f(A,E) \end{array}$$

such that for all $E \in \mathbb{C}^{n\times n}$

$$f(A+E) = f(A) + L_f(A,E) + o(\|E\|).$$

Therefore, the condition number $\mathrm{cond}(f, A)$ can be characterized as

$$\mathrm{cond}(f, A) = \frac{\|L_f(A)\|\,\|A\|}{\|f(A)\|}, \tag{1.8}$$

where

$$\|L_f(X)\| := \max_{Z \neq 0} \frac{\|L_f(X, Z)\|}{\|Z\|}. \tag{1.9}$$

We now define the eigenvalues of the Fréchet derivative. An eigenpair (λ, V) of $L_f(A)$ comprises a scalar λ, the eigenvalue, and a nonzero matrix $V \in \mathbb{C}^{n\times n}$, the eigenvector, such that $L_f(A, V) = \lambda V$. Since L_f is a linear operator

$$\mathrm{vec}(L_f(A, E)) = K(A)\mathrm{vec}(E) \tag{1.10}$$

for some $K(A) \in \mathbb{C}^{n^2\times n^2}$ that is independent of E. We refer to $K(A)$ as the Kronecker form of the Fréchet derivative. Recall that if we take $a = \mathrm{vec}(A)$, $y = \mathrm{vec}(f(A))$ and $f : a \mapsto y$ as a map from $\mathbb{C}^{n^2}$ to itself, then $K(A)$ is the *Jacobian matrix* of f with (i, j) entry equal to $(\partial f(a)/\partial a_{ij})$.

If (λ, V) is an eigenpair of $L_f(A)$ then $K(A)v = \lambda v$, where $v = \mathrm{vec}(V)$, so (λ, v) is an eigenpair of $K(A)$ in the usual matrix sense. For the rest of this section $\mathcal{D}$ denotes an open subset of $\mathbb{R}$ or $\mathbb{C}$. We now identify eigenpairs of $L_f(A)$.

Theorem 1.6 (eigenvalues of Fréchet derivative)**.** *Let f be $2n-1$ times continuously differentiable on $\mathcal{D}$ and let $A \in \mathbb{C}^{n\times n}$ have spectrum in $\mathcal{D}$. The eigenvalues of the Fréchet derivative of f at A are $f[\lambda_i, \lambda_j]$, $i, j = 1 : n$, where the λ_i are the eigenvalues of A and the divided difference $f[\lambda, \mu]$ is defined by*

$$f[\lambda, \mu] = \begin{cases} \dfrac{f(\lambda) - f(\mu)}{\lambda - \mu}, & \lambda \neq \mu, \\ f'(\lambda), & \lambda = \mu. \end{cases}$$

If u_i and v_j are nonzero vectors such that $Au_i = \lambda_i u_i$ and $v_j^T A = \lambda_j v_j^T$, then $u_i v_j^T$ is an eigenvector of $L_f(A)$ corresponding to $f[\lambda_i, \lambda_j]$.

Proof. See [72, Thm. 3.9]. ∎

Theorem 1.6 enables us to deduce when the Fréchet derivative is nonsingular.

Corollary 1.7 ([72, Cor. 3.10])**.** *Let f be $2n-1$ times continuously differentiable on $\mathcal{D}$. The Fréchet derivative L of f at a matrix $A \in \mathbb{C}^{n\times n}$ with eigenvalues $\lambda_i \in \mathcal{D}$ is nonsingular when $f'(\lambda_i) \neq 0$ for all i and $f(\lambda_i) = f(\lambda_j) \Rightarrow \lambda_i = \lambda_j$.*

1.2 Nonnegative matrices

We recall some background results from the theory of nonnegative matrices, which will be needed in Chapter 2. Recall that $A \in \mathbb{R}^{n\times n}$, $n \geq 2$, is *reducible* if there is a permutation matrix P such that

$$P^T A P = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}, \tag{1.11}$$

where A_{11} and A_{22} are square, nonempty submatrices. A is *irreducible* if it is not reducible. We write $X \geq 0$ ($X > 0$) to denote that the elements of X are all nonnegative (positive), and denote by $\rho(A)$ the spectral radius of A, $e = [1, 1, \ldots, 1]^T$ the vector of 1s, and e_k the unit vector with 1 in the kth position and zeros elsewhere.

In the next theorem we recall some key facts from Perron–Frobenius theory [12, Chap. 2], [78, Chap. 8], [96, Chap. 15].

Theorem 1.8 (Perron–Frobenius)**.** *If $A \in \mathbb{R}^{n\times n}$ is nonnegative then $\rho(A)$ is an eigenvalue of A with a corresponding nonnegative eigenvector. If, in addition, A is irreducible then*

(a) *$\rho(A) > 0$;*

(b) *there is an $x > 0$ such that $Ax = \rho(A)x$;*

(c) *$\rho(A)$ is a simple eigenvalue of A (that is, it has algebraic multiplicity 1).*

Let A be an irreducible nonnegative matrix and suppose that A has exactly h eigenvalues of modulus $\rho(A)$. The number h is called the *index of imprimitivity* of A. If $h = 1$, then the matrix A is said to be *primitive*; otherwise, it is *imprimitive* (we will investigate this particular structure in Section 2.6.7). For more background on the theory of nonnegative matrices, see Berman and Plemmons [12] and Minc [106].

Chapter 2

On pth Roots of Stochastic Matrices

2.1 Introduction

Discrete-time Markov chains are in widespread use for modelling processes that evolve with time. Such processes include the variations of credit risk in the finance industry and the progress of a chronic disease in healthcare, and in both cases the particular problem considered here arises.

In credit risk, a transition matrix records the probabilities of a firm's transition from one credit rating to another over a given time interval [114]. The shortest period over which a transition matrix can be estimated is typically one year, and annual transition matrices can be obtained from rating agencies such as Moody's Investors Service and Standard & Poor's. However, for valuation purposes, a transition matrix for a period shorter than one year is usually needed. A short term transition matrix can be obtained by computing a root of an annual transition matrix. A six-month transition matrix, for example, is a square root of the annual transition matrix. This property has led to interest in the finance literature in the computation or approximation of roots of transition matrices [85], [95]. Exactly the same mathematical problem arises in Markov models of chronic diseases, where the transition matrix is built from observations of the progression in patients of a disease through different severity states. Again, the observations are at an interval longer than the short time intervals required for study and the need for a matrix root arises [26]. An early discussion of this problem, which identifies the need for roots of transition matrices in models of business and trade, is that of Waugh and Abel [130].

A transition matrix is a stochastic matrix: a square matrix with nonnegative entries and row sums equal to 1. The applications we have described require a stochastic root of a given stochastic matrix A, that is, a stochastic matrix X such that $X^p = A$, where p is typically a positive integer. Mathematically, there are three main questions.

1. Under what conditions does a given stochastic matrix A have a stochastic pth root, and how many roots are there?

2. If a stochastic root exists, how can it be computed?

3. If a stochastic root does not exist, what is an appropriate approximate stochastic root to use in its place?

The focus of this chapter is on the first question, which has not previously been investigated in any depth. In Section 2.2 we recall known results on the existence of matrix pth roots and derive a new characterization of when a real matrix has a real pth root. With the aid of a lemma describing the pth roots of block triangular matrices whose diagonal blocks have distinct spectra, we obtain a classification of pth roots of possibly singular matrices. In Section 2.3 we derive a sufficient condition for a pth root of a stochastic matrix A to have unit row sums; we show that this condition is necessary for primary roots and that a nonnegative pth root always has unit row sums when A is irreducible. We use the latter result to connect the stochastic root problem with the problem of finding nonnegative roots of nonnegative matrices. Two classes of stochastic matrices are identified that have stochastic principal pth roots for all p: one is the inverse M-matrices and the other is a class of symmetric positive semidefinite matrices explicitly obtained from a construction of Soules. In Section 2.4 we demonstrate a wide variety of possible scenarios for the existence and uniqueness of stochastic roots of a stochastic matrix—in particular, with respect to whether a stochastic root is principal, primary, or nonprimary. Conditions for the existence of stochastic roots are investigated in Section 2.5. Given p, we denote by $\mathcal{P} \equiv \mathcal{P}(p)$ the set of stochastic matrices that have stochastic pth roots. The geometry of $\mathcal{P}$ is analyzed in Section 2.5.1, where we show that $\mathcal{P}$ is relative closed as a subset of all stochastic matrices and its relative interior is nonempty. In Section 2.5.2 we exploit results for the inverse eigenvalue problem for stochastic matrices in order to obtain necessary conditions that the spectrum of a stochastic matrix must satisfy in order for the matrix to have a stochastic pth root. Section 2.6 provides some results on the existence of stochastic roots for 2×2 and 3×3 matrices and stochastic matrices with certain structures.

The stochastic root problem is intimately related to the embeddability problem in discrete-time Markov chains, which asks when a nonsingular stochastic matrix A can be written $A = e^Q$ for some Q with $q_{ij} \geq 0$ for $i \neq j$ and $\sum_j q_{ij} = 0$, $i = 1\colon n$. (For background on the embeddability problem see Davies [36] or Higham [72, sec. 2.3] and the references therein.) In Section 2.7 we give a collection of known results in the literature on this problem and explore some facts on its relation to our stochastic roots problem. Finally, some conclusions are given in Section 2.8.

2.2 Theory of matrix pth roots

We are interested in the nonlinear equation $X^p = A$, where p is assumed to be a positive integer. In practice, p might be rational—for example if a transition matrix is observed for a five year time interval but the interval of interest is two years. If $p = r/s$ for positive integer r and s then the problem is to solve the equation $X^r = A^s$, and this reduces to the original problem with $p \leftarrow r$ and $A \leftarrow A^s$, since any positive integer power of a stochastic matrix is stochastic.

We can understand the nonlinear equation $X^p = A$ through the theory of functions of matrices. The following theorem classifies all pth roots of a nonsingular matrix [72, Thm. 7.1], [120] and will be exploited below.

Theorem 2.1 (classification of pth roots of nonsingular matrices)**.** *Let the nonsingular matrix $A \in \mathbb{C}^{n\times n}$ have the Jordan canonical form $Z^{-1}AZ = J = \operatorname{diag}(J_1, J_2, \dots, J_m)$, with Jordan blocks $J_k = J_k(\lambda_k) \in \mathbb{C}^{m_k\times m_k}$, and let $s \le m$ be the number of distinct eigenvalues of A. Let $L_k^{(j_k)} = L_k^{(j_k)}(\lambda_k)$, $k = 1\colon m$, denote the p pth roots of J_k given by*

$$L_k^{(j_k)}(\lambda_k) := \begin{bmatrix} f_{j_k}(\lambda_k) & f'_{j_k}(\lambda_k) & \cdots & \dfrac{f_{j_k}^{(m_k-1)}(\lambda_k)}{(m_k-1)!} \\ & f_{j_k}(\lambda_k) & \ddots & \vdots \\ & & \ddots & f'_{j_k}(\lambda_k) \\ & & & f_{j_k}(\lambda_k) \end{bmatrix}, \tag{2.1}$$

where $j_k \in \{1, 2, \dots, p\}$ denotes the branch of the pth root function $f(z) = \sqrt[p]{z}$. Then A has precisely p^s pth roots that are expressible as polynomials in A, given by

$$X_j = Z\operatorname{diag}(L_1^{(j_1)}, L_2^{(j_2)}, \dots, L_m^{(j_m)})Z^{-1}, \quad j = 1\colon p^s, \tag{2.2}$$

corresponding to all possible choices of $j_1, \dots, j_m$, subject to the constraint that $j_i = j_k$ whenever $\lambda_i = \lambda_k$. If $s < m$ then A has additional pth roots that form parametrized families

$$X_j(U) = ZU\operatorname{diag}(L_1^{(j_1)}, L_2^{(j_2)}, \dots, L_m^{(j_m)})U^{-1}Z^{-1}, \quad j = p^s + 1\colon p^m, \tag{2.3}$$

where $j_k \in \{1, 2, \dots, p\}$, U is an arbitrary nonsingular matrix that commutes with J, and for each j there exist i and k, depending on j, such that $\lambda_i = \lambda_k$ while $j_i \ne j_k$.

In the theory of matrix functions the roots (2.2) are called primary functions of A, and the roots in (2.3), which exist only if A is derogatory (that is, if some eigenvalue appears in more than one Jordan block), are called nonprimary functions [72, Chap. 1]. A distinguishing feature of the primary roots (2.2) is that they are expressible as polynomials in A, whereas the nonprimary roots are not, as discussed in Section 1.1. To give some insight into the theorem and the nature of nonprimary roots, we consider

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

which is already in Jordan form, and for which $m = 2$, $s = 1$. All square roots are given by

$$\pm \begin{bmatrix} 1 & \frac{1}{2} & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad \pm U \begin{bmatrix} 1 & \frac{1}{2} & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix} U^{-1},$$

where from the standard characterization of commuting matrices [72, Thm. 1.25] we find that U is an arbitrary nonsingular matrix of the form

$$U = \begin{bmatrix} a & b & d \\ 0 & a & 0 \\ 0 & e & c \end{bmatrix}.$$

While a nonsingular matrix always has a pth root, the situation is more complicated for singular matrices, as the following result of Psarrakos [113] shows.

Theorem 2.2 (existence of pth root)**.** *$A \in \mathbb{C}^{n\times n}$ has a pth root if and only if the "ascent sequence" of integers d_1, d_2, ... defined by*

$$d_i = \dim(\operatorname{null}(A^i)) - \dim(\operatorname{null}(A^{i-1})) \tag{2.4}$$

has the property that for every integer $\nu \geq 0$ no more than one element of the sequence lies strictly between $p\nu$ and $p(\nu+1)$.

For real A, the above theorems do not distinguish between real and complex roots. The next theorem provides a necessary and sufficient condition for the existence of a real pth root of a real A; it generalizes [78, Thm. 6.4.14], which covers the case $p = 2$, and [128, Cor. to Thm. 1], which applies to nonsingular A.

Theorem 2.3 (existence of real pth root)**.** *$A \in \mathbb{R}^{n\times n}$ has a real pth root if and only if it satisfies the ascent sequence condition* (2.4) *and, if p is even, A has an even number of Jordan blocks of each size for every negative eigenvalue.*

Proof. First, we note that a given Jordan canonical form J is that of some real matrix if and only if for every nonreal eigenvalue λ occurring in r Jordan blocks of size q there are also r Jordan blocks of size q corresponding to $\overline{\lambda}$; in other words, the Jordan blocks of each size for nonreal eigenvalues come in complex conjugate pairs. This property is a consequence of the real Jordan form and its relation to the complex Jordan form [78, sec. 3.4], [96, sec. 6.7].

($\Rightarrow$) If A has a real pth root then by Theorem 2.2 it must satisfy (2.4). Suppose that p is even, that A has an odd number, $2k+1$, of Jordan blocks of size m for some m and some eigenvalue $\lambda < 0$, and that there exists a real X with $X^p = A$. Since a nonsingular Jordan block does not split into smaller Jordan blocks when raised to a positive integer power [72, Thm. 1.36], the Jordan form of X must contain exactly $2k+1$ Jordan blocks of size m corresponding to eigenvalues μ_j with $\mu_j^p = \lambda$, which implies that each μ_j is nonreal since $\lambda < 0$ and p is even. In order for X to be real these Jordan blocks must occur in complex conjugate pairs, but this is impossible since there is an odd number of them. Hence we have a contradiction, so A must have an even number of Jordan blocks of size m for λ.

($\Leftarrow$) A has a Jordan canonical form $Z^{-1}AZ = J = \operatorname{diag}(J_0, J_1)$, where J_0 collects together all the Jordan blocks corresponding to the eigenvalue 0 and J_1 contains the remaining Jordan blocks. Since (2.4) holds for A it also holds for J_0, so J_0 has a pth root W_0, and W_0 can be taken real in view of the construction given in [113, sec. 3]. Form a pth root W_1 of J_1 by taking a pth root of each constituent Jordan block in such a way that every nonreal root has a matching complex conjugate—something that is possible because if p is even, the Jordan blocks of A for negative eigenvalues occur in pairs, by assumption, while the Jordan blocks for nonreal eigenvalues occur in complex conjugate pairs since A is real. Then, with $W = \operatorname{diag}(W_0, W_1)$, we have $W^p = J$. Since the Jordan blocks of W occur in complex conjugate pairs it is similar to a real matrix, Y. With $\sim$ denoting similarity, we have $Y^p \sim W^p = J \sim A$. Since Y^p and A are real and similar, they are similar via a real similarity [78, sec. 3.4]. Thus $A = GY^pG^{-1}$ for some real, nonsingular G, which can be rewritten as $A = (GYG^{-1})^p = X^p$, where X is real. ∎

The next theorem identifies the number of real primary pth roots of a real matrix.

Theorem 2.4. *Let the nonsingular matrix $A \in \mathbb{R}^{n\times n}$ have r_1 distinct positive real eigenvalues, r_2 distinct negative real eigenvalues, and c distinct complex conjugate pairs of eigenvalues. If p is even there are (a) $2^{r_1}p^c$ real primary pth roots if $r_2 = 0$ and (b) no real primary pth roots if $r_2 > 0$. If p is odd there are p^c real primary pth roots.*

Proof. By transforming A to real Schur form[1] R our task reduces to counting the number of real pth roots of the diagonal blocks, since a primary pth root of R has the same quasitriangular structure as R and its off-diagonal blocks are uniquely determined by the diagonal blocks [72, sec. 7.2], [120]. Consider a 2×2 diagonal block C, which contains a complex conjugate pair of eigenvalues. Let

$$Z^{-1}CZ = \operatorname{diag}(\lambda, \overline{\lambda}) = \theta I + i\mu K, \qquad K = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

Then $C = \theta I + \mu W$, where $W = iZKZ^{-1}$, and since $\theta, \mu \in \mathbb{R}$ it follows that $W \in \mathbb{R}^{2\times 2}$. The real primary pth roots of C are $X = ZDZ^{-1} = Z\operatorname{diag}(\alpha + i\beta, \alpha - i\beta)Z^{-1} = \alpha I + \beta W$, where $(\alpha + i\beta)^p = \theta + i\mu$, since the eigenvalues must occur in complex conjugate pairs. There are p such choices, giving p^c choices in total.

Every real eigenvalue must be mapped to a real pth root, and the count depends on the parity of p. There is obviously no real primary pth root if $r_2 > 0$ and p is even, while for odd p any negative eigenvalue $-\lambda$ must be mapped to $-\lambda^{1/p}$, which gives no freedom. Each positive eigenvalue λ yields two choices $\pm\lambda^{1/p}$ for even p, but only one choice $\lambda^{1/p}$ for odd p. This completes the proof. ∎

The next lemma enables us to extend the characterization of pth roots in Theorem 2.1 to singular A. We denote by $\Lambda(A)$ the spectrum of A.

Lemma 2.5. *Let*

$$A = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} \in \mathbb{C}^{n\times n},$$

where $\Lambda(A_{11}) \cap \Lambda(A_{22}) = \emptyset$. Then any pth root of A has the form

$$X = \begin{bmatrix} X_{11} & X_{12} \\ 0 & X_{22} \end{bmatrix},$$

where $X_{ii}^p = A_{ii}$, $i = 1, 2$ and X_{12} is the unique solution of the Sylvester equation $A_{11}X_{12} - X_{12}A_{22} = X_{11}A_{12} - A_{12}X_{22}$.

Proof. It is well known (see, e.g., [72, Prob. 4.3]) that if W satisfies the Sylvester equation $A_{11}W - WA_{22} = A_{12}$ then

$$D = \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix} = \begin{bmatrix} I & -W \\ 0 & I \end{bmatrix}^{-1} \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} \begin{bmatrix} I & -W \\ 0 & I \end{bmatrix} \equiv R^{-1}AR.$$

The Sylvester equation has a unique solution since A_{11} and A_{22} have no eigenvalue in common. It is easy to see that any pth root of $A = RDR^{-1}$ has the form $X = RYR^{-1}$,

[1] Here, R is block upper triangular with diagonal blocks either 1×1 or 2×2, and any 2×2 diagonal blocks have complex conjugate eigenvalues.

where $Y^p = D$. To characterize all such Y we partition Y conformably with D and equate the off-diagonal blocks in $YD = DY$ to obtain the nonsingular Sylvester equations $Y_{12}A_{22} - A_{11}Y_{12} = 0$ and $Y_{21}A_{11} - A_{22}Y_{21} = 0$, which yield $Y_{12} = 0$ and $Y_{21} = 0$, from which $Y_{ii}^p = A_{ii}$, $i = 1, 2$, follows. Therefore

$$X = RYR^{-1} = \begin{bmatrix} I & -W \\ 0 & I \end{bmatrix} \operatorname{diag}(Y_{11}, Y_{22}) \begin{bmatrix} I & -W \\ 0 & I \end{bmatrix}^{-1} = \begin{bmatrix} Y_{11} & Y_{11}W - WY_{22} \\ 0 & Y_{22} \end{bmatrix}.$$

The Sylvester equation for X_{12} follows by equating the off-diagonal blocks in $XA = AX$, and again this equation is nonsingular. □

We can now extend Theorem 2.1 to possibly singular matrices.

Theorem 2.6 (classification of pth roots)**.** *Let $A \in \mathbb{C}^{n\times n}$ have the Jordan canonical form $Z^{-1}AZ = J = \operatorname{diag}(J_0, J_1)$, where J_0 collects together all the Jordan blocks corresponding to the eigenvalue 0 and J_1 contains the remaining Jordan blocks. Assume that A satisfies the condition of Theorem 2.2. All pth roots of A are given by $A = Z\operatorname{diag}(X_0, X_1)Z^{-1}$, where X_1 is any pth root of J_1, characterized by Theorem 2.1, and X_0 is any pth root of J_0.*

Proof. Since A satisfies the condition of Theorem 2.2, J_0 does as well. It suffices to note that by Lemma 2.5 any pth root of J has the form $\operatorname{diag}(X_0, X_1)$, where $X_0^p = J_0$ and $X_1^p = J_1$. □

Among all pth roots the principal pth root is the most used in theory and in practice. For $A \in \mathbb{C}^{n\times n}$ with no eigenvalues on $\mathbb{R}^-$, the closed negative real axis, the principal pth root, written $A^{1/p}$, is the unique pth root of A all of whose eigenvalues lie in the segment $\{ z : -\pi/p < \arg(z) < \pi/p \}$ [72, Thm. 7.2]. It is a primary matrix function and it is real when A is real.

2.3 *p*th roots of stochastic matrices

We now focus on pth roots of stochastic matrices, and in particular the question of the existence of stochastic roots. We will need to exploit some standard properties of stochastic matrices contained in the following result. Recall that $e = [1, 1, \dots, 1]^T$ is the vector of 1s.

Theorem 2.7. *Let $A \in \mathbb{R}^{n\times n}$ be stochastic. Then*

(a) $\rho(A) = 1$*;*

(b) 1 *is a semisimple eigenvalue of A (that is, it appears only in 1×1 Jordan blocks in the Jordan canonical form of A) and has a corresponding eigenvector e;*

(c) *if A is irreducible, then* 1 *is a simple eigenvalue of A.*

Proof. The first part is straightforward. The semisimplicity of the eigenvalue 1 is proved by Minc [106, Chap. 6, Thm. 1.3], while the last part follows from Theorem 1.8. □

For a pth root X of a stochastic A to be stochastic there are two requirements: that X is nonnegative and that $Xe = e$. While $X^p = A$ and $X \geq 0$ together imply

that $\rho(X) = 1$ is an eigenvalue of X with a corresponding nonnegative eigenvector v (by Theorem 1.8), it does not follow that $v = e$. The matrices A and X in Fact 2.24 below provide an example, with $v = [1, 1, 2^{1/2}]^T$. The next result shows that a sufficient condition for a pth root of a stochastic matrix to have unit row sums is that every copy of the eigenvalue 1 of A is mapped to an eigenvalue 1 of X.

Lemma 2.8. *Let $A \in \mathbb{R}^{n\times n}$ be stochastic and let $X^p = A$, where for any eigenvalue μ of X with $\mu^p = 1$ it holds that $\mu = 1$. Then $Xe = e$.*

Proof. Since A is stochastic and so has 1 as a semisimple eigenvalue with corresponding eigenvector e, it has the Jordan canonical form $A = ZJZ^{-1}$ with $J = \mathrm{diag}(I, J_2, J_0)$, where $1 \notin \Lambda(J_2)$, $J_0 \in \mathbb{C}^{k\times k}$ contains all the Jordan blocks corresponding to zero eigenvalues, and $Ze_1 = e$. By Theorem 2.6 any pth root X of A satisfying the assumption of the lemma has the form $X = ZULU^{-1}Z^{-1}$, where $L = \mathrm{diag}(I, L_2, Y_0)$ with $Y_0^p = J_0$, and where $U = \mathrm{diag}(\widetilde{U}, I_k)$ with $\widetilde{U}$ an arbitrary nonsingular matrix that commutes with $\mathrm{diag}(I, J_2)$ and hence is of the form $\widetilde{U} = \mathrm{diag}(\widetilde{U}_1, \widetilde{U}_2)$. Then

$$Xe = ZULU^{-1}Z^{-1}e = ZULU^{-1}e_1 = Z\mathrm{diag}(I, \widetilde{U}_2L_2\widetilde{U}_2^{-1}, Y_0)e_1 = Ze_1 = e,$$

as required. □

The sufficient condition of the lemma for X to have unit row sums is not necessary, as the example $A = \left[\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right]$, $X = \left[\begin{smallmatrix}0&1\\1&0\end{smallmatrix}\right]$, $p = 2$, shows. However, for primary roots the condition is necessary, since every copy of the eigenvalue 1 is mapped to the same root ξ, and $Xe = \xi e$ (which can be proved using the property $f(ZJZ^{-1}) = Zf(J)Z^{-1}$ of primary matrix functions f; see Theorem 1.3), so we need $\xi = 1$. The condition is also necessary when A is irreducible, as the next corollary shows.

Corollary 2.9. *Let $A \in \mathbb{R}^{n\times n}$ be an irreducible stochastic matrix. Then for any nonnegative X with $X^p = A$, $Xe = e$.*

Proof. Since A is stochastic and irreducible, 1 is a simple eigenvalue of A, by Theorem 2.7. As noted just before Lemma 2.8, $X^p = A$ and $X \geq 0$ imply that $\rho(X) = 1$ is an eigenvalue of X, and this is the only eigenvalue μ of X with $\mu^p = 1$, since 1 is a simple eigenvalue of A. Therefore the condition of Lemma 2.8 is satisfied. □

The next result shows an important connection between stochastic roots of stochastic matrices and nonnegative roots of irreducible nonnegative matrices.

Theorem 2.10. *Suppose C is an irreducible nonnegative matrix with positive eigenvector x corresponding to the eigenvalue $\rho(C)$. Then $A = \rho(C)^{-1}D^{-1}CD$ is stochastic, where $D = \mathrm{diag}(x)$. Moreover, if $C = Y^p$ with Y nonnegative then $A = X^p$, where $X = \rho(C)^{-1/p}D^{-1}YD$ is stochastic.*

Proof. The eigenvector x necessarily has positive elements in view of the fact that C is irreducible and nonnegative, by Theorem 1.8. The stochasticity of A is standard (see [106, Chap. 6, Thm. 1.2], for example), and can be seen from the observation that, since $De = x$, $Ae = \rho(C)^{-1}D^{-1}Cx = \rho(C)^{-1}D^{-1}\rho(C)x = e$. We have $X^p = \rho(C)^{-1}D^{-1}Y^pD = \rho(C)^{-1}D^{-1}CD = A$. Finally, the irreducibility of C implies that of A, and hence the nonnegative matrix X has unit row sums, by Corollary 2.9. □

We can identify an interesting class of stochastic matrices for which a stochastic pth root exists for all p. Recall that $A \in \mathbb{R}^{n\times n}$ is a nonsingular M-matrix if $A = sI - B$ with $B \geq 0$ and $s > \rho(B)$. It is a standard property that the inverse of a nonsingular M-matrix is nonnegative [12, Chap. 6].

Theorem 2.11. *If the stochastic matrix $A \in \mathbb{R}^{n\times n}$ is the inverse of an M-matrix then $A^{1/p}$ exists and is stochastic for all p.*

Proof. Since $M = A^{-1}$ is an M-matrix, the eigenvalues of M all have positive real part and hence $M^{1/p}$ exists. Furthermore, $M^{1/p}$ is also an M-matrix for all p, by a result of Fiedler and Schneider [45]. Thus $A^{1/p} = (M^{1/p})^{-1} \geq 0$ for all p, and $A^{1/p}e = e$ follows from the comments following Lemma 2.8, so $A^{1/p}$ is stochastic. □

If $A \geq 0$ and we can compute $B = A^{-1}$ then it is straightforward to check whether B is an M-matrix: we just have to check whether $b_{ij} \leq 0$ for all $i \neq j$ [12, Chap. 6]. An example of a stochastic inverse M-matrix is given in Fact 2.21 below. Another example is the lower triangular matrix

$$A = \begin{bmatrix} 1 & & & \\ \frac{1}{2} & \frac{1}{2} & & \\ \vdots & \vdots & \ddots & \\ \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \end{bmatrix}, \tag{2.5}$$

for which

$$A^{-1} = \begin{bmatrix} 1 & & & & \\ -1 & 2 & & & \\ 0 & -2 & 3 & & \\ \vdots & \vdots & \ddots & \ddots & \\ 0 & 0 & \cdots & -(n-1) & n \end{bmatrix}.$$

Clearly, A^{-1} is an M-matrix and hence from Theorem 2.11, $A^{1/p}$ is stochastic for any positive integer p.

A particular class of inverse M-matrices is the strictly ultrametric matrices, which are the symmetric positive semidefinite matrices for which $a_{ij} \geq \min(a_{ik}, a_{kj})$ for all i, j, k and $a_{ii} > \min\{\, a_{ik} : k \neq i \,\}$ (or, if $n = 1$, $a_{11} > 0$). The inverse of such a matrix is a strictly diagonally dominant M-matrix [102], [107].

Using a construction of Soules [122] (also given in a different form by Perfect and Mirsky [112, Thm. 8]), a class of symmetric positive semidefinite stochastic matrices with stochastic roots can be built explicitly.

Theorem 2.12. *Let $Q \in \mathbb{R}^{n\times n}$ be an orthogonal matrix with first column $n^{-1/2}e$, $q_{ij} > 0$ for $i + j < n + 2$, $q_{ij} < 0$ for $i + j = n + 2$, and $q_{ij} = 0$ for $i + j > n + 2$. If $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$, $\lambda_1 > 0$, and*

$$\frac{1}{n}\lambda_1 + \frac{1}{n(n-1)}\lambda_2 + \frac{1}{(n-1)(n-2)}\lambda_3 + \cdots + \frac{1}{1\cdot 2}\lambda_n \geq 0 \tag{2.6}$$

then

(a) $A = \lambda_1^{-1} Q \mathrm{diag}(\lambda_1, \ldots, \lambda_n) Q^T$ *is a symmetric stochastic matrix;*

(b) *if* $\lambda_1 > \lambda_2$ *then* $A > 0$*;*

(c) *if* $\lambda_n \geq 0$ *then* $A^{1/p}$ *is stochastic for all* p*.*

Proof. (a) is proved by Soules [122, Cor. 2.4]. (b) is shown by Elsner, Nabben, and Neumann [44, p. 327]. To show (c), if $\lambda_n \geq 0$ then $\lambda_1^{1/p} \geq \lambda_2^{1/p} \geq \cdots \geq \lambda_n^{1/p}$ holds and (2.6) trivially remains true with λ_i replaced by $\lambda_i^{1/p}$ for all i and so $A^{1/p}$ is stochastic by (a). ∎

A family of matrices Q of the form specified in the theorem can be constructed as a product of Givens rotations G_{ij}, where G_{ij} is a rotation in the (i, j) plane designed to zero the jth element of the vector it premultiplies and produce a nonnegative ith element. Choose rotations G_{ij} so that

$$Ge := G_{12}G_{23}\dots G_{n-1,n}e = n^{1/2}e_1.$$

Then G has positive elements on and above the diagonal, negative elements on the first subdiagonal, and zeros everywhere else. We have $G^T e_1 = n^{-1/2}e$, and defining Q as G^T with the order of its rows reversed yields a Q of the desired form. For example, for $n = 4$,

$$Q = \begin{bmatrix} 0.5000 & 0.2887 & 0.4082 & 0.7071 \\ 0.5000 & 0.2887 & 0.4082 & -0.7071 \\ 0.5000 & 0.2887 & -0.8165 & 0 \\ 0.5000 & -0.8660 & 0 & 0 \end{bmatrix}.$$

There is a close relation between Theorems 2.11 and 2.12. If $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$ in Theorem 2.12 then A in Theorem 2.12 has the property that A^{-1} is an M-matrix and, moreover, A^{-k} is an M-matrix for all positive integers k [44, Cor. 2.4].

It is possible to generalize Theorem 2.12 to nonsymmetric stochastic matrices with positive real eigenvalues (using [27, sec. 3], for example) but we will not pursue this here.

Finally, we note a more specific result. Marcus and Minc [101] give a sufficient condition for the principal square root of a symmetric positive semidefinite matrix to be stochastic. We will discuss more about this in Section 2.6.

Theorem 2.13. *Let* $A \in \mathbb{R}^{n\times n}$ *be a symmetric positive semidefinite stochastic matrix with* $a_{ii} \leq 1/(n-1)$*,* $i = 1\colon n$*. Then* $A^{1/2}$ *is stochastic.*

Proof. See [101, Thm. 2] or [106, Chap. 5, Thm. 4.2]. ∎

2.4 Scenarios for existence and uniqueness of stochastic roots

Existence and uniqueness of pth roots under the requirement of preserving stochastic structure is not a straightforward matter. We present a sequence of facts that demonstrate the wide variety of possible scenarios. In particular, we show that if the principal pth root is not stochastic there may still be a primary stochastic pth root, and if there is no primary stochastic pth root there may still be a nonprimary stochastic pth root.

Fact 2.14. *A stochastic matrix may have no pth root for any p.* Consider the stochastic matrix $A = J_n(0) + e_n e_n^T \in \mathbb{R}^{n\times n}$, where $J_n(0)$, $n > 2$, is an $n \times n$ Jordan block with eigenvalue 0. The ascent sequence (2.4) is easily seen to be $n-1$ 1s followed by zeros. Hence by Theorem 2.2, A has no pth root for any $p > 1$.

Fact 2.15. *A stochastic matrix may have pth roots but no stochastic pth root.* This is true for even p because if A is nonsingular and has some negative eigenvalues then it has pth roots but may have no real pth roots, by Theorem 2.3. An example illustrating this fact is the stochastic matrix

$$A = \begin{bmatrix} 0.5000 & 0.3750 & 0.1250 \\ 0.7500 & 0.1250 & 0.1250 \\ 0.0833 & 0.0417 & 0.8750 \end{bmatrix}, \qquad \Lambda(A) = \{1, 3/4, -1/4\},$$

which has pth roots for all p but no real pth roots for any even p.

Fact 2.16. *A stochastic matrix may have a stochastic principal pth root as well as a stochastic nonprimary pth root.* Consider the family of 3×3 stochastic matrices [95]

$$X(p,x) = \begin{bmatrix} 0 & p & 1-p \\ x & 0 & 1-x \\ 0 & 0 & 1 \end{bmatrix},$$

where $0 < p < 1$ and $0 < x < 1$, and let $a = px$. The eigenvalues of $X(p,x)$ are 1, $a^{1/2}$, and $-a^{1/2}$. The matrix

$$A = X(p,x)^2 = \begin{bmatrix} a & 0 & 1-a \\ 0 & a & 1-a \\ 0 & 0 & 1 \end{bmatrix}$$

is stochastic. But there is another stochastic matrix $\widetilde{X}$ that is also a square root of A:

$$\widetilde{X} = \begin{bmatrix} a^{1/2} & 0 & 1-a^{1/2} \\ 0 & a^{1/2} & 1-a^{1/2} \\ 0 & 0 & 1 \end{bmatrix}.$$

Note that $\widetilde{X}$ is the principal square root of A (and hence a primary square root) while all members of the family $X(p,x)$ are nonprimary, since A is upper triangular but the $X(p,x)$ are not.

Fact 2.17. *A stochastic matrix may have a stochastic principal pth root but no other stochastic pth root.*

The matrix (2.5) provides an example.

Fact 2.18. *The principal pth root of a stochastic matrix with distinct, real, positive eigenvalues is not necessarily stochastic.*

This fact is easily verified experimentally. For a parametrized example, let

$$D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \alpha & 0 \\ 0 & 0 & \beta \end{bmatrix}, \qquad P = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & 0 \end{bmatrix}, \qquad 0 < \alpha,\ \beta < 1.$$

Then the matrix

$$X = PDP^{-1} = \frac{1}{4}\begin{bmatrix} 1+\alpha+2\beta & 1+\alpha-2\beta & 2-2\alpha \\ 1+\alpha-2\beta & 1+\alpha+2\beta & 2-2\alpha \\ 1-\alpha & 1-\alpha & 2+2\alpha \end{bmatrix} \tag{2.7}$$

has unit row sums, and $A = PD^2P^{-1}$ can be obtained by replacing α, β in (2.7) with α^2, β^2, respectively. Clearly, X is nonnegative if and only if $\beta \le (1+\alpha)/2$ while A is nonnegative if and only if $\beta \le ((1+\alpha^2)/2)^{1/2}$. If we let $(1+\alpha)/2 < \beta \le ((1+\alpha^2)/2)^{1/2}$ then A is stochastic and its principal square root $X = A^{1/2}$ is not nonnegative; moreover, for $\alpha = 0.5$, $\beta = 0.751$ (for example) it can be verified that none of the eight square roots of A is stochastic.

Fact 2.19. *A (row) diagonally dominant stochastic matrix (one for which $a_{ii} \ge \sum_{j\ne i} a_{ij}$ for all i) may not have a stochastic principal pth root.*

The matrix A of the previous example serves to illustrate this fact. For $\alpha = 0.99$, $\beta = 0.9501$,

$$A = \begin{bmatrix} 9.9005\times10^{-1} & 9.9005\times10^{-7} & 9.9500\times10^{-3} \\ 9.9005\times10^{-7} & 9.9005\times10^{-1} & 9.9500\times10^{-3} \\ 4.9750\times10^{-3} & 4.9750\times10^{-3} & 9.9005\times10^{-1} \end{bmatrix}, \tag{2.8}$$

which has strongly dominant diagonal. Yet none of the eight square roots of A is nonnegative.

Fact 2.20. *A stochastic matrix whose principal pth root is not stochastic may still have a primary stochastic pth root.* This fact can be seen from the permutation matrices

$$X = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \qquad A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} = X^2. \tag{2.9}$$

The eigenvalues of A are distinct (they are $-\frac{1}{2} \pm \frac{\sqrt{3}}{2}i$ and 1), so all roots are primary. The matrix X, which is not the principal square root (X has the same eigenvalues as A), is easily checked to be the only stochastic square root of A.

Fact 2.21. *A stochastic matrix with distinct eigenvalues may have a stochastic principal pth root and a different stochastic primary pth root.* As noted in [72, Prob. 1.31], the symmetric positive definite matrix M with $m_{ij} = \min(i, j)$ has a square root Y with

$$y_{ij} = \begin{cases} 0, & i+j \le n, \\ 1, & i+j > n. \end{cases}$$

For example,

$$\begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}^2 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 \\ 1 & 2 & 3 & 3 \\ 1 & 2 & 3 & 4 \end{bmatrix}.$$

It is also known that the eigenvalues of M are $\lambda_k = (1/4)\sec(k\pi/(2n+1))^2$, $k = 1\colon n$, so $\rho(M) = (1/4)\sec(n\pi/(2n+1))^2 =: r_n$ [47]. Since M has all positive elements

it has a positive eigenvector x corresponding to $\rho(M)$ (the Perron vector), and so we can apply Theorem 2.10 to deduce that the stochastic matrix $A = r_n^{-1}D^{-1}MD$, where $D = \mathrm{diag}(x)$, has stochastic square root $X = r_n^{-1/2}D^{-1}YD$, and X obviously has the same anti-triangular structure as Y. Since X is clearly indefinite, it is not the principal square root. However, since the eigenvalues of M, and hence A, are distinct, all the square roots of A are primary square roots. The stochastic square root X has $\lceil n/2 \rceil$ positive eigenvalues and $\lfloor n/2 \rfloor$ negative eigenvalues, which follows from the inertia properties of a 2×2 block symmetric matrix—see, for example, Higham and Cheng [75, Thm. 2.1]. However, X is not the only stochastic square root of A, as we now show.

Lemma 2.22. *The principal pth root of $A = r_n^{-1}D^{-1}MD$ is stochastic for all p.*

Proof. Because the row sums are preserved by the principal pth root, we just have to show that $A^{1/p}$ is nonnegative, or equivalently that $M^{1/p}$ is nonnegative. It is known that M^{-1} is the tridiagonal second difference matrix with typical row $[-1\ 2\ -1]$, except that the (n,n) element is 1. Since M^{-1} has nonpositive off-diagonal elements and M is nonnegative, M^{-1} is an M-matrix and it follows from Theorem 2.11 that $M^{1/p}$ is stochastic for all p. □

For $n = 4$, A and its two stochastic square roots are

$$\begin{bmatrix} 0.1206 & 0.2267 & 0.3054 & 0.3473 \\ 0.0642 & 0.2412 & 0.3250 & 0.3696 \\ 0.0476 & 0.1790 & 0.3618 & 0.4115 \\ 0.0419 & 0.1575 & 0.3182 & 0.4825 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 1.0000 \\ 0 & 0 & 0.4679 & 0.5321 \\ 0 & 0.2578 & 0.3473 & 0.3949 \\ 0.1206 & 0.2267 & 0.3054 & 0.3473 \end{bmatrix}^2$$
$$= \begin{bmatrix} 0.2994 & 0.2397 & 0.2315 & 0.2294 \\ 0.0679 & 0.3908 & 0.2792 & 0.2621 \\ 0.0361 & 0.1538 & 0.4705 & 0.3396 \\ 0.0277 & 0.1117 & 0.2626 & 0.5980 \end{bmatrix}^2.$$

Fact 2.23. *A stochastic matrix without primary stochastic pth roots may have non-primary stochastic pth roots.* Consider the circulant stochastic matrix

$$A = \frac{1}{3}\begin{bmatrix} 1-2a & 1+a & 1+a \\ 1+a & 1-2a & 1+a \\ 1+a & 1+a & 1-2a \end{bmatrix}, \qquad 0 < a \le \frac{1}{3}.$$

The eigenvalues of A are $1, -a, -a$. The four primary square roots X of A are all non-real, because in each case the two negative eigenvalues $-a$ and $-a$ are mapped to the same square root, which means that X cannot have complex conjugate eigenvalues. With $\omega = e^{-2\pi i/3}$, we have

$$A = Q^{-1}DQ, \qquad Q = \begin{bmatrix} 1 & 1 & 1 \\ 1 & \omega & \omega^2 \\ 1 & \omega^2 & \omega \end{bmatrix}, \quad D = \mathrm{diag}(1, -a, -a).$$

Let $X = Q^{-1}\mathrm{diag}(1, ia^{1/2}, -ia^{1/2})Q$. Then

$$X = \frac{1}{3}\begin{bmatrix} 1 & 1+(3a)^{1/2} & 1-(3a)^{1/2} \\ 1-(3a)^{1/2} & 1 & 1+(3a)^{1/2} \\ 1+(3a)^{1/2} & 1-(3a)^{1/2} & 1 \end{bmatrix},$$

which is a stochastic, nonprimary square root of A.

Fact 2.24. *A nonnegative pth root of a stochastic matrix is not necessarily stochastic.* Consider the nonnegative but non-stochastic matrix [99]

$$X = \begin{bmatrix} 0 & 0 & 2^{-1/2} \\ 0 & 0 & 2^{-1/2} \\ 2^{-1/2} & 2^{-1/2} & 0 \end{bmatrix}, \qquad \Lambda(X) = \{1, 0, -1\},$$

for which

$$A = X^{2k} \equiv \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad \Lambda(A) = \{1, 1, 0\}$$

is stochastic. Note that A is its own stochastic pth root for any integer p.

Fact 2.25. *A stochastic matrix may have a stochastic pth root for some, but not all, p.*

Consider again the matrix

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

appearing in Fact 2.20. We have $A^3 = I$, which implies $A^{3k+1} = A$ and $(A^2)^{3k+2} = A^4 = A$ for all nonnegative integers k. Hence A is its own stochastic pth root for $p = 3k + 1$ and A^2 is a stochastic pth root of A for $p = 3k + 2$. However, $\Lambda(A) = \{1, \omega, \overline{\omega}\}$ with $\omega = e^{-2\pi i/3}$, and the arguments in Section 2.5.2 show that A has no stochastic cube root (since ω lies outside the region Θ_3^3 in Figure 2.2). Hence A does not have a stochastic root for $p = 3k$. Note that A is irreducible and all three eigenvalues of A have modulus one, so A is an imprimitive stochastic matrix. We will deal further with examples of this type in Section 2.6.7.

2.5 A necessary condition for the existence of stochastic roots

2.5.1 The geometry of X^p

To investigate the conditions under which a stochastic matrix has stochastic roots, an intuitive, though not simple, method is to study the geometry of the set of *all* stochastic matrices that have stochastic pth roots. We begin our analysis with some definitions which will be needed in this section.

Definition 2.26. *Let S be a subset of $\mathbb{R}^n$. The affine hull of S, denoted by* $\mathrm{aff}(S)$*, is the set of all affine combinations of elements of S*

$$\mathrm{aff}(S) = \left\{ \sum_{i=1}^{k} \alpha_i x_i : \; x_i \in S, \; \alpha_i \in \mathbb{R}, \; \sum_{i=1}^{k} \alpha_i = 1, k = 1, 2, \ldots \right\}.$$

The convex hull of S, denoted by $\mathrm{conv}(S)$*, is the set of all convex combinations of elements of S that requires in the formula above that all α_i be nonnegative. The relative interior of S, denoted by* $\mathrm{ri}(S)$*, is the interior of S considered as a subset of* $\mathrm{aff}(S)$*. S is said to be relatively open if $S = \mathrm{ri}(S)$. S is said to be relatively closed if its complement $\mathbb{R}^n \backslash S$ is relatively open.*

We denote by $\mathcal{S}$ the set of all $n \times n$ stochastic matrices and $\mathcal{N}$ the set of all $n \times n$ nonnegative matrices. It is known that $\mathcal{S}$ is the convex hull of the set of n^n *elementary* stochastic matrices consisting of zeros and ones [63]. Thus, $\mathcal{S}$ is bounded, closed and hence compact in $\mathcal{N}$. Denote by $\mathcal{P} \equiv \mathcal{P}(p)$ the set of stochastic matrices which have stochastic pth roots. Here, we do not require the root to be unique. Since any positive integer power of a stochastic matrix is still stochastic, $\mathcal{P}$ is a subset of $\mathcal{S}$ given by

$$\mathcal{P} = \{X^p : \; X \in \mathcal{S}\}.$$

We have the following proposition.

Proposition 2.27. *$\mathcal{P}$ is relatively closed as a subset of $\mathcal{S}$.*

Proof. Assume we have a sequence $\{A_i\}$ with $A_i \in \mathcal{P}$, $i = 1, 2, \ldots$. We only need to show that, if $A_i \to A$ as $i \to \infty$ then $A \in \mathcal{P}$. For each $A_i \in \mathcal{P}$, there exists some $X_i \in \mathcal{S}$ such that $X_i^p = A_i$. Since $\mathcal{S}$ is closed and bounded in the set of all nonnegative matrices, the matrix sequence X_i is bounded and hence X_i has a convergent subsequence X_{i_k} with a limit X in $\mathcal{S}$. Since $f(X) = X^p$ is a continuous matrix function on $\mathcal{S}$ [72, Theorem 1.19], we have $f(X_{i_k}) \to f(X)$ and hence $A_{i_k} \to X^p = A$ which gives $A \in \mathcal{P}$. This proves our proposition. □

Since $\mathcal{S}$ is a convex set in $\mathcal{N}$, it is natural to ask whether $\mathcal{P}$, the image of $\mathcal{S}$ under the map $f(X) = X^p$ is also convex. For a 2×2 matrix A and even p, the answer is yes since in this situation the necessary and sufficient condition for $A \in \mathcal{P}$ is $\mathrm{trace}(A) \geq 1$ [64]. But even for odd p in the 2×2 case, $\mathcal{P}$ is not necessarily convex. To see this, let

$$X_1 = A_1 = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$$

and

$$X_2 = A_2 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Since $X_1^3 = A_1$ and $X_2^3 = A_2$, we have $A_1, \; A_2 \in \mathcal{P}$. Thus if $\mathcal{P}$ were convex in this case, it would contain

$$A = \frac{1}{2}(A_1 + A_2) = \begin{bmatrix} 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}.$$

However, the only real cube root of A is

$$X = \frac{1}{3}\begin{bmatrix} 1-2^{2/3} & 2+2^{2/3} \\ 1+2^{2/3} & 2-2^{2/3} \end{bmatrix}$$

which has a negative entry, so that $A \notin \mathcal{P}$.

We now consider the interior of $\mathcal{P}$ as a subset of $\mathcal{S}$. We will show that the interior of $\mathcal{P}$ is not empty by exploiting the fact that a homeomorphism maps an open set to an open set. Consider a map $F : \mathbb{R}^N \to \mathbb{R}^N$. It is a standard property that F is locally homeomorphic at $x \in \mathbb{R}^N$ if the corresponding Jacobian matrix is nonsingular [20, Thm. 3]. Now we consider the local homeomorphism of matrix functions. Recall the background knowledge in Section 1.1 on the Fréchet derivative of matrix functions. For a matrix function f to be locally homeomorphic at X, it is sufficient that the Fréchet derivative of f at X is nonsingular. We narrow the case to our functions of matrix powers.

Theorem 2.28. *The map $f : f(X) = X^p$ from $\mathbb{R}^{n\times n}$ into itself is locally homeomorphic except possibly when X has a zero eigenvalue, or a pair of eigenvalues differing by a nonzero multiple of ω where ω is a pth root of 1.*

Proof. As discussed before, f is a local homeomorphism at X if the Fréchet derivative of f at X is nonsingular. From Corollary 1.7, $L_f(X)$ is singular when there exists an eigenvalue λ of X such that $f'(\lambda) = p\lambda^{p-1} = 0$, or when there exists a pair of distinct eigenvalues λ_1, λ_2 with $\lambda_1^p = \lambda_2^p$. A simple calculation yields the result. ∎

Remark 2.29. *Let*

$$\mathcal{S}_0 = \{X \in \mathcal{S} : X = (x_{ij}),\ x_{ii} > 1/(1+\sin(\pi/p)),\ i = 1, 2, \ldots, n\}.$$

From Gershgorin's disk theorem, for any eigenvalue $\lambda = re^{i\alpha}$ of $X \in \mathcal{S}_0$, we have $|\lambda - x_{ii}| \le \sum_{j\neq i} x_{ij} = 1 - x_{ii}$. It follows that $|\sin\alpha| \le (1-x_{ii})/x_{ii} < \sin(\pi/p)$ and then $-\pi/p < \alpha < \pi/p$. Hence, for any $X \in \mathcal{S}_0$, X is nonsingular and no two distinct eigenvalues of X will differ by a multiple of w, where $w^p = 1$. Therefore, when restricted to the set $\mathcal{S}_0$, the map $f(X) = X^p$ is a local homeomorphism.

Proposition 2.30. *The relative interior of $\mathcal{P}$ as a subset of all stochastic matrices $\mathcal{S}$ is nonempty.*

Proof. Let $\mathcal{P}_0$ be the image of $\mathcal{S}_0$ under the map $f : X \mapsto X^p$. Since f is a local homeomorphism on $\mathcal{S}_0$ and $\mathcal{S}_0$ is relatively open as a subset of $\mathcal{S}$, $\mathcal{P}_0$ is relatively open in $\mathcal{P}$, which implies the relative interior of $\mathcal{P}$ as a subset of $\mathcal{S}$ is nonempty. ∎

The content above in this section is of more theoretical interest than numerical interest. The idea here can nevertheless be applied to investigating eigenvalues of the stochastic matrices that have stochastic pth roots and a necessary condition can thus be obtained for the existence of stochastic roots. This will be shown in the next section.

2.5.2 Necessary conditions based on inverse eigenvalue problem

Karpelevič [90] has determined the set Θ_n of all eigenvalues of all stochastic $n \times n$ matrices. This set provides the solution to the inverse eigenvalue problem for stochastic matrices, which asks when a given complex scalar is the eigenvalue of some $n \times n$ stochastic matrix. (Note the distinction with the problem of determining conditions under which a set of n complex numbers comprises the eigenvalues of some $n \times n$ stochastic matrix, which is called the inverse spectrum problem by Minc [106].)

The following theorem gives the main points of Karpelevič's theorem on the characterization of Θ_n; full details on the "specific rules" mentioned therein can be found in [90] and [106, Chap. 7, Thm. 1.8].

Theorem 2.31. *The set Θ_n is contained in the unit disk and is symmetric with respect to the real axis. It intersects the unit circle at points $e^{2i\pi a/b}$ where a and b range over all integers such that $0 \le a < b \le n$. For $n > 3$, the boundary of Θ_n consists of curvilinear arcs connecting these points in circular order. Any point λ on these arcs must satisfy one of the parametric equations*

$$\lambda^q(\lambda^s - t)^r = (1-t)^r, \tag{2.10}$$

$$(\lambda^b - t)^d = (1-t)^d\lambda^q, \tag{2.11}$$

where $0 \le t \le 1$, and b, d, q, s, r are positive integers determined from certain specific rules.

The set Θ_3 of eigenvalues of 3×3 stochastic matrices consists of points in the interior and on the boundary of an equilateral triangle of maximal size inscribed in the unit circle with one of its vertices at the point $(1, 0)$, as well as all points on the segment $[-1, 1]$; see Figure 2.1. The boundary of Θ_4 consists of curvilinear arcs determined by the parametric equations $\lambda^3+\lambda^2+\lambda+t = 0$ and $\lambda^3+\lambda^2-(2t-1)\lambda-t^2 = 0$, $0 \le t \le 1$, together with line segments linking $(1, 0)$ with $(0, 1)$, and $(1, 0)$ with $(0, -1)$, respectively, as can also be seen in Figure 2.1.

Denote by Θ_n^p the set of pth powers of points in Θ_n, i.e., $\Theta_n^p = \{\lambda^p : \lambda \in \Theta_n\}$. If A and X are stochastic $n \times n$ matrices such that $X^p = A$ then for any eigenvalue λ of X, λ^p is an eigenvalue of A. Hence, a necessary condition for A to have a stochastic pth root is that all the eigenvalues of A are in the set Θ_n^p. It can be shown that Θ_n^p is a closed set within the unit disk with boundary $\partial\Theta_n^p \subseteq \{\lambda^p : \lambda \in \partial\Theta_n\}$, where $\partial\Theta_n$ is the boundary of Θ_n, the points on which satisfy the parametric equation (2.10) or (2.11). Figure 2.2 shows the second to fifth powers of Θ_3 and Θ_4.

This approach provides necessary conditions for A to have a stochastic pth root. The conditions are not sufficient, because we are checking whether each eigenvalue of A is the eigenvalue of some pth power of a stochastic matrix, and not that every eigenvalue of A is an eigenvalue of the pth power of the same stochastic matrix.

To illustrate, consider the stochastic matrix

$$A = \begin{bmatrix} 1/3 & 1/3 & 0 & 1/3 \\ 1/2 & 0 & 1/2 & 0 \\ 10/11 & 0 & 0 & 1/11 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{bmatrix}. \tag{2.12}$$

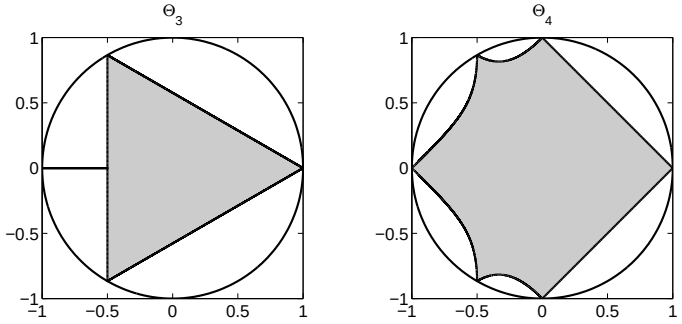

Figure 2.1: The sets Θ_3 and Θ_4 of all eigenvalues of 3×3 and 4×4 stochastic matrices, respectively.

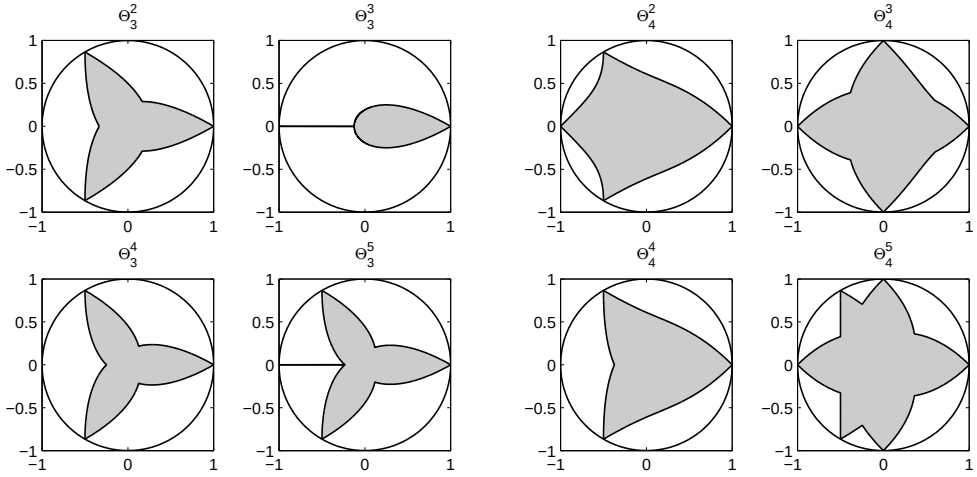

Figure 2.2: Regions obtained by raising the points in Θ_3 (left) and Θ_4 (right) to the powers 2, 3, 4, and 5.

From Figure 2.3 we see that A cannot have a stochastic 12th root, but may have a stochastic 52nd root. In fact, both $A^{1/12}$ and $A^{1/52}$ have negative elements and none of the 52nd roots is stochastic.

If $A \in \mathbb{R}^{n\times n}$ is stochastic then so is the matrix $\mathrm{diag}(A, 1)$ of order $n+1$, and it follows that $\Theta_3 \subseteq \Theta_4 \subseteq \Theta_5 \subseteq \ldots$. Moreover, the number of points at which the region Θ_n intersects the unit circle increases rapidly with n; for example, there are 23 intersection points for Θ_8 and 80 for Θ_{16}. As n increases the region Θ_n and its powers tend to fill the unit circle, so the necessary conditions given in this section are most useful for small dimensions. We emphasize, however, that small matrices do arise in practice; for example, in the model in [26] describing the progression to AIDS in an HIV-infected population the transition matrix[2] is of dimension 5.

[2]This matrix has one negative eigenvalue and a square root is required; that no exact stochastic square root exists follows from Theorem 2.3.

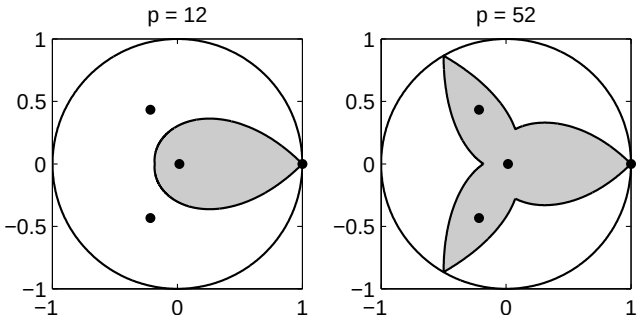

Figure 2.3: Θ_4^p for $p = 12$ and $p = 52$ and the spectrum (shown as dots) of A in (2.12).

2.6 Conditions for structural stochastic matrices

We start this section with general 2×2 and 3×3 stochastic matrices and then proceed to the stochastic matrices with particular structures, including rank 1 matrices, Pei matrix, circulant matrices, upper triangular matrices, irreducible imprimitive stochastic matrices, and symmetric positive definite matrices.

2.6.1 2×2 case.

He and Gunn [64] give all stochastic roots of 2×2 stochastic matrices explicitly. The results shown here are the same as in [64] but stated in a simpler way. A 2×2 stochastic matrix is of the form

$$A = \begin{bmatrix} a & 1-a \\ 1-b & b \end{bmatrix},$$

where $0 \le a, b \le 1$. If $a = b = 1$ then A is an identity matrix and A itself is a stochastic pth root for any integer p. Thus we assume further that there is at most one of a and b equal to 1. Hence, A has the following Jordan decomposition

$$A = \begin{bmatrix} 1 & a-1 \\ 1 & 1-b \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & a+b-1 \end{bmatrix} \begin{bmatrix} \frac{1-b}{2-a-b} & \frac{1-a}{2-a-b} \\ -\frac{1}{2-a-b} & \frac{1}{2-a-b} \end{bmatrix}.$$

Let $x^p = a+b-1$. The pth roots X of A that satisfy $Xe = e$, can be written explicitly

$$\begin{aligned} X &= \begin{bmatrix} 1 & a-1 \\ 1 & 1-b \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & x \end{bmatrix} \begin{bmatrix} \frac{1-b}{2-a-b} & \frac{1-a}{2-a-b} \\ -\frac{1}{2-a-b} & \frac{1}{2-a-b} \end{bmatrix} \\ &= \frac{1}{2-a-b} \begin{bmatrix} 1-b+(1-a)x & (1-a)(1-x) \\ (1-b)(1-x) & 1-a+(1-b)x \end{bmatrix}. \end{aligned} \tag{2.13}$$

Obviously, a necessary condition for X to be stochastic is that $t \equiv a+b-1$ has a real pth root. If p is even, then the necessary condition is $a+b-1 \ge 0$. Let $x = (a+b-1)^{1/p}$ be the principal pth root of $a+b-1$. Since $0 \le a, b \le 1$ implies $a+b-1 < 1$, we have $x < 1$ and hence X in (2.13) is nonnegative. Therefore, if p is even, the necessary and sufficient condition for A to have a stochastic root is

$a+b-1 \geq 0$, i.e., $\text{trace}(A) \geq 1$.

If p is odd, take x the real pth root of $a+b-1$. The pth root X in (2.13) is nonnegative if and only if

$$\begin{cases} 1-b+(1-a)x \geq 0, \\ 1-a+(1-b)x \geq 0, \end{cases}$$

which is equivalent to

$$\begin{cases} a+b-1 \geq -\left(\frac{1-b}{1-a}\right)^p, \\ a+b-1 \geq -\left(\frac{1-a}{1-b}\right)^p. \end{cases}$$

Hence, A has a stochastic root if and only if

$$a+b-1 \geq \max\left\{-\left(\frac{1-b}{1-a}\right)^p, -\left(\frac{1-a}{1-b}\right)^p\right\},$$

i.e., $\text{trace}(A) \geq \max\left\{-\left(\frac{1-b}{1-a}\right)^p, -\left(\frac{1-a}{1-b}\right)^p\right\}+1$.

2.6.2 3×3 case.

Though all stochastic roots of 2×2 stochastic matrices can be found explicitly, there is no similar result for 3×3 stochastic matrices due to the existence of infinitely many nonprimary roots. He and Gunn [64] investigate the primary roots for the 3×3 case where they drop the nonnegativity constraint of the original problem, express the primary pth roots as polynomials of A via the Hermite interpolating polynomial (see Definition 1.2) and identify the existence and number of the real primary pth roots with unit row sums. Let A be a 3×3 stochastic matrix with eigenvalues 1, λ_2 and λ_3 and B be a real matrix such that

$$A = B^p, \quad Be = e. \tag{2.14}$$

We summarize the results from [64] as follows. We make some corrections here and also comment on the existence of real nonprimary roots.

- In the case where $(\text{trace}(A)-1)^2 < 4\det(A)$, namely λ_2 and λ_3 are a pair of complex conjugates, there are in total p real (primary) pth roots B of A satisfying (2.14).
- In the case where $(\text{trace}(A)-1)^2 > 4\det(A)$, namely λ_2 and λ_3 are real and $\lambda_2 \neq \lambda_3$:
  - If p is odd, there is a unique real (primary) pth root B of A satisfying (2.14);
  - If p is even and $\det(A) < 0$, there is no real (primary) pth root B of A satisfying (2.14);
  - If p is even, $\det(A) \geq 0$ and $\text{trace}(A) < 1$, there is no real (primary) pth root B of A satisfying (2.14);
  - If p is even, $\det(A) \geq 0$ and $\text{trace}(A) \geq 1$, there are four real (primary) pth roots B of A satisfying (2.14).

- In the case where $(\text{trace}(A)-1)^2 = 4\det(A)$, namely $\lambda_2 = \lambda_3 = \alpha$:
  - If $\lambda_2 = \lambda_3 = 1$, the only real primary pth root of A is $A = I$ itself; there are infinitely many real nonprimary pth roots B of A satisfying (2.14);
  - If $A^2 - (1+\alpha)A + \alpha I \neq 0$, namely A is non-diagonalizable, and $\lambda_2 = \lambda_3 = \alpha = 0$, then there is no pth root of A;
  - If $A^2 - (1+\alpha)A + \alpha I \neq 0$, $1 > \lambda_2 = \lambda_3 = \alpha \neq 0$ and p is odd, there is a unique real (primary) pth root B of A satisfying (2.14);
  - If $A^2 - (1+\alpha)A + \alpha I \neq 0$, $1 > \lambda_2 = \lambda_3 = \alpha > 0$ and p is even, there are two real (primary) pth roots B of A satisfying (2.14);
  - If $A^2 - (1+\alpha)A + \alpha I \neq 0$, $\lambda_2 = \lambda_3 = \alpha < 0$ and p is even, there is no real (primary) pth root B of A satisfying (2.14);
  - If $A^2 - (1+\alpha)A + \alpha I = 0$, namely A is diagonalizable, and $1 > \lambda_2 = \lambda_3 = \alpha = 0$, then the only real primary pth root B of A satisfying (2.14) is $B = A$; we point out that there are infinitely many real nonprimary pth roots B of A satisfying (2.14);

  We mention in passing that, in the case where $A^2 - (1+\alpha)A + \alpha I = 0$ and $1 > \lambda_2 = \lambda_3 = \alpha \neq 0$, [64] wrongly states that, for either odd or even p, there are possibly p real primary pth roots B of A satisfying (2.14). We correct their results as follows.

  - If $A^2 - (1+\alpha)A + \alpha I = 0$, $1 > \lambda_2 = \lambda_3 = \alpha > 0$ and p is even, there are two real primary pth roots B of A satisfying (2.14); there are infinitely many real nonprimary pth roots B of A satisfying (2.14);
  - If $A^2 - (1+\alpha)A + \alpha I = 0$, $1 > \lambda_2 = \lambda_3 = \alpha > 0$ and p is odd, there is a unique real primary pth roots B of A satisfying (2.14); there are infinitely many real nonprimary pth roots B of A satisfying (2.14);
  - If $A^2 - (1+\alpha)A + \alpha I = 0$, $1 > \lambda_2 = \lambda_3 = \alpha < 0$ and p is even, there is no real primary pth root B of A satisfying (2.14); there are infinitely many real nonprimary pth roots B of A satisfying (2.14);
  - If $A^2 - (1+\alpha)A + \alpha I = 0$, $1 > \lambda_2 = \lambda_3 = \alpha < 0$ and p is odd, there are two real primary pth roots B of A satisfying (2.14); there are infinitely many real nonprimary pth roots B of A satisfying (2.14).

2.6.3 Rank 1 matrices

Let $A = ey^T$, where $y \geq 0$ and $y^T e = 1$. Then A is a stochastic rank 1 matrix. For any positive integer p, $A^p = A$, which means A is a stochastic pth root of itself.

2.6.4 Pei matrix

For the study of matrix inversion, Pei [111] provided a test matrix $T = J + \delta I$, where $J = ee^T$ and δ is a nonzero parameter. The Pei matrix can be generalized to

$$A = \begin{bmatrix} \alpha+\beta & \beta & \cdots & \beta \\ \beta & \alpha+\beta & \cdots & \beta \\ \vdots & \vdots & \ddots & \vdots \\ \beta & \beta & \cdots & \alpha+\beta \end{bmatrix} = \alpha I + \beta J, \tag{2.15}$$

where $\alpha \neq 0$ and $\beta \neq 0$. If we assume further that $\beta > 0$ and $\alpha + n\beta = 1$, then A is a stochastic matrix. We summarize some properties of a stochastic Pei matrix which follow immediately from its definition:

(a) A is symmetric and thus diagonalizable;

(b) A is a circulant matrix;

(c) A is positive definite if and only if $\alpha > 0$;

(d) α is an eigenvalue of multiplicity $n-1$ whose corresponding eigenvector is any vector whose entries sum to 0; the remaining eigenvalue is 1 with the corresponding eigenvector e.

We first show that if $\alpha > 0$, then the principal pth root of A is stochastic. Since the definition of the primary function of a matrix is independent of the particular Jordan canonical form that is used, we choose the following Jordan decomposition for A in the light of the properties of A's eigenvectors:

$$\begin{aligned} A &= \left[\begin{array}{c|c} 1 & e^T \\ \hline e & -I \end{array}\right] \left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & \alpha I \end{array}\right] \left[\begin{array}{c|c} 1 & e^T \\ \hline e & -I \end{array}\right]^{-1} \\ &= \frac{1}{n} \left[\begin{array}{c|c} 1 & e^T \\ \hline e & -I \end{array}\right] \left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & \alpha I \end{array}\right] \left[\begin{array}{c|c} 1 & e^T \\ \hline e & -nI + ee^T \end{array}\right]. \end{aligned} \tag{2.16}$$

Then the principal pth root of A is

$$\begin{aligned} A^{1/p} &= \left[\begin{array}{c|c} 1 & e^T \\ \hline e & -I \end{array}\right] \left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & \alpha^{1/p} I \end{array}\right] \left[\begin{array}{c|c} 1 & e^T \\ \hline e & -I \end{array}\right]^{-1} \\ &= \frac{1}{n} \begin{bmatrix} 1+(n-1)\alpha^{1/p} & (1-\alpha^{1/p})e^T \\ (1-\alpha^{1/p})e & n\alpha^{1/p} I + (1-\alpha^{1/p})ee^T \end{bmatrix} \end{aligned} \tag{2.17}$$

Since $0 < \alpha < 1$, $0 < \alpha^{1/p} < 1$ and thus $A^{1/p}$ is nonnegative. Together with $A^{1/p}e = e$, this implies that $A^{1/p}$ is a stochastic pth root of A.

If $\alpha < 0$, then (2.17) shows that the primary pth roots of A is nonreal for all even p, and hence not stochastic. However, when the multiplicity of the eigenvalue α is even, there may exist nonprimary stochastic pth roots, as can be seen from Fact 2.23. Unfortunately we can not get all the nonprimary roots by simply taking different branches of pth roots of α in (2.16) because the nonprimary pth roots are dependent on the Jordan canonical form (see Theorem 2.1).

If $\alpha < 0$ and p is odd, we can determine a condition under which A has a primary stochastic pth root. Let $\alpha^{1/p}$ be the real pth root of α in (2.17). Then $A^{1/p}$ is a real

matrix with row sums 1. To have a stochastic $A^{1/p}$ we need $A^{1/p}$ to be nonnegative, that is,

$$\begin{cases} 1+(n-1)\alpha^{1/p} \geq 0, \\ 1-\alpha^{1/p} \geq 0. \end{cases}$$

The second inequality is guaranteed by the assumption that $\alpha < 0$. For the first inequality, we have

$$\alpha \geq -\left(\frac{1}{n-1}\right)^p.$$

Thus we get the condition for A to have a primary stochastic pth root. This condition is nontrivial since the conditions for A to be stochastic only imply $\alpha > -\frac{1}{n-1}$.

2.6.5 Circulant stochastic matrices

We start with a Toeplitz matrix

$$\begin{bmatrix} a_0 & a_{-1} & a_{-2} & \cdots & a_{-(n-1)} \\ a_1 & a_0 & a_{-1} & \cdots & a_{-(n-2)} \\ a_2 & a_1 & a_0 & \cdots & a_{-(n-3)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n-1} & a_{n-2} & a_{n-3} & \cdots & a_0 \end{bmatrix},$$

where, to get a stochastic matrix, we assume

$$\begin{cases} a_0 + a_{-1} + a_{-2} + \cdots + a_{-(n-1)} = 1 \\ a_1 + a_0 + a_{-1} + \cdots + a_{-(n-2)} = 1 \\ a_2 + a_1 + a_0 + \cdots + a_{-(n-3)} = 1 \\ \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \\ a_{n-1} + a_{n-2} + a_{n-3} + \cdots + a_0 = 1 \\ a_i \geq 0 \quad i = 0, 1, \ldots, n-1 \\ a_{-i} \geq 0 \quad i = 1, \ldots, n-1 \end{cases}.$$

By subtracting every two successive equalities we have $a_i = a_{-(n-i)}$ for $i = 1, 2, \ldots, n-1$, which implies that any Toeplitz stochastic matrix is indeed a circulant matrix determined by a nonnegative vector $a = [a_0, a_1, \ldots, a_{n-1}]^T$, namely

$$A = \begin{bmatrix} a_0 & a_{n-1} & a_{n-2} & \cdots & a_1 \\ a_1 & a_0 & a_{n-1} & \cdots & a_2 \\ a_2 & a_1 & a_0 & \cdots & a_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n-1} & a_{n-2} & a_{n-3} & \cdots & a_0 \end{bmatrix}, \quad a \geq 0, \quad , \quad e^T a = 1. \tag{2.18}$$

Let F_n be the $n \times n$ discrete Fourier transform (DFT) matrix

$$F_n = \left(\omega^{(r-1)(s-1)}\right)_{r,s=1}^{n} = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega & \omega^2 & \cdots & \omega^{n-1} \\ 1 & \omega^2 & \omega^4 & \cdots & \omega^{2(n-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega^{n-1} & \omega^{2(n-1)} & \cdots & \omega \end{bmatrix},$$

where $\omega = e^{-2\pi i/n}$. It is well-known that circulant matrices can be diagonalized by the DFT matrix F_n [38, sec. 3.2], [70, sec. 23.2]

$$A = F_n^{-1} D F_n, \tag{2.19}$$

where $D = \operatorname{diag}(d)$ with $[d_0, \ldots, d_{n-1}]^T = F_n a$, i.e.,

$$\begin{cases} d_0 = 1 \\ d_k = \sum_{j=0}^{n-1} \omega^{kj} a_j, \quad k = 1, 2, \ldots, n-1. \end{cases} \tag{2.20}$$

Therefore the problem of computing pth roots of A reduces to computing pth roots of the diagonal matrix D. It can be verified that any primary pth root of A is still a circulant matrix. More generally, a primary matrix function of a circulant matrix is circulant. This follows from the fact that $f(A) = F_n^{-1} f(D) F_n$.

Due to the infinite number of nonprimary roots, we restrict our discussion to the primary pth root X of A. Note that the eigenvalues of X is

$$\sigma_k = f^{(j_k)}(d_k), \quad k = 0, 1, \ldots, n-1, \tag{2.21}$$

where $j_k \in \{1, 2, \ldots, p\}$ and $f^{(j_k)}(\cdot)$ denotes the j_kth branch of the pth root function. Since X is circulant, it is determined by its first column $x = F_n^{-1}\sigma$ with $\sigma = [\sigma_0, \ldots, \sigma_{n-1}]^T$. With a little algebraic manipulation, we have the elements of $x = [x_0, \ldots, x_{n-1}]^T$ given by

$$x_\ell = \frac{1}{n}\left(1 + \sum_{k=1}^{n-1} \omega^{-\ell k} f^{(j_k)}\left(\sum_{j=0}^{n-1} \omega^{kj} a_j\right)\right), \quad \ell = 0, 1, \ldots, n-1.$$

Therefore, if there exists a choice of the set $\{j_1, j_2, \ldots, j_{n-1}\}$, $j_k \in \{1, 2, \ldots, p\}$ such that $x_\ell \geq 0$ for all $\ell = 0, 1, \ldots, n-1$ and $\sum_{\ell=0}^{n-1} x_\ell = 1$, then A has a stochastic pth root. We make some further discussion on the choices of j_k. First, to have unit row sums in X, we should take pth root of 1 to be 1, namely $\sigma_0 = 1$. Note that the eigenvalues d_k (2.20) of A satisfy $d_k = \overline{d_{n-k}}$, $k = 1, 2, \ldots, n-1$, so for X to be a real matrix, j_k should be chosen such that $f^{(j_k)}(d_k) = \overline{f^{(j_{n-k})}(d_{n-k})}$, namely $\sigma_k = \overline{\sigma_{n-k}}$, $k = 1, 2, \ldots, n-1$. Then X is stochastic if and only if X is nonnegative.

2.6.6 Upper triangular matrices

Triangular matrices arise in Markov models of progressive diseases [32], where the health state of a patient can never improve. Consider a transition matrix for the progression of a progressive disease with five health states ordered from least to most

severe

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ 0 & a_{22} & a_{23} & a_{24} & a_{25} \\ 0 & 0 & a_{33} & a_{34} & a_{35} \\ 0 & 0 & 0 & a_{44} & a_{45} \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \tag{2.22}$$

Since any primary root of an upper triangular matrix is still upper triangular, in order to have a nonnegative primary root of A, one needs to choose the nonnegative branch of roots of the diagonal. It is clear that the only possible stochastic primary root of A is the principal root. However, it is not just the diagonal elements that determine whether there exits a primary stochastic root, as can be shown in the following example. The matrices

$$A = \begin{bmatrix} 0.4276 & 0.0843 & 0.4269 & 0.0148 & 0.0464 \\ 0 & 0.0075 & 0.3689 & 0.3942 & 0.2294 \\ 0 & 0 & 0.3691 & 0.3382 & 0.2927 \\ 0 & 0 & 0 & 0.3618 & 0.6382 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

and

$$B = \begin{bmatrix} 0.4276 & 0.0319 & 0.1945 & 0.0836 & 0.2620 \\ 0 & 0.0075 & 0.2947 & 0.2955 & 0.4023 \\ 0 & 0 & 0.3691 & 0.4655 & 0.1654 \\ 0 & 0 & 0 & 0.3618 & 0.6382 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

are stochastic matrices with the same diagonal. It can be verified that A has a stochastic principal square root while B does not. Since the diagonal elements are distinct, all the roots of A and B are primary. The situation is nevertheless more complicated when the matrix has nonprimary roots. As shown in Fact 2.16, a triangular stochastic matrix may have more than one stochastic nonprimary pth root.

2.6.7 Irreducible imprimitive stochastic matrices

The content in this section is from an unpublished note from Steve Kirkland [94]. Recall the background knowledge in Section 1.2. A *primitive* stochastic matrix is an irreducible stochastic matrix that has only one eigenvalue of modulus 1; otherwise it is called *imprimitive* (or *cyclic* [106, Chap. 3, Def.1.1]) and the number of eigenvalues with modulus 1 is called the *index* of A. Let A be an irreducible stochastic matrix with index $k \geq 2$. Then there exists a permutation matrix P such that PAP^T is of the form [106, Chap. 3, Thm. 3.1]

$$\begin{bmatrix} 0 & A_1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & A_2 & \cdots & 0 & 0 \\ \vdots & & \ddots & \ddots & & \vdots \\ 0 & 0 & & \cdots & 0 & A_{k-1} \\ A_1 & 0 & & \cdots & & 0 \end{bmatrix}, \tag{2.23}$$

where the zeros blocks along the main diagonal are square (not necessarily with the same size). To be exact, for each $i = 1:k$, we take A_i to be $m_i \times m_{i+1}$ with $m_{k+1} \equiv m_1$. If we partition the index set $\{1, \ldots, n\}$ as $S_1 \cup \cdots \cup S_k$ with $S_j = \left\{\sum_{\ell=1}^{j-1} m_\ell + 1, \ldots, \sum_{\ell=1}^{j} m_\ell\right\}$, $j = 1:k$ and let $S_{k+1} \equiv S_1$, then A has a nonzero entry in the (i,j) position only if there is some index ℓ such that $i \in S_\ell$, $j \in S_{\ell+1}$.

Markov chains with transition matrices of the form (2.23) possess the property that the minimum number of transitions that must be made on leaving any state to return that state, is a multiple of k. These models are called periodic Markov chains of period k [123]. Periodic Markov chains arise in a range of applications such as computer communication networks [21], [50, Chap. 6], economic fluctuations and business-cycle analysis [51].

The aim of this section is to investigate conditions on the existence of stochastic pth roots for irreducible imprimitive stochastic matrices. Without loss of generality we assume that stochastic matrix A is of the form (2.23) with $k \geq 2$. Assuming X is a stochastic pth root of A, we have the following facts and observations (we omit their proofs from here).

1. X is irreducible and periodic with period k and the eigenvalues of X of modulus 1 are $e^{2\pi jp/k}$, $j = 0:k-1$.

2. $\gcd(p,k) = 1$.

3. We can partition the index set $\{1, \ldots, n\}$ as $T_1 \cup \cdots \cup T_k$ such that, for some permutation σ of $\{1, \ldots, k\}$, for any indices i, j such that if X has a positive entry in the (i,j) position then necessarily there is an index ℓ such that $i \in T_{\sigma(\ell)}$, $j \in T_{\sigma(\ell+1)}$. We conclude that in fact the sets $S_1, \ldots, S_k$ and $T_1, \ldots, T_k$ yield the same partitioning of $1, \ldots, n$, i.e., the partitioning of X afforded by $T_1, \ldots, T_k$ coincides with the partitioning of A in (2.23). Moreover, together with the fact that $\gcd(p,k) = 1$, it follows that the partitioned form for X is given by

$$X = \begin{bmatrix} 0 & 0 & \cdots & 0 & X_{t+1} & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & X_{t+2} & \cdots & 0 \\ \vdots & \vdots & \cdots & 0 & 0 & \cdots & \ddots & \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & X_k \\ X_1 & 0 & \cdots & 0 & 0 & \cdots & \cdots & 0 \\ 0 & X_2 & \cdots & 0 & 0 & \cdots & \cdots & 0 \\ \vdots & & \ddots & & \vdots & & & \vdots \\ 0 & 0 & \cdots & X_t & 0 & \cdots & \cdots & 0 \end{bmatrix} \tag{2.24}$$

   Here, for each $j = 1, \ldots, k$, the submatrix X_j lies in the columns corresponding to the indices in S_j.

4. It follows from (2.24) that X^p can be written as

$$\begin{bmatrix} 0 & X_{t+1}X_{2t+1}\cdots X_{pt+1} & 0 & \cdots & 0 \\ 0 & 0 & X_{t+2}X_{2t+2}\cdots X_{pt+2} & \cdots & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & X_{t+k-1}X_{2t+k-1}\cdots X_{pt+k-1} \\ X_{t+k}X_{2t+k}\cdots X_{pt+k} & 0 & 0 & \cdots & 0 \end{bmatrix}, \tag{2.25}$$

where the subscripts on the X_j are to be taken modulo k. Hence, finding a stochastic pth root of A is equivalent to finding matrices $X_1, \ldots, X_k$ that are nonnegative with row sums 1 such that for each $j = 1 : k$, $A_j = X_{t+j} X_{2t+j} \cdots X_{pt+j}$.

5. Since $\gcd(p, k) = 1$, there is a unique pair of smallest positive integers t and s such that $tp - sk = 1$. We assume further that A_i, $i = 1 : k$ is square and invertible. Then there exists a pth root M of $(A_k^{-1} A_{k-1}^{-1} \cdots A_1^{-1})^s$ and hence

$$\begin{aligned} X_1 &= A_{k-t+1} \cdots A_k M, && (2.26)\\ X_j &= A_{k-t+j} \cdots A_k M A_1 \cdots A_{j-1},\ j = 2 : t, && (2.27)\\ X_{t+1} &= M A_1 \cdots A_t, && (2.28)\\ X_{t+j+1} &= A_j^{-1} \cdots A_1^{-1} M A_1 \cdots A_{t+j},\ j = 2 : k - t - 1. && (2.29) \end{aligned}$$

Based on the observations above, we now summarize the main results in the following theorem.

Theorem 2.32 ([94])**.** *Let A be an irreducible stochastic matrix that is imprimitive with index k, invertible, and given by* (2.23)*. Then A has a stochastic pth root X if and only if both of the following conditions holds:*

(a) $\gcd(p, k) = 1$*;*

(b) *there is a pth root M of $(A_k^{-1} A_{k-1}^{-1} \ldots A_1^{-1})^s$ such that the following inequalities hold*

$$\begin{aligned} A_{k-t+1} \cdots A_k M &\geq 0, && (2.30)\\ A_{k-t+j} \cdots A_k M A_1 \cdots A_{j-1} &\geq 0,\ j = 2 : t, && (2.31)\\ M A_1 \cdots A_t &\geq 0, && (2.32)\\ A_j^{-1} \cdots A_1^{-1} M A_1 \cdots A_{t+j} &\geq 0,\ j = 2 : k - t - 1, && (2.33) \end{aligned}$$

where t is defined by the condition that t and s is a pair of smallest positive integers satisfying $tp - sk = 1$.

In the event that conditions (a) *and* (b) *hold, then the matrix X given by* (2.24) *is a stochastic pth root of A, where the blocks $X_1, \ldots, X_k$ are given by* (2.26)–(2.29)*.*

2.6.8 Symmetric positive semidefinite matrices: An extension of Marcus and Minc's theorem

The content in this section from Ilse Ipsen [82], is an extension of Marcus and Minc's result (Theorem 2.13) on the existence of stochastic square root of stochastic Hermitian positive semidefinite matrix.

Denote by A^* the conjugate transpose of a matrix A.

Proposition 2.33. *If A is Hermitian positive semidefinite and $Av = \lambda v$, $\|v\|_2 = 1$, then*

$$a_{ii} \geq \lambda |v_i|^2, \quad i = 1, 2, \ldots, n.$$

Proof. Since A is Hermitian positive semidefinite, it has a Hermitian square root B, that is, $A = B^2$ and $B^* = B$. From B Hermitian follows

$$a_{ii} = e_i^* A e_i = e_i^* B^2 e_i = e_i^* B^* B e_i = \|Be_i\|_2^2.$$

It also implies $v^*B = \sqrt{\lambda}v^*$, so that $\|Be_i\|_2 \geq |v^*Be_i| = |\sqrt{\lambda}||v_i|$, where the first inequality is from the Cauchy-Schwarz inequality. □

Directly from $\sum_{i=1}^n a_{ii} = \text{trace}(A) = \sum_{i=1}^n \lambda_i \geq 1$, we have that at least one of the diagonal elements of a stochastic and symmetric positive semidefinite matrix should satisfy $a_{ii} \geq 1/n$. The following corollary shows that this inequality holds for all i.

Corollary 2.34. *If the $n \times n$ matrix A is stochastic and symmetric positive semidefinite, then $a_{ii} \geq 1/n$.*

Proof. Apply Proposition 2.33 with $\lambda = 1$ and $v = e/\sqrt{n}$. □

Corollary 2.34 tells us that, the diagonal elements of a stochastic symmetric positive semidefinite matrix can not be too small.

Theorem 2.35. *Let the $n \times n$ matrix A be nonnegative and symmetric positive semidefinite, with a maximal eigenvalue λ and maximal eigenvector v, i.e., $Av = \lambda v$, $\lambda \geq 0$, $v \geq 0$, $\|v\|_2 = 1$. If the diagonal elements of A satisfy*

$$a_{ii} \leq \frac{\lambda v_i^2}{1 - v_i^2}, \quad 1 \leq i \leq n \tag{2.34}$$

then A has a nonnegative square root $A^{1/2}$.

Proof. Let B be a symmetric positive semidefinite square root of A, i.e., $A = B^2$ and $B = B^T$. Then $v^T B = \sqrt{\lambda}v^T$. As in the proof of Proposition 2.33, $a_{ii} = \|Be_i\|_2^2$. Now suppose B is not nonnegative, so that $b_{\ell k} = b_{k\ell} < 0$ for some ℓ and k. Assume without loss of generality that $v_k \leq v_\ell < 1$. Let

$$w \equiv \begin{bmatrix} v_1 \\ \vdots \\ v_{\ell-1} \\ v_{\ell+1} \\ \vdots \\ v_n \end{bmatrix}, \quad c \equiv \begin{bmatrix} b_{1k} \\ \vdots \\ b_{\ell-1,k} \\ b_{\ell+1,k} \\ \vdots \\ b_{nk} \end{bmatrix}.$$

Then from $\sqrt{\lambda}v_k = b_{\ell k}v_\ell + w^T c$, we have $w^T c \geq \sqrt{\lambda}v_k \geq 0$. Hence

$$\|Be_k\|_2 > \|c\|_2 \geq |w^T c|/\|w\|_2 \geq \sqrt{\lambda}v_k/\sqrt{1 - v_\ell^2} \geq \sqrt{\lambda}v_k/\sqrt{1 - v_k^2},$$

where the second inequality is the Cauchy-Schwarz inequality. Therefore

$$a_{kk} = \|Be_k\|_2^2 > \lambda v_k^2/(1 - v_k^2),$$

which contradicts the upper bound on the diagonal elements. □

Remark 2.36. *Apply* Theorem 2.35 *with* $\lambda = 1$ *and* $v = e/\sqrt{n}$ *and then we can get Theorem* 2.13 *on the existence of a stochastic square root of a symmetric positive semidefinite stochastic matrix.*

2.7 Embeddability problem

The stochastic root problem is closely related to the embeddability problem in discrete-time Markov chains. Consider a time-homogeneous discrete-time Markov chain with a finite number n of states. The single-step transition probability matrix $P = (p_{ij})$ with

$$p_{ij} = \text{Prob}\{X_{k+1} = j | X_k = i\}, \quad i, j = 1, 2, \ldots, n,$$

is independent of k. The *embeddability* problem, first proposed by Elfving [43], is to determine whether there exists an *intensity matrix* Q such that $\exp(Q) = P$. Here the intensity matrix Q is a square matrix with $q_{ij} \geq 0$ for $i \neq j$ and $\sum_{j=1}^{n} q_{ij} = 0$, $i = 1\colon n$. The embeddability problem is indeed to determine whether the given process is a discrete manifestation of an underlying time-homogeneous continuous-time n-state Markov process. If there exists such a Q (which is called a *generator*), P is said to be *embeddable*, in which case the transition matrix $P(t)$ for arbitrary time periods is obtained $P(t) = \exp(Qt)$. For any intensity matrix Q, $\exp(Qt)$ is nonnegative for all $t \geq 0$ (see [72, Thm. 10.30]) and has unit row sums, so is stochastic. The following theorem by Kingman [93] fully describes the relation between the matrix root problem and the embeddability problem.

Proposition 2.37 ([93, Prop. 7])**.** *Let* P *be an* $n \times n$ *nonsingular stochastic matrix. If for each positive integer* m *there exists a stochastic matrix* Q_m *such that*

$$P = Q_m^m,$$

then there exists a generator for P*.*

This means the problem of embedding the chain in a continuous time process is equivalent to the problem of embedding it in a discrete time chain in which the unit of time is an arbitrary submultiple of that in the original chain. Iwanik and Shiflett [86] provide a slightly more general assertion than Proposition 2.37 when they analyze the existence of roots of stochastic operators on L^1-spaces: if a stochastic (doubly stochastic) matrix has stochastic (doubly stochastic) roots of all orders, then it is embeddable in a continuous one-parameter semigroup of stochastic (doubly stochastic) matrices. Here, the doubly stochastic matrix is a square nonnegative matrix with unit row and column sums.

According to Kingman [93], the embeddability problem is completely solved for 2×2 matrices case by Dendall: a 2×2 stochastic matrix P is embeddable if and only if $\det(P) > 0$. The sufficient and necessary conditions for embeddability of 3×3 matrices with distinct eigenvalues or positive multiple eigenvalues are given by Johansen in 1974 [87]. The case of 3×3 matrices with a negative eigenvalue of multiplicity 2 is solved by Carette in 1995 [25]. Johansen and Ramsey [88] and Frydman [48] give a necessary and sufficient condition for embeddability of a 3×3 stochastic matrix with at least one off-diagonal element equal to zero. By analyzing the geometry of the set of all embeddable matrices, Kingman [93] claims that no

simple necessary and sufficient conditions like the 2×2 matrices case can be found when the dimension is greater than 2. For more results on the structure of the set of all embeddable stochastic matrices, one can refer to [49] where the author shows that such a set is a Lipschitz manifold with boundary. In this section, we summarize some results on the general case of the embeddability problem.

2.7.1 Conditions for embeddability and uniqueness

The first theorem is a collection of some necessary conditions for the existence of a generator. Recall that a state j is accessible from state i if there is a sequence of states $k_0 = i, k_1, k_2, \ldots, k_m = j$ such that $a_{k_\ell k_{\ell+1}} > 0$ for each ℓ. We denote the $(i,\ j)$ entry of the matrix power P^m by $p_{ij}^{(m)}$.

Theorem 2.38. *Let P be an $n \times n$ transition matrix, and suppose that there is a generator Q for P. Then*

(a) (Kingman 1962 [93]) $\det(P) > 0$*;*

(b) (Goodman 1970 [55]) $\det(P) \leq \prod_i p_{ii}$*;*

(c) (Elfving 1937 [43]) *no eigenvalue of P other than 1 can satisfy $|\lambda| = 1$ and any negative eigenvalue must have even (algebraic) multiplicity;*

(d) (Chung 1967 [29], Grimmett and Stirzaker 1992 [58]) *for every pair of states i and j such that j is accessible from i, $p_{ij} > 0$;*

(e) (Chung 1967 [29]) *whenever $p_{ij} = 0$, then $p_{ij}^{(m)} = 0$, $m = 2, 3, \ldots$;*

(f) (Runnenberg 1962 [115]) *all eigenvalues of P must lie inside a heart-shaped region H_n in the complex plane whose boundary is the curve $x(v) + iy(v)$, where $0 \leq v \leq \pi / \sin(2\pi/n)$ and*

$$x(v) = \exp(-v + v\cos\frac{2\pi}{n})\cos(v\sin\frac{2\pi}{n}),$$

$$y(v) = \exp(-v + v\cos\frac{2\pi}{n})\sin(v\sin\frac{2\pi}{n}),$$

together with its symmetric image with respect to the real axis;

(g) (Singer and Spilerman 1976 [119], Israel et al. 2001 [85]) *if P has distinct eigenvalues, then each eigenvalue λ of Q satisfies $|\lambda| \leq |\log(\det(P))|$;*

(h) (Fuglede 1988 [49]) *there exist distinct indices i, j such that for all k*

$$p_{ik} = 0 \quad \textit{implies} \quad p_{jk} = 0,$$

and likewise distinct indices i', j' such that, for all k,

$$p_{ki'} = 0 \quad \textit{implies} \quad p_{kj'} = 0;$$

(i) (Israel et al. 2001 [85]) *the entries of P must satisfy*

$$p_{ik} \geq m^m r^r (m+r)^{-m-r} \sum_j (p_{ij} - b_m)(p_{jk} - b_r) \mathbf{1}_{p_{ij}>b_m,\ p_{jk}>b_r},$$

for any positive integers m and r. Here $b_m = \sum_{\ell=m+1}^{\infty} e^{-\sigma}\sigma^\ell/\ell!$ is the probability that $N' > m$, where N' is a Poisson random variable with mean $\sigma \equiv \max_i(-q_{ii})$. Furthermore $\mathbf{1}_B$ is the indicator function of the Boolean event B.

Some comments on these conditions are made in order. Condition (a) can be obtained as follows

$$\det(A) = \det(e^Q) = \exp(\text{trace}(Q)) > 0,$$

where Q is the generator of A. The second equality is from the fact that the eigenvalues of $f(A)$ are $f(\lambda_i)$, where the λ_i are the eigenvalues of A (Theorem 1.3 (d)); see also [72, Theorem 1.45]. If A is symmetric positive semidefinite matrix, the condition (b) is Hadamard's inequality: suppose $A = B^2$ where B is symmetric positive semidefinite matrix having b_i, $i = 1, \ldots, n$ as columns; then by Hadamard's inequality $\det(A) = \det(B)^2 \leq \prod_{i=1}^n \|b_i\|_2^2 = \prod_{i=1}^n a_{ii}$. Conditions (a) and (b) are the first known simple necessary conditions for embeddability of a stochastic matrix. Johansen and Ramsey [88] and Frydman [48] prove that (a) and (b) are also sufficient conditions for embeddability of a 3×3 stochastic matrix with at least one off-diagonal element equal to zero. It can be verified that the stochastic matrices satisfying conditions (a) and (b) form a closed subsemigroup of the semigroup of all stochastic matrices with positive determinant [49]. Condition (d) follows from the standard Lévy Dichotomy and (i) is a more quantitative version of (d). Condition (e) is given by Ornstein's theorem. The regions H_3, H_6, H_8 and H_{12} in Runnenberg's necessary condition (f) are visualized in Figure 2.4.

The following result identifies some cases in which there is a unique generator for a given transition matrix. Here, $\log P$ denotes the principal logarithm of P [72, Thm. 1.31], which is the unique logarithm whose spectrum lies in the strip $\{z : -\pi < \text{Im}(z) < \pi\}$.

Theorem 2.39. *Let P be a transition matrix.*

(a) (Israel et al. 2001 [85]) *If $\det(P) > 1/2$, then P has at most one generator.*

(b) (Israel et al. 2001 [85]) *If $\det(P) > 1/2$ and $\|P - I\| < 1/2$ (using any matrix norm), then the only possible generator for P is $\log P$.*

(c) (Cuthbert 1972 [34], Cuthbert 1973 [35]) *If P has distinct eigenvalues and $\det(P) > e^{-\pi}$, then the only possible generator for P is $\log P$.*

(d) (Singer and Spilerman 1976 [119]) *If P has real, positive, distinct eigenvalues, then the only real matrix Q such that $\exp(Q) = P$ is $\log P$.*

2.7.2 Relation to the stochastic *p*th root problem

Given a stochastic matrix A, Proposition 2.37 shows that the condition for the existence of a generator of A holds if and only if for every positive integer p there

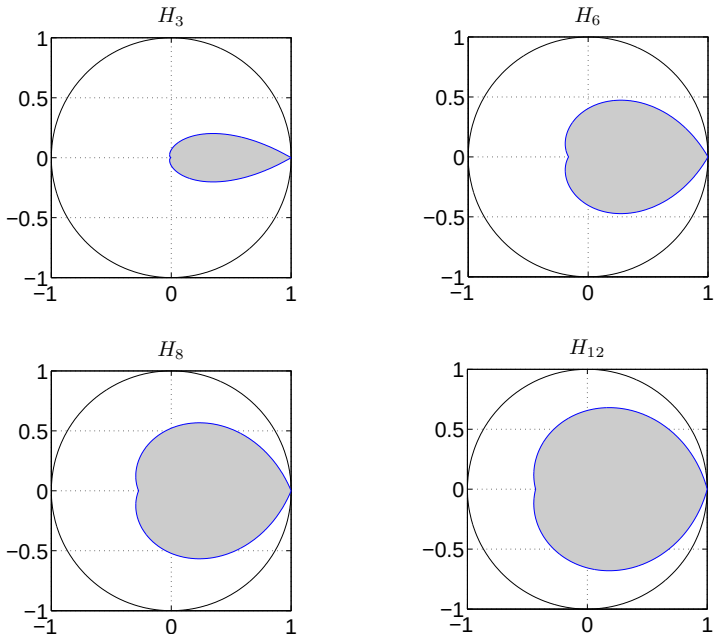

Figure 2.4: Region of Runnernberg's necessary condition for embeddability: H_3, H_6, H_8 and H_{12}.

exists some stochastic X_p such that $A = X_p^p$. (Thus the matrices identified in Theorems 2.11 and 2.12 form two classes of embeddable matrices.) The condition that A is embeddable is much stronger than the condition that A has a stochastic pth root for a particular p. This is emphasized by the following facts, which show that certain necessary conditions derived in the literature for A to be embeddable are not necessary for A to have a stochastic pth root for certain p. Moreover, A may of course be singular in the stochastic root problem, in which case it cannot be the exponential of any matrix.

Fact 2.40. $\det(A) > 0$ *is necessary for the embeddability of a stochastic matrix A; it is also necessary for the existence of a stochastic pth root when p is even, but it is not necessary when p is odd.*

The matrix

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

has $\det(A) = -1$, but A is its own stochastic pth root for any odd p.

Fact 2.41. $\det(A) \le \prod_i a_{ii}$ *is necessary for the embeddability of A* [55, Thm. 6.1]*, but it is not necessary for the existence of a stochastic pth root.* For example, let A be the matrix X in (2.9). Then $A^3 = I$, so $(A^2)^2 = A$ and A has a stochastic square root, but $\det(A) = 1 > 0 = a_{11}a_{22}a_{33}$.

Fact 2.42. *If there is a sequence $k_0 = i, k_1, k_2, \dots, k_m = j$ such that $a_{k_\ell k_{\ell+1}} > 0$ for each ℓ but $a_{ij} = 0$ then A is not embeddable* [58, sec. 6.10]*, but it is still possible for A to have a stochastic pth root for some p.* See the matrix A in (2.9), for which $a_{12} > 0$ and $a_{23} > 0$, while $a_{13} = 0$.

2.8 Further discussion and conclusions

In both the embeddability problem and the stochastic root problem it is difficult to identify conditions that guarantee the existence of a logarithm or root of the required form. For some further insight, consider a nonsingular upper triangular stochastic matrix T, taking $n = 3$ for simplicity. The equation $U^2 = T$ can be solved for U (assumed upper triangular) a diagonal at a time by a recurrence of Björck and Hammarling [17], [72, sec. 6.2]. This gives $u_{ii} = t_{ii}^{1/2}$, $i = 1\colon 3$ (since we require U nonnegative), $u_{i,i+1} = t_{i,i+1}/(u_{ii} + u_{i+1,i+1})$, $i = 1\colon 2$, and $u_{13} = (t_{13} - u_{12}u_{23})/(t_{11}^{1/2} + t_{33}^{1/2})$. Hence $u_{13} \geq 0$ when

$$t_{13} - \frac{t_{12}t_{23}}{(t_{11}^{1/2} + t_{22}^{1/2})(t_{22}^{1/2} + t_{33}^{1/2})} \geq 0.$$

If we assume that T is diagonally dominant, which implies $t_{ii} \geq 1/2$, $i = 1, 2$, and note that $t_{33} = 1$, we obtain the sufficient condition for nonnegativity that $(1 + 2^{1/2})t_{13} \geq t_{12}t_{23}$. But diagonal dominance alone is not sufficient to ensure nonnegativity. Thus even for diagonally dominant triangular matrices the stochasticity of the principal square root depends in a complicated way on the relationships between the matrix entries.

We can also consider general strictly diagonally dominant stochastic matrices, for which $a_{ii} > 1/2$ for all i. Let $m = \min_i a_{ii}$ and write $A = mI + E$. Then $E \geq 0$ and $Ee = (A - mI)e = (1 - m)e$, so $\|E\|_\infty = 1 - m$. Hence we can write $A = m(I + F)$, where $\|F\|_\infty = \|E\|_\infty/m = (1 - m)/m < 1$. Then the principal pth root can be expressed as

$$A^{1/p} = m^{1/p}(I + F)^{1/p} = m^{1/p}\big(I + \tfrac{1}{p}F + \tfrac{1}{2!}\tfrac{1}{p}(\tfrac{1}{p} - 1)F^2 + \cdots\big).$$

Unfortunately, it is difficult to obtain from this expansion useful sufficient conditions for $A^{1/p} \geq 0$. Nonnegativity is guaranteed if all the off-diagonal elements of F are positive and $\|F\|_\infty$ is sufficiently small, but as the matrix (2.8) shows, "small" here may have to be very small.

The existing literature on roots of stochastic matrices emphasizes computational aspects at the expense of a careful treatment of the underlying theory. We have used the theory of matrix functions to develop tools for analyzing the existence of stochastic roots of stochastic matrices. We have identified two classes of stochastic matrices for which the principal pth root is stochastic for all p. However, such matrices seem rare, and we have demonstrated a wide variety of possibilities for existence and uniqueness, in particular regarding primary versus nonprimary roots. We have also given some necessary spectral conditions for existence. We hope that as well as providing insight into what makes this interesting and practically important problem so difficult our work will prove useful for further development of theory and algorithms.

Chapter 3

Computing Short-interval Transition Matrices

3.1 Overview

As described in Chapter 2, the applications of finding a short term transition matrix require a stochastic root of a given stochastic matrix A. The focus therein is on the underlying theory of the stochastic roots problem. In this chapter we investigate numerical methods for computing approximate stochastic roots. We begin with surveying some techniques in statistics that are currently used to estimate the transition matrix for a required time period or, more generally, the transition rate matrix (also known as the generator of a Markov model in Section 2.7) based on a set of observation data.

3.1.1 Statistics techniques

The problem of estimating the transition rate or transition probability matrix of a Markov model is intensively investigated in statistics for a wide range of applications, such as computational physics [33], credit risk in the finance industry [7], [89], [97], and medical decision making in healthcare [18], [19], [22], [26], [131]. Different statistical techniques are intended for different models used and different kinds of data available: continuous-time Markov process versus discrete-time Markov chain; fully observed data versus partially observed data. For more about the underlying models in this problem, see [11], [105], [121] where practical guides on Markov models in medical decision making are given and [114] for its use in credit risk. Throughout this section, we only consider the time-homogeneous discrete-time Markov chains and continuous-time Markov models.

One of the advantages of the continuous-time Markov models is that they allow meaningful estimation of the probability of *rare* transitions, for example, a transition from a high rating category, say AAA in Moody's credit risk rating, to default [19], [97]. In a discrete-time model, if a single transition from AAA to default does not occur over a given time period, then the estimate of the corresponding probability is zero. However, if there are transitions from AAA to AA and from AA to default (possibly by other firms) then the estimator for transitions from AAA to default should not be zero because there is chance of defaulting within a certain time period (after

successive downgrades). A continuous-time model captures this transition probability whereas a discrete-time model does not. Another advantage of continuous-time Markov models is that the matrix of transition probabilities for any time period t can easily be obtained by $P = \exp(tQ)$, where Q is the transition rates matrix of the underlying Markov model. In a continuous-time Markov model, if a full record of all transitions is available, that is, observations are made continuously such that the exact time at which a transition takes place is known, then an explicit formula for the maximum likelihood estimator (MLE) of the transition rates is obtained [18]; see [97] for more details on this method and a comparison with estimators based on a discrete-time model. Welton and Ades [131] propose a Bayesian framework for estimating transition rates with fully observed data. However, it is more often that the observations are made at discrete time points. Bladt and Sørensen [18], [19] demonstrate that a continuous-time Markov model can also be used to analyse observations at discrete time points (which is referred as partially observed data in the continuous-time Markov model), where the expectation maximization (EM) algorithm and an EM approach employing a Markov Chain Monte Carlo (MCMC) technique are investigated to estimate the transition rate matrix. An MCMC approach within a Bayesian framework for estimations from partially observed data is also studied in [131]. Hence the advantages of a continuous-time model can be obtained without continuous-time data. See [103] for details on the EM algorithm, [31] for Bayesian modelling and [52] for the MCMC approach.

There is a distinction between discrete-time Markov chains and continuous-time Markov models. For the discrete-time Markov chains, we exploit transition probabilities directly instead of considering transition rates. Recall that a transition probability matrix describes probabilities of one step transition among different states where the step-size is known as the *cycle length* inherent to the Markov chain. In disease modeling, the cycle length is often set to an interval associated with medical follow-ups [105]. If the individuals are observed at an interval equal to the cycle length, the MLE of the transition probability matrix is easily obtained by a closed form [32]. Difficulties in estimation are being noted when the observation interval and the cycle length do not coincide (which is referred as partially observed data in discrete-time Markov model). For example, a cycle length of six month is desired while the observations are made at one-year intervals. A more complicated case arises when the observation intervals are not equal in length. Craig and Sendi [32] and Borg et al. [22] propose use of the EM algorithm to cope with these situations.

Some comments on the statistics techniques are in order:

1. An advantage of methods under the Bayesian framework is that information from multiple sources can be statistically combined into the currently used model.

2. The EM algorithm for estimating short-interval transition matrices with partially observed data in discrete-time Markov model works only when the interval of interest is a proper divisor of the observation interval.

3. All these methods require the acquisition of the *transition counts* (the number of transitions observed from one state to another). In many applications, nevertheless, the only available data is a transition matrix that is readily obtained

from the literature or from expert institutions, for example, Moody's Investor Service, Standard & Poor's rating agencies for credit risk and the Swiss HIV Cohort Study database for the study of AIDS.

4. Given a transition matrix, methods from the theory of matrices to get a valid short-interval transition matrix are also mentioned in [26], [32] but without further study. These methods compute a fractional root of the transition matrix by employing an eigendecomposition. We mention in passing that in both papers they wrongly take the nonnegativity of eigenvalues of the original transition matrix as a necessary and sufficient condition to get a valid short-interval transition matrix. However, as seen in Chapter 2, this condition is neither necessary nor sufficient for the existence of stochastic roots of a stochastic matrix.

3.1.2 Optimization techniques

As mentioned before, in many applications the transition matrix is readily obtained from the literature or from expert institutions. In this case, the problem of computing short-interval transition matrix reduces to computing a stochastic root of a stochastic matrix. Regarding the problem of computing matrix roots, various methods are available [14], [59], [60], [68], [72, Chap. 7], [79], [120], but there are currently no methods tailored to finding a stochastic root.

Current approaches are designed to find an appropriate *approximate* stochastic root. An immediate idea is to compute *some* pth root and perturb it to be stochastic [26], [85], [95]. By choosing the principal root of A, this idea can be formalized as

$$\min \|X - A^{1/p}\| \quad \text{subject to} \quad X \text{ a stochastic matrix.} \tag{3.1}$$

This is termed as *quasi-optimization of the root matrix* (QOM) in [85]. A very similar idea is to find the nearest intensity matrix G to $\log(A)$ and then an approximate stochastic root can be formed by $X = \exp(G/p)$. This is to solve the following *quasi-optimization of the generator* (QOG)

$$\min \|G - \log(A)\| \quad \text{subject to} \quad G \text{ an intensity matrix.} \tag{3.2}$$

A stochastic matrix X that minimizes $\|X - A^{1/p}\|$ may not minimize the residual $\|X^p - A\|$. This can be easily illustrated by an example where the principal pth root of A is not stochastic but there exists a stochastic matrix X that satisfies $X^p = A$, either primary or nonprimary (see Fact 2.20 and Fact 2.23). Similarly, an intensity matrix G that minimizes $\|G - \log(A)\|$ may not result in the matrix $X = \exp(G/p)$ that minimizes $\|X^p - A\|$. Relations between errors $\|X - A^{1/p}\|$, $\|G - \log(A)\|$ and the residual $\|X^p - A\|$ can be found in the following theorems.

Theorem 3.1. *Assume that $A \in \mathbb{C}^{n\times n}$ has no eigenvalues on $\mathbb{R}^-$, the closed negative real axis. If $\|X - A^{1/p}\| = \epsilon\|A^{1/p}\|$ then*

$$\|X^p - A\| \le \|A^{1/p}\|^p((1+\epsilon)^p - 1). \tag{3.3}$$

Proof. Let $B = A^{1/p}$ and $E = X - B$. Then we have

$$X^p \;=\; (B+E)^p = B^p + (B^{p-1}E + B^{p-2}EB + \cdots + EB^{p-1}) + \cdots + E^p.$$

It follows that

$$\begin{aligned}\|X^p - B^p\| &\leq p\|B\|^{p-1}\|E\| + \frac{p(p-1)}{2}\|B\|^{p-2}\|E\|^2 + \cdots + \|E\|^p \\ &= (\|B\|^p + p\|B\|^{p-1}\|E\| + \cdots + \|E\|^p) - \|B\|^p \\ &= \|B\|^p((1+\epsilon)^p - 1),\end{aligned}$$

since $\epsilon = \|E\|/\|B\|$. This completes the proof. □

Theorem 3.1 says that if the distance between X and $A^{1/p}$ is small then so is the distance between X^p and A. A similar result can be found for the matrix exponential. The following theorem is from [36] where $\|\cdot\|_\infty$ is used; the proofs there are nevertheless valid for any consistent matrix norm.

Theorem 3.2. *Assume that $A \in \mathbb{C}^{n\times n}$ has no eigenvalues on $\mathbb{R}^-$. If $\|G - \log(A)\| = \epsilon$ then*

$$\|A - e^G\| \leq \min\{2, e^\epsilon - 1\}. \tag{3.4}$$

Proof. See Davies [36, Thm. 13]. □

The minimal residual $\|X^p - A\|$ is defined by the following nonlinear programming

$$\min \|X^p - A\| \quad \text{subject to} \quad X \text{ a stochastic matrix.} \tag{3.5}$$

Due to the difficulty of solving the nonlinear programming (3.5) with n^2 variables, He and Gunn [64] propose an alternative to (3.5). Since for any positive integer k, A^k can be expressed in terms of $\{I, A, A^2, \ldots, A^{n-1}\}$ (by the Cayley-Hamilton theorem), any primary pth root X of A (and hence a polynomial of A) can be written as $X = h_0 I + h_1 A + \cdots h_{n-1}A^{n-1}$. So if we restrict the stochastic approximation to be a primary function of A, then problem (3.5) reduces to the following nonlinear programming with n variables

$$\begin{aligned}&\min \left\| \left(\sum_{i=0}^{n-1} h_i A^i \right)^p - A \right\| \\ &\text{subject to} \quad \sum_{i=0}^{n-1} h_i A^i \text{ a stochastic matrix.}\end{aligned} \tag{3.6}$$

A final idea is mentioned in [85] but has few numerical experiments in the literature. This is to modify the original stochastic matrix A first (either to make it an embeddable matrix or to make it admit a stochastic root) and then search for an *exact* generator or stochastic root. The aim of this chapter is to study the properties of the optimization problems described above and investigate numerical methods to solve them. In Section 3.2 we identify problems of interest where we state the available algorithms for finding the nearest stochastic matrix in (3.1) and the nearest intensity matrix in (3.2) with certain norms; we derive explicit formulae for the gradient and Hessian of the objective function in (3.5) and (3.6) with the Frobenius norm; we consider an active set method, an interior point method, a spectral projected gradient method (SPGM) and the sequential quadratic programming (SQP) method for both optimization problems. In Section 3.3 we give numerical experiments to compare

the performance of the methods. We also investigate different matrices to start the iteration. Finally, some conclusions are given in Section 3.4.

3.2 Problems of interest: properties and numerical methods

To have a differentiable objective function, we use the Frobenius norm $\|\cdot\|_F$ throughout this section.

3.2.1 The nearest stochastic matrix to $A^{1/p}$.

The problem of interest is

$$\text{minimize } f(X) = \|X - A^{1/p}\|_F^2 \tag{3.7a}$$

$$\text{subject to } X \in \Omega := \Big\{ X \in \mathbb{R}^{n\times n} : \sum_{j=1}^{n} x_{ij} = 1,\ i = 1\colon n, \quad x_{ij} \geq 0,\ i,j = 1\colon n \Big\}. \tag{3.7b}$$

Since both the objective function and the set Ω are convex, there is a global minimum to problem (3.7). This can essentially be found on a row-by-row basis by reducing it to n independent *distance minimization problems*

$$\min \|x - a\|_2 \quad \text{subject to} \quad x \in \mathbb{R}^n,\ x_i \geq 0,\ \sum_{i=1}^{n} x_i = 1, \tag{3.8}$$

where $a \in \mathbb{R}^n$ is a row vector of the matrix $A^{1/p}$. In the case where $A^{1/p}$ has nonreal numbers, let a be the real part of each row of $A^{1/p}$. An algorithm for solving distance minimization problem (3.8) is suggested by Merkoulovitch [104] and a corresponding iterative algorithm is provided in [95]. Now we state the algorithm.

Algorithm 3.3 (distance minimization algorithm)**.** *Given $a \in \mathbb{R}^n$ this algorithm computes a nonnegative vector x with $\|x\|_1 = 1$ that minimizes the distance $\|x - a\|_2$.*

```
1  if sum_{i=1}^n a_i = 1 & a >= 0, x = a, quit, end
2  while true
3        lambda = (sum_{i=1}^n a_i - 1)/n, x = a - lambda
4        if x >= 0, quit, end
5        for i = 1:n
6             x_i = max{0, x_i}
7        end
8        a = x
9  end
```

Note that the iterative algorithm stops after j steps where j does not exceed the size of the vector a [104]. The cost of Algorithm 3.3 is $O(n^2)$, so the cost of finding the nearest stochastic matrix in problem (3.7) is $O(n^3)$.

3.2.2 The nearest intensity matrix to log(A).

The problem of interest is

$$\text{minimize } f(X) = \|G - \log(A)\|_F^2 \tag{3.9a}$$

$$\text{subject to } G \in \Omega := \Big\{ G \in \mathbb{R}^{n\times n} : \sum_{j=1}^{n} g_{ij} = 0, i = 1\!:\! n, \quad g_{ij} \geq 0,\ i \neq j,\ i,j = 1\!:\! n \Big\}. \tag{3.9b}$$

Again since the objective function (3.9a) and the set Ω are convex, there is a global minimizer. In a similar manner as in problem (3.7), we solve (3.9) on a row-by-row basis. Define a standard cone in $\mathbb{R}^n$ as

$$K(n) = \Big\{ x \in \mathbb{R}^n :\ \sum_{i=1}^{n} x_{ij} = 0,\ x_1 \leq 0,\ x_i \geq 0,\ i = 2:n \Big\}. \tag{3.10}$$

By permuting each row vector of an intensity matrix, we can always represent it as a point in $K(n)$. Problem (3.9) can be reduced to n independent problems of projecting a point $a \in \mathbb{R}^n$ (each permuted row of the matrix $\log(A)$) onto the cone $K(n)$, i.e.,

$$\min \|g - a\|_2 \quad \text{subject to} \quad g \in K(n). \tag{3.11}$$

Kreinin and Sidelnikova [95] propose the following algorithm for solving (3.11). We mention that ℓ^* in line 3 of Algorithm 3.4 should be chosen among $1:n-1$ other than $2:n-1$ as stated in [95].

Algorithm 3.4 (distance minimization algorithm for the generator)**.** *Given $a \in \mathbb{R}^n$, this algorithm computes $g \in K(n)$ that minimizes the distance $\|g - a\|_2$.*

1 $\lambda = \sum_{i=1}^{n} a_i/n$, $a = a - \lambda$
2 $b = \sigma(a)$, σ is a permutation sorting a in descending order
3 find $\ell^* = \min_{1\leq\ell\leq n-1}\big\{\ell\colon\ b_{\ell+1} \geq (b_1 + \sum_{i=\ell+1}^{n} b_i)/(n-\ell+1)\big\}$
4 for $i = 2\!:\!\ell^*$, $g_i = 0$, end
5 for $i = 1, \ell^*+1\!:\!n$
6 $\quad g_i = b_i - (b_1 + \sum_{j=\ell^*+1}^{n} b_j)/(n-\ell^*+1)$
7 end
8 $g = \sigma^{-1}(g)$, where σ^{-1} is the inverse permutation of σ

Note that ℓ^* will be found within n steps of searching. The cost of Algorithm 3.4 is $O(n^2)$, so the cost of finding the nearest intensity matrix is $O(n^3)$.

3.2.3 Minimize the residual $\|X^p - A\|_F$

Now we consider the nonlinear programming problem

$$\text{minimize } f(X) = \|X^p - A\|_F^2 \tag{3.12a}$$

$$\text{subject to } X \in \Omega := \Big\{ X \in \mathbb{R}^{n\times n} : \sum_{j=1}^{n} x_{ij} = 1,\ i = 1\!:\! n, \quad x_{ij} \geq 0,\ i,j = 1\!:\! n \Big\}. \tag{3.12b}$$

The set Ω is convex; however, the objective function (3.12a) is nonconvex for $p > 1$. We can only expect to determine a local minimizer. We first derive the gradient of $f(X)$, i.e., $\nabla f(X) = (\partial f(X)/\partial x_{ij}) \in R^{n\times n}$.

Lemma 3.5. *For $f(X)$ in* (3.12a) *we have*

$$\nabla f(X) = 2\sum_{j=1}^{p}(X^T)^{j-1}(X^p - A)(X^T)^{p-j}. \tag{3.13}$$

Proof. For arbitrary $E \in \mathbb{R}^{n\times n}$ we have

$$\begin{aligned} f(X+E) &= \|(X+E)^p - A\|_F^2 \\ &= \mathrm{trace}(((X+E)^p - A)^T((X+E)^p - A)) \\ &= \mathrm{trace}((X^p - A)^T(X^p - A)) \\ &\quad + 2\mathrm{trace}\left(\sum_{j=1}^{p}(X^T)^{j-1}(X^p - A)(X^T)^{p-j}E^T\right) \\ &\quad + O(\|E\|_F^2). \end{aligned}$$

Then the expression of (3.13) follows using the definition of $\nabla f(X)$. □

Note that the Hessian H of f is an $n^2 \times n^2$ matrix that can be viewed as the representation of the Fréchet derivative $L_{\nabla f}$ of ∇f, that is, for any $E \in \mathbb{R}^{n\times n}$

$$\mathrm{vec}(L_{\nabla f}(X,E)) = H\mathrm{vec}(E). \tag{3.14}$$

Lemma 3.6. *For $f(X)$ in* (3.12a) *we have*

$$\begin{aligned} L_{\nabla f}(X,E) &= 2\sum_{j=1}^{p}\Bigg((X^T)^{j-1}(X^p - A)\sum_{l=1}^{p-j}(X^T)^{p-j-l}E^T(X^T)^{l-1} \\ &\qquad + (X^T)^{j-1}\sum_{k=1}^{p}X^{p-k}EX^{k-1}(X^T)^{p-j} \\ &\qquad + \sum_{i=1}^{j-1}(X^T)^{j-1-i}E^T(X^T)^{i-1}(X^p - A)(X^T)^{p-j}\Bigg). \end{aligned}$$

Proof. With the expression of $\nabla f(X)$ in (3.13), for arbitrary $E \in \mathbb{R}^{n\times n}$, we have

$$\begin{aligned} \nabla f(X+E) &= 2\sum_{j=1}^{p}(X^T + E^T)^{j-1}((X+E)^p - A)(X^T + E^T)^{p-j} \\ &= 2\sum_{j=1}^{p}\left((X^T)^{j-1}\sum_{i=1}^{j-1}(X^T)^{j-1-i}E^T(X^T)^{i-1}\right) \\ &\qquad \cdot\left(X^p + \sum_{k=1}^{p}X^{p-k}EX^{k-1} - A\right) \end{aligned}$$

$$\cdot \left((X^T)^{p-j} + \sum_{l=1}^{p-j} (X^T)^{p-j-l} E^T (X^T)^{l-1} \right) + O(\|E\|_F^2).$$

$L_{\nabla f}(X, E)$ is obtained immediately by finding the linear part in E of the expansion above. □

We consider several different numerical methods for this nonlinear optimization problem. Since the gradient and the Hessian are explicitly known, Newton's method can be used on problem (3.12). The function `fmincon` of the MATLAB Optimization Toolbox allows users to choose algorithms among an active set method, an interior point method and a sequential quadratic programming (SQP) method. We can also apply the routine `e04uc` of the NAG Toolbox for MATLAB [3], which implements an SQP method.

Recall that Algorithm 3.3 allows us to find the nearest stochastic matrix to a given matrix. This motivates us to use a spectral projected gradient method (SPGM) introduced by Birgin, Martínez, and Raydan [15, 16]. The method aims to minimize a continuously differentiable function f on a closed convex set in $\mathbb{R}^n$ by generating a sequence of vectors that is guaranteed to converge r-linearly to a stationary point of f. It generates vectors of the form $x_{k+1} = x_k + \alpha_k d_k$ with the spectral projected gradient direction $d_k = P(x_k - \lambda_k \nabla f(x_k)) - x_k$, where $\lambda_k > 0$ is some precomputed scalar, and with α_k being chosen by a nonmonotone line search strategy. The direction d_k is guaranteed to be descent direction [15, Lem. 2.1]. The method explicitly takes advantage of the possible simplicity of projections P onto the feasible set, which applies to our problem.

3.2.4 Minimize $\|X^p - A\|_F$ over all primary functions of A

As mentioned above (see (3.6)), we can solve the following problem to get a stochastic matrix which is a primary function of A and minimizes the residual

$$\text{minimize} \quad f(h) = \left\| \left(\sum_{i=0}^{n-1} h_i A^i \right)^p - A \right\|^2 \tag{3.15a}$$

$$\text{subject to } h \in \Omega := \big\{ h \in \mathbb{R}^n : e^T h = 1,\ Bh \geq 0, \tag{3.15b}$$
$$B = [\operatorname{vec}(I)\ \operatorname{vec}(A)\ \ldots\ \operatorname{vec}(A^{n-1})] \big\}.$$

Let $X(h) = \sum_{i=0}^{n-1} h_i A^i$. The constraint $Bh \geq 0$ in (3.15b) is to guarantee a nonnegative matrix $X(h)$ and $e^T h = 1$ is to ensure that $X(h)$ has unit row sums. The gradient of $f(h)$ is given in the following lemma.

Lemma 3.7. *For $f(h)$ in (3.15a), we have*

$$\nabla f(h) = 2 \left(\operatorname{vec}\Big(\sum_{j=1}^{p} (X(h)^T)^{j-1} (X(h)^p - A)(X(h)^T)^{p-j} \Big) \right)^T B. \tag{3.16}$$

Proof. Applying the chain rule, the result follows directly from Lemma 3.5 and the fact that $\dfrac{d\operatorname{vec}(X(h))}{dh} = B$. □

We consider the possibility of applying SPGM on the problem (3.15). The first ingredient required is the projection onto the set Ω in (3.15b). Note that Ω is a convex polyhedron, which is the intersection of a finite number of closed halfspaces. The problem of projecting a vector onto a convex polyhedron arises in many applications such as machine learning, pattern recognition [108], [118] and image restoration [100]. Nurminski [108] provides an efficient and stable algorithm to compute the projection, with an complexity of $O(mn^2)$, where n is the number of variables and m is the number of inequalities. In our case, however, the number of inequalities, i.e., the row number of B in (3.15b), is n^2, which results in a complexity of $O(n^4)$ for computing the projection onto the feasible region Ω in (3.15b). This prevents us from using SPGM except for very small n. Therefore, we will only apply the active set method, the interior point method and the SQP method on problem (3.15).

3.3 Numerical tests

Our experiments were performed in MATLAB R2010a using the NAG Toolbox for MATLAB Mark 22.0 on an Intel Dual-Core CPU (1.73GHz).

We first consider problem (3.12), which is to minimize $\|X^p - A\|_F^2$ over all stochastic matrices X. To encourage a fair comparison with all numerical methods, we use the same stopping criterion introduced in [16], [23] in all the algorithms employed in this section. The stopping criterion is

$$\|q(X)\|_F \leq \text{tol}, \tag{3.17}$$

where $q : \mathbb{R}^{n\times k} \mapsto \mathbb{R}^{n\times k}$ is defined by

$$q(X) = P(X - \nabla f(X)) - X.$$

Here, P is a projection onto the feasible set and f is the objective function. It can be shown that a point $X^* \in \Omega$ is a stationary point of problems (3.12) if and only if $q(X^*) = 0$ [42, (2.5)-(2.7)].

Now we consider several options to start the iteration for nonlinear programming (3.12). Recall that A is the given stochastic matrix.

- Ident: the $n \times n$ identity matrix I.
- StoRand: this matrix is a random matrix with elements from the uniform distribution on $[0, 1]$ which is then scaled to a stochastic matrix by dividing each element by its corresponding row sum.
- PrincRoot: this matrix is obtained by computing the principal pth root of A and getting the nearest stochastic matrix of $A^{1/p}$ (if it is not stochastic) by Algorithm 3.3. PrincRoot is the solution of problem (3.7). $A^{1/p}$ here is computed by a Schur algorithm [120].
- GenFro: this is to compute $\log(A)$ first, get the solution G of problem (3.9) by Algorithm 3.4 and construct GenFro by $\exp(G/p)$. During the computation, we use the inverse scaling and squaring method for the logarithm [72, sec. 11.5] and the scaling and squaring method for the exponential [5].

- GenInf: for this choice of starting point we compute the principal logarithm of A, $L = \log(A)$ and then adjust L as, for $i, j = 1 : n$,

$$\hat{\ell}_{ij} = \begin{cases} 0 & \ell_{ij} < 0 \text{ and } i \neq j, \\ \ell_{ij} & \text{otherwise.} \end{cases} \tag{3.18}$$

  The diagonal elements of $\widehat{L}$ are set $\hat{\ell}_{ii} = -\sum_{\substack{j=1 \\ j\neq i}}^{n} \ell_{ij}$ for $i = 1 : n$ so as to get an intensity matrix. GenInf is then $\exp(\widehat{L}/p)$. This is suggested by Stromquist [124] as an alternative method to get a generator. It is also discussed in [95] where it is called *diagonal adjustment*. Davis [36] proves that $\widehat{L}$ is actually the nearest intensity matrix to L where the distance is measured in the infinity norm, in contrast to the Frobenius norm in problem (3.9).

- GenWA: this is another way to get a near generator $\widehat{L}$ and then take $\exp(\widehat{L}/p)$ as a starting point. As for GenInf, we compute $L = \log(A)$ first and then adjust negative elements of L as in (3.18). In order to have all zero row sums, we further adjust all nonzero elements by the following *weighted adjustment* [95, 124]

$$\hat{\ell}_{ij} = \hat{\ell}_{ij} - |\hat{\ell}_{ij}| \sum_{j=1}^{n} \hat{\ell}_{ij} \Bigg/ \sum_{j=1}^{n} |\hat{\ell}_{ij}|\,, \quad \text{for} \quad i, j = 1 : n. \tag{3.19}$$

- UTri: this is an upper triangular matrix obtained by simply setting the diagonal with the real pth root of the corresponding diagonal element of A and then adjusting the last element of each row to get the unit row sums

$$X_0 = \begin{bmatrix} a_{11}^{1/p} & 0 & \cdots & 0 & 1 - a_{11}^{1/p} \\ 0 & a_{22}^{1/p} & \cdots & 0 & 1 - a_{22}^{1/p} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & a_{n-1,n-1}^{1/p} & 1 - a_{n-1,n-1}^{1/p} \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}. \tag{3.20}$$

  This starting point is motivated by the fact that in some applications the given stochastic matrix A is diagonally dominant and Utri is a rough approximation to a pth root of A.

- FullRow: this is another approximation of a pth root of diagonally dominant matrix A. It is a full matrix obtained by setting the diagonal elements in the same way as for UTri and then equally setting the off-diagonal elements for each row so as to get the unit row sums

$$X_0 = \begin{bmatrix} a_{11}^{1/p} & \dfrac{1 - a_{11}^{1/p}}{n-1} & \cdots & \dfrac{1 - a_{11}^{1/p}}{n-1} \\ \dfrac{1 - a_{22}^{1/p}}{n-1} & a_{22}^{1/p} & \cdots & \dfrac{1 - a_{22}^{1/p}}{n-1} \\ \cdots & \cdots & \ddots & \cdots \\ \dfrac{1 - a_{nn}^{1/p}}{n-1} & \dfrac{1 - a_{nn}^{1/p}}{n-1} & \cdots & a_{nn}^{1/p} \end{bmatrix}. \tag{3.21}$$

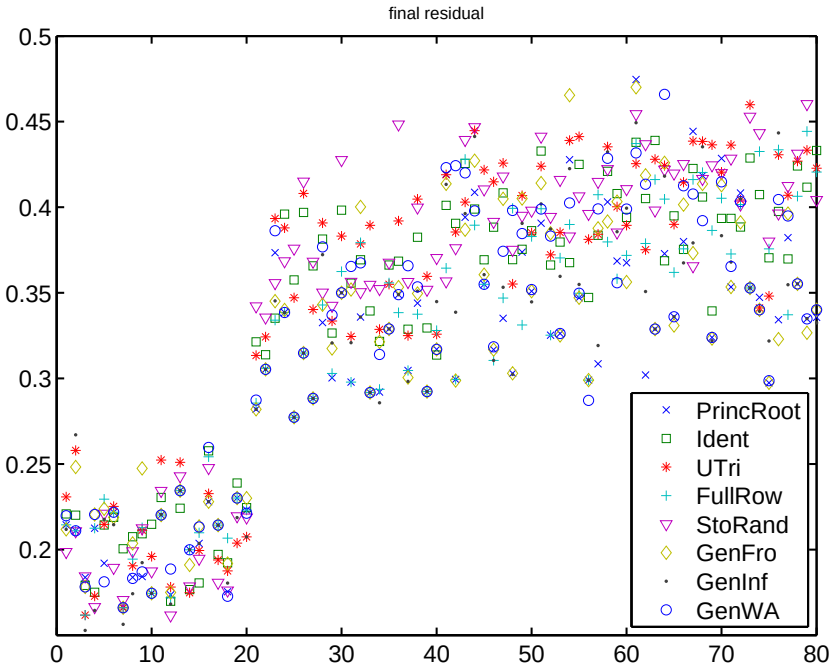

Figure 3.1: Final residual of each starting point.

We test these choices of starting matrix with the SQP method which is the most reliable method to solve the problem though it is expensive in computation. We used random matrices with elements from the uniform distribution on $[0, 1]$, which were then adjusted to stochastic matrices by dividing each element by its row sum. We test for $p = 2, 5, 7, 9$ with 20 instances of 12×12 random matrices for each p. Figure 3.1–3.3 reports the final residual, number of iterations and the computational time for each problem. To facilitate comparing the performance of different starting points, we show the performance profiles on these measures as well as the initial residual for each choice of starting point. A performance profile shows the proportion π of problems where the performance ratio of a method is at most α, where the performance ratio for a method on a problem is the measure, the error or residual say, of that method divided by the smallest value of the measure over all the methods (if we favor a method with a smaller value of that measure). For more on performance profiles, see [41] and [65, sec. 22.4]. Figure 3.4 shows the performance profiles for the starting points Ident, StoRand, GenFro and FullRow and Figure 3.5 shows that for PrincRoot, GenInf, GenWA, GenFro and FullRow. We omitted the performance profiles for UTri because it is the worst starting point under all measures we are using here. It is clear from Figure 3.4 that GenFro and FullRow outperform Ident and StoRand while from Figure 3.5 that PrincRoot outperforms GenFro, GenInf, GenWA and FullRow. PrincRoot has the best performance overall.

We do the remaining numerical experiments using the following sets of test matrices.

Set 1 Random 12×12 matrices with elements from the uniform distribution on $[0, 1]$ which are then scaled to a stochastic matrix by dividing each element by its corresponding row sum.

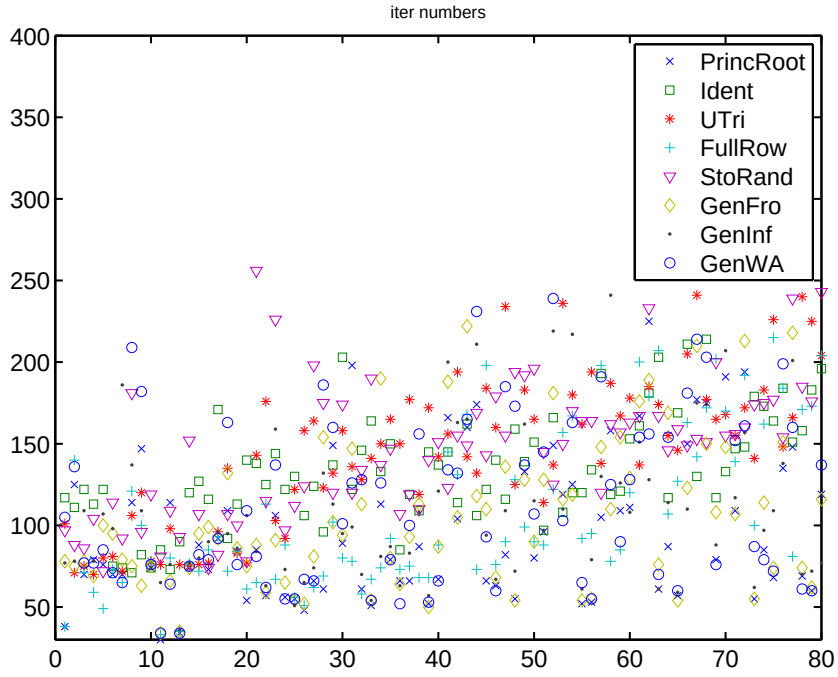

Figure 3.2: The number of iterations with each starting point.

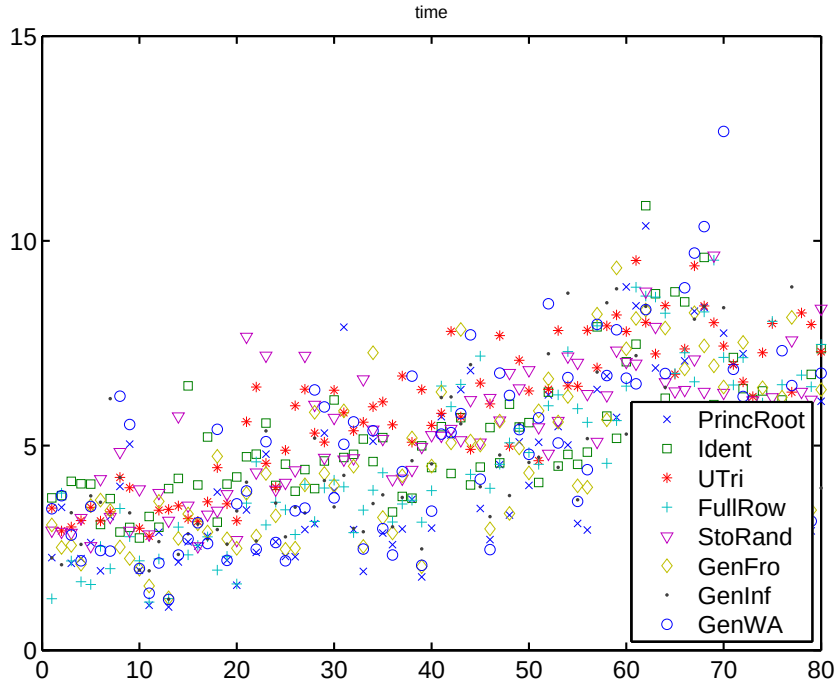

Figure 3.3: Computational time for each starting point.

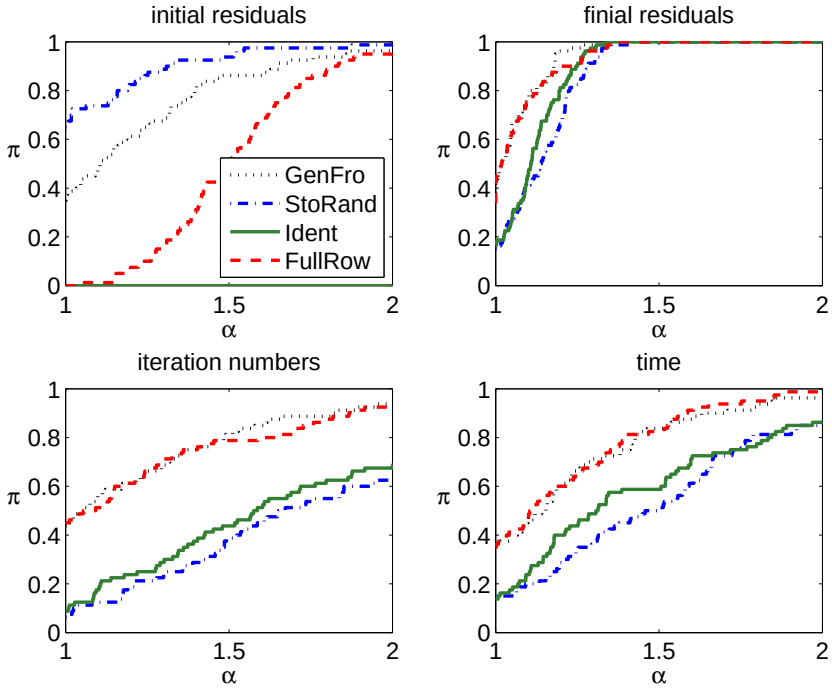

Figure 3.4: Performance profiles for Ident, StoRand, GenFro and FullRow. The legend for the first plot applies to all four plots.

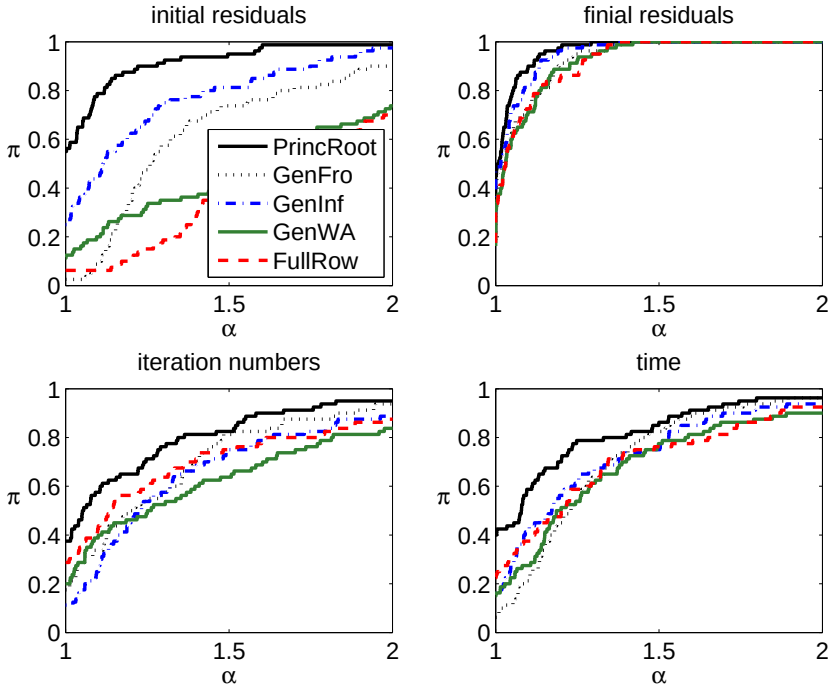

Figure 3.5: Performance profiles for PrincRoot, GenFro, GenInf, GenWA and FullRow. The legend for the first plot applies to all four plots.

Set 2 $A = X^p$ where X is a stochastic matrix generated in the same way as matrices in Set 1. Here the objective function f for (3.12) is zero at the global minimum.

Set 3 $A = \exp(Q)$ where Q is an intensity matrix obtained by generating a random 12×12 matrix with elements from the uniform distribution on $[0, 1]$ and then adjusting the diagonal elements such that each row sum is zero. In this case, the objective functions f in (3.12) and (3.15) are both zero at the global minimums for any p.

Set 4 Matrices from the literature on developing methods for roots of stochastic matrices. All are of dimension 10 or less, most of them arising from finance and healthcare applications.

Set 5 A 21×21 one year refined-rating transition matrix for year 2004 published in February 2005 by Moody's Global Structured Finance [1].

We computed with $p = \{2, 3, 4, 12\}$ and for each p we generated 10 matrices from Set 1–3. For problem (3.12), we tested with the active set method, the interior point method, the SQP method and SPGM using the stopping criterion $\|q(X)\|_F < \text{tol}$ with $\text{tol} = 10^{-3}$. We started the iteration with "PrincRoot". For problem (3.15), since it is expensive to compute the projection onto the feasible region, we use the default stopping criteria for each method from the software with the function tolerance 10^{-15} and constraints tolerance 10^{-8}. We report results averaged over 40 problems in Tables 3.1–3.3. Table 3.4 reports results with test matrices from Set 4. Table 3.5 shows results for the test matrix in Set 5 for each value of p where we omitted results for the interior point method and the active set method due to their poor performance in both accuracy and computational time.

The abbreviations for the methods and results reported are

- act-set: `fmincon` from MATLAB with option `'active-set'`.
- int-pt: `fmincon` from MATLAB with option `'interior-point'`.
- SQP: `e04uc`, NAG implementation of SQP method.
- SPGM: spectral projected gradient method.
- t: (mean) computational time (in seconds).
- it: (mean) number of iterations.
- it_{sd}: standard deviation of numbers of iterations.
- ires: (mean) initial residual $\|X^p - A\|_F$.
- res: (mean) final residual $\|X^p - A\|_F$.
- inq: (mean) initial value of $\|q(X)\|_F$.
- nq: (mean) final value of $\|q(X)\|_F$.

Several comments are made for results in Table 3.1–3.5.

Table 3.1: Results for matrices from Set 1.

Set 1	t	it	it_{sd}	res	nq
Prob. (3.7)	1.44e-2	–	–	6.64e-1	1.26
Prob. (3.9)	3.05e-2	–	–	8.96e-1	1.64
Prob. (3.12), ires = 6.64e-1, inq = 1.26 , tol = 1.00e-3					
act-set	5.27	4.82e1	2.75e1	3.29e-1	1.51e-3
int-pt	1.87e2	8.74e3	1.63e4	4.12e-1	7.51e-2
SQP	4.24e-1	5.02e1	2.53e1	2.85e-1	1.34e-3
SPGM	1.24e-1	1.63e2	1.36e2	2.82e-1	8.87e-4
Prob. (3.15), ires = 3.36 , inq = 4.01 , tol = 1.00e-3					
act-set	3.36e-1	7.32e1	6.01e1	5.26e-1	1.47e-1
int-pt	4.01	5.35e2	2.80e3	4.68e-1	1.95e-1
SQP	6.11e-2	1.51e2	3.58e2	5.23e-1	1.52e-1

1. The interior point method is not efficient in both the accuracy and the computational time for all test matrices and problems considered. SQP is more efficient than the active set method for both problems (3.12) and (3.15). For problem (3.12), SPGM is clearly the best method. Table 3.5 shows that an increased problem size ($n = 21$ for Set 5 and $n = 12$ for Set 1–4) gives a bigger time advantage of SPGM over SQP.

2. From Table 3.1, 3.3 and 3.5 we see that for each method, the computational time for solving problem (3.15) is less than that for problem (3.12). This is not surprising because there are n variables for the former problem and n^2 for the latter one. However, the same observation is not found in Table 3.2 for test matrices from Set 2. We point out that for each matrix from Set 2, there exists a stochastic root whereas the principal stochastic root is not stochastic. In this case, our experiments show that searching for an approximate primary root is less efficient than searching directly for a nearest stochastic root regardless of it being a primary function of the given stochastic matrix or not.

3. Matrices in Set 3 are all embeddable (see Section 2.7). A stochastic root is obtained by computing the principal root. Iterations starting with the principal root will stop after one iteration. Therefore the results for solving the problem (3.12) with starting matrix PrincRoot are omitted from Table 3.3. Though there exists a global minimum for problem (3.15), only a local minimum can be found here.

4. For the transition matrix in Set 5, we are unable to verify whether there exists a stochastic root for each p (though the matrix satisfies the necessary conditions for the existence of stochastic roots derived in Section 2.5.2). All the optimization techniques for solving problem (3.12) did not significantly reduce the residual from the starting point (which is the solution of problem (3.7)).

Table 3.2: Results for matrices from Set 2.

Set 2	t	it	it_{sd}	res	nq
Prob. (3.7)	1.38e-2	–	–	3.99e-2	7.89e-2
Prob. (3.9)	7.94e-3	–	–	3.83e-1	7.31e-1
Prob. (3.12), ires = 3.99e-2, inq = 7.89e-2, tol = 1.00e-3					
act-set	4.90e-1	9.55	1.15e1	3.00e-3	1.91e-3
int-pt	2.11	8.85e1	1.49e2	5.86e-3	6.26e-4
SQP	1.24e-1	8.88	1.15e1	3.21e-3	2.39e-3
SPGM	2.62e-2	1.02e1	1.34e1	3.15e-3	5.98e-4
Prob. (3.15), ires = 3.32 , inq = 3.79 , tol = 1.00e-3					
act-set	9.97e-2	2.22e1	2.71e1	1.60e-2	6.12e-3
int-pt	3.86e1	5.03e3	1.34e4	1.36e-2	4.59e-3
SQP	2.70e-1	9.27e2	5.07e3	1.41e-2	5.76e-3

Table 3.3: Results for matrices from Set 3.

Set 3	t	it	it_{sd}	res	nq
Prob. (3.7)	1.47e-2	–	–	1.11e-15	3.89e-15
Prob. (3.9)	6.46e-3	–	–	9.78e-16	2.24e-15
Prob. (3.15), ires = 3.31 , inq = 3.89 , tol = 1.00e-3					
act-set	1.09e-1	2.19e1	3.22e1	1.65e-2	3.50e-3
int-pt	5.85	7.31e2	1.52e3	8.45e-3	1.77e-3
SQP	2.58e-2	2.59e1	4.11e1	1.27e-2	3.51e-3

Table 3.4: Results for matrices from Set 4.

Set 4	t	it	it_{sd}	res	nq
Prob. (3.7)	5.92e-3	–	–	4.09e-2	3.49e-2
Prob. (3.9)	4.31e-3	–	–	4.85e-2	6.31e-2
Prob. (3.12), ires = 4.09e-2, inq = 3.49e-2, tol = 1.00e-3					
act-set	7.09e-1	1.38e1	3.89e1	6.35e-2	4.52e-4
int-pt	3.90e1	3.72e3	1.13e4	7.62e-1	9.34e-2
SQP	8.24e-2	1.69e1	1.86e1	3.92e-2	9.64e-3
SPGM	1.88e-2	1.25e1	5.82e1	3.91e-2	1.87e-4
Prob. (3.15), ires = 9.01e-1, inq = 2.60 , tol = 1.00e-3					
act-set	1.80e-1	4.41e1	2.93e1	1.35e-2	6.19e-2
int-pt	4.12e-1	5.43e1	7.88e1	2.29e-1	3.42e-1
SQP	2.23e-2	3.72e1	2.22e1	1.33e-2	6.11e-2

Table 3.5: Results for the matrix from Moody's in Set 5.

Set 5	t	it	res	nq	t	it	res	nq
	p = 2				p = 3			
Prob. (3.7)	8.66e-3	–	1.66e-3	4.38e-4	1.20e-2	–	2.19e-3	1.13e-3
Prob. (3.9)	1.36e-2	–	3.31e-3	1.76e-3	9.34e-3	–	3.31e-3	2.30e-3
Prob. (3.12)	ires = 1.66e-3, inq = 4.38e-4				ires = 2.19e-3, inq = 1.13e-3			
SQP	5.74	8.00	1.93e-3	2.55e-3	5.05	2.00e1	2.23e-3	2.11e-3
SPGM	1.86e-2	1.00	1.65e-3	3.10e-4	2.39e-2	1.00	2.19e-3	8.79e-4
Prob. (3.15)	ires = 8.72e-1, inq = 3.34				ires = 8.72e-1, inq = 4.24			
SQP	4.23e-2	2.00e1	2.52e-2	8.59e-2	6.98e-2	2.10e1	3.32e-2	1.45e-1
	p = 4				p = 12			
Prob. (3.7)	1.92e-2	–	2.47e-3	1.88e-3	1.23e-1	–	3.03e-3	7.22e-3
Prob. (3.9)	1.01e-2	–	3.31e-3	3.07e-3	8.96e-3	–	3.31e-3	8.16e-3
Prob. (3.12)	ires = 2.47e-3, inq = 1.88e-3				ires = 3.03e-3, inq = 7.22e-3			
SQP	5.71	2.10e1	2.48e-3	2.07e-3	9.56	1.78e2	3.01e-3	1.37e-3
SPGM	3.55e-2	2.00	2.45e-3	2.43e-4	1.81e-1	3.00	3.00e-3	5.07e-4
Prob. (3.15)	ires = 8.72e-1, inq = 4.76				ires = 8.72e-1, inq = 5.39			
SQP	6.62e-2	3.10e1	3.70e-2	1.91e-1	1.75e-1	2.50e1	4.45e-2	5.23e-1

3.4 Concluding remarks

In this chapter, we briefly surveyed some statistical methods for computing the short-interval transition matrices, where the existing literature emphasizes estimations of transition rate matrices. With a set of fully observed data where the exact dates on which transitions occur are known (for the continuous-time Markov process) or the observation intervals coincide with the inherent cycle length (for the discrete-time Markov process), an explicit formula for the maximum likelihood estimator of the transition rate matrix or the transition probability matrix, respectively, is obtained. The methods based on a Bayesian framework are also proposed for estimating the transition rates for fully observed data. More often, one needs to deal with the partially observed data: for a continuous-time Markov model this happens when the observations are made at discrete time points other than continuously; for a discrete-time model this is due to the fact that the observation intervals do not coincide with the cycle length of the model. The expectation maximum (EM) method is usually used in the case of partially observed data. However, the EM method for estimating short-interval transition matrices works only when the interval of interest is a proper divisor of the observation interval. Moreover, all the statistical techniques require the acquisition of the transition counts (number of transitions observed from one state to another over a certain time period).

Our main interest is in the case where the (long-term) transition matrix is readily obtained from the literature or the expert institutions. Here, a fractional root of a transition matrix is needed and thus methods based on the theory of matrices should be used. In the statistics literature or practical papers in financial applications an eigendecomposition is usually employed and then a fractional root is obtained by computing a root of the corresponding diagonal matrix; when an invalid transition matrix (with negative elements or even complex elements) results, it is perturbed to a nearest transition matrix under some measure of distance and then an approximate

short-term transition matrix is obtained. We have considered several methods to find an approximate stochastic root of a stochastic matrix. The first is to compute the principal root of the original matrix, and if it is not stochastic, perturb it to the nearest stochastic matrix in the sense of the Frobenius norm to get an approximate stochastic root. The second is to compute the principal logarithm of the given matrix, perturb it to the nearest intensity matrix (if it is not a valid one) in the sense of the Frobenius norm or the infinity norm and then compute an approximate stochastic root by the matrix exponential. Here the principal matrix root, the principal logarithm and the matrix exponential are computed with the best available methods when they are needed. We also took the perturbed principal root as a starting point and considered various optimization techniques for solving the nonlinear programming problem to minimize the residual $\|X^p - A\|_F$. Our experiments have shown that if the principal stochastic root is not stochastic then adjusting it to the nearest stochastic matrix gives a good choice of matrix to start the iteration. Despite the fact that all the optimization methods considered can only find a local minimum, the spectral projected gradient method is the most efficient method in terms of the computation time and final residual. A variant problem of finding an approximate stochastic root that is a primary function of A was also considered, where $\|X^p - A\|_F$ is minimized subject to X being stochastic and a primary function of A. The numerical experiments have shown that, though it reduces the number of variables from n^2 to n (n is the dimension of A), narrowing the feasible region to the set of the primary functions of A does not result in a significant reduction in cost while on the other hand it may result in a larger final residual compared with that from the optimization over all stochastic matrices regardless of them being primary functions of A or not. Our conclusion is that, in finding an approximate stochastic root, the spectral projected gradient method starting with the perturbed principal root of A to minimize the residual $\|X^p - A\|_F$ over all stochastic matrices is method of choice.

Chapter 4

A Schur–Padé Algorithm for Fractional Powers of a Matrix

4.1 Introduction

The need to compute fractional powers A^p of a square matrix A arises in a variety of applications, including Markov chain models in finance and healthcare [26], [85], fractional differential equations [81], discrete representations of norms corresponding to finite element discretizations of fractional Sobolev spaces [8], and the computation of geodesic-midpoints in neural networks [46]. Here, p is an arbitrary real number, not necessarily rational. Often, p is the reciprocal of a positive integer q, in which case $X = A^p = A^{1/q}$ is a qth root of A. Various methods are available for the qth root problem, based on the Schur decomposition and appropriate recurrences [57], [120], Newton or inverse Newton iterations [60], [79], Padé iterations [80], [98], or a variety of other techniques [14]; see [72, Chap. 7] and [74] for surveys. However, none of these methods is applicable for arbitrary real p.

Arbitrary matrix powers can be defined via the Cauchy integral [72, Def. 1.11]

$$A^p := \frac{1}{2\pi i} \int_\Gamma z^p (zI - A)^{-1} dz, \tag{4.1}$$

where Γ is a closed contour that encloses the spectrum $\Lambda(A)$. This definition yields many different matrices A^p, as the branch of the function z^p can be chosen independently around each eigenvalue. For practical purposes it is more useful to define A^p uniquely as follows.

Definition 4.1. *Let $A \in \mathbb{C}^{n\times n}$ have no eigenvalues on $\mathbb{R}^-$ except possibly for a semisimple zero eigenvalue, and let $p \in \mathbb{R}$. If A is nonsingular,*

$$A^p = \exp\left(p \log(A)\right), \tag{4.2}$$

where $\log(A)$ is the principal logarithm of A [72, Thm. 1.31]. Otherwise, write the Jordan canonical form of A as $A = Z\mathrm{diag}(J_1, 0)Z^{-1}$, where J_1 contains the Jordan blocks corresponding to the nonzero eigenvalues. Then

$$A^p = Z\mathrm{diag}(J_1^p, 0)Z^{-1}, \tag{4.3}$$

where J_1^p is defined by (4.2).

It follows from the theory of matrix functions that the matrix given by Definition 4.1 is independent of the particular choice of Jordan canonical form. Moreover, if A is real then A^p is real. For $p = 1/q$, with q a positive integer, A^p reduces to the principal qth root of A [72, Thm. 7.2]. For $0 < p < 1$, A^p can also be represented as the real integral [72, pp. 174, 187]

$$A^p = \frac{\sin(p\pi)}{p\pi} A \int_0^\infty (t^{1/p} I + A)^{-1}\, dt. \tag{4.4}$$

The aim of this work is to devise a reliable algorithm for computing A^p for arbitrary $p \in \mathbb{R}$. When A is diagonalizable, so that $A = XDX^{-1}$ for a diagonal $D = \mathrm{diag}(d_i)$ and nonsingular X, we can compute $A^p = XD^pX^{-1} = X\mathrm{diag}(d_i^p)X^{-1}$. Alternatively, for any A we can compute the Schur decomposition $A = QTQ^*$, with Q unitary and T upper triangular, from which $A^p = QT^pQ^*$. The matrix T^p has diagonal elements t_{ii}^p and we can obtain the superdiagonal elements from the Parlett recurrence if the t_{ii} are distinct [72, sec. 4.6], [109]. However, this approach breaks down when A is nonnormal with repeated eigenvalues.

The definition (4.2) suggests another way to compute A^p: to employ existing algorithms for the matrix exponential and the matrix logarithm. However, if we use the inverse scaling and squaring method for $X = \log(A)$ [28], [72, sec. 11.5], [91] followed by the scaling and squaring method for $\exp(pX)$ [5], [71], [73] then we are computing two Padé approximants: one of the logarithm and the other of the exponential. We expect benefits to accrue from employing a *single* Padé approximant, to $(1-x)^p$. In this work we develop an algorithm for computing A^p based on direct Padé approximation of $(1-x)^p$.

The rest of this chapter is organized as follows. We begin, in Section 4.2, by investigating the conditioning of fractional powers. Padé approximation of $(1-x)^p$, and in particular how to bound the error in the approximation at a matrix argument, is the subject of Section 4.3. Evaluation of the matrix Padé approximant is considered in Section 4.4, where we investigate the numerical stability of several possible methods. An algorithm for A^p with $p \in (-1, 1)$ that employs an initial Schur decomposition, matrix square roots, Padé approximation, and squarings, is developed in Section 4.5. In Section 4.6 we explain how to deal with general p not necessarily in the interval $(-1, 1)$ and negative integer p, while in Section 4.7 we extend our algorithm to handle singular matrices with a semisimple zero eigenvalue. Some alternative algorithms are considered in Section 4.8 and all the algorithms are compared in the numerical experiments of Section 4.9. Finally, some concluding remarks are given in Section 4.10.

4.2 Conditioning

We first investigate the sensitivity of A^p to perturbations in A. Recall that the Fréchet derivative of f at A in the direction E, denoted by $L_f(A, E)$, is a linear operator mapping E to $L_f(A, E)$ characterized by $f(A+E) = f(A) + L_f(A, E) + o(\|E\|)$. We

also recall the definition and characterization of condition number

$$\kappa_f(A) := \lim_{\epsilon\to 0} \sup_{\|E\|\le\epsilon\|A\|} \frac{\|f(A+E)-f(A)\|}{\epsilon\|f(A)\|} = \frac{\|L_f(A)\|\,\|A\|}{\|f(A)\|}, \tag{4.5}$$

where

$$\|L_f(X)\| := \max_{Z\neq 0} \frac{\|L_f(X,Z)\|}{\|Z\|}. \tag{4.6}$$

Let vec denote the operator that stacks the columns of a matrix into one long vector and let $\otimes$ denote the Kronecker product. For any f, we have $\mathrm{vec}(L_f(A,E)) = K_f(A)\mathrm{vec}(E)$ for a certain matrix $K_f(A) \in \mathbb{C}^{n^2\times n^2}$ called the Kronecker representation of the Fréchet derivative and, moreover, $\|L_f(A)\|_F = \|K_f(A)\|_2$ [72, (3.20)]. It follows that, in the Frobenius norm,

$$\kappa_f(A) = \frac{\|K_f(A)\|_2\,\|A\|_F}{\|f(A)\|_F}. \tag{4.7}$$

To obtain a formula for $K_{x^p}(A)$ we first apply the chain rule [72, Thm. 3.4] to the expression $A^p = \exp(p\log(A))$, to obtain

$$L_{x^p}(A,E) = p\,L_{\exp}\big(p\log(A), L_{\log}(A,E)\big). \tag{4.8}$$

Then, by applying the vec operator, we find that

$$\mathrm{vec}(L_{x^p}(A,E)) = p\,K_{\exp}(p\log(A))\mathrm{vec}(L_{\log}(A,E)) = p\,K_{\exp}(p\log(A))K_{\log}(A)\mathrm{vec}(E),$$

which implies

$$K_{x^p}(A) = p\,K_{\exp}(p\log(A))\,K_{\log}(A). \tag{4.9}$$

This matrix can be computed explicitly if n is small, or its norm can be estimated based on a few matrix–vector products involving $K_{x^p}(A)$ and its conjugate transpose [72, sec. 3.4].

We now derive some bounds for the condition number $\kappa_{x^p}(A)$ that give insight into its size. First, note that, since $(A+\epsilon I)^p = A^p + p\epsilon A^{p-1} + O(\epsilon^2)$ for sufficiently small ϵ (by a general result on the convergence of a matrix Taylor series [72, Thm. 4.7]), we have $L_{x^p}(A,I) = pA^{p-1}$ and hence $\|L_{x^p}(A)\| \ge |p|\,\|A^{p-1}\|/\|I\|$.

Since [72, (10.15)]

$$L_{\exp}(A,E) = \int_0^1 e^{A(1-s)}Ee^{As}\,ds, \tag{4.10}$$

we have, from (4.8),

$$\begin{aligned}
\|L_{x^p}(A,E)\| &= |p|\left\|\int_0^1 e^{p\log(A)(1-s)}L_{\log}(A,E)e^{p\log(A)s}ds\right\| \\
&\le |p|\,\|L_{\log}(A,E)\|\int_0^1 e^{|p|(1-s)\|\log(A)\|}e^{|p|s\|\log(A)\|}ds \\
&\le |p|e^{|p|\|\log(A)\|}\|L_{\log}(A)\|\;\|E\|,
\end{aligned}$$

and so $\|L_{x^p}(A)\| \le |p| e^{|p|\|\log(A)\|} \|L_{\log}(A)\|$. Thus we have the upper and lower bounds

$$\frac{|p|\,\|A^{p-1}\|}{\|I\|} \le \|L_{x^p}(A)\| \le |p| e^{|p|\,\|\log(A)\|}\ \|L_{\log}(A)\|. \tag{4.11}$$

We also have the following lower bound [72, Thm. 3.14, Cor. 3.16], with $f[\lambda,\mu]$ denoting the first divided difference of $f(x) = x^p$,

$$\|L_{x^p}(A)\| \ge \max_{\lambda,\mu\in\Lambda(A)} |f[\lambda,\mu]| = \max\left(\max_{\lambda\in\Lambda(A)} |p||\lambda^{p-1}|,\ \max_{\substack{\lambda,\mu\in\Lambda(A)\\ \lambda\ne\mu}} \frac{|\lambda^p-\mu^p|}{|\lambda-\mu|} \right), \tag{4.12}$$

which is an equality for the Frobenius norm when A is normal. When A is Hermitian the lower bounds in (4.11) and (4.12) are the same for the 2-norm; we will make use of the lower bound in this case in Section 4.6.

4.3 Padé approximation and error bounds

A $[k/m]$ Padé approximant of $(1-x)^p$ is a rational function $r_{km}(x) = p_{km}(x)/q_{km}(x)$ with $q_{km}(0) = 1$ such that

$$(1-x)^p - r_{km}(x) = O(x^{k+m+1}),$$

where p_{km} and q_{km} are polynomials of degree at most k and m, respectively. If a $[k/m]$ Padé approximant exists then it is unique [9, Thm. 1.1], [10, Thm. 1.4.3], [72, Prob. 4.2]. The aims of this section are to show the existence of Padé approximants of $(1-x)^p$ and to investigate the error in the Padé approximant at a matrix argument $X \in \mathbb{C}^{n\times n}$ with $\|X\| < 1$. Throughout this section the norm is assumed to be a subordinate matrix norm.

The scalar hypergeometric function is

$${}_2F_1(\alpha,\beta,\gamma,x) \equiv 1 + \frac{\alpha\beta}{\gamma}x + \frac{\alpha(\alpha+1)\beta(\beta+1)}{2!\gamma(\gamma+1)}x^2 + \cdots = \sum_{i=0}^{\infty} \frac{(\alpha)_i(\beta)_i}{i!(\gamma)_i}x^i, \tag{4.13}$$

where $\alpha,\beta,\gamma,x \in \mathbb{R}$, γ is not a nonpositive integer, $(a)_0 = 1$, and $(a)_i \equiv a(a+1)\dots(a+i-1)$ for $i \ge 1$. Replacing x in (4.13) with $X \in \mathbb{C}^{n\times n}$ we obtain the matrix hypergeometric function

$${}_2F_1(\alpha,\beta,\gamma,X) \equiv \sum_{i=0}^{\infty} \frac{(\alpha)_i(\beta)_i}{i!(\gamma)_i}X^i. \tag{4.14}$$

Since (4.13) converges if $|x| < 1$ [6, Thm. 2.1.1], the matrix series (4.14) converges if $\rho(X) < 1$ [72, Thm. 4.7], where ρ is the spectral radius. We are interested in the special case where $\alpha = -p,\ \beta = 1,\ \gamma = 1$, and $|x| < 1$:

$${}_2F_1(-p,1,1,x) = 1 - px + \frac{p(p-1)}{2}x^2 + \cdots = (1-x)^p.$$

The following lemma shows the existence of the Padé approximants of $(1-x)^p$ for

all $p \in \mathbb{R}$.

Lemma 4.2. *For $p \in \mathbb{R}$, the $[k/m]$ Padé approximant of $(1-x)^p$ exists for all nonnegative integers k and m.*

Proof. It is shown in [9, p. 65], [10, sec. 2.3] that for any $\alpha, \gamma \in \mathbb{R}$ the $[k/m]$ Padé approximant of the general hypergeometric function ${}_2F_1(\alpha, 1, \gamma, x)$ exists for $k-m+1 \geq 0$ and that the denominator $q_{km}(x)$ is given explicitly by

$$q_{km}(x) = \sum_{i=0}^{m} \frac{(-m)_i(-(\alpha+k))_i}{i!(1-(\gamma+k+m))_i} x^i \tag{4.15}$$

$$= {}_2F_1(-m, -(\alpha+k), 1-(\gamma+k+m), x). \tag{4.16}$$

Thus $[k/m]$ Padé approximants to $(1-x)^p$ exist for all $p \in \mathbb{R}$ for $k \geq m$. From $(1-x)^p = 1/(1-x)^{-p}$, and the duality property that the $[k/m]$ Padé approximant of the reciprocal of a function is the reciprocal of the $[m/k]$ Padé approximant of the function [10, Thm. 1.5.1], it follows that $(1-x)^p$ has a $[k/m]$ Padé approximant for $k \leq m$. ∎

We now state some properties of $q_{km}(x)$. The following result of Kenney and Laub bounds the condition number number of the matrix $q_{km}(X)$.

Lemma 4.3. *Let $q_{km}(x)$ be the denominator polynomial of the $[k/m]$ Padé approximant of ${}_2F_1(\alpha, 1, \gamma, x)$ where $0 < \alpha < \gamma$ and $k-m+1 \geq 0$. The zeros of $q_{km}(x)$ are all simple and lie in the interval $(1, \infty)$. Furthermore, for $X \in \mathbb{C}^{n\times n}$ with $\|X\| < 1$,*

$$\|q_{km}(X)\| \leq q_{km}(-\|X\|), \quad \|q_{km}(X)^{-1}\| \leq q_{km}(\|X\|)^{-1} \tag{4.17}$$

and hence

$$\kappa(q_{km}(X)) \leq \frac{q_{km}(-\|X\|)}{q_{km}(\|X\|)}. \tag{4.18}$$

Proof. See [92, Cor. 1 and Lem. 3], where $X \in \mathbb{R}^{n\times n}$ is assumed; the proofs there are nevertheless valid for complex X. ∎

Corollary 4.4. *Let $q_{km}(x)$ be the denominator polynomial of the $[k/m]$ Padé approximant of $(1-x)^p$ with $-1 < p < 1$ and $k-m \geq 0$. Then the zeros of $q_{km}(x)$ are all simple and lie in the interval $(1, \infty)$ and for $X \in \mathbb{C}^{n\times n}$ with $\|X\| < 1$, the matrix $q_{km}(X)$ satisfies (4.17) and (4.18). In particular, when $-1 < p < 0$ these conclusions hold for $k-m+1 \geq 0$.*

Proof. It is straightforward to show that $(1-x)^p = 1 - px \cdot {}_2F_1(1-p, 1, 2, x)$ and, moreover, that if $k \geq m$ then the $[k/m]$ Padé approximant of $(1-x)^p$ is $p_{km}/\widetilde{q}_{k-1,m} = 1 - px\widetilde{r}_{k-1,m}$, where $\widetilde{r}_{k-1,m} = \widetilde{p}_{k-1,m}/\widetilde{q}_{k-1,m}$ is the $[k-1/m]$ Padé approximant of ${}_2F_1(1-p, 1, 2, x)$.

Since $-1 < p < 1$ we have $0 < 1-p < 2$, and since also $(k-1)-m+1 \geq 0$ the properties of $\widetilde{q}_{k-1,m}(x)$ in Lemma 4.3 all hold. If $-1 < p < 0$, it follows from Lemma 4.3 with $\alpha = -p$ and $\gamma = 1$ that the conclusions hold for $k-m+1 \geq 0$. ∎

Denote by $E\big({}_2F_1(\alpha, 1, \gamma, \cdot), k, m, x\big)$ the error in the $[k/m]$ Padé approximant to ${}_2F_1(\alpha, 1, \gamma, x)$, that is,

$$E\big({}_2F_1(\alpha, 1, \gamma, \cdot), k, m, x\big) = {}_2F_1(\alpha, 1, \gamma, x) - r_{km}(x). \tag{4.19}$$

The following lemma provides a series expansion for this error.

Lemma 4.5. *For $|x| < 1$, $k - m + 1 \geq 0$, and α not a negative integer, the error* (4.19) *can be written*

$$E\big({}_2F_1(\alpha, 1, \gamma, \cdot), k, m, x\big) = \frac{q_{km}(1)}{q_{km}(x)} \sum_{i=k+m+1}^{\infty} \frac{(\alpha)_i (i - (k+m))_m}{(\gamma)_i (i + \alpha - m)_m} x^i. \tag{4.20}$$

Proof. See Kenney and Laub [92, Thm. 5]. The statement of Theorem 5 in [92] requires $0 < \alpha < \gamma$, but in fact only the condition that α is not a negative integer (and hence $(i + \alpha - m)_m$ is nonzero) is needed in the proof. □

We are now in a position to bound the error in Padé approximation of the matrix function $(I - X)^p = {}_2F_1(-p, 1, 1, X)$. The following result, which for $-1 < p < 0$ is a special case of [92, Cor. 4], shows that the error is bounded by the error of the same approximation at the scalar argument $\|X\|$.

Theorem 4.6. *For $k - m \geq 0$, $-1 < p < 1$, and $\|X\| < 1$,*

$$\|E\big((I - X)^p, k, m, X\big)\| \leq |E\big((1 - \|X\|)^p, k, m, \|X\|\big)|. \tag{4.21}$$

In particular, when $-1 < p < 0$, (4.21) *holds for $k - m + 1 \geq 0$.*

Proof. For any matrix X with $\|X\| < 1$, $(I - X)^p = {}_2F_1(-p, 1, 1, X)$ is defined and, by (4.20),

$$E((I - X)^p, k, m, X) = q_{km}(1) q_{km}(X)^{-1} \sum_{i=k+m+1}^{\infty} \frac{(-p)_i (i - (k+m))_m}{i!(i - p - m)_m} X^i, \tag{4.22}$$

where $q_{km}(x)$ is the denominator of the $[k/m]$ Padé approximant to $(1-x)^p$. We claim that every coefficient in the sum has the same sign, that is, the signs are independent of i for $i \geq k+m+1$. Indeed, $(-p)_i < 0$ for $0 < p < 1$ and $(-p)_i > 0$ for $-1 < p < 0$, and clearly $(i - (k+m))_m > 0$ and $(i - p - m)_m > 0$. Therefore, by Corollary 4.4 and the second inequality in (4.17), we have

$$\begin{aligned}
\|E((I - X)^p, k, m, X)\| &\leq \frac{|q_{km}(1)|}{q_{km}(\|X\|)} \sum_{i=k+m+1}^{\infty} \frac{|(-p)_i| (i - (k+m))_m}{i!(i - p - m)_m} \|X\|^i \\
&= \frac{|q_{km}(1)|}{q_{km}(\|X\|)} \left| \sum_{i=k+m+1}^{\infty} \frac{(-p)_i (i - (k+m))_m}{i!(i - p - m)_m} \|X\|^i \right| \\
&= \big|E\big((1 - \|X\|)^p, k, m, \|X\|\big)\big|.
\end{aligned}$$

If $-1 < p < 0$, the result holds for $k - m + 1 \geq 0$, since Corollary 4.4 shows that the required bound $\|q_{km}(X)^{-1}\| \leq q_{km}(\|X\|)^{-1}$ still holds in this case. □

In practice, we would like to select k and m to minimize the error for a given order of approximation. The following result of Kenny and Laub [92, Thm. 6] is useful in this respect.

Theorem 4.7. *Let $k - m + 1 \geq 0$ and $0 < \alpha \leq \gamma$, and let the subordinate matrix norm $\|\cdot\|$ satisfy $\|\widetilde{M}\| \leq \|M\|$ whenever $0 \leq \widetilde{M} \leq M$, where the latter inequalities are interpreted componentwise. Then, if $X \in \mathbb{R}^{n\times n}$ has nonnegative entries,*

$$\|E\big({}_2F_1(\alpha, 1, \gamma, \cdot), k, m, X\big)\| \leq \|E\big({}_2F_1(\alpha, 1, \gamma, \cdot), k+1, m-1, X\big)\|. \tag{4.23}$$

Applying Theorem 4.7 with $\alpha = -p \in (0, 1)$ and $\gamma = 1$, we obtain the corresponding result for $(I - X)^p$, where $-1 < p < 0$. For $0 < p < 1$, the inequality (4.23) holds for k, m satisfying $k - m \geq 0$; this can be proved in the same way as Theorem 4.7, using Corollary 4.4. We conclude that when X has nonnegative entries, the error is reduced as k and m approach the main diagonal ($k = m$) and first superdiagonal ($k + 1 = m$) of the Padé table. In the rest of the paper we will concentrate on the use of the diagonal Padé approximants $r_m \equiv r_{mm}$.

4.4 Evaluating Padé approximants of $(I - X)^p$

Just as for the logarithm [69], there are several possible methods for evaluation of Padé approximant $r_m(X)$ at $X \in \mathbb{C}^{n\times n}$:

1. Evaluation of the numerator and denominator in the representation $r_m(x) = p_m(x)/q_m(x)$ by Horner's method or the Paterson and Stockmeyer method [72, sec. 4.2], [110].

2. Evaluation of the continued fraction form of $r_m(X)$ in either top-down fashion or bottom-up fashion.

3. Evaluation of $r_m(x) = p_m(x)/q_m(x)$ using the representations of p_m and q_m as products of linear factors (the zeros of p_m and q_m are all real).

4. Evaluation of the partial fraction representation $r_m(x) = \alpha_0 + \sum_{j=1}^{m} \alpha_j/(\beta_j - x)$.

In this section we will give a detailed comparison of these possibilities with respect to numerical stability and computational cost to find the best method in the context of the algorithm to be developed in the next section.

4.4.1 Horner's method and the Paterson and Stockmeyer method

One class of methods is based on the rational representation $r_m(x) = p_m(x)/q_m(x)$ of the Padé approximant: evaluate the numerator and the denominator matrix polynomials $p_m(X)$ and $q_m(X)$, respectively, and then compute $Y = r_m(X)$ by solving $q_m Y = p_m$. Here, we use Horner's method and the Paterson Stockmeyer method [72, sec. 4.2] [110] to evaluate the polynomials. Let $p_m(X)$ be a matrix polynomial

$$p_m(X) = \sum_{k=0}^{m} b_k X^k. \tag{4.24}$$

Algorithm 4.8 (Horner's method)**.** *This algorithm evaluates the polynomial* (4.24) *by Horner's method.*

```
1  S_{m-1} = b_m X + b_{m-1} I
2  for k = m-2: -1: 0
3       S_k = X S_{k+1} + b_k I
4  end
5  p_m = S_0
```

Algorithm 4.9 (the Paterson and Stockmeyer method)**.** *This algorithm evaluates the polynomial* (4.24) *by the Paterson and Stockmeyer method, in which* $p_m(X)$ *is written as*

$$p_m(X) = \sum_{k=0}^{r} B_k \cdot (X^s)^k, \quad r = \lfloor m/s \rfloor, \tag{4.25}$$

where s *is an integer parameter and*

$$B_k = \begin{cases} b_{sk+s-1}X^{s-1} + \cdots + b_{sk+1}X + b_{sk}I, & k = 0 : r-1, \\ b_m X^{m-sr} + \cdots + b_{sr+1}X + b_{sr}I, & k = r. \end{cases}$$

```
1  Compute X^2, ..., X^s
2  Evaluate (4.25) by Horner's method with each B_k formed as needed
```

Van Loan's variant of the Paterson and Stockmeyer method is to compute p_m a column at a time, which reduces the storage required in the method but increases the cost of evaluating p_m.

Based on the standard model of floating point arithmetic with unit roundoff u, we now investigate the stability and accuracy of the evaluation of r_m with Algorithm 4.8 or Algorithm 4.9 to compute p_m and q_m. Let $\|\cdot\|_p$ denote any p-norm and let $\widehat{Y} = Y + \Delta Y$ denote the computed Y. The errors in obtaining Y from $q_m Y = p_m$ result from computing q_m and p_m and solving the system. The computed $\widehat{q}_m = q_m + \Delta Q$ and $\widehat{p}_m = p_m + \Delta P$ from Horner's method and the Paterson and Stockmeyer method satisfy [69, Lemma 3.1], [72, Thm. 4.5]

$$\begin{aligned} \|\Delta Q\| &\le m(n+1)u\widetilde{q}_m(\|X\|) + O(u^2), \\ \|\Delta P\| &\le m(n+1)u\widetilde{p}_m(\|X\|) + O(u^2), \end{aligned}$$

where $\widetilde{q}_m$ and $\widetilde{p}_m$ are polynomials corresponding to q_m and p_m in the form of (4.24) with the coefficient of each term replaced by its absolute value, respectively. Assume that the linear system solver is stable, so that [70, sec. 9]

$$\widehat{q}_m \widehat{Y} = \widehat{p}_m + R$$

where $\|R\| \le \gamma_n u \|\widehat{q}_m\| \|\widehat{Y}\|$ for some constant γ_n. Then from $q_m \Delta Y + \Delta Q Y = \Delta P + R + O(u^2)$, the overall forward error bound for $\widehat{Y}$ will be of the form

$$\frac{\|Y - \hat{Y}\|}{\|Y\|} \le d(m,n)u\kappa(q_m)\eta(X) + O(u^2), \tag{4.26}$$

where $d_j(m,n)$ denotes a constant depending on m and n and η is given by

$$\eta(X) = \left(\frac{\tilde{p}_m(\|X\|)}{\|q_m(X)\|\|Y\|} + \frac{\tilde{q}_m(\|X\|)}{\|q_m(X)\|} + \frac{\gamma_n}{d_1(m,n)} \right) \ge 1. \tag{4.27}$$

The stability of this method depends on the condition number $\kappa(q_m(X))$ which is bounded above by

$$\kappa(q_m(X)) \le \frac{q_m(-\|X\|)}{q_m(\|X\|)} \tag{4.28}$$

as shown in Lemma 4.3.

4.4.2 Continued fraction form

The Padé approximant $r_m(x)$ to $(1-x)^p$ has the continued fraction expansion [9, p. 66], [10, p. 174]

$$r_m(x) = 1 + \cfrac{c_1 x}{1 + \cfrac{c_2 x}{1 + \cfrac{c_3 x}{\cdots \cfrac{}{1 + \cfrac{c_{2m-1}x}{1 + c_{2m}x}}}}}, \tag{4.29}$$

where

$$c_1 = -p, \quad c_{2j} = \frac{-j+p}{2(2j-1)}, \quad c_{2j+1} = \frac{-j-p}{2(2j+1)}, \quad j = 1, 2, \ldots.$$

This expansion provides a convenient means to evaluate $r_m(X)$ for $X \in \mathbb{C}^{n\times n}$, either in top-down fashion or in bottom-up fashion. We will summarize both methods as follows.

Algorithm 4.10 (continued fraction, top-down)**.** *This algorithm evaluates the continued fraction* (4.29) *in top-down fashion at the matrix* $X \in \mathbb{C}^{n\times n}$*.*

```
1  P_{-1} = I, Q_{-1} = 0, P_0 = I, Q_0 = I
2  for j = 1: 2m
3        P_j = P_{j-1} + c_j X P_{j-2}
4        Q_j = Q_{j-1} + c_j X Q_{j-2}
5  end
6  r_m = P_{2m} Q_{2m}^{-1}
```

We now investigate the numerical stability of this recurrence. Since Algorithm 4.10 essentially computes r_m by converting the continued fraction to the rational form, the overall forward error bound (4.26) applies here with the constant $\eta(X)$ derived as follows.

The recurrence for the Q_j can be expressed as

$$\begin{aligned}\begin{bmatrix} Q_j \\ Q_{j-1} \end{bmatrix} &= \begin{bmatrix} I & c_j X \\ I & 0 \end{bmatrix} \begin{bmatrix} Q_{j-1} \\ Q_{j-2} \end{bmatrix} \\ &= \begin{bmatrix} I & c_j X \\ I & 0 \end{bmatrix} \cdots \begin{bmatrix} I & c_2 X \\ I & 0 \end{bmatrix} \begin{bmatrix} I \\ I \end{bmatrix}.\end{aligned}$$

From a standard error bound for matrix multiplication [70, Lem. 3.6] the errors in the computed $\widehat{Q}_{2m} = Q_{2m} + \Delta Q$ satisfy

$$\|\Delta Q\| \le d_2(m,n)u \prod_{j=2}^{2m}(1 + |c_j|\,\|X\|) + O(u^2).$$

Similarly, for the computed $\widehat{P}_{2m} = P_{2m} + \Delta P$,

$$\|\Delta P\| \le d_3(m,n)u \prod_{j=1}^{2m}(1 + |c_j|\,\|X\|) + O(u^2).$$

Again, assume that the solver for the linear systems $YQ_{2m} = P_{2m}$ is stable. Then

$$\widehat{Y}\widehat{Q}_{2m} = \widehat{P}_{2m} + R,$$

where $\|R\| \le \gamma_n u \|\widehat{Q}_{2m}\|\,\|\widehat{Y}\|$. Therefore, from $\Delta Y Q_{2m} + Y\Delta Q = \Delta P + R + O(u^2)$, we have the forward error bound (4.26) for the computed $\widehat{Y}$ with $\kappa(Q_{2m})$ in place of $\kappa(q_m)$ and η given by

$$\eta(X) = \frac{\prod_{j=2}^{2m}(1 + |c_j|\,\|X\|)}{\|Q_{2m}\|}\left(1 + \frac{1 + |c_1|\,\|X\|}{\|Y\|}\right) + \frac{\gamma_n}{d_4(m,n)}. \tag{4.30}$$

We now proceed to summarize the bottom-up evaluation of (4.29).

Algorithm 4.11 (continued fraction, bottom-up)**.** *This algorithm evaluates the continued fraction* (4.29) *in bottom-up fashion at the matrix* $X \in \mathbb{C}^{n\times n}$*.*

```
1  Y_{2m} = c_{2m} X
2  for j = 2m − 1: −1: 1
3      Solve (I + Y_{j+1}) Y_j = c_j X for Y_j
4  end
5  r_m = I + Y_1
```

We now investigate the numerical stability of this recurrence. Assume that $\|Y_j\| < 1$ for all j, and let $\widehat{Y}_j \equiv Y_j + \Delta Y_j$ denote the computed Y_j. The errors in obtaining Y_j from $(I + Y_{j+1})Y_j = c_j X$ result from forming the right-hand side and solving the system. We assume that the solver is stable, so that [70, sec. 9]

$$(I + \widehat{Y}_{j+1})\widehat{Y}_j = c_j X + F_j + R_j,$$

where $\|F_j\| \le u|c_j|\,\|X\|$ and $\|R_j\| \le \gamma_n u(1 + \|\widehat{Y}_{j+1}\|)\|\widehat{Y}_j\|$, for some constant γ_n, where u in the unit roundoff. Then $(I + Y_{j+1})\Delta Y_j = F_j + R_j - \Delta Y_{j+1}Y_j + O(u^2)$, which implies

$$\begin{aligned}\|\Delta Y_j\| \;\le\; & \frac{1}{1 - \|Y_{j+1}\|}\big(u|c_j|\,\|X\| + \gamma_n u(1 + \|Y_{j+1}\|)\,\|Y_j\| + \|Y_j\|\,\|\Delta Y_{j+1}\|\big) \\ & + O(u^2), \quad j = 2m-1:-1:1, \qquad \|\Delta Y_{2m}\| \le u|c_{2m}|\,\|X\|.\end{aligned} \tag{4.31}$$

We can bound $\|Y_j\|$ from the recurrence

$$\|Y_j\| \le \frac{|c_j|\|X\|}{1-\|Y_{j+1}\|}, \qquad j = 2m-1:-1:1, \qquad \|Y_{2m}\| = |c_{2m}|\|X\|. \tag{4.32}$$

Together, the recurrences (4.31) and (4.32) allow us to compute, to first order, a bound on $\|\Delta Y_1\|$ for any given $\|X\|$. An upper bound for the relative error can then be obtained by using $\|Y_1\| \ge |c_1|\|X\|/(1+\|Y_2\|)$ together with the upper bound for $\|Y_2\|$ from (4.32).

With the recurrence (4.32) we can therefore compute a bound on the condition number $\kappa(I+Y_j)$ for solving the linear systems

$$\kappa(I+Y_j) \le \frac{1+\|Y_j\|}{1-\|Y_j\|}. \tag{4.33}$$

4.4.3 Product form representation

This method is based on the product form representation of the denominator and numerator polynomials: $p_m(x) = \prod_{i=1}^m (s_i - x)/\prod_{i=1}^m s_i$ and $q_m(x) = \prod_{i=1}^m (t_i - x)/\prod_{i=1}^m t_i$, where s_i and t_i, $i = 1:m$, are the zeros of $p_m(x)$ and $q_m(x)$, respectively. Note that $p_m(0) = q_m(0) = 1$. Then we can rewrite r_m in the product form as

$$r_m(x) = c_m \prod_{i=1}^m \frac{s_i - x}{t_i - x}, \tag{4.34}$$

where $c_m = \prod_{i=1}^m t_i/\prod_{i=1}^m s_i$. The matrix $r_m(X)$ can be evaluated by solving m multiple right-hand side linear systems successively, as described in the following algorithm.

Algorithm 4.12 (product form)**.** *This algorithm evaluates the product form* (4.34) *at the matrix* $X \in \mathbb{C}^{n\times n}$*.*

```
1  Y_0 = I
2  for j = 1: m
3      Solve (t_j I - X)Y_j = (s_j I - X)Y_{j-1} for Y_j
4  end
5  r_m = c_m Y_m
```

With an idea from Swarztrauber [125], we can save the cost of one matrix multiplication for each j (one matrix-vector multiplication in Swarztrauber's case since Y_j's are vectors there) while solving the linear systems $(t_jI - X)Y_j = (s_jI - X)Y_{j-1}$. The idea is to rewrite the linear system as

$$(t_jI - X)(Y_j - Y_{j-1}) = (t_j - s_j)Y_{j-1}. \tag{4.35}$$

This essentially uses the partial fraction representation [24] for $j = 1:m$

$$\frac{x - s_j}{x - t_j} = 1 + \frac{t_j - s_j}{x - t_j}.$$

Then Algorithm 4.12 is implemented with line 3 replaced by

" 3 Solve $(t_jI - X)T_j = Y_{j-1}$ for T_j; $\quad Y_j = Y_{j-1} + (t_j - s_j)T_j$"

In order to reduce the amplification of the errors present in T_j, Swarztrauber [125] suggests ordering s_j and t_j such that $|t_j - s_j|$ is small for all j.

The product form method is based on the availability of the zeros of numerator and denominator polynomials. We now introduce a practical way of computing the zeros. As shown in the proof of Corollary 4.4, the denominator $q_m(x)$ of the $[m/m]$ Padé approximant of ${}_2F_1(-p, 1, 1, x) = (1-x)^p$ is that of the $[m-1/m]$ Padé approximant of ${}_2F_1(-p+1, 1, 2, x)$. Recall that $-1 < p < 1$ and thus $0 < -p + 1 < 2$. The following result is a special case of [92, (1.22) and Remark 2] that shows a well-known representation for the denominator q_m of the Padé approximant of ${}_2F_1(\alpha, 1, \gamma, x)$ with $0 < \alpha < \gamma$ in terms of orthogonal polynomials: we have

$$q_m(x) = x^m \psi_m(\frac{1}{x}), \quad q_m(0) = 1. \tag{4.36}$$

Here, the ψ_m are the orthogonal polynomials given by the Jacobi orthogonal polynomials over $(-1, 1)$ under the variable transformation $\widetilde{x} = 2x - 1$, which is given by

$$\psi_m(x) = cP_m^{(p,-p)}(2x - 1),$$

where c is a normalization constant and $P_m^{(a,b)}(\widetilde{x})$ is the mth degree orthogonal polynomial over $-1 < \widetilde{x} < 1$ with respect to the weight function $(1 - \widetilde{x})^a(1 + \widetilde{x})^b$ for $a, b > -1$ [4, sec. 22.7]. Now the problem reduces to computing the zeros x_i, $i = 1 : m$ of $P_m^{(p,-p)}(\widetilde{x})$ since for each x_i, $2/(1 + x_i)$ is a zero of $p_m(x)$. Golub and Welsch [54] propose an effective algorithm to compute the Gauss quadrature rules, where the zeros of an orthogonal polynomial are obtained from the computation of the eigenvalues of a tridiagonal matrix constructed from the three term recurrence relation of the orthogonal polynomials. The $P_i \equiv P_i^{(p,-p)}(x)$, $i = 0, 1, \ldots$ satisfy the following recurrence [4, sec. 22.7]

$$\begin{cases} P_0 = 1 \\ P_1 = p + x \\ P_{i+1} = a_{i+1}xP_i - b_{i+1}P_{i-1}, \quad i = 1, 2, \ldots, \end{cases} \tag{4.37}$$

with $a_i = \dfrac{2i-1}{i}$, $b_i = \dfrac{(i-1)^2 - p^2}{i(i-1)}$, $i = 2, 3, \ldots$. Then the computation of the zeros of $P_m^{(p,-p)}(\tilde{x})$ amounts to computing the eigenvalues of the $m \times m$ symmetric tridiagonal matrix

$$J_m = \begin{bmatrix} -p & \beta_1 & & & \\ \beta_1 & 0 & \beta_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \beta_{m-2} & 0 & \beta_{m-1} \\ & & & \beta_{m-1} & 0 \end{bmatrix}, \tag{4.38}$$

where $\beta_i = (i^2 - p^2)^{1/2}/(4i^2 - 1)^{1/2}$, $i = 1 : m - 1$. And then we obtain the zeros of the denominator polynomial q_m accordingly. To obtain the zeros of the numerator

polynomial p_m, we use again the duality property of Padé approximants [10, Theorem 1.5.1]. Recall that the numerator p_m of the $[m/m]$ Padé approximant of $(1-x)^p$ is the denominator of the $[m/m]$ Padé approximant of $(1-x)^{-p}$. Analogously to the above discussion, we have

$$p_m(x) = x^m \psi_m(\frac{1}{x}), \quad p_m(0) = 1 \tag{4.39}$$

with

$$\psi_m(x) = cP_m^{(-p,p)}(2x-1).$$

Furthermore, the zeros of $P_m^{(-p,p)}(\tilde{x})$ are exactly the eigenvalues of J_m (4.38) with the $(1,1)$ element $-p$ replaced by p.

To simplify the error analysis for the overall computation of the product form evaluation, we assume that the zeros t_j and s_j are exactly computed and that there are no errors in forming $t_jI - X$, for $j = 1:m$. So the errors in the computed $\widehat{r}_m$ result from solving the linear systems $(t_jI - X)T_j = Y_{j-1}$ and forming Y_j. Denote $\widehat{T}_j = T_j + \Delta T_j$ and $\widehat{Y}_j = Y_j + \Delta Y_j$ the computed T_j and Y_j, respectively. Then [70, sec. 9]

$$(t_jX - I)\widehat{T}_j = \widehat{Y}_{j-1} + R_j$$

where $\|R_j\| \le \gamma_n u \|t_jI - X\| \, \|\widehat{T}_j\|$. The computed $\widehat{Y}_j$ satisfies $\widehat{Y}_j = \widehat{Y}_{j-1} + (t_j - s_j)\widehat{T} + F_j$, where $\|F_j\| \le \gamma_n u \|\widehat{Y}_j\|$. From

$$\begin{cases} (t_jI - X)\Delta T_j = \Delta Y_{j-1} + R_j, \\ \Delta Y_j = \Delta Y_{j-1} + (t_j - s_j)\Delta T_j + F_j, \end{cases} \tag{4.40}$$

it follows that, for $j = 1:m$,

$$\begin{cases} \|\Delta T_j\| \le \|(t_jI - X)^{-1}\| \, \|\Delta Y_{j-1}\| + \gamma_n u \|t_jI - X\| \, \|(t_jI - X)^{-1}\| \, \|T_j\|, \\ \|\Delta Y_j\| \le \|\Delta Y_{j-1}\| + |t_j - s_j| \, \|\Delta T_j\| + \gamma_n u \|Y_j\|. \end{cases} \tag{4.41}$$

Therefore, the errors in the computed $\widehat{Y}_m = Y_m + \Delta Y_m$ can be obtained from the recurrence (4.41) with the inequalities $\|(t_jI - X)^{-1}\| \le 1/(t_j - \|X\|)$ and

$$\|T_j\| \le \frac{\|Y_{j-1}\|}{t_j - \|X\|}, \quad \|Y_j\| \le \frac{s_j + \|X\|}{t_j - \|X\|} \|Y_{j-1}\|$$

to bound $\|T_j\|$ and $\|Y_j\|$ above, where we have used the fact that $t_j > 1$ and $s_j > 1$ for all $j = 1{:}\, m$. An upper bound for the relative error can then be obtained by the recurrence $\|Y_j\| \ge (s_j + \|X\|)\|Y_{j-1}\|/(t_j + \|X\|)$.

The stability of the product form method is dependent on the condition of the linear systems to be solved, which is bounded by

$$\kappa(t_jI - X) \le \frac{t_j + \|X\|}{t_j - \|X\|}. \tag{4.42}$$

Table 4.1: Cost of evaluating $r_m(X)$. M denotes the cost of a matrix multiplication and D the cost of solving a linear system with n right-hand sides. The integer parameters $1 \le s \le m$ are used in the Paterson-Stockmeyer method with the optimal values being $\sqrt{2m}$ and $\sqrt{m}$, respectively. $f(s,m) = 1$ if s divides m and 0 otherwise.

Method	Computational cost	Storage
Horner	$2(m-1)M + D$	$3n^2$
Paterson-Stockmeyer	$(s + 2r - 1 - 2f(s,m))M + D$ $\gtrsim (2\sqrt{2}\sqrt{m} - 1)M + D$	$(s+2)n^2$
Continued fraction	top-down: $2(2m-2)M + D$	$5n^2$
	bottom-up: $(2m-1)D$	$3n^2$
Product form	mD	$3n^2$
Partial fraction	mD	$3n^2$

4.4.4 Partial fraction form

This method is based on the partial fraction representation

$$r_m(x) = \alpha_0 + \sum_{j=1}^{m} \frac{\alpha_j}{t_j - x}. \tag{4.43}$$

The coefficients α_j can be given by the zeros t_j for the denominator polynomial q_m and s_j for the numerator p_m as

$$\alpha_0 = \prod_{i=1}^{m} \frac{t_i}{s_i} \quad \text{and} \quad \alpha_j = \alpha_0 \frac{\prod_{i=1}^{m}(s_i - t_j)}{\prod_{i \ne j}(t_i - t_j)}, \qquad j = 1 : m.$$

An advantage of the partial fraction form over the product form is that the m linear systems in the former can be solved in parallel. The accuracy of the partial fraction method is dependent on the condition of the matrices $t_j I - X$. The normwise relative error is roughly bounded by $d(m,n)u\phi$ [69, (3.7)] where

$$\phi = \max_i [\alpha_i \kappa(t_i I - X)]. \tag{4.44}$$

Table 4.1, partially taken from [69], summarizes the cost of the methods discussed in this section.

4.4.5 Comparison and numerical experiments

We will show terms from the error analysis in the following tables for a range of $p \in (0,1)$ and $\|X\| \in (0,1)$. 2-norms are used here and throughout this section and the values of m, shown in Table 4.2, are chosen as the smaller of 100 and the minimal value for which

$$\|r_m(X) - (I - X)^p\| \le |(1 - \|X\|)^p - r_m(\|X\|)| \le u, \tag{4.45}$$

with $u = 2^{-53} \approx 1.1 \times 10^{-16}$, where the first inequality always holds by Theorem 4.6. Table 4.3 shows the bounds for $\kappa(q_m)$ from (4.28), $\max_j \kappa(I + Y_j)$ from (4.33) and $\max_j \kappa(t_j I - X)$ from (4.42), comparing the numerical stability of different methods. We use "–" in place of the negative outputs which should be positive theoretically. This is due to the rounding errors in finite precision computation. Table 4.4 shows the terms from the overall forward analysis: η in the error bounds for Horner's method (4.27) and the top-down continued fraction method (4.30), respectively, with $\gamma_n \equiv 1$ and $d(m, n) = m$; ϕ for the partial fraction method is defined in (4.44); d_1 and d_2 are the constants in the bound $\|\Delta Y\|/\|Y\| \leq du + O(u^2)$ from (4.32) and (4.41), respectively, with $\gamma_n \equiv 1$ (the bound scales roughly linearly with γ_n). For $Y = r_m(X)$, we approximated $\|Y\| \approx \|(I - X)^p\| \approx 1$ and $\|q_m\| \approx q_m(0) = 1$ when they were needed. "NaN" in the table stands for Not-a-Number in MATLAB, which is obtained as a result of dividing infinity by infinity. The infinity here is caused by overflow in computing the coefficients of the denominator and numerator of the rational representation of the Padé approximant with large m and certain values of p.

Table 4.5 gives the results of some numerical tests. The test matrices X are 8×8 random matrices with elements from the normal $(0, 1)$ distribution. We then scaled matrices X to get the desired values of norms. Table 4.5 shows the normwise relative errors $\|\widehat{Y} - Y\|/\|Y\|$ in $Y = (I - X)^p$ for a range of $p \in (0, 1)$. Here the "exact" matrix powers are computed using Algorithm 4.11 (which is stable and accurate anticipating the results from Table 4.3 and 4.4) at 100 digit precision with the VPA arithmetic of the Symbolic Math Toolbox.

Some observations can be made.

1. Horner's method, the Paterson-Stockmeyer method, and the continued fraction evaluated top-down can only be guaranteed to be stable if $\|X\|$ is much less than 1, below 0.25 say.

2. The factors from the error bounds for Horner's method, the top-down evaluation of the continued fraction and the constant in the error bound for the product form method grow rapidly as $\|X\|$ approaches 1. The factor for the partial fraction method increases as p approaches 1. The effect of rounding errors on the bottom-up evaluation of the continued fraction is negligible for all $\|X\|$ and p tested.

3. For the bottom-up evaluation of the continued fraction, the assumption $\|Y_j\| < 1$ was found to be satisfied in every case. The results show that as long as we keep $\|X\|$ below 0.9, say, the numerical stability of Algorithm 4.11 will be excellent. Table 4.5 confirms that the bottom-up evaluation of the continued fraction gives the best accuracy. In fact, in Algorithm 4.13, which is derived in the next section, with the bottom-up evaluation used in it we will limit $\|X\|$ to about 0.3, for other reasons.

4.5 Schur–Padé algorithm for A^p

Now we develop an algorithm for computing A^p for a real $p \in (-1, 1)$, where A has no nonpositive real eigenvalues. We can restrict p to $(-1, 1)$ without loss of generality,

Table 4.2: Minimal values of m for which (4.45) holds.

	p				
$\|X\|$	0.1	0.3	0.5	0.7	0.9
0.99	88	100	100	84	79
0.95	38	39	39	39	36
0.90	27	27	27	27	26
0.75	16	16	16	16	15
0.50	9	10	10	10	10
0.25	6	6	7	7	6
0.10	5	5	5	5	5

Table 4.3: Terms from the stability analysis, for different $\|X\| < 1$ and $p \in (0, 1)$.

	p				
$\|X\|$	0.1	0.3	0.5	0.7	0.9
			$\kappa(q_m)$ (4.28)		
0.99	7.26e15	NaN	NaN	2.45e16	–
0.95	4.00e18	3.11e18	–	–	–
0.90	5.60e13	4.17e13	3.11e13	2.33e13	5.33e12
0.75	1.03e6	8.50e5	7.00e5	5.77e5	1.99e5
0.50	1.10e2	1.68e2	1.50e2	1.35e2	1.21e2
0.25	4.45e0	4.23e0	5.18e0	4.92e0	3.63e0
0.10	1.63e0	1.60e0	1.57e0	1.54e0	1.51e0
			$\max_j \kappa(I + Y_j)$ (4.33)		
0.99	6.46e0	6.96e0	1.90e1	4.92e1	1.25e2
0.95	4.45e0	3.91e0	7.94e0	1.53e1	2.86e1
0.90	3.57e0	2.99e0	5.32e0	9.02e0	1.49e1
0.75	2.45e0	2.03e0	3.00e0	4.28e0	5.96e0
0.50	1.68e0	1.50e0	1.83e0	2.25e0	2.73e0
0.25	1.27e0	1.20e0	1.31e0	1.45e0	1.59e0
0.10	1.10e0	1.07e0	1.11e0	1.15e0	1.20e0
			$\max_j \kappa(t_j I - X)$ (4.42)		
0.99	1.95e2	1.95e2	1.94e2	1.91e2	1.89e2
0.95	3.82e1	3.80e1	3.78e1	3.76e1	3.71e1
0.90	1.86e1	1.85e1	1.84e1	1.83e1	1.82e1
0.75	6.86e0	6.82e0	6.79e0	6.75e0	6.67e0
0.50	2.93e0	2.93e0	2.91e0	2.90e0	2.88e0
0.25	1.63e0	1.63e0	1.63e0	1.62e0	1.60e0
0.10	1.21e0	1.21e0	1.20e0	1.20e0	1.20e0

Table 4.4: Terms from error analysis, for different $\|X\| < 1$ and $p \in (0,1)$. Here $\epsilon(p, \|X\|) := |(1 - \|X\|)^p - r_m(\|X\|)|$

		Approx. to			$\|\Delta Y\|/\|Y\| \le du + O(u^2)$	
$\|X\|$	$\epsilon(p, \|X\|)$	η in (4.27)	η in (4.30)	ϕ (4.44)	d from (4.32)	d from (4.41)
			$p = 0.1$			
0.99	0.00e0	3.83e14	8.12e16	2.52e2	4.24e2	4.21e59
0.95	0.00e0	1.89e6	1.10e7	6.42e2	8.79e1	1.03e22
0.90	0.00e0	2.35e4	5.84e4	3.07e2	4.25e1	5.94e13
0.75	0.00e0	2.43e2	2.48e2	1.00e2	1.49e1	1.22e6
0.50	1.11e-16	1.37e1	8.57e0	2.99e1	6.00e0	1.84e2
0.25	1.11e-16	3.99e0	2.26e0	1.30e1	3.26e0	1.47e1
0.10	1.11e-16	2.55e0	1.49e0	8.93e0	2.41e0	7.79e0
			$p = 0.3$			
0.99	3.89e-16	NaN	1.81e19	8.18e2	1.53e2	3.77e67
0.95	5.55e-17	2.72e6	1.85e7	1.30e4	4.32e1	2.77e22
0.90	1.11e-16	2.36e4	6.39e4	5.07e3	2.39e1	4.48e13
0.75	1.11e-16	2.44e2	2.68e2	1.35e3	1.01e1	1.02e6
0.50	0.00e0	1.69e1	1.14e1	4.17e2	4.80e0	3.07e2
0.25	0.00e0	3.99e0	2.31e0	1.20e2	2.94e0	2.33e1
0.10	0.00e0	2.55e0	1.50e0	7.73e1	2.31e0	1.42e1
			$p = 0.5$			
0.99	3.75e-16	NaN	1.94e19	4.28e3	5.27e1	2.24e67
0.95	8.33e-17	2.75e6	1.97e7	1.62e5	2.03e1	1.94e22
0.90	1.11e-16	2.38e4	6.80e4	5.47e4	1.29e1	3.38e13
0.75	0.00e0	2.45e2	2.83e2	1.18e4	6.67e0	8.57e5
0.50	0.00e0	1.70e1	1.18e1	3.05e3	3.80e0	3.71e2
0.25	0.00e0	4.48e0	2.62e0	1.11e3	2.64e0	6.71e1
0.10	0.00e0	2.55e0	1.51e0	4.36e2	2.22e0	3.34e1
			$p = 0.7$			
0.99	1.39e-17	8.83e13	1.70e16	9.29e4	1.68e1	1.68e56
0.95	5.55e-17	2.78e6	2.05e7	2.42e6	8.93e0	1.37e22
0.90	2.78e-17	2.41e4	7.08e4	7.06e5	6.60e0	2.55e13
0.75	5.55e-17	2.47e2	2.94e2	1.24e5	4.27e0	7.20e5
0.50	0.00e0	1.71e1	1.22e1	2.67e4	2.97e0	7.85e2
0.25	0.00e0	4.48e0	2.67e0	8.49e3	2.37e0	2.29e2
0.10	0.00e0	2.55e0	1.52e0	2.94e3	2.13e0	1.06e2
			$p = 0.9$			
0.99	8.67e-17	1.39e13	1.89e15	1.84e5	4.52e0	3.98e52
0.95	4.16e-17	9.56e5	5.81e6	8.10e7	3.48e0	1.72e20
0.90	2.78e-17	1.73e4	4.80e4	2.40e7	3.09e0	5.93e12
0.75	1.11e-16	1.86e2	2.13e2	3.12e6	2.61e0	2.78e5
0.50	0.00e0	1.72e1	1.25e1	7.10e5	2.29e0	5.22e3
0.25	1.11e-16	4.00e0	2.44e0	1.14e5	2.12e0	1.06e3
0.10	0.00e0	2.55e0	1.53e0	6.02e4	2.04e0	6.61e2

Table 4.5: Relative normwise errors $\|\widehat{Y} - Y\|/\|Y\|$ in $Y = (I - X)^p$ for a range of $p \in (0, 1)$.

$\|X\|$	m	Horner	Paterson-Stockmeyer	Continued fraction top-down	Continued fraction bottom-up	Product form	Partial fraction
				$p = 0.1$			
0.99	88	6.86e0	1.16e1	1.89e1	1.86e-17	3.64e-15	NaN
0.95	38	7.40e-8	9.74e-9	2.80e-8	1.05e-16	1.30e-15	7.55e-16
0.90	27	3.19e-10	2.48e-10	1.29e-10	1.42e-17	1.19e-15	1.22e-15
0.75	16	4.96e-14	4.06e-14	5.28e-14	2.12e-16	6.49e-16	8.45e-16
0.50	9	2.49e-15	1.48e-15	2.84e-15	1.04e-17	3.30e-16	3.26e-16
0.25	6	4.94e-16	6.24e-16	4.95e-16	3.73e-18	6.56e-16	4.37e-16
0.10	5	4.58e-16	4.85e-16	6.74e-16	1.96e-18	4.43e-16	2.52e-17
				$p = 0.3$			
0.99	100	NaN	NaN	1.44e2	7.96e-17	4.79e-15	NaN
0.95	39	3.72e-12	1.75e-12	3.30e-12	9.76e-17	1.02e-15	4.53e-15
0.90	27	1.76e-10	6.49e-11	1.12e-10	9.92e-17	1.14e-15	2.08e-15
0.75	16	3.52e-14	2.46e-14	6.36e-14	1.97e-16	5.24e-16	4.52e-15
0.50	10	3.08e-15	2.16e-15	4.77e-15	2.34e-17	4.92e-16	1.90e-15
0.25	6	4.92e-16	3.31e-16	7.04e-16	2.10e-16	4.22e-16	1.16e-15
0.10	5	4.39e-16	4.39e-16	4.43e-16	4.84e-18	4.36e-16	1.53e-15
				$p = 0.5$			
0.99	100	NaN	NaN	5.04e1	1.31e-16	2.74e-15	NaN
0.95	39	1.94e-7	6.75e-8	3.67e-8	1.07e-16	1.64e-15	4.26e-14
0.90	27	2.72e-10	1.43e-10	4.61e-10	5.46e-17	1.78e-15	1.55e-14
0.75	16	1.91e-14	1.24e-14	1.94e-14	1.02e-16	1.04e-15	1.30e-14
0.50	10	2.65e-15	2.17e-15	1.64e-15	9.93e-17	8.01e-16	9.53e-15
0.25	7	3.15e-16	4.84e-16	4.84e-16	1.03e-16	1.43e-15	5.29e-15
0.10	5	4.42e-16	4.37e-16	4.32e-16	8.04e-18	1.08e-15	2.15e-15
				$p = 0.7$			
0.99	84	6.21e2	1.18e1	1.46e1	1.59e-16	1.91e-15	NaN
0.95	39	2.00e-5	1.58e-5	6.05e-6	1.82e-16	1.43e-15	1.92e-13
0.90	27	3.03e-12	1.28e-12	1.17e-12	1.62e-16	1.15e-15	5.85e-14
0.75	16	1.99e-14	1.20e-14	3.00e-14	1.58e-16	1.92e-15	1.06e-13
0.50	10	1.50e-15	1.44e-15	2.72e-15	1.82e-16	1.84e-15	1.64e-14
0.25	7	3.20e-16	3.39e-16	5.01e-16	2.02e-16	1.71e-15	1.85e-14
0.10	5	3.24e-16	3.27e-16	8.66e-16	1.08e-16	1.29e-15	1.32e-14
				$p = 0.9$			
0.99	79	5.00e-1	3.57e-1	1.56e-2	1.68e-16	8.05e-15	NaN
0.95	36	2.76e-7	6.28e-7	2.40e-7	1.52e-16	6.32e-15	2.16e-12
0.90	26	9.15e-10	5.26e-10	7.85e-10	1.75e-16	1.04e-14	4.67e-13
0.75	15	5.14e-14	4.47e-14	4.99e-14	1.70e-16	1.03e-14	1.06e-12
0.50	10	1.17e-15	1.02e-15	1.61e-15	1.68e-16	9.44e-15	3.95e-13
0.25	6	4.26e-16	4.33e-16	5.10e-16	6.00e-17	1.09e-14	8.85e-14
0.10	5	4.48e-16	6.32e-16	6.23e-16	2.07e-16	1.07e-14	5.38e-14

Table 4.6: $\theta_m^{(p)}$, for $p = 1/2$ and selected m.

m	1	2	3	4	5	6	7	8	9
$\theta_m^{(1/2)}$	1.53e-5	2.25e-3	1.92e-2	6.08e-2	1.25e-1	2.03e-1	2.84e-1	3.63e-1	4.35e-1
m	10	11	12	13	14	15	16	32	64
$\theta_m^{(1/2)}$	4.99e-1	5.55e-1	6.05e-1	6.47e-1	6.84e-1	7.17e-1	7.44e-1	9.27e-1	9.81e-1

Table 4.7: Minimum values of $\theta_m^{(p)}$, for $p \in [-1, 1]$.

m	1	2	3	4	5	6	7	8	9
$\min_p \theta_m^{(p)}$	1.51e-5	2.24e-3	1.88e-2	6.04e-2	1.24e-1	2.00e-1	2.79e-1	3.55e-1	4.25e-1
m	10	11	12	13	14	15	16	32	64
$\min_p \theta_m^{(p)}$	4.87e-1	5.42e-1	5.90e-1	6.32e-1	6.69e-1	7.00e-1	7.28e-1	9.15e-1	9.76e-1

since in general we can compute $A^p = A^{p_1} A^{p_2}$ with $p_1 \in (-1, 1)$ and p_2 an integer. How best to choose p_1 and p_2 is considered in Section 4.6.

Our algorithm exploits the relation $A^p = (A^{1/2^k})^{p \cdot 2^k}$. We take square roots of A repeatedly until $A^{1/2^k}$ is close to the identity matrix. Then, with $X = I - A^{1/2^k}$, we can use the approximation $(A^{1/2^k})^p \approx r_m(X)$, where r_m is the $[m/m]$ Padé approximant to $(1 - x)^p$. We recover an approximation to the pth power of the original matrix from $A^p \approx r_m(X)^{2^k}$. This approach is analogous to the inverse scaling and squaring method for the matrix logarithm [28], [72, sec. 11.5], [91]. In order to facilitate the computation of the square roots we compute an initial Schur decomposition $A = QTQ^*$, so that the problem is reduced to that for a triangular matrix.

For any $p \in [-1, 1]$ and m we denote by $\theta_m^{(p)}$ the largest value of $\|X\|$ such that the second inequality holds in (4.45). With $u = 2^{-53}$, we determined $\theta_m^{(p)}$ empirically in MATLAB, using high precision computations with the Symbolic Math Toolbox. For $p = 1/2$ and a range of $m \in [1, 64]$. Table 4.6 reports the results to three significant figures. To see how the values of $\theta_m^{(p)}$ vary with p for a specific m, we show in Figure 4.1 the values of $\theta_m^{(p)}$ corresponding to 324 different values of p between -0.999 and 0.999, for a range of m. Table 4.7 reports the corresponding minimum values of $\theta_m^{(p)}$ over $p \in [-1, 1]$. For each m, $\theta_m^{(p)}$ tends to 1 as p tends to -1, 0 or 1. Our results show, however, that the relative variation of $\theta_m^{(p)}$ with p is slight, except when p is within distance about 10^{-4} of -1, 0, or 1. We therefore base our algorithm on the values

$$\theta_m = \min_{p \in [-1,1]} \theta_m^{(p)}, \tag{4.46}$$

and do not optimize the algorithm parameters separately for each particular p.

In designing the algorithm we minimize the cost subject to achieving the desired accuracy, adapting a strategy used within the inverse scaling and squaring algorithm for the matrix logarithm in [28], [72, sec. 11.5]. Computing a square root of a triangular matrix T by the Schur method of Björck and Hammarling [17], [72, Alg. 6.3] costs $n^3/3$ flops, while evaluating $r_m(T)$ by Algorithm 4.11 costs $(2m - 1)n^3/3$ flops.

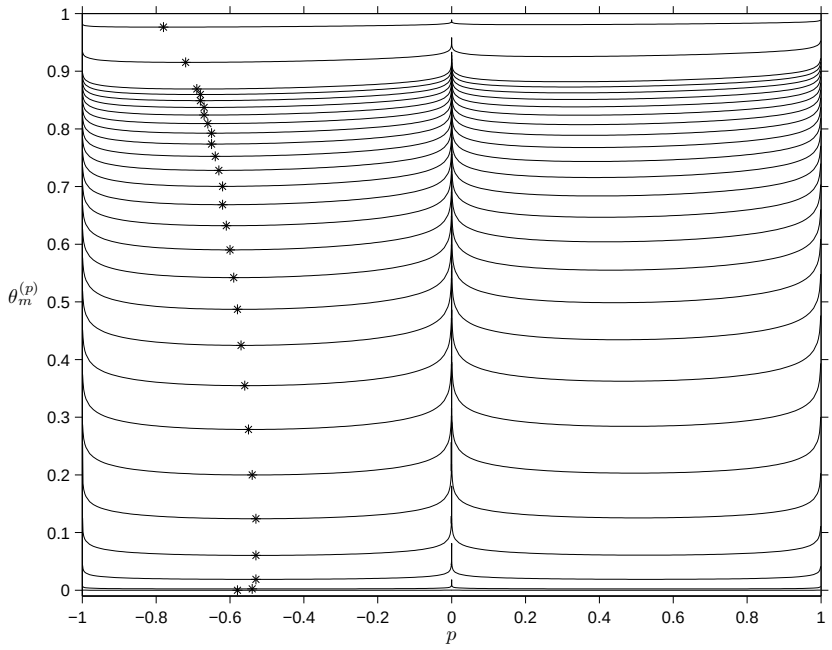

Figure 4.1: $\theta_m^{(p)}$ against p, for $m = 1\colon 25, 32, 64$; $m = 1$ is the lowest curve and $m = 64$ the highest curve. θ_m in (4.46) is marked as "$*$". The curves are not symmetric about $p = 0$.

Bearing in mind the squaring phase, it is therefore worthwhile to compute an extra square root if it allows a reduction in the Padé degree m by more than 1. Considering that

$$\|I - T^{1/2}\| = \|(I + T^{1/2})^{-1}(I - T)\| \approx \tfrac{1}{2}\|I - T\| \tag{4.47}$$

once $T \approx I$ and that, from Table 4.7, $\theta_m/2 < \theta_{m-2}$ for $m > 7$, the cost of computing T^p when $\|I - T\| > \theta_7$ will be minimized if we take square roots of T repeatedly until $\|I - T^{1/2^k}\| \le \theta_7$. Then it is worth taking one more square root if it reduces the required m by more than 1.

An important final ingredient of our algorithm is a special implementation of the squaring phase, obtained by adapting the approach suggested by Al-Mohy and Higham [5] for the matrix exponential. The squaring phase forms $r_m(I - T^{1/2^k})^{2^j} \approx T^{p/2^{k-j}}$, $j = 1\colon k$. But we can evaluate the diagonal and first superdiagonal elements of $T^{p/2^{k-j}}$ *exactly* from explicit formulae, and injecting these values into the recurrence should reduce the propagation of errors. The diagonal entries are computed in the obvious way. We now derive an appropriate formula for the first superdiagonal.

The (1,2) element of $F = \left[\begin{smallmatrix} \lambda_1 & t_{12} \\ 0 & \lambda_2 \end{smallmatrix}\right]^p$ is given by $f_{12} = t_{12}(\lambda_2^p - \lambda_1^p)/(\lambda_2 - \lambda_1)$ if $\lambda_1 \ne \lambda_2$, or $p\lambda_1^{p-1}t_{12}$ otherwise [72, sec. 4.6]. We need a way of evaluating the divided difference $(\lambda_2^p - \lambda_1^p)/(\lambda_2 - \lambda_1)$ accurately even when λ_1 and λ_2 are very close; this formula itself suffers from cancellation. We have

$$\begin{aligned} \frac{\lambda_2^p - \lambda_1^p}{\lambda_2 - \lambda_1} &= \frac{\exp(p\log\lambda_2) - \exp(p\log\lambda_1)}{\lambda_2 - \lambda_1} \\ &= \exp\left(\tfrac{p}{2}\big(\log\lambda_2 + \log\lambda_1\big)\right) \frac{\exp\left(\tfrac{p}{2}\big(\log\lambda_2 - \log\lambda_1\big)\right) - \exp\left(\tfrac{p}{2}\big(\log\lambda_1 - \log\lambda_2\big)\right)}{\lambda_2 - \lambda_1} \end{aligned}$$

$$= \exp\left(\tfrac{p}{2}\big(\log\lambda_2+\log\lambda_1\big)\right)\frac{2\sinh\left(\tfrac{p}{2}\big(\log\lambda_2-\log\lambda_1\big)\right)}{\lambda_2-\lambda_1}.$$

The remaining problem is to evaluate $w = \log\lambda_2 - \log\lambda_1$ accurately. To avoid cancellation we can rewrite [72, sec. 11.6.2]

$$w = \log\left(\frac{\lambda_2}{\lambda_1}\right) + 2\pi i\mathcal{U}(\log\lambda_2-\log\lambda_1) = \log\left(\frac{1+z}{1-z}\right) + 2\pi i\mathcal{U}(\log\lambda_2-\log\lambda_1),$$

where $z = (\lambda_2-\lambda_1)/(\lambda_2+\lambda_1)$ and $\mathcal{U}(z)$ is the unwinding number of $z\in\mathbb{C}$ defined by

$$\mathcal{U}(z) := \frac{z-\log(e^z)}{2\pi i} = \left\lceil\frac{\operatorname{Im} z-\pi}{2\pi}\right\rceil \in \mathbb{Z}. \tag{4.48}$$

Then, using the hyperbolic arc tangent $\operatorname{atanh}(z)$, defined by

$$\operatorname{atanh}(z) := \frac{1}{2}\log\left(\frac{1+z}{1-z}\right), \tag{4.49}$$

w can be expressed as

$$w = 2\operatorname{atanh}(z) + 2\pi i\mathcal{U}(\log\lambda_2-\log\lambda_1).$$

Hence

$$f_{12} = t_{12}\exp\left(\tfrac{p}{2}\big(\log\lambda_2+\log\lambda_1\big)\right)\frac{2\sinh\left(p\big(\operatorname{atanh}(z)+\pi i\mathcal{U}(\log\lambda_2-\log\lambda_1)\big)\right)}{\lambda_2-\lambda_1}. \tag{4.50}$$

Overall, we have the formula

$$f_{12} = \begin{cases} t_{12}p\lambda_1^{p-1}, & \lambda_1=\lambda_2,\\ t_{12}\dfrac{\lambda_2^p-\lambda_1^p}{\lambda_2-\lambda_1}, & |\lambda_1|<|\lambda_2|/2 \text{ or } |\lambda_2|<|\lambda_1|/2,\\ (4.50), & \text{otherwise,}\end{cases} \tag{4.51}$$

where we evaluate the usual divided difference if λ_1 and λ_2 are sufficiently far apart. Several comments are made here. Note that we say λ_1 and λ_2 are sufficiently far apart if $|\lambda_1| < |\lambda_2|/2$ or $|\lambda_2| < |\lambda_1|/2$. One might intuitively prefer the criterion $|\lambda_1-\lambda_2| \geq \max\{|\lambda_1|,|\lambda_2|\}$. However, the latter criterion does not work for some extreme cases. For example, for $\lambda_1 = 10^{14}$ and $\lambda_2 = 1$, the latter criterion is not satisfied, whereas λ_1 and λ_2 are clearly far apart. Therefore we discard that criterion. For the scalar function $\exp\left(\tfrac{p}{2}\big(\log\lambda_2+\log\lambda_1\big)\right)$, numerical experiments in MATLAB show that it is more accurate to evaluate it in the same way as it appears here by the scalar exponential and logarithm than to evaluate it by $(\lambda_1\lambda_2)^{p/2}$. We are assuming that accurate implementations of the scalar sinh and atanh functions are available. The definition (4.49) is that used in MATLAB; there is an alternative to (4.49) which necessitates modifications to (4.50) described in [72, sec. 11.6.2].

Now we state the overall algorithm.

Algorithm 4.13 (Schur–Padé algorithm)**.** *Given $A\in\mathbb{C}^{n\times n}$ with no eigenvalues*

on $\mathbb{R}^-$ and a nonzero $p \in (-1,1)$ this algorithm computes $X = A^p$ via a Schur decomposition and Padé approximation. It uses the constants $\theta_m := \min_p \theta_m^{(p)}$ in Table 4.7. The algorithm is intended for IEEE double precision arithmetic.

```
1  Compute a (complex) Schur decomposition A = QTQ*.
2  If T is diagonal, X = QT^pQ*, quit, end
3  T_0 = T
4  k = 0, q = 0
5  while true
6        τ = ||T − I||_1
7        if τ ≤ θ_7
8           q = q + 1
9           j_1 = min{ i: τ ≤ θ_i, i = 3:7 }
10          j_2 = min{ i: τ/2 ≤ θ_i, i = 3:7 }
11          if j_1 − j_2 ≤ 1 or q = 2, m = j_1, goto line 16, end
12       end
13       T ← T^{1/2} using the Schur method [72, Alg. 6.3].
14       k = k + 1
15 end
16 Evaluate U = r_m(I − T) using Algorithm 4.11.
17 for i = k: −1: 0
18     if i < k, U ← U^2, end
19     Replace diag(U) by diag(T_0)^{p/2^i}.
20     Replace first superdiagonal of U by first superdiagonal of T_0^{p/2^i}
       obtained from (4.51) with p ← p/2^i.
21 end
22 X = QUQ*
```

Cost: $25n^3$ flops for the Schur decomposition plus $(2k + 2m - 1)n^3/3$ flops for U and $3n^3$ to get X: about $(28 + (2k + 2m - 1)/3)n^3$ flops in total.

Note that line 2 simply computes T^p in the obvious way when T is diagonal, that is, when A is normal; there is no need for Padé approximation in this case.

If A is real, we could take the real Schur decomposition at line 1, and compute the square roots of the now quasitriangular T at line 13 using the real Schur method [68], [72, Alg. 6.7]. This would guarantee a real computed $\widehat{X}$ and could be faster due to the avoidance of complex arithmetic.

4.6 General $p \in \mathbb{R}$

In developing the Schur–Padé algorithm we assumed $p \in (-1,1)$. For a general noninteger $p \in \mathbb{R}$ there are two ways to reduce the power to the interval $(-1,1)$. We can write

$$p = \lfloor p \rfloor + p_1, \qquad p_1 > 0, \tag{4.52a}$$
$$p = \lceil p \rceil + p_2, \qquad p_2 < 0, \tag{4.52b}$$

where $p_1 - p_2 = 1$. To choose between these two possibilities we will concentrate on the computation of A^{p_1} and A^{p_2} and ask which of these computations is the better

conditioned. To make the analysis tractable we assume that A is Hermitian positive definite with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n > 0$ and we use the lower bound (4.12), which is now an equality for the Frobenius norm. Using the mean value theorem, we obtain, for $p \in (-1,1)$ and $f(x) = x^p$,

$$\begin{aligned} \|L_{x^p}(A)\|_F &= \max_{i\leq j} |f[\lambda_i,\lambda_j]| = \max_{i\leq j} |f'(\xi_{ij})|, \qquad \xi_{ij} \in [\lambda_i,\lambda_j] \\ &= |f'(\lambda_n)| = |p|\lambda_n^{p-1}. \end{aligned}$$

Hence, by (4.5) for the Frobenius norm,

$$\kappa_{x^p} = \frac{|p|\lambda_n^{p-1}\|A\|_F}{\|A^p\|_F} \approx \frac{|p|\lambda_n^{p-1}\|A\|_2}{\|A^p\|_2} = \begin{cases} |p|\kappa_2(A)^{1-p}, & p \geq 0, \\ |p|\kappa_2(A), & p \leq 0, \end{cases}$$

where $\kappa_2(A) = \|A\|_2\|A^{-1}\|_2 = \lambda_1/\lambda_n$. Since $p_1 > 0$ and $p_2 < 0$, in order to minimize the lower bound we should choose p_1 if $p_1\kappa_2(A)^{1-p_1} \leq -p_2\kappa_2(A) = (1-p_1)\kappa_2(A)$, that is, if $\kappa_2(A) \geq \exp(p_1^{-1}\log\bigl(p_1/(1-p_1)\bigr))$. Thus, for example, if $p_1 \leq 0.5$ then p_1 is always chosen, while if $p_1 = 0.75$ or $p_1 = 0.99$ then p_1 is chosen for $\kappa_2(A) \geq 4.3$ and $\kappa_2(A) \geq 103.7$, respectively.

Now we consider how to handle integer p. When p is positive, A^p should be computed by binary powering [72, Alg. 4.1]. When p is negative there are several possibilities, of which we state three. We write GEPP for Gaussian elimination with partial pivoting.

Algorithm 4.14. *This algorithm computes $X = A^p$ for $p = -k \in \mathbb{Z}^-$.*

1 $Y = A^k$ by binary powering
2 $X = Y^{-1}$ via GEPP

Algorithm 4.15. *This algorithm computes $X = A^p$ for $p = -k \in \mathbb{Z}^-$.*

1 $Y = A^{-1}$ via GEPP
2 $X = Y^k$ by binary powering

Algorithm 4.16. *This algorithm computes $X = A^p$ for $p = -k \in \mathbb{Z}^-$.*

1 Compute a factorization $PA = LU$ by GEPP.
2 $X_0 = I$
3 for $i = 0\colon k-1$
4 Solve $LX_{i+1/2} = PX_i$
5 Solve $UX_{i+1} = X_{i+1/2}$
6 end
7 $X = X_k$

Algorithms 4.14 and 4.15 have the same cost. Algorithm 4.16 is more expensive as it does not take advantage of binary powering. However, our main interest is in accuracy and a full rounding error analysis is given here for these three algorithms. Both Algorithm 4.14 and 4.15 involve inverting a full matrix via GEPP. There are

several methods to do this. For example, MATLAB's `inv` function takes the following steps [70, sec. 14.3]: compute the LU factorization $PA = LU$, compute U^{-1} by back substitution and then solve for X the equation $XL = U^{-1}$. Now we assume the matrix inversion required in Algorithm 4.14 and 4.15 is implemented in this way and that, for simplicity in deriving the bounds, $P = I$.

First consider Algorithm 4.14. We write the computed A^k as $\widehat{X} = fl(A^k)$. Then we have [70, Prob. 3.10]

$$\|\widehat{X} - A^k\|_2 \le (kn^2u + O(u^2))\|A\|_2^k, \tag{4.53}$$

where u is the unit roundoff. Let $\widehat{Y}$ be the computed inverse of $\widehat{X}$ via GEPP. Recall that $\widehat{X} = LU + \Delta X$ with $\|\Delta X\|_2 \le c_n u\|L\|_2\|U\|_2$ [70, Thm. 9.3], where we write the computed LU factors as L and U. Then it follows that [70, sec. 14.3.2]

$$\|\widehat{Y} - \widehat{X}^{-1}\|_2 \le c_n n^2 u\|L\|_2\|U\|_2\|\widehat{Y}\|_2\|\widehat{X}^{-1}\|_2 =: \delta_1\|\widehat{X}^{-1}\|_2. \tag{4.54}$$

Applying the triangle inequality, it follows from (4.53) and (4.54) that

$$\begin{aligned}\|\widehat{Y} - A^{-k}\|_2 &\le \|\widehat{Y} - \widehat{X}^{-1}\|_2 + \|\widehat{X}^{-1} - A^{-k}\|_2 \\ &\le \delta_1\|\widehat{X}^{-1}\|_2 + \|A^{-k}(\widehat{X} - A^k)A^{-k}\|_2 + O(u^2) \\ &\le \delta_1\|\widehat{X}^{-1}\|_2 + kn^2u\|A^{-k}\|_2^2\|A\|_2^k + O(u^2).\end{aligned}$$

Now we get the following lemma on the rounding errors in Algorithm 4.14.

Lemma 4.17. *Let $\widehat{Y}$ be the computed A^{-k} by Algorithm* 4.14. *Denote $\widehat{X} = fl(A^k)$ and let $\widehat{X} \approx LU$ be the computed LU factorization of $\widehat{X}$ by GEPP. Then we have*

$$\|\widehat{Y} - A^{-k}\|_2 \le \delta_1\|\widehat{X}^{-1}\|_2 + kn^2u\|A^{-k}\|_2^2\|A\|_2^k + O(u^2), \tag{4.55}$$

where $\delta_1 = c_n n^2 u\|L\|_2\|U\|_2\|\widehat{Y}\|_2$.

A rounding error bound for Algorithm 4.15 is given in the following lemma, which can be proved in a similar manner as for Algorithm 4.14.

Lemma 4.18. *Let $A \approx LU$ be the computed LU factorization of A by GEPP and $\widehat{Z} \approx A^{-1}$ be the computed inverse of A. Write the computed power $\widehat{Z}^k$ as $fl(\widehat{Z}^k)$. Then the computed A^{-k} by Algorithm* 4.15 *satisfies*

$$\|fl(\widehat{Z}^k) - A^{-k}\|_2 \le \delta_2\|A^{-1}\|_2^k + kn^2u\|\widehat{Z}\|_2^k + O(u^2), \tag{4.56}$$

where $\delta_2 = c_n n^2 u\|L\|_2\|U\|_2\|\widehat{Z}\|_2$.

Now we proceed to the error analysis for Algorithm 4.16. Let $\widehat{X}_{i+1/2} = X_{i+1/2} + \Delta X_{i+1/2}$, $\widehat{X}_{i+1} = X_{i+1} + \Delta X_{i+1}$ be the computed $X_{i+1/2}$ and X_{i+1}, respectively. Assume that the solver is stable, so we have [70, sec. 9]

$$\widehat{X}_{i+1/2}U = \widehat{X}_i + F_{i+1/2}, \quad \widehat{X}_{i+1}L = \widehat{X}_{i+1/2} + R_{i+1},$$

where $\|F_{i+1/2}\| \le \alpha_n u\|\widehat{X}_{i+1/2}\|\,\|U\|$ and $\|R_{i+1}\| \le \alpha_n u\|\widehat{X}_{i+1}\|\,\|L\|$ for some constant α_n. Then $\Delta X_{i+1} = \Delta X_i U^{-1}L^{-1} + F_{i+1/2}U^{-1}L^{-1} + R_{i+1}L^{-1} + O(u^2)$ and it follows

that

$$\|\Delta X_{i+1}\| \le \|\Delta X_i\|\,\|L^{-1}\|\,\|U^{-1}\| + \|F_{i+1/2}\|\,\|L^{-1}\|\,\|U^{-1}\| + \|R_{i+1}\|\,\|L^{-1}\|. \quad (4.57)$$

So $\|\Delta X_k\|$, rounding errors in X_k, can be bounded by the recurrence $\|X_{i+1/2}\| \le \|X_i\|\,\|U^{-1}\|$ and $\|X_{i+1}\| \le \|X_{i+1/2}\|\,\|L^{-1}\|$.

The forward error bounds from the above analysis are difficult to compare and do not provide any clear guidance on the choice of algorithm. Algorithm 4.14 inverts A^k, which is potentially a much more ill conditioned matrix than A. Intuitively, Algorithm 4.15 should therefore be preferred. Algorithm 4.16 does not explicitly invert a matrix but relies on triangular solves, and triangular systems are typically solved to higher accuracy than we might expect from conditioning considerations [70, Chap. 8]. We will use numerical experiments to guide our choice (see Experiment 7 in Section 4.9).

4.7 Singular matrices

Since our aim is to develop an algorithm of the widest possible applicability, we would like to extend Algorithm 4.13 so that it handles singular matrices with a semisimple zero eigenvalue. If A is singular then the Schur factor T will be singular. We reorder T (using unitary similarities) so that it has the form

$$T = \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix} \quad (4.58)$$

where T_{11} is nonsingular and T_{22} has zero diagonal. The zero eigenvalue is semisimple if and only if $T_{22} = 0$, by rank considerations. If $T_{22} = 0$ then $U = T^p$ is given by

$$U = \begin{bmatrix} U_{11} & T_{11}^{-1}U_{11}T_{12} \\ 0 & 0 \end{bmatrix}, \qquad U_{11} = T_{11}^p. \quad (4.59)$$

The diagonal blocks in this expression follow from the fact that any primary matrix function of a block triangular matrix is block triangular [72, Thm. 1.13], while the (1,2) block is obtained from the equation $TU = UT$. The conclusion is that we should obtain U_{11} from Algorithm 4.13 and compute U_{12} separately from the given formula.

Algorithm 4.19. *This algorithm is a modification of Algorithm* 4.13 *to handle singular matrices.*

1 Apply Algorithm 4.13 with the following changes.
2 if T has any zero eigenvalues
3 Just after line 2, reorder T into the form (4.58), where T_{11} is nonsingular and T_{22} has zero diagonal.
4 if $\|T_{22}\| \ge c_n u\|T\|$ for some suitable constant c_n
5 Quit with an error message that A^p is not defined.
6 else
7 Compute U in (4.59), obtaining U_{11} using lines 3–22 of Algorithm 4.13.
8 end
9 end

Algorithm 4.19 is the starting point for a practical algorithm but is flawed in its present form. In floating point arithmetic we are unlikely to obtain exact zeros on the diagonal of T. Consider, for example, the MATLAB matrix `A = gallery(5)`, which has integer entries and a Jordan form with one 5×5 Jordan block corresponding to the eigenvalue 0. The computed triangular Schur factor T has positive diagonal entries all of order 10^{-2}. The computed square root (for example) from Algorithm 4.13 has norm of order 10^{10}. Without further computations involving "difficult rank decisions" [53, sec. 7.6.5], which would effectively be the first stages of computing the Jordan form, it is not possible to determine whether it makes sense to compute A^p with $p \notin \mathbb{Z}$ when A is singular. We will therefore not pursue the development of a practical algorithm for the singular case.

4.8 Alternative algorithms

A number of alternatives to and variations of Algorithm 4.13 can be formulated. They are based on initial reduction to Schur form, the exp-log formula (4.2), and the Schur–Parlett algorithm of Davies and Higham [37], [72, Alg. 9.6]. The Schur–Parlett algorithm is designed for computing $f(A)$ for any f for which functions of arbitrary triangular matrices can be reliably computed. It employs a reordered and partitioned Schur triangular factor, computes $f(T_{ii})$ for the diagonal blocks T_{ii} by the given method and obtains the off-diagonal blocks by the block Parlett recurrence.

We summarize the main possibilities.

(a) `schur-pade`: Algorithm 4.13.

(b) `SP-Pade`: the Schur–Parlett method using Algorithm 4.13 on the diagonal blocks T_{ii}.

(c) `SP-ss-iss`: the Schur–Parlett method with evaluation of $\exp(p\log(T_{ii}))$ by the inverse scaling and squaring method for the logarithm [72, sec. 11.5] and the scaling and squaring method for the exponential [5].

(d) `tri-ss-iss`: reduction to Schur form T with evaluation of $\exp(p\log(T))$ by the inverse scaling and squaring method for the logarithm applied to the whole matrix T and the scaling and squaring method for the exponential.

(e) `powerm`: the algorithm discussed in Section 4.1 based on an eigendecomposition, which is implemented in the MATLAB function of Figure 4.2.

Note that a variant of `tri-ss-iss` that works directly on A instead of reducing to Schur form is not competitive in cost with `tri-ss-iss`, since computing square roots of full matrices is relatively expensive [72, Chap. 6].

We make some brief comments on the relative merits of these methods.

For the methods that employ a Schur decomposition the cost will be dominated by the cost of computing the Schur decomposition unless $\|A\|$ is large. If the matrix is already triangular then `schur-pade` and `tri-ss-iss` have similar cost, and in particular require approximately the same number of square roots.

`SP-Pade` differs from `schur-pade` in that it applies Padé approximation to each diagonal block of T (possibly with a different degree for each block) rather than to T as a whole. It is possible for the partitioning to be the trivial one, $T \equiv T_{11}$, in which case `SP-Pade` and `schur-pade` are identical.

```
function X = powerm(A,p,str)
%POWERM    Arbitrary power of matrix.
%   POWERM(A,p) computes the p'th power of A for a nonsingular,
%   diagonalizable matrix A and an abritrary real number p.
%   POWERM(A,p,'nobalance') performs the computation with balancing
%   disabled in the underlying eigendecomposition.

if nargin == 3 && strcmp(str,'nobalance')
   [V,D] = eig(A,'nobalance');
else
   [V,D] = eig(A);
end
X = V*diag(diag(D).^p)/V;
```

Figure 4.2: MATLAB function `powerm`.

An advantage in cost of `SP-Pade` and `SP-ss-iss` over `schur-pade` is that large elements of T do not affect the number of square roots computed, and hence the cost, as long as they lie in the superdiagonal blocks T_{ij} of the Schur–Parlett partitioning of T.

In the next section we compare these methods numerically.

4.9 Numerical experiments

Our numerical experiments were carried out in MATLAB R2010a, for which the unit roundoff $u = 2^{-53} \approx 1.1 \times 10^{-16}$. Our implementations of `SP-Pade` and `SP-ss-iss` are obtained by modifying the MATLAB function `funm`. For all methods except `powerm` we evaluate powers of 2×2 triangular matrices directly, using the formula (4.51).

Relative errors are measured in the Frobenius norm. For the "exact" solution we take the matrix computed using `powerm` at 100 digit precision with the VPA arithmetic of the Symbolic Math Toolbox; thus we can compute relative errors only when A is diagonalizable.

When $q = 1/p$ is an integer, another measure of the quality of a computed solution X is its relative residual,

$$\rho(X) = \frac{\|A - X^q\|}{\|X\|\,\eta(X)},$$

where $\eta(X) = \left\|\sum_{i=0}^{q-1} \left(X^{q-1-i}\right)^T \otimes X^i\right\|$ if $p > 0$ and $\eta(X) = \left\|\sum_{i=1}^{-q} \left(X^{-i}\right)^T \otimes X^{i+q-1}\right\|$ if $p < 0$, with $\otimes$ denoting the Kronecker product. This is a more practically useful definition of relative residual than $\|A - X^q\|/\|X^q\|$, as explained in [60], [72, Prob. 7.16].

Experiment 1. We computed the pth power of the matrix

$$A(\epsilon) = \begin{bmatrix} 1 & 1 \\ 0 & 1+\epsilon \end{bmatrix}, \tag{4.60}$$

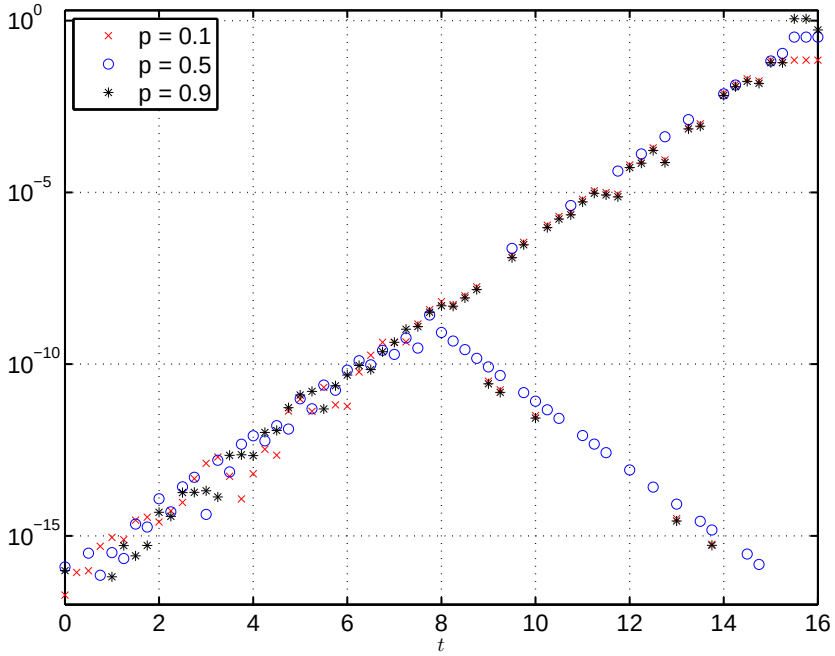

Figure 4.3: Experiment 1: relative errors for `powerm` on matrix (4.60) with $\epsilon = 10^{-t}$.

for $p \in \{0.1, 0.5, 0.9\}$ and $\epsilon = 10^{-t}$ with 65 equally spaced values of $t \in [0, 16]$. The condition number $\kappa_{x^p}(A(\epsilon))$ is of order 1 for all these ϵ and p. The relative errors for `powerm` are shown in Figure 4.3. Clearly, the errors deteriorate as t increases and $A(\epsilon)$ approaches a defective matrix; the reason for the "bifurcation" in the error curves is not clear. The other methods defined in Section 4.8 all produce results with relative error less than $4u$ in all cases.

Experiment 2. In this experiment we formed 50 random 50×50 matrices with elements from the normal (0,1) distribution; any matrix with an eigenvalue on $\mathbb{R}^-$ was discarded and another random matrix generated. Then we reduced A to Hessenberg form using the MATLAB function `hess` and computed $A^{1/3}$ by all five methods as well as by `powerm_nb`, the latter denoting `powerm` with the `'nobalance'` argument, which inhibits the use of balancing in the eigendecomposition. The results, with 2-norms used in the residuals, are shown in Figure 4.4. The improved performance of `powerm_nb` over `powerm` shows that it is the balancing that is affecting the numerical stability of `powerm` in this example. This is not surprising, because Watkins [129] has pointed out that for upper Hessenberg matrices balancing can seriously degrade accuracy in the eigendecomposition and should not be automatically used.

We note that using `powerm_nb` in place of `powerm` makes no difference to the results in Experiment 1, as balancing has no effect in that example.

Experiment 3. In this experiment we use a selection of 10×10 nonsingular matrices taken from the MATLAB `gallery` function and from the Matrix Computation Toolbox [66]. Any matrix found to have an eigenvalue on $\mathbb{R}^-$ was squared. We computed A^p for $p \in \{1/52,\ 1/12,\ 1/3,\ 1/2\}$, these values being ones likely to occur in applications where roots of transition matrices are required [72, sec. 2.3], [76], as well as the negatives of these values. This gives 376 problems in total. We omit `tri-ss-iss` from this test, as it is generally outperformed by `SP-ss-iss` (as can

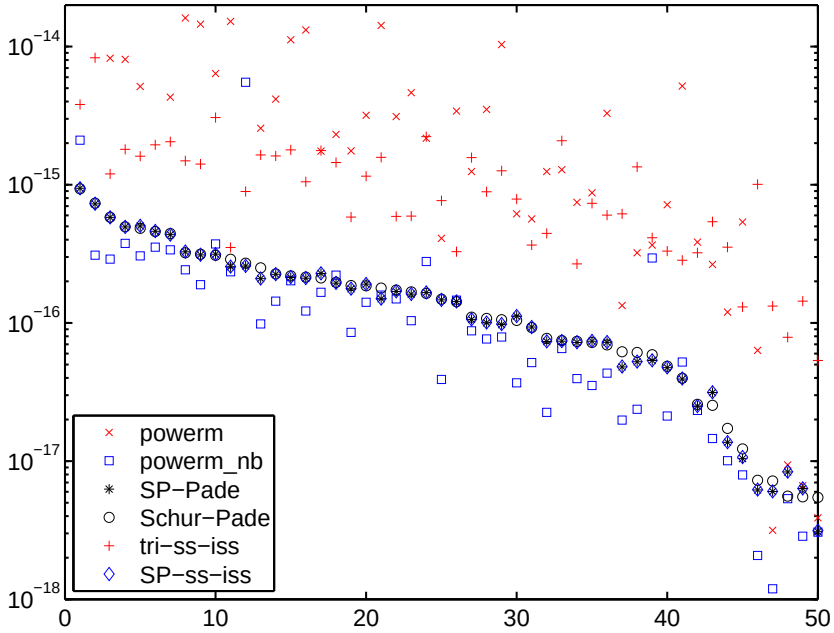

Figure 4.4: Experiment 2: relative residuals for 50 random Hessenberg matrices.

be seen in Experiment 2). Figures 4.5 and 4.7 show the relative errors and relative residuals. The solid line in Figure 4.5 is $\kappa_{x^p}(A)u$, where κ_{x^p} is computed via (4.7) and (4.9) using codes from the Matrix Function Toolbox [67] that compute $K_{\exp}$ and $K_{\log}$; the problems are sorted by decreasing condition number. Figures 4.6 and 4.8 show performance profiles. A performance profile shows the proportion π of problems where the performance ratio of a method is at most α, where the performance ratio for a method on a problem is the error or residual of that method divided by the smallest error or residual over all the methods. The errors and residuals lead to the same conclusions. First, `powerm` often produces very good results but is sometimes very unstable. Second, `schur-pade` `SP-Pade` and `SP-ss-iss` perform similarly, with `schur-pade` having a slight edge overall.

Experiment 4. This experiment is identical to the previous one except that we use the upper triangular QR factor R of each matrix and replace every negative diagonal element of R by its absolute value. The errors and residuals and their performance profiles are shown in Figures 4.9–4.12. For this class of matrices `schur-pade` is clearly greatly superior to the other methods. The performance profiles are qualitatively similar if we use the Schur factor instead of the QR factor.

Experiment 5. In this experiment we compute the three bounds in (4.11), (4.12) as well as the true norm of the Fréchet derivative $\|L_{x^p}(A)\|$ for the same matrices and values of p as in Experiment 3, using the Frobenius norm. The computed upper bound, which sometimes overflowed, was set to the minimum of 10^{30} and itself. The results are plotted in Figure 4.13. The results show that the lower bounds are sharper than the upper bounds and that they are often correct to within a couple of orders of magnitude, being less reliable for the very ill conditioned problems.

Experiment 6. In this experiment, we test our proposed choice of the fractional part of p when $p \notin [-1, 1]$. For $\kappa_2(A)$ we use the lower bound $\max_i |t_{ii}| / \min_i |t_{ii}|$

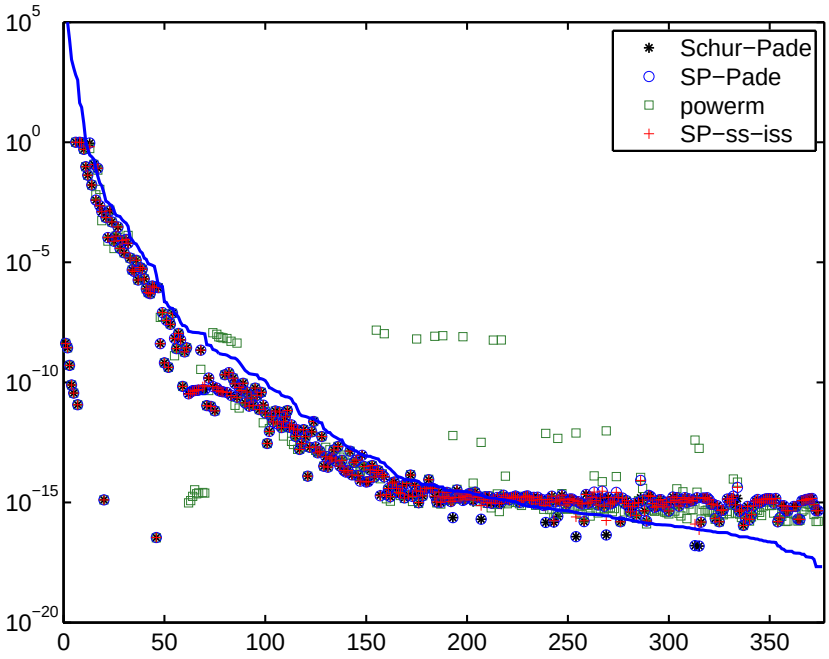

Figure 4.5: Experiment 3: relative errors for a selection of 10×10 matrices and several p.

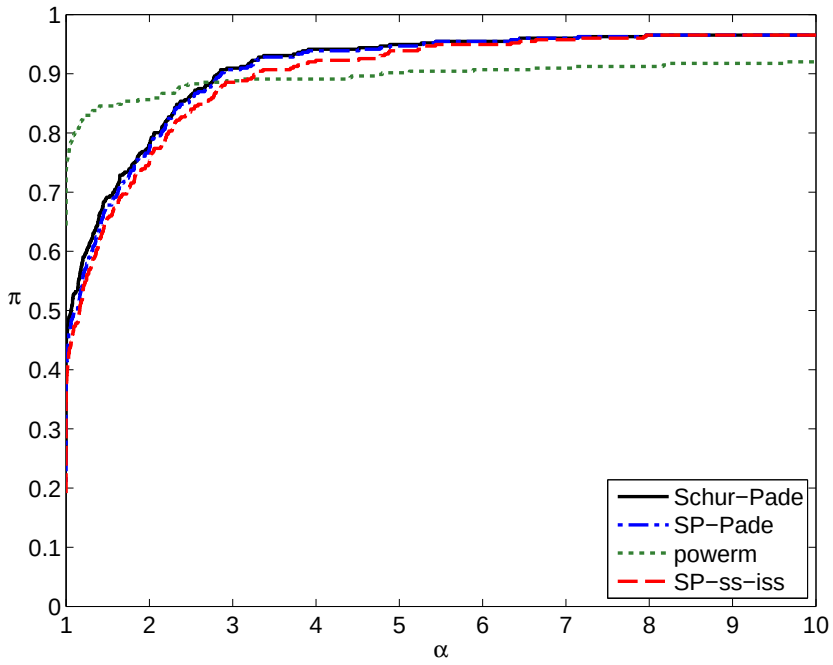

Figure 4.6: Experiment 3: performance profile of relative errors.

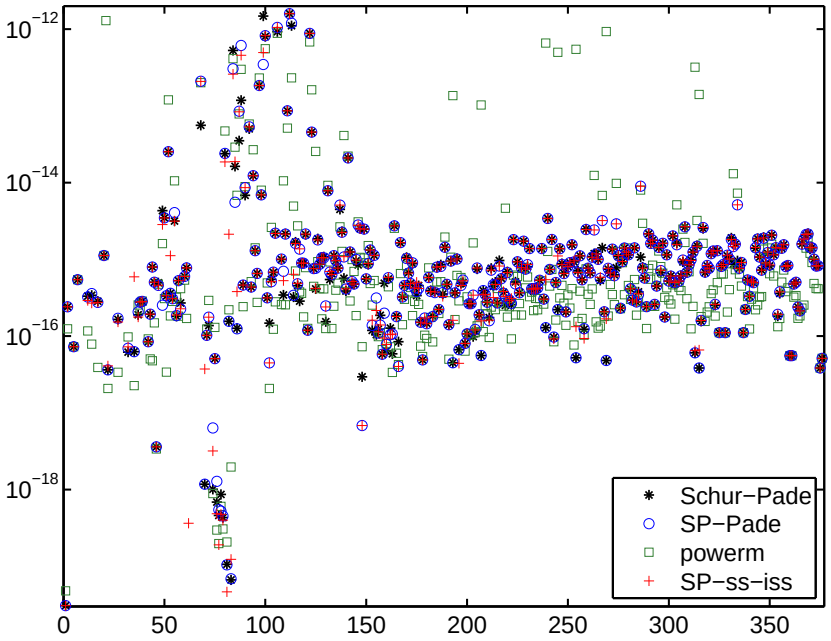

Figure 4.7: Experiment 3: relative residuals for a selection of 10×10 matrices and several p.

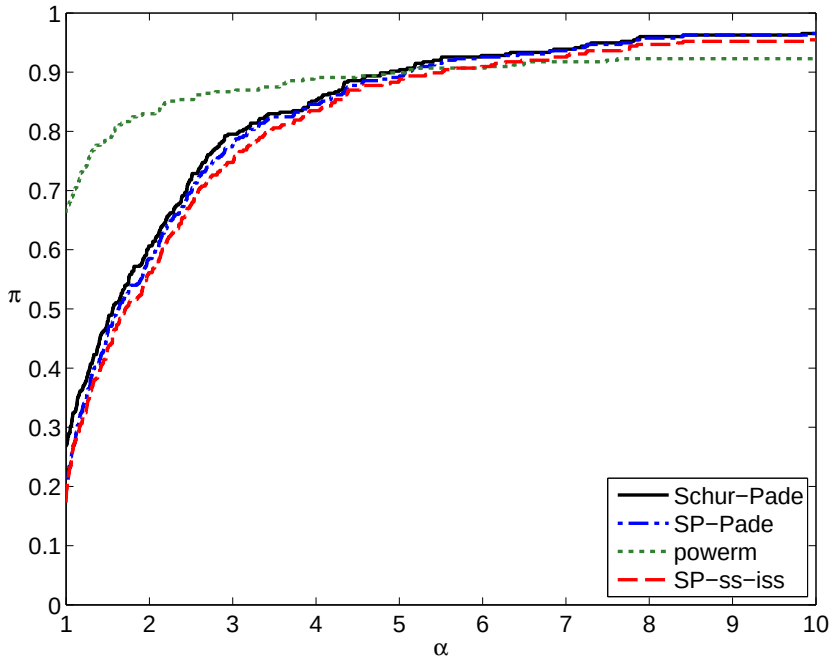

Figure 4.8: Experiment 3: performance profile of relative residuals.

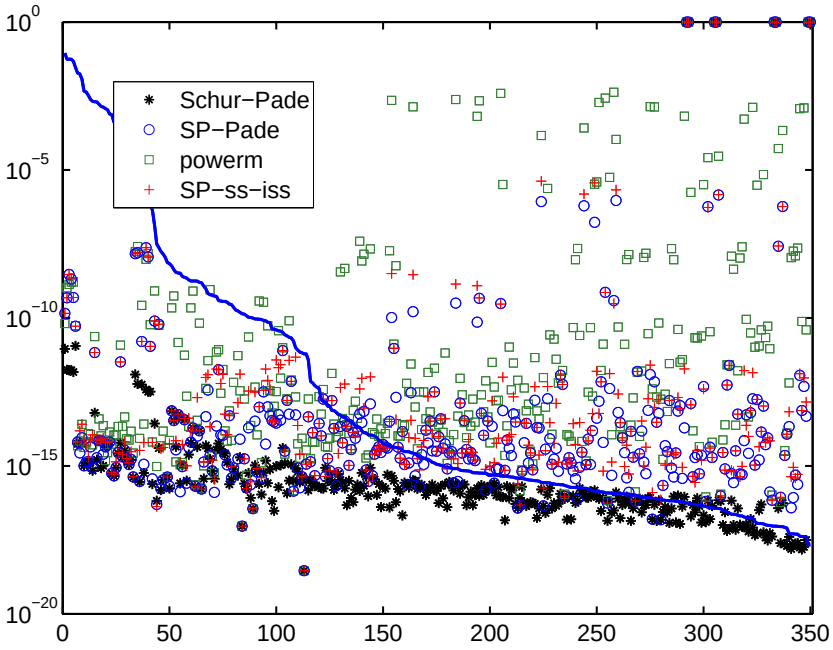

Figure 4.9: Experiment 4: relative errors for a selection of 10×10 triangular matrices and several p.

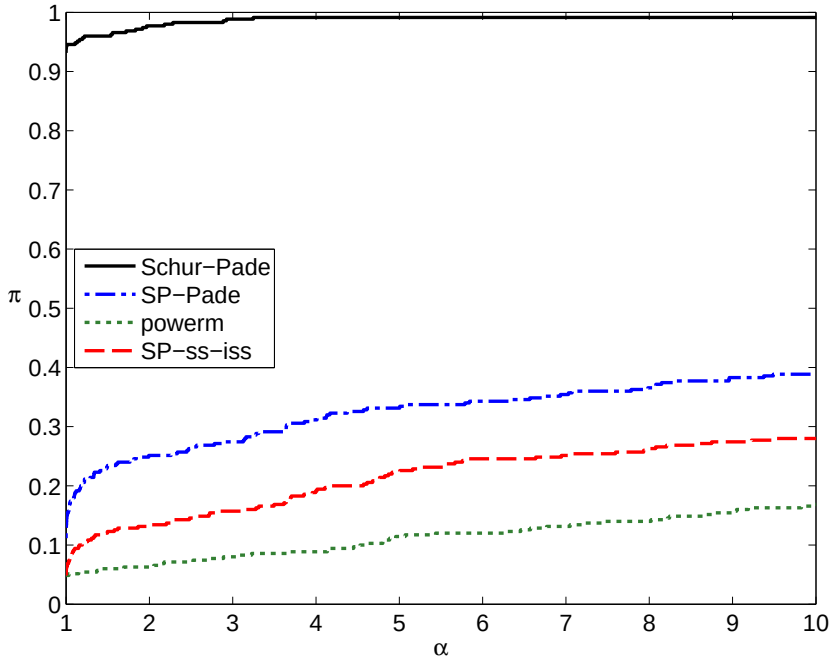

Figure 4.10: Experiment 4: performance profile of relative errors.

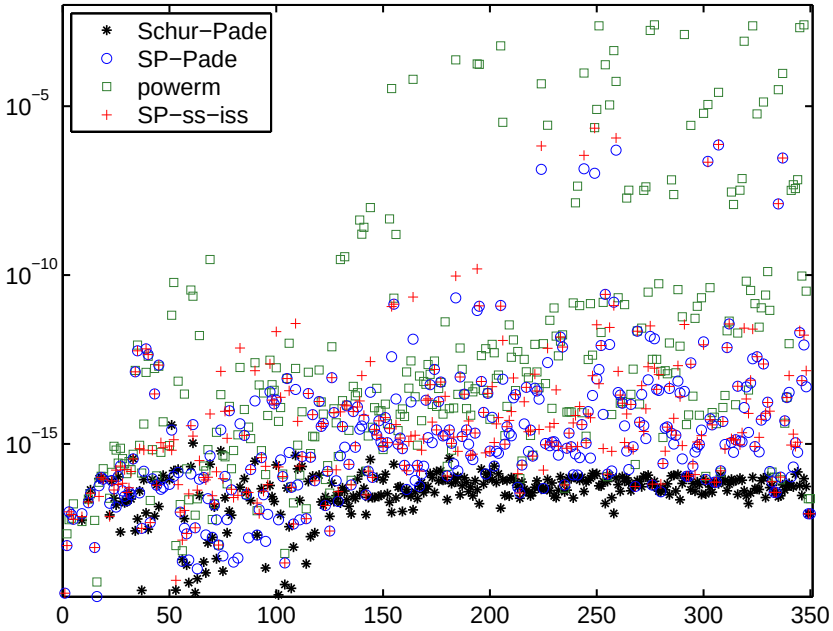

Figure 4.11: Experiment 4: relative residuals for a selection of 10×10 triangular matrices and several p.

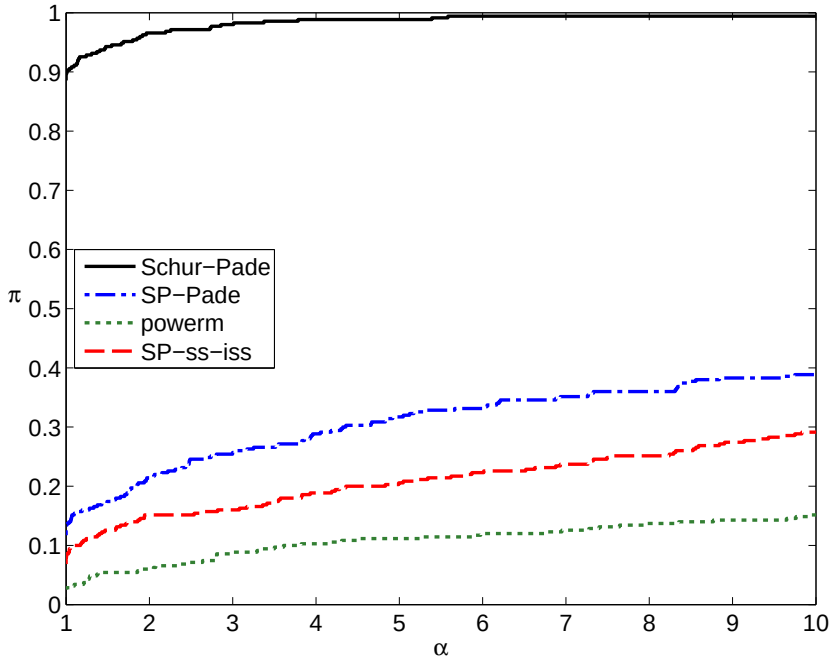

Figure 4.12: Experiment 4: performance profile of relative residuals.

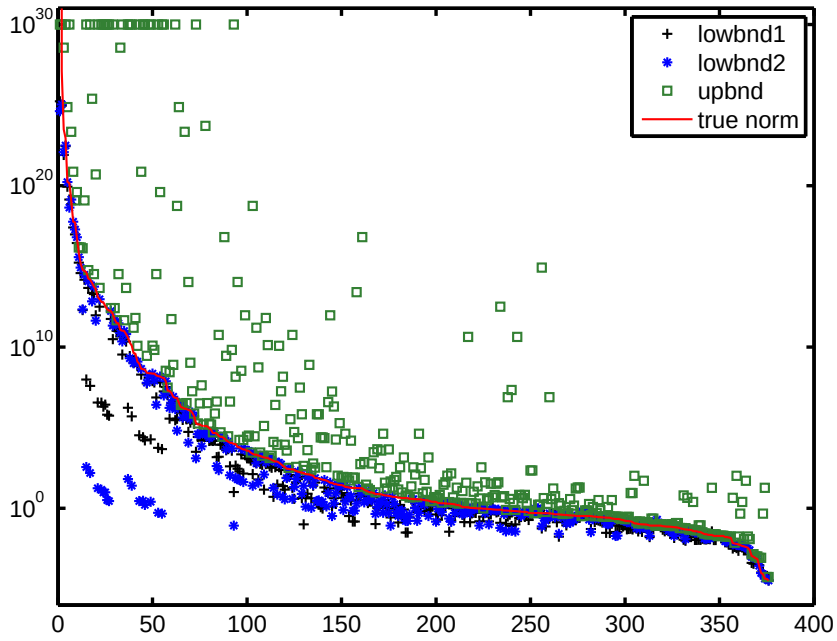

Figure 4.13: Experiment 5: the lower bounds `lowbnd1` in (4.11) and `lowbnd2` in (4.12), the upper bound `upbnd` in (4.12), and the true norm $\|L_{x^p}(A)\|_F$, for the matrices in Experiment 3.

in the prescription of Section 4.6, where T is the triangular Schur factor. We use the same matrices as in Experiment 3 and compute A^p for $p = 3.9, 3.7, 3.3, 3.1$. The performance profiles of the relative errors are shown in Figure 4.14. Our strategy chose p_1 in 169 of the 197 cases in this experiment. Indeed, p_1 is almost as good a choice as the "optimal" choice, as can be seen in two ways. First, the performance profile curve for p_1 is almost indistinguishable from that for the "optimal" choice and so is omitted from the figure. Second, the maximum and minimum values of the relative error for p_1 divided by that for p_2 were 3.2 and 1.3×10^{-16}, respectively.

Experiment 7. In this final experiment we compare Algorithms 4.14, 4.15, and 4.16, all of which compute A^p where $p = -k$ is a negative integer. We test the algorithms on the same set of matrices as in Experiment 3 for $p = -3, -5, -7, -9$. The results are shown in Figures 4.15 and 4.16. Algorithms 4.15 and 4.16 clearly produce much more accurate results than Algorithm 4.14, as we expected. There is little to choose between Algorithms 4.15 and 4.16; we favour the former in view of its lower computational cost.

4.10 Concluding remarks

We have derived a new algorithm (Algorithm 4.13) for computing arbitrary powers A^p of a matrix, based on diagonal Padé approximants of $(1-x)^p$ and the Schur decomposition. The algorithm performs in a generally numerically stable fashion in our tests, with relative error usually less than the product of the condition number of the problem and the unit roundoff. Our experiments demonstrate the superiority of this approach over alternatives based on separate approximation of the exponential

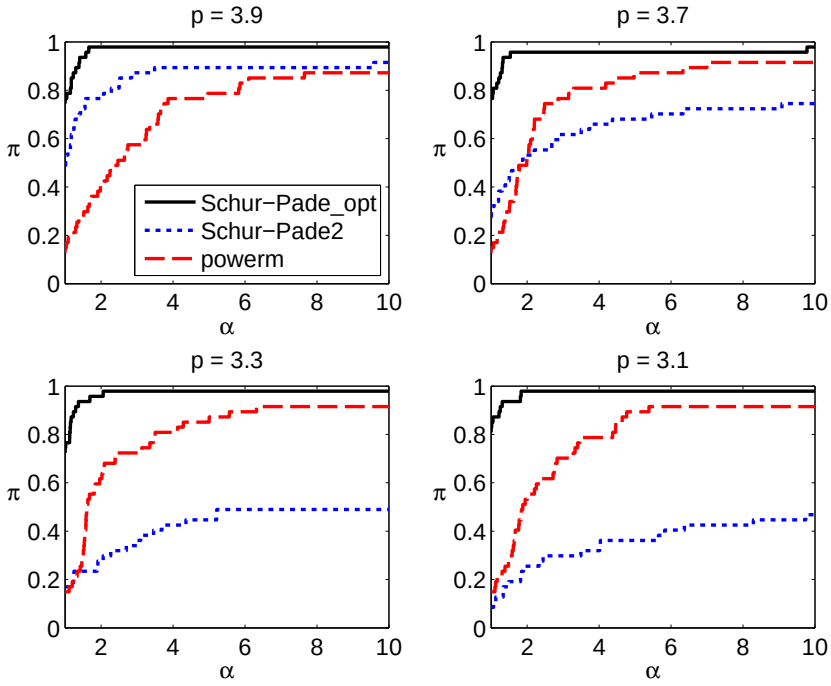

Figure 4.14: Experiment 6: performance profile of relative errors. The legend for first plot applies to all four plots. `Schur-Pade2` uses p_2 in (4.52b) and `Schur-Pade_opt` uses the choice defined in Section 4.6.

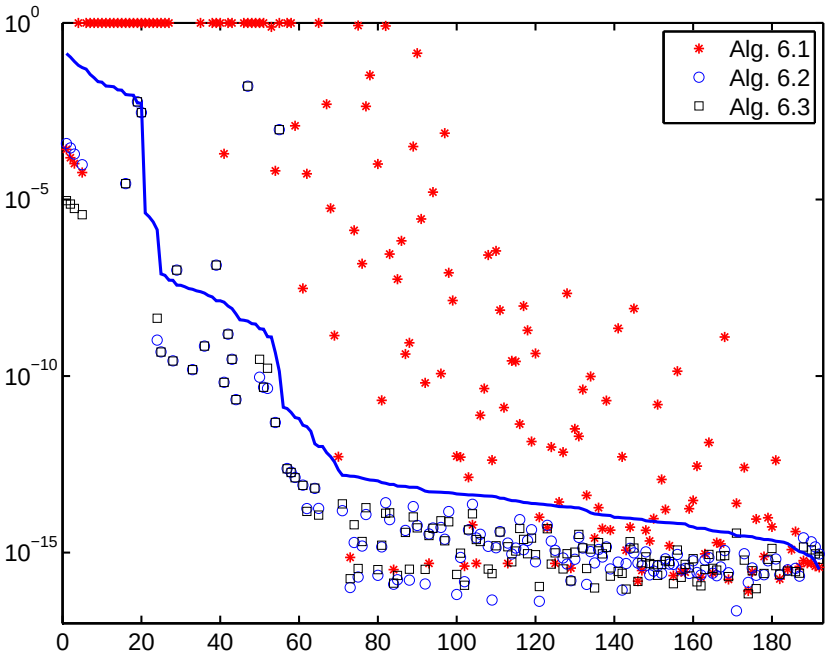

Figure 4.15: Experiment 7: relative errors for Algorithms 4.14, 4.15, and 4.16 for a selection of 10×10 matrices and several negative integers p.

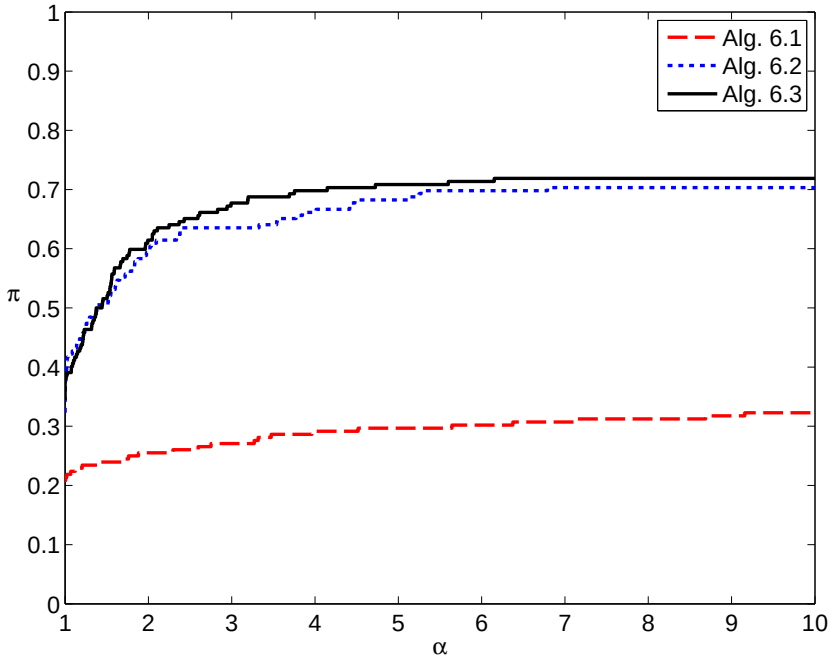

Figure 4.16: Experiment 7: performance profile of relative errors .

and logarithm in the formula $A^p = \exp(p \log(A))$ using the best available methods. The use of Algorithm 4.13 within the Schur–Parlett algorithm (to compute T_{ii}^p for the diagonal blocks T_{ii} of the blocked and re-ordered triangular Schur factor) merits consideration as it is generally faster than applying it to the whole T, but Algorithm 4.13 is significantly more accurate in our tests with triangular matrices (Experiment 4).

MATLAB has a built-in function `mpower` for which the function call `mpower(A,p)` is equivalent to the syntax `A^p`. In our tests with MATLAB R2010a, `mpower` performs identically to our `powerm` function for noninteger p, and in particular performs badly on matrices that are defective or nearly defective. For negative integer p, `mpower` performs identically to Algorithm 4.14 in our tests.

Chapter 5

Conclusions and Future Work

On the problem of roots of stochastic matrices, we started with a careful treatment of the underlying theory where we have used the theory of matrix functions to develop tools for analyzing the existence of stochastic roots of stochastic matrices. We have identified two classes of stochastic matrices for which the principal pth root is stochastic for all p and demonstrated a wide variety of possibilities for existence and uniqueness. We have also given some necessary spectral conditions for existence.

On the computational side, we emphasized finding an approximate stochastic root by solving the nonlinear programming problem of minimizing the residual $\|X^p - A\|_F$. A spectral projected gradient method starting with the perturbed principal root is found efficient in the sense of the computational time and the final residual.

We also considered a more general problem of matrix powers A^α where $A \in \mathbb{C}^{n\times n}$ and α is an arbitrary real number. We have derived a new algorithm for computing A^α based on diagonal Padé approximants of $(1-x)^\alpha$ and the Schur decomposition. The algorithm performs in a generally numerically stable fashion in our tests and shows its superiority over alternatives based on separate approximation of the exponential and logarithm in the formula $A^\alpha = \exp(\alpha \log(A))$ using the best available methods and that based on the Schur–Parlett algorithm with our new algorithm applied to the diagonal blocks.

The problem of the existence of stochastic roots is still open. We have not yet given a full characterization of all stochastic matrices that have stochastic pth roots for a given p. One of the problems that is closely related to the stochastic roots problem is the inverse eigenvalue problem that determines conditions under which a set of n complex numbers comprises the eigenvalues of some $n \times n$ stochastic matrix (which is called the inverse spectrum problem by Minc [106]). Different from the necessary condition derived in Section 2.5.2 where we check whether each eigenvalue of A is the eigenvalue of some pth power of a stochastic matrix, a refined necessary condition can be derived by checking whether every eigenvalue of A is an eigenvalue of the pth power of the same stochastic matrix. This can be done with a full understanding of the inverse spectrum problem. Though it has been completely solved for the 3×3 case, the inverse spectrum problem for stochastic matrices with a set of arbitrary n complex numbers remains open. Note that deriving a necessary and sufficient condition for the existence of stochastic roots can be quite difficult since the nonprimary roots of a derogatory matrix can not be identified from its spectrum alone.

It is worthwhile to be aware of a more general setting of functions preserving nonnegativity of matrices. Bharali and Holtz [13] characterize entire functions $f(A)$ that preserve nonnegativity of two classes of structured matrices: triangular and block-triangular matrices and circulant matrices. One can consider the characterizations of matrix functions (which may not be entire functions) that preserve nonnegativity of matrices with or without certain structures. For a specific matrix function $f(A)$, the conditions under which $f(A)$ preserves the nonnegativity of A could also be looked at.

About the computational matter of finding an approximate stochastic root, since the methods currently considered to minimize $\|X^p - A\|$ can only guarantee a local minimum, one can consider the global optimization techniques, for example, the multilevel coordinate search currently used in the NAG Toolbox for MATLAB [3] and the genetic algorithm and pattern search used in Global Optimization Toolbox [2].

Finally, a more general class of functions that arise in the applications of fractional differential equations is the *Mittag-Leffler function* defined by

$$E_{k_1,k_2}(z) := \sum_{j=0}^{\infty} \frac{z^j}{\Gamma(jk_1 + k_2)}, \quad k_1, k_2 > 0, \tag{5.1}$$

whenever the series converges. These functions are of fundamental importance in the analysis of fractional differential equations [40, Chap. 4], [62]. Note that Mittag-Leffler functions are generalizations of the ψ functions defined by $\psi_k = \sum_{j=0}^{\infty} z^j/(j+k)!$, which are closely related to the exponential:

$$\psi_0(z) = e^z, \quad \psi_1(z) = \frac{e^z - 1}{z}, \quad \psi_2(z) = \frac{e^z - 1 - z}{z^2}, \ldots.$$

We have $E_{1,k}(z) = \psi_k(z)$ for integers $k > 0$. The need to evaluate Mittag-Leffler functions at a matrix argument arises. Recall that the evaluation of ψ_k, $k = 0, 1, \ldots$, at a matrix argument can be done via an analogue of the scaling and squaring method for the matrix exponential [72, sec. 10.7.4]. Even for scalar arguments it is nontrivial to evaluate $E_{k_1,k_2}(z)$ accurately, on which some work has been done based on the integral representation of $E_{k_1,k_2}(z)$ [39], [56], [117]. However, no methods have yet been proposed on evaluating Mittag-Leffler functions at matrix arguments.

Bibliography

[1] Moody's Investors Service. `http://www.moodys.com/`.

[2] Global Optimization Toolbox. MathWorks. `http://www.mathworks.com/products/global-optimization/index.html`.

[3] NAG Toolbox for MATLAB. NAG Ltd., Oxford. `http://www.nag.co.uk/`.

[4] Milton Abramowitz and Irene A. Stegun, editors. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover Publications Inc., New York, 1992. Reprint of the 1972 edition.

[5] Awad H. Al-Mohy and Nicholas J. Higham. A new scaling and squaring algorithm for the matrix exponential. *SIAM J. Matrix Anal. Appl.*, 31(3):970–989, 2009.

[6] George E. Andrews, Richard Askey, and Ranjan Roy. *Special Functions*. Cambridge University Press, 2000.

[7] Anatoliy Antonov and Yanka Yanakieva. Transition matrix generation. In *CompSysTech 04: Proceedings of the 5th international conference on Computer systems and technologies*, pages 1–6, New York, NY, USA, 2004. ACM.

[8] Mario Arioli and Daniel Loghin. Discrete interpolation norms with applications. *SIAM J. Matrix Anal. Appl.*, 47(4):2924–2951, 2009.

[9] George A. Baker, Jr. *Essentials of Padé Approximants*. Academic Press, New York, 1975.

[10] George A. Baker, Jr. and Peter Graves-Morris. *Padé Approximants*, volume 59 of *Encyclopedia of Mathematics and Its Applications*. Cambridge University Press, second edition, 1996.

[11] J. R. Beck and S. G. Pauker. The Markov process in medical prognosis. *Medical Decision Making*, 3(4):419–458, 1983.

[12] Abraham Berman and Robert J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1994. Corrected republication, with supplement, of work first published in 1979 by Academic Press.

[13] Gautam Bharali and Olga Holtz. Functions preserving nonnegativity of matrices. *SIAM J. Matrix Anal. Appl.*, 30(1):84–101, 2008.

[14] Dario A. Bini, Nicholas J. Higham, and Beatrice Meini. Algorithms for the matrix pth root. *Numerical Algorithms*, 39(4):349–378, 2005.

[15] Ernesto G. Birgin, José Mario Martínez, and Marcos Raydan. Nonmonotone spectral projected gradient methods on convex sets. *SIAM J. Optim.*, 10(4):1196–1211, 2000.

[16] Ernesto G. Birgin, José Mario Martínez, and Marcos Raydan. Algorithm 813: SPG—Software for convex-constrained optimization. *ACM Trans. Math. Software*, 27(3):340–349, 2001.

[17] Åke Björck and Sven Hammarling. A Schur method for the square root of a matrix. *Linear Algebra Appl.*, 52/53:127–140, 1983.

[18] M. Bladt and M. Sørensen. Statistical inference for discretely observed Markov jump processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(3):395–410, 2005.

[19] M. Bladt and M. Sørensen. Efficient estimation of transition rates between credit ratings from observations at discrete time points. *Quantitative Finance*, 9(2):147–160, 2009.

[20] N. A. Bobylev, S. A. Ivanenko, and I. G. Ismailov. Several remarks on homeomorphic mappings. *Mathematical Notes*, 60(4):442–445, 1996.

[21] F. Bonhoure, Y. Dallery, and W. J. Stewart. On the use of periodicity properties for the efficient numerical solution of certain Markov chains. *Numerical Linear Algebra with Applications*, 1(3):265–286, 1994.

[22] S. Borg, U. Persson, T. Jess, O. Ø. Thomsen, T. Ljung, L. Riis, and P. Munkholm. A maximum likelihood estimator of a Markov model for disease activity in Crohn's disease and ulcerative colitis for annually aggregated partial observations. *Medical Decision Making*, 30(1):132–142, 2010.

[23] Rüdiger Borsdorf, Nicholas J. Higham, and Marcos Raydan. Computing a nearest correlation matrix with factor structure. *SIAM J. Matrix Anal. Appl.*, 31(5):2603–2622, 2010.

[24] D. Calvetti, E. Gallopoulos, and L. Reichel. Incomplete partial fractions for parallel evaluation of rational matrix functions. *J. Comput. Appl. Math.*, 59:349–380, 1995.

[25] Philippe Carete. Characterizations of embeddable 3×3 stochastic matrices with a negative eigenvalue. *New York J. Math.*, 1:120–129, 1995.

[26] Theodore Charitos, Peter R. de Waal, and Linda C. van der Gaag. Computing short-interval transition matrices of a discrete-time Markov chain from partially observed data. *Statistics in Medicine*, 27:905–921, 2008.

[27] Mei Q. Chen, Lixing Han, and Michael Neumann. On single and double Soules matrices. *Linear Algebra Appl.*, 416:88–110, 2006.

[28] Sheung Hun Cheng, Nicholas J. Higham, Charles S. Kenney, and Alan J. Laub. Approximating the logarithm of a matrix to specified accuracy. *SIAM J. Matrix Anal. Appl.*, 22(4):1112–1125, 2001.

[29] K. L. Chung. *Markov Chains with Stationary Transition Probabilities*. Springer-Verlag, second edition, 1967.

[30] A. R. Collar. The first fifty years of aeroelasticity. *Aerospace (Royal Aeronautical Society Journal)*, 5:12–20, February 1978.

[31] Peter Congdon. *Bayesian Statistical Modelling*. Wiley Series in Probability and Statistics. Wiley, Chichester, UK, second edition, 2006.

[32] B. A. Craig and P. P. Sendi. Estimation of the transition matrix of a discrete-time Markov chain. *Health Economics*, 11(1):33–42, 2002.

[33] D. T. Crommelin and E. Vanden-Eijnden. Fitting timeseries by continuous-time Markov chains: A quadratic programming approach. *Journal of Computational Physics*, 217(2):782–805, 2006.

[34] James R. Cuthbert. On uniqueness of the logarithm for Markov semi-groups. *J. London Math. Soc.*, 4:623–630, 1972.

[35] James R. Cuthbert. The logarithmic function for finite-state Markov semi-groups. *J. London Math. Soc.*, 6:524–532, 1973.

[36] E. B. Davies. Embeddable Markov matrices. *Electronic Journal of Probability*, 15:1474–1486, 2010.

[37] Philip I. Davies and Nicholas J. Higham. A Schur–Parlett algorithm for computing matrix functions. *SIAM J. Matrix Anal. Appl.*, 25(2):464–485, 2003.

[38] Philip J. Davis. *Circulant Matrices*. Wiley, New York, 1979.

[39] K. Diethelm, N. J. Ford, A. D. Freed, and Y. Luchko. Algorithms for the fractional calculus: a selection of numerical methods. *Computer Methods in Applied Mechanics and Engineering*, 194(5):743–773, 2005.

[40] Kai Diethelm. *The Analysis of Fractional Differential Equations*. Lecture Notes in Mathematics. Springer-Verlag, Berlin, 2010.

[41] Elizabeth D. Dolan and Jorge J. Moré. Benchmarking optimization software with performance profiles. *Math. Programming*, 91:201–213, 2002.

[42] J. C. Dunn. Global and asymptotic convergence rate estimates for a class of projected gradient processes. *SIAM J. Control Optim.*, 19:368–400, 1981.

[43] G. Elfving. Zur theorie der markoffschen ketten. *Acta Social Sci. Fennicae n.*, A.2.(8):1–17, 1937.

[44] L. Elsner, R. Nabben, and M. Neumann. Orthogonal bases that lead to symmetric nonnegative matrices. *Linear Algebra Appl.*, 113:93–112, 1986.

[45] Miroslav Fiedler and Hans Schneider. Analytic functions of M-matrices and generalizations. *Linear and Multilinear Algebra*, 13:185–201, 1983.

[46] Simone Fiori. Leap-frog-type learning algorithms over the Lie group of unitary matrices. *Neurocomputing*, 71(10-12):2224–2244, 2008.

[47] J. Fortiana and C. M. Cuadras. A family of matrices, the discretized Brownian bridge, and distance-based regression. *Linear Algebra Appl.*, 264:173–188, 1997.

[48] Halina Frydman. The embedding problem for Markov chains with three states. *Math. Proc. Cambridge Philos. Soc.*, 87(2):285–294, 1980.

[49] B. Fuglede. On the imbedding problem for stochastic and doubly stochastic matrices. *Probability Theory and Related Fields*, 80:241–260, 1988.

[50] F. Gebali. *Analysis of Computer and Communication Networks*. Springer Verlag, 2008.

[51] E. Ghysels. On the periodic structure of the business cycle. *Journal of Business & Economic Statistics*, 12(3):289–298, 1994.

[52] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman and Hall, 1996.

[53] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, MD, USA, third edition, 1996.

[54] Gene H. Golub and John H. Welsch. Calculation of Gauss quadrature rules. *Math. Comp.*, 23:221–230, 1969.

[55] G. S. Goodman. An intrinsic time for non-stationary finite Markov chains. *Z. Wahrscheinlichkeitstheorie*, 16:165–180, 1970.

[56] Rudolf Gorenflo, Joulia Loutchko, and Yuri Luchko. Computation of the Mittag-Leffler funtion $E_{\alpha,\beta}(z)$ and its derivative. *Fract. Calc. Appl. Anal.*, 5(4):419–518, 2002. Erratum: Frac. Calc. Appl. Anal., 6(1):111–112, 2003.

[57] Federico Greco and Bruno Iannazzo. A binary powering Schur algorithm for computing primary matrix roots. *Numerical Algorithms*, 55(1):59–78, 2010.

[58] Geoffrey R. Grimmett and David R. Stirzaker. *Probability and Random Processes*. Oxford University Press, New York, third edition, 2001.

[59] Chun-Hua Guo. On Newton's method and Halley's method for the principal pth root of a matrix. *Linear Algebra Appl.*, 432:1905–1922, 2010.

[60] Chun-Hua Guo and Nicholas J. Higham. A Schur–Newton method for the matrix pth root and its inverse. *SIAM J. Matrix Anal. Appl.*, 28(3):788–804, 2006.

[61] Ernst Hairer, Christian Lubich, and Gerhard Wanner. *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*. Springer-Verlag, Berlin, 2002.

[62] H. J. Haubold, A. M. Mathai, and R. K. Saxena. Mittag-Leffler functions and their applications. Technical report, 2009. Available from: `http://arxiv.org/abs/0909.0230v2`.

[63] Michiel Hazewinkel, editor. *Encyclopaedia of Mathematics*, volume 9. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1993.

[64] Qi-Ming He and Eldon Gunn. A note on the stochastic roots of stochastic matrices. *Journal of Systems Science and Systems Engineering*, 12:210–223, 2003.

[65] Desmond J. Higham and Nicholas J. Higham. *MATLAB Guide*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, second edition, 2005.

[66] Nicholas J. Higham. The Matrix Computation Toolbox. `http://www.ma.man.ac.uk/~higham/mctoolbox`.

[67] Nicholas J. Higham. The Matrix Function Toolbox. `http://www.ma.man.ac.uk/~higham/mftoolbox`.

[68] Nicholas J. Higham. Computing real square roots of a real matrix. *Linear Algebra Appl.*, 88/89:405–430, 1987.

[69] Nicholas J. Higham. Evaluating Padé approximants of matrix logarithm. *SIAM J. Matrix Anal. Appl.*, 22(4):1126–1135, 2001.

[70] Nicholas J. Higham. *Accuracy and Stability of Numerical Algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, second edition, 2002.

[71] Nicholas J. Higham. The scaling and squaring method for the matrix exponential revisited. *SIAM J. Matrix Anal. Appl.*, 26(4):1179–1193, 2005.

[72] Nicholas J. Higham. *Functions of Matrices: Theory and Computation*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2008.

[73] Nicholas J. Higham. The scaling and squaring method for the matrix exponential revisited. *SIAM Rev.*, 51(4):747–764, December 2009.

[74] Nicholas J. Higham and Awad H. Al-Mohy. Computing matrix functions. *Acta Numerica*, 19(1):159–208, 2010.

[75] Nicholas J. Higham and Sheung Hun Cheng. Modifying the inertia of matrices arising in optimization. *Linear Algebra Appl.*, 275–276:261–279, 1998.

[76] Nicholas J. Higham and Lijing Lin. On pth roots of stochastic matrices. *Linear Algebra Appl.*, In Press, 2010. DOI: 10.1016/j.laa.2010.04.007.

[77] Nicholas J. Higham, D. Steven Mackey, Niloufer Mackey, and Françoise Tisseur. Functions preserving matrix groups and iterations for the matrix square root. *SIAM J. Matrix Anal. Appl.*, 26(3):849–877, 2005.

[78] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, UK, 1985.

[79] Bruno Iannazzo. On the Newton method for the matrix pth root. *SIAM J. Matrix Anal. Appl.*, 28(2):503–523, 2006.

[80] Bruno Iannazzo. A family of rational iterations and its application to the computation of the matrix pth root. *SIAM J. Matrix Anal. Appl.*, 30(4):1445–1462, 2008.

[81] M. Ilić, I. W. Turner, and D. P. Simpson. A restarted Lanczos approximation to functions of a symmetric matrix. *IMA J. Numer. Anal.*, 30(4):1044–1061, 2010.

[82] Ilse Ipsen. Private communication, May 15, 2008.

[83] Arieh Iserles, Hans Z. Munthe-Kaas, Syvert P. Nørsett, and Antonella Zanna. Lie-group methods. *Acta Numerica*, 9:215–365, 2000.

[84] Arieh Iserles and Antonella Zanna. Efficient computation of the matrix exponential by generalized polar decompositions. *SIAM J. Numer. Anal.*, 42(5):2218–2256, 2005.

[85] Robert B. Israel, Jeffrey S. Rosenthal, and Jason Z. Wei. Finding generators for Markov chains via empirical transition matrices, with applications to credit ratings. *Mathematical Finance*, 11(2):245–265, 2001.

[86] A. Iwanik and R. Shiflett. The root problem for stochastic and doubly stochastic operators. *Journal of Mathematical Analysis and Applications*, 113:93–112, 1986.

[87] S. Johansen. Some results on the imbedding problem for finite Markov chains. *J. London Math. Soc.*, 8(2):345–351, 1974.

[88] Søren Johansen and Fred L. Ramsey. A bang-bang representation for 3×3 embeddable stochastic matrices. *Z. Wahrsch. Verw. Gebiete*, 47(1):107–118, 1979.

[89] Matthew T. Jones. Estimating Markov transition matrices using proportions data: an application to credit risk. *IMF Working Paper*, pages 1–27, 2005. Available at SSRN: `http://ssrn.com/abstract=888088`.

[90] F. Karpelevič. On characteristic roots of matrices with nonnegative elements. *Izvestia Akademii Nauk SSSR, Mathematical Series*, 15:361–383 (in Russian), 1951. English Translation appears in Amer. Math. Soc. Trans., Series 2, 140, 79-100, 1988.

[91] Charles S. Kenney and Alan J. Laub. Condition estimates for matrix functions. *SIAM J. Matrix Anal. Appl.*, 10(2):191–209, 1989.

[92] Charles S. Kenney and Alan J. Laub. Padé error estimates for the logarithm of a matrix. *Internat. J. Control*, 50(3):707–730, 1989.

[93] J. F. C. Kingman. The imbedding problem for finite Markov chains. *Z. Wahrsch.*, 1:14–24, 1962.

[94] Steve Kirkland. Note on stochastic pth roots for irreducible nonprimitive stochastic matrices. Private communication, March 25, 2010.

[95] Alexander Kreinin and Marina Sidelnikova. Regularization algorithms for transition matrices. *Algo Research Quarterly*, 4(1/2):23–40, 2001.

[96] Peter Lancaster and Miron Tismenetsky. *The Theory of Matrices*. Academic Press, London, second edition, 1985.

[97] D. Lando and T. M. Skødeberg. Analyzing rating transitions and rating drift with continuous observations. *Journal of Banking & Finance*, 26(2-3):423–444, 2002.

[98] Beata Laszkiewicz and Krystyna Ziętak. A Padé family of iterations for the matrix sector function and the matrix pth root. *Numerical Linear Algebra with Applications*, 16(11-12):951–970, 2009.

[99] David London. Nonnegative matrices with stochastic powers. *Israel J. Math.*, 2:237–244, 1964.

[100] F. Malgouyres. Estimating the probability law of the codelength as a function of the approximation error in image compression. *Comptes Rendus Mathematique*, 344(9):607–610, 2007.

[101] Marvin Marcus and Henryk Minc. Some results on doubly stochastic matrices. *Ameri. Math. Soc.*, 13(4):571–579, 1962.

[102] Servet Martínez, Gérard Michon, and Jaime San Martín. Inverse of strictly ultrametric matrices are of Stieltjes type. *SIAM J. Matrix Anal. Appl.*, 15(1):98–106, 1994.

[103] G. J. McLachlan, T. Krishnan, and Ebooks Corporation. *The EM Algorithm and Extensions*. Wiley Series in Probability and Statistics. Wiley, New York, second edition, 2008.

[104] L. Merkoulovitch. The projection on the standard simplex. 2000. Algorithmics Inc., Working Paper.

[105] D. K. Miller and S. M. Homan. Determining transition probabilities. *Medical Decision Making*, 14(1):52–58, 1994.

[106] Henryk Minc. *Nonnegative Matrices*. Wiley, New York, 1988.

[107] Reinhard Nabben and Richard S. Varga. A linear algebra proof that the inverse of a strictly ultrametric matrix is a strictly diagonally dominant Stieltjes matrix. *SIAM J. Matrix Anal. Appl.*, 15(1):107–113, 1994.

[108] E. A. Nurminski. Projection onto polyhedra in outer representation. *Computational Mathematics and Mathematical Physics*, 48(3):367–375, 2008.

[109] Beresford N. Parlett. A recurrence among the elements of functions of triangular matrices. *Linear Algebra Appl.*, 14(2):117–121, 1976.

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

[111] M. L. Pei. A test matrix for inversion procedures. *Commun. ACM*, 5(10):508, 1962.

[112] Hazel Perfect and L. Mirsky. Spectral properties of doubly-stochastic matrices. *Monatshefte für Mathematik*, 69(1):35–57, 1965.

[113] Panayiotis J. Psarrakos. On the mth roots of a complex matrix. *The Electronic Journal of Linear Algebra*, 9:32–41, 2002.

[114] D. Lando R. A. Jarrow and S. M. Turnbull. A Markov model for the term structure of credit risk spreads. *Rev. Financial Stud.*, 10:481–523, 1997.

[115] J. Th. Runnenberg. On Elfving's problem of imbedding a time-discrete Markov chain in a continuous time one for finitely many states. In *Proceedings, Koninklijke Nederlandse Akademie van Wetenschappen, France*, volume 65 of *series A, Math. Sci.*, pages 536–541, 1962.

[116] Hans Schwerdtfeger. *Les Fonctions de Matrices. I. Les Fonctions Univalentes.* Number 649 in Actualités Scientifiques et Industrielles. Hermann, Paris, France, 1938.

[117] Honsjörg Seybold and Rudolf Hilfer. Numerical algorithm for calculating the generalized Mittag-Leffler function. *SIAM J. Numer. Anal.*, 47(1):69–88, 2008.

[118] S. Shalev-Shwartz and Y. Singer. Efficient learning of label ranking by soft projections onto polyhedra. *The Journal of Machine Learning Research*, 7:1567–1599, 2006.

[119] Burton Singer and Seymour Spilerman. The representation of social processes by Markov models. *Amer. J. Sociology*, 82(1):1–54, 1976.

[120] Matthew I. Smith. A Schur algorithm for computing matrix pth roots. *SIAM J. Matrix Anal. Appl.*, 24(4):971–989, 2003.

[121] F. A. Sonnenberg and J. R. Beck. Markov models in medical decision making. *Medical decision making*, 13(4):322–338, 1993.

[122] George W. Soules. Constructing symmetric nonnegative matrices. *Linear and Multilinear Algebra*, 13:241–251, 1983.

[123] William J. Stewart. *Introduction to the Numerical Solution of Markov Chains.* Princeton University Press, Princeton, NJ, 1994.

[124] W. Stromquist. Roots of transition matrices. Practical paper, Daniel H. Wagner Associates, 1997.

[125] Paul N. Swarztrauber. A direct method for the discrete solution of separable elliptic equations. *SIAM J. Matrix Anal. Appl.*, 11(6):1136–1150, 1974.

[126] J. J. Sylvester. On the equation to the secular inequalities in the planetary theory. *Philosophical Magazine*, 16:267–269, 1883. Reprinted in [127, pp. 110–111].

[127] *The Collected Mathematical Papers of James Joseph Sylvester*, volume IV (1882–1897). Chelsea, New York, 1973.

[128] G. ten Have. Structure of the nth roots of a matrix. *Linear Algebra Appl.*, 187:59–66, 1993.

[129] D. S. Watkins. A case where balancing is harmful. *Electronic Transactions on Numerical Analysis*, 23:1–4, 2006.

[130] Frederick V. Waugh and Martin E. Abel. On fractional powers of a matrix. *J. Amer. Statist. Assoc.*, 62:1018–1021, 1967.

[131] Nicky J. Welton and A. E. Ades. Estimation of Markov chain transition probabilities and rates from fully and partially observed data: Uncertainty propagation, evidence synthesis, and model calibration. *Medical Decision Making*, 25(6):633–644, 2005.